BIOS 755: Fixed versus Random effect models

Alexander McLain

February 6, 2023

Introduction

- Today we are going to be talking about time-varying and time-invariant covariates.
- Let $\boldsymbol{X}_{ij}$ denote the time-varying covariates and $\boldsymbol{W}_{ij} = \boldsymbol{W}_i$ the time-invariant covariates.
- To analyze such data we could use

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{W}_i\gamma + \alpha_i + e_{ij}$$

where $\boldsymbol{e}_i \sim N(0, \sigma^2)$.

Fixed effect model

- What if we didn't assume α_i was random, but rather estimated it from the data. What kind of model would we have?
- An issue with this model is that we couldn't estimate the time-invariant covariate effects:
  - **can't estimate both γ and the α_i's.**

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{W}_i\gamma + \alpha_i + e_{ij}$$

Fixed effect model

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \boldsymbol{W}_i\gamma + \alpha_i + e_{ij}$$

- We could estimate $\boldsymbol{\beta}$
- For example, if we only had two observations:

$$(Y_{i2} - Y_{i1}) = (\boldsymbol{X}_{i2} - \boldsymbol{X}_{i1})'\boldsymbol{\beta} + e^*_{i2}$$

  which could be fitted using OLS of $(Y_{i2} - Y_{i1})$ on $(\boldsymbol{X}_{i2} - \boldsymbol{X}_{i1})$.
- Notice that this model removes the potential for bias due to confounding by all measured and unmeasured time-invariant characteristics of individuals (as long as the effect is constant over time). Wow.

Fixed effect model

- This approach can be expanded with the mean-centered model

$$Y_{ij}^{*} = \boldsymbol{X}_{ij}^{*}\boldsymbol{\beta} + e_{ij}$$

where $Y_{ij}^{*} = Y_{ij} - \bar{Y}_{i}$ and $\boldsymbol{X}_{ij}^{*} = \boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_{i}$.
- Or with the *first difference* model for $j = 2, \ldots, n_i$

$$Y_{ij}^{\dagger} = \boldsymbol{X}_{ij}^{\dagger}\boldsymbol{\beta}^{\dagger} + e_{ij}^{\dagger}$$

where $Y_{ij}^{\dagger} = Y_{ij} - Y_{i1}$ and $\boldsymbol{X}_{ij}^{\dagger} = \boldsymbol{X}_{ij} - \boldsymbol{X}_{i1}$.

Fixed effect model

- This model removes all the variation due to time-invariant covariates.
- If the random intercept model is correct, the correlation e_{ij} can be ignored
  - assuming only time-invariant covariates are causing the dependence between measurements.
- To fit this model we'll use `proc glm` with an independent correlation matrix.

GO TO EXAMPLE

Random effect versus Fixed effect model

Some technical differences:

- The fixed effect model allows for correlation between α_i and $\boldsymbol{X}_{ij}$, and α_i and $\boldsymbol{W}_i$.
- The random effect model does not allow for this correlation and biases in $\boldsymbol{\beta}$ arise when violated.
- The random effect model allows for estimation of time-invariant fixed-effects, the fixed effect model does not.
- The random effect model is more efficient than the fixed effect model.
- Individuals must have more than 1 observation in the fixed effect model.

Longitudinal vs Cross-sectional effects

Introduction

- It is possible to allow for **longitudinal** and **cross-sectional** effects in longitudinal analyses.
- Such an approach acknowledges the two distinct sources of variation in a covariate:
  - one based on within-subject variation, and
  - one based on between-subject variation.
- Such a model recognized that longitudinal data provide information about
  - how individuals differ at any one occasion, and
  - how an individuals response varies over time.
- Commonly these effects are erroneously combined.

Longitudinal and cross-sectional information

- Assessment of within-subject changes in the response due to aging (for example) can only be achieved within a longitudinal study design.
- What would happen with a cross-sectional design?
- Recall the Muscatine Coronary Risk Factor (MCRF) study which had five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15.
- Goal: determine whether the risk for obesity increased with age.
- Measurements were taken in 1977, 1979 and 1981.
- Could we measure the effect of age using only data from 1977?

Model

- To combine longitudinal and cross-sectional effects we will include both effects in the model.
- For example, we can use the linear mixed effects model:

$$Y_{ij} = \boldsymbol{X}^*_{ij}\beta^{(L)} + \boldsymbol{X}'_{i1}\beta^{(C)} + \boldsymbol{W}'_{i}\gamma + \boldsymbol{Z}'_{ij}b_i + e_{ij}$$

  where $\boldsymbol{X}^*_{ij} = \boldsymbol{X}_{ij} - \boldsymbol{X}_{i1}$.
- Here, $\beta^{(C)}$ represent the cross-section effects while $\beta^{(L)}$ are the longitudinal effects.
- This is one example, another example is using $\boldsymbol{X}^*_{ij} = \boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_{i}$ and $\bar{\boldsymbol{X}}_{i}$

Example

- Let A_{ij} be the age of person i at measurement j. One option is to fit the model

$$Y_{ij} = \beta_0 + \beta_1 A_{ij} + b_{0i} + \epsilon_{ij}$$

- Separating out the cross-sectional and longitudinal effects of age we have

$$Y_{ij} = \beta_0^* + \beta_1^{(L)} A_{ij}^* + \beta_1^{(C)} A_{i1} + b_{0i}^* + \varepsilon_{ij}$$

  where $A_{ij}^* = A_{ij} - A_{i1}$ is the change in age from baseline.
- When there is not a well defined baseline measurement, I prefer to the average model.

Interpretation

- For the baseline measurement, it's straightforward to show that

$$E(Y_{i1}) = \boldsymbol{X}'_{i1}\beta^{(C)} + \boldsymbol{W}'_i\gamma$$

so $\beta^{(C)}$ is the expected difference in the average baseline outcome for a 1 unit change in $\bar{\boldsymbol{X}}_i$

- Further, the model for the within-subject changes is

$$\begin{aligned} E(Y_{ij} - Y_{i1}) &= \boldsymbol{X}^*_{ij}\beta^{(L)} + \boldsymbol{X}'_{i1}\beta^{(C)} + \boldsymbol{W}'_i\gamma - \left(\boldsymbol{X}^*_{i1}\beta^{(L)} + \boldsymbol{X}'_{i1}\beta^{(C)} + \boldsymbol{W}'_i\gamma\right) \\ &= (\boldsymbol{X}^*_{ij} - \boldsymbol{X}^*_{i1})'\beta^{(L)} \end{aligned}$$

so $\beta^{(L)}$ is the expected within person difference in the outcome for a 1 unit change in $\boldsymbol{X}^*_{ij}$

Example

- Separating out the cross-sectional and longitudinal effects of age we have

$$Y_{ij} = \beta_0^* + \beta_1^{(L)} A_{ij}^* + \beta_1^{(C)} \bar{A}_i + b_{0i}^* + \varepsilon_{ij}$$

where $A_{ij}^* = A_{ij} - \bar{A}_i$ and $\bar{A}_i$ is the persons average age in the study.
- For subjects with a baseline age that was 1 unit higher, we expect their baseline outcome to be $\beta_1^{(C)}$ units larger.
- During our study, individuals aging 1 year from baseline is associated with a $\beta_1^{(L)}$ unit increase in the outcome.

$\beta^{(L)}$ versus $\beta^{(C)}$

- Notice that when $\beta^{(L)} = \beta^{(C)} = \beta$

$$\begin{aligned} Y_{ij} &= \boldsymbol{X}^{*}_{ij}\beta^{(L)} + \boldsymbol{X}'_{i1}\beta^{(C)} + \boldsymbol{W}'_{i}\gamma + \boldsymbol{Z}'_{ij}b_i + e_{ij} \\ &= \boldsymbol{X}'_{ij}\beta + \boldsymbol{W}'_{i}\gamma + \boldsymbol{Z}'_{ij}b_i + e_{ij} \end{aligned}$$

  which is the "standard" mixed effects model
- So the hypothesis test $H_0 : \beta^{(L)} = \beta^{(C)}$ tests whether longitudinal and cross section effects are equal.

Longitudinal and cross-sectional information

- Could we measure the effect of age using only data from 1977?
- Differences in $\beta^{(L)}$ versus $\beta^{(C)}$ can arise when there are cohort or period effects.
- When we assume $\beta^{(L)} = \beta^{(C)} = \beta$, i.e., use the "standard" linear mixed effects model, we get an estimate for β that is a combination of $\beta^{(L)}$ and $\beta^{(C)}$.