Nebojte se logistické regrese*

BLANKA ŘEHÁKOVÁ**
Sociologický ústav AV ČR, Praha

Introducing Logistic Regression

Abstract: The aim of this article is to acquaint Czech sociologists who are interested in quantitative methods of data analysis with logistic regression. The logistic regression model has been in use in statistical analyses for many years. It is the standard method for regression analysis of categorical data in many fields. Unfortunately, in the Czech Republic the method is less known, and consequently rarely used. The content of the article is divided into several parts according to the type of the explained variable. The article begins with the simplest case, in which the explained variable is dichotomous. It continues with the case, in which the explained variable is polytomous and nominal, or is treated as being nominal. The article ends with the case, in which the explained variable is ordinal. Special attention is devoted to the manner of treating categorical explaining variables, to the assessment of the adequacy of the fitted model, and to the interpretation of the coefficients of the logistic regression model. All these steps are demonstrated on real data set through the use of the SPSS and GoldMineR software packages.
Sociologický časopis, 2000, Vol. 36 (No. 4: 475-492)

Úvod

Regresní metody patří mezi nejčastěji využívané přístupy k analýze dat nejrůznější povahy. Při vyslovení slova regrese se nám nejčastěji vybaví regrese lineární, méně často nelineární nebo logistická, i když právě logistická regrese je už nejméně dvě desetiletí standardní metodou v západoevropské a americké vědě včetně společenské. Cílem analýzy, která využívá metodu regrese, je nalézt co nejlepší, nejúspornější a současně věcně smysluplný model, který popíše vztah mezi závislou (vysvětlovanou, predikovanou) proměnnou a skupinou nezávislých (vysvětlujících, predikujících) proměnných. Je-li vysvětlovaná proměnná spojitá, obracíme se k regresi lineární, není-li spojitá, pak ke slovu přichází regrese logistická.

Hned na počátku je třeba zdůraznit, že metoda logistické regrese není omezená jen na případ, kdy vysvětlovaná proměnná je binární (dichotomická, alternativní), to znamená, že může nabývat jen dvou hodnot. Pravda je, že pro tuto situaci byla logistická regrese původně vyvinuta a že je interpretačně, ale i jinak nejsnazší. Již dlouho však existují metody a také programy, které si poradí s případy, kdy kategorizovaná závislá proměnná není binární, a dokonce dokáží i respektovat požadavek, aby ji považovaly za ordinální. Pokud si situace žádá rozlišení, používá se termínu binární logistická regrese či logistická

*) Tato práce vznikla v rámci grantu „Mezinárodní program sociálního výzkumu“, který vede v Sociologickém ústavu AV ČR Klára Vlachová. Grant byl udělen Grantovou agenturou České republiky a má číslo 403/99/1129.

**) Veškerou korespondenci posílejte na adresu: RNDr. Blanka Řeháková, CSc., Sociologický ústav AV ČR, Jilská 1, 110 00 Praha 1, e-mail rehakova@soc.cas.cz

regrese jen pro případ binární závislé proměnné. Jinak se hovoří o polytomické nebo multinomické logistické regresi, případně o ordinální (logistické) regresi, ale není to dogma. Vysvětlující proměnné mohou být ve všech případech jak kategorizované (faktory, kategorizované kovariáty), tak spojité (kovariáty). To je stejné jako u lineární regrese včetně skutečnosti, že faktory vyžadují speciální zacházení.

Cílem tohoto článku je seznámit čtenáře se základními pojmy logistických regresních modelů, s modelem jako takovým, s posuzováním vhodnosti modelu, s interpretací regresních koeficientů, s různými možnostmi zacházení s faktory a s nejrůznějšími záludnostmi, které mohou v aplikacích nastat. Začneme s nejjednodušším případem binární logistické regrese, pak přejdeme k případu, kdy vysvětlovaná proměnná nabývá tří hodnot a nejprve ji budeme nazírat jako nominální a později jako ordinální. Vše budeme ilustrovat na reálných datech z výzkumu *Sociální nerovnosti a spravedlnost* (ISSP 1999). K výpočtům využijeme programů LOGISTIC REGRESSION pro dichotomickou závislou proměnnou, NOMREG pro nominální závislou proměnnou s více než dvěma kategoriemi a PLUM pro ordinální závislou proměnnou ze SPSS 9.0 nebo 10.0 [viz *SPSS Regression…* 1999, *SPSS Advanced…* 1999].[1] Pro predikci ordinální závislé proměnné použijeme také programu GOLDMineR 2.0 [viz Magidson 1998].[2] Použitá metodologie vychází z Magidsonových prací, zejména však z výsledků Leo Goodmana, zatímco metodologie programu PLUM je založena na pracích Mc Cullagha.

1. Binární závislá proměnná

1. 1 Pravděpodobnost, šance, logit

Předpokládejme, že binární závislá proměnná Y nabývá hodnot 0 a 1.[3] Nechť $Y = 1$, jestliže u sledovaného případu (jednotky, respondenta) nastal jev J a $Y = 0$, jestliže jev J nenastal, tj. jestliže nastal jev non J. Zajímá nás, zda lze klasifikovat případy do dvou kategorií závislé proměnné na základě skupiny nezávislých proměnných. Místo toho, abychom se snažili predikovat libovolně zvolené hodnoty sloužící k označení dvou kategorií binární proměnné, v našem případě 0 a 1, zaměříme se na problém predikce pravděpodobnosti, že případ patří do jedné kategorie závislé proměnné. Známe-li totiž $P(Y = 1)$, známe i $P(Y = 0)$, protože $P(Y = 0) = 1 – P(Y = 1)$.

Mohli bychom zkusit modelovat pravděpodobnost, že $Y = 1$ jako

$$P(Y = 1) = \alpha + \beta_1 X_1 + \ldots + \beta_K X_K, \qquad (1.\ 1)$$

ale při řešení této regresní rovnice bychom narazili na numerické problémy, protože pravděpodobnost jevu je číslo, které leží mezi nulou a jedničkou, a rovnicí predikované hodnoty by tuto podmínku nemusely splňovat. Prvním krokem k odstranění tohoto nedostatku je záměna pravděpodobnosti jevu šancí jevu. Šance, že nastal jev J, tj. šance, že $Y = 1$, psáno šance(J) nebo šance($Y = 1$), je definovaná jako podíl pravděpodobnosti, že $Y = 1$ a pravděpodobnosti, že $Y \neq 1$, tedy

$$\text{šance}(Y = 1) = P(Y = 1) / [1 – P(Y = 1)], \qquad (1.\ 2)$$

můžeme to také zapsat jako

[1]) První dvě procedury jsou i v nižších verzích SPSS, PLUM je novinka v SPSS 10.0.

[2]) GOLDMineR čte systémové soubory SPSS.

[3]) Numerické hodnoty binární proměnné jsou arbitrární, je to věc dohody, nemají žádný skutečný význam.

$$\text{šance}(J) = \text{P}(J) / \text{P}(\text{non } J). \quad (1.\ 3)$$

Šance nemá žádnou pevnou maximální hodnotu, ale její minimální hodnota je nula. Provedeme tedy ještě jednu transformaci, a sice přirozený logaritmus šance. Tato proměnná se nazývá logit a je definovaná pomocí vztahu

$$\text{logit}(Y) = \ln \{\text{P}(Y = 1) / [1 \ (\ \text{P}(Y = 1)]\}. \quad (1.\ 4)$$

Hodnoty logitu se pohybují od minus nekonečna do plus nekonečna. Použijeme-li tedy logit(Y) jako závislou proměnnou, zbavíme se problémů, které jsme měli v případě pravděpodobnosti a šance. Regresní rovnice bude mít tvar

$$\text{logit}(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_K X_K. \quad (1.\ 5)$$

Logit můžeme převést zpět na šanci pomocí exponenciální funkce jako

$$\text{šance}(Y = 1) = \exp[\text{logit}(Y)] = \exp(\alpha + \beta_1 X_1 + \ldots + \beta_K X_K) = \quad (1.\ 6)$$

$$\exp(\alpha) \times \exp(\beta_1 X_1) \times \ldots \times \exp(\beta_K X_K).$$

Od šance se dostaneme zpět k pravděpodobnosti pomocí formule

$$\text{P}(Y = 1) = \text{šance}(Y = 1) / [1 + \text{šance}(Y = 1)] = \quad (1.\ 7)$$

$$\exp(\alpha + \beta_1 X_1 + \ldots + \beta_K X_K) / [1 + \exp(\alpha + \beta_1 X_1 + \ldots + \beta_K X_K)].$$

Pravděpodobnost, šance a logit jsou tedy tři různé způsoby vyjádření téhož v tom smyslu, že jsou na sebe vzájemně převoditelné. Pro interpretaci jsou snadněji pochopitelné, a proto vhodnější pravděpodobnosti a šance než logity. Na tomto místě je vhodné upozornit na nešvar, s kterým se můžeme čas od času v pracích, které používají logistickou regresi k analýze dat, setkat. Autoři interpretují hodnoty šancí, ale dávají jim v textu význam pravděpodobností. Z výše uvedených definic a vztahů by mělo být každému naprosto jasné, že šance nejsou pravděpodobnosti.

1. 2 Kategorizované nezávislé proměnné

Jak již bylo řečeno, nezávislé proměnné mohou být spojité, jako je například příjem, věk, počet let studia, součtové indexy, průměry položkových hodnot nebo kategorizované, jako například nejvyšší dosažený stupeň vzdělání, rodinný stav, pohlaví. Pokud je kategorizovaná proměnná nominální, tj. mezi jejími kategoriemi nejsou žádné relace (například uspořádání, vzdálenost), pak je nepřípustné vstoupit do regrese (a nejen logistické!) s kódy těchto kategorií, ale je třeba místo této jedné proměnné s I kategoriemi vytvořit $I - 1$ nových kontrastních proměnných, které určitým způsobem korespondují s původními kategoriemi.[4] Předpokládejme, že poslední nezávislá proměnná X_K je kategorizovaná a že má I kategorií. Označíme-li $I - 1$ nových proměnných D_{K1}, D_{K2}, …, $D_{K,I-1}$, pak rovnice modelu je

$$\text{logit}(Y) = \alpha + \beta_1 X_1 + \ldots + \beta_{K-1} X_{K-1} + \Sigma\, \beta_{Ki} D_{Ki}. \quad (1.\ 8)$$

Nejobvyklejším způsobem jak se vyrovnáváme s nominálními nezávislými proměnnými, je vytvoření $I - 1$ tzv. indikátorových proměnných. Každá z těchto proměnných odpovídá jedné z $I - 1$ kategorií nezávislé proměnné, vynechaná kategorie se nazývá referenční a můžeme si ji zvolit. Program LOGISTIC REGRESSION nabízí dalších šest standardních

[4]) Neobávejte se, že budete muset tyto nové proměnné vytvářet sami, program LOGISTIC REGRESSION to udělá za vás, musíte mu jen sdělit, které nezávislé proměnné jsou kategorizované.

způsobů vytvoření kontrastních proměnných[5] a umožňuje analytikovi i realizaci vlastního způsobu. Na zvoleném způsobu pak závisí interpretace regresních koeficientů.

Jsou-li některé nezávislé kategorizované proměnné ordinální, máme několik možností, jak s nimi zacházet. Má-li ordinální proměnná dostatečný počet kategorií, říkává se alespoň sedm, můžeme s ní pracovat jako se spojitou. Vždy je ovšem možný výše popsaný způsob vytvoření nových proměnných. Pro ordinální proměnné je zajímavý způsob označený v programu LOGISTIC REGRESSION jako Polynomial. Z nezávislé ordinální proměnné, která má například čtyři kategorie, vzniknou tři proměnné, přičemž regresní koeficient u první z nich charakterizuje lineární vztah mezi nezávislou proměnnou a logitem, koeficient u druhé z nich kvadratický vztah a koeficient u třetí z nich charakterizuje vztah kubický. Ještě jiný způsob práce s ordinálními proměnnými nabízí program GOLDMineR a je popsán v části 2. 5.

Pokud je mezi nezávislými proměnnými binomická proměnná, například pohlaví, můžeme si v logistické regresi vybrat, zda pro ni vytvoříme novou proměnnou, nebo zda ji budeme považovat za spojitou. V obou případech dostaneme jen jeden regresní koeficient, i když se numericky mohou lišit. Hosmer a Lemeshow [1989: 45-47] doporučují, aby kategorie všech dichotomických proměnných byly pro logistickou regresi kódovány nulou a jedničkou a aby se s těmito proměnnými pracovalo jako se spojitými.

1. 3 Interakce mezi nezávislými proměnnými

Do rovnice logistické regrese lze mezi nezávislé proměnné zahrnout i interakce jednotlivých proměnných. Jsou-li X a Z spojité, lze uvažovat například X^2, $X \times Z$, Z^3 a nezpůsobíme si v zásadě žádné velké komplikace, neboť význam interakčních členů je zřejmý. Komplikovanější situace nastává, pokud se jedná o interakci jedné kategorizované proměnné a jedné či více spojitých, a ještě složitější případ představuje interakce dvou či více kategorizovaných proměnných. Jak jsme uvedli výše, kategorizovaná proměnná vstupuje do logistické regrese v podobě sady nových proměnných a význam interakcí je tak závislý na způsobu jejich vytvoření. Proto v každém konkrétním případu je nutné si rozebrat, co daná interakce vlastně reprezentuje, a to nemusí být vždy jednoduché. Vstup interakčních členů do logistické regrese klade na interpretaci modelu a významu jednotlivých proměnných i jinak dodatečné nároky. V tomto článku se s tímto problémem nebudeme dále zabývat a zájemcům doporučuji již mnohokrát citovanou knihu Hosmera a Lemeshowa [Hosmer a Lemeshow 1989: 63-71].

1. 4 Statistiky pro ohodnocení logistického regresního modelu

Pracuje náš model dobře? Můžeme si být jisti, že mezi souborem vysvětlujících proměnných a vysvětlovanou proměnnou je vztah? Existuje-li tento vztah, jak je silný? Statistik, které dávají určitým způsobem odpovědi na vznesené otázky, existuje celá řada. Zde se omezíme na ty, které nacházíme ve výstupech výše zmíněného programu. Za prvé je to statistika –2LL (–2 log likelihood), která má asymptoticky rozdělení χ^2. Tato statistika nabývá kladných hodnot a větší hodnoty indikují horší predikci závislé proměnné. Nejprve se určí hodnota této statistiky pro model, který obsahuje jenom konstantu α, potom pro model, který obsahuje zvolenou skupinu K vysvětlujících proměnných. Jejich rozdíl se nazývá χ^2 modelu a poskytuje test nulové hypotézy, že v logistickém regresním mode-

[5]) Angličtina má pro tyto nové proměnné termín „dummy variables“. Někteří autoři používají ale tento termín jen pro indikátorové proměnné.

lu $\beta_1 = \beta_2 = \ldots = \beta_K = 0$. Je-li dosažená hladina významnosti (P-hodnota) menší než předem zvolená hladina významnosti nebo jí je rovna, pak zamítáme tuto nulovou hypotézu a vyvozujeme, že informace o nezávislých proměnných umožňuje lepší predikci závislé proměnné, než by byla možná bez této informace. Za hladinu významnosti, ke které vztahujeme P-hodnotu, obvykle volíme číslo 0,05.

Jiný způsob, jak ocenit adekvátnost modelu, spočívá v porovnání pozorovaných a modelem predikovaných zařazení do kategorií binární vysvětlované proměnné, které je vyjádřeno klasifikační tabulkou. Klasifikační tabulka má dva řádky a dva sloupce. Číslo na průsečíku řádku *r* a sloupce *s* udává, u kolika případů s pozorovanou hodnotou závislé proměnné *r*, byla predikována hodnota *s* (*r*, *s* = 0, 1). Případ je zařazen do kategorie s označením 1, jestliže modelem predikovaná $P(Y = 1) \geq 0{,}5$.[6] Součet případů na hlavní diagonále udává, kolik případů bylo klasifikováno správně. Může se snadno stát, že model je dobrý, ale jeho diskriminační síla je slabá, což vyjeví právě klasifikační tabulka. Je-li hlavním účelem logistického regresního modelu predikce, pak musíme hledat další vysvětlující proměnné, které zlepší jeho diskriminační sílu.

Pro logistickou regresi bylo navrženo mnoho analogů ke koeficientu determinace R^2, který je znám z lineární regrese a jehož stonásobek má onu příjemnou interpretovatelnost jako procento variability v závislé proměnné vysvětlené uvažovanými nezávislými proměnnými [viz např. Knoke a Burke 1980, Hosmer a Lemeshow 1989, Agresti 1990, Menard 1995]. Ve výstupech z SPSS se setkáme s R^2 Coxové a Snella a s R^2 Nagelkerka. Určitý nedostatek prvního z koeficientů je v tom, že nemůže dosáhnout maximální hodnoty 1. Proto Nagelkerke navrhl modifikaci, která tento nedostatek odstraňuje. Interpretace těchto dvou koeficientů je analogická interpretaci koeficientu determinace v lineární regresi, i když variabilita v logistickém regresním modelu musí být definována jinak [viz *SPSS Regression…* 1999: 45-46, Nagelkerke 1991].

Test dobré shody regresního modelu s daty skýtá další možnost ocenění přiměřenosti modelu. Pro logistickou regresi takový test navrhli Hosmer a Lemeshow [1989: 140-145]. Smysluplné využití výsledku tohoto testu je však možné jen tehdy, máme-li dostatečně velký výběrový soubor. To však ještě samo o sobě nestačí. Soubor dat je pro účely tohoto testu rozdělen podle určitého kritéria do deseti přibližně stejně velkých skupin a v každé z těchto skupin se zjišťuje skutečný a očekávaný počet případů, u kterých sledovaný jev *J* nastal či nenastal. Aby byl výsledek testu použitelný, je nutné, aby všechny očekávané četnosti nebyly menší než 1 a aby většina z nich byla větší než 5.

1. 5 Interpretace regresních koeficientů

Z rovnice 1. 5 vyplývá, že logistický koeficient β_k lze interpretovat jako změnu logitu spojenou s jednotkovou změnou hodnoty nezávislé proměnné X_k za předpokladu, že hodnoty ostatních nezávislých proměnných se nezmění.[7] Z rovnice 1. 6 plyne, že $\exp(\beta_k)$ je násobek, o který se změní šance, jestliže hodnota nezávislé proměnné X_k se změní o jednotku a hodnoty ostatních nezávislých proměnných se nezmění. Je-li $\beta_k > 0$, šance se zvětší, je-li $\beta_k < 0$, šance se zmenší, je-li $\beta_k = 0$, šance se nezmění.

---

[6]) Je možné zvolit si i jiné číslo než 0,5, musí to ale být číslo mezi nulou a jedničkou s výjimkou těchto krajních hodnot.

[7]) Proto je velmi důležité uvádět v publikacích, v jakých jednotkách jsou proměnné měřené. Hodnoty koeficientu například u příjmu budou vypadat jinak pro příjem udávaný v korunách nebo v tisících korun.

Interpretace regresních koeficientů u sady nových proměnných reprezentujících kategorizovanou proměnnou závisí na způsobu jejich konstrukce. Tak například koeficienty u indikátorových proměnných reprezentují efekt každé kategorie v porovnání s kategorií referenční. Regresní koeficient pro referenční kategorii je roven nule. Koeficienty u nových proměnných vytvořených pomocí polynomických kontrastů z určité kategorizované proměnné zase reprezentují sílu lineárního, kvadratického, kubického atd. vztahu mezi logitem a danou kategorizovanou proměnnou. Více k tomuto problému lze nalézt například u Hosmera a Lemeshowa [Hosmer a Lemeshow 1989: 47-56] a Menarda [Menard 1995: 50-52].

Hodnota regresního koeficientu β_k resp. $\exp(\beta_k)$ sama o sobě nestačí k vyslovení závěru, že nezávislá proměnná X_k je významná pro predikci či vysvětlení závislé proměnné. K tomu je třeba testovat hypotézu, že $\beta_k = 0$. Test je založen na Waldově statistice, která má asymptoticky rozdělení χ^2 , proto je test vhodný jen pro dostatečně velké výběrové soubory.[8] Je-li dosažená hladina významnosti menší než například 0,05, zamítáme hypotézu, že $\beta_k = 0$, v opačném případě nemáme důvod tuto hypotézu zamítnout. Waldova statistika má tu nepříjemnou vlastnost, že pro regresní koeficienty s velkou absolutní hodnotou, a tudíž i s velkou standardní chybou nabývá malých hodnot, takže nulová hypotéza není zamítnuta, i když by měla být. Kdykoliv se tedy vyskytne regresní koeficient s velkou absolutní hodnotou, neměli bychom na Waldovu statistiku spoléhat. Místo toho se doporučuje vytvořit model s odpovídající proměnnou a model bez ní a test významnosti založit na změně v hodnotě –2LL [viz Hauck a Donner 1977].

1. 6 Příklad

Nejprve popíšeme závislou proměnnou a čtyři nezávislé proměnné, z nichž tři jsou kategorizované a jedna spojitá. Závislou proměnnou nazveme *Přínos*. Kategorie s kódem 1 (resp. 0) znamená, že ekonomický systém, který je nyní v České republice, přinesl respondentovi a jeho rodině více špatného než dobrého (resp. více dobrého než špatného). Nezávislé proměnné jsou *Živur* (životní úroveň), *Polor* (politická orientace), *Třída* (společenská skupina) a *Příjos* (průměrný celkový čistý měsíční příjem domácnosti na osobu). Kategorie proměnné *Živur* jsou *nižší*, *střední*, *vyšší*. Kategorie proměnné *Polor* jsou *levice*, *střed*, *pravice*. Kategorie proměnné *Třída* jsou *nižší nebo dělnická*, *nižší střední*, *střední nebo vyšší střední nebo vyšší*. Hodnoty proměnné *Příjos* jsou 1, 2, …, 30, jsou to totiž kvantily. Hodnotu 1 nabývá tato proměnná pro přibližně 3,3 % respondentů s nejnižšími příjmy domácnosti na osobu a hodnotu 30 pro přibližně 3,3 % respondentů s nejvyššími příjmy domácnosti na osobu. Počet případů, které vstupují do analýzy je 848.

Živur, *Polor* a *Třída* jsou kategorizované proměnné o třech kategoriích. Do logistické regrese vstoupí tři sady indikátorových proměnných: *Živur*(1), která odpovídá kategorii *nižší*, *Živur*(2), která odpovídá kategorii *střední*, *Polor*(1), která reprezentuje kategorii *levice*, *Polor*(2), která reprezentuje kategorii *střed*, *Třída*(1), která odpovídá kategorii *nižší nebo dělnická*, *Třída*(2), která odpovídá kategorii *nižší střední*. Referenční

[8]) V programu LOGISTIC REGRESSION se pod Waldovou statistikou míní druhá mocnina zlomku, který má v čitateli odhad b_k regresního koeficientu β_k a ve jmenovateli jeho standardní chybu. Takto definovaná statistika má asymptoticky rozdělení χ^2. Ale například Hosmer a Lemeshow [1989] míní pod Waldovou statistikou týž zlomek, ale bez druhé mocniny. Takto definovaná Waldova statistika má asymptoticky standardní normální rozdělení.

kategorií je u všech třech kategorizovaných proměnných ta poslední.[9] S proměnnou *Příjos* budeme pracovat jako se spojitou. Rovnice uvažovaného modelu je

$$\text{logit}(Přínos) = \alpha + \beta_{11}Živur(1) + \beta_{12}Živur(2) + \beta_{21}Polor(1) + \\ + \beta_{22}Polor(2) + \beta_{31}Třída(1) + \beta_{32}Třída(2) + \beta_4Příjos. \quad (1.\ 9)$$

Pomocí metody maximální věrohodnosti a hodnot všech v rovnici zahrnutých proměnných hledá program odhady koeficientů α, β_{11}, β_{12}, β_{21}, β_{22}, β_{31}, β_{32}, β_4. Současně počítá statistiky pro ohodnocení modelu (viz část 1. 4). Na základě těchto údajů nejprve zjistíme, zda je model adekvátní. Pokud není, hledáme jiný, pokud je, prohlídneme si odhady koeficientů a jejich významnost a přistoupíme k interpretaci.

Podle všech dostupných statistik se zdá, že model odpovídá datům dobře, neboť

a) χ^2 modelu je 365,532 a této hodnotě při sedmi stupních volnosti odpovídá dosažená významnost 0,000, takže zamítáme hypotézu, že $\beta_{11} = \beta_{12} = \beta_{21} = \beta_{22} = \beta_{31} = \beta_{32} = \beta_4 = 0$. To znamená, že informace o hodnotách nezávislých proměnných umožňuje lepší predikci závislé proměnné, než by byla možná bez této informace.

b) Z klasifikační tabulky se dozvídáme, že do kategorie závislé proměnné s kódem nula bylo správně zařazeno 67,0 % případů, do kategorie s kódem jedna 88,4 % případů a celkově bylo správně zařazeno 80,8 % případů, což ukazuje na docela dobrou diskriminační sílu modelu.

c) Hodnota R^2 Coxové a Snella je 0,350, hodnota R^2 Nagelkerka je 0,481. Podle posledního údaje tedy vyvozujeme, že model vysvětluje 48 % „variability“ v závislé proměnné.

d) Test dobré shody Hosmera a Lemeshowa je použitelný, neboť z kontingenční tabulky pro tento test, která je součástí výstupu zjišťujeme, že žádná z dvaceti očekávaných četností není menší než jedna a jen dvě jsou menší než pět. Hodnota statistiky χ^2 je 6,749, což při osmi stupních volnosti dává dosaženou hladinu významnosti 0,481. To znamená, že nezamítáme nulovou hypotézu, která postuluje, že mezi pozorovanými a modelem predikovanými hodnotami není žádný rozdíl.

Model je tedy přijatelný a na řadu přicházejí odhady koeficientů. Napíšeme si jeho logitovou formu a též jeho vyjádření pomocí šance:

$$\text{logit}(Přínos) = -1{,}400 + 1{,}946Živur(1) + 0{,}532Živur(2) + 2{,}474Polor(1) + \\ + 1{,}683Polor(2) + 0{,}983Třída(1) + 0{,}687Třída(2) - 0{,}037\text{Příjos}, \quad (1.10)$$

$$\text{šance}(Přínos = 1) = \text{P}(Přínos = 1) / \text{P}(Přínos = 0) = \\ = \text{P}(\textit{převaha špatného}) / \text{P}(\textit{převaha dobrého}) = \exp(-1{,}400 + 1{,}946Živur(1) + \\ + 0{,}532Živur(2) + 2{,}474Polor(1) + 1{,}683Polor(2) + \\ + 0{,}983Třída(1) + 0{,}687Třída(2) - 0{,}037Příjos). \quad (1.\ 11)$$

Všechny koeficienty jsou statisticky významné. Dosažená hladina významnosti pro koeficient u proměnné *Živur*(2) je 0,013, u *Třída*(2) to je 0,003, u *Příjos* 0,001 a u všech ostatních 0,000. To znamená, že všechny uvažované proměnné mají významný vliv na predikci či na vysvětlení závislé proměnné.

S interpretací začněme u spojité proměnné *Příjos*. Jestliže se hodnota této proměnné změní o jednotku a hodnoty ostatních nezávislých proměnných zůstanou beze změny,

[9]) Nemusí to být vždy ta poslední, v programu LOGISTIC REGRESSION si ji můžeme zvolit.

pak se logit změní o hodnotu koeficientu u *Příjos*, tj. o –0,037. Násobek, o který se změní šance převahy špatného, jestliže hodnota *Příjos* se změní o jednotku a hodnoty ostatních nezávislých proměnných se nezmění je exp(–0,037) = 0,964. To znamená, že se šance převahy špatného zmenší. S růstem příjmu domácnosti na osobu tedy klesá šance názoru, že ekonomický systém přinesl respondentovi a jeho rodině více špatného než dobrého.

Koeficient u proměnné *Živur*(1) charakterizuje změnu v logitu, když porovnáme nižší životní úroveň s vyšší, koeficient u proměnné *Živur*(2) charakterizuje změnu v logitu, když porovnáme střední životní úroveň s vyšší. Obě hodnoty koeficientu jsou kladné (1,946 a 0,532), to znamená, že vztaženo k vyšší životní úrovni, nižší a střední životní úroveň jsou spojeny s růstem logitu převahy špatného. Navíc z hodnot usuzujeme, že nižší životní úroveň zvětšuje logit více než střední životní úroveň. V terminologii šancí vyjadřuje hodnota exp(1,946) = 6,998 poměr šancí převahy špatného pro nižší životní úroveň vzhledem k vyšší za předpokladu, že hodnoty ostatních nezávislých proměnných se nezmění. Obdobně hodnota exp(0,532) = 1,702 je poměr šancí převahy špatného pro střední životní úroveň vzhledem k vyšší životní úrovni za předpokladu, že hodnoty ostatních nezávislých proměnných se nezmění. Šance názoru, že současný ekonomický systém přinesl respondentovi a jeho rodině více špatného než dobrého, je tedy téměř sedmkrát větší u respondentů s nižší životní úrovní než u respondentů s vyšší životní úrovní a 1,7 krát větší u respondentů se střední životní úrovní než u respondentů s vyšší životní úrovní.

Interpretace dalších dvou kategorizovaných proměnných jsou analogické, neboť i pro ně byly použity indikátorové proměnné. Závěr plynoucí z hodnot koeficientů u těchto proměnných je následující: Šance názoru, že současný ekonomický systém přinesl respondentovi a jeho rodině více špatného než dobrého, je skoro dvanáctkrát větší u respondentů levicově orientovaných než u respondentů pravicově orientovaných a 5,4 krát větší u respondentů středových než u respondentů pravicových. Šance téhož názoru je 2,7 krát větší u respondentů hlásících se k nižší nebo dělnické třídě než u respondentů hlásících se ke střední nebo vyšší střední nebo vyšší třídě a skoro dvakrát větší u respondentů hlásících se k nižší střední vrstvě než u respondentů hlásících se ke střední, vyšší střední nebo vyšší třídě.

Kategorizované proměnné *Živur*, *Polor* a *Třída* jsou ordinální, a proto na nich můžeme demonstrovat i postup jejich záměny sadou kontrastních polynomiálních proměnných. Protože každá ze zmíněných proměnných má tři kategorie, bude reprezentována dvěma novými proměnnými. *Živur*(1), *Polor*(1) a *Třída*(1) nyní charakterizují lineární vztah mezi životní úrovní nebo politickou orientací nebo společenskou skupinou a logitem, *Živur*(2), *Polor*(2) a *Třída*(2) kvadratický vztah. Na základě hodnot logistických regresních koeficientů u proměnných *Živur*(1), …, *Třída*(2) lze testovat, zdali vztah mezi logitem a jednotlivými kategorizovanými nezávislými proměnnými má významné lineární a/nebo kvadratické složky. Na základě předchozích výsledků lze předpokládat významnost pouze lineární složky jen u proměnné *Třída*. Na tomto místě je třeba upozornit, že jiná volba kontrastů mění pouze hodnoty logistických regresních koeficientů, statistiky pro ohodnocení modelu se nemění.

Dosažená hladina významnosti logistických regresních koeficientů v modelu daném rovnicí (1. 9) s polynomiálními kontrasty pro kategorizované proměnné je 0,000 u *Živur*(1), 0,018 u *Živur*(2), 0,000 u *Polor*(1), 0,020 u *Polor*(2), 0,000 u *Třída*(1), 0,343 u *Třída*(2), 0,001 u *Příjos*. To znamená, že u proměnných *Živur* a *Polor* jsou významné jak

složky lineární, tak kvadratické, zatímco u proměnné *Třída* je významná jen složka li neární.

Při výše popsaném využití polynomiálních kontrastů jsme předpokládali, že vzdálenost kategorií jednotlivých kategorizovaných nezávislých proměnných je stejná. To znamená například, že vzhledem ke zkoumanému problému je levice od středu stejně vzdálená jako střed od pravice. Program LOGISTIC REGRESSION nám umožňuje tento předpoklad opustit, musíme však vzdálenosti kategorií určitým způsobem zadat [viz např. SPSS *Regression…* 1999: 182]. Jiným způsobem řeší problematiku vzdálenosti kategorií ordinálních proměnných program GOLDMineR 2.0, jak ukážeme v části 2. 5.

2. Polytomická závislá proměnná[10]

2. 1 Nominální závislá proměnná

Máme-li kategorizovanou závislou proměnnou, která má více než dvě kategorie, pak problém nelze řešit opakovaným použitím logistické regrese, ale pomocí multinomické (polytomické) logistické regrese, která je rozšířením binárního logistického regresního modelu. Rozšíření je vedeno takto: Jedna z hodnot závislé proměnné je považována za referenční (v programu NOMREG je to vždy poslední kategorie[11]) a pravděpodobnost příslušnosti ke každé jiné kategorii je porovnávána s pravděpodobností příslušnosti k referenční kategorii. Má-li závislá proměnná M kategorií, pak je třeba pro popis vztahu mezi závislou proměnnou a K nezávislými proměnnými řešit $M - 1$ rovnic

$$\ln[\mathrm{P}(\mathrm{kategorie}_\mathrm{m}) / \mathrm{P}(\mathit{kategorie}_M)] = \alpha_\mathrm{m} + \beta_{\mathrm{m}1}X_1 + \ldots + \beta_{\mathrm{m}K}X_\mathrm{K}$$
$$\mathrm{m} = 1, \ldots, M - 1. \qquad (2.\ 1)$$

Pro každý z $M - 1$ logitů[12] tak dostáváme zvláštní sadu regresních koeficientů. Pro referenční kategorii jsou všechny koeficienty rovny nule.

O způsobu zacházení s nezávislými proměnnými platí totéž, co bylo řečeno v částech 1. 2 a 1. 3. Program NOMREG ale nemá volbu tvorby kontrastních proměnných, pracuje pouze s indikátorovými proměnnými a referenční kategorií je vždy ta poslední, stejně jako u závislé proměnné.[13] Interpretace regresních koeficientů je stejná jako v části 1. 5. Platí i vše, co bylo řečeno v téže části o testování hypotéz týkajících se regresních koeficientů. Zde jsou navíc k dispozici testy poměrem věrohodnosti efektů jednotlivých nezávislých proměnných.

Některé statistiky pro ohodnocení multinomického logistického modelu, které poskytují výstupy programu NOMREG, jsou analogické těm, o kterých jsme hovořili v části

[10]) Česká terminologie není ustálená, takže se můžeme setkat i s jiným pojmenováním kategorizované proměnné, která má více než dvě hodnoty. Někteří autoři používají například označení množná proměnná.

[11]) Pro tuto referenční kategorii je v NOMREG používáno termínu „baseline category“. Pokud chceme za referenční kategorii zvolit jinou než poslední, musíme bohužel závislou proměnnou přetransformovat tak, aby v nové proměnné byla vybraná kategorie na posledním místě.

[12]) Striktně řečeno nejde o logity, ale o zobecněné logity, neboť levá část rovnice (2. 1) nemá tvar obyčejného logitu, tj. $\ln[\mathrm{P}(J) / \mathrm{P}(\mathrm{non}\ J)]$.

[13]) Program NOMREG nabízí ve srovnání s programem LOGISTIC REGRESSION zase například explicitní specifikaci hnízdových (nested) modelů. Tento program lze použít i pro binární logistickou regresi. Srovnání možností obou programů lze nalézt například v [*SPSS Regression…* 1999: 1-2].

1. 4, některé jsou jiné, některé chybějí. Setkáme se zde opět s klasifikační tabulkou, která ovšem má tentokrát *M* řádků a *M* sloupců, ale nenajdeme tu test dobré shody Hosmera a Lemeshowa. Najdeme tu R^2 Coxové a Snella, R^2 Nagelkerka a navíc ještě R^2 McFaddena [viz McFadden 1973, Agresti 1990: 110, *SPSS Regression…* 1999: 75].

O kvalitě modelu dále vypovídá výsledek testu, který zkoumá, zda konečný model je významně lepší než model, který obsahuje jen konstantu. Ve výstupech ho najdeme pod hlavičkou Model Fitting Information. Jestliže dosažená hladina významnosti je malá, například menší než 0,05, pak výsledný model je významně lepší než model, který obsahuje jen konstantu. Testy dobré shody jsou zde dva, a sice Pearsonova statistika χ^2 a odchylka χ^2 [viz *SPSS Regression…* 1999: 73-74]. Je-li dosažená hladina významnosti těchto dvou testů malá (například menší než 0,05), pak zamítáme hypotézu, že model odpovídá pozorovaným datům. V opačném případě nemáme důvod tuto hypotézu zamítnout. Tyto statistiky jsou velmi užitečné pro modely s malým počtem kategorizovaných nezávislých proměnných. Jsou však velmi citlivé na prázdná pole v odpovídající mnohorozměrné tabulce. Je-li mezi nezávislými proměnnými spojitá proměnná, pak velmi často nastává situace s mnoha prázdnými políčky. Tehdy se na dosažené hladiny významnosti nemůžeme spoléhat.

2. 2 Příklad

Nezávislé proměnné jsou stejné jako v příkladu z části 1. 6, závislá proměnná *NPřínos* má ale nyní tři kategorie: kód 0 (resp. 1, resp. 2) znamená, že ekonomický systém, který je nyní v České republice, přinesl respondentovi a jeho rodině více špatného než dobrého (resp. stejně špatného jako dobrého, resp. více dobrého než špatného).[14] Počet případů vstupujících do analýzy je 1351. Pro popis vztahu mezi závislou proměnnou *NPřínos* a nezávislými proměnnými *Živur*(1), *Živur*(2), *Polor*(1), *Polor*(2), *Třída*(1), *Třída*(2) a *Příjos* potřebujeme dvě rovnice

$$\begin{aligned}\ln[\mathrm{P}(\textit{převaha špatného}) / \mathrm{P}(\textit{převaha dobrého})] = \alpha_1 + \beta_{111}\textit{Živur}(1) + \\ + \beta_{112}\textit{Živur}(2) + \beta_{121}\textit{Polor}(1) + \beta_{122}\textit{Polor}(2) + \\ + \beta_{131}\textit{Třída}(1) + \beta_{132}\textit{Třída}(2) + \beta_{14}\textit{Příjos},\end{aligned} \qquad (2.\ 2)$$

$$\begin{aligned}\ln[\mathrm{P}(\textit{stejně špatného jako dobrého}) / \mathrm{P}(\textit{převaha dobrého})] = \alpha_2 + \beta_{211}\textit{Živur}(1) + \\ + \beta_{212}\textit{Živur}(2) + \beta_{221}\textit{Polor}(1) + \beta_{222}\textit{Polor}(2) + \\ + \beta_{231}\textit{Třída}(1) + \beta_{232}\textit{Třída}(2) + \beta_{24}\textit{Příjos}.\end{aligned} \qquad (2.\ 3)$$

Nejprve je třeba zkontrolovat vhodnost modelu. Dosažená hladina významnosti u statistik dobré shody Pearsonův χ^2 a odchylka χ^2 je příznivá závěru, že model dobře odpovídá datům (významnosti jsou 0,138 a 0,102), ale protože mezi nezávislými proměnnými je jedna spojitá proměnná, nelze na tyto testy zcela spoléhat.[15] Dále se ukazuje, že lze zamítnout hypotézu, že všechny regresní koeficienty β jsou rovny nule, z čehož usuzujeme, že model je významně lepší než model, který by obsahoval jen konstantu. Hodnota R^2 Coxové a Snella je 0,266, Nagelkerka 0,302 a McFaddena 0,145. Tyto hodnoty jsou nižší než v předchozím příkladu, ale lze je ještě považovat za uspokojivé. Predikční síla modelu je slabá, protože v klasifikační tabulce bylo do kategorie *převaha špatného* správně

[14]) Jde tedy vlastně o ordinální proměnnou, ale zde budeme tuto vlastnost ignorovat.

[15]) Pokud bychom z modelu vynechali spojitou proměnnou *Příjos*, dostali bychom významnosti 0,960 u Pearsona a 0,924 u odchylky, což by ukazovalo velmi dobrou shodu modelu a pozorovaných dat.

zařazeno 71,2 % případů, do kategorie *stejně špatného jako dobrého* 43,9 % a do kategorie *převaha dobrého* 43,2 %, což dává celkové procento správných predikcí 54,8 %. Model tedy nelze používat pro predikci, ale jen pro objasnění typu a síly závislostí. Odhady regresních koeficientů jsou v následujících rovnicích:

$$\begin{aligned} \text{P}(\textit{převaha špatného}) / \text{P}(\textit{převaha dobrého}) = \exp(&-1{,}465 + 1{,}945\textit{Živur}(1) + \\ &+ 0{,}561\textit{Živur}(2) + 2{,}532\textit{Polor}(1) + 1{,}727\textit{Polor}(2) + \\ &+ 1{,}033\textit{Třída}(1) + 0{,}708\textit{Třída}(2) - 0{,}037\textit{Příjos}), \end{aligned} \tag{2. 4}$$

$$\begin{aligned} \text{P}(\textit{stejně špatného jako dobrého}) / \text{P}(\textit{převaha dobrého}) = \exp(&-0{,}299 + 1{,}032\textit{Živur}(1) + \\ &+ 0{,}568\textit{Živur}(2) + 1{,}396\textit{Polor}(1) + 0{,}877\textit{Polor}(2) + \\ &+ 0{,}350\textit{Třída}(1) + 0{,}304\textit{Třída}(2) - 0{,}021\textit{Příjos}). \end{aligned} \tag{2. 5}$$

Všechny koeficienty u nezávislých proměnných v rovnici (2. 4) jsou statisticky významné. Dosažená hladina významnosti koeficientu u proměnné *Živur*(2) je 0,005, u *Třída*(2) 0,001, u všech ostatních je to 0,000. Srovnáme-li rovnici (2. 4) s rovnicí (1. 11) vidíme, že numerické hodnoty odpovídajících si regresních koeficientů jsou si velmi podobné, takže závěry z těchto dvou rovnic plynoucí jsou prakticky stejné. V rovnici (2. 5) jsou statisticky významné koeficienty u proměnných *Živur*(1), *Živur*(2), *Polor*(1), *Polor*(2) a *Příjos*. Dosažené hladiny významnosti jsou po řadě 0,000, 0,001, 0,000, 0,000, 0,028. Nevýznamné jsou koeficienty u proměnných *Třída*(1) a *Třída*(2), když dosažené hladiny významnosti jsou 0,081 a 0,115. Porovnáme-li hodnoty regresních koeficientů u nezávislých proměnných v rovnicích (2. 4) a (2. 5) vidíme, že mají stejná znaménka a že až na *Živur*(2) jsou v rovnici (2. 5) vždy slabší než v rovnici (2. 6).

2. 3 Ordinální závislá proměnná: kumulativní logit

V této části probereme postup ordinální logistické regrese, kdy závislá proměnná je ordinální a tato její vlastnost je respektována a s nezávislými kategorizovanými proměnnými se pracuje jako s nominálními, i kdyby mezi nimi byly i ordinální. Program PLUM [viz *SPSS Advanced*... 1999: 63-70, 241-260], který budeme využívat k výpočtům, vytváří ke každé kategorizované nezávislé proměnné sadu indikátorových proměnných, stejně jako program NOMREG, jinou volbu neumožňuje.

Model je založen na předpokladu, že existuje latentní spojitá proměnná a manifestní ordinální proměnná vzniká diskretizací tohoto skrytého kontinua do M uspořádaných skupin. Hodnoty na tomto kontinuu, které definují kategorie, jsou odhadovány pomocí prahů $\theta_1, \ldots, \theta_{M-1}$. Práh θ_m závisí pouze na tom, která z kategorií závislé proměnné je predikována, nikoliv na hodnotách nezávislých proměnných $X_1, \ldots, X_K$ a regresní koeficienty $\beta_1, \ldots, \beta_K$ jsou stejné pro všechna m, tedy nezávisí na m. Rovnice modelu je

$$\ln[\gamma_m / (1 - \gamma_m)] = \theta_m - (\beta_1 X_1 + \ldots + \beta_K X_K), \tag{2. 6}$$

kde γ_m je kumulativní pravděpodobnost kategorie m, $m = 1, \ldots, M - 1$, tj.

$$\gamma_m = \text{P}(\text{kategorie}_1) + \ldots + \text{P}(\text{kategorie}_m). \tag{2. 7}$$

Práh θ_m v rovnici (2. 6) koresponduje s interceptem α v lineární regresi. Označíme-li J_m jev, že se respondent zařadil do kategorie$_1$ ∪ … ∪ kategorie$_m$ (symbol ∪ znamená konjunkci čili nebo), pak můžeme rovnici (2. 6) přepsat do tvaru

$$\ln[\text{P}(J_m) / \text{P}(\text{non } J_m)] = \theta_m - (\beta_1 X_1 + \ldots + \beta_K X_K) \tag{2. 8}$$

a následně

$$\text{P}(J_m) / \text{P}(\text{non } J_m)) = \exp[\theta_m - (\beta_1 X_1 + \ldots + \beta_K X_K)]. \tag{2. 9}$$

Levá strana rovnice (2. 9) má význam šance jevu J_m, tedy šance, že se respondent zařadil do kategorie$_1$ □ … □ kategorie$_m$.

Předpoklad modelu, že regresní koeficienty β_1, …, β_K jsou stejné pro všechna *m*, lze v programu PLUM otestovat pomocí testu rovnoběžnosti. Předpoklad je přijatelný, jestliže dosažená hladina významnosti příslušné testové statistiky je velká, řekněme alespoň větší než 0,05. Pokud tento předpoklad není splněn, pak je lépe pracovat se závislou proměnnou, jako kdyby byla nominální. Zjišťování adekvátnosti modelu (2. 6) se děje stejnými prostředky, které byly popsány v části 2. 1. Připomeneme si je v následujícím příkladu.

2. 4 Příklad

Proměnné jsou až na jednu výjimku stejné jako v příkladu z části 2. 2. Zde totiž vynecháme spojitou proměnnou *Příjos*, abychom se nedostávali do problémů při testech využívajících statistiku χ^2. Přítomnost této proměnné vyvolává velký počet prázdných polí v mnohorozměrné tabulce dat a tím jsou zpochybněny výsledky testů založených na statistice χ^2.[16] Počet případů je nyní 1631.

Nejprve zjistíme, jak je to s adekvátností modelu. Hypotézu $\beta_1 = \beta_2 = \ldots = \beta_K = 0$ zamítáme, protože dosažená hladina významnosti je 0,000. Testy dobré shody Pearsonův χ^2 a odchylka χ^2 ukazují na dobrou shodu dat a modelu, protože dosažená hladina významnosti prvého je 0,976 a druhého 0,946. R^2 Coxové a Snella, Nagelkerka a McFaddena jsou po řadě 0,257, 0,291, 0,138. Predikční síla modelu je slabá, protože do první kategorie závislé proměnné bylo správně zařazeno 62,9 % respondentů, do druhé 52,5 % a do třetí 42,3 %. Celkem bylo správně zařazeno jen 54,1 % respondentů. Pokud bychom chtěli model využívat k predikci, museli bychom hledat další nezávislé proměnné nebo zkusit jinou transformační funkci než logit, což program PLUM umožňuje, pak bychom už ovšem nemluvili o logistické regresi. Dosažená hladina významnosti testu rovnoběžnosti je 0,812, a to znamená, že nelze zamítnout hypotézu, že koeficienty β_1, …, β_K jsou nezávislé na *m*.

Odhady regresních koeficientů jsou uvedeny v následující rovnici modelu, ve které J_1 označuje skutečnost, že se respondent řadí do kategorie *převaha špatného* a J_2 skutečnost, že se respondent řadí do kategorií *převaha špatného* nebo *stejně špatného jako dobrého*.

$$\begin{aligned} P(J_m) / P(\text{non } J_m)) = \exp[\theta_m + 1{,}329\textit{Živ}ur(1) + 0{,}349\textit{Živ}ur(2) + \\ + 1{,}592\textit{Polor}(1) + 1{,}142\textit{Polor}(2) + 0{,}964\textit{Třída}(1) + 0{,}638\textit{Třída}(2)], \qquad (2.\ 10) \\ m = 1,\ 2,\ \theta_1 = -2{,}583,\ \theta_2 = -0{,}447. \end{aligned}$$

Dosažená hladina významnosti u proměnné *Živ*ur(2) je 0,003, u všech ostatních 0,000. To znamená, že všechny uvedené nezávislé proměnné jsou důležité. Šance názoru *převaha špatného* pro lidi s nižší (střední) životní úrovní je 3,8 krát (1,4 krát) větší než pro lidi s vyšší životní úrovní. Tatáž šance pro lidi politicky orientované vlevo (středově) je 4,9 krát (3,1 krát) větší než pro lidi s pravicovou orientací. A konečně stejná šance pro lidi řadící se do nižší nebo dělnické třídy (nižší střední třídy) je 2,6 krát (1,9 krát) větší než

[16]) Vliv této proměnné by byl sice statisticky významný, ale hodnoty statistik pseudo R^2 a predikční sílu modelu zlepšuje jen nepatrně. Proto si ji dovolím vynechat.

pro lidi řadící se do vyšších tříd.[17] Z modelu (2. 10) ovšem plyne, že totéž lze zopakovat i pro šanci názoru *převaha špatného* nebo *stejně špatného jako dobrého*.

2. 5 Ordinální závislá proměnná: monotónní regresní model

V minulých dvaceti letech se objevila řada prací zejména Leo A. Goodmana,[18] které svědčily o existenci jednoho regresního supermodelu, který je vhodný pro závislou proměnnou ať už dichotomickou, ordinální nebo spojitou. Tento supermodel v sobě obsahuje lineární a logistický regresní model jako speciální případy. GOLDMineR (Grafical Ordinal Logit Displays based on Monotonic Regression)[19] je první specializovaný program, který umožňuje analyzovat dichotomické, ordinální a spojité závislé proměnné. Program má velmi široké možnosti zacházení se závislými i nezávislými proměnnými (například různé způsoby vytváření kontrastních proměnných včetně vlastního), bohaté výstupy, zejména grafické. Naším účelem zde není seznámit čtenáře s celou škálou možností programu,[20] ale využít ho pro případ, kdy závislá i nezávislé proměnné jsou ordinální.

Abychom se vyhnuli složitému značení a indexování, budeme bez újmy na obecnosti předpokládat, že máme pouze tři nezávislé proměnné U, V a W a závislou proměnnou Y. Tyto proměnné nabývají po řadě hodnot $u_1, \ldots, u_I, v_1, \ldots, v_L, w_1, \ldots, w_S, y_1, \ldots, y_M$. Tyto hodnoty mohou být kódy kategorií, nebo jinak pevně zvolená čísla, nebo to mohou být zatím neznámé hodnoty, které budou odhadovány spolu s regresními koeficienty. Aby byly tyto neznámé hodnoty odhadnutelné, předpokládá se, že mají jednotkové rozpětí a že jejich součin s vahami kategorií je roven nule. Dále se omezíme jen na práci s referenčními kategoriemi. V tom případě je váha referenční kategorie rovna jedničce, váhy ostatních kategorií jsou nuly.[21] Opět jen pro zjednodušení zápisu, avšak bez újmy na obecnosti budeme předpokládat, že referenční kategorie jsou ty poslední. Rovnice modelu za vyjmenovaných předpokladů jsou

$$\ln[\mathrm{P}(Y = y_m) / \mathrm{P}(Y = y_M)] = \\ = \alpha_m + [\beta_1(U - u_I) + \beta_2(V - v_L) + \beta_3(W - w_S)] \times (y_m - y_M), \qquad (2.\ 11)$$
$$m = 1, \ldots, M - 1.^{22}$$

Odhadují se regresní koeficienty $\alpha_1, \ldots, \alpha_{M-1}, \beta_1, \beta_2, \beta_3$, případně také skóry kategorií jednotlivých proměnných a koeficienty β specifické pro jednotlivé kategorie jako $\beta_{1i} = \beta_1 \times u_i$, $\beta_{2l} = \beta_2 \times v_l$, $\beta_{3s} = \beta_3 \times w_s$, $i = 1, \ldots, I - 1$, $l = 1, \ldots, L - 1$, $s = 1, \ldots, S - 1$.

---

[17]) Uvedené hodnoty jsou exp(β). Například 3,8 = exp(1,329), 1,4 = exp(0,349), 4,9 = exp(1,592).

[18]) Tyto práce zde nebudu citovat, neboť se jedná o články ryze matematické a pro společenského vědce by nemusely být zcela srozumitelné. Čtenář najde v seznamu literatury odkaz na manuál GOLDMineR 2.0, kde se dozví o metodě dost na to, aby ji mohl používat, a kde je také uvedená pro případného čtenáře s matematicko-statistickým vzděláním bohatá literatura na dané téma.

[19]) Monotónní regrese je obecnější pojem než lineární regrese. Predikce závislé proměnné Y jakožto funkce prediktoru X_k je monotónní pro každé k = 1, …, K, jestliže kdykoliv se k-tý prediktor zvětší, zatímco ostatní zůstávají beze změny, predikovaná hodnota Y se zvětší, zmenší nebo zůstává konstantní v závislosti na tom, zda je $\beta_k > 0$, $\beta_k < 0$ nebo $\beta_k = 0$.

[20]) To by ani nebylo možné, neboť některé aplikace vyžadují znalosti, které nelze běžně předpokládat a jejichž výklad je nad rámec této stati.

[21]) GOLDMineR umožňuje kromě dalších dvou předvoleb typu vah i vstup vlastních vah.

[22]) Jedná se o zobecněný logit stejně jako v případě nominální polytomické proměnné.

Validita modelu se testuje pomocí množství asociace, kterou model nevysvětluje. Ve výstupu je značené *Residual* L^2. Analýza asociace mezi nezávislými proměnnými a závislou proměnnou totiž rozděluje celkovou asociaci (*Total* L^2) na množství vysvětlené modelem (*Explained* L^2) a na množství, které zůstává nevysvětlené (*Residual* L^2). Je-li dosažená hladina významnosti (p-value) u *Residual* L^2 velká (například alespoň větší než 0,05), pak model akceptujeme, v opačném případě nikoliv. Dosažená hladina významnosti příslušná k *Explained* L^2 se používá k testu hypotézy $R^2 = 0$ a také k testu hypotézy $\varphi = 0$. Pokud je malá (například menší než 0,05), pak zamítáme hypotézu, že $R^2 = 0$ (resp. $\varphi = 0$). V opačném případě nemáme důvod tyto hypotézy zamítnout. Statistika R^2, přesněji monotónní R^2 leží mezi nulou a jedničkou (obou krajních hodnot může nabýt) a její stonásobek měří procento variability v Y, které je vysvětlené monotónním regresním modelem. Není tak užitečná jako R^2 v lineární regresi, protože její hodnota může být docela nízká, když závislá proměnná není spojitá a / nebo když je hodně šikmá, a to i když model obsahuje vysoce významné prediktory. Statistika φ je nezáporná a měří stupeň asociace mezi pozorovanou a predikovanou závislou proměnnou.

Použití statistiky *Residual* L^2 je vázáno na dostatečně velké soubory vzhledem k počtu polí v mnohorozměrné tabulce četností, kterou chceme analyzovat. Jsou-li data řídká, pak bychom tuto statistiku neměli používat. GOLDMineR nám v této situaci nabízí alternativní statistiku *Decile* L^2, která je ve výstupech značená jako *Decile Fit*. Jde o rozšíření testu Hosmera a Lemeshowa, o kterém jsme hovořili v části 1. 4. Abychom si byli jisti, kdy lze věřit statistice *Residual* L^2 a kdy už je třeba použít *Decile* L^2, objevuje se na požádání ve výstupech ještě Pearsonova statistika *Residual* χ^2. Jestliže si hodnoty *Residual* L^2 a *Residual* χ^2 nejsou blízké, pak je třeba použít statistiku *Decile* L^2, jinak se doporučuje používat *Residual* L^2.

GOLDMineR 2.0 obsahuje ještě další způsob ohodnocení validity modelu, a to adjustovaná rezidua. Pro každé pole tabulky $U \times V \times W$ počítá standardizovaný rozdíl mezi predikovaným skórem proměnné Y a pozorovaným průměrným skórem proměnné Y. Adjustované reziduum, které má velikost větší nebo rovnou 1,96 respektive menší nebo rovnou (–1,96) indikuje, že rozdíl mezi predikovaným a pozorovaným průměrným skórem proměnné Y v příslušném poli je významný na hladině 0,05.

2. 6 Příklad

Závislá proměnná je *NPřínos* s třemi kategoriemi jako v části 2. 2. Jen stručně zopakuji, že kód 0 značí kategorii *převaha špatného*, kód 1 značí kategorii *stejně špatného jako dobrého* a kód 2 značí kategorii *převaha dobrého*. Nezávislé proměnné jsou *Živur* s kategoriemi *nižší*, *střední*, *vyšší* s kódy 0, 1, 2, *Polor* s kategoriemi *levice*, *střed*, *pravice* s kódy 0, 1, 2, *Třída* s kategoriemi *nižší nebo dělnická, nižší střední*, *střední nebo vyšší střední nebo vyšší*, zkráceně *střední+*, s kódy 0, 1, 2. Všechny zúčastněné proměnné jsou ordinální a my s nimi nejprve budeme pracovat způsobem, jako kdyby jejich kategorie byly od sebe stejně vzdálené. Kategoriím přiřadíme čísla 0, 0,5, 1.[23] Za referenční kategorii budeme považovat u všech proměnných tu poslední, které je zde vždy přiřazeno číslo 1.[24] Rovnice modelu jsou

[23]) Mohli bychom jim přiřadit i jiná čísla, třeba kódy 0, 1, 2, výsledná interpretace by se nezměnila. Podstatné je to, že čísla jsou od sebe stejně vzdálená.

[24]) Jak se ukáže později, není to ta nejšikovnější volba, snadněji se pracuje v situaci, kdy referenční kategorie má kód 0.

$$\ln[P(\textit{převaha špatného}) / P(\textit{převaha dobrého})] = \\ = \alpha_1 + [\beta_1(\textit{Živur} - 1) + \beta_2(\textit{Polor} - 1) + \beta_3(\textit{Třída} - 1)] \times (-1), \qquad (2.\ 12)$$

$$\ln[P(\textit{stejně špatného jako dobrého}) / P(\textit{převaha dobrého})] = \\ = \alpha_2 + [\beta_1(\textit{Živur} - 1) + \beta_2(\textit{Polor} - 1) + \beta_3(\textit{Třída} - 1)] \times (-0{,}5). \qquad (2.\ 13)$$

V rovnici (2. 12) činitel (–1) = (0 – 1), kde 0 je číslo přiřazené kategorii *převaha špatného* a 1 je číslo přiřazené referenční kategorii *převaha dobrého*. V rovnici (2. 13) činitel (–0,5) = (0,5 – 1), kde 0,5 je číslo přiřazené kategorii *stejně špatného jako dobrého* a 1 je číslo přiřazené referenční kategorii *převaha dobrého*. Za *Živur* dosazujeme do obou rovnic číslo 0 pro *nižší životní úroveň*, číslo 0,5 pro *střední životní úroveň* a číslo 1 pro *vyšší životní úroveň*. Obdobně dosazujeme do obou rovnic za *Polor* číslo 0 pro *levici*, číslo 0,5 pro *střed*, číslo 1 pro *pravici* a za *Třída* číslo 0 pro kategorii *nižší nebo dělnická*, číslo 0,5 pro kategorii *nižší střední* a číslo 1 pro kategorii *střední+*.

Nejprve posoudíme, zda je model kvalitní podle následujících ukazatelů, o jejichž významu jsme hovořili výše. *Explained* L^2 = 458,14, což při třech stupních volnosti dává dosaženou hladinu významnosti 5,6e-99. *Residual* L^2 = 57,36, čemuž při 49 stupních volnosti odpovídá dosažená hladina významnosti 0,19, takže model je přijatelný. Monotónní R^2 = 0,252 a φ = 0,638. Protože dosažená hladina významnosti u statistiky *Explained* L^2 je prakticky nulová, zamítáme hypotézu, že $R^2 = 0$, resp. že φ = 0. Monotónní regresní model vysvětluje 25,2 % variability v závislé proměnné. Hodnota Pearsonovy statistiky *Residual* χ^2 je 57,52. Je tedy velmi blízká hodnotě statistiky *Residual* L^2 a proto můžeme statistice *Residual* L^2 věřit.[25] Mezi 27 adjustovanými rezidui je sice pět nad 1,96 nebo pod (–1,96), konkrétně jsou to hodnoty 1,99, 2,04, 2,05, –2,19, –2,38, ale tím se nemusíme příliš znepokojovat.[26]

Rovnice (2. 12) a (2. 13) už s odhady regresních parametrů jsou

$$\ln[P(\textit{převaha špatného}) / P(\textit{převaha dobrého})] = \\ = -2{,}25 + [1{,}99(\textit{Živur} - 1) + 2{,}47(\textit{Polor} - 1) + 1{,}54(\textit{Třída} - 1)] \times (-1), \qquad (2.\ 14)$$

$$\ln[P(\textit{stejně špatného jako dobrého}) / P(\textit{převaha dobrého})] = \\ = -0{,}59 + [1{,}99(\textit{Živur} - 1) + 2{,}47(\textit{Polor} - 1) + 1{,}54(\textit{Třída} - 1)] \times (-0{,}5). \qquad (2.\ 15)$$

Dosažená hladina významnosti je u všech koeficientů prakticky nulová, jsou tedy všechny významně odlišné od nuly. Šance názoru *převaha špatného* u respondenta s nižší (resp. střední) životní úrovní je 7,3 krát (resp. 2,7 krát) větší než tatáž šance u respondenta s vyšší životní úrovní. Stejná šance u respondenta levicově (resp. středově) orientovaného je 11,8 krát (resp. 3,4 krát) větší než u respondenta pravicově orientovaného. Stejná šance u respondenta hlásícího se k nižší nebo dělnické třídě (resp. nižší střední třídě) je 4,7 krát (resp. 2,2 krát) větší než u respondenta, který se hlásí k vyšší třídě, než je nižší střední.[27]

Šance názoru *stejně špatného jako dobrého* u respondenta s nižší (resp. střední) životní úrovní je 2,7 krát (resp. 1,6 krát) větší než u respondenta s vyšší životní úrovní.

---

[25]) Pracujeme s 1631 případem a máme jen tři prediktory o třech kategoriích, nejde proto o řídká data.

[26]) Jednotlivě sice tato adjustovaná rezidua významná jsou, ale při simultánním testování by nebylo významné žádné z nich. Nicméně to můžeme brát jako výzvu k hledání lepšího modelu.

[27]) Činitel 7,3 = exp(1,99), činitel 2,7 = exp(0,5 × 1,99). Činitel 11,8 = exp(2,47), činitel 3,4 = exp(0,5 × 2,47). Činitel 4,7 = exp(1,54), činitel 2,2 = exp(0,5 × 1,54).

Stejná šance je u respondenta levicově (resp. středově) orientovaného 3,4 krát (resp. 1,9 krát) větší než u respondenta pravicově orientovaného. Stejná šance u respondenta řadícího se do nižší nebo dělnické třídy (resp. do nižší střední třídy) je 2,2 krát (resp. 1,5 krát) větší než u respondenta, který se řadí do vyšší třídy, než je nižší střední.[28]

Nyní budeme na stejných datech demonstrovat situaci, kdy kategoriím žádné ze zúčastněných proměnných nepřiřadíme skóry, ale necháme na metodě, aby je odhadla sama.[29] Za těchto podmínek je *Explained* L^2 = 482,49, což dává při sedmi stupních volnosti dosaženou hladinu významnosti 4,6e-100. *Residual* L^2 = 33,01, čemuž při 45 stupních volnosti odpovídá dosažená hladina významnosti 0,91, takže model je velmi dobrý, lepší než předchozí. *Monotónní* R^2 = 0,264, φ = 0,659. Protože dosažená hladina významnosti u statistiky *Explained* L^2 je skoro nulová, zamítáme hypotézu, že $R^2 = 0$ a rovněž hypotézu, že φ = 0. Z hodnoty R^2 vyvozujeme, že monotónní model vysvětluje 26,4 % variability v závislé proměnné. Žádné z adjustovaných reziduí nepřevyšuje v absolutní hodnotě číslo 1,96.

Kategoriím závislé proměnné byly přiřazeny tyto skóry: –1,00 (*převaha špatného*), –0,49 (*stejně špatného jako dobrého*), 0,00 (*převaha dobrého*). Kategoriím proměnné *Živur* byly přiřazeny tyto skóry: –1,00 (*nižší*), –0,25 (*střední*), 0,00 (*vyšší*). Kategoriím proměnné *Polor* byly přiřazeny skóry: –1,00 (*levice*), –0,72 (*střed*), 0,00 (*pravice*). Kategoriím proměnné *Třída* byly přiřazeny skóry: –1,00 (*nižší nebo dělnická*), –0,66 (*nižší střední*), 0,00 (*střední+*). Kategorie závislé proměnné jsou od sebe vzdáleny téměř stejně, což ale neplatí pro kategorie nezávislých proměnných.

Rovnice modelu jsou[30]

$$\ln[\mathrm{P}(\textit{převaha špatného}) / \mathrm{P}(\textit{převaha dobrého})] = \\ = -2{,}27 + [2{,}04\textit{Živur} + 2{,}37\textit{Polor} + 1{,}45\textit{Třída}] \times (-1), \qquad (2.\ 16)$$

$$\ln[\mathrm{P}(\textit{stejně špatného jako dobrého}) / \mathrm{P}(\textit{převaha dobrého})] = \\ = -0{,}55 + [2{,}04\textit{Živur} + 2{,}37\textit{Polor} + 1{,}45\textit{Třída}] \times (-0{,}49). \qquad (2.\ 17)$$

Odhady regresních parametrů z rovnic (2.16) a (2,17) jsou všechny vysoce významné, dosažená hladina významnosti u všech je prakticky nulová. Vysoce významné jsou i koeficienty β specifické pro prostřední kategorie nezávislých proměnných. Jejich významnost či nevýznamnost se zde musí speciálně zjišťovat, neboť neplyne automaticky z výsledků pro koeficienty uvedené v rovnicích (2. 16) a (2. 17). To je způsobeno tím, že skóry prostředních kategorií se nyní odhadují.

Když dosadíme do rovnice (2.16) po řadě skóry kategorií nezávislých proměnných *Živur*, *Polor* a *Třída* zjistíme, že šance názoru *převaha špatného* u respondenta s nižší (resp. střední) životní úrovní je 7,7 krát (resp. 1,7 krát) větší než tatáž šance u respondenta s vyšší životní úrovní. Stejná šance u respondenta levicově (resp. středově) orientovaného je 10,7 krát (resp. 5,5 krát) větší než u respondenta pravicově orientovaného. Stejná šance

[28]) Činitel 1,6 = exp(0,5 × 0,5 × 1,99). Činitel 1,9 = exp(0,5 × 0,5 × 2,47). Činitel 1,5 = exp(0,5 × 0,5 × 1,54). Výpočet činitelů 2,7, 3,4 a 2,2 je vysvětlen už v předchozí poznámce.

[29]) De facto půjde jen o odhad prostředních kategorií (viz odstavec nad rovnicí (2. 11)). Krajní hodnoty jsou voleny jako 0 a 1 nebo (–1) a 0.

[30]) Srovnejte s rovnicí (2. 11) a uvědomte si, že referenční kategorie jsou vždy ty poslední, kterým metoda přiřadila skór nula.

u respondenta hlásícího se k nižší nebo dělnické třídě (resp. nižší střední třídě) je 4,3 krát (resp. 2,6 krát) větší než u respondenta, který se hlásí k vyšší třídě, než je nižší střední.[31]

Když dosadíme do rovnice (2.17) po řadě skóry kategorií nezávislých proměnných *Živur*, *Polor* a *Třída* zjistíme, že šance názoru *stejně špatného jako dobrého* u respondenta s nižší (resp. střední) životní úrovní je 2,7 krát (resp. 1,3 krát) větší než u respondenta s vyšší životní úrovní. Stejná šance je u respondenta levicově (resp. středově) orientovaného 3,2 krát (resp. 2,3 krát) větší než u respondenta pravicově orientovaného. Stejná šance u respondenta řadícího se do nižší nebo dělnické třídy (resp. do nižší střední třídy) je 2,0 krát (resp. 1,6 krát) větší než u respondenta, který se řadí do vyšší třídy, než je nižší střední.[32][33]

Závěr

Snahou této stati bylo přesvědčit čtenáře o síle analýzy dat pomocí logistické regrese a přispět tak k jejímu většímu rozšíření i u nás. Možnosti použitých programů jsou širší, než zde bylo prezentováno, to se týká zejména programu GOLDMineR. Analytik jistě uvítá například možnost smysluplného spojování kategorií na základě blízkosti přiřazených skórů a regresních koeficientů, nebo možnost statistického porovnávání adekvátnosti různých modelů. Článek se také snažil upozornit na úskalí použití metody na řídká data, neboť bezmyšlenkovité a mechanické použití logistické regrese je stejně nebezpečné jako bezmyšlenkovité a mechanické použití lineární regrese či jakékoliv jiné techniky. Ne všechny možné problémy zde byly zmíněny. Nehovořili jsme například o multikolinearitě, která znehodnocuje výsledky logistické regrese stejně jako lineární regrese a kterou dokáže odhalit z použitých programů jen GOLDMineR. Jen letmo jsme se dotkli problematiky volby různých typů kontrastů a interakcí mezi nezávislými proměnnými. Jejich zahrnutí patří už mezi pokročilé aplikace, rozhodně bychom s nimi neměli začínat dříve, než zvládneme základy a než dobře pochopíme jejich smysl. Pak mohou být ovšem velmi užitečné.

BLANKA ŘEHÁKOVÁ *je vědeckou pracovnicí Sociologického ústavu AV ČR, kde je členkou výzkumného týmu „Sociální stratifikace". Věnuje se zejména problematice nerovností a proměnám vztahu mezi sociální třídou a volebním chováním.*

Literatura

Agresti, A. 1990. *Categorical Data Analysis.* New York: John Wiley and Sons.

Hauck, W. W., A. Donner 1977. „Wald's Test as Applied to Hypotheses in Logit Analysis." *Journal of the American Statistical Association* 72: 851-853.

Hosmer, D. W., Jr., S. Lemeshow 1989. *Applied Logistic Regression.* New York: John Wiley and Sons.

[31]) Činitel 7,7 = exp(2,04), činitel 1,7 = exp(2,04 × 0,25). Činitel 10,7 = exp(2,37), činitel 5,5 = exp(2,37 × 0,72). Činitel 4,3 = exp(1,45), činitel 2,6 = exp(1,45 × 0,66).

[32]) Činitel 2, 7 = exp(2,04 × 0,49), činitel 1,3 = exp(2,04 × 0,25 × 0,49). Činitel 3,2 = exp(2,37 × 0,49), činitel 2,3 = exp(2,37 × 0,72 × 0,49). Činitel 2,0 = exp(1,45 × 0,49), činitel 1,6 = exp(1,45 × 0,66 × 0,49).

[33]) Porovnáte-li výsledky modelu při stejně vzdálených kategoriích a modelu při nestejně vzdálených kategoriích, zjistíte určité numerické rozdíly v poměrech šancí, ale na základní interpretaci bez čísel se nic nemění.

Knoke, D., P. J. Burke 1980. „Log-Linear Models.“ *Sage University Paper series on Quantitative Applications in the Social Sciences*, 07-020. Beverly Hills, CA: Sage.

Magidson, J. 1998. *GOLDMineR™ 2.0. User's Guide.* Belmont, MA: Statistical Innovations Inc.

McFadden, D. 1973. „Conditional Logit Analysis of Qualitative Choice Behavior.“ Pp. 105-142 in *Frontiers in Econometrics*, ed. by P. Zarembka. New York: Academic Press.

Menard, S. 1995. „Applied Logistic Regression Analysis.“ *Sage University Paper series on Quantitative Applications in the Social Sciences*, 07-106. Thousand Oaks, CA: Sage.

Nagelkerke, N. J. D. 1991. „A Note on General Definition of the Coefficient of Determination.“ *Biometrika* 78: 691-692.

SPSS Regression Models™ 9.0. 1999. Chicago, IL: SPSS Inc.

SPSS Advanced Models™ 10.0. 1999. Chicago, IL: SPSS Inc.