Panel Data and Multilevel Models for Categorical Outcomes: Introduction

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

Overview. This course is going to focus on categorical outcomes (e.g. panel data and multilevel logistic regression models) but many of the same ideas will also apply to linear models. Because the results from categorical outcome models can often be difficult to interpret, I will also talk about how *adjusted predictions* and *marginal effects* can often make the substantive meaning of results clearer.

Many statistical analyses are done on samples of individuals (or institutions, or countries, etc.) that are sampled independently of each other with each case measured at only one point in time. For such data, methods like OLS regression, logistic regression, t-tests, and Poisson regression may be appropriate.

Other samples, hower, can be much more complex.

- With *panel data*, the same individuals or units are measured at multiple points in time. For example, the Wisconsin Longitudinal Study has periodically collected data from a sample of 1957 Wisconsin high school graduates in an effort to see what happens to them over their lifetimes. Economists sometimes collect information on the annual earnings of corporations, while political scientists might examine countries measured at multiple points over time.
- *Multilevel data* are collected from units organized or observed within units at a higher level (from which data are also obtained). For example, we might have data on students (level 1) who are clustered within classrooms (level 2). Or, we could have siblings clustered in families. In the United States, several studies have followed samples of students who started off in the same classrooms and then followed them as they moved through early grades into early adulthood.
- Panel data are actually a special type of multilevel data – records from multiple time points are clustered by individual.

Panel/ multilevel data offer special challenges. At a minimum the analysis must take into account that the records are not all independent of each other. An individual's response at time 1 will generally not be unrelated to his or her response at time 3. Forty students from the same school will share more in common than forty students from forty different schools. If, say, we had 200 individuals, each of whom was measured at 5 points in time, and we acted as though we had a sample of 1,000 independent cases, our standard errors would be too low and we would overstate the statistical significance of our results.

Getting the data set up correctly can also be a challenge. Data might be in wide format – one record for each case, with a caseid variable and different variables for each time point (e.g. inc1990, inc1992, inc1994). Such data will often have to be restructured to long format, with one record for each individual at each time point. The variables might then be caseid, year, and inc. Wide format will have fewer records but more variables. With long format, the software will have to know how the cases are connected, e.g. what the id variable is.

Beyond that, though, panel/multilevel data offer several unique analytical opportunities. A few that I will focus on include

- The effects of omitted variables can sometimes be controlled for. With *fixed effects models* and *conditional logit models*, individuals basically serve as their own controls. So, for example, if an important variable like gender or race is not included in the data set, you may be able to control for its effects anyway.
- With *multilevel models*, you can examine effects on a case's outcomes from each level, e.g. the parental Socio-Economic status of a child's parents and characteristics of the school that s/he attends. You can also examine interaction effects between levels, e.g. maybe the effect of parent's SES varies by the school attended. You will also often hear these referred to as *random-effects models*, *random-coefficients models*. *Mixed-effects models*, or *hierarchical linear models*.
- Sometimes we are interested more in the timing of events than whether or not the event occurs. For example, everybody dies sooner or later, but what causes some people to die more quickly than others? With the right kind of longitudinal data, you can use regular logistic regression for this. These are called *Discrete Time Methods for the Analysis of Event Histories*.

Course Outline

The first few handouts will focus on the analysis of panel data. With panel data, the same individuals (or countries, or businesses, etc.) are measured at multiple points in time. We will then transition into multilevel and (time permitting) event history models. As Hedeker notes, multilevel data are collected from units organized or observed within units at a higher level (from which data are also obtained). For example, we might have data on students (level 1) who are clustered within classrooms (level 2). Or, we could have siblings clustered in families. As we will see, the same techniques that are used for panel data can often be used with multilevel data, and vice-versa.

Course Web Page. The page may have additional materials not included here, including suggested readings, additional or revised handouts, and Stata do files used in these handouts.

https://www3.nd.edu/~rwilliam/Taiwan2018/index.html