Longitudinal/Panel Data Analysis: Lecture 2

Raymond Duch

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

May 5, 2010

Readings

- Stata 10.0 Manual Longitudinal/Panel Data, xtabond, xtabond postestimation, xtdpdsys, xtivreg
- Green, Donald P. and David H. Yoon. 2002. "Reconciling Individual and Aggregate Evidence Concerning Partisan Stability: Applying Time-Series Models to Panel Survey Data." Political Analysis 10:1-24.
- Wawro, Gregory. 2002. "Estimating Dynamic Panel Data Models in Political Science." Political Analysis 10:25-48.
- Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." Annual Review of Sociology. 30:507-44.
- Beck, Nathaniel and Jonathan N. Katz. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." Political Analysis. 15: 182-195.

Readings

- Wilson, Sven E. and Daniel M. Butler. 2007. "A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications." Political Analysis. 15:101-123.
- Plumper, Thomas and Vera E. Troeger. 2007. "Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects." Political Analysis 15:124-139.
- 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.

Panel Data, Unobservables and Causality

- Leveraging panel data to solve problems of causal inference
- Problems of unobservables in non-randomized studies
  - time-invariant unit-specific unobservables
  - time-varying unit-specific unobservables that represent transitory and idiosyncratic forces acting upon units

Cross-sectional Data

Consider the following data and treatment condition:

- cross-sectional data ($t = 1$)
- causal variable (or treatment) d_i scored 1 for the treatment group and 0 for the control group
- the treatment occurs a period, τ, prior to $t = 1$
- y and d are the only observed variables
- γ is a parameter for the causal effect
- δ_1 is a period effect common to all units
- $\theta_{i|d}$ is a time-invariant unit-specific effect that captures unobserved unit heterogeneity and is conditional on d
- $\epsilon_{i1|d}$ is white noise unique to the ith unit at $t = 1$ conditional on d and θ_i

Cross-sectional Data

The control group:

$$y_{i1|d=0} = \theta_{i|d=0} + \delta_1 + \epsilon_{i1|d=0} \tag{1}$$

The treatment group:

$$y_{i1|d=1} = \gamma + \theta_{i|d=1} + \delta_1 + \epsilon_{i1|d=1} \tag{2}$$

The treatment effect?

$$E(y_{i1|d=1} - y_{i1|d=0}) = \gamma + E(\theta_{i|d=1} - \theta_{i|d=0}) + E(\epsilon_{i1|d=1} - \epsilon_{i1|d=0}) \tag{3}$$

Cross-sectional Data: Important Assumptions

- δ_1 drops out
- mean of disturbances is independent of d: $E(\epsilon_{1t|d=1} - \epsilon_{i1|d=0}) = 0$
- unobserved heterogeneity is mean independent of the causal variable: $[E(\theta_{i|d=1}) = E(\theta_{i|d=0})]$

which implies the OLS regression:

$$y_{i1} = \alpha + \gamma d_i + \mu_{i1} \tag{4}$$

where $\mu_{i1} = \theta_{i|d} + \epsilon_{i1|d}$

The control group:

$$y_{i0|d=0} = \theta_{i|d=0} + \delta_0 + \epsilon_{i0|d=0} \tag{5}$$

$$y_{i1|d=0} = \theta_{i|d=0} + \delta_1 + \epsilon_{i1|d=0} \tag{6}$$

The treatment group:

$$y_{i0|d=1} = \gamma + \theta_{i|d=1} + \delta_0 + \epsilon_{i0|d=1} \tag{7}$$

$$y_{i1|d=1} = \gamma + \theta_{i|d=1} + \delta_1 + \epsilon_{i1|d=1} \tag{8}$$

Panel Design

The treatment effect?

$$E(y_{i1|d=1} - y_{i1|d=0}) - E(y_{i0|d=1} - y_{i0|d=0}) = \gamma +$$

$$\mathrm{E}(\epsilon_{i1|d=1} - \epsilon_{i0|d=1}) - E(\epsilon_{i1|d=1} - \epsilon_{i0|d=0})$$

(9)

- unit effects that were source of heterogeneity bias are eliminated
- identification does not require assumption that period effects are temporally stable
- exogeneity identification restriction holds:
- $E(\epsilon_{i1|d=1} - \epsilon_{i0|d=1}) - E(\epsilon_{i1|d=1} - \epsilon_{i0|d=0}) = 0$

Regression form of the posttest and pretest models

The control and treatment groups:

$$y_{i0} = \delta_0 + \theta_{i|d} + \epsilon_{i0} \tag{10}$$

$$y_{i1} = \delta_1 + \gamma d_i + \theta_{i|d} + \epsilon_{i1} \tag{11}$$

Differencing then gives you:

$$(y_{i1} - y_{i0}) = (\delta_1 - \delta_0) + \gamma d_i + (\epsilon_{i1} - \epsilon_{i0}) \tag{12}$$

Least squares gives an estimate of treatment effect:

$$\hat{\gamma}_{dd} = (\bar{y}_{.1|d=1} - \bar{y}_{.1|d=0}) - (\bar{y}_{.0|d=1} - \bar{y}_{.0|d=0}) \tag{13}$$

Fixed Effects Estimator

Deviations from within-unit means:

$$\bar{y}_{i.} = (\delta_1 + \delta_0)/2 + \gamma\bar{d}_{i.} + \theta_i\hat{\epsilon}_i \tag{14}$$

Consistent fixed effects estimator:

$$(y_{it} - \bar{y}_i) = (\delta_0 - \delta_1)/2 + (\delta_1 - \delta_0)p_1 + \gamma(d_{it} - \bar{d}_i) + (\epsilon_{it} - \bar{\epsilon}_i) \tag{15}$$

Static Panel Models

$$y_{it} = \alpha + \sum_k \beta_k \omega_{kit} + \sum_p \phi_p z_{pi} + \gamma x_{jt} + \theta_i + \epsilon_{it} \qquad (16)$$

- $i = 1, ..., N$ t=1,...,T

- θ_i is a term for unit effects
- causal variable of interest is x_{it}
- ω_{kit} $(k = 1, ..., K)$ consists of additional explanatory variables that vary over time
- z_{pi} consists of additional explanatory variables that do not vary over time

Estimation

- Should θ_i be treated as random or fixed?
- if unobserved θ_i are uncorrelated with regressors random effects is reasonable

- serial correlation of composite errors suggests efficiency gains from GLS
- but if unit effects correlated with explanatory variables estimators may be biased and inconsistent

First Differencing and Time-Demeaning

$$(y_{it} - y_{it-1}) = \sum_k \beta_k(\omega_{kit} - \omega_{kit-1}) + \gamma(x_{it} - x_{it-1}) + (\epsilon_{it} - \epsilon_{it-1}) \tag{17}$$

$$(y_{it} - \bar{y_i}) = \sum_k \beta_k(\omega_{kit} - \bar{\omega_{ki}}) + \gamma(x_{it} - \bar{x_i}) + (\epsilon_{it} - \bar{\epsilon_i}) \tag{18}$$

If the disturbances in the transformed equations are constant variance and serially uncorrelated then both fixed effects and difference estimators are efficient for a fixed effects model.

Time-Invariant Explanatory Variables

$$y_{i1} = \delta_1 + \phi_1 z_i + \gamma(x_{i1}) + \theta_i + \epsilon_{i1} \quad (19)$$

$$y_{i2} = \delta_2 + \phi_2 z_i + \gamma(x_{i2}) + \theta_i + \epsilon_{i2} \quad (20)$$

$$(y_{i2} - y_{i1}) = (\delta_2 - \delta_1) + (\phi_2 - \phi_1)z_i + \gamma(x_{i2} - x_{i1}) + (\epsilon_{i2} - \epsilon_{i1}) \quad (21)$$

The change in $(\phi_2 - \phi_1)$ is estimable as are time-invariant variables interacted with time trends or periods.

Hausman Test

$$y_{it} = \alpha + \sum_k \beta_k \omega_{kit} + \sum_p \phi_p z_{ip} + \gamma x_{it} + \alpha_i + \epsilon_{it} \qquad (22)$$

We can estimate γ employing either a fixed effect (γ_{fe}) estimator or a random effect (γ_{re}) estimator. Hausman showed that $(\hat{\gamma_{fe}} - \hat{\gamma_{re}})$ could be used to test the null hypothesis that the unit effects and the explanatory variables are uncorrelated.

Small values of Hausman Test indicate a failure to reject null hypothesis that the unit effects and the explanatory variables are uncorrelated and favour the random effects estimation of fixed effects models.

$$y_{it} = \phi_1 z_{1i} + \phi_2 z_{2i} + \gamma_1 x_{1it} + \gamma_2 x_{2it} + \theta_i + \epsilon_{it} \tag{23}$$

- all explanatory variables are strictly exogenous
- z_{2i} and x_{2it} (but not z_{1i} and x_{1it}) are correlated with θ_i
- unbiased estimates of ϕ_1 and γ_1 is unproblematic because z_{1i} and x_{1it} are exogenous and therefore act as their own instruments
- $(x_{2it} - \bar{x_2}_i)$ is an instrument for x_{2it}
- $\bar{x_1}_i$ is an instrument for z_2

```
. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed, endog(exp exp2 wks ms union ed)

Hausman-Taylor estimation                       Number of obs      =      4165
Group variable: id                              Number of groups   =       595
                                                Obs per group: min =         7
                                                               avg =         7
                                                               max =         7
Random effects u_i ~ i.i.d.                     Wald chi2(12)      =   6891.87
                                                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TVexogenous  |
         occ |  -.0207047   .0137809    -1.50   0.133    -.0477149    .0063055
       south |   .0074398    .031955     0.23   0.816    -.0551908    .0700705
        smsa |  -.0418334   .0189581    -2.21   0.027    -.0789906   -.0046761
         ind |   .0136039   .0152374     0.89   0.372    -.0162608    .0434686
TVendogenous |
         exp |   .1131328    .002471    45.79   0.000     .1082898    .1179758
        exp2 |  -.0004189   .0000546    -7.67   0.000    -.0005259   -.0003119
         wks |   .0008374   .0005997     1.40   0.163    -.0003381    .0020129
          ms |  -.0298508     .01898    -1.57   0.116    -.0670508    .0073493
       union |   .0327714   .0149084     2.20   0.028     .0035514    .0619914
TIexogenous  |
         fem |  -.1309236    .126659    -1.03   0.301    -.3791707    .1173234
         blk |  -.2857479   .1557019    -1.84   0.066    -.5909179    .0194221
TIendogenous |
          ed |    .137944   .0212485     6.49   0.000     .0962977    .1795902
             |
       _cons |   2.912726   .2836522    10.27   0.000     2.356778    3.468674
-------------+----------------------------------------------------------------
     sigma_u |  .94180304
     sigma_e |  .15180273
         rho |  .97467788   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Note:  TV refers to time varying; TI refers to time invariant.
```