Panel Data and Multilevel Models for Categorical Outcomes: Hybrid Models [DRAFT]

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Institute for Political Methodology, Taiwan, July 17 & 18, 2018

These notes borrow very heavily from Paul Allison's book, *Fixed Effects Regression Models for Categorical Data*, and from Schunk and Perales "Within- and between-cluster effects in generalized linear mixed models: A discussion of approaches and the xthybrid command" The Stata Journal Volume 17 Number 1: pp. 89-115 The Stata XT and ME manuals are also good references.

As we have seen, Fixed Effects Models can, under the right conditions, control for the effects of time-invariant variables with time invariant effects, whether those variables are explicitly measured or not. Unfortunately, while such effects can be controlled for, they cannot be estimated in a fixed effects model. For example,

```
. use https://www3.nd.edu/~rwilliam/statafiles/teenpovxt, clear
. xtlogit pov i.mother i.spouse i.school hours i.year age i.black, 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: age omitted because of no within-group variance.
note: 1.black 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
             |
         age |          0  (omitted)
     1.black |          0  (omitted)
------------------------------------------------------------------------------
```

Despite their many virtues, fixed effects models have attracted criticisms and concerns.

- As noted, they can help avoid omitted variable bias, by controlling for time-invariant variables that may not have even be measured.

- BUT, the tradeoff is that, while these variables can be controlled for, their effects cannot be estimated.
- As Schunk and Perales note, in multilevel analysis this is often a major concern, because (p. 94) “the interest often lies in these effects, for example, how the characteristics of neighborhoods, schools, workplaces, or geographical areas influence individuals’ outcomes.”
- Another example: suppose your dissertation examined the effects of gender on earnings, and the model you were using did not allow you to estimate the effects of gender!
- Schunk and Perales add that “Because the fixed-effects approach discards all contextual (level-two) information, some argue that it is generally less preferable than the random-effects approach for multilevel analysis.” [Emphasis added.]
- Put another way, some researchers would prefer to put up with some omitted-variable bias if, in exchange, they could examine the effects of critical variables they were especially interested in.

A *hybrid model* may be a compromise. A hybrid model makes it possible to get unbiased estimates for some variables (indeed, estimates are nearly identical to estimates from an FE model) while at the same time being able to estimate effects for time-invariant or group-invariant variables like gender. Schunk and Perales write the model as

$$g(\mu_{ij}) = \beta_W(x_{ij} - \overline{x}_i) + \beta_B\overline{x}_i + \gamma c_i + u_i$$

Conceptually, the procedure is as follows:

- Within each group, calculate the mean for each independent time-varying variable (the x_{ij} variables, or level 1 variables). The means will represent the between-group differences (i.e. group means will differ between clusters but not within them).
- Then, again within each group, subtract the mean for the group from each time-varying variable. These deviations from the group mean will represent the within-group variability.
- Estimate an RE (not FE) model that includes both the means of the variables and the difference-from-the-means variables.
- Unlike a regular FE model, you can also include time-invariant/level 2 variables) variables (the c_i variables) like gender and estimate their effects.
- Allison (2009) shows how you can write Stata code yourself:

```
*** Hybrid Model by hand
use https://www3.nd.edu/~rwilliam/statafiles/teenpovxt, clear
tab year, gen(yr)
gen mysample = !missing(pov, mother, age, black, spouse, school, hours, yr2-yr5, id)
foreach var of varlist mother spouse school hours yr2-yr5 {
       egen m`var' = mean(`var') if mysample, by (id)
}

foreach var of varlist mother spouse school hours yr2-yr5 {
       gen d`var' = `var' - m`var' if mysample
}
xtlogit pov dmother-dyr5 mmother-myr5 age i.black , nolog re
```

However, life is much simpler if you use Perales and Schunk’s `xthybrid` command, available from SSC, which automates the whole process. For this problem,

```
. xthybrid pov age black, use(mother spouse school hours yr2-yr5) ///
>        family(binomial) link(logit) clusterid(id) star


Hybrid model. Family: binomial. Link: logit.

+--------------------------------------+
|             Variable |     model     |
|----------------------+---------------|
| pov                  |               |
|               R__age |    -0.1233*   |
|             R__black |     0.5719*** |
|            W__mother |     0.5939*** |
|            W__spouse |    -0.8068*** |
|            W__school |     0.2754*   |
|             W__hours |    -0.0210*** |
|               W__yr2 |     0.3329**  |
|               W__yr3 |     0.3296**  |
|               W__yr4 |     0.4307*** |
|               W__yr5 |     0.3913**  |
|            B__mother |     1.0797*** |
|            B__spouse |    -2.1469*** |
|            B__school |    -1.3625*** |
|             B__hours |    -0.0468*** |
|               B__yr2 |  (omitted)    |
|               B__yr3 |  (omitted)    |
|               B__yr4 |  (omitted)    |
|               B__yr5 |  (omitted)    |
|                _cons |     2.1900**  |
|----------------------+---------------|
|        var(_cons[id])|               |
|                _cons |     1.2488*** |
|----------------------+---------------|
| Statistics           |               |
|                   ll | -3363.5329    |
|                 chi2 |   334.1950    |
|                    p |     0.0000    |
|                  aic |  6759.0657    |
|                  bic |  6865.5909    |
+--------------------------------------+
   legend: * p<.05; ** p<.01; *** p<.001
Level 1: 5755 units. Level 2: 1151 units.
```

- You are primarily interested in the variables that start with R_ (the coefficients for the time-invariant variables) and those that start with W_ (which show the effects of within-group variability).
- If the assumptions of the random effects model are true, the coefficients for the B_ variables (between-group) should equal the coefficients for the corresponding W_ variables. xthybrid has a test option that lets you test whether or not the assumptions hold.
- If you don’t like the way the results are displayed, xthybrid has options for changing their appearance.

- In this example, the estimates for the W (within) variables are very similar to the corresponding estimates from the FE model. Furthermore, you now get estimates for the time-invariant variables. Blacks are significantly more likely to be in poverty, while those who were older at the time of the first interview are somewhat less likely.

The `xthybrid` command has some limitations that might sometimes make you prefer to compute all the necessary variables yourself.

- Factor variable notation (e,g. i.gender) is not supported. You need to create any dummy variables yourself.
- Temporary variables are created but then deleted. As a result, some post-estimation commands (e.g. predict) will not work.

More importantly, hybrid models themselves have some limitations.

- You may not be able to estimate marginal effects correctly with them (however, estimating marginal effects after any FE model can be problematic)
- Other post-estimation commands available with fe models may or may not work correctly.
- Schunk (Stata Journal, 2013) notes various other limitations, e.g. including interaction terms can be cumbersome.
- Nevertheless, Schunk concludes "[hybrid] models are useful extensions to the standard random-effects and fixed-effects approaches."