Week 10. Panel Models 1

Introduction

- So far, we discussed statistical models to analyze a cross-sectional dataset. This week we will discuss statistical models to analyze panel data.
- Panel data, also known as longitudinal data or cross-sectional time series data in some special cases, is data that is derived from observations over time on a number of cross-sectional units like individuals, households, firms, or governments.
- Panel methods are more close to causality implications than cross-sectional methods because panel models estimate how the changes over time in explanatory variables are associated with the (sometimes lagged) change over time in dependent variable. By doing so, we can rule out the possibility that unobserved heterogeneity across cross-sectional units is a driving force which makes an independent variable statistically significant even though the real association between independent and dependent variables is null.
- Panel data allows you to control for variables you cannot observe or measure like cultural factors or difference in business practices across companies; or variables that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc.). In case of individual, you can control for unobservable time-invariant individual characteristics (e.g., IQ or some attitude). This is, it accounts for individual heterogeneity.
- With panel data you can include variables at different levels of analysis suitable for multilevel or hierarchical modeling. For example, if you're analyze income growth of family members using growth curve models, your level 1: individual-time, level 2: individual, and level 3: family.
- The most commonly estimated panel models are probably fixed effects models (FEM) and random effects models (REM).
- REM will yield statistically significant results more likely than FEM (we will discuss why that is the case later). But the assumptions of REM is much stronger than FEM and unfortunately those assumptions are rarely met with real data. Therefore, FEM is almost always a better choice than REM.
- In the early 2000s, you may be able to publish a paper in a top sociological journal with REM, but now it is impossible. Halaby's (2004, ARS) paper is basically to say that "don't use REM."

Panel Data Structure

- Typical data structure

id	year	y	x1	x2	x3
1	2010	10.0	4	25	0
1	2011	10.2	4	26	0
1	2012	10.5	4	27	1
2	2010	9.1	2	30	1
2	2011	9.2	2	31	1
2	2012	9.0	2	32	0
2	2013	9.3	2	33	1
3	2010	10.5	5	28	0
3	2011	10.6	5	29	1
3	2012	10.7	5	30	1

- You need to reshape the data to "long" if your data format is "wide." Use the Stata command, `reshape`.
- Example 1: wide format

```
        famid   faminc96   faminc97   faminc98
 1.         3      75000      76000      77000
 2.         1      40000      40500      41000
 3.         2      45000      45400      45800
```

- Example 1: long format

```
. reshape long faminc, i(famid) j(year)

        famid       year     faminc
 1.         1         96      40000
 2.         1         97      40500
 3.         1         98      41000
 4.         2         96      45000
 5.         2         97      45400
 6.         2         98      45800
 7.         3         96      75000
 8.         3         97      76000
 9.         3         98      77000
```

- In the 1st example, the unit of observation for the long format is "family-year". The unit of observation for the wide format is family.
- Example 2: wide format

```
         famid      birth        ht1        ht2
 1.         1          1        2.8        3.4
 2.         1          2        2.9        3.8
 3.         1          3        2.2        2.9
 4.         2          1          2        3.2
 5.         2          2        1.8        2.8
 6.         2          3        1.9        2.4
 7.         3          1        2.2        3.3
 8.         3          2        2.3        3.4
 9.         3          3        2.1        2.9
```

- Example 2: long format

```
. reshape long ht, i(famid birth) j(age)

        famid      birth        age         ht
 1.         1          1          1        2.8
 2.         1          1          2        3.4
 3.         1          2          1        2.9
 4.         1          2          2        3.8
 5.         1          3          1        2.2
 6.         1          3          2        2.9
 7.         2          1          1          2
 8.         2          1          2        3.2
 9.         2          2          1        1.8
10.         2          2          2        2.8
11.         2          3          1        1.9
12.         2          3          2        2.4
13.         3          1          1        2.2
14.         3          1          2        3.3
15.         3          2          1        2.3
16.         3          2          2        3.4
17.         3          3          1        2.1
18.         3          3          2        2.9
```

- In the 2nd example, the unit of observation for the long format is "each birth within family (individual)-age". The unit of observation for the wide format is individual (within family).
- Using the `reshape` command, you can convert the long format into the wide format.

Fixed Effects Model (FEM): In Essence

- FEM is basically OLS with many dummy variables which identify each cross-sectional unit of observation.
- OLS

$$y_{it} = \alpha + \sum \beta_j x_{ijt} + \sum \gamma_k z_{ik} + e_{it} \tag{1}$$

  where x_{ijt} is a vector of time-variant variable j at time t for individual i and z_{ik} is a vector of time invariant variable k for individual i. Each individual i has t number of observations.
- Fixed Effects Model

$$y_{it} = \alpha_i + \sum \beta_j x_{ijt} + e_{it} \tag{2}$$

  where α_i is constant for individual i. OLS has only one intercept, but FEM has i number of intercepts ($\alpha + u_i$). Because all time-invariant characteristics for individual i is accounted for by α_i, the effects of time-invariant characteristics z_{ik} are dropped from the model.
- FEM is basically OLS with many dummy variables which identify each cross-sectional unit of observation. If you run OLS with individual specific fixed effects (i.e., add all dummies which identify individuals), you will get the identical results with FEM. This OLS method is usually called the dummy variable regression.
- Interpretation: As x_j changes 1 unit over time, y is expected to change by β_j net of other covariates.

An Example: Fixed Effects Model Result

```
. reg   lnwage age female i.educ

      Source |       SS       df       MS              Number of obs =     951
-------------+------------------------------           F(  5,   945) =   98.23
       Model |  99.9338627     5  19.9867725            Prob > F      =  0.0000
    Residual |  192.269618   945  .203459913            R-squared     =  0.3420
-------------+------------------------------           Adj R-squared =  0.3385
       Total |  292.203481   950  .307582611            Root MSE      =  .45107

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0027996   .0015268     1.83   0.067    -.0001966    .0057959
      female |  -.3702238   .0302035   -12.26   0.000    -.4294974   -.3109502
             |
        educ |
          2  |   .2517716   .0407223     6.18   0.000      .171855    .3316883
          3  |   .3042206   .0554722     5.48   0.000     .1953575    .4130836
          4  |   .6852116   .0444734    15.41   0.000     .5979335    .7724898
             |
       _cons |   4.457611   .0810778    54.98   0.000     4.298498    4.616725
------------------------------------------------------------------------------
```

```
. xtset pid year
. xtreg lnwage age female i.educ, fe
note: female omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       951
Group variable: pid                             Number of groups   =       229

R-sq:  within  = 0.1029                         Obs per group: min =         2
       between = 0.0006                                        avg =       4.2
       overall = 0.0095                                        max =         8

                                                F(4,718)           =     20.60
corr(u_i, Xb)  = -0.6218                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0430487   .0052149     8.25   0.000     .0328103    .0532871
      female |          0  (omitted)
             |
        educ |
          2  |   .1104229   .1140319     0.97   0.333    -.1134529    .3342987
          3  |   .3808312   .1816478     2.10   0.036      .024207    .7374555
          4  |   .3967386   .1690044     2.35   0.019     .0649366    .7285405
```

```
             |
       _cons |   2.855121    .229647    12.43   0.000     2.404261    3.305981
-------------+----------------------------------------------------------------
     sigma_u |  .65866485
     sigma_e |  .29513481
         rho |  .83279476   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(228, 718) =     8.07            Prob > F = 0.0000
```

- `R-sq: within`: The proportion of the variance within group over t accounted for by explanatory variables.
- `R-sq: between`: The proportion of the variance across group i accounted for by explanatory variables. For FEM, this is fixed.
- `R-sq: overall`: The proportion of the overall variance accounted for by explanatory variables.
- `corr(u_i, Xb`: individual fixed effects u_i are correlated with the regressors. In REM, this is assumed to be zero.
- `Number of obs`: the total number of observations (e.g., individual-time)
- `Number of groups`: the total number of groups (e.g., individual)
- `Obs per group: min`: the minimum number of t.
- `Obs per group: avg`: the average number of t.
- `Obs per group: max`: the maximum number of t.
- `sigma_u`: standard deviation of u_i. That is, the standard deviation of the difference across group i.
- `sigma_e`: standard deviation of e_i. That is, the standard deviation of the residuals across it after accounting for the difference across group i.
- `rho`: $\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$, the proportion of the total variance that can be accounted for by the difference across group is.

Fixed Effects Model (FEM): More Detailed Explanation

- FEM. Fixed effect estimates are a way to remove group i specific fixed effects from your estimates. It can be done by doing fixed effects transformation as follows:

  1.
  $$y_{it} = \alpha_i + \sum \beta_j x_{ijt} + e_{it} \tag{3}$$

2. Now, for each i, average this equation over time. We get

$$\bar{y}_i = \alpha_i + \sum \beta_j \bar{x}_{ij} + \bar{e}_i \tag{4}$$

3. If we subtract equation 4 from equation 3 for each t, we wind up with:

$$\begin{aligned} y_{it} - \bar{y}_i &= (\alpha_i - \alpha_i) + \sum \beta_j (x_{ijt} - \bar{x}_{ij}) + (e_{it} - \bar{e}_i) \\ \ddot{y}_{it} &= \sum \beta_j \ddot{x}_{ijt} + \ddot{e}_{it} \end{aligned} \tag{5}$$

Now we estimate with "demeaned-data".

- This fixed effects transformation is also called within transformation. Equation 5 is the same as OLS. This OLS is based on the time-demeaned variables is called the fixed effects estimator or the *within* estimator. That is, it is OLS which uses the time variation in y and x within each cross-sectional observation.
- The *between* estimator is obtained as the OLS estimator on the cross-sectional equation 4.
- With FEM, we are not interested in the *between* estimator because it is biased when α_i is correlated with $\bar{x}_i$. If not, we can use REM.
- The fixed effects estimator allows for arbitrary correlation between α_i and the explanatory variables in any time period. Because of this, any explanatory variable that is constant over time for all i gets swept away by the fixed effects transformation. Put differently, all time-invariant individual heterogeneity are taken into account in FEM.
- Beyond this, FEM has the same assumptions with OLS. The idiosyncratic error e_it should be uncorrelated with each explanatory variable across all time periods. The errors are homoskedastic and serially uncorrelated.
- Regarding df, in OLS we have total $n = i*t$, thus $df = i*t-k$, but in FEM, $df = i*t-i-k$.
- R^2 of equation 4 is between-R^2. R^2 of equation 5 is within-R^2.
- Standard deviation of $\bar{e}_i$ of equation 4 is `sigma_u`. Standard deviation of $\ddot{e}_{it}$ of equation 5 is `sigma_e`.
- The overall intercept of FEM is actually the average of the individual-specific intercepts, which is an unbiased, consistent estimator of $\alpha = E(\alpha_i)$.
- Although time-constant variables cannot be included by themselves in a fixed effects model, they can be interacted with variables that change over time and, in particular, with year dummy variables.
- If you transform all time-varying independent variable by doing group-specific mean-centering, you can estimate OLS with both time-invariant covariates and time-varying covariates. The coefficients estimated for group-specific mean-centered time-varying covariates will be identical with those of FEM.

Fixed Effects Model (FEM) with Time-invariant Covariates

- If you transform all time-varying independent variable by doing group-specific mean-centering, you can estimate OLS with both time-invariant covariates and time-varying covariates. The coefficients estimated for group-specific mean-centered time-varying covariates will be identical with those of FEM.
- $$y_{it} = \alpha + \sum \beta_j (x_{ijt} - \bar{x}_{ij}) + \sum \gamma_k z_{ik} + e_{it} \quad (6)$$

```
. egen mage    = mean(age), by(pid)

. tab educ, gen(educ)
. egen meduc1  = mean(educ1), by(pid)
. egen meduc2  = mean(educ2), by(pid)
. egen meduc3  = mean(educ3), by(pid)
. egen meduc4  = mean(educ4), by(pid)

. gen cage = age-mage
. gen ceduc1 = educ1 - meduc1
. gen ceduc2 = educ2 - meduc2
. gen ceduc3 = educ3 - meduc3
. gen ceduc4 = educ4 - meduc4

. reg lnwage cage female ceduc2-ceduc4

      Source |       SS       df       MS              Number of obs =     951
-------------+------------------------------           F(  5,   945) =   40.38
       Model |  51.4446079     5  10.2889216           Prob > F      =  0.0000
    Residual |  240.758873   945  .254771294           R-squared     =  0.1761
-------------+------------------------------           Adj R-squared =  0.1717
       Total |  292.203481   950  .307582611           Root MSE      =  .50475

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cage |   .0430487   .0089188     4.83   0.000     .0255458    .0605516
      female |  -.4323907   .0328027   -13.18   0.000    -.4967654   -.3680161
      ceduc2 |   .1104229   .1950209     0.57   0.571    -.2723013    .4931471
      ceduc3 |   .3808312   .3106597     1.23   0.221    -.2288314    .9904938
      ceduc4 |   .3967385   .2890366     1.37   0.170    -.1704894    .9639665
       _cons |   4.899105   .0224388   218.33   0.000     4.855069     4.94314
------------------------------------------------------------------------------
```

- In the above result, the constant is the expected y for men when all x's are set at their group specific means. The effect of being female is the (dis)advantage of being female when all x's are set at their group specific means.

First Difference Model (FDM)

- If there are two observations for 2 time points 1 and 2 for the same individual (group) i, then we can estimate the following model:

$$\begin{aligned}(y_{i2} - y_{i1}) &= \gamma_0 + \sum \beta_k (x_{ik2} - x_{ik1}) + (e_{i2} - e_{i1}) \\ \Delta y_i &= \gamma_0 + \sum \beta_k \Delta x_{ik} + \Delta e_i\end{aligned} \tag{7}$$

  where γ_0 is a fixed effect of time 2.

- Equation 7 is identical with the following equation:

$$y_{it} = \beta_0 + \gamma_0 + \sum \beta_k x_{ikt} + (\alpha_i + u_{it}) \tag{8}$$

- That is, FDM is OLS when all dependent and independent variables are transformed into the differences between time 1 and time 2.

- FDM is identical with FEM when the number of time is two. When $T \geq 3$, FDM is not equal to FEM.

- FDM can be used for data with $T \geq 3$.

$$\Delta y_{it} = \sum \gamma_t T_t + \sum \beta_k \Delta x_{ikt} + \Delta e_{it} \tag{9}$$

  where T_t is a set of dummy variables of time t.

- The key assumption is that the idiosyncratic errors are uncorrelated with the explanatory variable in each time period: $Cov(x_{ijt}, e_{is}) = 0$ for all, t and j. This assumption is not necessarily met with real data. If not, there can be serious biases. Thus, when $T \geq 3$, FEM is better than FDM.

Random Effects Model (REM)

- REM is the same as FEM but we assume that the unobserved individual (group) specific effect α_i is uncorrelated with each explanatory variable: $Cov(x_{ijt}, \alpha_i) = 0$. That is, there is one more assumption on top of all assumptions of FEM.

- In fact, if this assumption is met, we simply can estimate OLS. No panel technique is necessary.

- However, because α_i is in the composite error (that is, $\alpha_i + e_{it}$ constitutes the total error v_{it}, $v_{it} = \alpha_i + e_{it}$) in each time period, total errors are serially correlated across time.

- Thus, under the random effects assumptions, $Corr(v_{it}, v_{is}) = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$. This (necessarily) positive serial correlation in the error term can be substantial, and, because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will the usual test statistics.

- To fix this problem, we need to transform all variables by $\theta = 1 - \sqrt{(\sigma_e^2/(\sigma_e^2 + T\sigma_a^2)}$ which is between 0 and 1.

- New equation becomes

$$y_{it} - \theta \bar{y}_i = \beta_0(1 - \theta) + \sum \beta_k (x_{ikt} - \theta \bar{x}_{ik}) + (v_{it} - \theta \bar{v}_i) \tag{10}$$

- This is using quasi-demeaned data. Recall that FEM is using demeaned data.
- The random effects transformation subtracts a fraction of that time average, where the fraction depends on σ_e^2, σ_a^2, and the number of time periods, T.
- The transformation in equation 10 allows for explanatory variables that are constant over time, and this is one advantage of random effects (RE) over either fixed effects or first differencing. This is possible because RE assumes that the unobserved effect is uncorrelated with all explanatory variables, whether the explanatory variables are fixed over time or not. Thus, in our model, we can include a variable such as education even if it does not change over time. But we are assuming that education is uncorrelated with a_i, which contains ability and family background (of course this is unrealistic). In many applications, the whole reason for using panel data is to allow the unobserved effect to be correlated with the explanatory variables.
- Example: REM

```
. xtreg lnwage age female i.educ, re

Random-effects GLS regression                   Number of obs      =       951
Group variable: pid                             Number of groups   =       229

R-sq:  within  = 0.0364                         Obs per group: min =         2
       between = 0.3724                                        avg =       4.2
       overall = 0.3306                                        max =         8

                                                Wald chi2(5)       =    160.90
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0074483   .0024211     3.08   0.002     .0027031    .0121935
      female |  -.3697987   .0522792    -7.07   0.000     -.472264   -.2673333
             |
        educ |
          2  |     .23775   .0628876     3.78   0.000     .1144926    .3610074
          3  |   .3923649   .0885583     4.43   0.000     .2187938    .5659361
          4  |   .6595233   .0735716     8.96   0.000     .5153256     .803721
             |
       _cons |   4.252047   .1289119    32.98   0.000     3.999385     4.50471
-------------+----------------------------------------------------------------
     sigma_u |  .34085686
     sigma_e |  .29513481
         rho |  .57152131   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

OLS vs. REM vs. FEM

- You can do Hausman test to determine between FEM and REM.

```
. xtreg lnwage age female i.educ, fe
. estimates store fe
. xtreg lnwage age female i.educ, re
. estimates store re
. hausman fe re


             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0430487     .0074483        .0356004        .0046189
        educ |
          2  |    .1104229       .23775       -.1273271        .0951232
          3  |    .3808312     .3923649       -.0115337         .158598
          4  |    .3967386     .6595233       -.2627847        .1521503
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(4) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       69.20
                Prob>chi2 =      0.0000
```

When the null hypothesis is fail to reject, we can use REM.

- In reality, that happens very rarely.
- By the way, the coefficient estimates with REM is somewhere middle between OLS and FEM.