# Homework 2

Biost 540

## General Instructions

- Students may discuss with each other but each of you will be required to submit the work in your own writing.
- Grading will be based on completion (3 pts), accuracy (3 pts), work shown (3 pts), and neatness (1 pt).
- Be sure to show work for all problems. `R` code should not appear in the main body of the homework; however, the code should appear at the end of the assignment as an Appendix. It should be possible for someone to use the code to reproduce any figures or numeric results.

## Problem 1: Six Cities

### The Dataset

This data set comes from the Six Cities Study of Air Pollution and Health. The data are from a cohort of 300 school-age female children living in Topeka, Kansas. Most children enrolled in the first or second grade (between the ages of six and seven). They were then measured annually until high school graduation or loss to follow-up.

### Questions

a) Conduct a short exploratory data analysis. Specifically, create the following figures

   i. Histogram of age at entry
   ii. Scatterplot of log(FEV1) at entry (y-axis) vs age at entry
   iii. Scatterplot of log(FEV1) (y-axis) vs age (x-axis)
   iv. Scatterplot of log(FEV1) (y-axis) vs time from entry (x-axis)

b) Write a short paragraph commenting on what you notice from a).

c) Fit the following models using subject-specific random intercept and produce plots of the fitted curves for each:

   i. Step-function: (where AGE($k$): dummy variable for age category $k$)

$$\text{logFEV1}_{ij} = \beta_0 + b_{0i} + \beta_1 \text{AGE}(8-10) + \beta_2 \text{AGE}(10-12) + \cdots + \beta_5 \text{AGE}(16+) + \epsilon_{ij}$$

   ii. Linear trend:

$$\text{logFEV1}_{ij} = \beta_0 + b_{0i} + \beta_1 \text{age}_{ij} + \epsilon_{ij}$$

   iii. Quadratic:

$$\text{logFEV1}_{ij} = \beta_0 + b_{0i} + \beta_1 \text{age}_{ij} + \beta_2 \text{age}_{ij}^2 + \epsilon_{ij}$$

   iv. Linear spline: (where $(\text{age}_{ij} - \text{age}_k)_+ = \max(\text{age}_{ij} - \text{age}_k, 0)$)

$$\text{logFEV1}_{ij} = \beta_0 + b_{0i} + \beta_1 \text{age}_{ij} + \beta_2 (\text{age}_{ij} - 10)_+ + \beta_3 (\text{age}_{ij} - 14)_+ + \epsilon_{ij}$$

v. Cubic spline: (where $(\text{age}_{ij} - \text{age}_k)_+ = \max(\text{age}_{ij} - \text{age}_k, 0)$: linear spline based on "knot" $\text{age}_k$)

$$\begin{aligned}\text{logFEV1}_{ij} = \beta_0 + b_{0i} + \beta_1 \text{age}_{ij} + \beta_2 \text{age}_{ij}^2 + \beta_3 \text{age}_{ij}^3 + \\ \beta_4\{(\text{age}_{ij} - 10)_+\}^3 + \beta_5\{(\text{age}_{ij} - 14)_+\}^3 + \epsilon_{ij}\end{aligned}$$

d) Create a table that has the estimates of mean log FEV1 at ages 9, 11, 13, and 15 for each of the models.

e) Comment on your observations about the figures and table. Which models allow for non-linear mean curves and non-constant rates of change?

f) For this data we are considering a single group of subjects. In general, we will usually like to consider differences among groups over time. Let's consider the linear spline model and for simplicity use only a single knot at age 14:

$$E[\text{logFEV1}_{ij}|age_{ij}] = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2(\text{age}_{ij} - 14)_+$$

Suppose we have 2 groups and let $\text{group}_i$ be an indicator for the $i$th individual being in the "exposed" group (i.e. the reference group is the "unexposed" group)

i. First consider this model:

$$E[\text{logFEV1}_{ij}|\text{age}_{ij}, \text{group}_i] = \beta_0 + \beta_1 \text{group}_i + \beta_2 \text{age}_{ij} + \beta_3(\text{age}_{ij} - 14)_+$$

- What is the rate of change in the unexposed group prior to age 14? After age 14?
- What is the rate of change in the exposed group prior to age 14? After age 14?

ii. Now consider this model:

$$E[\text{logFEV1}_{ij}|\text{age}_{ij}, \text{group}_i] = \beta_0 + \beta_1 \text{group}_i + \beta_2 \text{age}_{ij} + \beta_3 \text{age}_{ij} \times \text{group}_i + \beta_4(\text{age}_{ij} - 14)_+$$

- What is the rate of change in the unexposed group prior to age 14? After age 14?
- What is the rate of change in the exposed group prior to age 14? After age 14?

iii. Finally, consider this model:

$$\begin{aligned}E[\text{logFEV1}_{ij}|\text{age}_{ij}, \text{group}_i] = \beta_0 + \beta_1 \text{group}_i + \beta_2 \text{age}_{ij} + \beta_3 \text{age}_{ij} \times \text{group}_i + \\ \beta_4(\text{age}_{ij} - 14)_+ + \beta_5(\text{age}_{ij} - 14)_+ \times \text{group}_i\end{aligned}$$

- What is the rate of change in the unexposed group prior to age 14? After age 14?
- What is the rate of change in the exposed group prior to age 14? After age 14?

g) From f), which model assumes the difference between the groups is constant over time? The difference between the groups is increasing (or decreasing) linearly over time? The difference between the groups is changing in a non-linear fashion over time?

## Problem 2: UK Understanding Soceity Study

### The Dataset

The data set is a subset of UK Understanding Society Study. Below are the description of the variables.

- pidp: subject identifier
- wave: data collection wave (1-9), annual follow up
- age_dv: age at data collection time
- sex_dv: sex (1: female)
- scghq1_dv: a well-being measure at each follow up

**Questions**

a) Produce a spaghetti plot with age as the x-axis and the well-being as the y-axis.
b) Produce a variogram summarizing the dependence in the well-being measures as a function of difference in study waves, after accounting for age and sex.
c) Fit three linear models with the same mean model (age and sex as covariates) but a different covariance structure:

i. Linear model without accounting for any within-subject correlation.
ii. Linear mixed model with a subject-specific random intercept.
iii. Linear mixed model with a subject-specific random intercept and serial correlation over data collection waves.

d) Compare the estimates and hypothesis tests based on the three models, for similarities and differences. To test whether age or sex is associated with well-being, which model would you choose and why? Interpret the findings of your chosen model.