Statistics Learning, Module 2, 2021

Syndicate Problem Set # 3

All responses are to be neatly typed up. This includes any equations that you choose to present. The page setup should have at least 2cm margins on all sides, with all texts formatted to 12pt font size and at least 1.5 spacing. Present your numerical results neatly in tables and plots. R screenshots are not acceptable unless explicitly requested.

The submitted responses should not exceed 6 pages. Graphs and tables may be presented in an Appendix, with the length of the appendix not exceeding 4 pages.

The assignment is due on Monday July 2nd at 9PM. You are expected to work through these problems over the weeks prior, as indicated below.

CD4 percentage level in child with HIV (To be completed during Week 6)

CD4 cells are carried in the blood and are part of the human immune system. One of the effects of a HIV infection is that these cells die. The count of CD4 cells is used in determining the onset of advanced AIDS in a patient. To study the effectiveness of a new treatment on HIV, 226 HIV-positive child patients had their CD4 counts recorded and were then put on a treatment course with a drug. After taking the drug, their CD4 counts were again recorded in several visits. The aim of the experiment was to test whether or not patients taking the drug had increased CD4 counts.

The data files "hiv.csv" contains the records of the CD4 percentage (CD4PCT) in each child patient for some visits including the initial one, along with the age at a given visit and type of treatment received.

1. Construct a multilevel linear model that allows for random effects on the intercept across different child patients. For the time being, use the variable *time* only as a predictor. Present and analyse this model. Your response should not exceed 1 page.

2. Construct a multilevel linear model so that it also allows for *treatment* and the child's age at initial visit *baseage* to explain the random intercept. Present the models and discuss the results. How do the multilevel models compare now that the child-specific variables have been accounted for in the effect? Analyse the effect of child-specific variables on the baseline intercept as part of your discussion. Your response should not exceed 1 page.

3. Discuss how the multilevel model is different to the multiple linear regression model in this context. Analyse the data using multiple linear regression for a comparison. In the context of the problem, would you prefer to use the random effect model or the linear regression? Your response to this should not exceed 1 page.

Segments in the Credit data (To be completed during Week 7)

The file "Credit.csv" contains the data you explored during weeks 1 and 5.

1. Construct a latent mixture linear model for Balance with two components. Use the following base linear model:

   Balance~Income+Rating+Cards+Age+Education+Student+Married+Gender+Ethnicity

   Present your model.

   Using the "stepFlexmix" command, estimate the latent mixture model for k=1:5. How many segments do you think is present in this data? Make sure you specify your choice of the number of segments and provide justification for it. Your response to this should not exceed 1 page.

2. Discuss the chosen model from 1.). Specifically, you should focus on the differences/similarities between the components implied by your chosen model. What are the key differences between this chosen model and the model you constructed in weeks 1 and 5?

   Your response to this should not exceed 1.5 pages.