PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #17: Fixed Effects and Random Effects Models

I. Fixed Effects Transformation

- In using panel data, first differencing is just one way to eliminate the fixed effects, a_i. An alternative method is the *fixed effects* transformation.

- Consider a univariate model:

  - $y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}, \; t = 1, 2, \text{K}, T.$

- Now, for each *i*, we can average this equation over time:

  - $\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i, \; where \; \bar{y}_i = T^{-1} \sum_{i=1}^{T} y_{it}, \; etc.$

- If we subtract the second equation from the first equation, the fixed effects (and the intercept term) drop out:

  - $y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i), \; t = 1, 2, \text{K}, T.$

- In other words, the data have been *time-demeaned*. This *fixed effects transformation* is also called the *within transformation*. The unobserved "a" effect disappears. Conducting OLS on the time-demeaned data is called the *fixed effects estimator* or *within estimation*.

- Adding more independent variables to the equation is a straightforward extension in which each independent variable is also time-demeaned. Notice that any independent variable that are constant over time (e.g., race, gender) will get eliminated by the fixed effects transformation. If we have

year intercepts, variables that change at the same time also get absorbed in the year intercepts. We can interact time-invariant variables with variables that do change over time. Also, the transformation could reduce the variation in the independent variables resulting in high standard errors.

- The fixed effects estimator eliminates the problem of bias stemming from correlation between unobserved time-invariant effects and the independent variables. However, fixed effects estimates are unbiased only if the idiosyncratic error term (u_{it}) is uncorrelated with each independent variable across *all* time periods. We also typically assume that the idiosyncratic error term is homoskedastic and serially uncorrelated across time periods.

- The between estimator is obtained as the OLS estimator on the cross-sectional data. Use the time averages for both Y and X to run a cross-sectional regression. We do not study the between estimator in more detail because it is biased when a_i is correlated with X_i. The between group estimator ignores important information on how variables change over time.

- Even though we have N*T observations and k independent variables, the degrees of freedom after the fixed effects transformation are not (N*T)-k. For each cross-sectional observation *i*, we can figure out the value of the nth observation based on the other n-1 observations, so we lose one degree of freedom for each *i*. Therefore, d.f. = (N*T)-k-N = N(T-1)-k.

II. Dummy Variable Estimation of Fixed Effects

- Rather than running OLS on the time-demeaned data, it is easier to compute the fixed effects estimator by giving each cross-sectional unit its own intercept term. This gives us exactly the same coefficient estimates and standard errors as the within estimation discussed above. The dummy variable approach also automatically provides the proper degrees of freedom because it includes another variable in the regression for each cross-sectional unit.

- Including cross-sectional dummy variables (as well as the time dummy variables) will result in a high R^2. This is to be expected. We can use this R^2 estimate to conduct F-tests, such as testing the joint significance of the cross-sectional dummy variables. It is likely that the null hypothesis will fail.

- The coefficient estimates for the cross-sectional dummy variables might also be of interest if we want to see whether a certain cross-sectional unit is above or below average holding all else equal.

- It is a bit tedious to use the dummy variable approach, especially when N is large. It you don't care about the coefficient estimates for the cross-sectional dummy variables, then Stata provides a shortcut.

Example: Effect of Job Training on Firm Scrap Rates

3 years of data (1987, 1988, 1989) on 54 firms. No grants were received prior to 1988. In 1988, 19 firms received grants and in 1989 10 firms received grants.

Dependent Variable: log (scrap rate)

Year 88	-.080 (.109)
Year 89	-.247 (.133)
GRANT	-.252 (.151)
$GRANT_{-1}$	-.422 (.210)

N = 162; df = 104; R^2

1. The lagged effect of the training grant is larger than the contemporaneous effect: job training has an effect one year later. Obtaining a grant in 1988 is predicted to lower the scrap rate in 1989 by 34.4% (exp (-.422) – 1) = -.344

2. Coefficient on Year 89 indicates that the scrap rate was substantially lower in 1989 than in the base year, even in the absence of the job training grants.

III. Fixed Effects with Unbalanced Panels

Panel data sets with missing years for some cross-sectional units results in an unbalanced panel. However, cross-sectional units with only one observation drop out.

One needs to be careful if the time data are missing for systematic reasons. If the reasons for missing data for some observations is uncorrelated with the idiosyncratic errors, then the unbalanced panel causes no problem.

The difficulty arises if we have data on individuals, families or firms. For example, suppose we obtain a random sample of firms in 1990 and want to test if unionization affects firm profitability. Use panel data to control for worker and management characteristics that might affect the fraction of the firm's work force that is unionized.

Collect data in subsequent years and some firms are lost because they went out of business. The question is whether FE applied to the unbalanced panel will yield unbiased estimates.
If the reason that a firm leaves the sample (attrition) is correlated with the error term (those unobserved factors that change over time and affect profits) then sample selection can cause biased estimators.

IV. Fixed Effects or First Differencing?

- The two methods are identical if T=2 for all cross sections, but they yield different results when T>2.

- Both methods result in unbiased and consistent estimates.

- When the idiosyncratic error term is serially uncorrelated, then fixed effects is more efficient than first differencing. If the idiosyncratic error term is serially correlated and follows a random walk, then first differencing is better. For other forms of serially correlation, it is unclear which is better.

- First differencing is typically better when T is large and N is small. But fixed effects is less sensitive to violations of the strict exogeneity assumption.

IV. Random Effects Models

- If we think that the fixed effects captured in a_i are uncorrelated with each independent variable (big assumption!), then eliminating them through first-differencing or the fixed effects model results in inefficient (though unbiased) estimates.

- If a_i is uncorrelated with the independent variables, then we can just do a pooled OLS regression. However, this ignores the serial correlation of our composite error term, v_{it} (= a_i + u_{it}). Put simply, the unobservable characteristics that influence cross section *i* over time are not independent draws. Ignoring this serial correlation will lead to incorrect standard errors.

- Recall from Quant. II that one can address serial correlation by Generalized Least Squares (GLS). That is, we can transform the data to eliminate the serial correlation. For the random effects model, the GLS transformation is as follows:

- $y_{it} - \lambda \bar{y}_i = \beta_0(1-\lambda) + \beta_1(x_{it1} - \lambda\bar{x}_{i1}) + \Lambda\ \beta_k(x_{itk} - \lambda\bar{x}_{ik}) + (v_{it} - \lambda\bar{v}_i),$

  where $\lambda = 1 - [\sigma_u^2 / (\sigma_u^2 + T\sigma_a^2)]^{1/2}$.

- This transformation is similar to our time-demeaned transformation, except that the time mean is multiplied by a fraction that depends on the variance of the idiosyncratic error term and the variance of the time-invariant error term. Note that we cannot know these variance terms. However, we can run a pooled OLS regression and use the resulting residuals to estimate $\hat{\lambda}$. Running OLS on the above equation (using our predicted estimate for the transformation term) eliminates the serial correlation and results in consistent estimates and accurate standard errors.

- Another advantage of the random effects model over the fixed effects model is that we can obtain coefficient estimates for time-invariant independent variables (though we are assuming that these characteristics are uncorrelated with the unobservable time-invariant variables).

- Note also that the closer $\hat{\lambda}$ is to 0, the closer the estimates are to the pooled OLS regression results. In this case, the unobserved effect is relatively unimportant. It is more common for σ^2_a to be large relative to σ^2_u, so $\hat{\lambda}$ is closer to 1. In this case, the estimates are closer to the fixed effects regression results.

V. Random Effects versus Fixed Effects

- If we think that a_i is uncorrelated with the independent variables, then the random effects model is the appropriate

estimation strategy. Random effects uses up fewer degrees of freedom and has broad conceptual appeal as a broad characterization of the sources of error in a large dataset with substantial time-series and cross-sectional variation.

- If we think that a_i is correlated with the independent variables, then the fixed effects model (or first differencing) is the appropriate estimation strategy. The fixed-effects approach allows the researcher to analyze the extent to which the dependent variable for each cross-section unit differs from the overall cross-section mean. Fixed effects do not require the assumption that the individual effects that are incorporated into the error term are uncorrelated with the explanatory variables in the model. This assumption may not be valid and may cause parameter estimates to be inconsistent.

- Comparing the two results can serve as a test of whether there is correlation between a_i and the independent variables, assuming that the idiosyncratic errors and independent variables are uncorrelated across all time periods.

- Example 1: estimate a wage equation for men using fixed effects, random effects and pooled OLS. With fixed effects time invariant characteristics drop out of the model. Show fixed effects using two alternative approaches.

```
> There are 545 men in the sample, each sampled from 1980
> through 1987.
> */
>
>
> /*
> Stata has two commands for conducting fixed effects regressions.
> The first is areg, which can be used when there are many
> dummy variables that make up category.
>
> This command is similar to regression, except
> we must tell Stata which variable
> contains the cross-sectional identifiers (dummies).
> */
>
> areg lwage expersq married union d81 d82 d83 d84 d85 d86 d87, a(nr);

                                                      Number of obs =    4360
                                                      F( 10,  3805) =   83.85
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.6209
                                                      Adj R-squared =  0.5657
                                                      Root MSE      =  .35099

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     expersq |  -.0051855   .0007044    -7.36   0.000    -.0065666   -.0038044
     married |   .0466804   .0183104     2.55   0.011     .0107811    .0825796
       union |   .0800019   .0193103     4.14   0.000     .0421423    .1178614
         d81 |   .1511912   .0219489     6.89   0.000     .1081584     .194224
         d82 |   .2529709   .0244185    10.36   0.000     .2050963    .3008454
         d83 |   .3544437   .0292419    12.12   0.000     .2971125    .4117749
         d84 |   .4901148   .0362266    13.53   0.000     .4190894    .5611402
         d85 |   .6174823   .0452435    13.65   0.000     .5287784    .7061861
         d86 |   .7654966   .0561277    13.64   0.000     .6554532    .8755399
         d87 |   .9250249   .0687731    13.45   0.000     .7901893    1.059861
       _cons |   1.426019   .0183415    77.75   0.000     1.390058    1.461979
-------------+----------------------------------------------------------------
          nr |      F(544, 3805) =      9.157   0.000         (545 categories)

. /*
> We can show that this is the same as typing out
> a dummy variable for each cross section.
> */
>
> /*First, we create the 545 dummy variables.*/
> tab nr, gen(r) nofreq;

. /*
> Now we can run the regression in which we include
> each dummy variable.

> *reg lwage d81 d82 d83 d84 d85 d86 d87 r2-r545 expersq married union;
. /*
> The other Stata command for fixed effects is xtreg.
> This is actually the command used for fixed or random effects models
> of panel data.
>
> All we need to do is tells stata whether we want fixed or random
> effects, and to tell stata which variable contains
> the cross-sectional identifiers (dummies)
> */
```

```
> /*Fixed effects model*/
> xtreg lwage expersq married union d81 d82 d83 d84 d85 d86 d87, fe i(nr);

Fixed-effects (within) regression               Number of obs      =      4360
Group variable (i): nr                          Number of groups   =       545

R-sq:  within  = 0.1806                         Obs per group: min =         8
       between = 0.0286                                        avg =       8.0
       overall = 0.0888                                        max =         8

                                                F(10,3805)         =     83.85
corr(u_i, Xb)  = -0.1222                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     expersq |  -.0051855   .0007044    -7.36   0.000    -.0065666   -.0038044
     married |   .0466804   .0183104     2.55   0.011     .0107811    .0825796
       union |   .0800019   .0193103     4.14   0.000     .0421423    .1178614
         d81 |   .1511912   .0219489     6.89   0.000     .1081584     .194224
         d82 |   .2529709   .0244185    10.36   0.000     .2050963    .3008454
         d83 |   .3544437   .0292419    12.12   0.000     .2971125    .4117749
         d84 |   .4901148   .0362266    13.53   0.000     .4190894    .5611402
         d85 |   .6174823   .0452435    13.65   0.000     .5287784    .7061861
         d86 |   .7654966   .0561277    13.64   0.000     .6554532    .8755399
         d87 |   .9250249   .0687731    13.45   0.000     .7901893    1.059861
       _cons |   1.426019   .0183415    77.75   0.000     1.390058    1.461979
-------------+----------------------------------------------------------------
     sigma_u |  .39176195
     sigma_e |  .35099001
         rho |  .55472817   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3805) =     9.16           Prob > F = 0.0000


. /*Random effects model*/
> xtreg lwage educ black hisp exper expersq married union d81 d82 d83 d84 d85 d
> 86 d87, re i(nr);

Random-effects GLS regression                   Number of obs      =      4360
Group variable (i): nr                          Number of groups   =       545

R-sq:  within  = 0.1799                         Obs per group: min =         8
       between = 0.1860                                        avg =       8.0
       overall = 0.1830                                        max =         8

Random effects u_i ~ Gaussian                   Wald chi2(14)      =    957.77
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0918763   .0106597     8.62   0.000     .0709836    .1127689
       black |  -.1393767   .0477228    -2.92   0.003    -.2329117   -.0458417
        hisp |   .0217317   .0426063     0.51   0.610    -.0617751    .1052385
       exper |   .1057545   .0153668     6.88   0.000     .0756361    .1358729
     expersq |  -.0047239   .0006895    -6.85   0.000    -.0060753   -.0033726
     married |    .063986   .0167742     3.81   0.000     .0311091    .0968629
       union |   .1061344   .0178539     5.94   0.000     .0711415    .1411273
         d81 |    .040462   .0246946     1.64   0.101    -.0079385    .0888626
         d82 |   .0309212   .0323416     0.96   0.339    -.0324672    .0943096
         d83 |   .0202806    .041582     0.49   0.626    -.0612186    .1017798
         d84 |   .0431187   .0513163     0.84   0.401    -.0574595    .1436969
         d85 |   .0578155   .0612323     0.94   0.345    -.0621977    .1778286
         d86 |   .0919476   .0712293     1.29   0.197    -.0476592    .2315544
         d87 |   .1349289   .0813135     1.66   0.097    -.0244427    .2943005
       _cons |   .0235864   .1506683     0.16   0.876     -.271718    .3188907
-------------+----------------------------------------------------------------
     sigma_u |  .32460315
     sigma_e |  .35099001
         rho |  .46100216   (fraction of variance due to u_i)
```

```
------------------------------------------------------------------------------

. /*OLS model*/
> reg lwage educ black hisp exper expersq married union d81 d82 d83 d84 d85 d86
>  d87;

      Source |       SS       df       MS              Number of obs =    4360
-------------+------------------------------           F( 14,  4345) =   72.46
       Model |  234.048277    14  16.7177341           Prob > F      =  0.0000
    Residual |  1002.48136  4345  .230720682           R-squared     =  0.1893
-------------+------------------------------           Adj R-squared =  0.1867
       Total |  1236.52964  4359  .283672779           Root MSE      =  .48033

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0913498   .0052374    17.44   0.000     .0810819    .1016177
       black |  -.1392342   .0235796    -5.90   0.000    -.1854622   -.0930062
        hisp |   .0160195   .0207971     0.77   0.441    -.0247535    .0567925
       exper |   .0672345   .0136948     4.91   0.000     .0403856    .0940834
     expersq |  -.0024117     .00082    -2.94   0.003    -.0040192   -.0008042
     married |   .1082529   .0156894     6.90   0.000     .0774937    .1390122
       union |   .1824613   .0171568    10.63   0.000     .1488253    .2160973
         d81 |     .05832   .0303536     1.92   0.055    -.0011886    .1178286
         d82 |   .0627744   .0332141     1.89   0.059    -.0023421    .1278909
         d83 |   .0620117   .0366601     1.69   0.091    -.0098608    .1338843
         d84 |   .0904672   .0400907     2.26   0.024      .011869    .1690654
         d85 |   .1092463   .0433525     2.52   0.012     .0242533    .1942393
         d86 |   .1419596    .046423     3.06   0.002     .0509469    .2329723
         d87 |   .1738334    .049433     3.52   0.000     .0769194    .2707474
       _cons |   .0920558   .0782701     1.18   0.240    -.0613935    .2455051
------------------------------------------------------------------------------
```

Points to Note:

1. The coefficients on educ, hisp and black are similar for the pooled OLS and random effects. The pooled OLS standard errors underestimate the true standard errors because they ignore the positive serial correlation.

2. Comparing RE and pooled OLS shows that the experience profile is somewhat different and both the marriage and union premiums fall notably in RE.

3. With Fixed effects, the marriage premium is 4.7% compared to 6.4% under RE and 10.8% under pooled OLS. The drop in the marriage premium variable is consistent with the idea that men who are more able—as captured by a higher unobserved heterogeneity effect—are more likely to be married.

4. In the RE estimation, the estimate of λ is .643. This explains in part why on the time-varying variables, the RE estimates lie closer to the FE estimates than to pooled OLS estimates.

$$\lambda = 1 - [\sigma_u^2 / (\sigma_u^2 + T\sigma_a^2)]^{1/2}$$

$$Y_{it} = \alpha + \beta X_{it} + u_i + e_{it}$$

In Stata $u_i = a_i$ and $e_{it} = u_{it}$

$\sigma_u = .324$ and $\sigma_e = .35$

$$\lambda = 1 - [\sigma_e^2 / (\sigma_e^2 + T\sigma_u^2)]^{½}$$

$$= 1 - [(.35)^2 / ((.35)^2 + 8(.324)^2)]^{½}$$

$$= 1 - [.1225 / (.1225 + 8(.105)]^{½}$$

$$= 1 - [.1225 / .9625]^{½}$$

$$= 1 - (.1273)^{½}$$

$$= 1 - (.357) = .643$$

Example 2: Patent Applications and Spending on Research and Development

The relationship between the log of the number of patent applications (*P)* and the log of the number of R & D expenditures (*RND*) was evaluated using panel data for 45 firms over a 7 year period. The R & D data are lagged 5 years to reflect the long interval that passes before research is translated into actual patent applications.

Basic regression model is:

$$P_{it} = \beta_0 + \beta_1 \text{RND}_{i,t-5} + e_{it}$$

where i refers to firms and t refers to time. The OLS estimates based on the pooled time series data (315 observations).

$$P_{it} = 1.438 + .845 \text{RND}_{i,t-5}$$
$$t = \quad (14.01) \quad (24.17)$$

$R^2 = .65$

OLS results show that on average patent applications increased by .845% for every 1% increase in R & D expenditures.

Fixed Effects—includes time period and firm specific dummies

$$P_{it} = .195 \text{RND}_{i,t-5} \quad R^2 = .94$$
$$t = \quad (2.38)$$

FE suggests that patent applications increase by .195% for every 1% increase in R & D expenditures. The researchers conducted an F test of the null hypothesis that all the coefficients on the firm specific dummies are jointly equal to zero. They rejected this null hypothesis that the firm specific dummies are jointly equal to zero.

Random Effects—includes a composite error term. Unobserved heterogeneity is part of the composite error term.

$$P_{it} = 2.299 + .519 \text{RND}_{i,t-5}$$
$$t = \quad (12.13) \quad (8.78)$$

$R^2 = .92$

Random effects estimate of .519 is much larger than FE estimate of .195. RE suggests that patent applications increase by .513% for every 1% increase in R & D expenditures.

Which is better—RE or FE? The assumption that the errors are uncorrelated with the explanatory variables can be tested using a Hausman specification test. With this test, one compares the parameters of the fixed-effects model to the parameters obtained using the random-effects model. Null hypothesis of a random effects model is compared to the alternative hypothesis of fixed-effects. Results yield a Chi-square of 32.62 with 1 degree of freedom, significant at 5% level. Conclusion: fixed effects model most accurately characterizes the relationship between R & D and patent applications.

Code in STATA for testing whether fixed effects or random effects is the more appropriate model.

(1) First you need to run a panel regression with FE, following by the command "estimates store fixed."

```
xtreg DepVar IndVar1 IndVar2 ...., fe
estimates store fixed
```

(2) Then you need to run the same panel regression with RE, followed by "estimates store random."

(3) Finally, you run the Hausman test, with the following code: "hausman fixed random"

All together, the code looks like this:

```
xtreg DepVar IndVar1 IndVar2 ...., fe
estimates store fixed
xtreg DepVar IndVar1 IndVar2 ...., re
estimates store random

hausman fixed random
```