POLS/CS&SS 503:
Advanced Quantitative Political Methodology

INTRODUCTION

June 2, 2015

Jeffrey B. Arnold

UNIVERSITY of WASHINGTON
DEPARTMENT OF POLITICAL SCIENCE

CENTER for STATISTICS and the SOCIAL SCIENCES

Panel Data

- What is Panel Data?
- Why use it?
- What are the problems?
- What methods address them?
- Causal inference interpretations?

What is Panel Data?

Example: Garrett (1998) government composition and economic indicators in OECD countries

countryname	year	gdp	infl	unem	capmob	corp
US	1966	5.11	2.90	3.80	0	1.80
US	1967	2.28	2.80	3.80	0	1.81
...						
US	1990	0.90	5.40	5.41	0	2.01
Canada	1966	6.80	3.70	3.60	0	2.27
Canada	1967	2.92	3.60	4.10	0	2.30
...						
Canada	1990	0.40	4.80	8.06	0	1.71
UK	1966	1.88	3.90	1.50	1	2.14
UK	1967	2.26	2.50	2.30	1	2.13
...						
UK	1990	0.80	9.50	5.47	0	2.89
...						

$N = 14$ OECD countries, $T = 25$ years (1966–1990).

What is Panel Data?

- Data (and models) structured into units and periods units

$$y_{i,t} = x_{i,t}\beta + \epsilon_{i,t}$$

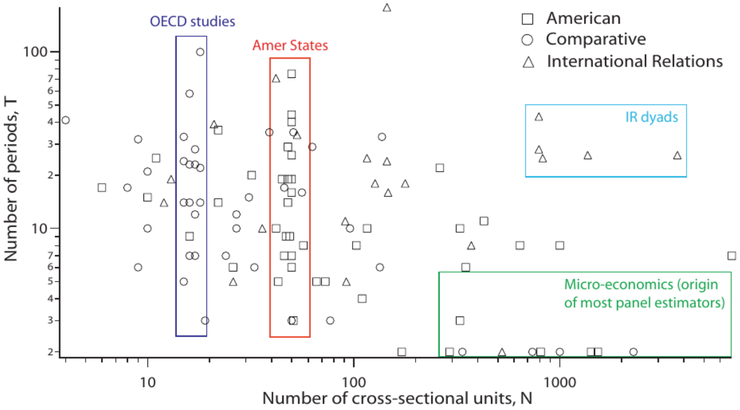
- units $i = 1, \ldots, N$ each observed over $t = 1, \ldots, T$, for a total of $N \times T$ observations.
- balanced data: all units i have same number of observations T
- unbalanced data: units have different values of T (missingness, sample selection)
- some methods may require adjustments if using unbalanced data

What is Panel Data?

Many different names, sometimes different things

- Other names
  - panel data
  - longitudinal
  - time-series cross-section (TSCS)
- But can mean different things with different appropriate methods depending on the size of N and T.

Different N and T in different contexts



What is Panel Data?

Different things

- Size of dimensions can influence which methods are appropriate:
  - Big N, small T (e.g. panel surveys)
  - Small(er) N, big T (e.g. country time series, financial)
- Some methods emphasize unit differences (fixed/random effects, PCSE)
- Others emphasize time (lagged dependent variables, serial correlation)

Why use Panel Data?

- More data, which might make inference more precise (at least if we believe β is the same or similar across units)
- Can help with omitted variables, especially if they are time invariant
- Some analysis only possible with panel data; e.g., if variables don't change much over time, like institutions
- Heterogeneity is interesting! As long as we can specify a general DGP for whole panel, can parameterize and estimate more substantively interesting relationships

Difficulties of Panel Data?

- More complex to conceptualize and model
- Need to worry about issues in *time* and *space*
- Needs more powerful or flexible estimation tools

A Pooled Time Series Model

Example with GDP data

$$\texttt{gdp}_{i,t} = \alpha + \beta_1 \texttt{corp}_{i,t} + \beta_2 \texttt{leftlab}_{i,t} + \beta_3 \texttt{leftlab} \times \texttt{corp}_{i,t} + \beta_4 \texttt{demand}_{i,t}$$

- The model is **pooled** because it assumes β are the same between all countries
- Ignores heterogeneity between units
- Almost always overestimates precision
- However, some amount of pooling is always necessary in a model, the question is how much.

Varying Intercepts Models

$$\texttt{gdp}_{i,t} = \alpha_i + \beta_1 \texttt{corp}_{i,t} + \beta_2 \texttt{leftlab}_{i,t} + \beta_3 \texttt{leftlab} \times \texttt{corp}_{i,t}$$
$$+ \beta_4 \texttt{c}$$

Fixed Effects

No stochastic component of intercepts

$$\alpha_i = \alpha_i^*$$

Random Effects

Intercepts are modeled as coming from a distribution

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

Fixed effects

$$\texttt{gdp}_{i,t} = \alpha_i + \beta_1 \texttt{corp}_{i,t} + \beta_2 \texttt{leftlab}_{i,t} + \beta_3 \texttt{leftlab} \times \texttt{corp}_{i,t}$$

$$+ \beta_4 \texttt{c}$$

- $\alpha_i = \alpha_i^*$ are individual for each country
- α_i can be correlated with $x_{i,t}$. Controls for *all* (known and unknown) time-invariant variables
- Cost: we're purging the cross-sectional variation from the analysis
- Assuming change in $x_{i,t}$ has same response in each series
- Uses over-time variation in covariates to estimate parameters

Estimating Fixed Effects

Within Estimator

$$y_{i,t} - \bar{y}_i = (x_{i,t} - \bar{x}_i)\beta + (\epsilon_{i,t} - \bar{\epsilon}_i)$$

- Differencing absorbs (removes) fixed effects
- Cannot include time-varying
- The “between” estimator

$$\bar{y}_i = \bar{x}_i\beta + \epsilon_i \tag{1}$$

- Does not estimate the fixed effects; only removes them
- Errors are now correlate and standard errors need to adjusted

Estimating Fixed Effects

Dummy Variable Estimator (LSDV)

$$y_{i,t} = x_{i,t}\beta + \alpha_i + u_{i,t}$$

- estimates α_i, which may be useful (suggest omitted variables)
- For large T, similar to within estimator
- For small T, estimates of α will be poor.

Time Varying Covariates and Fixed Effects

- The fixed effects absorb all time-varying covariates so you cannot get separate estimates of them (perfect collinearity).
- Can include *interactions* of time-invariate variables? Estimate how these time-invariate variables *mediate* the effects of other variables.
- Methods that decompose fixed effects in to known and unknown covariates (Plumper and Troeger 2007)
- Use random effects

Random Effects

$$
\begin{aligned}
\texttt{gdp}_{i,t} &= \beta_0 + \beta x_{i,t} + \alpha_i + \epsilon_{i,t} \\
\alpha_i &\sim N(0, \sigma^2_\alpha) \\
\epsilon_{i,t} &\sim N(0, \sigma^2_\epsilon)
\end{aligned}
$$

- Error variance is $\sigma^2_\alpha + \sigma^2_\epsilon$
- Random effects are another part of the error
-

Fixed Effects or Random Effects?

- Random effects are more efficient if Cor$(\alpha_i, x) = 0$, but inconsistent if Cor$(\alpha_i, x) \neq 0$
- Fixed effects are consistent, but less efficient if random effects model is efficient.
- Run Hausmann test on random effects and f (R function `phtest`). H_a is one test is inconsistent (random effects) and means to use fixed effects.
- Use random effects if you want to estimate effects of time-invariant variables.
- Can include group level averages or time-invariant variables in random effects model to approx the fixed effects part.

Implementations

- **plm** R package for panel estimation. Includes random effects, fixed effects.
- **lme4** R package for fixed and random effects. From a statistics background, not specific to panel data.

What makes a panel dynamic?

Static panel model:

$$y_{i,t} = X_{i,t}\beta + \epsilon_{i,t}$$

Dynamic panel model (Lagged dependent variable)

$$y_{i,t} = \phi y_{i,t-1} + X_{i,t}\beta + \epsilon_{i,t}$$

Lagged Dependent Variable

Equivalent to geometrically decaying independent variable

$$
\begin{aligned}
y_{i,t} &= X_{i,t}\beta + \epsilon_{i,t} + \phi y_{i,t-1} \\
&= X_{i,t}\beta + \epsilon_{i,t} + \phi(X_{i,t-1}\beta + \epsilon_{i,t-1} + y_{i,t-1}) \\
&= \sum_{k=0}^{1} \phi^k X_{i,t-k}\beta + \sum_{k=0}^{1} \phi^k \epsilon_{i,t-k} + \phi(X_{i,t-2}\beta + \epsilon_{i,t-2} + y_{i,t-2}) \\
&= \sum_{k=0}^{2} \phi^k X_{i,t-k}\beta + \sum_{k=0}^{2} \phi^k \epsilon_{i,t-k} + \phi(X_{i,t-3}\beta + \epsilon_{i,t-3} + y_{i,t-3}) \\
&\vdots \\
&= \sum_{k=0}^{\infty} \phi^k X_{i,t-k}\beta + \sum_{k=0}^{\infty} \phi^k \epsilon_{i,t-k}
\end{aligned}
$$

- OLS is optimal if $\epsilon_{i,t}$ are IID
- Important that $|\phi| < 1$ (**stationarity**). What would happen if $|\phi_i| > 1$?
- OLS inconsistent if $\epsilon_{i,t}$ are serially correlated.
- If $\epsilon_{i,t}$ are serially correlated, can estimate with appropriate method (Cochrane-Orcutt, MLE)

Lagged Dependent Variables with Fixed Effects

- Lagged DV + fixed effects: **estimates are biased**
- Methods exist to correct for that bias. IV methods of Anderson and Hsiao, Arellano and Bond. Rely on asymptotics. Variance of those estimators much higher.
- However, in most TSCS research, the bias of the RMSE of LS is better than or not much worse than the more complicated estimators.

Autoregressive Distributed Lag

$$y_{i,t} = \beta x_{i,t} + \phi y_{i,t-1} + \gamma x_{i,t-1} + \epsilon_{i,t}$$

- Extremely flexible: nests many different time-series specifications
- Beck and Katz (2011), De Boef and Keele (2008) suggest it a "default" model for TSCS.
- Does not account for fixed effects; these could be added with previous caveats
- Works with stationary and non-stationary data
- Equivalent to another model: **error correction model**
- Can usually estimate with OLS

Serial Correlation

$$y_{i,t} = X\beta_{i,t} + \epsilon_{i,t} + \rho\epsilon_{i,t-1}$$

- Causes issues with standard errors in OLS
- Can model it directly with other methods (Prais-Winsten, Cochrane-Orcutt), or include lagged y (see Beck-Katz 2011, p. 339)
- If serial correlation and lag DV, then add y_{t-2} (Beck-Katz 2011)

Panel-corrected standard errors

- What PCSE account for:
  - Heteroskedasticity between units,

$$\mathsf{V}(\epsilon_{USA}) \neq \mathsf{V}(\epsilon_{CAN}) \tag{2}$$

  - Contemporaneous correlation between units,

$$\mathsf{Cor}(\epsilon_{USA,1990}, \epsilon_{CAN,1990}) \neq 0 \tag{3}$$

- They do not account for serial correlation or non-contemporaneous correlations.

$$\mathsf{Cor}(\epsilon_{USA,1990}, \epsilon_{USA,1991}) = 0$$
$$\mathsf{Cor}(\epsilon_{USA,1990}, \epsilon_{CAN,1991}) = 0$$

Panel-corrected standard errors

- Suggest using OLS with PCSE and lagged DV as a baseline model
- Many think that fixed effects should also be used
- PCSE (and other error corrections) are 2nd order to getting lag structure and including fixed effects where appropriate
- Implementations: R packages **pcse**, **plm** (vcovBK)

Panel-corrected Standard Errors

How to adjust the standard errors?

- Replace variance-covariance matrix used in calculating standard errors

$$\mathsf{c}(\beta) = (X'X)^{-1}(X'\Omega)(X'X)^{-1}$$

- Linear regression with classical SE, $\Omega = \sigma^2 I_N$, so

$$\mathsf{c}(\beta) = \sigma^2(X'X)^{-1}$$

- In PCSE, Ω is $NT \times NT$ block-diagonal matrix with $N \times N$ matrix Σ of contemporaneous covariances on the diagonal.

PCSE

What is the variance-covariance matrix?

$$\Sigma_N = \begin{bmatrix} \sigma^2_{\epsilon_1} & \sigma_{\epsilon_1,\epsilon_2} & \dots & \sigma_{\epsilon_1,\epsilon_N} \\ \sigma_{\epsilon_1,\epsilon_2} & \sigma^2_{\epsilon_2} & \dots & \sigma_{\epsilon_1,\epsilon_N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\epsilon_1,\epsilon_N} & \sigma_{\epsilon_2,\epsilon_N} & \dots & \sigma^2_{\epsilon_N} \end{bmatrix}$$

$$\Omega_{NT\times NT} = \begin{bmatrix} \Sigma_N & 0_N & \dots & 0_N \\ 0_N & \Sigma_N & \dots & 0_N \\ \vdots & \vdots & \ddots & \vdots \\ 0_N & 0_N & \dots & \Sigma_N \end{bmatrix} = \Sigma_N \otimes I_T$$

How to estimate the matrix Σ?

- Suppose that the panel is balanced,
- Estimate OLS, and then use residuals to estimate Σ,

$$\hat{\Sigma}_{i,j} = \sum_{t=1}^{T} \frac{E_{i,t}E_{j,t}}{T}$$

- Plug in $\hat{\Sigma}$ to calculate the covariance matrix
- This is possible, but notation more tedious in unbalanced panels.

- Old-school Beck and Katz (1995): lagged dependent variable + PCSE
- New-school Beck and Katz (2011), De Boef and Keele (2008)
    - ADL or ECM model
    - Try fixed effects (OLS will probably be fine as long as T not too small)
    - Try not to use error corrections to avoid thinking about dynamics
- Angrist and Pischke:
    - lagged dependent variable and fixed effects bound the effect of X: try both

There is no advice

- Know your data
- Know your model
- Ensure your results are robust
- Think!

References

- Christopher Adolph, “Introduction to Panel Data Analysis” [lecture slides], POLS 503, Spring 2014. |http://faculty.washington.edu/cadolph/503/topic9.pw.pdf|.
- De Boef, Suzanna, and Luke Keele. 2008. “Taking Time Seriously.” *American Journal of Political Science* 52(1): 184–200. |http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5907.2007.00307.x/abstract|.
- Beck, Nathaniel, and Jonathan N. Katz. 2011. “Modeling Dynamics in Time-Series–Cross-Section Political Economy Data.” *Annual Review of Political Science* 14(1): 331–52. |http://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-071510-103222|.