Panel Data and Multilevel Models for Categorical Outcomes: Fixed effects and conditional logit models

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*. The Stata XT and ME manuals are also good references. See Allison's book for a more detailed explanations of why assertions made here are true and what the technical details behind the models are.

Overview. 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 binary dependent variables, this can be done via the use of *conditional logit/fixed effects logit models*. 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. This does not control for time-varying variables, but such variables can be explicitly included in the model, e.g. employment status, income.

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.

Allison notes there are two conditions for using fixed effects methods.

- The dependent variable must be measured on at least two occasions for each individual.
- The independent variables must change across time for some substantial portion of the individuals. Fixed effects models are not much good for looking at the effects of variables that do not change across time, like race and sex.

There are several other points to be aware of with fixed effects logit 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.
- 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.
- 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.
    - 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.
    - 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.
    - 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."
- Conditional logit/fixed effects models can be used for things besides Panel Studies. For example, Long & Freese show how conditional logit models can be used for alternative-specific data. If you read both Allison's and Long & Freese's discussion of the `clogit` command, you may find it hard to believe they are talking about the same command!

Example. Here is an example from Allison's 2009 book *Fixed Effects Regression Models*. Data are from the National Longitudinal Study of Youth (NLSY). The data set has 1151 teenage girls who were interviewed annually for 5 years beginning in 1979. The data have already been reshaped and xtset so they can be used for panel data analysis. That is, each of the 1151 cases has 5 different records, one for each year of the study. The variables are

- id is the subject id number and is the same across each wave of the survey
- year is the year the data were collected in. 1 = 1979, 2 = 1980, etc.
- pov 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 is coded 1 if the respondent currently has at least 1 child, 0 otherwise.
- spouse is coded 1 if the respondent is currently living with a spouse, 0 otherwise.
- school is coded 1 if the respondent is currently in school, 0 otherwise.
- hours is the hours worked during the week of the survey.

We can use either Stata's `clogit` command or the `xtlogit, fe` command to do a fixed effects logit analysis. Both give the same results. (In fact, I believe `xtlogit, fe` actually calls `clogit`.) First we will use `xtlogit` with the `fe` option.

```
. use https://www3.nd.edu/~rwilliam/statafiles/teenpovxt, clear
. xtlogit pov i.mother i.spouse i.school hours i.year, fe nolog
note: multiple positive outcomes within groups encountered.
note: 324 groups (1,620 obs) dropped because of all positive or
      all negative outcomes.

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

Here is how we interpret the results. The note "multiple positive outcomes within groups encountered" is a warning that you may need to check your data, because with some analyses there should be no more than one positive outcome. In the present case, that is not a problem, i.e. there is no reason that respondents cannot be in poverty at multiple points in time.

The note "324 groups (1620 obs) dropped because of all positive or all negative outcomes" means that 324 subjects were either in poverty during all 5 time periods or were not in poverty during all 5 time periods. Fixed-effects models are looking at the determinants of within-subject variability. If there is no variability within a subject, there is nothing to examine. Put another way, in the 827 groups that remained, sometime during the 5 year period the subject went from being in poverty to being out of poverty; or else switched from being out of poverty to being in poverty. If poverty status were something that hardly ever changed across time, or if very few people were ever in poverty, there would not be many cases left for a fixed effects analysis. Even as it is, more than a fourth of the sample has been dropped from the analysis. (Other techniques, like `xtreg, fe`, won't cost you so many cases.)

In terms of interpreting the coefficients, it may also be helpful to have the odds ratios.

```
. xtlogit, or

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 |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    1.mother |   1.790388   .2857157     3.65   0.000     1.309513    2.447848
    1.spouse |   .4734266   .0830137    -4.26   0.000     .3357355    .6675871
    1.school |    1.31241   .1479521     2.41   0.016     1.052231    1.636923
       hours |   .9805456   .0030891    -6.24   0.000     .9745098    .9866189
             |
        year |
          2  |   1.393447   .1415223     3.27   0.001     1.141931    1.700359
          3  |   1.397909   .1513231     3.09   0.002     1.130672    1.728308
          4  |   1.541515   .1796087     3.71   0.000      1.22679    1.936979
          5  |   1.495561   .1907255     3.16   0.002     1.164802    1.920242
------------------------------------------------------------------------------
```

The OR for mother is 1.79. This means that, if a girl switches from not having children to having children, her odds of being in poverty are multiplied by 1.79. Remember, these are teenagers at the start of the study, so having a baby while you are still very young is not good in terms of avoiding poverty. Conversely, if a girl switches from being unmarried to married, her odds of being in poverty get multiplied by .47, i.e. getting married helps you to stay out of poverty. Being in school multiplies the odds of poverty by 31 percent, while each additional hour you work reduces the odds of poverty by 2 percent. The year coefficients are all comparisons with year 1 and are all positive and significant; on an all other things equal basis, teens are more likely to be in poverty in the later years.

Notice that we did NOT include the time-invariant variables for age and black. Let's see what happens when we do.

```
. 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. [Rest of output deleted]
```

The two variables get dropped because their values do not vary within each group. Something that is a constant cannot explain variability in a dependent variable. (Allison, however, demonstrates that interactions between time-varying and time-constant variables can be included in the model.)

To do the same thing with `clogit`,

```
. use https://www3.nd.edu/~rwilliam/statafiles/teenpovxt, clear
. xtset, clear
. clogit pov i.mother i.spouse i.school hours i.year, group(id) nolog
note: multiple positive outcomes within groups encountered.
note: 324 groups (1,620 obs) dropped because of all positive or
      all negative outcomes.

Conditional (fixed-effects) logistic regression

                                                Number of obs     =      4,135
                                                LR chi2(8)        =      97.28
                                                Prob > chi2       =     0.0000
Log likelihood = -1520.1139                     Pseudo R2         =     0.0310

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

I did not need to clear the xtsettings; but I did so to illustrate that with `clogit`, it isn't necessary to `xtset` the data. Instead, the panelvar is specified by using the `group` option. Further, with neither method was the timevar actually needed. Instead of years, these could have been children within schools. The `xt` labeling of commands can be deceptive in that you do not necessarily need to have longitudinal data to use some of the commands.

WARNING!!! As I will explain later, marginal effects and adjusted predictions can often provide a great way to make the results from Categorical outcomes models more interpretable. But, Marginal effects and predicted values after `xtlogit, fe` and `clogit` can be problematic. By default, margins is giving you "the probability of a positive outcome assuming that the fixed effect is zero." This may be an unreasonable assumption. For a discussion of the problem and possible solutions, see Steve Samuels' comments at

http://www.statalist.org/forums/forum/general-stata-discussion/general/1304704-cannot-estimate-marginal-effect-after-xtlogit