BIOS 6612 Homework 6: Longitudinal analysis with mixed models

Background: The Genetic Epidemiology of COPD (COPDGene) Study is a multi-center case-control study designed to identify genetic factors associated with COPD and to characterize COPD-related phenotypes. The study recruited COPD cases and smoking controls that are non-Hispanic whites (NHW) and African Americans (AA) ages 45 to 80 with at least 10 pack-years of smoking history. Forced expiratory volume in 1 second (FEV1) was taken before and after bronchodilator use.

Question of interest: Three investigators want to determine if Pre and Post bronchodilator FEV1 are jointly associated with emphysema adjusting for pack-years of smoking history, current smoking status, race, height and BMI in the first 1,000 subjects in the COPDGene study.

Investigators

- **Investigator 1** decided to not adjust for the two measurements per person (i.e., pre and post bronchodilator FEV1). He decided to run a linear regression on post and pre bronchodilator FEV1 adjusting for bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI using `file2_FEV1.csv`. This means that investigator 1 is pretending that he has 2,000 independent subjects when he in fact has 1,000 subjects with 2 measurements each.

- **Investigator 2** decided to run a linear regression on the difference of post minus pre bronchodilator FEV1 adjusting for baseline FEV1, pack-years of smoking history, current smoking status, emphysema, race, height and BMI using `file1_FEV1.csv`.

- **Investigator 3** decided to run 2 models. She first ran a random intercept and random slope model on FEV1 adjusting for bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI. She decided also to run a random intercept model on FEV1 adjusting for bronchodilator use, pack-years of smoking history, current smoking status, emphysema, race, height and BMI. She fits both of these models using `file2_FEV1.csv`.

Data sets

- **Data set 1**: Investigator 2 used `file1_FEV1.csv`. File 1 has 10 columns and 1,000 rows. `delta_FEV1` is the difference in post bronchodilator FEV1 (`post_FEV1`) minus

pre bronchodilator FEV1 (`pre_FEV1`). ID is the subjects ID. For current smoking status, `current_smoker=1` if the subject is a current smoker, and `current_smoker=0` if the subject is a former smoker. For race, `race=0` if NHW and `race=1` if AA.

- **Data set 2**: Investigators 1 and 3 used `file2_FEV1.csv`. File 2 has 9 columns and 2000 rows. Each of the 1,000 subjects appears twice: once before bronchodilator use (`trt=0`) and once post bronchodilator use (`trt=1`). `FEV1` is the post bronchodilator FEV1 if `trt=1` and `FEV1` is the pre bronchodilator FEV1 if `trt=0`. Note: the random slope model refers to a random slope for `trt`, which is equivalent to an indicator variable for two time points, pre and post.

Questions

Answer the following questions based on the above information and your analyses of the supplied data sets. Provide all code used for analysis, **but do not provide raw output**.

1. Fit each of the 4 models run by the 3 investigators (recalling that investigator 3 ran 2 models). Determine if emphysema is significantly associated with the outcome of interest adjusting for the provided confounders. Clearly state the outcome of interest and confounders for each investigator. (Hint: if you are using R, try fitting the random effects models with both 'lme4' and 'nlme'.)

2. (a) Explain how the model fit by investigator 2 is testing a different question of interest than the models fit by investigator 1 and 3.
   (b) Which investigators are directly answering the question of interest?

3. (a) Should investigator 3 choose the random intercept and random slope model or the random intercept model as her final model to perform inference on the fixed effects? Justify your answer.
   (b) Give two reasons why a likelihood ratio test comparing investigator 3's models using the χ^2_2 reference distribution is inappropriate.

4. Based on the models run by the 3 investigators, which model do you feel is most appropriate in this specific scenario to answer the question of interest that pre and post bronchodilator FEV1 are jointly associated with emphysema adjusting for the given confounders? Justify your answer.

5. The covariance structure for $\mathbf{Y}_i$ for a general mixed effects model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

can be written as

$$\operatorname{Var}\mathbf{Y}_i = \operatorname{Var}\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i\right) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{R}.$$

For the random intercept model fitted by investigator 3,

$$\mathbf{Z}_i = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{G} = \sigma_0^2, \quad \mathbf{R} = \begin{pmatrix} \sigma_\epsilon^2 & 0 \\ 0 & \sigma_\epsilon^2 \end{pmatrix},$$

so that

$$\text{Var}\,\mathbf{Y}_i = \begin{pmatrix} \sigma_0^2 + \sigma_\epsilon^2 & \sigma_0^2 \\ \sigma_0^2 & \sigma_0^2 + \sigma_\epsilon^2 \end{pmatrix}$$

(a) Give $\mathbf{Z}_i$ for the random intercept and slope model.

(b) Give $\mathbf{G}$ for the random intercept and slope model.

(c) Find the covariance structure of $\mathbf{Y}_i$ for the random intercept and slope model.

(d) Using these results, explain why the random intercept and slope model did not converge.