Statistics 600 Problem Set 4

Due November 17th at midnight

This problem set focuses on data analysis using least squares techniques. You will work with the NHANES (National Health and Nutrition Examination Survey) dataset. There are many waves of data, you should use the data for the 2015-2016 wave. It is easy to find the data by searching online. Be sure that you get the data from the CDC website (the URL begins `wwwn.cdc.gov`).

The files are in XPT format. You will need to find a reader for XPT files in your language. Then, download the demographics (`DEMO_I.XPT`), body measurements (`BMX_I.XPT`) and blood pressure (`BPX_I.XPT`) files. Finally, merge the files on the `SEQN` variable (which indicates distinct subjects).

The exercises below focus on explaining the variance in systolic blood pressure (`BPXSY1`) using up to three explanatory variables. The explanations below should be self-contained, but if you wish you can consult the data dictionaries that are available from the same web pages where you obtained the data.

For each question below, submit any numerical or graphical results, along with a few sentences describing your findings. You do not need to submit your code.

1. Consider linear models for blood pressure in terms of age (`RIDAGEYR`) and sex (`RIAGENDR`, coded 2=female and 1=male). Using techniques discussed in this course, consider (i) is there any relationship between linear age and blood pressure?; (ii) is there any relationship between sex and blood pressure? If you find evidence for both associations, consider whether they are additive.

2. Use splines to assess for a non-linear relationship between age and blood pressure. If a non-linear relationship exists, consider whether it is additive with respect to sex.

3. Graph the conditional mean of blood pressure with respect to age, for each sex. These conditional means should be estimated using the best model that you discovered in parts 1-2. Do not stratify the data by sex and fit separate sex-specific models.

4. Consider the contrast $E[\text{BP}|\text{age}, \text{sex} = F] - E[\text{BP}|\text{age}, \text{sex} = M]$. Esti-

mate this contrast using the best model that you discovered in parts 1-2, then graph the estimated contrast against age. Then use the Scheffeé approach to put a 95% simultaneous coverage band around this function.

5. Consider BMI (`BMXBMI`) as a third predictor of blood pressure. Use methods discussed in this course to find an appropriate model using all three predictors, or a subset if warranted.

6. Find the best-fitting additive model for blood pressure in terms of age, sex, and BMI. Consider non-linearities in age and BMI as appropriate. Using this model, determine the partial R^2 for each of the three variables when added to a model that already contains the other two.

7. Based on the additive model that you constructed in part 6, make a partial residual plot of blood pressure in relation to BMI.

8. Create an added variable plot by residualizing blood pressure against age and sex, and residualizing BMI against age and sex. These residuals should incorporate non-linear and non-additive effects of age and sex where appropriate. Note the R^2 for each of the two models that you use to residualize. Then make a scatterplot of the residuals against each other.

9. Take the residuals from the model you constructed in part 5. Make a scatterplot of the absolute values of the residuals against age and against BMI. Use OLS to regress the absolute residuals (or their logarithms) against age, sex, and BMI in an additive model. Consider what this analysis reveals about heteroscedasticity.

10. Using the definitions of total, direct, and indirect effects from problem set 3, calculate these effects in the NHANES data. Specifically, consider the direct effect of age on blood pressure, and the indirect effect of age on blood pressure through BMI. Perform this analysis separately for women and for men.