Panel Data 4: Fixed Effects vs Random Effects Models

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised March 20, 2018

These notes borrow very heavily, sometimes verbatim, from Paul Allison's book, *Fixed Effects Regression Models for Categorical Data*. The Stata XT manual is also a good reference. This handout tends to make lots of assertions; Allison's book does a much better job of explaining why those assertions are true and what the technical details behind the models are.

Overview. With panel/cross sectional time series data, the most commonly estimated models are probably fixed effects and random effects models. Population-Averaged Models and Mixed Effects models are also sometime used. In this handout we will focus on the major differences between fixed effects and random effects models.

Several considerations will affect the choice between a fixed effects and a random effects model.

1. *What is the nature of the variables that have been omitted from the model?*
   a. If you think there are no omitted variables – or if you believe that the omitted variables are uncorrelated with the explanatory variables that are in the model – then a random effects model is probably best. It will produce unbiased estimates of the coefficients, use all the data available, and produce the smallest standard errors. More likely, however, is that omitted variables will produce at least some bias in the estimates.
   b. If there are omitted variables, and these variables are correlated with the variables in the model, then fixed effects models may provide a means for controlling for omitted variable bias. In a fixed-effects model, subjects serve as their own controls. The idea/hope is that whatever effects the omitted variables have on the subject at one time, they will also have the same effect at a later time; hence their effects will be constant, or "fixed." HOWEVER, in order for this to be true, the omitted variables must have time-invariant values with time-invariant effects.
      i. By time-invariant values, we mean that the value of the variable does not change across time. Gender and race are obvious examples, but this can also include things like the Educational Level of the Respondent's Father.
      ii. By time-invariant effects, we mean the variable has the same effect across time, e.g. the effect of gender on the outcome at time 1 is the same as the effect of gender at time 5.
      iii. If either of these assumptions is violated, we need to have explicit measurements of the variables in question and include them in our models. In the case of time-varying effects, we can include things like the interaction of gender with time. We also need explicit measurements of time-invariant variables if they are thought to interact with other variables in the model, e.g. we think the effect of SES differs by race.
2. *How much variability is there within subjects?*
   a. If subjects change little, or not at all, across time, a fixed effects model may not work very well or even at all. There needs to be within-subject variability in the variables if we are to use subjects as their own controls. If there is little variability

within subjects then the standard errors from fixed effects models may be too large to tolerate.

   b. Conversely, random effects models will often have smaller standard errors. But, the trade-off is that their coefficients are more likely to be biased.

3. *Do we wish to estimate the effects of variables whose values do not change across time, or do we merely wish to control for them?*
   a. With fixed effects models, we do not estimate the effects of variables whose values do not change across time. Rather, we control for them or "partial them out." This is similar to an experiment with random assignment. We may not measure variables like SES, but whatever effects those variable have are (subject to sampling variability) assumed to be more or less the same across groups because of random assignment.
   b. Random effects models will estimate the effects of time-invariant variables, but the estimates may be biased because we are not controlling for omitted variables.

Fixed effects models. Allison says "In a fixed effects model, the unobserved variables are allowed to have any associations whatsoever with the observed variables." Fixed effects models control for, or partial out, the effects of time-invariant variables with time-invariant effects. This is true whether the variable is explicitly measured or not. Exactly how they do so varies by the statistical technique being used. The optional appendix discusses these methods further. Unfortunately, the effects of time-invariant variables that are measured cannot be estimated.

```
. use https://www3.nd.edu/~rwilliam/statafiles/teenpovxt, clear
. *fixed effects
. xtlogit pov i.mother i.spouse i.school hours i.year i.black age, fe nolog
note: multiple positive outcomes within groups encountered.
note: 324 groups (1,620 obs) dropped because of all positive or
      all negative outcomes.
note: 1.black omitted because of no within-group variance.
note: age omitted because of no within-group variance.

Conditional fixed-effects logistic regression   Number of obs     =      4,135
Group variable: id                              Number of groups  =        827

                                                Obs per group:
                                                              min =          5
                                                              avg =        5.0
                                                              max =          5

                                                LR chi2(8)        =      97.28
Log likelihood  = -1520.1139                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.mother |   .5824322   .1595831     3.65   0.000      .269655    .8952094
    1.spouse |  -.7477585   .1753466    -4.26   0.000    -1.091431   -.4040854
    1.school |   .2718653   .1127331     2.41   0.016     .0509125    .4928181
       hours |  -.0196461   .0031504    -6.24   0.000    -.0258208   -.0134714
             |
        year |
          2  |   .3317803   .1015628     3.27   0.001      .132721    .5308397
          3  |   .3349777   .1082496     3.09   0.002     .1228124     .547143
          4  |   .4327654   .1165144     3.71   0.000     .2044013    .6611295
          5  |   .4025012   .1275277     3.16   0.002     .1525514     .652451
             |
     1.black |          0  (omitted)
         age |          0  (omitted)
------------------------------------------------------------------------------
```

Random Effects Models. Quoting Allison, “In a random effects model, the unobserved variables are assumed to be uncorrelated with (or, more strongly, statistically independent of) all the observed variables.” That assumption will often be wrong but, for the reasons given above (e.g. standard errors may be very high with fixed effects, RE lets you estimate effects for time-invariant variables), an RE model may still be desirable under some circumstances. RE models can be estimated via Generalized Least Squares (GLS). Here is an example of a random effects logistic regression model.

```
. *random effects
. xtlogit pov i.mother i.spouse i.school hours i.year i.black age, re nolog

Random-effects logistic regression              Number of obs     =      5,755
Group variable: id                              Number of groups  =      1,151

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          5
                                                              avg =        5.0
                                                              max =          5

Integration method: mvaghermite                 Integration pts.  =         12

                                                Wald chi2(10)     =     266.60
Log likelihood  = -3403.7655                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.mother |   1.009877     .118372     8.53   0.000     .7778724    1.241882
    1.spouse |  -1.171833    .1512544    -7.75   0.000    -1.468286   -.8753802
    1.school |  -.1145721    .0990775    -1.16   0.248    -.3087604    .0796163
       hours |  -.0259014    .0028771    -9.00   0.000    -.0315403   -.0202624
             |
        year |
          2  |   .2830958    .1000437     2.83   0.005     .0870138    .4791778
          3  |    .213423    .1040523     2.05   0.040     .0094842    .4173618
          4  |   .2415184    .1090094     2.22   0.027     .0278639     .455173
          5  |   .1447937    .1161395     1.25   0.212    -.0828355     .372423
             |
     1.black |   .6093942    .0975653     6.25   0.000     .4181698    .8006186
         age |  -.0627952    .0472163    -1.33   0.184    -.1553373     .029747
       _cons |  -.0045847    .7620829    -0.01   0.995     -1.49824     1.48907
-------------+----------------------------------------------------------------
    /lnsig2u |   .3086358    .1008833                      .1109083    .5063634
-------------+----------------------------------------------------------------
     sigma_u |   1.166862    .0588584                      1.057021    1.288117
         rho |   .2927197    .0208864                      .2535175    .3352612
------------------------------------------------------------------------------
LR test of rho=0: chibar2(01) = 327.62                 Prob >= chibar2 = 0.000
```

Among other things, according to this model, blacks are significantly more likely to be in poverty than are whites. The highly significant likelihood ratio test at the end tells us it would not be appropriate to use regular logistic regression instead. Note too that there are some major differences in the coefficients for the fixed and random effects models, which might reflect the importance of omitted variable bias in the latter.

Mixed Effects Model. Give or take a few decimal places, a mixed-effects model (aka multilevel model or hierarchical model) replicates the above results. Again, it is ok if the data are `xtset` but it is not required. We will (hopefully) explain mixed effects models more later.

```
. * Equivalent mixed-effects model
. xtset, clear
. melogit pov i.mother i.spouse i.school hours i.year i.black age || id:, nolog

Mixed-effects logistic regression               Number of obs     =      5,755
Group variable:              id                 Number of groups  =      1,151

                                                Obs per group:
                                                              min =          5
                                                              avg =        5.0
                                                              max =          5

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(10)     =     266.64
Log likelihood = -3403.7637                     Prob > chi2       =     0.0000
------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
             |
    1.mother |   1.009935   .1183721     8.53   0.000     .7779301     1.24194
    1.spouse |  -1.171859   .1512457    -7.75   0.000    -1.468295   -.8754231
    1.school |   -.114617   .0990711    -1.16   0.247    -.3087927    .0795587
       hours |  -.0259016   .0028769    -9.00   0.000    -.0315403   -.0202629
             |
        year |
          2  |   .2830838   .1000419     2.83   0.005     .0870052    .4791624
          3  |   .2134042     .10405     2.05   0.040       .00947    .4173385
          4  |   .2414921   .1090061     2.22   0.027      .027844    .4551401
          5  |    .144759   .1161351     1.25   0.213    -.0828617    .3723796
             |
     1.black |   .6094854   .0975621     6.25   0.000     .4182672    .8007036
         age |  -.0628037   .0472134    -1.33   0.183    -.1553403     .029733
       _cons |  -.0045483   .7620352    -0.01   0.995     -1.49811    1.489013
-------------+----------------------------------------------------------------
id           |
   var(_cons)|   1.361483   .1371712                      1.117513    1.658715
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 327.62      Prob >= chibar2 = 0.0000
```

Appendix (Optional): Estimation methods for fixed-effects models

Fixed effects models control for, or partial out, the effects of time-invariant variables with time-invariant effects. This is true whether the variable is explicitly measured or not. Exactly how they do so varies by the statistical technique being used. Some of the methods used include

- *Demeaning variables*. The within-subject means for each variable (both the Xs and the Y) are subtracted from the observed values of the variables. Hence, within each subject, the demeaned variables all have a mean of zero. For time-invariant variables, e.g. gender, the demeaned variables will have a value of 0 for every case, and since they are constants they will drop out of any further analysis. This basically gets rid of all between-subject variability (which may be contaminated by omitted variable bias) and leaves only the within-subject variability to analyze. This method works for linear regression models but does not work for things like logistic regression.
- *Unconditional maximum likelihood*. With UML, dummy variables are created for each subject (except one) and included in the model. So, for example, if you had 2000 subjects each of whom was measured at 5 points in time, you would include 1,999 dummy variables in the model. Needless to say, this can be pretty time consuming, and can produce a lot of coefficients that you aren't really interested in! However, Allison argues that it is better to use `nbreg` with UML than it is to use Stata's `xtnbreg, fe`. The latter, he claims, uses a flawed approach and does not, in fact control for all stable predictors. UML can also be used for linear regression but produces biased estimates with logistic regression.
- *Conditional maximum likelihood*. This is used for logistic regression and some other statistical techniques. Quoting Allison (p. 32; α_i refers to the fixed effects parameters),

> The solution is to do conditional maximum likelihood, which *conditions* the α_i parameters out of the likelihood function (Chamberlain, 1980). This is accomplished by conditioning the likelihood function on the total number of events observed for each person. In effect, each person's contribution to the likelihood function is the answer to a question such as the following: Given that a girl was in poverty for 2 out of the 5 years, what is the probability that this happened in, say, Years 2 and 4 (when it actually occurred) rather than in one of the nine other possible pairs of years? These conditional probabilities do not contain the α_i parameters. This conditioning approach only works for the logistic regression model for dichotomous response variables, not for other "link" functions such as probit or complementary log-log.

Note that, with the conditional logit model, for all subjects where the dependent variable is a constant (e.g. at all five time periods the subject has a value of 1 on the dependent variable, or a value of zero) the case is dropped from the statistical analysis. Basically, there is no alternative possibility to compare to, e.g. the only way you can have 5 ones is by being a one at every time period.

Before proceeding, we will show examples of UML (the dummy variable for each case approach). This will show that `regress` using UML gives the same results as `xtreg,fe` but different results when using `logit` and `xtlogit, fe`. The data sets used here are also used in Allison's book.

```
. set more off
. use https://www3.nd.edu/~rwilliam/statafiles/nlsy.dta, clear
. des anti* self* pov*

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------
---------------------------------------------------------------
anti90          byte   %8.0g                  child antisocial behavior in 1990
anti92          byte   %8.0g                  child antisocial behavior in 1992
anti94          byte   %8.0g                  child antisocial behavior in 1994
self90          byte   %8.0g                  child self-esteem in 1990
pov90           byte   %8.0g                  family poverty status in 1990
[some output deleted]

. gen id=_n
. reshape long anti pov self, i(id) j(year)
(note: j = 90 92 94)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                      581   ->    1743
Number of variables                  17   ->      12
j variable (3 values)                     ->   year
xij variables:
                   anti90 anti92 anti94   ->   anti
                      pov90 pov92 pov94   ->   pov
                   self90 self92 self94   ->   self
-----------------------------------------------------------------------------

. xtset id year
       panel variable:  id (strongly balanced)
        time variable:  year, 90 to 94, but with gaps
                delta:  1 unit

. * UML works fine with linear regression model
. xtreg anti self pov i.year, fe

Fixed-effects (within) regression               Number of obs      =      1743
Group variable: id                              Number of groups   =       581

R-sq:  within  = 0.0331                         Obs per group: min =         3
       between = 0.0418                                        avg =       3.0
       overall = 0.0359                                        max =         3

                                                F(4,1158)          =      9.92
corr(u_i, Xb)  = 0.0683                         Prob > F           =    0.0000
```

```
------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0551514   .0105258    -5.24   0.000    -.0758031   -.0344997
         pov |   .1124749   .0934099     1.20   0.229    -.0707967    .2957464
             |
        year |
         92  |   .0443934    .058584     0.76   0.449    -.0705493     .159336
         94  |   .2107366   .0587978     3.58   0.000     .0953744    .3260987
             |
       _cons |   2.637156   .2173038    12.14   0.000     2.210803     3.06351
-------------+----------------------------------------------------------------
     sigma_u |  1.3218868
     sigma_e |  .99707353
         rho |  .63737335   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(580, 1158) =     5.16           Prob > F = 0.0000

. set matsize 2000
. reg anti self pov i.year i.id

      Source |       SS       df       MS              Number of obs =    1743
-------------+------------------------------           F(584,  1158) =    5.48
       Model |  3181.88311   584  5.44842999           Prob > F      =  0.0000
    Residual |  1151.23221  1158  .994155619           R-squared     =  0.7343
-------------+------------------------------           Adj R-squared =  0.6003
       Total |  4333.11532  1742  2.48743704           Root MSE      =  .99707

------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0551514   .0105258    -5.24   0.000    -.0758031   -.0344997
         pov |   .1124749   .0934099     1.20   0.229    -.0707967    .2957464
             |
        year |
         92  |   .0443934    .058584     0.76   0.449    -.0705493     .159336
         94  |   .2107366   .0587978     3.58   0.000     .0953744    .3260987
             |
          id |
          2  |  -.8875251   .8194485    -1.08   0.279    -2.495295    .7202448
          3  |   4.130859   .8194591     5.04   0.000     2.523068    5.738649
 [Rest of coefficients for dummy variables for ids are deleted]

. * UML does not work fine with logit -- Need conditional model instead
. xtlogit pov mother spouse school hours i.year, fe nolog
note: multiple positive outcomes within groups encountered.
note: 324 groups (1620 obs) dropped because of all positive or
      all negative outcomes.

Conditional fixed-effects logistic regression   Number of obs      =      4135
Group variable: id                              Number of groups   =       827

                                                Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

                                                LR chi2(8)         =     97.28
Log likelihood  = -1520.1139                    Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mother |   .5824322   .1595831     3.65   0.000      .269655    .8952094
      spouse |  -.7477585   .1753466    -4.26   0.000    -1.091431   -.4040854
      school |   .2718653   .1127331     2.41   0.016     .0509125    .4928181
       hours |  -.0196461   .0031504    -6.24   0.000    -.0258208   -.0134714
             |
        year |
          2  |   .3317803   .1015628     3.27   0.001      .132721    .5308397
          3  |   .3349777   .1082496     3.09   0.002     .1228124     .547143
          4  |   .4327654   .1165144     3.71   0.000     .2044013    .6611295
          5  |   .4025012   .1275277     3.16   0.002     .1525514     .652451
------------------------------------------------------------------------------

. logit pov mother spouse school hours i.year i.id, nolog
note: 141.id != 0 predicts failure perfectly
      141.id dropped and 5 obs not used
note: 298.id != 0 predicts success perfectly
      298.id dropped and 5 obs not used

[Other similar warnings deleted – these are the 324 cases where the outcome is the
same at all 5 time periods for the case]

Logistic regression                               Number of obs   =       4135
                                                  LR chi2(834)    =     998.93
                                                  Prob > chi2     =     0.0001
Log likelihood = -2304.2196                       Pseudo R2       =     0.1781

------------------------------------------------------------------------------
         pov |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      mother |   .7341873    .179498     4.09   0.000     .3823778    1.085997
      spouse |  -.9407072   .1971326    -4.77   0.000     -1.32708   -.5543344
      school |   .3410341   .1264389     2.70   0.007     .0932184    .5888497
       hours |  -.0246849   .0035439    -6.97   0.000    -.0316308   -.0177391
             |
        year |
          2  |   .4196558   .1142231     3.67   0.000     .1957827     .643529
          3  |   .4218788    .121389     3.48   0.001     .1839608    .6597968
          4  |   .5452897   .1306011     4.18   0.000     .2893163    .8012631
          5  |   .5071969   .1427835     3.55   0.000     .2273463    .7870475
             |
          id |
         75  |   -.107972   1.592235    -0.07   0.946    -3.228695    3.012751
         92  |   1.206116   1.476275     0.82   0.414     -1.68733    4.099562
[Coefficients for other id dummies not shown]
```