Panel Data for Linear Models: Very Brief Overview

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised March 19, 2018

These notes borrow very heavily, often verbatim, from Paul Allison's book, *Fixed Effects Regression Models for Categorical Data*. I strongly encourage people to get their own copy. The Stata XT manual is also a good reference, as is *Microeconometrics Using Stata, Revised Edition*, by Cameron and Trivedi. Allison's book does a much better job of explaining why assertions made here are true and what the technical details behind the models are.

Overview. We often have data where variables have been measured for the same subjects (or countries, or companies, or whatever) at multiple points in time. These are typically referred to as Panel Data or as Cross-Sectional Time Series Data. We need special techniques for analyzing such data, e.g. it would be a mistake to treat 200 individuals measured at 5 points in time as though they were 1,000 independent observations, since doing so would usually result in standard error estimates that were too small. Therefore, Stata has an entire manual and suite of XT commands devoted to panel data, e.g. `xtreg`, `xtlogit`, `xtpoisson`, etc. Panel Data offer some important advantages over cross-sectional only data, only a very few of which will be covered here.

The Linear Regression Panel Model. (Adapted heavily from Allison pp. 6-7) Suppose we have a continuous dependent variable that is linearly dependent on a set of predictor variables. We have a set of individuals who are measured at two or more points of time. Allison notes that the model can be written as

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it}$$

- μ_t is an intercept term that can be different for each time period, e.g. it might be 7 at time 1 and 3 at time 2. μ_t does NOT vary across cases, it only varies across time.
- x stands for the independent variables whose values can vary across time, e.g. income, marital status. We say that these are *time-varying variables*.
- z stands for the independent variables whose values do NOT change across time, e.g. race, gender. We can say that these variables have *time-invariant values* or measure *stable characteristics*.
- β and γ are the coefficients for the xs and zs. As written, *the model assumes that these effects are time-invariant*, e.g. the effect of x1 is the same at same 1 as it is at time 4 (although the value of x1 can be different at different time periods). Interactions with time can be added if the effects of the xs or zs are thought to vary with time (e.g. maybe race is thought to have less effect at time 1 than it does at time 4).
- α_i and ε_{it} are both error terms. ε_{it} is different for each individual at each point in time. α_i only varies across individuals but not across time. We can think of α_i as representing the effects of all the time invariant/stable variables that have NOT been included in the model. So, for example, if data from 4 time periods were collected for each case, then the four records for Case 1 would all have the same value for α_1, the four records for case 2 would all have the same value for α_2, etc. But, ε_{it} is free to be different for every case at every time period.

- The assumptions we make about α_i help to determine what kind of panel model we should estimate. Remember that in the past, we have said that error terms should be uncorrelated with the explanatory variables in the model. This assumption might be violated if, say, relevant variables have been omitted from the model. If we believe that α_i is correlated with the xs (the time-varying explanatory variables) then we can estimate what is known as a *fixed effects model*.
    - The fixed effects method controls for time-invariant variables that have not been measured but that affect y. For example, it could control for the effect of race if information on race was not available in the data set.
    - However, while the effects of time-invariant variables (measured or unmeasured) can be controlled for, their effects cannot actually be estimated, i.e. we cannot estimate the γs for the model.
- If, on the other hand, α_i is uncorrelated with the xs (e.g. because no time-invariant variables are omitted, or because the variables that are omitted are not correlated with the variables that are in the model) then a *random effects model* can provide unbiased estimates of both the βs and the γs, and will generally have lower standard errors than a fixed effects model.

Fixed Effects Models. In experimental research, unmeasured differences between subjects are often controlled for via random assignment to treatment and control groups. Hence, even if a variable like Socio-Economic Status is not explicitly measured, because of random assignment, we can be reasonably confident that the effects of SES are approximately equal for all groups. Of course, random assignment is usually not possible with most survey research. If we want to control for the effect of a variable, we must explicitly measure it. If we don't measure it, we can't control for it. In practice, there will almost certainly be some variables we have failed to measure (or have measured poorly), so our models will likely suffer from some degree of omitted variable bias.

Allison notes, however, that when we have panel data (the same subjects measured at two or more points in time) another alternative presents itself: we can use the subjects as their own controls. With panel data we can control for stable characteristics (i.e. characteristics that do not change across time) whether they are measured or not. These include such things as sex, race, and ethnicity, as well as more difficult to measure variables such as intelligence, parents' child-rearing practices, and genetic makeup. The idea is that, whatever effect these variables have at one point in time, they will have the same effect at a different point in time because the values of such variables do not change.

We can do this via *fixed effects models.* Such models do not control for time-varying variables, but such variables can be explicitly included in the model, e.g. employment status, income. Also, they do not control for unmeasured stable characteristics whose effects change across time (e.g. the effect of gender on learning might be different at different ages).

Examples (from Allison): Suppose you want to know whether marriage reduced recidivism among chronic offenders. We could compare an individual's arrest rate when he is married with his arrest rate when he is not. The difference in arrest rates between the two periods is an estimate of the marriage effect for that individual. Or, you might see how a child's performance

in school differs depending on how much time s/he spends playing video games. So, you could compare how the child does when not spending much time on video games versus when s/he does.

Estimation of fixed effects models when T = 2. As Allison notes (p. 7) it is especially easy to see how and why fixed effects models work in the 2 period case (T = 2). The equations for the two periods can be written as

$$y_{i1} = \mu_1 + \beta x_{i1} + \gamma z_i + \alpha_i + \varepsilon_{i1}$$
$$y_{i2} = \mu_2 + \beta x_{i2} + \gamma z_i + \alpha_i + \varepsilon_{i2}$$

If we subtract the first equation form the second, we get

$$\begin{aligned} y_{i2} - y_{i1} &= (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\gamma z_i - \gamma z_i) + (\alpha_i - \alpha_i) + (\varepsilon_{i2} - \varepsilon_{i1}) \\ &= (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1}) \end{aligned}$$

which can be rewritten as

$$\Delta y_i = \Delta\mu + \beta\Delta x_i + \Delta\varepsilon_i$$

where Δ indicates a difference score. Note that both α_i (which represents the effects of all the time invariant/stable variables that have NOT been included in the model) and z_i (the time invariant variables that have been included) have been differenced out. *We therefore no longer have to worry about the effects of omitted time-invariant variables.* We can do this because, whatever effect the time invariant variables have, it is the same at both time 1 and time 2. At the same time, we can't estimate the effects of the time invariant variables (the γs) even when those variables are measured.

In practice, you can do this via something like

```
gen ydif = y2 – y1
gen xdif = x2 – x1
reg ydif xdif
```

The differencing approach doesn't work when there are more than 2 periods. Instead you can use one of the other methods described below.

Estimation of fixed effects models when T >= 2. 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 it does so varies by the statistical technique being used.

In the case of quantitative dependent variables analyzed in linear regression models, a commonly used approach is *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 quantitative variables in linear regression models but does not work for things like logistic regression. This is the procedure used by Stata's `xtreg` command. Methods used for other types of statistical problems (e.g. logistic regression, count models) include *Unconditional Maximum Likelihood* (UML) and *Conditional Maximum Likelihood*. The example below illustrates the demeaning approach while Appendix B illustrates UML.

Fixed Effects Example. Allison (starting on p. 7 of his book) gives an example using the National Longitudinal Survey of Youth. This subset of the data set has 581 children who were interviewed in 1990, 1992, and 1994. The numbers at the ends of some variable names reflect the time period the variable refers to (90 = 1990, 92 = 1992, 94 = 1994.) Variables without numbers in the names do not vary across time. Variables used in this example include

- id is the subject id number and is the same across each wave of the survey
- $anti_t$ is Antisocial behavior (scale ranges from 0 to 6)
- $self_t$ – Self esteem (scale ranges from 6 to 24)
- pov_t – coded 1 if family is in poverty, 0 otherwise
- black is coded 1 if the child is black, 0 otherwise
- hispanic is coded 1 if the child is Hispanic, 0 otherwise
- childage is child's age in 1990
- married is coded 1 if the child's mother was currently married in 1990, 0 otherwise
- gender is coded 1 if the child is female, 0 if male
- momage is the mother's age at birth of child
- momwork is coded 1 if the mother was employed in 1990, 0 otherwise

In this example, we are first interested in examining how antisocial behavior is affected by self esteem, poverty and the year the data were collected in. The data are currently in wide format, e.g. there is one record per case, and the time varying variables have names like anti90, anti92, anti94. The data needs to be restructured into long format first, i.e. there is one record for each case for each time period. Appendix A describes in more detail how to set up the data.

```
. 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
```

If the `xtreg` command did not exist, we could estimate a fixed effects model by using OLS regression with the demeaning approach. We compute the mean of each variable for each case, subtract the mean from the original variable, and then run a regression on the demeaned variables. So, for example, if the three values of anti for case 1 were 2, 4, and 6, the mean of anti for case 1 would be 4, and the demeaned values of anti would then be -2, 0, and 2. The following code shows how to do this:

```
. egen antix = mean(anti), by(id)
. egen selfx = mean(self), by(id)
. egen povx = mean(pov), by(id)
. gen antidif = anti - antix
. gen selfdif = self - selfx
. gen povdif = pov - povx
. reg antidif selfdif povdif i.year

      Source |       SS       df       MS              Number of obs =    1743
-------------+------------------------------           F(  4,  1738) =   14.88
       Model |  39.4344602     4  9.85861505           Prob > F      =  0.0000
    Residual |   1151.2322  1738   .66238907           R-squared     =  0.0331
-------------+------------------------------           Adj R-squared =  0.0309
       Total |  1190.66666  1742  .683505547           Root MSE      =  .81387

------------------------------------------------------------------------------
     antidif |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     selfdif |  -.0551514   .0085918    -6.42   0.000    -.0720027   -.0383001
      povdif |   .1124749   .0762469     1.48   0.140    -.0370704    .2620202
             |
        year |
          92 |   .0443934   .0478199     0.93   0.353    -.0493972    .1381839
          94 |   .2107366   .0479944     4.39   0.000     .1166037    .3048694
             |
       _cons |  -.0850433   .0338549    -2.51   0.012    -.1514439   -.0186428
------------------------------------------------------------------------------
```

Substantive Explanation & Interpretation. Before going on, let's think about the logic behind the demeaning approach. After demeaning, all variables for all cases have a mean of 0. That means that all the between-subject variability has been eliminated. All that is left is the within-subject variability. So, with a fixed effects model, we are analyzing what causes individual's values to change across time. Variables whose values do not change (like race or gender) cannot cause changes across time (unless their effects change across time as well). However, whatever effect they have at one time is the same effect that they have at other times, so the effects of such stable characteristics are controlled.

To think of it another way: when describing the effects of variables in a regression model, we have often used phrases like "Suppose we had two otherwise identical individuals where one of them had one more year of education than the other did. How would their expected income values differ?" We have also said things like "Suppose an individual got one more year of education. How would we expect his/her income to change?" Both statements are legitimate interpretations of effects. But fixed effects models are basically doing the latter: They estimate how changes within individuals across time affect their outcomes.

The regression approach gives the right coefficient estimates (except for the constant), *but the standard errors are wrong* because the estimation does not take into account the fact that the cases are not independent of each other. Luckily, the `xtreg` command does exist. It will demean the variables for you and estimate the standard errors correctly.

```
. 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
```

Incidentally, note that we have not included any of the time invariant variables, i.e. the variables whose values do not change across time (e.g. black, gender), in the fixed effects model. Let's see what happens when we do.

```
. xtreg anti self pov i.year black hispanic childage married gender momage momwork, fe
note: black omitted because of collinearity
note: hispanic omitted because of collinearity
note: childage omitted because of collinearity
note: married omitted because of collinearity
note: gender omitted because of collinearity
note: momage omitted because of collinearity
note: momwork omitted because of collinearity

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
             |
       black |          0  (omitted)
    hispanic |          0  (omitted)
    childage |          0  (omitted)
     married |          0  (omitted)
      gender |          0  (omitted)
      momage |          0  (omitted)
     momwork |          0  (omitted)
       _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
```

The time invariant variables get dropped and no effects are estimated (likewise with all the other time-invariant variables). Why? Realize that, once you demean a variable like black (i.e. subtract each case's mean for black from the case's value for black) the demeaned variable always has a value of 0 and is hence a constant.

So, the bad news is that the effects of time-invariant variables like black cannot be estimated in a fixed effects model. The good news is that, so long as the effects of the time-invariant variables are also time invariant (e.g. black has the same effect in 1990 as in 1992 as in 1994) those variables are controlled for *whether we have measured them or not*. (For stable characteristics that are measured, I could include interactions with time if I thought the effects were not time-invariant.)

Random effects models. Another popular approach is to use random effects models. Linear Random effects models are estimated via Generalized Least Squares (GLS) which I won't try to explain here. If there are no omitted variables (or if the omitted variables are uncorrelated with the variables that are in the model) then a random effects model is preferable to fixed effects because (a) the effects of time-invariant variables like race or gender can be estimated, rather than just controlled for, and (b) standard errors of estimates tend to be smaller. However, if relevant time-invariant variables have been omitted from the model, coefficients may be biased. We will start by estimating a random effects model that only includes the time varying variables, i.e. the random effects version of the fixed effects model we have been estimating.

```
. xtreg anti self pov i.year, re

Random-effects GLS regression                   Number of obs      =      1743
Group variable: id                              Number of groups   =       581

R-sq:  within  = 0.0309                         Obs per group: min =         3
       between = 0.0580                                        avg =       3.0
       overall = 0.0458                                        max =         3

                                                Wald chi2(4)       =     65.18
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0597404   .0095401    -6.26   0.000    -.0784387   -.0410422
         pov |   .2959055   .0774595     3.82   0.000     .1440877    .4477234
             |
        year |
          92 |   .0469942    .058749     0.80   0.424    -.0681516    .1621401
          94 |    .215774   .0589213     3.66   0.000     .1002903    .3312577
             |
       _cons |   2.667697   .2038331    13.09   0.000     2.268191    3.067202
-------------+----------------------------------------------------------------
     sigma_u |   1.169244
     sigma_e |  .99707353
         rho |  .57897725   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Compared to the earlier fe results, the most striking difference is that the effect of poverty is both larger and much more statistically significant than before (.112 with a standard error of .09 in the fe model, compared to .296 with a standard error of .08 in the re model). When such discrepancies occur, Allison says that one possible explanation is that the standard errors are higher in the fe model. In this case, they don't differ by that much. The other possible explanation is that the magnitudes of the estimated efforts are different, and in this case the estimated re effect for pov is more than twice as large as the estimated fe effect. Allison says (p. 22) that "The most plausible explanation is that there are unobserved [time invariant] variables that explain away the observed association between poverty and antisocial behavior. When these variables are controlled, via fixed effects, the relationship disappears." Such time invariant variables might include gender, race, parental characteristics, the way in which the respondent was raised, etc. [NOTE: Appendix C shows how to test whether a random effects model is justified, or whether you should stick with a fixed effects model instead].

Finally, let's see what happens when the time invariant variables are added to the re model.

```
. xtreg anti self pov i.year black hispanic childage married gender momage momwork, re

Random-effects GLS regression                   Number of obs      =      1743
Group variable: id                              Number of groups   =       581

R-sq:  within  = 0.0320                         Obs per group: min =         3
       between = 0.1067                                        avg =       3.0
       overall = 0.0853                                        max =         3

                                                Wald chi2(11)      =    104.53
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0620586    .009518    -6.52   0.000    -.0807135   -.0434036
         pov |    .246818   .0804041     3.07   0.002     .0892288    .4044072
             |
        year |
          92 |   .0473322   .0587008     0.81   0.420    -.0677193    .1623836
          94 |   .2163669   .0588738     3.68   0.000     .1009763    .3317575
             |
       black |   .2268535   .1255617     1.81   0.071     -.019243    .4729499
    hispanic |  -.2181591   .1380795    -1.58   0.114      -.48879    .0524718
    childage |   .0884583   .0909947     0.97   0.331     -.089888    .2668047
     married |   -.049499   .1262863    -0.39   0.695    -.2970156    .1980176
      gender |  -.4834304   .1064056    -4.54   0.000    -.6919815   -.2748793
      momage |  -.0219284   .0252608    -0.87   0.385    -.0714386    .0275818
     momwork |   .2612145   .1145722     2.28   0.023     .0366571     .485772
       _cons |   2.531237   1.094669     2.31   0.021     .3857254    4.676749
-------------+----------------------------------------------------------------
     sigma_u |  1.1355938
     sigma_e |  .99707353
         rho |  .56467881   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

We see that the effects of the stable characteristics can now be estimated. Females have significantly lower scores on the antisocial behavior scale. This model again produces a larger and more statistically significant effect for pov than the fixed effects model does. But, this could still be due to the failure to control for omitted variables (although it is hopefully less likely now since so many more variables have been included).

Summary

- With fixed effects models, the good thing is that the effects of stable characteristics, such as race and gender, are controlled for, whether they are measured or not. The bad thing is that the effects of these variables are not estimated. Again, it is similar to an experiment with random assignment. The effects of variables not explicitly measured are controlled for (because random assignment makes the groups more or less similar on these characteristics) but their effects are not estimated.
    - Keep in mind, however, that fixed effects doesn't control for unobserved variables that change over time. So, for example, a failure to include income in the model could still cause fixed effects coefficients to be biased.
    - Fixed effects models also won't control for the effects of omitted time invariant variables that have time-varying effects, e.g., you will have problems if race is not measured and the effect of race is different at different time periods.

    - However, if measures of such time invariant variables exist in the data, you can include interactions of them with time to estimate their time varying effects.
- Other methods (e.g. random effects) can be used when we want to estimate the effects of variables like sex and race, but then the method is no longer controlling for omitted variables.
    - If you are confident that there are no omitted variables, then random effects may be fine and actually preferable. The effects of both the time varying and time invariant variables can be estimated and the standard errors will be lower than with fixed effects.
    - See Appendix C for details on how to decide whether use of a random effects model is justified.
- Substantive interests are also going to drive your choice of model. If you are really interested in the effects of race and gender, a fixed effects model really isn't an option.
- Why do fixed effects models have higher standard errors? Fixed effects estimates *use only within-individual differences*, essentially discarding any information about differences *between* individuals. If predictor variables vary greatly across individuals but have little variation over time for each individual, then fixed effects estimates will be imprecise and have large standard errors. Put another way, fixed effects methods basically discard a lot of information; hence the standard errors are higher.
    - Why tolerate the higher errors? Allison says there is a trade-off between bias and efficiency. Other methods, e.g. random effects, will suffer from omitted variable bias; fixed effects methods help to control for omitted variable bias by having individuals serve as their own controls.
    - Allison likes fixed effects models because they are less vulnerable to omitted variable bias. But he cautions that "in applications where the within-person variation is small relative to the between-person variation, the standard errors of the fixed effects coefficients may be too large to tolerate."

Appendix A: Setting up the data

In order to use Stata's XT commands, the data set needs to be properly structured. This will sometimes require that the data be restructured from wide to long. In wide format, a data set has one record for each subject. This record has several variables, e.g. income1, income2, income3, where each of the income variables gives the value of income at a different time point. In long format, the data are restructured to have one record for each subject for each time point. I am going to give some examples of how to do this, but if in doubt be sure to read the Stata documentation for help on setting up your data.

Here is an example from Allison's 2009 book *Fixed Effects Regression Models*. Data are from the National Longitudinal Study of Youth (NLSY). This subset of the data set (which is different than the subset used earlier) has 1151 teenage girls who were interviewed annually for 5 years beginning in 1979. Here is a listing of the values for the first three cases:

```
. use https://www3.nd.edu/~rwilliam/statafiles/teenpov.dta, clear
. rename inschool* school*
. list in 1/3
```

1.

id	pov1	mother1	spouse1	school1	hours1	pov2	mother2	spouse2	school2	hours2	pov3	mother3	spouse3	school3
22	1	0	0	1	21	0	0	0	1	15	0	0	0	1

hours3	pov4	mother4	spouse4	school4	hours4	age	black	pov5	mother5	spouse5	school5	hours5
3	0	0	0	1	0	16	0	0	0	0	1	0

2.

id	pov1	mother1	spouse1	school1	hours1	pov2	mother2	spouse2	school2	hours2	pov3	mother3	spouse3	school3
75	0	0	0	1	8	0	0	0	1	0	0	0	0	1

hours3	pov4	mother4	spouse4	school4	hours4	age	black	pov5	mother5	spouse5	school5	hours5
0	0	0	0	1	4	17	0	1	0	0	1	0

3.

id	pov1	mother1	spouse1	school1	hours1	pov2	mother2	spouse2	school2	hours2	pov3	mother3	spouse3	school3
92	0	0	0	1	30	0	0	0	1	27	0	0	0	1

hours3	pov4	mother4	spouse4	school4	hours4	age	black	pov5	mother5	spouse5	school5	hours5
24	1	1	0	0	31	16	0	1	1	0	0	0

The numbers at the ends of some variable names reflect the time period the variable refers to (1 = 1979, 2 = 1980, etc.) Variables without numbers in the names do not vary across time.

- id is the subject id number and is the same across each wave of the survey
- pov_t is coded 1 if the subject was in poverty during that time period, 0 otherwise.
- age is the age at the first interview.
- black is coded 1 if the respondent is black, 0 otherwise.
- $mother_t$ is coded 1 if the respondent currently has at least 1 child, 0 otherwise.
- $spouse_t$ is coded 1 if the respondent is currently living with a spouse, 0 otherwise.
- $school_t$ is coded 1 if the respondent is currently in school, 0 otherwise.
- $hours_t$ is the hours worked during the week of the survey.

The data are currently in wide format. There is one record per case with multiple variables representing values at different points in time. We need to get the data into long format instead. In Stata, we can do this with the `reshape` command.

```
. reshape long pov mother spouse school hours, i(id) j(year)
(note: j = 1 2 3 4 5)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                     1151   ->    5755
Number of variables                  28   ->       9
j variable (5 values)                     ->   year
xij variables:
                      pov1 pov2 ... pov5   ->   pov
             mother1 mother2 ... mother5   ->   mother
             spouse1 spouse2 ... spouse5   ->   spouse
             school1 school2 ... school5   ->   school
                hours1 hours2 ... hours5   ->   hours
-----------------------------------------------------------------------------
```

The `reshape long` part of the command told Stata we wanted to reshape the data from wide to long. (There is also a `reshape wide` command for going from long to wide.) The variable list that followed was the list of variables (actually the stubnames of the variables) that varied across time (you should use a consistent naming convention, e.g. pov1, mother1, etc. pov79, mother79, pov80, mother80, would have also been ok. Be careful about doing something like inc2, inc79, inc80, inc81, where inc2 = income squared; Stata will think inc2 is another of the time-varying variables.) The variables not listed are those that do not vary across time; their values will be copied on to each of the new records for the case. `i(varlist)` specifies the variables whose unique values denote a logical observation. i() is required. In this case only `i(id)` was needed but in other cases multiple variables might define a case. `j(varname)` specifies the variable whose unique values denote a subobservation. Here is what the reshaped data for the first 3 (now 15) cases looks like.

```
. list in 1/15
```

	id	year	age	black	pov	mother	spouse	school	hours
1.	22	1	16	0	1	0	0	1	21
2.	22	2	16	0	0	0	0	1	15
3.	22	3	16	0	0	0	0	1	3
4.	22	4	16	0	0	0	0	1	0
5.	22	5	16	0	0	0	0	1	0
6.	75	1	17	0	0	0	0	1	8
7.	75	2	17	0	0	0	0	1	0
8.	75	3	17	0	0	0	0	1	0
9.	75	4	17	0	0	0	0	1	4
10.	75	5	17	0	1	0	0	1	0
11.	92	1	16	0	0	0	0	1	30
12.	92	2	16	0	0	0	0	1	27
13.	92	3	16	0	0	0	0	1	24
14.	92	4	16	0	1	1	0	0	31
15.	92	5	16	0	1	1	0	0	0

Each of the original cases now has 5 records, one for each year of the study. The value of year varies from 1 to 5. The values of age (age at first interview) and black have been duplicated on each of the 5 records. Instead of 5 poverty variables, we have 1, whose value can differ across the five records (e.g. the original value of pov2 for id 22 is now the value of pov for id 22 year 2). The same is true for the other time-varying variables.

The next thing we want to do is xtset the data. The `xtset` command tells Stata that these are Panel data. The usual format is

```
xtset panelvar
xtset panelvar timevar
```

That is, we must tell Stata what the panelvar is; in this case it is id. The timevar is optional and may or may not be necessary depending on our analysis. In the current case the timevar is year. `xtset` typed with no parameters tells us how the data are xtset.

```
. xtset id year
       panel variable:  id (strongly balanced)
        time variable:  year, 1 to 5
                delta:  1 unit

. xtset
       panel variable:  id (strongly balanced)
        time variable:  year, 1 to 5
                delta:  1 unit
```

NOTE (copied verbatim from the Stata 12 Manual): "The terms balanced and unbalanced are often used to describe whether a panel dataset is missing some observations. If a dataset does not contain a time variable, then panels are considered balanced if each panel contains the same number of observations; otherwise, the panels are unbalanced. When the dataset contains a time variable, panels are said to be strongly balanced if each panel contains the same time points, weakly balanced if each panel contains the same number of observations but not the same time points, and unbalanced otherwise."

A data set might be unbalanced because data are missing for some years. If you were, say, analyzing countries, it might even be that the country did not exist during some time periods. Strongly balanced data are best but my understanding is that Stata can generally do a good job with unbalanced data.

Once the data are xtset, several commands are available to us; see `help xt`. For example, you can use the `xtsum` command, which is similar to the `summarize` command but contains some additional information.

```
. xtsum

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
id       overall |  6016.672   3298.064         22      12539 |     N =    5755
         between |             3299.211         22      12539 |     n =    1151
         within  |                   0   6016.672   6016.672 |     T =       5
                 |                                          |
year     overall |         3   1.414336          1          5 |     N =    5755
         between |                   0          3          3 |     n =    1151
         within  |             1.414336          1          5 |     T =       5
                 |                                          |
age      overall |  15.64639    1.04682         14         17 |     N =    5755
         between |             1.047184         14         17 |     n =    1151
         within  |                   0   15.64639   15.64639 |     T =       5
                 |                                          |
black    overall |  .5742832   .4944942          0          1 |     N =    5755
         between |             .4946661          0          1 |     n =    1151
         within  |                   0   .5742832   .5742832 |     T =       5
                 |                                          |
pov      overall |  .3768897    .484649          0          1 |     N =    5755
         between |             .3100424          0          1 |     n =    1151
         within  |             .3725925  -.4231103    1.17689 |     T =       5
                 |                                          |
mother   overall |  .1986099   .3989883          0          1 |     N =    5755
         between |             .3253864          0          1 |     n =    1151
         within  |             .2310605  -.6013901   .9986099 |     T =       5
                 |                                          |
spouse   overall |  .0992181   .2989806          0          1 |     N =    5755
         between |             .2206498          0          1 |     n =    1151
         within  |             .2018338  -.7007819   .8992181 |     T =       5
                 |                                          |
school   overall |  .6304083   .4827361          0          1 |     N =    5755
         between |              .32013          0          1 |     n =    1151
         within  |             .3614169  -.1695917   1.430408 |     T =       5
                 |                                          |
hours    overall |  8.671764   14.54341          0         90 |     N =    5755
         between |             9.363817          0       52.4 |     n =    1151
         within  |             11.13062  -43.72824   72.07176 |     T =       5
```

The different values for the standard deviations can sometimes be useful. For id, age and black, the within standard deviation is 0. This is because, within each subject, the value of these variables does not vary, i.e. for each of the five records the case has, the values of these variables are the same. For year, the between subjects standard deviation is 0. This is because all subjects have the same set of values on year. For poverty, the between and within standard deviations are nearly the same. This tells us that the variation in poverty across women is nearly equal to that observed within a woman over time. That is, if you were to draw two women randomly from the data, the difference in poverty is expected to be nearly equal to the difference for the same woman in two randomly selected years.

Some techniques, such as fixed effects models, work much better when there is a lot of within-subject variability (or conversely, they don't work well when subjects change little across time). There are many advantages to fixed-effects models, but some types of data are friendlier to them than are others.

Appendix B: Unconditional Maximum Likelihood Estimation

The `regress` command we used earlier produced correct estimates of the coefficients; but the standard errors were wrong because regress did not know that the cases were not independent of each other. But if your life depends on it, and you have plenty of time and a powerful enough computer, `regress` can provide coefficient estimates and standard errors that are correct. From a pedagogical standpoint, this approach may also help to clarify the idea of a fixed effect.

This can be done via unconditional maximum likelihood. With unconditional maximum likelihood, a dummy variable is added for each case (except 1). So, for example, if you have 581 cases, 580 dummy variables will be added to the model. The coefficients for the dummy variables represent the fixed effects. Going back to our earlier example,

```
. use https://www3.nd.edu/~rwilliam/statafiles/nlsyxt.dta, clear
. 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
[output deleted…]
         580 |   .4090406   .8194723     0.50   0.618    -1.198776    2.016857
         581 |  -.7946292   .8153047    -0.97   0.330    -2.394269    .8050105
             |
       _cons |    2.05258   .6297253     3.26   0.001       .81705    3.288111

. reg anti self pov i.year i.id black hispanic childage married gender momage momwork
note: black omitted because of collinearity
note: hispanic omitted because of collinearity
[Rest of output deleted]
```

Comparing this to the earlier fe results, you see we now get the correct coefficients and standard errors. We still can't estimate effects for the time invariant variables. However, we also get hundreds of coefficients that we probably don't care about. Plus, when there are a large number of cases, the model can be slow or impossible to estimate. Finally, while the UML approach works ok for linear regression, it gives incorrect results when used with logistic regression and various other techniques.

Appendix C: Testing the assumptions of the random effects model

The random effects model assumes that α_i is uncorrelated with the xs (perhaps because no time-invariant variables are omitted, or because the variables that are omitted are not correlated with the variables that are in the model) then a random effects model can provide unbiased estimates of both the betas and the gammas, and will generally have lower standard errors than a fixed effects model. If the assumption is violated though, parameter estimates will be biased. It would be nice if the assumptions of the re model were met, but how can we do a formal test of whether they are or not?

As the Stata 12 manual points out, the Hausman specification test compares an estimator that is known to be consistent/unbiased (in this case, the fixed effects model) with an estimator that is known to be efficient (in this case, the random effects model) under the assumption being tested. Here, it is testing whether the fixed effects estimates are the same as the random effects estimates. If the estimates do not significantly differ, then the random effects estimator is preferable because its standard errors are lower. In the case of the NLSY example we used earlier, Allison says the fairest test is to compare the fixed effects model with a random effects model that controls for several of the time invariant variables that are available in the data set.

In Stata, you can conduct the test as follows. (I am avoiding the use of factor variables since the `hausman` command gives erroneous albeit harmless warning message when they are used. Also, I am using a version of the NLSY data that has already been xtset.)

```
. use https://www3.nd.edu/~rwilliam/statafiles/nlsyxt.dta, clear
. quietly tab1 year, gen(yr)
. quietly xtreg anti self pov yr2 yr3, fe
. estimates store fixed
. quietly xtreg anti self pov yr2 yr3 black hispanic childage married gender momage momwork, re
. estimates store random
. hausman fixed random

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+------------------------------------------------------------
        self |   -.0551514    -.0620586        .0069072        .0044943
         pov |    .1124749     .246818       -.1343431        .0475455
         yr2 |    .0443934    .0473322       -.0029388               .
         yr3 |    .2107366    .2163669       -.0056303               .
------------------------------------------------------------------------------
                           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)
                          =       10.01
                Prob>chi2 =      0.0403
                (V_b-V_B is not positive definite)
```

Because the p value is .04, Allison says there is "some evidence against the random effects model and in favor of the fixed effects model." Allison also suggests some alternative tests that may be better than the Hausman test.