Ekonometrická analýza panelových dat s aplikací na vybavenost domácností[#]

Zuzana Fíglová[*]

1. Úvod

Panelová data představují specifický typ pozorování, která kombinují průřezová a časově uspořádaná data. Znamená to, že několik subjektů, např. domácnosti, firmy nebo státy, jsou pozorovány v několika časových obdobích. Průřezová složka je označována indexem $i = 1, 2,\ldots, N$ a časová složka je rovna $t = 1, 2,\ldots, T$. Panelová data tak umožňují zvětšit datový soubor, snižují kolinearitu mezi vysvětlujícími proměnnými a umožňují testovat i složitější ekonometrické modely. Ekonometrický model, ve kterém je endogenní proměnná diskrétního charakteru (kategoriální nebo kvalitativní veličinou), nazýváme modelem diskrétní volby (Hušek, 2003). Často jde o dichotomickou endogenní proměnnou nabývající pouze hodnot 1 nebo 0, které odpovídají výsledkům binární volby, kdy rozhodující se individuální subjekt může volit pouze mezi dvěma alternativami.

Uvažujme model binární volby panelových dat

$$y_{it}^{*} = X_{it}^{T}\beta + \mu_i + v_{it} \qquad i = 1, 2,\ldots, N, \quad t = 1, 2,\ldots, T, \tag{1}$$

kde

y_{it}^{*} představuje index užitečnosti i-tého individuálního subjektu v čase t při nějaké akci, například při nákupu automobilu, volbě povolání, apod.,
X_{it} $-1\times K$ vektor exogenních proměnných,
μ_i – efekt individuálního subjektu,
v_{it} – náhodná složka s normálním rozdělením.

Náhodná složka v_{it} zachytává vliv vynechaných proměnných v modelu, které mohou zodpovídat různým faktorům:

- faktorům, které jsou typické jenom pro jednotlivá časová období,
- faktorům, které zachytávají rozdíly mezi jednotlivými subjekty, ale nemění se v čase,
- faktorům, které jsou typické jak pro jednotlivé subjekty, tak pro různá časová období.

[#] Článek byl zpracován jako jeden z výstupů výzkumného projektu Ekonometrická analýza mikroekonomických procesů pomocí modelů panelových dat, aplikace v ekonomickém prostředí ČR registrovaného u Grantové agentury České republiky pod evidenčním číslem 402/04/0756.

[*] Ing. Zuzana Fíglová – asistentka; Katedra ekonometrie, Fakulta informatiky a statistiky, VŠE v Praze, zuzanafiglova@hotmail.com.

Hodnota μ_i představuje individuální efekt subjektu, což znamená, že dvě pozorování určitého subjektu v různém období si budou více podobná než dvě pozorování o různých subjektech ve stejném období. Když je tento individuální efekt μ_i korelován s X_{it} jedná se o model s fixními efekty panelových dat, se kterým budeme dále pracovat. V případě nekorelovanosti μ_i s X_{it} jde o model s náhodnými efekty panelových dat.

Breusch a **Pagan** (Greene, 2003) navrhli test, který slouží k určení, zda datový soubor vyhovuje spíše modelům s fixními či náhodnými efekty. Tento test je založený na reziduích získaných pomocí metody nejmenších čtverců. Testovaná hypotéza o tom, že meziskupinový rozptyl komponent je nulový, je ve tvaru

$$\mathrm{H}_0: \quad \sigma_e^2 = 0$$
$$\mathrm{H}_1: \quad \sigma_e^2 \neq 0$$

Testovací statistika založená na Lagrangeově multiplikátoru má tvar

$$LM = \frac{NT}{2(T-1)}\left[\frac{\sum_{i=1}^{N}\left[\sum_{t=1}^{T} e_{it}\right]^2}{\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}^2} - 1\right]^2 \tag{2}$$

Testovací statistika LM má χ^2 rozdělení s jedním stupněm volnosti. V případě nezamítnutí nulové hypotézy se jedná o model s fixními efekty, v opačném případě jde o model s náhodnými efekty.

Pro endogenní proměnnou y_{it} platí

$$y_{it} = 1 \quad \text{když } y_{it}^* > 0 \qquad \text{a } y_{it} = 0 \quad \text{když } y_{it}^* \leq 0. \tag{3}$$

U modelu binární diskrétní volby s fixními efekty tedy platí, že výběrová pravděpodobnost Y_i nabývá hodnoty 1 a $1 - p_i$ je opačná pravděpodobnost, že Y_i má hodnotu 0

$$\mathrm{P}(y_{it} = 1) = \mathrm{p}(y_{it}^* > 0) = \mathrm{P}(v_{it} > -X_{it}^T\beta - \mu_i) = F(X_{it}^T\beta + \mu_i), \tag{4}$$

kde $F(.)$ je distribuční funkce. U logitového modelu binární diskrétní volby předpokládáme logistickou distribuční funkci a označíme ji jako Λ, u probitového modelu uvažujeme normální distribuční funkci s označením Φ.

Logitový model binární diskrétní volby s fixními efekty má tvar

$$\mathrm{P}(y_{it} = 1 | X_{it}) = \frac{e^{X_{it}^T\beta + \mu_i + v_{it}}}{1 + e^{X_{it}^T\beta + \mu_i + v_{it}}}. \tag{5}$$

2. Metody odhadu modelů binární volby panelových dat

Jednou z metod odhadu modelu binární volby s fixními efekty je **metoda nepodmíněné maximální věrohodnosti** pro logitový model definovaná ve tvaru

$$L^N = \prod_{i=1}^{N} \prod_{t=1}^{T} \Lambda(X_{it} + \mu_i)^{Y_{it}} [1 - \Lambda(X_{it} + \mu_i)]^{1-Y_{it}}. \tag{6}$$

Odhadová funkce modelu binární volby s fixními efekty je **metodou podmíněné maximální věrohodnosti** (Chamberlain, 1980) ve tvaru

$$L^P = \prod_{i=1}^{N} \mathrm{P}\left(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, ..., Y_{iT} = y_{iT} \middle| \sum_{t=1}^{T} Y_{it} \right). \tag{7}$$

Pro případ, kdy datový soubor obsahuje velký počet pozorování a $T = 2$, můžou nastat tyto čtyři možnosti $(Y_{i1}, Y_{i2}) = (0,0), (0,1), (1,0)$ a $(1,1)$, a tedy pro

$$\mathrm{P}(Y_{i1} = 0 \text{ a } Y_{i2} = 0 \mid \sum Y_{it} = 0) = 1,$$

$$\mathrm{P}(Y_{i1} = 1 \text{ a } Y_{i2} = 1 \mid \sum Y_{it} = 2) = 1,$$

takže v případě, když se hodnoty těchto pozorování zlogaritmují, jsou rovny nule. Avšak pro další dvě možnosti platí

$$\mathrm{P}(Y_{i1} = 0 \text{ a } Y_{i2} = 1 \mid \sum Y_{it} = 1) = \frac{\mathrm{P}(0,1)}{\mathrm{P}(0,1) + \mathrm{P}(1,0)}, \tag{8}$$

z čehož vyplývá, že podmíněná pravděpodobnost pro tyto dvě pozorování je po náležité úpravě rovna podílu

$$\frac{e^{X_{i2}^T \beta}}{e^{X_{i1}^T \beta} + e^{X_{i2}^T \beta}}.$$

Podobně to platí pro případ, kdy $Y_{i1} = 1$ a $Y_{i2} = 0$.

Mezi další metody odhadu modelu binární diskrétní volby patří metoda tzv. **„maximálního skóre“** (Manski, 1975), v níž maximalizujeme počet správných předpovědí pomocí funkce

$$MS(\beta) = N^{-1} \sum_{i=1}^{N} (2y_{it} - 1)\{\mathrm{X}_{it}^T \beta \geq 0\}. \tag{9}$$

Další neparametrickou metodou odhadu je tzv. „**Kernelova odhadová funkce“**, kde je funkční hodnota F vypočítaná podle

$$F(z^*) \approx \frac{\sum_{i=1}^{N} w_i(z^*) y_i}{\sum_{i=1}^{N} w_i(z^*) y_i}, \tag{10}$$

kde

$$w_i(z^*) = K[(z^* - z_i)/(\lambda s)],$$

$$K(r_i) = P(r_i)[1 - P(r_i)],$$

$$\mathrm{P(ri)} = [1 + \exp(-\mathrm{cri})]^{-1}.$$

Konstanta $c = (\pi / \sqrt{3})^{-1} \approx 0{,}5513$ se používá ke standardizaci logistického rozdělení. Parametr λ je vyrovnávající parametr. Problematiku odhadování modelů binární diskrétní volby panelových dat je třeba chápat jako otevření problému a článek je prezentován především jako metodický.

Podobně jako u lineárního regresního modelu panelových dat, tak i u nelineárních modelů je potřeba testovat přítomnost heterogenity v datovém souboru. **Hausmanův test specifikace** (1978) v případě nezamítnutí nulové hypotézy o homogenitě je odhad metodou nepodmíněné i podmíněné maximální věrohodnosti (MV) konzistentní, ale neefektivní u metody podmíněné MV. V případě nezamítnutí alternativní hypotézy o heterogenitě je odhad metodou nepodmíněné MV nekonzistentní a odhad metodou podmíněné MV je konzistentní a efektivní. Testovací statistika Husmanova testu specifikace je rovna

$$\chi^2 = (\hat{\beta}_{\mathrm{PMV}} - \hat{\beta}_{\mathrm{NMV}})^T (Var[\mathrm{PMV}] - Var[\mathrm{NMV}])^{-1} (\hat{\beta}_{\mathrm{PMV}} - \hat{\beta}_{\mathrm{NMV}}). \tag{11}$$

Testovací statistika má χ^2 rozdělení s počtem stupňů volnosti odpovídajícímu počtu odhadnutých parametrů v modelu.

3. Aplikace modelů diskrétní volby panelových dat

K ilustraci modelů panelových dat byla použita data *Českého statistického úřadu ze Statistiky rodinných účtů* za roky 2000–2004 ze 3 270 domácností. Tyto údaje byly roztříděny do deseti skupin se stejnou četností podle výše ročního hrubého peněžního příjmu v korunách na jednoho člena domácnosti, tj. datový soubor tvořilo 50 pozorování. Ve všech deseti příjmových skupinách byla zjištěna vybavenost domácností předměty dlouhodobé spotřeby. Endogenní proměnnou v modelu byla vybavenost domácností osobním počítačem a exogenní proměnné představovaly čistý příjem (*Příjem*) na osobu a jednotlivé pozorované roky (*Rok 2000* až *Rok 2004*). Časovou proměnnou *Rok 2000* při specifikaci všech modelů vynecháme, abychom zabránili vzniku perfektní multikolinearity.

Pro porovnání bylo použito následujících statických modelů panelových dat:

- lineární pravděpodobnostní model panelových dat s fixními efekty (LPMPDFE), nazývaný i „model nejmenších čtverců s binární závislou proměnnou“ anebo „model založený na analýze rozptylu“. Na základě předpokladu o náhodní složce

v_{it} víme, že metoda nejmenších čtverců („odhadová funkce v rámci skupin“) dává nejlepší lineární nevychýlený odhad neznámých parametrů tohoto modelu. Podle testu Breusche a Pagana je hodnota *LM* statistiky rovna 53,68, což vedlo k zamítnutí nulové hypotézy o modelu s náhodnými efekty.

- logitový model binární diskrétní volby panelových dat s fixními efekty (LOGPDFE), odhadnutý metodou nepodmíněné maximální věrohodnosti (5),
- probitový model binární diskrétní volby panelových dat s fixními efekty (PROBPDFE) odhadnutý metodou nepodmíněné maximální věrohodnosti (5).

Výsledky odhadnutých parametrů spolu s hodnotami t-statistiky (χ^2-statistiky) všech modelů jsou uvedeny v tabulce č. 1. Uvedené odhadové metody jsou součástí softwarového produktu *Limdep*, s jehož pomocí byly provedeny tyto odhady.

Odhady parametrů modelů binární diskrétní volby panelových dat

	LPMPDFE		LOGPDFE		PROBPDFE	
			F i x n í e f e k t y			
Konstanta	0,1660	(7,889)				
Příjem	$0{,}4567*10^{-5}$	(3,308)	$0{,}3197*10^{-5}$	(0,056)	$0{,}1762*10^{-5}$	(0,051)
Rok 2001	0,0343	(1,586)	0,2081	(0,133)	0,1244	(0,135)
Rok 2002	0,0651	(3,009)	0,2792	(0,111)	0,1594	(0,106)
Rok 2003	0,1195	(5,497)	0,6170	(0,427)	0,3685	(0,432)
Rok 2004	0,1725	(7,905)	0,8389	(0,608)	0,5055	(0,616)
			M a r g i n á l n í e f e k t y			
Příjem	$0{,}4567*10^{-5}$	(3,308)	$0{,}1434*10^{-5}$	(0,026)	$0{,}3060*10^{-5}$	(0,027)
Rok 2001	0,0343	(1,586)	0,1597	(0,103)	0,0341	(0,119)
Rok 2002	0,0651	(3,009)	0,1815	(0,082)	0,0387	(0,082)
Rok 2003	0,1195	(5,497)	0,4790	(0,357)	0,1022	(0,382)
Rok 2004	0,1725	(7,905)	0,6926	(0,373)	0,1477	(0,567)

U modelu LPMPDFE koeficient vícenásobné determinace dosáhl poměrně vysokou hodnotu $R^2 = 0{,}6977$ a na základě *F*-testu ($F_{(5,44)} = 20{,}31$) zamítáme nulovou hypotézu o statistické nevýznamnosti R^2. Všechny odhadnuté parametry tohoto modelu kromě proměnné *Rok 2001* jsou statisticky významné na pětiprocentní hladině významnosti. Je splněn i předpoklad o kladných znaménkách odhadnutých parametrů, například, zvýšení příjmu o jednu příjmovou skupinu zvyšuje pravděpodobnost, že tato domácnost bude vybavena osobním počítačem. Podobně lze interpretovat i další odhadnuté parametry. V modelu LPMPDFE zamítáme hypotézu o homoskedasticitě pomocí Whiteova testu, kde testovací statistika $\chi^2_{(6)} = 17{,}017$ s *p*-hodnotou rovnou 0,009 a $F_{(6,37)} = 3{,}182$ *p*-hodnotou rovnou 0,013. Rovněž nezamítáme hypotézu

o autokorelaci prostřednictvím testu Lagrangeova multiplikátoru s testovací statistikou $\chi^2_{(1)} = 23{,}580$ s p-hodnotou rovnou 0,000 a $F_{(1,43)} = 38{,}376$ s p-hodnotou 0,000.

U modelů LOGPDFE a PROBPDFE je splněn předpoklad o kladných znaménkách odhadnutých parametrů, avšak všechny odhadnuté parametry jsou statisticky nevýznamné. Porovnání všech modelů navzájem lze přenásobením vhodnou hodnotou (Greene, 2003). V další části tabulky jsou uvedeny hodnoty marginálních efektů modelů binární volby V případě LPMPDFE jsou tyto hodnoty rovny přímo hodnotám odhadnutých parametrů. U modelů LOGPDFE a PROBPDFE jsou hodnoty marginálních efektů vypočítány vždy pro jednotlivou umělou nula-jednotkovou proměnnou za podmínky, že za hodnoty ostatních umělých nula-jednotkových proměnných dosadíme jejich průměry. Suma marginálních efektů každé proměnné je vždy rovna jedné a změna představuje rozdíl mezi zahrnutím a nezahrnutím dané proměnné do modelu.

4. Závěr

Tato práce obsahuje přehled metod odhadu nelineárních modelů binární volby panelových dat, a to zejména logitového a probitového modelu. Byly uvedeny čtyři základní metody odhadu pro případ modelu panelových dat s fixními efekty. Část těchto metod byla aplikovaná na konkrétní data z oblasti vybavenosti domácností předměty dlouhodobé spotřeby v závislosti na exogenních proměnných příjem a čas, které byly statisticky významné jenom u lineárního pravděpodobnostního modelu binární diskrétní volby panelových dat. Odhady byly provedeny pomocí softwaru *Limdep*.

Literatura

[1] CHAMBERLAIN, G., 1980: Analysis of Covariance with Qualitative Data. *The Review of Economic Studies 47/1*, Econometrics Issue, str. 225–238, 1980.

[2] GREENE, W. H, 2003.: *Econometric Analysis.* New Jersey, Prentice Hall, 2003.

[3] HSIAO, C., 2003: *Analysis of Panel Data.* Cambridge University Press, 2003.

[4] HUŠEK, R., 2003: *Aplikovaná ekonometrie: teorie a praxe.* Praha: Professional Publishing, 2003.

[5] MADDALA, G. S., 1987: Limited Dependent Variable Models Using Panel Data. *The Journal of Human Resources 22/3,* 1987, str. 307–338.

[6] MANSKI, C. F., 1975: Maximum Score Estimation of the Stochastic Utility Model of Choice. *Journal of Econometrics 3,* 1975, str. 205–228.

[7] WOOLDRIDGE, J. 2002: *Econometric Analysis of Cross Section and Panel Data.* MIT Press, 2002.

Ekonometrická analýza panelových dat s aplikací na vybavenost domácností

Zuzana Fíglová

Abstrakt

Panelová data můžeme definovat jako specifický typ dat, které kombinují časové řady a průřezová data, tj. víc subjektů je pozorováno v několika časových obdobích. Použití této kombinace při analýze dat přináší mnoho výhod v podobě zvětšeného datového souboru, snížení kolinearity mezi vysvětlujícími proměnnými a umožňuje testovat i složitější ekonometrické modely. K ilustraci modelů panelových dat byla použita data *Českého statistického úřadu ze Statistiky rodinných účtů* za roky 2000–2004. Data o domácnostech obsahovala demografické charakteristiky, údaje o domácnosti, čistý příjem, hodnocení vlastní socioekonomické situace. Endogenní proměnnou v modelu byla vybavenost domácnosti osobním počítačem a její vývoj během dané periody pomocí tří ekonometrických modelů.

Klíčová slova: model binární volby; panelová data; statistika rodinných účtů; předměty dlouhodobé spotřeby.

Econometric analysis of panel data applied to household characteristics

Abstract

Panel data are specific data where cases are observed at two or more time periods. This approach brings many advantages: larger dataset, decreasing collinearity between exogenous variables and using advanced econometric models. The panel data models were applied to data from the Household Budget Surveys 2000–2004 carried out by the Czech Statistical Office in order to analyze choice behavior of households. The data on households included demographic characteristics of individuals, housing, household amenities, net income, and opinions of households about their own socioeconomic situation. We analyzed the role of income as a determinant of PC ownership and its development through the observed period by using three static econometric models with panel data.

Key words: binary choice model; panel data; household budget survey; durable goods.

JEL classification: C23, C25