Problem set 3

S670 Spring 2020

Upload a HTML or PDF document with your code, graphs, and write-up to Canvas by 11:59 pm, Thursday 6th February.

The problem

High systolic blood pressure is a strong predictor of heart attacks and strokes. A health researcher wants to know how average systolic blood pressure varies with:

- Age
- Height
- Weight

For ease of interpretation, she does not wish to transform systolic blood pressure, but she is willing to consider interpretable transformations of the explanatory variables. She thinks the trends should be relatively smooth, but not necessarily linear. She suspects that for at least some of these variables, the trends may be different for men and women. In addition to estimating the trends, she wants to know how close observations typically are to the trend and whether the models might have any explanatory value. However, she is not interested in making predictions for individuals. She is not interested in formal inference right now, though she may be in the future. She knows some R, so you may include R code in your report, but she can't read your mind, so label your graphs.

The questions

Use the NHANES data in the NHANES package to explore the researcher's questions. The relevant variables are:

- BPSysAve (the average of three measurements of systolic blood pressure)
- Age (in years; 80 or older is recorded as 80)
- Weight (in kilograms)
- Height (in centimeters)
- Gender (male or female)

Write a document in three sections, giving the relationship of average systolic blood pressure with age, height, and weight respectively. (You can also include an introduction and conclusion if you really want to.) Each section should include approximately TWO graphs (a set of faceted plots counts as one graph) examining the trend and the residuals. Including many more graphs than this may be penalized. In each section, include a brief justification of your modeling choices (type of model, transformations or lack of transformations) and a verbal description of the differences you see between men and women. Some (sensibly rounded) quantitative measures will probably be useful, but you do not (and should not) list every single statistic you can think of.

Tips

- A safe approach would be to fit separate models for men and women for each explanatory variable (though other approaches are possible.)
- The group and color arguments within aes() can be used to distinguish between men and women.
- Because there's a lot of data, you might have to play around with the plot settings to get legible graphs. Google the help pages for the individuals geoms (e.g. `geom_point()`) to learn about aesthetic arguments for those functions.
- If the default axis limits don't look nice, you can choose them with `+ xlim()` and `+ ylim()`.
- If you see any weird results, try to explain them.