Logitové modely: analýza vlivu exogenních faktorů u kategorizovaných dat

JAN ŘEHÁK*
BLANKA ŘEHÁKOVÁ
Sociologický ústav ČSAV

Logit models: analyzing the influence of exogenous factors in categorical data

Abstract: the primary objective of the article is an introduction to the logit model and its use in methods for modeling the relationships between a dichotomous dependent variable and a set of independent categorical variables. the logit model is viewed here as a special case of the log-linear approach. The deviation and polynomial contrasts are explained as to their interpretational meaning. Examples from migration studies are used to illustrate model-selecting strategies and quantitative outputs.

Sociologický časopis, 1992, Vol. 28 (No. 1: 84-102)

Moderní statistická metodologie podstatně rozšiřuje možnosti modelování vztahů mezi kategorizovanými proměnnými. Samostatný komplex metod tvoří logitové modely, které umožňují řešit úlohy obdobné k regresní analýze resp. analýze rozptylu a to v situaci, v níž vystupuje dvouhodnotová proměnná jako závislá a několik dvou- nebo vícehodnotových nezávislých kategorizovaných proměnných; cílem je odhalit vliv jednotlivých nezávislých proměnných, jejich kategorií i kombinací kategorií na rozdělení závislé proměnné.[1] Tak je možno modelovat kauzální hypotézy o vztazích mezi vysvětlujícími kategorizovanými proměnnými a zvolenou vysvětlovanou proměnnou. Jsou to hypotézy vyjadřující nejen vliv proměnných (tj. celých klasifikací), ale též vlivy jednotlivých kvalit (kategorií) těchto klasifikací a identifikující interakční působení (tj. zesilování resp. zeslabování asociačního vztahu) při kombinování různých kvalit. Zároveň je možné porovnávat vlivy kvalit a jejich kombinací na závislou proměnnou. U vysvětlujících proměnných můžeme

*) Veškerou korespondenci posílejte na adresu RNDr. Blanka Řeháková, CSc., resp. doc. RNDr. Jan Řehák, Sociologický ústav ČSAV, Jilská 1, 110 00 Praha 1. Telefax (02) 235 78 88, E-mail soc@cspgas11.bitnet.

[1]) Takto formulovaný základní problém logitové analýzy lze rozšířit dvěma směry:
- závislá proměnná může mít více hodnot, pak mluvíme o vícehodnotovém logitovém modelu (multinomial logit model, též multinomial response model) a závislých proměnných může být více;
- mezi nezávislými proměnnými mohou být také číselné spojité proměnné.

K modelovým prostředkům logitové analýzy patří také možnost některé kombinace kategorií proměnných z analýzy vypustit (v modelu blokovat).

zavádět do modelu také ordinalitu, tj. respektovat a v analýze využívat uspořádání kategorií nebo kardinalitu pomocí kvantifikací jejich kategorií.[2]

O logitových modelech pojednává celá řada monografií a to na různých úrovních. Jmenujme alespoň Haberman [1978], Knoke a Burke [1980], Aldrich a Nelson [1984].

Postup analytické práce u logitových modelů lze shrnout do několika kroků:

1. Formulace hypotézy pro vybraný seznam proměnných.
   a) Určení závisle proměnné.
   b) Určení nezávislých proměnných a předpokládaných interakčních vlivů.
   c) Určení komparací jednotlivých vlivů na závislou proměnnou.
   d) Určení specifických vlivů skupin kategorií na závisle proměnnou.
2. Testování platnosti formulovaného modelu.
3. Úprava modelu, pokud je třeba.
   a) Studium reziduí, tj. odchylek vstupních dat od očekávaných četností odpovídajících vstupní hypotéze.
   b) Identifikace chybějících složek modelu, resp. vypuštění těch, které se ukázaly jako nevýznamné.
   c) Reformulace hypotézy, tj. přechod k bodu 1.

V tomto postupu je možno cyklicky pokračovat až do vyladění interpretovatelného modelu, který dostatečně dobře odpovídá empirickým datům.[3]

Logit a logitový model

Studium faktorů vysvětlujících vznik rozdělení četností u dichotomických jevů se zjednodušuje tím, že obě četnosti jsou komplementární a tudíž stačí matematicky modelovat vztah k jedné z nich. Logitová metoda, která je současným nejrozšířenějším přístupem k analýze v popsaných situacích vychází ze dvou předpokladů:

a) Místo přímých četností se používá poměr obou komplementárních hodnot rozdělení p/(1-p), který se nazývá šancí. Vyjadřuje zastoupení jedné kategorie vzhledem ke druhé (při prezentování dat tyto údaje vyjadřujeme též jako stonásobky, tj. jako zastoupení jedné kategorie v procentech druhé). Šance je ovšem přirozenější srovnávat spíše poměrově než rozdílově, protože to jsou čísla poměrová.

---

[2]) Takto formulovaná úloha logitové analýzy je obecnější, než klasická úloha použití logitové nebo probitové analýzy, která studuje pravděpodobnost jevu v závislosti na postupně se zvyšujících úrovních číselné nezávislé proměnné, tak jak je běžně aplikovaná v biologických a farmaceutických experimentech.

[3]) Některé počítačové programy umožňují také automatizované vyhledávání modelů a to většinou zpětným procesem tak, že vycházejí z předpokladu kompletních významných vztahů všech vysvětlujících proměnných a všech jejich kombinací k závisle proměnné a pak postupně vypouštějí složité interakce tak dlouho, až je model co nejjednodušší a přitom vykazuje dobrou shodu s daty. Postup má ovšem limity, jak interpretační, tak v možnostech využití komparační složky kauzální hypotézy.

b) Do analýzy vstupují místo šancí jejich přirozené logaritmy ln(p/(1-p)), nazývan logity (log odds).[4]

Přístup má čtyři zpracovatelské výhody:

- Využívá dobře interpretovatelného poměru šancí (odds ratio), který i ve složitýc tabulkách vyšších stupňů umožňuje rychlou vizuální orientaci.
- Přechod k logaritmům dovoluje převedení poměrových srovnání na rozdílové, t umožňuje využít klasické součtové formulace obdobné k analýze rozptylu a k re gresní analýze a tím zjednodušuje tvar modelu i jeho teoretickou a numericko řešitelnost; též interpretace různých výstupních parametrů modelu je tak jedno dušší a odpovídá zvyklostem.
- Logaritmováním se dosahuje důležitá statistická vlastnost - přiblížení se k norma litě.
- Na rozdíl od přímého lineárního modelování vztahů mezi četnostmi se aplikac logitového modelu vyvarujeme nepříjemného jevu, který spočívá v překročen hranic (0,1) modelovými pravděpodobnostmi odvozenými z přijatých vztahovýc rovnic.

Logitem nazýváme hodnotu

$$\text{logit}(p) = \ln(p/(1-p)) = \ln(m/(n-m)), \quad (1)$$

kde p je relativní četnost resp. pravděpodobnost jevu (např. jedna kategorie dicho tomického znaku), m je absolutní četnost výskytu jevu v souboru o rozsahu n. Tat funkce není definovaná pro krajní případy p=0, p=1 a může mít nepříjemné nepřijatelné matematické důsledky pro hodnoty velmi blízké k nule nebo k jednič ce. Proto se v případech, u nichž nastává toto nebezpečí (vícerozměrné tabulk s prázdnými poli, analýza vzácně se vyskytujících jevů) používá upravený vzorec v němž se k čitateli i k jmenovateli přičítají vhodně zvolené konstanty, aby extrémní případy nemohly nastat. V praxi se používá několik typů takových korek cí, nejběžnější je logit(p) = ln((m+0,5)/(n-m+0,5)). Směrodatná odchylka logitu empirických četností se spočte podle vzorce

$$s = 1/(mp(1-p)). \quad (2)$$

Logitové hodnoty se mění pro kombinace kategorií vysvětlujících proměnných Modely předpokládají, že empiricky zjištěná logitová variabilita je systematicky vysvětlitelná vlivem těchto nezávislých proměnných, které mají význam faktorů.[5] Předpokládá se, že hodnota logitu je vytvářena

- celkovou úrovní výskytu jevu v dané populaci (charakteristika populace);
- přímým vlivem příslušné kategorie každé nezávisle proměnné (charakteristika vli vu kategorií);

---

[4]) Podle tradičních zvyklostí je přirozený logaritmus o základu e = 2,71828 značen v tomto článku ln (logaritmus naturalis). V literatuře se používá i značení lg nebo log.

[5]) Jde o pojem faktoru explicitně vyjádřeného obdobně jako u analýzy rozptylu, tj. jde o známou vysvětlující proměnnou, nikoli o pojem faktoru z modelu faktorové analýzy, kde jde o latentní proměnnou definovanou pouze svými důsledky.

- interakčním vlivem, který vzniká kombinováním příslušných kategorií dvojic nezávislých proměnných (specifické působení dvojic kategorií, které není obsaženo v přímém vlivu každé z nich);
- obdobným interakčním působením trojic, čtveřic atd. kategorií.

Formální zápis modelu ukážeme na příkladu dichotomické závislé proměnné M a tří nezávislých proměnných N, P, V, které mají po řadě 2, 2 a 4 hodnoty, přičemž jevem, k němuž se vztahuje pravděpodobnost je první kategorie proměnné M. Tak např. pro hodnoty 2, 1, 3 proměnných N, P, V vypadá zápis takto:

$$\begin{aligned} \text{logit}(p^{M\,NPV}{}_{1.213}) &= v \\ &+ v^{N}{}_{2} + v^{P}{}_{1} + v^{V}{}_{3} \\ &+ v^{NP}{}_{21} + v^{NV}{}_{23} + v^{PV}{}_{13} \\ &+ v^{NPV}{}_{213}. \end{aligned} \qquad (3)$$

Zde v znamená příspěvek, tj. kvantifikovaný vliv (efekt) kategorie nebo kombinace kategorií. Horní index označuje ty proměnné a dolní index ty jejich kategorie, které vliv vytvářejí. Tak např. $v^{PV}{}_{13}$ je speciální vliv kombinace první kategorie proměnné P a třetí kategorie proměnné V. Obdobně pro každou jinou kombinaci kategorií.[6] Číselné hodnoty těchto (v modelu neznámých) parametrů, které odhadneme výpočtem z dat, charakterizují stupeň i směr všech vlivů. Kladná hodnota parametru znamená vliv na zvýšení logitu, tj. i na (p/(1-p)), a tudíž i na p. Záporná hodnota ovlivňuje logit tak, že se snižuje p, tedy se zvyšuje četnost (1-p) v komplementární kategorii. Nulová hodnota značí absenci vlivu.

V rovnici (3) je hodnota logitu určena všemi možnými existujícími vlivy, které jsou v dané situaci myslitelné. Takový model nazýváme saturovaný. Vede k tomu, že jsme schopni plně vysvětlit dané empirické relativní četnosti a ztotožnit je tak s modelovými pravděpodobnostmi. Závěr o úplném vysvětlení dat však většinou odporuje našim představám o faktorovém působení jednotlivých proměnných a jejich kategorií a na druhé straně odráží nepřirozený předpoklad toho, že v datech zachytíme pravděpodobnosti výskytu zcela přesně pomocí empirických relativních četností. Proto reálné analytické modely neobsahují všechny členy rovnice, ale pouze ty, které vyjadřují naši konkrétní hypotézu faktorového působení.

Dalším aspektem modelu je jeho zpracovatelnost, která závisí na tom, kolik nezávislých rovnic a kolik neznámých parametrů máme. V našem případě je neznámých více, a proto musíme provést další opatření, které spočívá v určení dodatečných podmínek pro neznámé parametry, např. ve stanovení počátku pro hodnoty parametrů nebo v nové formulaci modelu, který by obsahoval menší počet smysluplných a interpretovatelných neznámých parametrů (může jít např. o stu-

---

[6]) Můžeme se setkat i se zápisem

$\text{logit}(p^{M\,NPV}{}_{1.213}) = v^{M}{}_{1} + v^{MN}{}_{12} + v^{MP}{}_{11} + v^{MV}{}_{13} + v^{MNP}{}_{121} + v^{MNV}{}_{123} + v^{MPV}{}_{113} + v^{MNPV}{}_{1213}$,

kde je v každém členu zvýrazněno, že závislá proměnná je M a že logitová pravděpodobnost se týká první kategorie proměnné M. Zápis je ale zbytečně komplikovaný.

dium relativních vlivů kategorií vzhledem k jedné pevně předem zvolené kategorii - tzv. prosté kontrasty).[7]

Příklady z problematiky migrací

První z dále uvedených příkladů ilustruje využití základního modelu, druhý je ukázkou zahrnutí specifického typu závislosti (lineární vliv ordinální vysvětlující proměnné).

A. Mimokrajová migrační výměna

Při analýze migrací Severočeského kraje (rok 1987) byly odhaleny tři základní nezávislé dimenze výměnného migračního procesu: výměna věku, vzdělání a důvodů migrace [Řehák et al. 1989]. Při hledání specifických korelátů první dimenze byly studovány výměnné procesy v jednotlivých věkových skupinách [Řeháková, Řehák 1990]. Odtud je převzata následující ilustrace.

Pro skupinu migrantů ve věku 15-24 let byl model určen následujícím způsobem:

závislá proměnná = směr migrace M (do SVČK, ze SVČK),

nezávislé proměnné = pohlaví P (muž, žena), národnost N (česká, jiná), vzdělání V (základní, bez maturity, s maturitou, vysokoškolské).

Nejprve byl formulován jednoduchý logitový model zahrnující jen přímé vlivy vysvětlujících proměnných.

$$\text{logit}(p_{1.ijk}) = v + v^N_i + v^P_j + v^V_k, \qquad (4)$$

kde p znamená pravděpodobnost jevu migrace do SVČK a p/(1-p) má význam koeficientu migračního přílivu. Z rovnice (4) se určí šance jako

$$\begin{aligned} p_{1.ijk}/p_{2.ijk} &= \exp(v + v^N_i + v^P_j + v^V_k) \qquad (4') \\ &= \exp(v)\exp(v^N_i)\exp(v^P_j)\exp(v^V_k) \\ &= w\, w^N_i\, w^P_j\, w^V_k, \end{aligned}$$

což je součinový zápis modelu (4).

Z předchozí části víme, že model bez dalších úprav nelze přímo řešit, a protože se zajímáme o vlivy jednotlivých kategorií N, P, V, zvolíme jednoduchou úpravu spočívající v použití odchylkových kontrastů. Půjde o dodatečný požadavek, aby parametry pro kategorie každé jedné nezávislé proměnné vyjadřovaly odchylku od celkové průměrné úrovně logitů v tabulce (tj. parametru v), tedy aby součet vlivů pro každou proměnnou byl roven nule.[8]

V tabulce 1 je uvedeno třídění čtvrtého stupně mezi zmíněnými proměnnými. V předposledním sloupci je přílivový poměr do/ze pro každou kombinaci kategorií vysvětlujících proměnných, v posledním sloupci je jeho přirozený logaritmus. Je to právě heterogenita čísel v posledním resp. předposledním sloupci, kterou se

[7]) Oba postupy jsou matematicky ekvivalentní, liší se pouze formulací v konkrétních případech.

[8]) Procedura LOGLINEAR v systému SPSS, ve které logitové modely určujeme, předpokládá automaticky splnění uvedené podmínky a příslušné předvolené kontrasty označuje DEVIATION.

snažíme logitovým modelem vysvětlit. Na první pohled je vidět, že kategorie 'jiná národnost' má vyšší poměry než národnost 'česká' prakticky ve všech kombinacích pohlaví a vzdělání. Rovněž muži mají značně vyšší přílivové poměry než ženy, a to jak u 'jiné', tak u 'české národnosti'. To už samo naznačuje význam vlivu kategorií 'muž', 'žena' a kategorií 'česká' a 'jiná'. U kategorií vzdělání však nějaká pravidelnost není zdaleka jasná.

Základní informací, kterou poskytuje model, jsou číselné hodnoty odhadnutých efektů, které určují intenzitu i směr působení jednotlivých složek na studovanou četnost. Tyto efekty jsou smysluplné a interpretovatelné pouze tehdy, když se můžeme spolehnout na platnost modelu. Proto před vlastní interpretací efektů je nutno zhodnotit výsledky testů dobré shody a informaci o vztahu pozorovaných a modelem odhadnutých četností, na jejichž základě lze rozhodnout o validitě modelu.

1. Testy dobré shody

Shodu modelu a empirických dat testujeme obvykle jedním ze dvou chí-kvadrátových testů dobré shody, které pracují na tomto principu:

odhadnou se parametry modelu, pomocí nich se vypočítají modelové (očekávané) četnosti výskytů v polích tabulky (v našem příkladu čtyřrozměrné) a poté se testuje významnost rozdílu mezi těmito modelově očekávanými a skutečně pozorovanými četnostmi. Zpracovatelské programy tisknou obvykle Pearsonovu statistiku dobré shody χ^2 a věrohodnostní statistiku G^2 spolu s počtem stupňů volnosti (df) a dosaženou hladinou významnosti (P). Model zamítáme, když dosažená hladina významnosti je menší než předem určená hodnota. Dalším kritériem shody, vhodným zejména pro porovnání modelů, je poměr G^2/df resp. χ^2/df, u něhož očekáváme hodnotu 1. Čím vyšší hodnoty nabývá, tím horší je model. Hranice tohoto kritéria pro přijetí či odmítnutí modelu není přesně stanovena a vlastní rozhodnutí závisí na subjektivní zkušenosti z analytické práce. Při velikých souborech nastává často situace, v níž nelze přijmout (z důvodů velkého n a vysoké citlivosti testů dobré shody na detekci odchylek) žádný model a všechny uvedené testy jsou významné. Aby bylo možno porovnávat modely i v takových situacích, navrhl A. E. Raftery [1986] bayesovský informační koeficient BIC = G^2 - (df)ln n. Je-li pro daný model M hodnota BIC záporná[9], pak bychom ho měli preferovat před saturovaným. Porovnáváme-li více modelů mezi sebou, pak bychom měli preferovat ten, který vykazuje nejmenší hodnotou koeficientu BIC. Model (4) pro tabulku 1 vykazuje $G^2 = 17{,}10$ ($\chi^2 = 16{,}04$), df = 10, P = 0,072 (P = 0,099), $G^2/df = 1{,}7$, $\chi^2/df = 1{,}6$. Protože situace nevyžaduje BIC, nejsou jeho numerické hodnoty ani uváděny. Dosažená hladina významnosti nepodkročila tradiční hranici 0,05, je však poměrně blízko. To spolu s vysokým poměrem G^2/df (χ^2/df) model

[9]) A. E. Raftery [1985] vyšel z podílu pravděpodobností B = P(M je správný model za daných dat)/P(saturovaný model je správný model za daných dat) a ukázal, že pro velká n přibližně platí -2 ln B = G^2 (df)*ln(n). Tuto hodnotu nazval BIC, kde G^2 je hodnota věrohodnostní statistiky pro daný model, df je jeho počet stupňů volnosti a n je velikost souboru.

zpochybňuje. Ač je tedy model přijatelný na hladině 0,05, budeme mít zřejmě zájem na jeho zlepšení.

2. Odchylky pozorovaných a odhadnutých četností

Z numericky odhadnutých parametrů modelu odhadujeme četnosti nejen pro potřeby testu dobré shody, ale též pro posouzení platnosti odvozeného vztahu pro jednotlivá pole. V tabulce 2 jsou uvedeny modelové odhadnuté četnosti, jejich rozdíly od pozorovaných četností z tabulky 1, adjustovaná rezidua a odhadnutý přílivový poměr, který vychází jako výsledek vyhlazení četností modelem. Prosté rozdíly pozorovaných a modelových četností jsou informativní jen ve výjimečných případech, neboť jsou nesrovnatelné pro různě obsazená pole. Jejich standardizace je provedena ve sloupcích adjustovaných reziduí (upravené odchylky), které jsou přepočteny tak, aby měly přibližně standardizované normální rozdělení. Upravené odchylky mohou být posuzovány jako z-skóry:

a) jednotlivě za významnou odchylku na hladině 0,05 prohlásíme to pole, v němž je adjustované reziduum v absolutní hodnotě větší než 1,96 (v tabulce 2 jsou významné čtyři odchylky);

b) pokud nás zajímá významná struktura některé skupiny polí nebo tabulka celá, použijeme simultánní inferenci obdobnou postupům u znaménkového schématu (viz [Řehák, Řeháková 1986], Dodatek III).

Postup je však nestabilní u málo obsazených polí a v takovém případě se musí používat s velkou opatrností. Odhadnuté přílivové poměry mohou též, alespoň opticky, sloužit pro posouzení vhodnosti modelu, neboť by se neměly příliš lišit od skutečných poměrů (kromě nestabilních polí s malými četnostmi). Porovnáním přílivových poměrů z tabulek 1 a 2 zjistíme větší odlišnosti ve čtyřech případech (dva z nich se týkají málo obsazených polí). To je další důvod pro modifikaci modelu.

Přijmeme-li model, mají očekávané četnosti a všechny jejich funkce (poměry šancí, logity, procenta atp.) také predikční charakter. Model oddělil systematickou (nenáhodnou) a náhodnou (šumovou) složku. Očekávané četnosti reprezentují vyhlazenou složku po odečtení fluktuací. Proto u přijatých modelů prezentujeme míry založené na očekávaných četnostech jako relevantní informaci.

3. Parametry modelových rovnic

Za předpokladu, že model odpovídá datům (ale jen za tohoto předpokladu), přinášejí podstatnou informaci numerické hodnoty, které jsme získali jako odhad neznámých parametrů modelových rovnic. Jsou uvedeny v tabulce 3.[10] Sloupec 'efekt' obsahuje odhadnutou numerickou hodnotu parametru. Pomocí standardní chyby

---

[10]) Vzhledem k podmínkám, kterými byly omezeny parametry (součet do nuly), počítá program jen tzv. nezávislé parametry. Některé programy (SPSS) také pouze tyto parametry tisknou, zbývající musíme dopočítat.

U obecných procedur log-lineárních modelů, v jejichž rámci se logitové modely specifikují jako zvláštní případ (též např. SPSS) se tisknou ve výstupech poloviční hodnoty koeficientů i jejich standardních chyb, z-skóry vycházejí správně.

(třetí sloupec) lze tento koeficient převést na z-skór (čtvrtý sloupec). Koeficient odpovídá síle vlivu kategorie (nebo ve složitějších případech kombinace kategorií) na logit. Z-skór má, obdobně jako adjustovaná rezidua, vlastnost přibližného standardního normálního rozdělení, která umožňuje statistickou inferenci o vlivu kategorií či jejich kombinací. Můžeme jej použít následujícími způsoby:

- Standardním normálním testem rozhodneme o významnosti jednotlivých předem určených efektů (tj. o každém zvlášť bez ohledu na ostatní).
- V případech, kdy má proměnná více kategorií než dvě, tj. kdy závěr o jedné kategorii není zároveň závěrem o kategorii komplementární, nebo u interakčních efektů, do nichž vstupují více než dvouhodnotové proměnné, můžeme pomocí simultánní inference určit strukturu vlivů v rámci zvolené proměnné nebo kombinace proměnných (odhalení dílčích struktur vlivů).
- Posouzením všech hodnot současně (tj. celého sloupce z-skórů kromě komplementárních hodnot u dichotomií) můžeme simultánní inferencí určit celkovou statisticky prokazatelnou strukturu vlivů na závislou proměnnou.
- Ty proměnné či jejich kombinace, u kterých se neprokázal významný efekt (podle zvoleného postupu), mohou být případně z modelu vypuštěny a tím model zjednodušen (efekty musí být však znovu přepočítány).

4. Závěr a hodnocení modelu, případné modifikace

Charakteristiky vhodnosti modelu (4) ukazují, že předpokládané rovnice odrážejí vlivy vysvětlujících proměnných nedostatečně a měly by být modifikovány. Studium reziduí a přílivových poměrů naznačuje, že dalším členem by měl být interakční vliv národnosti a vzdělání, a proto upravíme model takto:

$$\text{logit}(p_{1.ijk}) = v + v^{N}_{i} + v^{P}_{j} + v^{V}_{k} + v^{NV}_{ik}. \qquad (5)$$

Výsledky modelu jsou shrnuty v tabulkách 4 a 5. Jak testové statistiky, tak adjustovaná rezidua vykazují velmi dobrou shodu modelu a dat. Nově vstupující parametry zdůrazňují interakci národnosti a vzdělání v tom, že osoby nečeské národnosti s vysokoškolským vzděláním významně zvyšují přílivový poměr, obdobně osoby české národnosti s nižším vzděláním. Zatímco v modelu (4) se vliv vzdělání na příliv projevoval pouze u nižších kategorií a neukazoval na přímou či nepřímou úměru se stupněm vzdělání, projevuje se v modelu (5) vliv vzdělání na přílivový poměr v přímé úměře ovšem s tím, že existují ještě dodatečné odchylkové efekty specificky ovlivňující příliv pro českou národnost a jinou národnost. Tyto interakční vlivy se v modelu (4) zprůměrovaly, a tak způsobily velká rezidua. Jelikož proměnná "pohlaví" nevstupuje do interakcí ani se vzděláním ani s národností, jsou odhadnuté poměry šancí $(m_{i1k1}/m_{i1k2}):(m_{i2k1}/m_{i2k2})$ konstantní pro všechna i=1, 2 a k=1, 2, 3, 4 (indexy zleva doprava odpovídají postupně národnosti, pohlaví, vzdělání, směru migrace; písmenko m znamená odhadnutou očekávanou četnost při platnosti modelu, která je vyhlazením empirických absolutních četností). Konstanta se rovná $\exp(v_1^P - v_2^P) = \exp(0{,}204) = 1{,}2$. Přílivový poměr pro muže je 1,2 krát větší než pro ženy a to bez ohledu na národnost a vzdělání. V obrázku 1 je navržen způsob grafického znázornění významnosti a směru jednotlivých efektů z modelu (5).

B. Záměr opustit Severočeský kraj

Ve výzkumu Vztah ke kraji (L. Michalová a kolektiv, SEÚ ČSAV, 1989) byla položena otázka na záměr vystěhovat se z kraje. V tabulce 6 je zachyceno třídění třetího stupně, které vyjadřuje vztah věku A (1 = do 40 let, 2 = 41 a více let) a vzdělání E (1 = bez maturity, 2 = s maturitou, 3 = vysokoškolské) k uvedenému záměru M (1 = nebudu se stěhovat z kraje, 2 = budu se stěhovat z kraje). Jednoduchý model

$$\text{logit}(p_{2.ij}) = v + v^{A}_{i} + v^{E}_{j}, \qquad (6)$$

zahrnující obě vysvětlující proměnné vyhovuje datům velmi dobře (G^2 = 0,213, df = 2, P = 0,899, G^2/df = 0,11, maximální adjustované reziduum = ± 0,46). Jednotlivé efekty byly zjištěny jako v = -1,426 (z = -12,546), v^{A}_{1} = 0,306 (z = 3,300), v^{A}_{2} = -0,306 (z = -3,300), v^{E}_{1} = -0,346 (z = -2,602), v^{E}_{2} = -0,034 (z = -0,245), v^{E}_{3} = 0,380 (z = 1,980). Vycházíme-li z modelu (6), můžeme si položit dvě otázky, které ovšem interpretačně divergují:

1. Vzhledem k tomu, že z-skóry u vzdělání vykazují nižší významnost než u věku, můžeme se ptát, zda vliv věku není pro výklad variability záměru stěhování z kraje postačující. Pro zodpovězení této otázky zkusíme model

$$\text{logit}(p_{2.ij}) = v + v^{A}_{i}. \qquad (7)$$

Výpočet ukazuje značný pokles u dosažené hladiny významnosti (G^2 = 6,818, df = 4, P = 0,146), který však ještě indikuje možnou shodu modelu a dat. Velký skok u poměru G^2/df z 0,11 na 1,70 však již varuje před vypuštěním vlivu vzdělání, i když se všechna adjustovaná rezidua pohybují v hranicích statistické nevýznamnosti (jsou mezi hodnotami -1,71 a 1,71) při $\alpha = 0{,}05$.

2. Data a odhadnuté parametry v modelu (6) indikují ordinální vztah vzdělání a záměru odejít pro první i druhou věkovou kategorii. Zahrnutí ordinality do modelu můžeme provést několika způsoby. Nejjednodušším přístupem je předpoklad lineárního působení. Přestože do úlohy vstupujeme s kódováním kategorií vzdělání 1, 2, 3, pro výpočty se provede normalizační transformace na skóry t_j (j = 1,2,3) tak, aby jejich součet byl nula a součet jejich čtverců byl jedna (v našem příkladu vzniknou skóry -0,7071, 0, 0,7071). Model, který zahrnuje lineární vliv vzdělání má tvar

$$\text{logit}(p_{2.ij}) = v + v^{A}_{i} + \beta t_j. \qquad (8)$$

Tento model má ještě lepší ukazatele dobré shody než model (6) zahrnující nominalitu vzdělanostních kategorií; výsledky jsou v tabulce 7.[11] Všechny parametry jsou vysoce významné a tudíž se účastní výkladového schématu pro variabilitu četností u záměru odejít z kraje. Model (8) je oproti (6) jednodušší, místo dvou nezávislých parametrů u proměnné vzdělání obsahuje pouze jeden (β), který ukazuje přírůstek při přechodu mezi první a druhou, resp. druhou a třetí kategorií vzdělání a to pro kvantifikaci t_j (j = 1,2,3) a logaritmickou škálu poměru šancí. Abychom přešli k vlastnímu přírůstku podél původní škály, je nutné násobit β délkou inter-

[11]) V proceduře LOGLINEAR se člen β t_j modelu (8) zadává pomocí kontrastu POLYNOMIAL pro vzdělání E a pomocí specifikace u příslušného členu v modelu, ve kterém je v závorce jednička pro označení mnohočlenu prvního stupně (M BY E(1)).

valu, tj. hodnotou 0,7071; výsledek je 0,346. Abychom převedli efekt na podílovou změnu u původních poměrů šancí, přejdeme k exponenciále exp 0,346 = 1,41; při přechodu mezi sousedními kategoriemi vzdělání se tedy poměr změní 1,41 krát. Model (8) v sobě nese ještě další důležitou informaci. Vzhledem k velmi vysoké míře shody modelu a dat je zřejmé, že jeho další zlepšování už nemá věcný smysl. Rovnice (8) vyjadřuje nejen, že vliv vzdělání je lineární, ale také, že přírůstky jsou stejné pro obě věkové skupiny. Jako další zesložitění modelu bychom totiž mohli zahrnout možnost dvou různých lineárních trendů v různých věkových skupinách.

Z modelů (6), (7), (8) je poslední nejlepší. Plyne z něj závěr, že

a) na záměr odejít z kraje působí silně věk i vzdělání a to nezávisle na sobě,

b) vliv vzdělání je lineární a působí 141 % vzrůst poměru šancí mezi sousedními kategoriemi vzdělání, čím vyšší vzdělání, tím vyšší zastoupení těch, kteří chtějí z kraje odejít,

c) nižší věková kategorie vykazuje významně vyšší záměr odejít než kategorie vyšší a to s 1,84 (= $\exp(v^{A}_{1} - v^{A}_{2})$) násobkem poměru šancí.

Závěr

Logitové modely jsou obdobou analýzy rozptylu a také analýzy kovariance pro dichotomické závislé proměnné. Vysvětlují variabilitu procenta výskytu jevu pro různé hodnoty a kombinace hodnot nezávislých kategorizovaných i číselných proměnných. V článku byly na příkladech ilustrovány a vysvětleny dva prakticky nejčastěji se vyskytující typy modelů, u nichž jsou vysvětlujícími faktory kategorizované proměnné. Další varianty aplikace zahrnují číselné faktory (obdoba regresního členu v analýze kovariance) a složitější možnosti využití různých typů kontrastů, blokování polí a různých vah pro pole tabulky. Každá z těchto možností znamená však zkomplikování modelu a zasluhuje proto samostatný výklad. Důležitým analytickým prostředkem pro rozšíření modelových možností je opakování analýzy a odlaďování modelů pro různé podsoubory a jejich komparace. Tím je vnesen do interpretace nový závažný kvalitativní prvek, výpověď o různém fungování příčin v různých kontextech.

Logitové modely mají také některé nedostatky, kterých si musíme být vědomi a které limitují využitelnost formálních matematických výsledků.

a) Metoda vychází z asymptotické teorie, která vyžaduje větší počet pozorování a to ve všech polích tabulky. Je velmi citlivá na nulová a málo obsazená pole, zvláště je-li jich větší počet. Jak adjustovaná rezidua, tak odhady parametrů příslušných k málo obsazeným polím jsou nestabilní a nelze se na jejich hodnoty spoléhat.

b) Metoda vyžaduje dodržení určitých pravidel (ortogonalita vlivů, normalizace škál, dimenzionalita), což znamená, že vstupní model je pro výpočty upraven. Ne u všech programů jsou spočteny všechny parametry v původních významech (např. je třeba dopočítat závislé parametry a jejich směrodatné odchylky, přepočítat efekty lineárních kontrastů na původní škálu ap.). Poslední uvedený aspekt patří však spíše na vrub autorů programů.

c) K jistým nevýhodám patří i relativní obtížnost formalizování výzkumných hypotéz, která roste velmi rychle při zahrnutí dalších aspektů (blokování polí, větší počet kontrastů). Též zavedení interakcí vyšších řádů do modelu vede obvykle k interpretačním nejasnostem. Ani zadávání modelů v počítačových programech nebývá jednoduché a v případě komplexnějších modelů si vyžaduje speciální školení.

Z těchto důvodů je nejlépe pronikat do logitových modelů od nejjednodušších (avšak výzkumně zajímavých) výkladových schémat a postupně obohacovat používanou metodiku o další možnosti. A to bylo také účelem této stati.

Příloha: Zadání příkladů v programu SPSSX

Významy symbolů v zadání jsou uvedeny v textu.

Mimokrajová migrační výměna

```
LOGLINEAR M (1,2) BY N P (1,2) V (1,4)
/ * MODEL (4):
/DESIGN = M M BY N M BY P M BY V
/ * MODEL (5):
/DESIGN = M M BY N M BY P M BY V M BY N BY V
```

Záměr opustit Severočeský kraj

```
RECODE M (1=2) (2=1) INTO MIG
/* překódování zajišťující požadavek programu, aby "1" odpovídala analyzovanému
  jevu s pravděpodobností v čitateli logitu
VARIABLE LABELS MIG 'MIGRACE Z KRAJE'
VALUE LABELS MIG 1 'ANO' 2 'NE'
LOGLINEAR MIG (1,2) BY A (1,2) E (1,3)
/ * MODEL (6):
/DESIGN = MIG MIG BY A MIG BY E
/ * MODEL (7):
/DESIGN = MIG MIG BY A
/ * MODEL (8):
/CONTRAST (E) = POLYNOMIAL
/DESIGN = MIG MIG BY A MIG BY E(1)
```

Tabulková příloha

Tabulka 1. Mimokrajová migrační výměna Severočeského kraje v roce 1987. Četnosti migračních přílivů a odlivů v kombinacích vzdělání národnosti a pohlaví, přílivové poměry a logity.

národnost	pohlaví	vzdělání	do SVČK	ze SVČK	do/z	ln(do/z)
česká	muž	zákl.	144	134	1.07	0.07197
		bez mat.	445	301	1.48	0.39096
		s mat.	265	209	1.27	0.23739
		vys.	94	72	1.31	0.26663
	žena	zákl.	196	196	1.00	0.00000
		bez mat.	410	323	1.27	0.23850
		s mat.	397	418	0.95	-0.05154
		vys.	91	80	1.14	0.12883
jiná	muž	zákl.	94	47	2.00	0.69315
		bez mat.	69	23	3.00	1.09861
		s mat.	52	15	3.47	1.24319
		vys.	32	3	10.67	2.36712
	žena	zákl.	79	44	1.79	0.58526
		bez mat.	48	33	1.45	0.37469
		s mat.	50	21	2.38	0.86750
		vys.	9	2	4.50	1.50408

Tabulka 3. Mimokrajová migrace Severočeského kraje v roce 1987. Parametry modelu (4) a (4'); jejich významnost.

parametr		efekt	std.chyba	z-skór	znam.sch.	součinový efekt
	v	0.523	0.050	10.372	+++	1.687
nár.	v_1	-0.349	0.048	-7.283	---	0.705
	v_2	0.349	0.048	7.283	+++	1.418
pohl.	v_1	0.106	0.031	3.400	+++	1.112
	v_2	-0.106	0.031	-3.400	---	0.899
vzděl.	v_1	-0.154	0.060	-2.552	- -	0.857
	v_2	0.115	0.051	2.264	+	1.112
	v_3	-0.070	0.053	-1.335	0	0.932
	v_4	0.110	0.082	.1.317	0	1.116

Poznámka: Jedno, dvě resp. tři znaménka značí významnost efektu na hladině 0.1 , 0.05 resp. 0.01 simultánně v blocích.

Tabulka 2. Mimokrajová migrace Severočeského kraje v roce 1987.
Odhadnuté četnosti modelu (4), absolutní a adjustované odchylky,
a odhady migračních poměrů

národnost	pohlaví	vzdělání	odhadnuté četnosti		pozorov. - odhad.		adjustovaná rezidua		odhad poměru
			do SVČK	ze SVČK	do SVČK	ze SVČK	do SVČK	ze SVČK	
česká	muž	zákl.	147.70	130.30	-3.70	3.70	-0.59	0.59	1.13
		bez mat.	445.76	300.24	-0.76	0.76	-0.09	0.09	1.48
		s mat.	261.76	212.24	3.24	-3.24	0.43	-0.43	1.23
		vys.	98.97	67.03	-4.96	4.96	-1.09	1.09	1.48
	žena	zákl.	187.56	204.44	8.44	-8.44	1.29	-1.29	0.92
		bez mat.	400.05	332.95	9.95	-9.95	1.23	-1.23	1.20
		s mat.	407.12	407.88	-10.12	10.12	-1.29	1.29	1.00
		vys.	93.09	77.91	-2.09	2.09	-0.46	0.46	1.19
jiná	muž	zákl.	97.99	43.01	-3.99	3.99	-0.88	0.88	2.28
		bez mat.	68.91	23.09	0.09	-0.09	0.02	-0.02	2.98
		s mat.	47.74	19.26	4.26	-4.26	1.25	-1.25	2.48
		vys.	26.18	8.82	5.82	-5.82	2.42	-2.42	2.97
	žena	zákl.	79.75	43.25	-0.75	0.75	-0.17	0.17	1.84
		bez mat.	57.28	23.72	-9.28	9.28	-2.52	2.52	2.41
		s mat.	47.38	23.62	2.62	-2.62	0.73	-0.73	2.00
		vys.	7.77	3.23	1.23	-1.23	0.84	-0.84	2.40

Tabulka 4. Mimokrajová migrační výměna Severočeského kraje v roce 1987. Výsledky modelu (5): očekávané četnosti, adjustovaná rezidua a odhady přílivových poměrů.

národnost	pohlaví	vzdělání	odhad četností		adjust. rezidua		odhad poměru do/z
			do SVČK	ze SVČK	do SVČK	ze SVČK	
česká	muž	zákl.	149.38	128.62	-0.92	0.92	1.16
		bez mat.	449.68	296.32	-0.61	0.61	1.52
		s mat.	258.72	215.28	0.86	-0.86	1.20
		vys.	95.39	70.61	-0.32	0.32	1.35
	žena	zákl.	190.62	201.38	0.92	-0.92	0.95
		bez mat.	405.32	327.68	0.61	-0.61	1.24
		s mat.	403.28	411.72	-0.86	0.86	0.98
		vys.	89.61	81.39	0.32	-0.32	1.10
jiná	muž	zákl.	95.44	45.56	-0.38	0.38	2.09
		bez mat.	64.15	27.85	1.61	-1.61	2.30
		s mat.	50.88	16.12	0.44	-0.44	3.16
		vys.	31.37	3.63	0.68	-0.68	8.64
	žena	zákl.	77.56	45.44	0.38	-0.38	1.71
		bez mat.	52.85	28.15	-1.61	1.61	1.88
		s mat.	51.12	19.88	-0.44	0.44	2.57
		vys.	9.63	1.37	-0.68	0.68	7.03

$G^2 = 4.84$, df = 7 , P = 0.679 , $G^2/df = 0.69$
$X^2 = 4.86$, df = 7 , P = 0.677 , $X^2/df = 0.69$

Tabulka 5. Mimokrajová migrační výměna Severočeského kraje v roce 1987. Parametry modelu (5).

parametr		efekt	std.chyba	z-skór	znam.sch.	součinový efekt
	v	0.639	0.072	8.907	+++	1.894
nár.	v_1	-0.478	0.072	-6.661	---	0.620
pohl.	v_1	0.102	0.031	3.282	+++	1.107
vzděl.	v_1	-0.297	0.089	-3.318	---	0.743
	v_2	-0.116	0.094	-1.231	0	0.890
	v_3	-0.075	0.101	-0.738	0	0.928
	v_4	0.488	0.186	2.613	+ +	1.629
nár. * vzděl.	v_1	0.184	0.089	2.053	+	1.202
	v_2	0.270	0.094	2.874	+ +	1.310
	v_3	-0.004	0.101	-0.042	0	0.996
	v_4	-0.450	0.186	-2.410	- -	0.638

Poznámka: 1. Jedno, dvě resp. tři znaménka značí významnost efektu na hladině 0.1 , 0.05 resp. 0.01 simultánně v blocích.
2. Komplementární parametry nejsou uvedeny.

Tabulka 7. Záměr opustit Severočeský kraj. Parametry modelu (8).

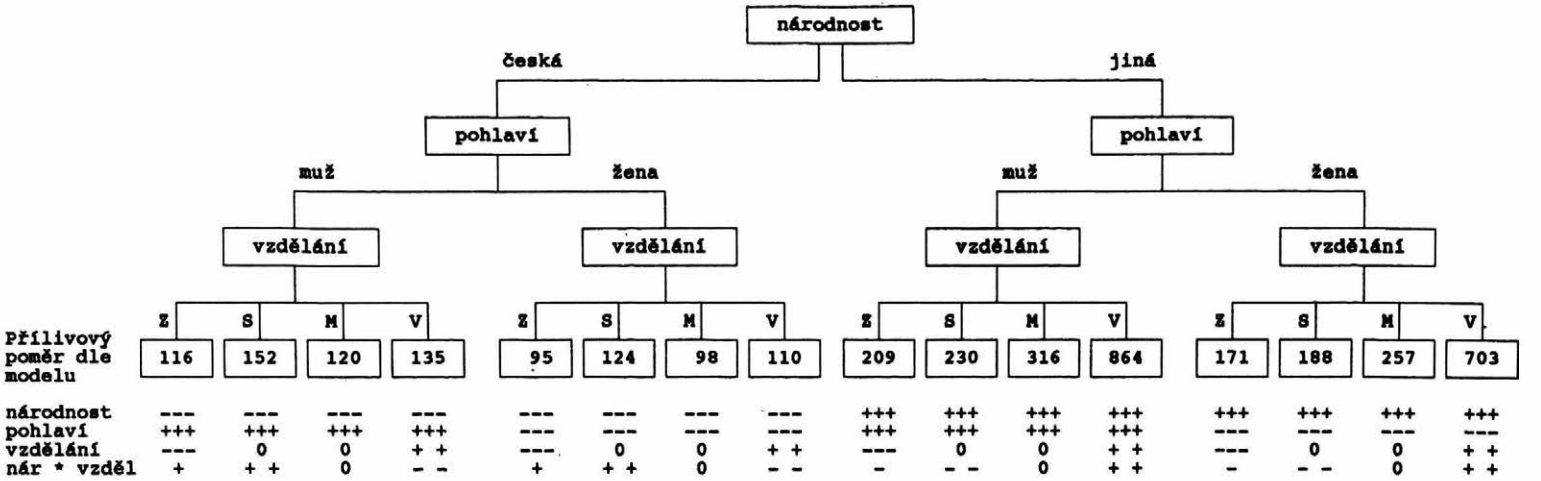
parametr		efekt	std.chyba	z-skór	znam.sch.	součinový efekt
	v	-1.436	0.107	-13.419	---	0.238
věk	v_1	0.306	0.093	3.298	+++	1.358
	v_2	-0.306	0.093	-3.298	---	0.736
vzděl.	β	0.489	0.189	2.588	+++	1.631
	βt_1	-0.346	0.134	-2.588	---	0.708
	βt_2	0	/	/	/	1
	βt_3	0.346	0.134	2.588	+++	1.143

Tabulka 6. Záměr opustit Severočeský kraj.
Výsledky modelu (8).

věk	vzdělání	stěhování		odhad stěhování		adjust. rezidua		poměr ano/ne	odhad poměru ano/ne
		ne	ano	ne	ano	ne	ano		
do 40	bez mat.	197	47	198.61	45.39	-0.52	0.52	0.23	0.23
	s mat.	131	40	129.25	41.75	0.43	-0.43	0.30	0.32
	vys.	28	13	28.15	12.85	-0.06	0.06	0.46	0.46
41+	bez mat.	227	27	225.98	28.02	0.36	-0.36	0.12	0.12
	s mat.	116	21	116.57	20.43	-0.18	0.18	0.18	0.17
	vys.	22	6	22.44	5.56	-0.24	0.24	0.27	0.25

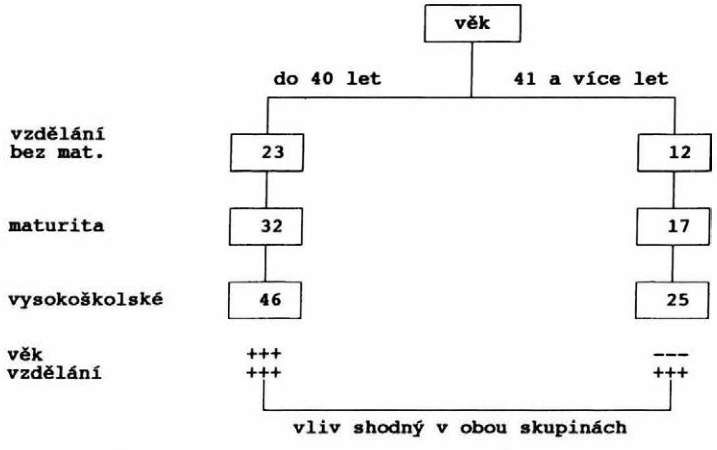
$G^2 = 0.274$, df = 3 , P = 0.965 , $G^2/df = 0.09$
$X^2 = 0.274$, df = 3 , P = 0.965 , $X^2/df = 0.09$

Obrázek 1. Efekty vzdělání, pohlaví a národnosti na výměnný migrační proces Severočeského kraje v roce 1987
(výkladové schéma logitového modelu pro přílivový poměr zahrnuje přímé vlivy a interakci vzdělání a národnosti)



Poznámky: a/ znaménka označují intenzitu statistické významnosti odvozenou Holmovou metodou simultánní inference pro každý blok parametrů (tj. každý řádek ve schematu)
b/ jedno, dvě resp. tři znaménka značí významnost efektu postupně na hladinách 0.1, 0.05 resp. 0.01; nula značí nevýznamnost efektu (tj. absenci vlivu)
c/ uvedené přílivové poměry jsou odhady poskytované modelem (tj. plynoucí z předpokladu platnosti ověřených vlivů)
d/ $X^2 = 4.84$, df = 7, P = 0.679, $X^2/df = 0.69$, max.adj.rez. = 1.61
e/ kategorie vzdělání: Z = základní, S = střední bez maturity, M = maturita, V = vysokoškolské

Obrázek 2. Efekty věkových skupin a lineární působení stupně vzdělání v logitovém modelu vysvětlujícím záměr stěhovat se ze Severočeského kraje (schema obsahuje stonásobky poměrů ano/ne)



Poznámka: a/ znaménka vyjadřují významnost na hladině 0.01
b/ vliv vzdělání se ve skupinách neliší - v sousední vyšší kategorii vzdělání je podíl (ano/ne) vždy 1.413-krát vyšší
c/ uvedené poměry jsou výsledkem modelových rovnic, tudíž vyjadřují hodnoty po odečtu náhodných složek

Literatura

Aldrich, J. H., F. D. Nelson, 1984. *Linear Probability, Logit, and Probit Models.* Newbury Park California: Sage Publications, Inc.

Haberman, S. J. 1978. *Analysis of Qualitative Data. (Vol 1).* New York: Academic Press.

Knoke, D., P. J. Burke 1980. *Log-Linear Models.* Newbury Park, California: SAGE Publications, Inc.

Raftery, A. E. 1985. "A note on Bayes factor for log-linear contingency table models with vague prior information." *Journal of the Royal Statistical Society*, Series B 47.

Raftery, A. E. 1986. "Choosing models for cross-classifications." *American Sociological Review* 51: 145-146.

Řehák, J. et al. 1989. *Migrační toky Severočeského kraje. Mimokrajová výměna.* Ústí n. L.: SEÚ ČSAV.

Řeháková, B., J. Řehák 1990. *Výměna věku v mezikrajových migracích ekonomicky aktivních v Severočeském kraji v r. 1987.* Ústí n. L.: SEÚ ČSAV.

SPSS-X User's Guide, 3rd ed. 1989. Chicago: SPSS Inc.

Summary

The logit models serve to analyze the relationships of dependent categorical variable on categorical and numerical explanatory variables. The paper centers on the dichotomous dependent variable and the model is treated as a special case of the log-linear approach. The model formulation, its solution, and the meaning of output information (goodness-of-fit testing, residuals, parameters) are illustrated on two examples of migration data. The results are interpreted as well. The deviation and the polynomial contrasts are explained and applied. The model-selecting strategies are based on tests and measures of goodness-of-fit and on adjusted residuals. The appendix contains the command sets for the SPSS runs which were used in the analysis.