Práce s panelovými daty[#]

Václava Pánková*

Úvod

Panelová data vznikají opakovaným pozorováním skupiny jednotek, např. domácností, firem nebo i států, majících určitou společnou charakteristiku (např. země EU15, tranzitivní ekonomiky). V kontextu ekonometrických analýz jsou svébytnou kategorií, která umožňuje nahlédnout současně do struktury i dynamiky studovaných ekonomických jevů. Představují větší souhrn detailních informací a umožňují tak lépe postihnout měnící se ekonomickou strukturu i příčiny takových změn. V rámci panelových dat a technik pro jejich zpracování je uspokojivě řešena i otázka krátkých časových řad, což může být užitečné při analýze dat ČR (viz např. Pánková, oba texty 2005). Mohou se tak stát i pomůckou při zkoumání událostí, ke kterým chybí dostatečně dlouhé časové řady, avšak vyskytují se paralelně v podobných vývojových situacích. To je příklad tranzitivních ekonomik. Převážně se jedná o data průřezová, přičemž je možné je zjistit opakovaně, avšak ne v příliš dlouhém časovém horizontu. Krátké časové řady ukazatelů, zejména ročních, neumožňují kvalitní individuální zkoumání, avšak sdružením údajů z několika analogických ekonomik vytvoříme datový soubor, který dovoluje provést rozumnou statistickou verifikaci výsledků.

Velká část empirických aplikací koresponduje buď s typem modelu obsahujícího pouze náhodné vlivy, nebo s modelem obsahujícím systematické vlivy; tomu pak odpovídá volba odhadových metod. Rozhodnutí o výběru mezi oběma typy modelů je možné podpořit aplikací Hausmanova testu.

Panelová data jako východisko pro formulaci modelu

V obecném případě pracujeme s datovou strukturou

$$y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \quad X_i = \begin{bmatrix} X_{i1}^1 & X_{i1}^2 & .. & X_{i1}^k \\ X_{i2}^1 & X_{i2}^2 & .. & X_{i2}^k \\ \vdots & \vdots & & \vdots \\ X_{iT}^1 & X_{iT}^2 & .. & X_{iT}^k \end{bmatrix} \quad \varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix},$$

kde

y_{it} je vysvětlovaná proměnná příslušející jednotce i a času t

[#] 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.

[*] Doc. RNDr. Václava Pánková, CSc.; Katedra ekonometrie, Fakulta informatiky a statistiky, VŠE v Praze, pankova@vse.cz.

X_{it}^{j} ..je hodnota *j*-té vysvětlující proměnné (*j* = 1, 2, …, *k*) pro *i*-tou jednotku v čase *t*

ε_{it} …je náhodná složka rovnice pro jednotku *i* v čase *t* $\quad i = 1, 2, \ldots, n, \; t = 1, 2, \ldots, T.$

Pro stručnost zápisu použijeme obvyklé značení

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{1}$$

kde *y* je vektor rozměru *nT*, *X* matice $nT \times k$, ε vektor s počtem složek *nT*.

Relevantní standardní lineární model pak formulujeme vztahem

$$y = X\beta + \varepsilon \tag{2}$$

když

$$\beta = \begin{bmatrix} \beta_1 & \beta_2 & .. & \beta_k \end{bmatrix}'$$

Pokud náhodná složka modelu (2) splňuje odpovídající předpoklady, můžeme vektor parametrů β odhadnout i metodou nejmenších čtverců (MNČ). Znamená to ovšem ignorovat skutečnost, že *n* individuálních pozorování *T*-krát není totéž jako *nT* individualit.

Metodicky je třeba rozlišovat případy, kdy *T* je podstatně větší než *n* (krajní případ *n*=1 znamená standardní časové řady) od situace, kdy větším z obou rozměrů je *n* (limitně pak *T*=1 znamená čistě průřezová data). Při analýzách vztahujících se k české ekonomice je obvyklejší druhý z obou případů. Dále popsané metody jsou na VŠE k dispozici v rámci softwarového vybavení PcGive.

Metody pro odhad parametrů

Vzhledem k dostupnosti dat popisujících ekonomiku ČR je především zájem o metody týkající se případu *n*>*T*. V zájmu snadné orientace ve školním softwaru VŠE nebudou jejich anglické názvy překládány.

Bereme-li v úvahu panelovou strukturu dat, je vhodné náhodnou složku chápat jako součet

$$\varepsilon_{it} = \alpha_i + \eta_{it},$$

kde η_{it} není korelováno s X_{it} a α_i reprezentuje individuální efekt. Tím je myšlena skutečnost, že dvě časově různá pozorování téže jednotky si budou více podobná než údaje o dvou různých jednotkách ve stejném čase.

Individuální efekt se obecně rozlišuje na dva případy

(i) α_i není korelováno s X_{it}
(ii) α_i je korelováno s X_{it}.

Metody spojené s případem (i)

Postup typu „between“

Vektor parametrů β je odhadován pomocí MNČ z rovnice

$$\bar{y}_i = \bar{X}_i \beta + u_i^B ,$$

kde $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ a $\bar{X}_i$ je definováno analogicky. Pro maticový zápis bude vhodné zavést $nT \times n$ – rozměrnou matici $D = I_n \otimes i_T$, v níž i_T je T-rozměrný vektor jedniček, která obsahuje nula – jedničkové (dummy) proměnné korespondující s každou z n jednotek souboru. Parametr typu *„between"* je pak odhadnut pomocí estimátoru

$$\hat{\beta}_B = (X'P_D X)^{-1} X'P_D y , \tag{3}$$

kde $P_D = D(D'D)^{-1}D'$ je symetrická idempotentní matice.

Každá jednotka je tedy v regresi zastoupena průměrnými hodnotami ze svých T pozorování; MNČ je pak aplikována na data v počtu n.

Postup typu „within"

Jedná se o estimátor

$$\hat{\beta}_W = (X'M_D X)^{-1} X'M_D y , \tag{4}$$

kde $M_D = I_{nT} - D(D'D)^{-1}D'$ je rovněž symetrická idempotentní matice. Výraz (4) je shodný s aplikací MNČ na rovnici

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + u_{it}^W$$

(podrobnosti např. v Johnston, DiNardo, 1997).

Do regrese tedy za každou jednotku vstupují jen odchylky od jejích průměrných ukazatelů; celkem nT údajů pro každou proměnnou.

Odhady zobecněnou MNČ

Tato metoda (v PcGive GLS) probíhá ve dvou krocích: v prvním je odhadována kovarianční matice vektoru náhodných složek modelu, ve druhém je znalost kovarianční struktury použita k vyjádření odhadu vektoru parametrů β modelu.

K provedení prvního kroku nejprve formalizujeme

$$E(\eta) = 0 , \quad E(\eta\eta') = \sigma_\eta^2 I_{\eta T} , \quad E(\alpha_i) = 0 , \quad E(\alpha_i \alpha_j) = 0 \quad \text{pro} \quad \text{i} \neq \text{j},$$
$$E(\alpha_i \alpha_i) = \sigma_\alpha^2 , \; E(\alpha_i \eta_{jt}) = 0 .$$

Za těchto předpokladů má každá z n jednotek kovarianční matici rozměru $T \times T$

$$\Sigma = E(\varepsilon_i \varepsilon_i') = \sigma_\eta^2 I_T + \sigma_\alpha^2 i_T i_T' = \begin{bmatrix} \sigma_\eta^2 + \sigma_\alpha^2 & \sigma_\alpha^2 & .. & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\eta^2 + \sigma_\alpha^2 & .. & \sigma_\alpha^2 \\ \vdots & \vdots & & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & .. & \sigma_\eta^2 + \sigma_\alpha^2 \end{bmatrix} .$$

Datům ze schématu a modelu pak bude odpovídat kovarianční matice

$$\Omega = I_n \otimes \Sigma = E(\varepsilon\varepsilon') = \begin{bmatrix} \Sigma & 0 & .. & 0 \\ 0 & \Sigma & .. & 0 \\ : & : & & : \\ 0 & 0 & .. & \Sigma \end{bmatrix}.$$

Pro inverzi blokově diagonální matice Ω pak bude třeba zjistit

$$\Sigma^{-1/2} = \frac{1}{\sigma_\eta}\left[I_T - \left(\frac{1-\theta}{T} ii' \right) \right],$$

kde

$$\theta = \sqrt{\frac{\sigma_\eta^2}{T\sigma_\alpha^2 + \sigma_\eta^2}} \tag{5}$$

je veličina, kterou v prvním kroku najdeme odhadem.

V druhém kroku pak použijeme MNČ na data $\tilde{y}$ a $\tilde{X}$ získaná transformací

$$\tilde{y}_{it} = y_{it} - \bar{y}_i + \hat{\theta}\bar{y}_i, \quad \tilde{X}_{it} = X_{it} - \bar{X}_i + \hat{\theta}\bar{X}_i . \tag{6}$$

Pokud máme k dispozici odhady typu *„between“* s rezidui $\hat{u}^B$ a odhady typu *„within“* s rezidui $\hat{u}^W$, můžeme do (5) dosadit

$$\hat{\sigma}_\eta^2 = \frac{1}{nT - nk - n}\hat{u}'^W\hat{u}^W, \quad \hat{\sigma}_\alpha^2 = \hat{\sigma}_B^2 - \frac{\hat{\sigma}_\eta^2}{T}, \quad \hat{\sigma}_B^2 = \frac{\hat{u}'^B\hat{u}^B}{n-k},$$

čímž získáme hledané $\hat{\theta}$.

Všimněme si ještě že estimátor získaný pomocí MNČ z (2) můžeme rozepsat jako

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}(X'M_D y + X'P_D y) = (X'X)^{-1}X'M_D X\hat{\beta}_W + (X'X)^{-1}X'P_D X\hat{\beta}_B .$$

Je to tedy vážený součet estimátorů *„within“* a *„between“*.

Vztahy (6) pak ukazují, že $\sigma_\alpha^2 = 0$ (neexistence nekorelovaných individuálních komponent) znamená θ=1 a zobecněná MNČ je pak pouze MNČ.

Metody spojené s případem (ii)

Nyní budeme pracovat s modelem

$$y_{it} = X_{it}\beta + \alpha_i + \eta_{it} \tag{7}$$

v němž musíme připustit, že $\mathrm{cov}(X_{it}, \alpha_i) \neq 0$. Důsledkem je, že hodnoty α_i ze vztahu (7) je třeba zjistit odhadem jako další parametry. Odhady těchto parametrů však nemohou být konzistentní, protože jejich počet je roven *n* a s rostoucím rozsahem souboru se tedy i počet parametrů stále zvyšuje. Abychom provedli konzistentní odhad alespoň pro parametry β, lze postupovat v souladu s větou Frishe, Waugha a Lovella (viz např. Davidson, McCinnon, 1993) a vyjít z modifikace (7)

$$y = X\beta + D\alpha + \eta ,$$

kde $D = I_n \otimes i_T$, a odhadnout pomocí MNČ $\hat{\beta} = (X'M_D X)^{-1}X'M_D y$, když $M_D = I - D(D'D)^{-1}D'$, Snadno zjistíme, že takto byl vlastně proveden odhad typu

within, tak jak byl popsán pro případ (i). Oprávněnost tohoto přístupu nahlédneme, když uvážíme, že je

$$\bar{y}_i = \bar{X}_i\beta + \bar{\alpha}_i + \bar{\eta}_i = \bar{X}_i\beta + \alpha_i + \bar{\eta}_i$$

a tedy

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + (\eta_{it} - \bar{\eta}_i)\,. \tag{8}$$

Ze vztahu (8) je zřejmé, že individuální vlivy α_i byly takto eliminovány. Stejného efektu lze dosáhnout, bude li se pracovat s diferencemi původních dat. Je ale třeba upozornit, že touto cestou z modelu vyloučíme i takové proměnné, které jsou individuálním a časově invariantním ukazatelem jednotek v souboru (podrobnosti např. v Johnston, DiNardo, 1997).

Pro úplnost zmíníme ještě dva přístupy k datovým souborům s malým počtem jednotek a dostatečně dlouhými časovými řadami.

Metoda LSDV (= Least Squares with Dummy Variables) ošetřuje individuální vlivy zavedením nula – jedničkových proměnných v počtu *n*-1, kterými jsou rozlišeny jednotky souboru. Individuální efekt se tak projevuje v různých hodnotách konstanty.

Odlišení ve všech parametrech můžeme dosáhnout zpracováním, které je aplikací simultánní soustavy rovnic. Počet rovnic se rovná počtu sledovaných jednotek a formálně koresponduje s filozofií zdánlivě nesouvisejících regresních rovnic, které jsou propojeny prostřednictvím určitých vlastností svých náhodných složek.

Rovnice uvažujme opět ve tvaru

$$y_i = X_i\beta_i + u_i\,, \quad \text{i} = 1,..., \text{n},$$

kde n je počet rovnic a také datových jednotek. Jsou spojeny prostřednictvím náhodných složek, pro které

$$E(u_i) = 0\,, \;\; E(u_i u_i') = \omega_{ii} I_m\,, \;\; E(u_{it}u_{jt}) = \omega_{ij}\,, \text{ pro i, j} = 1,...,\text{n},$$

přičemž $t = 1,...,T$ je počet pozorování v čase. Předpokládáme, že $E(u_{it}u_{j\tau}) = 0$ pro $t \neq \tau$. Veškerá informace týkající se korelačních vztahů mezi náhodnými složkami tak může být popsána maticí Ω rozměru $n \times n$, která na místě ij má prvek ω_{ij}. Dále je možné předpokládat, že každá rovnice může mít jiný počet vysvětlujících proměnných a tudíž i parametrů. Počet parametrů v i-té rovnici označme k_i a $\sum_{i=1}^{n} k_i = k$.

Celou soustavu rovnic můžeme popsat vztahem

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & X_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

nebo kompaktněji jako

$$y = X\beta + u\,, \tag{9}$$

když y je vektor délky Tn, X je matice rozměru $Tn \times k$ a β je vektor délky k. Vektor náhodných složek o délce Tn má kovarianční matici

$$V = E(uu') = \begin{pmatrix} E(u_1u_1') & E(u_1u_2') & \dots & E(u_1u_n') \\ E(u_2u_1') & E(u_2u_2') & \dots & E(u_2u_n') \\ \vdots & \vdots & & \vdots \\ E(u_nu_1') & E(u_nu_2') & \dots & E(u_nu_n') \end{pmatrix} = \Omega \otimes I$$

rozměru $Tn \times Tn$.

Parametry modelu (9) odhadujeme zobecněnou metodou nejmenších čtverců. V případě, že předpokládané vztahy mezi náhodnými složkami existují, tedy není-li $\omega_{ij} = 0$ při $i \neq j$, budou odhady parametrů vydatnější než při použití metody nejmenších čtverců na každou rovnici zvlášť.

Hausmanův test

Případy (i) a (ii) odlišující vlastnosti individuálního efektu představují v praxi krajnost, které zpravidla není dosaženo. Empirické poznatky ukazují, že vyjdeme-li z předpokladu (i), mohou být odhadnuté parametry vychýlené směrem nahoru, zatímco předpoklad (ii) spíše povede k vychýlení odhadů směrem dolů. Rozhodování mezi (i) a (ii) je možné podpořit provedením Hausmanova testu, který stručně popíšeme.

Předpokládejme, že odhad provedený za předpokladu (i) poskytl vektor parametrů $\hat{\beta}_{(i)}$ a kovarianční matici $\Sigma_{(i)}$, zatímco předpoklady (ii) vyústily v $\hat{\beta}_{(ii)}$ a $\Sigma_{(ii)}$. Testovat budeme statistiku

$$H = (\hat{\beta}_{(i)} - \hat{\beta}_{(ii)})'(\Sigma_{(ii)} - \Sigma_{(i)})^{-1}(\hat{\beta}_{(i)} - \hat{\beta}_{(ii)}),$$

která asymptoticky má χ^2 test s k (= počet sloupců v matici X) stupni volnosti. Nulovou hypotézou je, že platí předpoklad (i). Podrobněji viz např. Green (2003).

Budeme-li mít k dispozici $\hat{\theta}$ jako odhad ze vztahu (5), lze postupovat i takto (viz Johnston, DiNardo, 1997). Proměnné y, X transformujeme na

$$\tilde{y}_{it} = y_{it} - \bar{y}_i + \hat{\theta}\bar{y}_i , \quad \tilde{X}_{it} = X_{it} - \bar{X}_i + \hat{\theta}\bar{X}_i ,$$

dále definujme $\tilde{\tilde{X}}_{it} = X_{it} - \bar{X}_i$. Hausmanův test pak lze provést jakožto F – test parametru γ ve vztahu

$$\tilde{y} = \tilde{X}\beta + \tilde{\tilde{X}}\gamma + u .$$

Testována je hypotéza, zda vynechání individuálního efektu má vliv na konzistentnost odhadů, pokud tyto byly provedeny metodami typu (i).

Literatura

[1] DAVIDSON, R. – MacKINNON, J., 1993: *Estimation and Inference in Econometrics*. Oxford Univ. Press, 1993.

[2] GREEN, W. H., 2003: *Econometric Analysis. Pearson Education Ltd*. New Persey, 2003.

[3] JOHNSTON, J. – DiNARDO, J., 1997: *Econometric Methods*. McGrawHill, 1997.

[4] PÁNKOVÁ, V., 2005: Poptávka po kapitálu v tranzitivních ekonomikách. *Ekonomický časopis*, 2005, roč. 53, č. 2, s. 119–128. ISSN 0013-3035

[5] PÁNKOVÁ, V., 2005: Tobinovo Q – teorie a aplikace. *Politická ekonomie*, 2005, roč. LIII, č. 5, s. 601–608. ISSN 0032-3233

Práce s panelovými daty

Václava Pánková

Abstrakt

Panelová data vznikají opakovaným pozorováním určité skupiny jednotek, např. domácností či firem, ale také celých ekonomik s některými společnými charakteristikami jako třeba země EU15, tranzitivní ekonomiky, apod. Získáme tak více detailních informací, které nám dovolí analyzovat a zdůvodnit změny zkoumané ekonomické struktury. Z řady možných odhadových technik chce tento článek upozornit na takové, které umožňují práci s velmi krátkým časovým rozměrem panelových dat. Pro Českou republiku a ostatní stále ještě relativně nové trhy jsou právě tyto techniky velmi vhodné.

Většina aplikací je spojena s rozlišováním vlastností individuálního efektu a v důsledku toho pak s volbou adekvátní metody pro odhad. Rozlišit modely v tomto smyslu umožňuje např. Hausmanův test.

Klíčová slova: panelová data; model s náhodnými/systematickými vlivy; krátké časové řady; Hausmanův test.

Econometric models with panel data

Abstract

Panel data are a result of repeating observations of a group of units, e.g. households, firms, but also whole economies with some common characteristics as EU15, transition economies a . s. o. So, more details are available enabling to analyze a changing economic structure and its reasoning. Specific techniques can be chosen to deal with short time series what in case of Czech Republic, and other relatively new markets, can be very helpful.

Most part of empirical applications corresponds with random or fixed effect models, respective. To each of this type appropriate methods relate. An exact choice between both effects can be done by the help of Hausman test.

Key words: panel data; random / fixed effects model; short time series; Hausman test.

JEL classification: C23