Analýza panelových dat

Petr Novák[*]

Úvod

V posledních desetiletích výrazně roste zájem o problematiku analýzy panelových dat. A to jak ve výzkumu sociálních vazeb a způsobů chování obyvatelstva na straně jedné, tak při řešení otázek ve sféře podnikové, národní či mezinárodní. Aplikace tedy nalezneme na mikro i makroúrovni. V ekonomii je analýza panelových dat užívána například ke studiu chování firem a mezd zaměstnanců v průběhu určitého časově ohraničeného období. V politických vědách se můžeme setkat s panelovým výzkumem volebních preferencí, změn politické příslušnosti. Je využívána v psychologii, sociologii, zdravotnictví, bankovnictví, pojišťovnictví, školství.

Analýzu panelových dat bychom mohli definovat jako studium jednotlivých subjektů (jednotlivců, domácností, podniků, regionů, států, ...) a jejich vzájemných vztahů, u kterých periodicky provádíme zjišťování charakteristických znaků a jejich následné hlubší prozkoumávání. Podobná definice může být následující: panelová data, někdy také nazývaná jako longitudinální data nebo taktéž věcně-prostorová data zjišťovaná opakovaně za určitý časový úsek, jsou data, kde jsou charakteristiky za jednotlivá pozorování (lidé, firmy, země,...) zjišťovány za dvě a více časových období.

Panelová data poskytují oproti prostým věcně prostorovým datům (tzn. získaných pouze v jednom časovém okamžiku nebo za jeden časový interval) a datům v časových řadách několik nesporných výhod. Především získáme velké množství pozorování, která nejsou v konvenčních časových řadách dostupná. Panelová data nejsou obvykle příliš agregovaná jako typická data v časových řadách, proto můžeme analyzovat a testovat komplikovanější hypotézy dynamiky a vzájemného chování. To se nám nepodaří v případě použití věcně prostorových dat získaných právě pouze v čase *t*. Konečně využití panelových dat může taktéž sloužit k dokonalejší analýze skrytých, nepozorova(tel)ných, náhodných skutečností v ekonometrické (pokud provádíme výzkum na makroúrovni), sociologické a další struktuře vztahů mezi jednotkami.

Panelem (angl. ekvivalent – „panel data set“) se rozumí soubor jednotek, které si jsou nějakou charakteristickou vlastností velmi podobné nebo příbuzné (osoby, domácnosti, firmy, geografické oblasti atp.), na kterých se provádí kontinuální (v čase se opakující) výzkum. Zmíněným souborem může být například jak celá populace, tak soubor náhodným způsobem vygenerovaný a původní generaci dobře reprezentující. Nutnou podmínkou pro možnost definování panelu a následné analýzy panelových dat je zejména ta skutečnost, že soubor jednotek se v čase nemění, „vypadnuté“ jednotky se nenahrazují novými.

[*] Ing. Petr Novák – doktorand; Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, VŠE v Praze, novakp@vse.cz .

Zkoumání panelových dat využívá modelového způsobu řešení, ve kterém se objevují jak prvky analýzy časových řad, tak prvky regresní analýzy. Panelová analýza tedy v podstatě představuje další stupeň modelace, která mnohonásobně zhodnocuje obvykle draze získané informace o nějaké skutečnosti. Panelová analýza není doposud detailně teoreticky popsána, jednak z důvodu krátké historie a jednak z důvodu náročnosti zkoumání. Nicméně s rozvojem nových softwarových produktů, bez nichž je v současnosti analýza panelů nepředstavitelná, se můžeme dočkat nových poznatků objevujících se ve vědeckých pracích exponenciálním tempem.

Model a symbolika

Předpokládejme, že získáme pomocí výběrových technik hodnoty u K+1 znaků u N pozorovaných objektů přes T časových období. Jinými slovy, provedeme T-krát opakovaný záznam odpovědí na K+1 otázek, které jsme položili N respondentům. Nyní si zavedeme nutnou symboliku, která nám pomůže v přehledném zápisu vztahů a modelů.

Symbolem y_{it}, $i = 1, 2, ..., N$; $t = 1, 2, ..., T$ budeme označovat hodnotu vysvětlované proměnné u i-tého pozorování v čase t závislé na K exogenních, vysvětlujících proměnných, tedy $\mathbf{x}_{it} = (x_{1it},...,x_{Kit})'$, kde x_{jit}; $j = 1, 2, ..., K$, $i = 1, 2, ..., N$, $t = 1, 2, ..., T$ vyjadřuje hodnotu j-té nezávislé, vysvětlující, proměnné u i-tého pozorování v čase t.

V případě, že $T = 1$, získáme čistě průřezová data, pak můžeme použít obvyklé regresní či jiné techniky analýzy průřezových dat. Pokud $N = 1$, dostaneme v čase opakovaně získané údaje týkající se jednoho pozorování, tedy časovou řadu hodnot, potom použijeme pro zkoumání této řady nejrůznější techniky analýzy časových řad.

Uvažujme lineární regresní model ve tvaru

$$y_{it} = \mu + \sum_{j=1}^{K} \beta_j x_{jit} + u_{it} \; ; j = 1, 2, ..., K, i = 1, 2, ..., N, t = 1, 2, ..., T \qquad (1)$$

Tento model předpokládá, že vliv vypuštěných, v čase měnlivých, proměnných u jednotlivých pozorování je nevýznamný, nicméně pro model jako celek již ho jako významný chápat lze. Všechny tyto vypuštěné „individuální“ vlivy jsou zahrnuty v parametru μ. Parametry směrnic β_j, $j = 1, 2, ..., K$, jsou stejné pro všechna pozorování. Pokud bychom chtěli tento model použít k popisu nějaké skutečnosti, dospějeme k závěru, že za těchto předpokladů je nepoužitelný. Například porovnáváme-li důchody na osobu v několika zemích, zjistíme, že změny v těchto důchodech nereagují stejnou měrou na změny populačního růstu nebo na změny kapitálových statků. Jinými slovy individuální efekty a efekty času nebudou tak nedůležité. Mohou být korelovány s vysvětlujícími proměnnými. Proto je potřeba vytvořit model, který by zachytil tuto heterogenitu.

Problém heterogenity v panelových modelech řeší dva typy modelů: modely fixních efektů a modely náhodných efektů.

Modely fixních a náhodných efektů

Model fixních efektů („Fixed Effects Model“)

Nechť y_{it} závisí na souboru K vysvětlujících proměnných, $\mathbf{x}_{it} = (x_{1it},...,x_{Kit})'$ a konstanty jsou specifické pro i-tou jednotku v čase t, ve stejném čase jsou ale konstantní, potom

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}' \mathbf{x}_{it} + u_{it} \; ; \; i = 1, 2, ..., N, \; t = 1, 2, ..., T. \tag{2}$$

$\boldsymbol{\beta}'$ je vektor konstant rozměru 1xK a α_i^* je konstanta reprezentující efekty těch proměnných, které jsou příznačné (charakteristické) i-tému pozorování. Chybová složka u_{it} reprezentuje efekty nevýznamných proměnných příznačných i-tým pozorováním a danému časovému intervalu. Dále o této složce předpokládáme, že je nekorelovaná s vektorem $\mathbf{x}_{it}$, pro všechna i a t, a pochází z nezávisle identického rozdělení s nulovou střední hodnotou a konstantním rozptylem, symbolicky

$$u_{it} \sim IID(0; \sigma_u^2) . \tag{3}$$

Tento model je často nazýván také jako model kovarianční analýzy („analysis-of-covariance model“) nebo jako základní model reprezentující strukturu panelových dat.

Model náhodných efektů („Random Effects Model“)

V regresní analýze se považuje za standardní praxi předpokládat, že velká část faktorů, které mají dopad na chování závislé proměnné a přitom nejsou explicitně součástí nezávislých proměnných, jsou zahrnuty do faktoru vyjadřujícího náhodné výkyvy. Pokud provedeme v čase opakované zjišťování u N objektů, předpokládá se často, že některé proměnné budou reprezentovat faktory, které jsou příznačné jak jednotlivým objektům, tak jednotlivým časovým úsekům. Jiné proměnné budou odrážet individuální rozdíly, které mají v průběhu času sklon ovlivňovat získané hodnoty jednotlivých objektů víceméně stejným způsobem. Konečně poslední část proměnných bude odrážet faktory, které jsou vlastní specifickým časovým úsekům, ale mají podobný dopad na chování jednotlivých objektů panelu. Proto zavedeme rezidua υ_{it}, která spojují výše popsané charakteristiky proměnných, ve tvaru

$$\upsilon_{it} = \alpha_i + \lambda_t + u_{it} , \tag{4}$$

kde

$$E(\alpha_i) = E(\lambda_t) = E(u_{it}) = 0 , \tag{5}$$

$$E(\alpha_i \lambda_t) = E(\alpha_i u_{it}) = E(\lambda_t u_{it}) = 0 , \tag{6}$$

$$E(\alpha_i \alpha_i) = \sigma_\alpha^2 , \; E(\alpha_i \alpha_j) = 0 \text{ pro } i \neq j , \tag{7}$$

$$E(\lambda_t \lambda_t) = \sigma_\lambda^2 , \; E(\lambda_t \lambda_s) = 0 \text{ pro } t \neq s , \tag{8}$$

$$E(u_{it} u_{it}) = \sigma_u^2 , \; E(u_{it} u_{js}) = 0 \text{ pro } i \neq j \text{ nebo } t \neq s . \tag{9}$$

Dále se předpokládá

$$E(\alpha_i \mathbf{x}'_{it}) = E(\lambda_t \mathbf{x}'_{it}) = E(u_{it} \mathbf{x}'_{it}) = \mathbf{0}' . \tag{10}$$

Rozptyl proměnné y_{it} je potom dán jako součet rozptylů jednotlivých složek, tedy

$$\sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2 \tag{11}$$

Rozptyly σ_α^2, σ_λ^2 a σ_u^2 jsou označovány jako složkové rozptyly („variance components“): každý z rozptylu má svou typičnost a je složkou rozptylu σ_y^2. Proto se tento model ve tvaru

$$y_{it} = \mu + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + \lambda_t + u_{it} \tag{12}$$

označuje jako model složek rozptylu („variance-components model“) nebo také jako model komponentních chyb („error-components model“).

Dynamické modely panelových dat

Většina ekonomických veličin jsou ve své podstatě dynamickými procesy, proto by bylo vhodné je i jako dynamické modelovat. V případě použití panelu pro modelaci jako datové základny nám časový rozměr umožňuje popsat v čase probíhající proces přizpůsobování.

Tímto se dostáváme k popisu dynamických modelů typu

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i^* + \lambda_t + u_{it} , \; i = 1, 2, ..., N, \; t = 1, 2, ..., T, \tag{13}$$

kde $\mathbf{x}_{it}$ je Kx1 rozměrný vektor vysvětlujících proměnných, $\boldsymbol{\beta}'$ je 1xK rozměrný vektor konstant, α_i^* jsou (nepozorované) individuální efekty a λ_t časově specifické efekty, o kterých se předpokládá, že zůstávají pro i-té pozorování v čase konstantní; a u_{it} reprezentuje složku nepozorovatelných proměnných přes indexy i a t. Dále požadujeme některé vlastnosti týkající náhodných chyb:

1. $E(u_{it}) = 0$, tedy nulovou střední hodnotu náhodných chyb,
2. $E(u_{it} u_{js}) = \sigma_u^2$ pro i = j a t = s, tzn. pro všechna pozorování v čase konstantní rozptyl a
3. $E(u_{it} u_{js}) = 0$ pro $i \neq j$ a/nebo $t \neq s$, tedy nulové autokovariance náhodných chyb.

Při použití tohoto modelu, ačkoliv vypadá podobně jako statický panelový model, se objeví komplikace s odhady. Problémy se týkají konzistence a nezkreslenosti odhadu $\hat{\beta}$. Z tohoto důvodu byly navrženy odhadové techniky používající:

- metodu maximální věrohodnosti („Maximum Likelihood Estimator“ – MLE),
- metodu zobecněných nejmenších čtverců („Generalized Least-Squares Estimator“ – GLS),

- instrumentální proměnné („Instrumental-Variable Estimator“ – IV),
- zobecněnou momentovou metodu („Generalized Method of Moments Estimator“ – GMM).

Testy jednotkových kořenů panelových dat („panel unit root tests“)

Jak uvádí například Breitung, Pesaran (2005), jeden z hlavních důvodů stojících za aplikacemi testů jednotkových kořenů a podobně i testů kointegrace panelových dat bylo dosažení statistické síly a případně zvýšení síly jejich existujících jednorozměrných protějšků. Toho bylo docíleno aplikacemi testů jednotkových kořenů tzv. první generace na časových řadách typu reálný měnový kurz, celkový výstup a inflace.

Bohužel testování hypotéz o existenci jednotkových kořenů a kointegrace za použití panelových dat oproti testům v rámci jednorozměrných časových řad přináší dodatečné komplikace. Za prvé, panelová data obecně vnášejí do modelů podstatné množství nepozorované heterogenity, která je spodobněná ve specifických parametrech jednotlivých objektů. Za druhé, v mnoha empirických aplikacích, zejména v aplikacích typu reálných měnových kurzů, se neadekvátně předpokládá nezávislost průřezových dat. K překonání těchto problémů byly vyvinuty a doposud jsou vyvíjeny variantní techniky testování panelových jednotkových kořenů uplatnitelných v rozličných formách meziobjektových závislostí. Za třetí, je často velmi obtížné interpretovat výsledky určitého panelového modelu v případě zamítnutí hypotézy neexistence jednotkových kořenů nebo neexistence kointegračních vztahů mezi proměnnými v modelu. Závěr uskutečněný na základě výsledků testů nemůže tedy obvykle vypadat následovně: „statisticky významná část objektů panelu je stacionární nebo kointegrovaná“.

V porovnání s testy panelových jednotkových kořenů, analýza kointegrace v panelech je stále na počátku hledání odpovědí.

Práce současných autorů navrhují tedy testy jednotkových kořenů panelových dat, které mají větší sílu než testy jednotkových kořenů používaných pro ověřování stacionarity jednorozměrných časových řad. Lze zmínit testy autorů

- Levin, Lin a Chu (2002) – test LLC
- Breitung (2000)
- Im, Pesaran a Shin (2003) – test IPS
- Maddala a Wu (1999), Choi (2001) – Fisher-ADF test a Fisher-PP test
- Hadri (2000).

Z hlediska testování panelových jednotkových kořenů je nutné přijmout určité předpoklady týkající se parametrů γ_i. Uvažuje se případ, kdy jsou všechny autoregresní parametry γ_i identické pro všechny objekty, tedy $\gamma_i = \gamma$, pro všechna $i = 1, 2, ..., N$. Za tohoto předpokladu lze použít testy Levina, Lina a Chua (test LLC), Breitunga a Hadriho. Alternativně lze parametry γ_i označit jako individuálně specifické autoregresní koeficienty, potom použijeme testy Ima, Pesarana a Shina (IPS test), Fisherův ADF test nebo Fisherův PP test.

Tabulka: Panelové testy jednotkových kořenů

Test	Nulová hypotéza	Alternativní hypotéza	Deterministické komponenty v modelu	Autokorelační korekční metoda
Levin, Lin a Chu	Jednotkové kořeny jsou	Nejsou jednotkové kořeny	N, F, T	Zpoždění
Breitung	Jednotkové kořeny jsou	Nejsou jednotkové kořeny	N, F, T	Zpoždění
IPS	Jednotkové kořeny jsou	Některé objekty mají jednotkové kořeny	F, T	Zpoždění
Fisher-ADF	Jednotkové kořeny jsou	Některé objekty mají jednotkové kořeny	N, F, T	Zpoždění
Fisher-PP	Jednotkové kořeny jsou	Některé objekty mají jednotkové kořeny	N, F, T	Kernel
Hadri	Nejsou jednotkové kořeny	Jednotkové kořeny jsou	F, T	Kernel

Pozn.: N – bez exogenních proměnných, F – fixní efekty, T – individuální efekty a individuální trendy, Kernel – jádrová metoda odhadu parametrů

Počítačové aplikace – efektivní zpracování panelových dat

V současné době je již nemyslitelné analyzovat datové soubory v rámci panelové analýzy bez použití statisticko-ekonometrických počítačových produktů. Mezi hlavní představitele uveďme programy jako EVIEWS, LIMDEP, PcGive (modul GiveWin2), SHAZAM, STATA. Pracujeme-li se specifickým modelem panelových dat nebo máme-li zájem zkoumat panelová data atypickým způsobem, lze kromě výše uvedených programů využít doplňkové zpracování pomocí ekonometrických programovacích jazyků (GAUSS, OX).

Závěr

Cílem této práce bylo seznámit čtenáře se základy problematiky panelových modelů. Chtěl jsem ukázat, že práce s analyzováním panelových dat není v žádném případě jednoduchá záležitost. U analytika klade vysoké nároky jak teoretického charakteru, tak i na zkušenosti s modelováním skutečných jevů a procesů. Nedílnou součástí musí být samozřejmě znalosti týkající se schopností výpočetních technik různých aplikačních softwarů, bez kterých je již nepředstavitelné s panelovými daty pracovat.

Literatura

[1] NOVÁK, P., 2006: *Analýza panelových dat*. Diplomová práce, 2006.

[2] ARELLANO, M., 2003: *Panel Data Econometrics*. Oxford University Press, 2003.

[3] BALTAGI, B. H., 1995: *Econometric Analysis of Panel Data.* England, J. Wiley, 1995.

[4] BOND, S. R., 2002: Dynamic Panel Data Models: A Guide to Micro Data and Practice. *Cemmap Working Paper Series,* 2002, No. CWP09/02, Institute for Fiscal Studies, London.

[5] BREITUNG, J., PESARAN, M. H., 2005: Unit Roots and Cointegration in Panels. *IEPR woring paper*, 2005.

[6] BREITUNG, J., 2000: The Local Power of Some Unit Root Tests for Panel Data. In B. Baltagi (ed.), *Advances in Econometrics*, 2000, Vol. 15: Nonstationary Panels, Panel Cointegration, and Dynamic Panels, Amsterdam: JAI Press, p. 161–178.

[7] HADRI, K., 2000: Testing for Stationarity in Heterogeneous Panel Data. *Econometric Journal*, 3, 148–161.

[8] HARRIS, R., SOLLIS, S., 2003: *Applied Time Series Modelling and Forecasting*. England, J. Wiley, 2003.

[9] HSIAO, Ch., 2003: *Analysis of panel data.* Cambridge University Press, 2nd ed, 2003,.

[10] CHOI, I., 2001: Unit Root Tests for Panel Data, *Journal of International Money and Finance*, 2001, 20, 249–272.

[11] IM, K. S., PESARAN, M. H., SHIN, Y., 2003: Testing for Unit Roots in Heterogeneous Panels, *Journal of Econometrics*, 2003, 115, 53–74.

[12] LEVIN, A., LIN, C. F., CHU, C., 2002: Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties, *Journal of Econometrics*, 2002, 108, 1–24.

[13] MADDALA, G. S., WU, S., 1999: A Comparative Study of Unit Root Tests with Panel Data and A New Simple Test, *Oxford Bulletin of Economics and Statistics*, 1999, 61, 631–52.

Analýza panelových dat

Petr Novák

Abstrakt

Tento článek má za cíl seznámit zájemce se základními prvky analýzy panelových dat, modely fixních a náhodných efektů, dynamickými modely panelových dat. Poslední část této práce je věnována otázkám možnosti testování jednotkových kořenů modelů panelových dat s poznámkou ohledně používaného aplikačního softwaru a programovacích jazyků při práci s panelovými daty.

Klíčová slova: analýza panelových dat; dynamický model panelových dat; jednotkové kořeny panelových dat.

Panel data analysis

Abstract

This article takes focus on the main basic elements of panel data analysis, fixed effects and random effects models, dynamic panel data models. The last part of this article is about possibilities of testing panel data unit roots with a notice about the usage of the application software and special languages in the area of panel data.

Key words: panel data analysis; dynamic panel data model; panel data unit roots.

JEL classification: G30