

Longitudinal/Panel Data Analysis: Lecture 4

Raymond Duch

University of Oxford
Nuffield College
raymond.duch@nuffield.ox.ac.uk
raymond Duch.com/class/2011/paneldata

May 23, 2011

- Week 1: Introduction to panel data
- Week 2: Longitudinal Panels, Fixed, Random Effects, Dynamic Models
- Week 3: Hierarchical nested data sets
- Week 4: Cross section time series

- Stata 10.0 Manual Longitudinal/Panel Data, xtabond, xtabond postestimation, xtdpdsys, xtivreg
- Beck, Nathaniel and Jonathan N. Katz. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments" Political Analysis 15: 182-195.
-
-
- Plumper, Thomas, Troeger, Vera E. and Philip Manow 2005: Panel Data Analysis in Comparative Politics. Linking Method to Theory: European Journal of Political Research 44: 327-354.
- Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data" Political Analysis 15: 165-181.

Two critical issues in TSCS models.

- Dynamics
- Heterogeneity

TSCS data are a particular type of multilevel data where years are the level 1 units and group (e.g., countries, states, etc...) are the level 2 units

- Lots of times, with annual aggregate data, we are unable to take full advantage of all that time-series has to offer us. This is not to say that problems don't exist, rather it is to say that with relatively small T , we will be less sure about how to fix these things
- Some problems that are “easy” to deal with in single-time-series are not nearly as easily dispensed with in TSCS applications (e.g., stationarity)

We will focus mainly on the lagged outcome variable (LDV) model.

- As Beck and Katz (2004) suggest, there is generally little reason to prefer an AR(1) model to the LDV model

- One problem for inference with time-series data is serial correlation in the errors.
- We can deal with this if we properly model the serial correlation either through an autoregressive parameter (AR) or include a lagged outcome variable on the RHS of the model.
- Testing for serial correlation - you can use a Lagrange Multiplier Test by doing the following:
 - 1 Run OLS
 - 2 Compute residuals
 - 3 Regress residuals on all explanatory variables and the lagged residual
 - 4 If the coefficient on the lagged residual is significant, we can reject the null of independent errors

- Fixed-effects models are easy to run in stata
- You can get the results you want either by using the “demeaning” feature of the FE option in Stata
- It is also possible to include a lagged outcome variable if you like.

TSCS Fixed Effect with Lag in Stata

```
. xtreg balance l.balance pop1564 electpred developed, fe  
note: developed omitted because of collinearity
```

```
Fixed-effects (within) regression      Number of obs   =   1939  
Group variable: cc                    Number of groups =    68
```

```
R-sq:  within = 0.4487                Obs per group: min =    5  
        between = 0.9630              avg =           28.5  
        overall = 0.6302              max =           40
```

```
corr(u_i, Xb) = 0.5748                F(3,1868)       =   506.81  
                                                Prob > F        =    0.0000
```

```
-----  
      balance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]  
-----+-----  
      balance |  
      L1.     |   .6485425   .016657    38.94  0.000    .6158743   .6812107  
      |  
      pop1564 |   .0257448   .020502     1.26  0.209   -.0144645   .0659541  
      electpred |  -.311099   .1432059   -2.17  0.030   -.5919594  -.0302385  
      developed | (omitted)  
      _cons   |   -2.3161    1.25083    -1.85  0.064   -4.769272   .1370717  
-----+-----  
      sigma_u |   .82391353  
      sigma_e |   2.1868295  
      rho     |   .12430442   (fraction of variance due to u_i)
```

```
-----  
F test that all u_i=0:      F(67, 1868) =    2.86      Prob > F = 0.0000
```

Fixed Effects with “Sluggish” variables

- One of the big problems people have with fixed effects models is that any heterogeneity that is a function of non-time-varying variables (or slowly changing variables) will be captured by the unit-specific intercepts (i.e., the fixed effects)
- Until recently, there has been little work done on how we might remedy this problem, the advice was either accept it, or use random effects
- I've collected a fair amount of anecdotal evidence suggesting that this is the main reason that people choose random effects over fixed effects, regardless of how reasonable the assumptions are
- Plümper and Troeger (2007) suggest an estimator that will essentially permit an estimation of the fixed effects and more importantly, the effect that non-time-varying variables have on the fixed effects

- Let's assume we have the following DGP:

$$y_{it} = \alpha + \sum_{k=1}^K \beta_k x_{kit} + \sum_{m=1}^M \gamma_m z_{mi} + c_i + \varepsilon_{it}$$

where the z_{mi} 's are non-time-varying variables.

- It might go without saying, though I won't let it, that if the c_i 's were a perfect linear combination of the z_{mi} 's, then we wouldn't need to estimate the c_i 's.
 - That is to say, that if we could account for all of the differences between units with observed variables, we wouldn't have to worry about fixed effects (i.e., we would have no *unobserved* heterogeneity, it would all be observed)

Fixed Effects Vector Decomposition (FEVD)

The model is estimated in three stages

- Estimate the standard fixed effects model to obtain estimates of the unit effects:

$$y_{it} = \sum_{k=1}^K \beta_k^{FE} x_{it} + e_{it}$$

$$\hat{c}_i = \bar{y}_i - \sum_{k=1}^K \beta_k^{FE} \bar{x}_{ki} - \bar{e}_i$$

- Regress the estimated unit effects \hat{c}_i on the non-time-varying and slowly changing variables:

$$\hat{c}_i = \sum_{m=1}^M \gamma_m z_{mi} + h_i$$

where h_i is that part of the unit effect that cannot be explained by the $n - t - v$ variables

- In the third stage, we add back in the RHS from stage 2:

$$y_{it} = \alpha + \sum_{k=1}^K \beta_k x_{kit} + \sum_{m=1}^M \gamma_m z_{mi} + \delta h_i + \varepsilon_{it}$$

- We do this for two reasons:
 - 1 This provides the right number of degrees of freedom for statistical tests
 - 2 Also, this model allows for fixing problems with autocorrelation (by adding a lagged DV) and/or heteroskedasticity (by using a robust variance estimator in the form of a White HCCM or PCSE's)

- Fixed effects allow arbitrary correlation between the unit effects and the x_{it} which makes them a less restrictive form of estimation
- Through recent developments, the effects of non-time-varying variables can be captured through the fixed-effects vector decomposition
- With TSCS data of any kind, pains must be taken to “fix” any problems in the data such as serial correlation and unmodelled heterogeneity

- Random effects do not model c_i directly, rather they estimate the mean and variance of a distribution for c_i .
- Wooldridge (2002) suggests that if the units are exchangeable (i.e., their names are irrelevant), then random effects often make sense. If the units are not exchangeable, they are interesting in their own right, then random effects are probably not the right model
- The general model takes this form:

$$y_{it} = x_{it} + \nu_{it} \text{ where } \nu_{it} = c_i + u_{it}$$

- In random effects models, we're basically acknowledging the fact that the errors are not *iid*, but we're not interested in estimating the unit effects

Structure of Error Variance

- We are assuming that the variance-covariance matrix of the errors Ω is not diagonal, rather there is some non-independence in the following form:

$$\Omega = E(v_i v_i') = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & \dots & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{pmatrix}$$

- Since c_j is in the error term, this requires the zero correlation between x_{it} and c_j
- What we are gaining from this more restrictive assumption is a lot of degrees of freedom.

Estimating the Random Effects Model

- We can use feasible generalized least squares (FGLS) [not to be confused with GLM] to estimate the random effects model
- The FGLS estimator of β is:

$$\hat{\beta}_{RE} = \left(\sum_{i=1}^N X_i' \hat{\Omega}^{-1} X_i \right)^{-1} \left(\sum_{i=1}^N X_i' \hat{\Omega}^{-1} y_i \right)$$

- This is akin to weighted least squares, but the “weight” matrix is not diagonal
- You could also do a less restrictive FGLS estimation where you are attempting to estimate all of the off-diagonal elements of Ω to be different, but this takes a *lot* of data.

Hausman Test for FE vs RE

- Hausman proposed a test of the difference between the RE and FE estimates.
- Before we get to the statistic, there are 2 caveats:
 - 1 Strict exogeneity is assumed
 - 2 The test assumes that under the null - zero covariance between x_i and c_i , so it is not actually testing this assumption.
- The test looks at the difference between the two coefficient vectors as FE is consistent even if the x_i and c_i are correlated, but RE is more efficient when they are not.

$$H = (b - B)' (V(b) - V(B))^{-1} (b - B)$$

where b and $V(b)$ come from the FE model, B and $V(B)$ are from the RE model, and $H \sim \chi^2$ with k (length of b) degrees of freedom

TSCS Random Intercept with Lag in XTMIXED

```
. xtmixed balance l.balance pop1564 electpred developed || cc:
Mixed-effects REML regression      Number of obs      =      1939
Group variable: cc                 Number of groups   =         68

                                   Obs per group: min =         5
                                   avg      =      28.5
                                   max      =         40
                                   Wald chi2(4)    =    2037.72
                                   Prob > chi2     =         0.0000

Log restricted-likelihood = -4317.3869
```

```
-----+-----
      balance |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
      balance |
      L1.     |   .6964841   .015464   45.04   0.000   .6661753   .726793
      |
      pop1564 |   .0198841   .0153968   1.29   0.197   -.0102931   .0500613
      electpred |  -.2782937   .142441   -1.95   0.051   -.5574729   .0008854
      developed |  -.073381   .2134494   -0.34   0.731   -.4917342   .3449723
      _cons    | -1.832513   .9125965   -2.01   0.045   -3.621169   -.0438564
-----+-----
```

```
-----+-----
      Random-effects Parameters |   Estimate   Std. Err.    [95% Conf. Interval]
-----+-----
cc: Identity
      sd(_cons) |   .6299414   .0893588    .4770403   .8318505
-----+-----
      sd(Residual) |   2.189215   .0359145    2.119943   2.26075
-----+-----
```

```
LR test vs. linear regression: chibar2(01) =    42.03 Prob >= chibar2 = 0.0000
```

Fixed Effects Vector Decomposition (FEVD)

```
egen avgbalance = mean(balance), by(cc)
egen avgpop1564 = mean(pop1564), by(cc)
egen avgelectpred = mean(electpred), by(cc)

xtreg balance l.balance pop1564 electpred, fe
predict e_hat, e
egen avgresidual = mean(e_hat), by(cc)

predict yhat
gen c_hat=avgbalance - (_b[_cons] + _b[l.balance]*avgbalance + _b[pop1564]*avgpop1564 + _b[electpred]*avg

[OR PREDICT C_HAT, U]

regr c_hat developed

predict h_hat, residual

reg balance l.balance pop1564 electpred developed h_hat
```

Fixed Effects Vector Decomposition (FEVD)

```
. reg balance l.balance pop1564 electpred developed h_hat
```

Source	SS	df	MS	Number of obs =	1939
Model	17688.3343	5	3537.66686	F(5, 1933) =	758.25
Residual	9018.59567	1933	4.66559528	Prob > F =	0.0000
Total	26706.93	1938	13.7806656	R-squared =	0.6623
				Adj R-squared =	0.6614
				Root MSE =	2.16

balance	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
balance						
L1.	.6475123	.0164023	39.48	0.000	.6153442	.6796804
pop1564	.0212201	.0115756	1.83	0.067	-.0014818	.043922
electpred	-.2947447	.1386743	-2.13	0.034	-.5667117	-.0227777
developed	-.0703401	.1264835	-0.56	0.578	-.3183985	.1777184
h_hat	1.02207	.0765381	13.35	0.000	.8719643	1.172176
_cons	-2.01946	.6772957	-2.98	0.003	-3.347767	-.6911529