Pooled Time-Series and Cross-Sectional Data

Introduction
Fixed and Random Effects

What is Panel/Pooled data?

- We will be dealing with data that follows a given sample of units (individuals, countries, dyads, etc), i = 1, 2,…, N, over time, t = 1, 2,…,T, so that we have multiple observations ($N*T$) on each unit over time.
- The convention is to refer to this data as either panel data or pooled cross sectional time series data.

Panel Data

- Panel data often refers to a data set where the observations are dominated by large numbers of units (i) relative to time periods (t). These units are (typically) a random sample – the idiosyncratic differences across individuals are not of interest (the features of person j and k are assumed to be identical).
- The most commonly known panel data in Political Science is probably the National Election study. These studies observe over 2000 individuals over three (at this point in time) time points.
- Key idea is that asymptotics hold as T approaches infinity as N is thought of as fixed.

Pooled Time Series and Cross Sectional Data

- PTSCS data is either dominated by time OR simply has fewer units than the typical panel data set relative to the number of time periods.
- Examples include studies of dyads, countries, states observed over periods of time that are longer relative to the number of units.
- Key idea is that we think of *N* as fixed and the asymptotics are in *T*
- But…there are specific considerations where the PTSCS data look more like panel data (short and wide data)

Other Language

- Repeated Measures Data: usually used in biostatistics; can mean either panel or ptscs data
- Longitudinal Data: usually means very wide, very short data. Used in sociology/psychology in reference to survey data.

Organization of Data

- Easy way to think about it: think of a simple cross section for all units *i* at time *t*.
- Take these cross-sections and "stack" them on top of one another.
- Note:
  - The cross sections do not have to have identical units
  - The distance between times *t* does not have to be identical
  - You can have variables that are constant for unit *i* over time.
- In stata this is known has having the data in "long" format

Example: Globalization and Human Development blmt5.dta

- Dataset is from a joint project with Mewhinney, Teets and Brown
- Focus is on role that debt plays in the ability of governments to provide public goods to their citizens
  - Dependent variables measures illiteracy, health, water quality, etc
  - Independent variables of interest: measures of external debt
  - Control variables include domestic political and economic conditions
- Variables are country averages for four five-year periods from 1980-2000 for between 80 & 185 countries

- list code quin TOTALDE DPTimm devdum
- tsset code quin

```
•  839.  ZAF      1          .       74.5         0
•  840.  ZAF      2          .       71.6         0
•        ---------------------------------------------
•  841.  ZAF      3    16.2505       77.2         0
•  842.  ZAF      4   18.13934       74.6         1
•  843.  ZAR      1   47.16431          .         1
•  844.  ZAR      2   106.0707          .         1
•  845.  ZAR      3   142.8818          .         1
•        ---------------------------------------------
•  846.  ZAR      4   234.1696          .         0
•  847.  ZMB      1   111.9932         54         1
•  848.  ZMB      2   273.0919       76.2         1
•  849.  ZMB      3     226.05       85.8         1
•  850.  ZMB      4   209.5058      87.18         1
•        ---------------------------------------------
•  851.  ZWE      1   23.03842       49.5         1
•  852.  ZWE      2   40.14788       78.8         1
```

Why Use PTSCS Data?

1. Structure of the question: Often we are interested in explicit comparisons: how are nations different? Examining these differences over time allow for dynamic comparisons.
2. We can increase our theoretical leverage on a question with PCSTS data. It may be more appropriate to generalize to a population by pooling units over time.
3. We can increase our statistical leverage. Often events of interest are relatively rare events (on the right hand side). Pooling increases our degrees of freedom though at a cost (and benefit) of increased heterogeneity

Variation in TSCS

- Variation in TSCS data can occur over units, over time, or over both. In the case of our example, variation in debt can occur within a country over time, across countries at a single point in time or both.

```
. xtsum TOTALDEBT

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
TOTALD~p overall |  81.30874   122.9437   .1736625    2094.39 |     N =     469
         between |             91.24547   5.236095   773.5627 |     n =     133
         within  |             80.16914  -558.9961   1402.136 | T-bar = 3.52632
```

This says that the sd of debt between countries is larger than the variation within countries over time. More on this later.

OLS and Pooled Designs

- Consider a simple pooled model

$$y_{it} = \alpha + \mathbf{x}_{it}\beta + \varepsilon_{it}$$

- This model assumes:
  - All the usual OLS assumptions are not violated
  - The constant is constant across all units *i*
  - That the effect of any given X on Y is constant across observations (assuming, of course, that there are no interactions in **X**).
- These last two items are crucial; they are at the heart of specification problems/omitted variable bias. In TSCS models they are likely to be a problem because we have heterogeneity across units and over time.

Variable Intercepts

- One possible violation of the above assumptions is that the intercepts vary. The easiest way to write this is as a model where the units have individual intercepts:

$$y_{it} = \sum_{i=1}^{I} \alpha_{0i} + \mathbf{x}_{it}\beta + \varepsilon_{it}$$

- The slopes over each unit are the same but the intercepts are different. We can also write this in such a way that the intercepts vary over time

$$y_{it} = \sum_{t=1}^{T} \alpha_{0t} + \mathbf{x}_{it}\beta + \varepsilon_{it}$$

- We can also write this so that the intercepts vary over time and unit. The key is that if the data are really generated by either of the above equations and we estimate a model with homogenous intercepts then we can get biased estimates.

Variable Slopes

- The other possibility is that we have a constant intercept but that the effects of **X** on Y differs across either units or time

$$y_{it} = \alpha_0 + \mathbf{x}_{it} \sum_{i=1}^{I} \beta_i + \varepsilon_{it}$$

- We can also have variation in the slopes over time

$$y_{it} = \alpha_0 + \mathbf{x}_{it} \sum_{t=1}^{T} \beta_t + \varepsilon_{it}$$

- We can also have slopes that vary over both units and time.
- We can have slopes and intercepts that vary over both dimensions – but…WHAT?

The Error Term

- All the above models assume that the error term is homoscedastic and uncorrelated both (a) within *i* and (b) across *t*.
- This assumption is violated all the time.
- Dealing with these violations are at the heart of tscs models.
- Approaches include
  - Fixed and random effects
  - GLS and PCSEs
  - Dynamic panel models
  - Panel models for non-normal dependent variables

A Little Stata

- Tell stata that you have tscs data
  - `tsset i t   */i=numeric variable identifying unit/*`

```
.  tsset
       panel variable:  code, 2 to 215
        time variable:  quin, 1 to 4
```

- `sort` command: sorts the data by any variable
- `expand` command: creates multiple copies of the observations already in memory. This is useful if you are adding observations where some of the variables do not vary over time.
- `reshape` command: allows conversion between wide and long formats.
- `stack` command: allows you to `stack' existing variables into a single column.

Dealing with (modeling?) Heterogeneity

- Consider the model with individual (unit) effects; the variable intercept model

$$y_{it} = \sum_{i=1}^{I} \alpha_{0i} + \mathbf{x}_{it}\beta + \varepsilon_{it}$$

  - this is the called (Hsiao 2002) the 'variable intercepts' model and can be interpreted in the context that the conditional mean of y varies across units (or time if we subscript with t).
- A variable intercepts model can be motivated by reference to an underlying model of individual heterogeneity…or, as a nod to 'controlling' for omitted variable bias. Hsiao argues that we can think of this unmeasured heterogeneity arising from three sources:
  - unit-varying, time-constant variables (γ**V**)
  - unit-constant, time-varying variables (δ**W**)
  - variables that vary over both time and unit (β**X**)

- If we do not have variables to measure **V** and **W** we can consider their "combined" (or average) effect. This leads to the following model with time and unit specific intercepts:

$$y_{it} = \alpha + \mathbf{x}_{it}\beta + \gamma_i + \delta_t + \varepsilon_{it}$$

- We can estimate this model in a few different ways.

Fixed Effects Models

- Treating the unit effects as a fixed value is the simplest thing we can do. We can proceed by including N-1 separate indicator (dummy) variables for each unit along with the **x**s.
  - Note: this is identical to analysis of covariance and is the same as ANOVA if we drop the **x**s. If we add both unit and time effects then we have two-way ANCOVA.
  - Note: this is also called least squares with dummy variables (LSDV)
- In panel data if N is large relative to T then we have lots (and lots) of individual intercepts to estimate; consequently they will not be estimated very accurately (large se) but that should not matter as they can be thought of as "nuisance parameters."

Estimating LSDV Models 1

- Include a set of unit dummy variables

```
tab code, gen(code_dum)
```

-this will generate a set of N dummy variables; one corresponding to each unit

-include them in a regression (stata will drop one automatically so that it can estimate a constant)

```
reg y x code_dum*
```

- Interpretation of the dummy variables is straight-forward: each intercept says that the unit's average value of y is higher or lower than that of the omitted unit.
- The same can be done for time

Estimating LSDV Models 2

- We can remove the unit-specific effect from the data prior to estimation as well. We can do this by recoding (rescoring) each variable as a deviation from the unit average.
  - Think of it this way: regress y on the set of unit intercepts and collect the residuals. These residuals will not have the average value of the units included.
  - If we do this for y and the **x**s then unit-specific heterogeneity will be cleaned removed from the data.
- Stata can do this in two ways

xtreg, fe

```
xtreg with the fe option
.  xtreg  illiteracyrateTOTAL TOTALD GNPC , fe

Fixed-effects (within) regression               Number of obs      =       392
Group variable (i): code                        Number of groups   =       109

R-sq:  within  = 0.0312                         Obs per group: min =         1
       between = 0.0004                                        avg =       3.6
       overall = 0.0005                                        max =         4
                                                   F(2,281)           =      4.52
corr(u_i, Xb)  = -0.0266                        Prob > F           =    0.0117

------------------------------------------------------------------------------
illiteracy~L |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TOTALDEBTgnp |  -.0094478   .0032245    -2.93   0.004     -.015795   -.0031006
      GNPCAP |  -.0001512    .000181    -0.84   0.404    -.0005075    .0002051
       _cons |   33.76344   .4695798    71.90   0.000      32.8391    34.68779
-------------+----------------------------------------------------------------
     sigma_u |  24.756631
     sigma_e |  5.4462951
         rho |  .95383706   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(108, 281) =    63.24           Prob > F = 0.0000
```

areg

```
. areg  illiteracyrateTOTAL TOTALD GNPC , a(code)

                                                       Number of obs =     392
                                                       F(  2,   281) =    4.52
                                                       Prob > F      =  0.0117
                                                       R-squared     =  0.9656
                                                       Adj R-squared =  0.9522
                                                       Root MSE      =  5.4463

------------------------------------------------------------------------------
illiteracy~L |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TOTALDEBTgnp |  -.0094478   .0032245    -2.93   0.004     -.015795   -.0031006
      GNPCAP |  -.0001512    .000181    -0.84   0.404    -.0005075    .0002051
       _cons |   33.76344   .4695798    71.90   0.000      32.8391    34.68779
-------------+----------------------------------------------------------------
        code |       F(108, 281) =     63.237   0.000         (109 categories)
```

xtreg v areg

- Both commands absorb or condition out the "nuisance" parameters which (a) makes estimation easier and (b) improves the consistency of the estimated effects.
- One disadvantage is that the intercepts are useful from a diagnostic point of view; they may indicate that there are outliers.
- All three approaches (LSDV included) do provide F-tests for the joint significance of the unit effects.

Advantages and Disadvantages of LSDV

- Advantages
  - if you do not then you may end up with specification (omitted variable) bias; something that does not have a statistical fix
  - unit effects have a simple and intuitive explanation and can, as noted above, be useful to help you learn about your data
  - they are widely used and it does not take fancy math to explain and/or justify
- Disadvantages
  - they can be HIGHLY collinear with **x** variables that vary very little or are constant over time. (see Green's IO article)
  - inefficiency: fixed effects eat up lots of degrees of freedom which has consequences for all estimated standard errors

Random Effects Models

- We can rewrite the basic linear model and break down the error term into separate components resulting from our three sources of variation: time, unit or both

$$y_{it} = \mathbf{x}_{it}\beta + \varepsilon_{it}$$
$$\varepsilon_{it} = \alpha_i + \lambda_t + \nu_{it}$$

- The α captures the specific unit effects; the λ captures the time effects and the ν captures the unmeasured time and unit effects.
- Consider the unit effects; this is like the random error ν except that we have a single draw from the distribution that contributes to the error during each period (more on this later).

Assumptions of the RE Model

$$E(\alpha_i) = E(\lambda_t) = E(\nu_{it}) = 0$$
$$E(\alpha_i\lambda_t) = E(\lambda_t\nu_{it}) = E(\nu_{it}\alpha_i) = 0$$
$$E(\alpha_i\alpha_i) = \sigma_\alpha^2 \text{ if i=j, 0 otherwise}$$
$$E(\lambda_t\lambda_t) = \sigma_\lambda^2 \text{ if t=s, 0 otherwise}$$
$$E(\nu_{it}\nu_{it}) = \sigma_\nu^2 \text{ if i=j, t=s, 0 otherwise}$$
$$E(\alpha_i\mathbf{x}_{it}) = E(\lambda_t\mathbf{x}_{it}) = E(\nu_{it}\mathbf{x}_{it}) = 0$$

- If these conditions hold then the variance of y_{it} conditional on the **x**s is

$$\sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_\nu^2$$

This is also written as a variance components model as each element is a component of ε_{it}

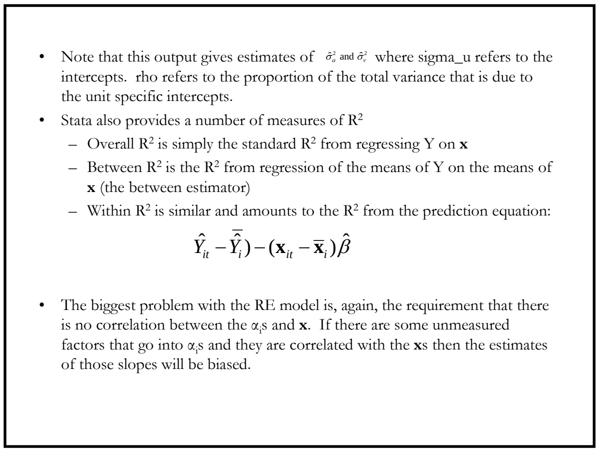
- The traditional way of thinking of random effects is to say that, instead of the α_is being fixed and us estimating them, that we treat them as a random draw from single distribution. We can then estimate the parameters of that distribution which (in almost every case) reduces the number of estimable parameters significantly.
- This is not necessarily assuming away the ballgame because the α_is were included because we were ignorant of the unit (or time) specific heterogeneity.

- Lets assume for now that λ_t=0; that there are no time-specific effects. (we will generalize later)
- We typically assume that α_i and ν_{it} are drawn from a normal distribution. We want to estimate $\hat{\sigma}_{\alpha}^{2}$ and $\hat{\sigma}_{\nu}^{2}$
- This means that we need to separate out the unit-specific error component from the unit-and-time specific part.
- How can we do this? OLS estimates will be unbiased and consistent in terms of the slopes but the standard errors will be significantly underestimated because we are acting as if we have information on N*T separate observations rather than on T observations on N units…this is analogous to serial correlation
- We need to account for the fact that the within unit errors are correlated. The simple way to proceed is via GLS (recall that we use GLS to deal with a similar problem when we have heteroscedastic errors.

- Of course, to use GLS we need to have an estimate of the variances to begin with which we do not. So, as in the heteroscedastic case, we use feasible GLS (FGLS).
- One key concern with FGLS is that we are assuming that the unit specific effects (the α_is) are uncorrelated with the exogenous variables; if this is not the case then our estimates will be biased.
- We do not need to make this assumption for the fixed effects model.

Estimating RE Models in STATA

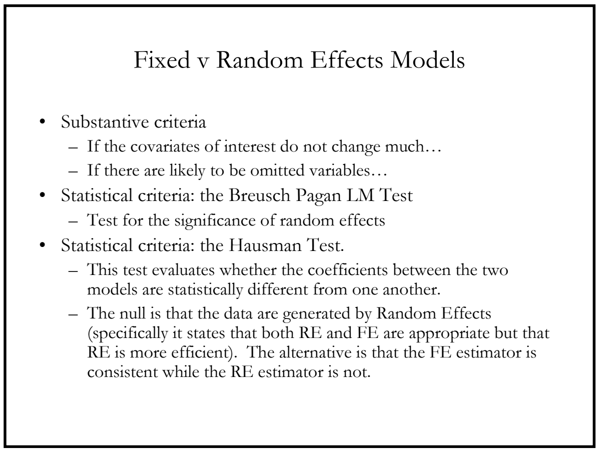
```
. xtreg  illiteracyrateTOTAL TOTALD GNPC

Random-effects GLS regression                   Number of obs      =       392
Group variable (i): code                        Number of groups   =       109

R-sq:  within  = 0.0286                         Obs per group: min =         1
       between = 0.0219                                        avg =       3.6
       overall = 0.0186                                        max =         4

Random effects u_i ~ Gaussian                   Wald chi2(2)       =      9.00
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0111

------------------------------------------------------------------------------
illiteracy~L |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TOTALDEBTgnp |  -.0085579   .0032842    -2.61   0.009    -.0149948   -.0021211
      GNPCAP |  -.0002998    .000183    -1.64   0.101    -.0006585    .0000588
       _cons |   32.02346    2.24976    14.23   0.000     27.61401    36.43291
-------------+----------------------------------------------------------------
     sigma_u |  22.313335
     sigma_e |  5.4462951
         rho |  .94377349   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```



- Note that this output gives estimates of $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\nu^2$ where sigma_u refers to the intercepts. rho refers to the proportion of the total variance that is due to the unit specific intercepts.
- Stata also provides a number of measures of R^2
  - Overall R^2 is simply the standard R^2 from regressing Y on **x**
  - Between R^2 is the R^2 from regression of the means of Y on the means of **x** (the between estimator)
  - Within R^2 is similar and amounts to the R^2 from the prediction equation:

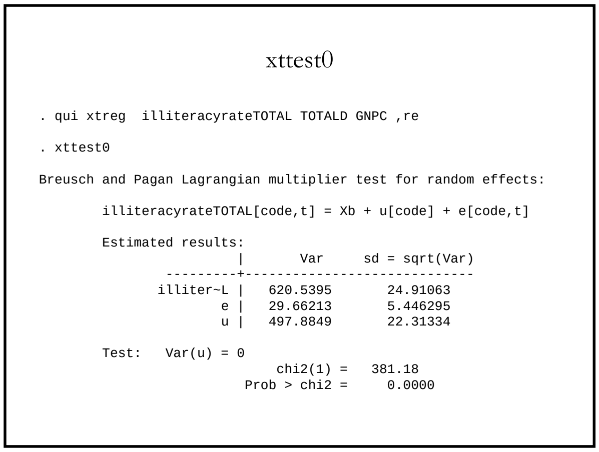
$$\hat{Y}_{it} - \bar{\hat{Y}}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\hat{\beta}$$

- The biggest problem with the RE model is, again, the requirement that there is no correlation between the α_is and **x**. If there are some unmeasured factors that go into α_is and they are correlated with the **x**s then the estimates of those slopes will be biased.



Fixed v Random Effects Models

- Substantive criteria
  - If the covariates of interest do not change much…
  - If there are likely to be omitted variables…
- Statistical criteria: the Breusch Pagan LM Test
  - Test for the significance of random effects
- Statistical criteria: the Hausman Test.
  - This test evaluates whether the coefficients between the two models are statistically different from one another.
  - The null is that the data are generated by Random Effects (specifically it states that both RE and FE are appropriate but that RE is more efficient). The alternative is that the FE estimator is consistent while the RE estimator is not.



xttest0

```
. qui xtreg  illiteracyrateTOTAL TOTALD GNPC ,re

. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects:

        illiteracyrateTOTAL[code,t] = Xb + u[code] + e[code,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
               illiter~L |   620.5395       24.91063
                       e |   29.66213       5.446295
                       u |   497.8849       22.31334

        Test:   Var(u) = 0
                              chi2(1) =   381.18
                          Prob > chi2 =     0.0000
```

xthaus

```
. qui xtreg  illiteracyrateTOTAL TOTALD GNPC ,re


Hausman specification test

                ---- Coefficients ----
             |      Fixed       Random
illiteracy~L |    Effects      Effects       Difference
-------------+-----------------------------------------
TOTALDEBTgnp |  -.0094478    -.0085579        -.0008899
      GNPCAP |  -.0001512    -.0002998         .0001486

    Test:  Ho:  difference in coefficients not systematic

              chi2(  2) = (b-B)'[S^(-1)](b-B), S = (S_fe -S_re)
                           =     0.00
                 Prob>chi2 =     1.0000
```

Last Thoughts

- We have not talked about heteroscedasciticy within units yet
- Most FE and RE models are concerned with unit heterogeneity; not time effects
- As N→∞ the FE and RE estimators will converge; assuming of course, no systematic omitted variable bias.