Notes on Twisk(2003)

Jae-Woo Kim

Applied Longitudinal Data Analysis for Epidemiology, Cambridge Press.

Longitudinal analysis for binary outcome variables

- According to Twisk(2003), there are two approaches for panel data: GEE (generalized estimating equations) and Random coefficient analysis. Both of them are available in the recent version of Stata: xtgee and xtlogit.

GEE vs Random coefficient analysis

- The basic idea in GEE is to incorporate a new variable representing the within-subject correlation to handle the dependence of cases in panel data. It might be independent, exchangeable, m-dependent, autoregressive correlation, or unstructured correlation.
- To tackle the same problem in longitudinal data, the latter approach suggests that regression coefficients differ among subjects. Furthermore, it is able to allow intercepts to be different among subjects.

- “When an exchangeable correlation structure is not appropriate, GEE analysis with a different correlation structure can be used. When an exchangeable correlation structure is appropriate, and there is no random variation in one of the estimated regression coefficients (except the intercept), GEE analysis and random coefficient analysis are equally appropriate. When there is significant and relevant random variation in one (or more) of the regression coefficients, random coefficient analysis can be used, with the additional possibility of allowing other coefficients to vary between subjects” (Twisk, 2003: 91).

- Another issue is how to model the effect of time more generally, without assuming a linear function. One good idea is to employing a set of dummy variables. This approach is more acceptable as long as the number of repeated measurements is low compared to the number of subjects.
- However, the interpretation of coefficients is twofold: 1) between-subjects; 2) within-subjects. In other word, it is impossible to determine the contribution of each part in either GEE or random coefficient analysis although the overall error variance is divided into two parts, one from the intercept and the other from the slope.

- It is also a good idea to consider both time-lag effects and autoregression. Time-lag models are reasonable when the time periods between subsequent measurements are short. Autoregressive models are useful in the sense that they allow to control the previous value of Y. Put together, I will take into account the time-lagged effect of X_{t-1} on Y_t while controlling for the effect of Y_{t-1} on Y_t. (However, Twisk points out it would be much more complicated to incorporate either of them in the model with binary outcomes because changes in a dichotomous variable between two time-points lead to a categorical variables with four groups, but without more detailed discussion. See p.143).

Longitudinal logistic regression

- It does not seem like difficult to apply both of approaches to the data with binary outcome variables.
- However, Twisk does not seem to be clear about how to define the type of correlation structure for binary outcomes. “It is not possible to use the correlation structure of the observed data as a guide for the choice of ‘working correlation structure’ (Ibid: 131). For example, what does the exchangeable correlation structure look like?

- Another issue is that the regression coefficients from GEE analysis are ‘population averaged,’ while they from random coefficient analysis are ‘subject specific’ (Ibid: 140-1). This is the main reason the two approaches produce different results.
- What Twisk suggests: “If one is interested in the relationship between a dichotomous outcome variable and several other predictor variables, GEE analysis will probably provide the most ‘valid’ results. However, if one is interested in the individual development over time of a dichotomous outcome variable, random coefficient analysis will probably provide the most ‘valid’ results” (Ibid: 142).