Longitudinal Data Analysis Exercises

Joschka Hüllmann

11 Januar 2019

Exercise One

The goal of this manuscript is to reproduce Marinov's analysis and compare its underlying model to other model choices. In particular, I reproduce the exact model and compare it to other conditional unit effects models, i.e. a random effects model and **???**, and to **[number]** general equation estimation (GEE) models.

Marinov's article explores the effect of various covariates, e.g. the presence of international sanctions, and economic wealth and growth, on the a leader's ability to retain power. He bases his analysis on a data set aggregated from various sources. The data set includes 160 countries from the year 1947 to 1999 ($N = 160; T = 52$; **check actual data**). The dependent variable Y is binary with the leader either surviving in office ($Y = 0$) or giving up his position ($Y = 1$). The author is interested in what covariates reduce the time to event of a leader giving up his position. Although this suggests a a survival model, Marinov, following Beck, Katz and Tucker (1998), argues that a logit model can be equivalent to a cox survival model (c.f. slides 4-pm, p. 42 **put in equation appendix**). Besides three cubic splines to capture the three phases of a leader's stability early, middle and late phase of his leadership, as well as the time t in office, interacting with regime type, no explicit time variance on the other covariates is assumed. Not looking into any dynamic effects, the author models unit effects conditional on the countries, i.e. assuming institutional or other effects inherent to each country.

In summary, the assumptions of the author for the data generating process are as follows: - panel type data with $N >> T$ - binary outcome $Y = 1$ or $Y = 0$ - no time variance on covariates other than years as a covariate interacting with regime dummies - three cubic splines to deal with differences in the three phases of leadership. - conditional unit effects on the countries

Strictly following the assumptions, it is clear, that a conditional unit effects model is to be specified with binary outcome, thus suggesting a logit model conditional on the country. The authors choose a fixed effect logit model over random effects, which I guess is due to the panel data $N >> T$ (c.f. slides 1-am, p. 51). The specified model shoud look like this, f being the *logit* function and α_i being the country-level intercept:

$$Y_{it} = f(X_{it}\beta + \alpha_i + u_{it})$$

Alternatively to the proposed model by the authors, we can specify a naive estimator without unit-level effects, which would remove the α_i from the equation. The variance would end up in the coefficients and the error. As a result, not considering existing unit-level effects may bias the estimates and lead to an inconsistent estimator (**TODO: Not sure if correct, look up**). Conversely, we can model random effects, if we assume that the unit level effects are uncorrelated to the covariates. In this case, the α_i would be a second error term instead of fixed effect, so the error looks like $u_{it} = \alpha_i + \eta_{it}$. If the independence assumptions holds, it is more efficient than the fixed effects model. However, if the assumption does not hold, it leads to an inconsistent model. Analytically, this can be tested with the Hausman test. For completeness, the fixed effects model could also be specified using an LSDV model ("brute force"). With $N = 160; T = 52$, the asymptotics should be fine for a fixed effects model, so we do not encounter the incidental parameter problem (**TODO: check this, look up**).

Contrary to a conditional unit effect model, we can change our assumptions about the influence and correlation between the outcomes and covariates, and in particular the dependence between Y and time t, and specify a marginal population average model, i.e. generalised equation estimation model (GEE). At the core of a GEE is the working correlation matrix, which describes the structure of covariance (correlation?) and is specified

by the analyst. In general the formula as follows:

$$U(\beta) = \sum_{i=1}^{N} D_i' V_i^{-1} [Y_i - \mu_i] = 0$$

with the variance (after the standard GLM assumption) decomposed into:

$$V_i = \frac{(A_i^{1/2}) R_i(\alpha) (A_i^{1/2})}{\phi}$$

Typical specifications for the working correlation matrix ϕ include: independent, exchangeable, autoregressive, stationary, unstructured (empirical).

TODO:

- rework first paragraph about conditional unit effect models and fix todos
- more detail in GEEs and how underlying assumptions change, go into detail for each working correlation matrix
- provide calculations for all of the given models

Open Questions and possible TODO:

- Is there a relationship between dynamic panel data and GEE?
- transform into survival model