MODEL LOGISTICKEJ REGRESIE PRE LONGITUDINÁLNE ÚDAJE

LOGISTIC REGRESSION MODEL FOR LONGITUDINAL DATA

Viera Labudová, Martina Lakatová

Abstract: *The objective of this paper is to describe particularity of longitudinal data and methods which can be used to analyse them. The assumption of usual tools used for analysis is the independence of observations. In order to analyse of longitudinal data, we have to make provisions for their particularity, which is the dependence of observations. Therefore, while we analyse them, we must employ methods that are adjusted to that dependence. Several approaches have been proposed to model binary outcomes that arise from longitudinal studies. Most of the approaches can be grouped into two classes: the population-averaged and subject-specific approaches. The generalized estimating equations (GEE) method is used to estimate population averaged effects. In this paper, we investigate the Generalized Estimating Equation (GEE) capabilities of PROC GENMOD for correlated outcome data to fit models using unspecified (unstructured) correlation structure. The data from EU SILC was used to find out how material deprivation of households in the Slovak Republic (material deprivation: yes (1), no (0)) is linked to their available characteristics.*

Keywords: *Longitudinal Data Analysis, Material Deprivation, Generalized Estimating Equation Model, EU SILC.*

JEL Classification*: C10, M31, O10.*

Úvod

Dizajn mnohých experimentálnych výskumov je založený na meraniach, ktoré sa opakujú na výskumnej vzorke v oddelených časoch, tzv. vlnách. Takýto druh merania má názov longitudinálny výskum. V rámci longitudinálnej štúdie sa zbierajú dáta aspoň v dvoch rozdielnych časových obdobiach. Predmetom skúmania môžu byť jednotlivci alebo skupiny, v rámci jednotlivých cyklov to môžu byť tí istí jednotlivci alebo aspoň im podobní. Cieľom takejto štúdie je zachytenie a skúmanie zmeny v čase a porovnanie medzi jednotlivými cyklami (Basl, 2007). Aplikácia časovej dimenzie je faktorom, ktorý longitudinálny výskum stavia do opozície k prierezovému výskumu (Babbie, 2010).

Modelovanie longitudinálnych dát si vyžaduje použitie iných prístupov ako pri analýze prierezových údajov. Hlavným cieľom tohto príspevku je ukázať možnosti modelovania hodnôt binárnej závislej premennej na základe údajov, ktoré boli získané longitudinálnym prieskumom.

Longitudinálne (panelové) údaje začali používať v akademickom výskume P. F. Lazarsfeld a M. Fiske (Lazarsfeld a Fiske, 1938; Lazarsfeld, 1940) v 40. rokoch minulého storočia a to v oblasti výskumu verejnej mienky. Potreba analyzovať longitudinálne údaje viedla k vývoju modelovacích techník, ktoré zohľadňujú ich špecifický charakter. Pri modelovaní binárnej závislej premennej sa používa zovšeobecnený lineárny model, ktorý bol prvý krát predstavený v práci Neldera a Wedderburna (1972). Možnosti použitia tohto modelu pri skorelovaných dátach jeho

rozšírením o tzv. pracovnú korelačnú maticu, ktorá kvantifikuje závislosti medzi pozorovaniami, opísali ako prví Liang a Zeger (1986) a Zeger a Liang (1986). Do modelovania závislej premennej zaviedli prístup, ktorý sa uvádza pod skratkou GEE – *generalized estimating equations*, alebo sa používa pomenovanie marginálny model, resp. populačne-spriemerovaný model. Ďalšie podrobnosti nielen o triede populačne-spriemerovaných modelov, ale aj o objektovo-špecifických modeloch prinášajú Zeger et al. (1988) a Neuhaus et al. (1991). Opis logistického modelu s náhodnými efektmi, ktorý sa používa pri odhade vplyvu hlavných efektov v objektovo-špecifických modeloch možno nájsť v prácach autorov Stiratelli et al. (1984), Wong a Mason (1985), Lee a Nelder (1996) a Hu et al. (1998). Autori Snijders a Bosker (1999) a Mason (2001) poukázali, v súvislosti s analýzou longitudinálnych údajov, na dôležitosť použitia takých modelovacích techník, ktoré zohľadňujú závislosť meraní. Svojím výskumom potvrdili, že nerešpektovanie podmienky nezávislosti pri aplikácii analytických nástrojov vedie k neefektívnym a často aj ku skresleným odhadom parametrov modelu. Rôzne prístupy k analýze panelových dát v štatistickom programe SAS uvádza vo svojej práci Paul Allison (1999), ktorý praktickými ukážkami ilustruje hlavné rozdiely medzi marginálnymi modelmi, modelmi s fixovanými efektmi a modelmi s robustnými štandardnými chybami.

1 Formulácia problematiky

1.1 Dizajn longitudinálneho výskumu

Longitudinálny výskum má tri druhy dizajnov: opakovaný prierezový výskum, panelový výskum a kohortný výskum.

Opakovaný prierezový výskum (repeated cross-sectional analysis), resp. výskum trendu sa uskutočňuje v rôznych časových obdobiach a to vždy na inej vzorke. Vzhľadom na to, že jeho zameraním je porovnanie výsledkov získaných v rôznych časových obdobiach, viaže sa k jednej téme, má teda to isté zameranie (Ruspini, 2002; Šubrt, 2013).

Kohortný výskum (cohort analysis) je analýza udalostí, ktoré nastali v tej istej kohorte alebo generácii.

Panelový výskum (panel analysis) je špeciálny dizajn longitudinálneho výskumu. Jeho špecifickým znakom je to, že sa pri ňom opakovane zbierajú informácie na tej istej výberovej vzorke v rôznych časových obdobiach. Longitudinálne údaje získané z panelových štúdií možno použiť na analýzu krátkodobej dynamiky zmien, napr. pohybov do a von z trhu práce, prechodov do a von z chudoby alebo na sledovanie procesu demografických zmien. Možno ich použiť aj na skúmanie dlhodobých účinkov, napr. vplyvu vzdelávania na trh práce, závislosti študijných výsledkov detí v škole a neskôr v ich samostatnom živote v závislosti od rodinného zázemia, sledovanie vzťahu medzi zdravotným stavom človeka a jeho spôsobom života (Laurie, 2013).

Okrem týchto troch dizajnov longitudinálneho výskumu možno uskutočniť aj hybridný výskum, ktorý je kombináciou panelového výskumu a opakovaného prierezového zisťovania (Kalvas, 2003; Lechnerová, 2009).

1.2 Longitudinálne dáta, panelové dáta

V praxi sa často používa ako ekvivalent longitudinálneho prieskumu jeho dizajnová podoba panelový prieskum. Toto nahradenie všeobecného tvaru jednou z jeho

špecifických podôb spôsobuje množstvo nedorozumení a nejasností. Na druhej strane sa akceptuje použitie pojmu longitudinálne dáta ako ekvivalentu k pojmu panelové dáta. V ďalšej časti budeme používať pojem longitudinálne, resp. panelové dáta ako výsledok panelového zisťovania.

Longitudinálne, niekedy nazývané panelové dáta, sú údaje, ktoré boli získané na tom istom výberovom súbore v rôznych časových obdobiach. Jednotkami súboru môžu byť jednotlivci, domácnosti, podniky, atď.

Na longitudinálne dáta možno nazerať tiež ako na zhluky dát, resp. špeciálny prípad zhlukovo skorelovaných dát (cluster-correlated data). Zhluky môžu byť výsledkom prirodzenej hierarchickej štruktúry populácie, alebo sú výsledkom dizajnu štúdie, prípadne sa pri vytváraní zhlukov v dátach aplikujú obidva aspekty.

Longitudinálne údaje majú v porovnaní s prierezovými údajmi, alebo s údajmi časových radov niekoľko výhod. Umožňujú presnejšie odhady parametrov modelu a to z dôvodu väčšieho počtu pozorovaní. Modely využívajúce panelové dáta umožňujú kontrolovať vplyv vynechaných, resp. nepozorovaných premenných, ktoré sa nemenia ani v čase ani v rámci prierezových dát (Hsiao, 2005).

Za hlavné nevýhody modelov, ktoré sú založené na longitudinálnych údajoch, možno považovať predovšetkým problémy, ktoré sú spojené so zberom údajov. Tie sú dané samotnou charakteristikou panelu zisťovania a hlavne podmienkou, aby sa počas celého obdobia zisťovania nemenil súbor štatistických jednotiek. Ďalším dôležitým predpokladom longitudinálneho výskumu je aj vytvorenie stabilného výskumného tímu, ktorý musí ostať zachovaný počas celej doby výskumu (Bijleveld, 1998).

2 Metódy

2.1 Metódy analýzy longitudinálních údajov

Pre analýzu longitudinálnych údajov (údajov so zhlukovou štruktúrou) existuje niekoľko rôznych analytických metód. Opakované merania možno analyzovať pomocou analýzy rozptylu. Jej použitie však naráža na množstvo obmedzujúcich podmienok (Ballinger, 2004).

Pri modelovaní hodnôt závislej premennej na základe longitudinálnych údajov treba brať do úvahy skutočnosť, že výsledky opakovaných meraní na tom istom objekte majú tendenciu byť skorelované (Zeger, Liang, Albert, 1988).

Pri modelovaní sa využívajú dva rozdielne prístupy: objektovo-špecifický (subject-specific) a populačne spriemerovaný prístup (population-averaged) (Zeger, Liang, Albert, 1988). Objektovo-špecifický prístup používa model s náhodnými efektami a modely zmiešaných efektov (mixed-effect models) (Crowder, 1995). Model zmiešaných efektov zachytáva závislosti medzi pozorovaniami, ktoré boli získané na tej istej jednotke. Populačne spriemerovaný prístup využíva GEE metóda (The generalized estimating equation) (Crowder, 1995). GEE metóda poskytuje priemerné hodnoty odhadov parametrov pre celú populáciu (Crowder, 1995, Zorn, 2001).

GEE metóda (GEE prístup) bola vyvinutá a opísaná autormi Liang a Zeger (Liang, Zeger, 1986), (Zeger, Liang, 1986). Jej využitie súviselo s testovaním vplyvu faktorov (nezávislých premenných) na premenné s rozdelením pravdepodobnosti patriacim do množiny rozdelení exponenciálneho typu (Ballinger, 2004). GEE modelovanie je rozšírením zovšeobecneného lineárneho modelu (Nelder, Wedderburn, 1972) a to o

zahrnutie korelácií medzi hodnotami závislej premennej. Na základe tohto prístupu bol vyvinutý marginálny model, resp. model populačného priemeru.

V článku najskôr veľmi zjednodušene predstavíme model logistickej regresie pre dáta, v ktorých predpokladáme nezávislosť pozorovaní. V ďalšej časti zavedieme jeho tvar pre prípad longitudinálnych údajov.

2.1.1 Zovšeobecnený lineárny model GLZ (Generalized Linear Model)

Model logistickej regresie je špeciálnym prípadom zovšeobecneného lineárneho modelu (Generalized Linear Model). Ten možno použiť v prípade, ak má vysvetľovaná premenná aj iné ako normálne rozdelenie pravdepodobnosti, napr. binomické, Poissonovo, exponenciálne, gama rozdelenie.

Zovšeobecnený lineárny model sa skladá z troch častí: z náhodnej zložky, identifikujúcej rozdelenie pravdepodobnosti závislej premennej, zo systematickej zložky, ktorá špecifikuje lineárnu funkciu vysvetľujúcich premenných a z väzbovej funkcie, ktorá opisuje funkčný vzťah medzi náhodnou zložkou a systematickou zložkou.

Náhodná zložka zovšeobecneného lineárneho modelu je tvorená vektorom nezávislých náhodných premenných $\mathbf{Y}=\left(Y_1, Y_2,...Y_n\right)$. Každá nezávislá premenná Y_i, $i=1, 2,...n$, má rozdelenie pravdepodobnosti zo skupiny rozdelení exponenciálneho typu.

Systematickou zložkou je vektor $\boldsymbol{\eta}=\left(\eta_1,\eta_2,...\eta_n\right)^T$, ktorý môžeme vyjadriť takto:

$$\boldsymbol{\eta}=\mathbf{X}\boldsymbol{\beta} \tag{1}$$

kde $\mathbf{X}$ je matica typu $n\times\left(p+1\right)$ obsahujúca pozorované hodnoty vysvetľujúcich premenných $X_1, X_2,... X_p$ na objektoch $\mathrm{O}_1, \mathrm{O}_2,...\mathrm{O}_n$, pričom prvý stĺpec matice obsahuje iba jednotky a $\boldsymbol{\beta}$ je $\left(p+1\right)$-prvkový vektor neznámych parametrov modelu $\boldsymbol{\beta}=\left(\beta_0,\beta_1,...\beta_k\right)^T$. Vektor $\boldsymbol{\eta}$ sa nazýva lineárny prediktor.

Väzbová, resp. spojovacia funkcia, spájajúca náhodnú a systematickú zložku modelu, je poslednou časťou GLZ. Väzbovou funkciou je funkcia g podmienenej strednej hodnoty vysvetľovanej náhodnej premennej $E\left(y_i\right)=\mu_i=E\left(Y|\mathbf{x}_i\right)$, ktorá vyhovuje nasledujúcej požiadavke

$$g\left(E\left(y_i\right)\right)=g\left(\mu_i\right)=\boldsymbol{\eta}_i=\mathbf{x}_i\boldsymbol{\beta},\ i=1, 2,...n \tag{2}$$

kde $\mathbf{x}_i$ je vektor hodnôt vysvetľujúcich premenných zodpovedajúcich objektu O_i. Použitím väzbovej funkcie logit dostávame model logistickej regresie

$$g\left(\mu_i\right)=ln\frac{\mu_i}{1-\mu_i}=\mathbf{x}_i\boldsymbol{\beta} \tag{3}$$

2.1.2 Rozšírenie zovšeobecneného lineárneho modelu pre longitudinálne dáta

V ďalšom budeme predpokladať, že merania hodnôt nezávislých premenných $X_1, X_2,... X_p$ [1] a hodnôt závislej premennej Y boli opakované t_i krát, $1\le t_i\le t$.

[1] Vstupné premenné môžu byť časovo nezávislé (napr. pohlavie), alebo ich hodnoty sa môžu meniť s časom (napr. príjem, počet členov domácnosti).

Výsledkom sú pozorovania $\left(y_{ij}, \mathbf{x}_{ij}\right)$ pre objekty $i = 1, 2,...n$ a obdobia t_{ij} , $j = 1, 2,...t_i$ [2], kde y_{ij} je hodnota závislej premennej meraná na i-tom objekte v čase j a $\mathbf{x}_{ij} = \left(\mathbf{x}_{ij1}, \mathbf{x}_{ij2}...\mathbf{x}_{ijp}\right)^T$ je $\left(p \times 1\right)$- rozmerný vektor hodnôt vysvetľujúcich premenných pre objekt i v čase j. Na každom objekte O_i je teda uskutočnených t_i meraní. Nech $\mathbf{y}_i$ je $\left(t_i \times 1\right)$-rozmerný vektor $\left(y_{i1}, y_{i2},...y_{it_i}\right)^T$ a $\mathbf{x}_i$ je $\left(t_i \times p\right)$-rozmerná matica $\mathbf{x} = \left(\mathbf{x}_{i1}, \mathbf{x}_{i2}...\mathbf{x}_{it_i}\right)^T$, opisujúca hodnoty nezávislých premenných pre i-ty objekt.

Nech je i-ty objekt opísaný vektorom hodnôt závislej premennej $\mathbf{y_i} = \left(y_{i1}, y_{i2},...y_{it_i}\right)^T$ a prislúchajúcim vektorom stredných hodnôt $\boldsymbol{\mu}_\mathbf{i} = \left(\mu_{i1}, \mu_{i2},...\mu_{it_i}\right)^T$, kde μ_{ij} je stredná hodnota závislej premennej Y_i v čase j. Premenné Y_i sú nezávislé medzi jednotlivými prípadmi a skorelované medzi jednotlivými obdobiami (vnútri objektu). GLM model pre longitudinálne dáta sa líši od modelu pre merania, medzi ktorými nie je korelácia len tým, že sa musí navyše odhadovať kovariančná (resp. korelačná) štruktúra skorelovaných meraní. Marginálny model špecifikujúci vzťah medzi funkciou strednej hodnoty μ_{ij} a vektorom hodnôt vysvetľujúcich premenných $\mathbf{x}_{ij}$ má tvar (Allison, 2009; 2012)

$$g\left(\mu_{ij}\right) = \mathbf{x}_{ij}^T \boldsymbol{\beta} \tag{4}$$

kde g je známa väzbová funkcia (v prípade alternatívnej závislej premennej je to funkcia logit) a $\boldsymbol{\beta}$ je $\left(p \times 1\right)$- rozmerný vektor hodnôt neznámych parametrov modelu.

3 Rozbor problému

Na Slovensku, podobne ako v ďalších krajinách Európskej únie, sa vykonáva každý rok štatistické zisťovanie EU SILC, ktoré je zamerané na príjmy a životné podmienky domácností. Zisťovanie v sebe zahŕňa okrem prierezovej zložky aj longitudinálnu zložku. V praktickej časti článku je využitá práve longitudinálna zložka tohto zisťovania na identifikáciu faktorov, ktoré štatisticky významne vplývajú na výskyt materiálnej deprivácie v domácnostiach Slovenska.

Za materiálne deprivovanú domácnosť sa v Európskej únii považuje domácnosť, ktorá čelí vynútenému nedostatku aspoň v troch z deviatich nasledovných deprivačných položiek, ktoré si nemôže finančne dovoliť: čeliť neočakávaným výdavkom, ísť raz za rok na 1 týždeň dovolenky mimo domov, uhrádzať nedoplatky spojené s hypotékou alebo nájomným, úhradou za energie alebo splácaním nákupov na splátky a iných pôžičiek, jesť jedlo s mäsom, kurčaťom alebo rybou každý druhý deň, udržiavať primerané teplo v byte, alebo si nemôže finančne dovoliť, aj keby chcela (ide o tzv. vynútený nedostatok): práčku, farebný televízor, telefón alebo automobil.

Pri analýze sme použili databázu individuálnych údajov zisťovania EU SILC (UDB verzia 23/09/2016), ktorú nám pre potreby výskumných analýz poskytol Štatistický úrad Slovenskej republiky. Z longitudinálnej zložky tejto databázy sme použili údaje

[2] Ak je počet meraní pre všetky objekty rovnaký a vzdialenosti medzi jednotlivými časovými obdobiami sú rovnaké, potom t je celkový počet období, v ktorých bolo uskutočnené meranie.

o domácnostiach, ktoré boli opakovane sledované a opytované v rokoch 2012, 2013, 2014 a 2015.

EU SILC sa v rámci Slovenska realizuje ako integrovaná („rotačná“) forma zisťovania so štyrmi čiastkovými súbormi (Obr. 1). Na začiatku (rok 2012) bola vybraná reprezentatívna vzorka domácností a tá sa rozdelila na štyri podsúbory (replikácie 1, 2, 3, 4). Každý z nich reprezentuje celý základný súbor a to tak, že štruktúra každého z týchto podsúborov je podobná štruktúre základného súboru. V nasledujúcich rokoch zisťovania (roky 2013, 2014 a 2015) sa vždy jeden z podsúborov nahradil novým podsúborom. Na obrázku (Obr. 1) sú nové podsúbory zobrazené ako svetlo vyfarbené bunky. Čísla v jednotlivých bunkách určujú, koľký krát je daný podsúbor súčasťou výberovej vzorky. Podsúbor (replikácia 4), na ktorom bolo robené zisťovanie vo všetkých štyroch rokoch je označený na obrázku oválnym obdĺžnikom.

Obr. 1: Rotačný dizajn EU SILC

Replikácia vzorky	2012	2013	2014	2015
1	1			
2	1	2		
3	1	2	3	
4	1	2	3	4
5		1	2	3
6			1	2
7				1

Zdroj:EU statistics on income and living conditions (EU-SILC) methodology – sampling, Gerbery (2011), upravené autormi

Dátový súbor, ktorý sme následne analyzovali, obsahoval 1 234 domácností, čo predstavuje 4 936 pozorov aní za štyri sledované roky.

Na identifikáciu faktorov ovplyvňujúcich stav materiálnej deprivácie domácností sme využili model logistickej regresie. Ako vstupné premenné boli v tomto modeli použité hlavne premenné, ktoré opisujú vlastnosti domácnosti[3]: pod hranicou rizika chudoby[4] (CHUDOBA – ARPT60i), stupeň urbanizácie[5] (URBAN – DB100), región (REGION – DB040)[6], typ domácnosti (TYP_DOM – HT) a vlastnícky status (VLASTN – HH021)[7]. K osobe stojacej na čele domácnosti sa vzťahovala iba jedna vstupná premenná pohlavie[8] (POHLAVIE – RB090). Pôvodné hodnoty premenných región, typ domácnosti a vlastnícky status boli pred samotnou analýzou upravené zlúčením a prekódovaním kategórií. Upravené hodnoty premennej typ domácnosti sú v Tab. 1.

[3] V zátvorke sú uvedené pracovné skratky premenných, ktoré boli využité pri práci so softvérom a za pomlčkou je označenie premenných prebraté z použitej databázy.

[4] U binárnej premennej pod hranicou rizika chudoby označovala hodnota 1 stav, keď je domácnosť ohrozená rizikom finančnej chudoby.

[5] Premenná stupeň urbanizácie nadobúdala hodnotu 1,ak domácnosť žila na území s hustým osídlením, hodnotu 2 pre územia s mierne hustým osídlením a hodnotu 3 pre územia s riedkym osídlením.

[6] Pri premennej región bolo použité takéto kódovanie: Východné Slovensko (1) , Stredné Slovensko (2) a Bratislavský kraj a Západné Slovensko (3).

[7] Premenná vlastnícky status má v tejto analýze takéto kategórie: ubytovanie poskytované bezplatne (1), nájomník/ podnájomník platiaci bežné nájomné alebo nájomné za trhovú cenu a ubytovanie prenajímané za zníženú cenu (2), majiteľ a vlastník platiaci hypotéku (3).

[8] Binárna premenná *pohlavie* nadobúda hodnotu 1, ak stojí na čele domácnosti muž a hodnotu 0, ak je na čele domácnosti žena.

Tab. 1: Kódovanie zlúčených kategórií premennej typ domácností

Typ domácnosti	**kódovanie**
domácnosť 2 dospelých bez závislých detí - obaja vo veku pod 65 rokov, domácnosť 2 dospelých bez závislých detí - aspoň jeden dospelý vo veku 65 rokov a viac	1
ostatné domácnosti bez závislých detí, ostatné domácnosti so závislými deťmi	2
domácnosť s 1 rodičom a s 1 alebo viac závislými deťmi, domácnosť 2 dospelých s 3 alebo viac závislými deťmi	3
domácnosť 2 dospelých s 1 závislým dieťaťom, domácnosť 2 dospelých s 2 závislými deťmi	4
jednočlenná domácnosť	5

Zdroj: EU SILC – verzia 26/09/2016,vlastné spracovanie

Do analýz boli zahrnuté aj premenné obdobie (OBDOBIE – RB010) a identifikačné číslo domácnosti (ID – RB0303). Modelovanou premennou bola premenná materiálna deprivácia (MD), ktorá nadobúdala dve obmeny: MD = 1, ak bola domácnosť materiálne deprivovaná, MD = 0, ak domácnosť nebola materiálne deprivovaná.

Odhady hodnôt parametrov modelu logistickej regresie sme uskutočnili v programe SAS Base pomocou procedúry GENMOD. Uvedená procedúra vyžaduje definovanie pracovnej korelačnej matice obsahujúcej hodnoty koeficientov korelácie medzi hodnotami premennej *materiálna deprivácia*, ktoré boli zistené v rôznych rokoch. GEE metódou boli odhadnuté dva modely. V jednom bola použitá autoregresná a v druhom neštruktúrovaná korelačná pracovná matica. Prvky týchto matíc vyjadrujú silu vzťahu medzi hodnotami premennej materiálna deprivácia v jednotlivých rokoch zisťovania (Tab. 2).

Tab. 2: Výstup procedúry GENMOD – neštruktúrovaná pracovná korelačná matica (UN), autoregresná pracovná korelačná matica (AR)

Working Correlation Matrix (UN)				
obdobie	**2012**	**2013**	**2014**	**2015**
2012	1	0,5697	0,5533	0,4479
2013	0,5697	1	0,6446	0,5429
2014	0,5529	0,6446	1	0,6707
2015	0,4479	0,5433	0,6707	1

Working Correlation Matrix (AR)				
obdobie	**2012**	**2013**	**2014**	**2015**
2012	1	0,6189	0,3831	0,2371
2013	0,6189	1	0,6189	0,3831
2014	0,3831	0,6189	1	0,6189
2015	0,2371	0,3831	0,6189	1

Zdroj: EU SILC – verzia 26/09/2016, SAS Base,vlastné spracovanie

Hodnoty odhadnutých parametrov modelu obsahuje Tab. 3.

Tab. 3: Výstup procedúry GENMOD pre dva typy pracovnej matice (TYPE = UN), (TYPE = AR)

Parameter		Neštruktúrovaná matica (TYPE = UN)					Autoregresná matica (TYPE = AR)				
		Est	Standard Error	Z	Pr > |Z|	OR	Est	Standard Error	Z	Pr > |Z|	OR
Intercept		-0,3116	0,1981	-1,57	0,1158	0,73	-0,1834	0,1991	-0,92	0,3568	0,83
REGION	1	0,2734	0,1313	2,08	0,0373	1,31	0,2626	0,1319	1,99	0,0465	1,30
REGION	2	0,4052	0,1317	3,08	0,0021	1,50	0,3921	0,1326	2,96	0,0031	1,48
REGION	3	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
TYP_DOM	1	-0,5949	0,1415	-4,20	0,0001	0,55	-0,5744	0,1459	-3,94	0,0001	0,56
TYP_DOM	2	-1,0001	0,1428	-7,01	0,0001	0,38	-1,0393	0,1446	-7,19	0,0001	0,35
TYP_DOM	3	-0,5475	0,1887	-2,90	0,0037	0,58	-0,5692	0,1926	-2,96	0,0031	0,57
TYP_DOM	4	-1,1732	0,1627	-7,21	0,0001	0,31	-1,2187	0,1665	-7,32	0,0001	0,30
TYP_DOM	5	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
VLASTN	1	0,4527	0,1755	2,58	0,0099	1,57	0,4641	0,1788	2,59	0,0095	1,59
VLASTN	2	0,2647	0,2662	0,99	0,3199	1,30	0,3292	0,2925	1,13	0,2604	1,39
VLASTN	3	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
CHUDOBA	0	-0,6095	0,1250	-4,88	0,0001	0,54	-0,7264	0,1244	-5,84	0,0001	0,48
CHUDOBA	1	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
POHLAVIE	0	0,8306	0,1192	6,97	0,0001	2,29	0,8140	0,1201	6,78	0,0001	2,26
POHLAVIE	1	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
URBAN	1	-0,5190	0,1588	-3,27	0,0011	0,60	-0,5119	0,1605	-3,19	0,0014	0,60
URBAN	2	-0,1726	0,1212	-1,42	0,1543	0,84	-0,1682	0,1205	-1,40	0,1626	0,85
URBAN	3	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00
OBDOBIE	1	0,3890	0,0699	5,56	0,0001	1,48	0,3862	0,0705	5,47	0,0001	1,47
OBDOBIE	2	0,2853	0,0629	4,53	0,0001	1,33	0,2815	0,0634	4,44	0,0001	1,33
OBDOBIE	3	0,2052	0,0537	3,83	0,0001	1,23	0,2036	0,0541	3,76	0,0002	1,23
OBDOBIE	4	0,0000	0,0000	.	.	1,00	0,0000	0,0000	.	.	1,00

Zdroj: EU SILC – verzia 26/09/2016, SAS Base, vlastné spracovanie

V tabuľke (Tab. 3) sú okrem hodnôt bodových odhadov (Est), hodnoty štandardných chýb odhadu týchto parametrov (Standard Error), hodnota testovacej štatistiky (Z) a p-hodnota (Pr > |Z|) testu štatistickej významnosti parametrov.

Pri interpretácii hodnôt odhadnutých parametrov sa v prípade modelu logistickej regresie využívajú pomery šancí (Odds Ratio). Tie vyjadrujú aký je pomer šance domácnosti byť deprivovanou, ak je zaradená podľa niektorej sledovanej vlastnosti (napríklad typ domácnosti) do príslušnej kategórie v porovnaní so šancou byť deprivovanou, ak by bola na základe tejto vlastnosti zaradená do tzv. referenčnej kategórie tejto premennej[9]. Ak zoberieme do úvahy výsledky modelu, ktorý bol vytvorený použitím autoregresnej pracovnej korelačnej matice, potom môžeme konštatovať, že najväčšie rozdiely sú medzi kategóriou jednočlenných domácností (referenčná kategória) a ostatnými kategóriami, do ktorých boli jednotlivé domácnosti zaradené na základe ich zloženia (Tab.1). Šanca domácností dvoch dospelých s jedným alebo dvoma závislými deťmi je 3,33-krát vyššia (1/0,30) a v skupine

[9] Referenčnou kategóriou bola kategória, ktorej číselné kódovanie malo najvyššiu hodnotu. Predpokladáme pritom, že hodnoty ostatných premenných sú u porovnávaných skupín domácností rovnaké.

domácností dvoch dospelých bez závislých detí je 1,79-krát vyššia (1/0,56) ako šanca materiálnej deprivácie u jednočlenných domácností. Ostatné domácnosti bez ohľadu na to, či v nich žijú nezaopatrené deti majú šancu byť materiálne deprivovanými, ktorá je na úrovni 0,35 takejto šance vyčíslenej pre jednočlenné domácnosti.

Pozornosť si zaslúži porovnanie šancí materiálnej deprivácie domácností, ktorých rôzny sociálny status sa prejavuje aj cez ich podmienky alebo možnosti bývania. Domácnosti, ktorým je ubytovanie poskytované bezplatne majú šancu byť materiálne deprivovanými 1,59-krát vyššiu ako domácnosti, ktoré vlastnia nehnuteľnosť určenú na bývanie (referenčná kategória). Skupina domácností, ktoré žijú v podnájme zahŕňa aj nájomníkov žijúcich v sociálnych bytoch, ktorým je poskytované bývanie za znížený nájom. Šanca domácností žijúcich v podnájme, že budú materiálne deprivované, je 1,39-krát vyššia ako u vlastníkov bývania.

Miery materiálnej deprivácie by mali byť komplementárne ku mieram chudoby. Príjmová chudoba a materiálna deprivácia sú totiž dva koncepty, ktorých spojením možno analyzovať životné podmienky domácností z rôznych aspektov. Ak totiž vnímame chudobu ako stav deprivácie, potom meranie chudoby možno považovať za meranie deprivácie. Vzťah medzi obidvoma fenoménmi možno identifikovať aj na základe parametra, ktorý kvantifikuje rozdiely medzi rodinami, ktoré sú ohrozené rizikom príjmovej chudoby a tými, ktoré ohrozenými nie sú vzhľadom na ich stav materiálnej deprivácie. Domácnosti, ktoré sú ohrozené rizikom chudoby majú približne 2-krát (1/0,48) vyššiu šancu, že budú materiálne deprivované, ako domácnosti, ktoré nie sú rizikom chudoby ohrozené.

Štatisticky významný je aj rozdiel medzi šancou domácností, na čele ktorých stojí žena a šancou domácností, na čele ktorých stojí muž. Šanca byť materiálne deprivovanou je u domácností, na čele ktorých stojí žena 2,26-krát vyššia ako u domácností na čele s mužom. Do skupiny domácností, ktoré uvádzajú, že ich prednostom je žena, patria totiž jednočlenné domácnosti a domácnosti osamelo žijúcich matiek s deťmi, ktoré patria z hľadiska výskytu materiálnej deprivácie k najviac ohrozeným.

Vplyv oblasti, v ktorej žijú domácnosti sme sledovali prostredníctvom premenných región a urbanizácia. Až 1,48-krát vyššiu šancu, že budú materiálne deprivované, majú domácnosti Stredného Slovenska a 1,3-krát vyššiu šancu domácnosti Východného Slovenska v porovnaní s domácnosťami žijúcimi v Bratislavskom kraji alebo v regióne Západného Slovenska.

Šanca materiálnej deprivácie u domácností, ktoré žijú v oblastiach s hustým osídlením je približne 1,67-krát vyššia (1/0,60) ako u domácností žijúcich v oblastiach s riedkym osídlením.

Získané výsledky podporujú naše hypotézy o skupinách domácností, najviac ohrozených materiálnou depriváciou, ktoré sme získali predchádzajúcimi analýzami.

4 Diskusia

Výsledky modelu logistickej regresie vytvoreného na longitudinálnych údajoch, ak odhliadneme od možnosti rôzneho kódovania hodnôt vstupných premenných, závisia od spôsobu odhadu parametrov. V našom prípade boli tieto možnosti limitované aj typmi procedúr, ktoré možno použiť na odhad hodnôt parametrov v programe SAS. Ďalšie rozdiely môžu byť spôsobené typom použitej pracovnej korelačnej matice.

V príspevku sú zverejnené výsledky odhadu pre model s využitím neštruktúrovanej matice a model, vytvorený použitím autoregresnej matice. Okrem týchto pracovných korelačných matíc, ktorých definovanie podmieňuje GEE metóda odhadu, možno použiť nezávislú, konvertibilnú alebo Toeplitzova maticu (podrobnosti napr. v Liang, Zeger, 1986). Rozdiely vo výsledkoch odhadov závisia aj od použitej procedúry štatistického programu. V práci bola použitá procedúrou GENMOD programu SAS Base. Okrem nej možno v SASe použiť aj procedúry LOGISTIC a SURVEYLOGISTIC. Pri interpretácii výsledkov modelu sme uprednostnili model, pri odhade ktorého bola použitá autoregresná korelačná štruktúra.

Hodnotenie kvality modelov odhadnutých GEE metódou je založené na iných kritériách ako sú tie, ktoré sa štandardne používajú pri modeloch logistickej regresie, odhadnutých na prierezových údajoch. Pri GEE metóde sa nepoužíva vierohodnostná funkcia, preto nemôžeme na posúdenie kvality modelu použiť Akaikeho informačné kritérium alebo Bayesovské informačné kritérium. Alternatívou k týmto mieram je QIC kritérium (Quasilikelihood under the Independence model Criterion), alebo QICu kritérium. Podrobnosti o spôsobe ich výpočtu a odporúčaniach pre ich použitie uvádzajú napr. Hardin a Hilb (2003). Lepší je model s nižšou hodnotou QIC. Vzhľadom na hodnoty tohto kritéria pre model s autoregresnou korelačnou štruktúrou (QIC = 5 276,65) a model s neštruktúrovanou korelačnou maticou (QIC = 5 295,95) bol ako lepší model, ktorého parametre boli následne interpretované, vybraný model s autoregresnou korelačnou štruktúrou.

Záver

Metodika odhadu parametrov lineárneho regresného modelu, v prípade číselnej spojitej premennej, alebo modelu logistickej regresie, ak je závislá premenná binárna, sa štandardne používa pri práci s prierezovými zložkami databáz. Problémy vznikajú vtedy, ak sa tie isté metódy odhadu používajú aj pri práci s longitudinálnymi údajmi. Ignorovanie toho, že výsledky opakovaných meraní na tom istom objekte majú tendenciu byť skorelované a nezačlenenie tejto skutočnosti do modelu môže viesť k odhadom parametrov, ktoré nie sú, hlavne pri silných závislostiach, dostatočne výdatné. Ambíciou tohto článku bolo poukázať na špecifiká odhadu parametrov modelu logistickej regresie, vytvoreného na základe longitudinálnych údajov.

V článku sme sa venovali technikám, ktoré sa používajú na modelovanie binárnej premennej s využitím longitudinálnych údajov. V praxi sa pri takýchto analýzach používajú dva rozdielne prístupy: objektovo-špecifický, alebo populačne-spriemerovaný prístup. Populačne- spriemerovaný prístup využíva GEE metóda, ktorá poskytuje priemerné hodnoty odhadov parametrov modelu pre celú populáciu. Túto metódu sme použili na testovanie vplyvu faktorov (charakteristík slovenských domácností a osôb stojacich na čele týchto domácností) na premennú materiálna deprivácia. Údaje pochádzali z longitudinálnej zložky databázy EU SILC pre roky 2012 až 2015. Analýza bola urobená v štatistickom programe SAS Base.

Hoci sa zisťovanie EU SILC realizuje na Slovensku už od roku 2005, nie sú k dispozícii analytické štúdie, ktoré by využívali pri modelovaní longitudinálnu zložku tejto databázy. Analýzu tejto databázy poskytuje len práca *Pretrvávajúca chudoba – analýza longitudinálnej databázy EU SILC* (Gerbery, 2011). Tá však využíva len nástroje opisnej štatistiky. Za hlavný prínos tohto článku preto považujeme tak teoretický výklad problematiky, ktorá je v našich podmienkach ešte stále na okraji

záujmu analytikov, ako aj prezentáciu výsledkov získaných praktickou aplikáciou opísaných metód. Model, ktoré sme získali špeciálnou metódou odhadu parametrov modelu logistickej regresie, zohľadňujúc osobitný charakter longitudinálnych údajov, poskytuje zaujímavé výsledky, ktoré umožňujú vytvoriť profil slovenských domácností vzhľadom na pretrvávajúci stav materiálnej deprivácie v rokoch 2012 až 2015.

Poďakovanie

Tento článok bol spracovaný s podporou výskumného projektu: VEGA č. 1/0770/17: Dostupnosť bývania na Slovensku.

Referencie

Allison, P. (2009). *Fixed Effects Regression Methods for Longitudinal Data Using SAS*. 2. vyd. USA: SAS Institute Inc.

Allison, P. (2012). *Logistic Regression Using SAS*. 1. vyd. USA: SAS Institute Inc.

Babbie, E. R. (2010). *The Practice of Social Research*. 12th ed. Belmont: Wadsworth.

Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. *Article in Organizational Research Methods*, 7, (2), 127 – 150. DOI: 10.1177/1094428104263672

Basl, J. (2007). British Household Panel Survey jako příklad longitudinálního výzkumu [online]. *Socioweb*, (6), 8 – 9. Dostupné z: http://www.socioweb.cz. [14. 07. 2017].

Bijleveld, C. et al. (1998). *Longitudinal Data Analysis: Designs, Models and Method.* UK: SAGE Publications Ltd.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, 82, (2), 407 – 410. DOI: org/10.1093/biomet/82.2.407

EU statistics on income and living conditions (EU-SILC) methodology – sampling [online]. (2017). Luxembourg: Eurostat. Dostupné z: http://ec.europa.eu [10. 05. 2017].

Gerbery, D. (2011). *Pretrvávajúca chudoba – analýza longitudinálnej databázy EU SILC* [online]. Bratislava: Inštitút pre výskum práce a rodiny. Dostupné na: http://www.ivpr.gov.sk/IVPR/images/IVPR/vyskum/2011/Gerbery/gerbery_2251.pdf [03. 07. 2017].

Hardin, J.W., Hilbe, J. M. (2003). *Generalized Estimating Equations.* New York: Chapman & Hall.

Hsiao, Ch. (2005). *Why Panel Data? IEPR Working Paper 05.33* [online]. University of Southern California: Institute of Economic Policy Research. Dostupné na: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=820204 [14. 08. 2017].

Hu, F. B., Goldberg, J., Hedeker, D. et al. (1998). Comparison of population-averaged and subject-spe

cifc approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, 147 (7), 694 – 703.

Kalvas F. (2003). Zkoumání sociální změny: zaostřeno na panelové šetření. *SDA Info*, 2003, 5 (1), 6 – 9.

Laurie, H. (2013). *Panel Studies* [online].

DOI: http://dx.doi.org/10.1093/obo/9780199756384-0108 [21. 08. 2017]

Lazarsfeld, P. F., Fiske, M. (1938). The Panel as a New Tool for Measuring Opinion. *Public Opinion Quarterly*, 2 (4), 596–612. DOI: 10.1086/265234

Lazarsfeld, P. F. (1940). Panel studies. *Public Opinion Quarterly*, 4 (1), 122 – 128. DOI: 10.1086/265373

Lee, Y., Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society*, Series B, 58 (4), 619 – 678.

Lechnerová, Z. (2009). Charakteristika panelového výzkumu a jeho vývoj. *Data a výzkum – SDA Info,* 3 (1), 31 – 52.

Liang, K. Y., Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73 (1),13 – 22. DOI: 10.2307/2336267

Nelder, J. A., Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135 (3), 370 – 384. DOI: 10.2307/2344614

Neuhaus, J. M., Kalbfleisch, J. D., Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review,* 59 (1), 25 – 35. DOI: 10.2307/1403572

Ruspini, E. (2002). *Introduction to Longitudinal Research*. London: Routledge.

Snijders, T., Bosker, R. (1999). *Multilevel Analysis. An introduction to basic and advanced multilevel modeling*. London, Thousand Oaks, New Delhi: SAGE.

Stiratelli, R., Laird, N.,Ware, J. H. (1984). Random-Effects Models for Serial Observations with Binary Response. *Biometrics*, 40 (4), 961 – 971. DOI:· 10.2307/2531147

Šubrt, J. a kol. (2013). *Soudobá sociologie V (Teorie sociální změny)*. Vyd. 1. Praha: Karolinum.

Veselý, A. (2013). *Přístupy k analýze a explanaci sociální změny. Soudobá sociologie V (Teorie sociální změny).* Vyd. 1. Praha: Karolinum, 2013.

Wong, G. Y., Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association*, 80 (391), 513 – 524. DOI:10.1080/01621459.1985.10478148

Zeger, S. L., Liang, K. Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, 42 (1), 121 – 130. DOI: 10.2307/2531248

Zeger, S. L., Liang, K. Y., Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, 44 (4), 1049 – 1060. DOI: 10.2307/2531734

Zorn, Ch. J. W. (2001). Generalized Estimating Equation Models for Correlated Data: A Review with Applications [online]. *American Journal of Political Science*, 45 (2), 470 – 490. URL: http://www.jstor.org/stable/2669353 [20. 06. 2017]

Kontaktná adresa

doc. RNDr. Viera Labudová, PhD.
Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky
Dolnozemská 1, 852 35 Bratislava, SR
E-mail: viera.labudova@euba.sk
Telefónne číslo: +421267295733

Ing. Martina Lakatová
Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky
Dolnozemská 1, 852 35 Bratislava, SR
E-mail: martina.lakatova@gmail.com
Telefónne číslo: +421267295733

Received: 01. 09. 2017, reviewed: 28. 02. 2018
Approved for publication: 27. 06. 2018