Linear Mixed Effects Models II

The simplest mixed effect model is a random intercept model where $Z_i = 1$;

$$Y_i = X_i\beta + Z_i b_i + e_i$$

The random intercept model can be interpreted as the effect of all unobserved subject-specific variables (b_i) on the linear predictor.

Random slopes of time-varying covariates (δ) can be interpreted as interaction of unobserved subject specific covariates with observed time-varying covariates.

We can also include a random effect with a time-invariate covariate b_i (e.g. treatment group) to produce a heteroscediastic random intercept $Z_{i0} + Z_{i1}b_i$.

Two-Stage Formulation

An alternative formulation of the mixed effects models is a two-stage formulation, in which we first specify a within subject or level-1 model where the occasion specific linear predictor is a function of time-varying covariates. Then a between subject or level 2 model is specified:

$$\begin{aligned} y_{ij} &= X_{ij}\boldsymbol{\delta}_i \qquad \text{Level 1} \\ \boldsymbol{\delta}_i &= \Gamma W_i + b_i \qquad \text{Level 2} \end{aligned}$$

Some elements of δ_i are constant and some depend only on the observed covariates.

Consider the following with a random intercept and a random slope for time:

$$y_{ij} = \delta_{i0} + \delta_{i1} t_{ij} + \delta_{i2} x_{ij}$$

x_{ii} is a time-varying covariate and w_i is a time-invariant covariate

$$\begin{aligned} \delta_{i0} &= \gamma_{00} + \gamma_{01} w_i + b_{i0} \\ \delta_{i1} &= \gamma_{10} + \gamma_{11} w_i + b_{i1} \\ \delta_{i2} &= \gamma_{20} \end{aligned}$$

The redirected form is obtained by substituting the level-2 model into level-1 model. We then substitute level-2 into level-1 to produce the reduced form with a "cross-level interaction" term $\gamma_{11} w_i t_{ij}$

- The two formulations are equivalent but it can have an impact on the types of models being considered
- The two-stage formulation encourages inclusion of many cross-level interactions and few same-level interactions
- Due to an abundance of interactions in the two-stage formulation, we usually center around the mean all variables except time. In our example, centering of w_i makes γ_{10} interpretable as the mean effect of t_{ij} when t_{ij} when w_i takes it mean value.

Unobserved Confounders

In longitudinal studies, it is often said the "subjects serve as their own controls" when considering time-varying covariates. This seems to imply that all subject-level observed and unobserved covariates have been controlled for, but this is NOT true since this omitted covariates may correlate and hence be confounded with the time-varying variables of interest. If Cov(x_{ij}, u_{ij}) != 0, then we will have omitted variable bias; We say that x_{ij} is *endogenous* or correlated with the random intercept b_{i0}

Centering

Cov(x_{ij}, u_{ij}) = 0 if x_hat_i = x_hat_j for any i,j; We can then avoid omitted variable bias by subject-mean centering x_{ij}, forming the instrumental variable x_{ij}^{d} = x_{ij} - x_hat_i and running the following model:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1(x_{ij} - \bar{x}_i) + \beta_2\bar{x}_i + b_{i0} + e_{ij} \\ &= \beta_0 + \beta_1 x_{ij} + (\beta_2 - \beta_1)\bar{x}_i + b_{i0} + e_{ij} \end{aligned}$$

β_1 can be interpreted as a purely within-subject (or longitudinal) effect and β_2 as between-subject (or cross-sectional) effect. A test of the null hypothesis $\beta_1 = \beta_2$ is equivalent to the Durbin-Wu-Hausman test for exogeneity.

Linear Fixed Effects Models

Fixed effects linear models are formulated as:

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

where X denotes a q*1 vector of time-varying covariates, Wi denotes a (p-q)*1 vector of time-invariant covariates and α_i are fixed effects representing the time-invariant unobserved confounders, and e_{ij} remains the random within-subject errors.

This accounts for all observed/unobserved time-invariate confounders*, making this a model likely less prone to bias with the obvious drawback of way more terms to estimate (loss of power/degrees of freedom)
*assuming that the effects on the response remain constant over time

Although this looks very similar to the random intercepts model considered last class, note that in the mixed effects formulation α_i are considered *random* while now they are considered *fixed*.

Properties of Fixed Effect Models

- X_{ij} is assumed to be strictly exogenous; i.e. current values of the response Y_{ij} given X_{ij} do not predict the subsequent value of X_{ij+1}
- The fixed effects α_i can be correlated with X_{ij} and W_i unlike mixed effects where α_i are assumed independent of X_{ij} and W_i
- Fixed effects models cannot estimate the effects of time-invariant covariates. Since α_i and $W'_{i\gamma}$ are perfectly colinear, the time-invariant covariates are effectively removed from the analysis
- Fixed effects models remove bias when there are unmeasured but stable characteristics of the subjects that are correlated with time-varying covariates of main scientific interest

Bias-Variance Trade-Off

Although, under the conditions we discussed the fixed effects model may provide unbiased estimates of the time-varying covariates, they will generally have larger standard errors for those estimated effects than those produced by mixed effects models

Fixed effects base estimation exclusively on the within-subject variation and ignore any between-subject variation, while mixed effects utilize both sources of variability resulting in smaller standard errors.

The greater the proportion of between subject variation in a time-varying covariate, the larger the differences in the magnitudes of the standard errors between mixed and fixed effects models.

Comparison of Fixed and Mixed Effects Estimates

Consider the simple random effect model:

$$Y_{ij} = \beta_1 + \beta_2 X_{ij} + b_i + e_{ij}$$

It can be shown that the random effects estimator β_2^{RE} can be written as:

$$\hat{\beta}_2^{RE} \approx (1 - w)\hat{\beta}_2^{FE} + w\hat{\beta}_2^{B}$$

Where β_2^{FE} is the fixed effects estimator and β_2^{B} is an estimator based only on between-subject variation and is obtained from the regression of $\bar{Y}_i$ on $\bar{X}_i$. w is the weight given by:

$$w = \frac{(1 - \rho_y)\rho_x}{(1 - \rho_y) + n\rho_y(1 - \rho_x)}$$

Where $\rho_y = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}$ is the proportion of variability in the response due to between subject variation and ρ_x is the corresponding proportion of variability of X_{ij} that is due to between-subject variation.

- When within-subject variation in the response is small then ρ_y -> 1 which results to w -> 0 and $\beta_2^{RE} = \beta_2^{FE}$

- When within-subject variation in the covariate is large then ρ_x -> 1 which results to w -> 0 and $\beta_2^{RE} = \beta_2^{FE}$

Residual Analyses and Diagnostics

We define a vector of residuals for each individual as $r_i = Y_i - X_i\beta$

From there a simple scatter plot of the residuals against the predicted mean response or covariates can be observed to see if there are is systemic pattern. The residuals are correlated and usually do not have constant variance.

The standardized residuals are defined as: $r^*_i = L_i^{-1} r_i$ where L_i is a lower triangular matrix so that:

$$\hat{\Sigma}_i = L_i L_i'$$

Effectively a L_i is a 'square root' [but for matrices] of the var/covar matrix.

r_i *are uncorrelated and have unit variance*. r^*_{ik} is an estimate of:

$$\frac{Y_{ik} - \mathrm{E}(Y_{ik} | Y_{i1}, \ldots, Y_{ik-1})}{\sqrt{\mathrm{Var}(Y_{ik} | Y_{i1}, \ldots, Y_{ik-1})}}$$

We can use the standardized residuals to detect outlying observations:

- Scatterplot of standardized residuals against teh predicted means response
- Scatterplot of standardized residuals against covariates
- Normal QQ plot to assess the assumption of normallity
- Compute Mahalanobis distance to detect *outlying individuals:*

  $$d_i = r_i^{*\prime} r_i^*$$

  d_i ~ chi-square distributed with degrees of freedom equal to the number of measurements for individual i

Semi-Variogram

We use the semi-variogram to assess the adequacy of a selected model for the covariance.

$$\gamma(h_{ijk}) = \frac{1}{2}\mathrm{E}(r_{ij} - r_{ik})^2$$

Where h_{ijk} = the time elapsed between repeated measurement j and k on the ith individual. Since $E(r_{ii}) = 0$ then:

$$\begin{aligned}
\gamma(h_{ijk}) &= \frac{1}{2}\mathrm{E}(r_{ij} - r_{ik})^2 \\
&= \frac{1}{2}\mathrm{E}(r_{ij}^2 + r_{ik}^2 - 2 r_{ij} r_{ik}) \\
&= \frac{1}{2}\mathrm{Var}(r_{ij}) + \frac{1}{2}\mathrm{Var}(r_{ik}) - \mathrm{Cov}(r_{ij}, r_{ik})
\end{aligned}$$

So since each r* has a variance of 1 and covariance of 0, the semi-variogram for r^*_{ij} is:

$$\gamma(h_{ijk}) = \frac{1}{2}\mathrm{Var}(r^*_{ij}) + \frac{1}{2}\mathrm{Var}(r^*_{ik}) - \mathrm{Cov}(r^*_{ij}, r^*_{ik}) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0$$

Thus, in a correctly specified model for the covariance the plot of the semi-variogram for the standardized residuals vs. time should fluctuate randomly around a horizontal line centered at 1.

Model Selection Strategy

We can use the following guidelines for selecting a model for correlated data:

1. Fit a saturated model such as profile analysis
2. For a full model select an appropriate covariance structure using REML. You can use likelihood ratio when appropriate, or AIC/BIC when LRT is not possible
3. For an appropriate covariance structure, select an appropriate set of covariates using ML. As previously seen, use LRT, AIC or BIC
4. For an appropriate set of covariates, select an appropriate covariance structure using REML
5. Repeat steps 3 and 4

Note that REML is used to compare covariance structures to ensure unbiased estimators, ML is used to compare different combinations of covariates because it's more powerful.

SAS Code

```
/***********************************************
* Six cities study of air pollution and health  *
***********************************************/
data fev1;
set s857.fev1;
lgfevht=Log_FEV1_-2*logHgt;
run;

*Durbin-Wu-Hausman test;
*Computes mean age by ID, spits it into another dataset;
proc means data=fev1 ;
*where id ne 197;
by id;
var age ;
output out=mean_age;
run;
```

```
data mean_age;
set mean_age;
if _STAT_='MEAN' then;else delete;
rename age=age_mean;
keep id age;
run;
data fev1;
merge fev1 mean_age;
by id;
run;
data fev1;
set fev1;
*Creating mean-centered age;
*Prevents confounding with random slope;
age_dev=age-age_mean;
run;
proc means data=fev1 mean var;
var age age_dev;
run;
proc means data=mean_age;
var age_mean;
run;

*RE model;
proc mixed data=fev1 method=ml covtest;
*where id ne 197;
class id;
model lgfevht=age age_mean/s chisq;
random intercept/type=un subject=id G V ;
run;
*Fixed Effects model;
proc glm data=fev1;
*where id ne 197;
class id;
model lgfevht=id age/solution;
run;
quit;
proc mixed data=fev1 method=ml covtest;
*where id ne 197;
class id;
model lgfevht=age /s chisq;
```

```
random intercept/type=un subject=id G V;
run;

*To compute rho of Y (the proportion due to within-subject variation in the repsonse;
*Sigma2(b) is UN(1, 1) = 0.01144, Sigma2(e) is Residual = 0.01391;
proc mixed data=fev1 method=ml covtest;
*where id ne 197;
class id;
model lgfevht=/s chisq;
random intercept/type=un subject=id G V ;
run;
proc mixed data=fev1 method=ml covtest;
*where id ne 197;
class id;
model age=/s chisq;
random intercept/type=un subject=id G V ;
run;
/***********************************************
Study of influence of menarche on changes in body fat
***********************************************/

proc sgplot data=s857.fat  noautolegend ;
   * spaghetti plot;
   yaxis min = 0 max = 50;
   reg x=time_men y=Perc_BF
    / group = id nomarkers LINEATTRS = (COLOR= gray PATTERN = 1 THICKNESS = 1) ;
   * overall spline;
   reg x=time_men y=Perc_BF
    / nomarkers LINEATTRS = (COLOR= red PATTERN = 1 THICKNESS = 3) ;
run;
quit;

ods graphics on;

proc loess data=s857.fat plots=all;
   model Perc_BF = time_men;
run;
ods graphics off;

data fat;
set s857.fat;
```

```
knot=max(time_men,0);
run;
ods graphics;
proc mixed data=fat covtest plots=all;
model Perc_BF = time_men knot/s chisq outpred=yhat outpm = pred1f residual;
random intercept time_men knot/type=un subject=id G V ;
run;

/*****************************************
*     Mahalanobis Distance          *
*****************************************/
data pred1f;
set pred1f;
sqStudentResid=StudentResid**2;
visits=1;
run;
proc means data=pred1f noprint;
*sum command gives sums in order given (sum of sqStudentResid is output as distance, visits as novisits);
var sqStudentResid visits;
output out=mahalanobis (drop=_type_ _freq_)
sum(sqStudentResid visits)=distance novisits;
by id;
run;
data mahalanobis;
set mahalanobis;
pvalue=1-probchi(distance,novisits);
run;
proc print data=mahalanobis;
*Bonferroni correction to p value based on number of tests;
where pvalue<0.05/162;
run;
/*****************************************/

proc means data=fat;
var time_men;
run;
proc variogram data=pred1f outv=outv noprint;
*Computing measurement differences within subjects with measurement differences (i.e., when they were
taken) of up to 10 with iterative values of 1;
*first plot (fit plot) shows average within-subject residuals for different lags. Should be randomly distributed
around 0, but def shouldn't have a pattern;
```

```
compute lagd=1 maxlag=10;
coord xc=time_men yc=visits;
by id;
var StudentResid;
run;
ods graphics on;

proc loess data=outv plots=all;
   model variog = distance;
run;
ods graphics off;



/*ALternative model*/
data fat;
set fat;
t=time_men;

run;
*Outpred gives FE (population-level) predictions, and outpm gives subject-level predictions;
proc mixed data=fat covtest plots=all;
class t;
model Perc_BF = time_men knot/s chisq outpred=yhat2 outpm = pred2f residual;
random intercept time_men knot/type=un subject=id G V;
*The added repeated statement assumes the error terms don't have constant variances and is looking for that
structure;
repeated t/type=sp(exp)(t) subject=id;
run;
data pred2f;
set pred2f;
sqStudentResid=StudentResid**2;
visits=1;
run;
proc variogram data=pred2f outv=outv2 noprint;
compute lagd=1 maxlag=10;
coord xc=time_men yc=visits;
by id;
var StudentResid;
run;
ods graphics on;
```

```
proc loess data=outv2 plots=all;
   model variog = distance;
run;
ods graphics off;
```

Revision #5
Created 16 February 2023 19:14:06 by Elkip
Updated 19 February 2023 21:42:22 by Elkip