STAT UN2103 Homework 6 [200 pts]

Due TBA

Final Graded Assignment

Homework 6 is an open ended case study intended to be a capstone project for this class. The assignment is weighted 200 points (2 homeworks).

Data Description

The data comprise of roughly 25,000 records for males between the age of 18 and 70 who are full time workers. A variety of variables are given for each subject: years of education and job experience, college graduate (yes, no), working in or near a city (yes, no), US region (midwest, northeast, south, west), commuting distance, number of employees in a company, and race (African America, Caucasian, Other). The response variable is weekly wages (in dollars). The data are taken many decades ago so the wages are low compared to current times. The data set `salary.txt` is posted on Canvas.

Research Question

A government official is interested in whether the average male wages are statistically different for the three race classes. Specifically, the government official wants you to answer the following research questions:

1. Do African American males have statistically different wages compared to Caucasian males?
2. Do African American males have statistically different wages compared to *all* other males?

The goal of this case study is two fold:

i. Come up with a linear regression model that incorporates all relevant variables, interactions and functional forms of the covariates.

ii. Using your final model, test the two research questions above.

Write up

Students are required to type up the final report. The final report should be broken up into the following sections.

I. **Introduction:** Include a brief description of the goals of this analysis coupled with some exploratory data analysis. Keep the exploratory analysis brief, including a few plots and summary statistics to support the research question. Be creative on the exploratory analysis and only include plots that you feel informative.

II. **Statistical Model:** In this section, clearly state your final model along with the `R` summary output. Be sure to describe all interactions, functional forms and transformations of your model. Also include the statistics AIC, R^2 and R_a^2.

III. **Research Question:** Perform the relevant testing procedures to answer the two research questions. Also include a brief written summary of your results.

IV. **Appendix**

a. **Model Selection:** Here you will explain in detail what interactions, functional forms and variables you decided to include in the model. Describe if and why a transformation is applied to the response variable. You will also explain how you arrived at your final model based on a handful of model candidates. Without overwhelming me, include relevant `R` output and plots that helped you arrive at your final model.

b. **Diagnostics and Model Validation:** Include all relevant diagnostic plots on the final model. Also include the computed $MSPR$ and how it compares to the computed MSE.

c. **Anything you Feel Necessary:**

R Code

- Students should prepare an organized `R` script file that complements the written report. Please include comments that help describe your model building process. Also describe the exploratory analysis, relevant diagnostics and how you tested the research questions.
- **Do Not** copy and paste the `R` code into your appendix. Please upload the `R` script file on Canvas by the due date.

Further Considerations

- Always include the main effects when including an interaction.
- Include any lower order terms when including a polynomial functional form in the model, i.e., do not include only x_1^3 without also including x_1 and x_1^2.
- If you believe that another covariate interacts with race, you must test race in conjunction with the interaction, i.e., you cannot just test race by itself if you believe it statistically interacts with another predictor. An between race and other covariates is not required and does make the research question more difficult to answer.
- In order to reduce collinearity, make sure to center any covariates that have a polynomial functional form **Or** use orthogonal polynomials, i.e., the function `poly()`.
- Feel free to include any functional forms you fell necessary other than polynomials and interactions (piecewise functions, square roots, reciprocals... etc).
- For diagnostics, there is no need to perform any formal testing procedures to validate assumptions, i.e., do not perform the F lack-of-fit test, Levene's homogeneity of variance test and kolmogorov-Smirnov normality test. All of the regression assumptions can be validated by inspecting the residual plots.
- Use studentized deleted residuals for all diagnostic plots. (`rstudent(model)`)

- You should initially split the data set up into a *model building* data set and a *model validation* data set. Use the *model building* data set to construct your model and use the *model validation* for validation. Given the size of the data set, I recommend letting the validation set be roughly 20% of the model building data set. When validating the model, only look at the *mean square prediction error* $MSPE$ and compare this value to MSE. So everyone uses the same data sets, I will provide the `R` code for extracting the correct building and validation data sets.
- I recommend using AIC and C_p for model selection.

Grading

- This homework assignment will be graded on completeness, correctness and organization/neatness. I want to see a nice organized final report. It must be typed with graphs labeled. **Please do not make the report too long!** Maybe 10-20 pages.
- Also try to get R^2 above 30%.