P8157:

Analysis of Longitudinal Data
Midterm

Your solutions should reflect your own work and you are not to collaborate with anyone else. Throughout, make sure your responses are clear, legible and neat. No late exams will be accepted. As appropriate, append your R code. As you do, make sure that it is well-documented an easy to read.

Question 1:

In Homework #1 you conducted EDA for a sub-sample of K=300 girls from the Six Cities study, specifically from Topeka. In this question we are going to return to these data and fit a series of linear mixed models investigating the relationship between lung function and age. Towards adjusting for height we are going to consider the response:

$$Y_{ki} = \frac{\text{FEV}_{ki}}{\text{height}^2_{ki}}.$$

With this response in mind, consider two linear mixed models with the following (induced) marginal means:

$$Y_{ki} = \beta_0 + \beta_1 \text{Age}_{ki} \tag{2}$$

$$Y_{ki} = \beta_0 + \beta_1 \text{Age}_{ki} + \beta_2 \text{Age}^2_{ki} + \beta_3 \text{Age}^3_{ki} \tag{3}$$

The ultimate goal is to evaluate whether the more parsimonious model (2) can be adopted, or if there is sufficient evidence in the data to suggest that the more complex model (3) is appropriate. For the purposes of this question, restrict attention to those girls in the dataset with at least 5 measurements.

(a) Produce a figure of the response, Y_{ki} as a function of age. On the figure indicate the individual trajectories for a random sample of 4 girls.

(b) Use the `lme()` function to fit model (3) with the following dependence structures:

(i) independent, homoskedastic errors

(ii) random intercepts plus independent, homoskedastic errors

(iii) random intercepts/slopes plus independent, homoskedastic errors

(iv) random intercepts plus auto-regressive errors

(v) random intercepts plus exponential spatial errors

(vi) random intercepts plus exponential spatial errors and independent, homoskedastic errors

(vii) random intercepts plus independent, heteroskedastic errors

(viii) random intercepts/slopes plus independent, heteroskedastic errors

Note for (vii) and (viii), consider heterskedasticity by age. Report the results in two succinct tables. Specifically, as in the notes, produce a table reporting the log-likeihood and AIC commenting on which two models provide the best fit of the data. For these two models, write out the models in full notation, and report point and standard error estimates from the fits.

(c) Produce another figure of the response as a function of age (as in part (a)) but instead of indicating individual trajectories, augment the figure with two fitted regression curves The first is the fitted regression curve (i.e. the induced marginal mean model) using the best-fitting dependence structure from part (b). The second is the fitted regression curve using the same best-fitting dependence structure but with the mean model given by (2).

(d) Using at most 3-4 sentences, provide an explanation/interpretation of the curve for model (3) that you could use with a non-biostatistician collaborator.

(e) Using the output from the model fits in part (c), conduct a formal evaluation of whether model (2) can be adopted in favor of model (3). Explain in detail how you do this and what you conclude.

(f) Given the (somewhat artificial) choice between models (2) and (3), explain which model you could advise a collaborator to adopt and why.