HW6 (Due May 16)

1. Reproduce S359-364 and comment.

Slide 395 is the code, specifying random intercept and slope for the time effect

Slide 360:

matrix

Random effects variance covariance matrix

(Intercept) day

(Intercept) 9.547100 0.053333

day 0.053333 0.026657

The within subject error or residual variance:

Residual 0.8283737

The variance is this quantity squared: 0.686203

From this we see that most of the unexplained variability is explained by the random (unaccounted for) variability between subjects.

Fit statistics:

AIC BIC logLik

647.98 673.0192 -315.99

AIC and BIC are slightly different than the ones on the slide. An R/SAS thing, I suppose.

Slide 361 is the solution for fixed effects:

Fixed effects: strength ~ group.f \* day

Value Std.Error DF t-value p-value

(Intercept) 81.23965 0.6909881 134 117.57026 0.0000

group.f1 -1.23486 1.0500206 35 -1.17603 0.2475

day 0.17292 0.0426646 134 4.05291 0.0001

group.f1:day -0.03774 0.0636998 134 -0.59254 0.5545

Here we see that group differences are not significant at baseline, and that the group seems to have no effect in strength gain (group by time interaction is not significant).

Slide 362 are the formulas for obtaining Cov(Yi), the marginal covariance of Yi

Slide 363 are the corresponding matrices Cov(Yi) and Corr(Yi)

Cov(Yi)

Marginal variance covariance matrix

1 2 3 4 5

1 10.2330 9.7604 9.8671 9.9738 10.187

2 9.7604 11.0860 10.7200 11.0400 11.680

3 9.8671 10.7200 11.8330 11.5730 12.426

4 9.9738 11.0400 11.5730 12.7930 13.173

5 10.1870 11.6800 12.4260 13.1730 15.352

Here we see that, as for most longitudinal studies, Var(Yi) is the lowest at baseline and then increases as time goes by.

Corr(Yi)

1 2 3 4 5

1 1.0000000 0.9163564 0.8966754 0.8717077 0.8127599

2 0.9163564 1.0000000 0.9359646 0.9270353 0.8952902

3 0.8966754 0.9359646 1.0000000 0.9406538 0.9219719

4 0.8717077 0.9270353 0.9406538 1.0000000 0.9399771

5 0.8127599 0.8952902 0.9219719 0.9399771 1.0000000

Here we see that the correlation estimated through the random effect structure of the model decreases as time separation between observations increases.

Slide 364 is the interpretation of the fixed effects, which for the group effect I think are backwards, as the reference group throughout the analyses is group 2. The substantive part though (group X time interaction) remains valid, that the rates of increase of strength in both groups is not significantly different.

1. Reproduce S408-411 and comment.

Slide 408 is specifying the model’s code in SAS (requesting the solution for random effects “S”).

Slide 409 are the empirical Baye’s estimates of bi’s:

(Intercept) day

1 -1.01112331 -0.0318196576

2 3.37720596 0.1604247697

3 1.28109893 -0.0193501567

4 1.02810647 -0.0768505069

5 0.16950818 0.2365153105

6 -3.89366129 -0.1784566425

7 1.55135158 0.1692003024

8 -2.26901883 0.1484178009

9 5.44718316 0.0376307554

10 -4.07251633 0.1198988918

.

.

.

R does not produce standard errors or inferences about the random effects. I found in some forums that this is because the sampling distribution of random effects/variance estimates is usually strongly asymmetric. Douglas Bates says “Technically, the random effects don't have standard errors because they are not parameters in the model”. Now, if this is truly the case, I wouldn’t know.

Slide 410 has the code for estimating subject specific response profile

Slide 411: Predicted response profiles

id group time strength predicted residual

1 1 1 0 79 78.99367 0.006334439

3 1 1 2 79 79.40707 -0.407071058

4 1 1 3 80 79.61377 0.386226193

5 1 1 4 80 79.82048 0.179523445

7 1 1 6 80 80.23388 -0.233882052

8 2 1 0 83 83.38199 -0.381994826

10 2 1 2 85 84.56438 0.435621968

11 2 1 3 85 85.15557 -0.155569636

12 2 1 4 86 85.74676 0.253238761

14 2 1 6 87 86.92914 0.070855555

.

.

.

Was not able find out how to get SE’s of predictions. Perhaps when in need I will. But are they really necessary?

1. Reproduce Table 39 and 40 and comment.

Table 39:

Value Std.Error DF t-value p-value

(Intercept) 21.361383 0.5645443 885 37.83828 0.0000

time 0.417113 0.1571548 885 2.65415 0.0081

time0 2.047138 0.2279646 885 8.98007 0.0000

These are the solutions for the regression coefficients (fixed effects). Here we see that before menarche increase in % body fat is very low (0.4% increase per year), while after menarche, the increase in % body fat is almost 5 times more (2% increase per year).

Table 40 is the solution for the random effects. What I could get is this:

Random effects variance covariance matrix

(Intercept) time time0

(Intercept) 45.9390 2.5260 -6.1091

time 2.5260 1.6310 -1.7504

time0 -6.1091 -1.7504 2.7494

Standard Deviations: 6.7778 1.2771 1.6581

Which is the covariance matrix for the random effects, essentially contains the same, except for the SE’s of the variances and covariances of the random effects.

Again, I was not able to find a way to get the standard errors of the elements of this matrix in R. They say though “the sampling distribution of variance estimates is in general strongly asymmetric: the standard error may be a poor characterization of the uncertainty” (Douglas Bates)

This problem could be overcome using and MCMC approach for estimating the model, and then get 95% HPD interval for the parameters of the covariance matrix.

1. Fit a mixed effects model for log(FEV1) with random intercept, while decomposing between- and within-subject effects. What are key points and how to interpret results? [Hint: the codes are in Ch9 in authors’ website.]

Since I rather know what I’m doing instead of just running the code, I had to read the whole chapter 9 to get to the point of answering problem 4 and 6. What this model is doing is decomposing the age effect into between subjects effect (the effect of mean age for each subject) and within subject effects (the effect mean-centered age at each measuring occasion within subjects).

Here are the model results:

Value Std.Error DF t-value p-value

(Intercept) -0.3483213 0.03517432 1693 -9.90272 0

mage 0.0292337 0.00290142 297 10.07565 0

cage 0.0298226 0.00047964 1693 62.17656 0

From the model results, we can see that they’re virtually the same. So, when testing the hypothesis βLongitudinal = βCross-sectional via a contrast we get the following results:

Linear Hypotheses:

Estimate Std. Error z value Pr(>|z|)

1 == 0 0.0005889 0.0029408 0.2 0.841

Which show that the two estimates are not significantly different from each other, meaning there’s no evidence of any conflict between the longitudinal and cross-sectional information. This provides justification for the combined estimate of the age effect obtained through a mixed effects model where the covariate is age. The estimate of the effect of age in such a model is a weighted average of both cross-sectional and longitudinal effects presented before.

1. Run the following codes with “dental” dataset and describe what the “noint” option does and any usefulness?

The noint option runs models without intercept. It is useful if you don’t want to assume (or makes no sense so assume) that one of the covariates reach a value of zero. For the dental data example, the interpretation of the intercept makes no sense (i.e. the value of distance when all other covariates equal zero), specifically for age. I suppose that if you’re unhappy with this you can always center the value of age, instead running a no intercept model.

With such a model you can also get direct estimates for all levels of categorical variables, although estimating them from a model with an intercept is not such a big hassle either.

proc mixed method=ml data=dent1;

class gender child;

model distance = gender gender\*age / noint solution ;

repeated / type = cs subject = child r rcorr; run;

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Solution for Fixed Effects** | | | | | | |
| **Effect** | **gender** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** |
| **gender** | **F** | 17.3727 | 1.1615 | 25 | 14.96 | <.0001 |
| **gender** | **M** | 16.3406 | 0.9631 | 25 | 16.97 | <.0001 |
| **age\*gender** | **F** | 0.4795 | 0.09231 | 79 | 5.20 | <.0001 |
| **age\*gender** | **M** | 0.7844 | 0.07654 | 79 | 10.25 | <.0001 |

proc mixed method=ml data=dent1;

class gender child;

model distance = gender age gender\*age / solution chisq;

repeated / type = cs subject = child r rcorr; run;

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Solution for Fixed Effects** | | | | | | |
| **Effect** | **gender** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > |t|** |
| **Intercept** |  | 16.3406 | 0.9631 | 25 | 16.97 | <.0001 |
| **gender** | **F** | 1.0321 | 1.5089 | 25 | 0.68 | 0.5003 |
| **gender** | **M** | 0 | . | . | . | . |
| **age\*gender** | **F** | 0.4795 | 0.09231 | 79 | 5.20 | <.0001 |
| **age\*gender** | **M** | 0.7844 | 0.07654 | 79 | 10.25 | <.0001 |

1. There is a popular model, called, “fixed effect (FE) model” in Economics, which is a special case of the FE model we learned. Run a FE model using the “fev” dataset and comment on this model, e.g., what is doing, assumption, usefulness, etc. [Authors’ website/book has sample codes for SAS and R.]

After reading chapter 9, I learned that FE models:

What they’re doing: they’re estimated the regression coefficients for time varying covariates (Xij) using OLS, while controlling for time invariant characteristics of individuals potentially associated with the Xijs (potential confounders). For doing this, they estimate subject effects as fixed effects. This OLS estimation of β is an unbiased estimate, even if αi are correlated with Xij.

Assumptions: 1) Xij is assumed to be strictly exogenous (independent of random errors). 2) αi (the subject fixed effects) are allowed to be correlated with Xij (this is opposed to mixed effect models, where they are assumed to be independent), and this is where the usefulness of the model comes from.

Usefulness: they control time invariant characteristics of individuals (potential confounders) whose effect on the response are assumed to remain constant over time. They have the potential to remove bias when there are unmeasured, but stable, characteristics of the subjects that are correlated with time-varying covariates of main scientific interest (and so the focus should be on the latter)

Limitations: they can’t estimate the effects of time invariant covariates. However, you can still estimate interactions between time invariant and time varying covariates. Also, if most of the variability of the time varying covariates is between subjects, this model produces estimates of effect that are less precise for these variables (as opposed to mixed models).

Here are the results from the model (first 10 subjects)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

age 0.0298226 0.0004798 62.154 < 2e-16 \*\*\*

factor(id)1 -0.3987398 0.0252212 -15.810 < 2e-16 \*\*\*

factor(id)2 -0.3021976 0.0237526 -12.723 < 2e-16 \*\*\*

factor(id)3 -0.2430880 0.0222739 -10.914 < 2e-16 \*\*\*

factor(id)4 -0.4051946 0.0212142 -19.100 < 2e-16 \*\*\*

factor(id)5 -0.4240729 0.0252776 -16.777 < 2e-16 \*\*\*

factor(id)6 -0.3295989 0.0203574 -16.191 < 2e-16 \*\*\*

factor(id)7 -0.5259333 0.0252687 -20.814 < 2e-16 \*\*\*

factor(id)8 -0.3577286 0.0224608 -15.927 < 2e-16 \*\*\*

factor(id)9 -0.4090951 0.0222210 -18.410 < 2e-16 \*\*\*

factor(id)10 -0.1147699 0.0213233 -5.382 8.38e-08 \*\*\*

The estimated coefficient for age, 0.0298, indicates that each one year increase in age is associated with an e0.0298 = 1.030 or 3.0% relative increase in FEV1 (adjusted for height^2).

1. Conduct a sample N/power calculation assuming you are designing a longitudinal study. [You can use textbook formula or other valid formula in dropbox or other reliable sources.]

I thought I would be able to do the one in Ch20 for binary outcomes based on GEE estimation of beta, but that seems intractable at the moment.

Hence, I’ll just do the one based on a continuous outcome and a two stage estimation of mixed effect models. Here’s the setting:

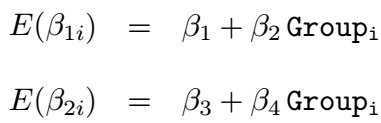
The goal of the study is to compare 2 treatments, A and B, for controlling obesity in overweight and obese people (BMI ≥ 25), through a randomized clinical trial, randomizing equal number of subjects, N, to two groups, and taking repeated measurements of the response variable, BMI. The two groups will be compared in terms of changes of the mean response over duration of the study, and it will be assumed the trends in BMI changes are linear over time in both groups. The model to estimate the mean effect will be a two stage mixed effect model, where the only covariate evaluated for explaining the differences in the time effect will be the treatment group.

For stage 1 the model for each subject will be:



Where is the subject i specific intercept and is the subject specific time effect, and is the within subject error, which are assumed to be iid N(0,)

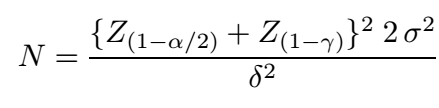
For stage 2 we will model each subject’s as a function of the treatment received:



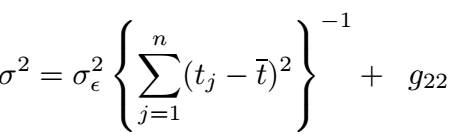
Where is 1 for treatment A and 0 for treatment B

The effect of interest will be the difference in slopes for time due to treatment effect, . H0: .

The sample size formula will be based on the following formula:



Where is



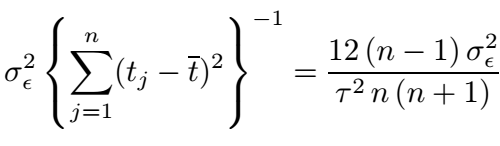
Where is the within subject variability obtained from stage 1 and is the between subject (unexplained) variability for

7 measures will be taken, one at baseline and 6 every 2 months, for a total follow up of 1 year.

We want to detect a minimum difference in the annual rate of change of

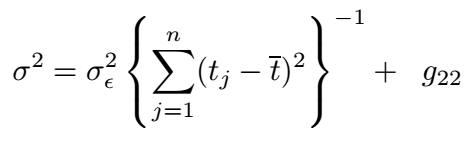
Based on historical data we consider reasonable estimates of Var( and of . The desired power is 80% when conducting a 2-sided test at the 5% significance level ( and ).

Given the above we have that



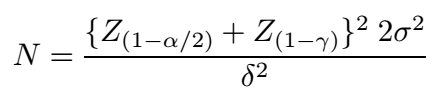
Translates into

And



Is equal to 7.7 + 3 = 10.3

Then, the N projected for each group is



So we would need a total of 52 subjects.