Assignment 3, SDGB 7840

Instructor: Prof. Nagaraja

Due: 4/12 in class

Submit two files: (a) .Rmd R Markdown file with explanations and code and (b) Word document of knitted R Markdown file. Email both files to fordhamRcomputing@gmail.com by the start of class with the subject line "HW3-[Full Name]-[Class Time]" and include HW 3 and your name in the file names (time of class is either 3:30 PM or 5:45 PM). Please email your solutions only once! Comment your code for full credit and complete the assignments individually. Use the materials in the "Assignments" folder for more on how to use RMarkdown and how to set up your assignment. Note: There are two different data sets in this assignment.

1. We will model the prestige level of occupations using variables such as education and income levels. This data was collected in 1971 by Statistics Canada (the Canadian equivalent of the U.S. Census Bureau or the National Bureau of Statistics of China)[1]. The data is in the file "prestige.dat" and the variables are described below:

variable	description
prestige (y)	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s
education	average education of occupational incumbents, years, in 1971
income	average income of incumbents, dollars, in 1971
women	percentage of incumbents who are women
census	Canadian Census occupational code
type	type of occupation: "bc"=blue collar, "prof"= professional/managerial/technical, "wc"=white collar

   (a) Do some internet research and write a short paragraph in your own words about how the Pineo-Porter prestige score is computed. Include the reference(s) you used. Do you think this score is a reliable measure? Justify your answer.

[1] Source: Canada (1971) *Census of Canada*. Vol. 3, Part 6. Statistics Canada; 19-1–19-21.

(b) Create a scatterplot matrix of all the quantitative variables and describe what you see. Use a different symbol for each profession type: no type (`pch=3`), "bc" (`pch=6`), "prof" (`pch=8`), and "wc" (`pch=0`) when making your plot. For the remainder of this question, we will use the explanatory variables: income, education, and type. Does restricting our regression to only these variables make sense given your exploratory analysis? Justify your answer.

(c) Which professions are missing "type"? Since the other variables for these observations are available, we could group them together as a fourth professional category to include them in the analysis. Is this advisable or should we remove them from our data set? Justify your answer.

(d) Visually, does there seem to be an interaction between type and education and/or type and income? Justify your answer.

(e) Fit a model to predict prestige using: income, education, type, and any interaction terms based on your answer to part (d). Evaluate your model by checking the regression assumptions (including collinearity/multicollinearity) and provide details and conclusions for any hypothesis tests. Include relevant output. Use your answer to part (c) to determine which observations to use in your analysis.

(f) Create a histogram of income and a second histogram of log(income) (i.e., natural logarithm). How does the distribution change?

(g) Fit the model in (e) but this time use log(income) (i.e., natural logarithm) instead of income. Evaluate your model by checking the regression assumptions and details and conclusions for any hypothesis tests. Include relevant output.

(h) Is the model in (e) or (g) better? Justify your answer. Why can't we use a partial F-test here?

2. In this question, you will be modeling medical expenditures of individuals who lived in Vietnam in 1997. The data set is in the file "VietNam.csv" and was taken from the Vietnam World Bank Living Standards Survey. The variables are:

- ID number of person
- number of direct pharmacy visits
- log of total medical expenditures (y)
- age of household head
- gender (male/female)
- 1 if person is married; 0 if single
- number of years of education
- number of illnesses experienced in the past 12 months
- 1 = if injured when the individual completed the survey; 0 if not

- number of illness days
- number of days of limited activity
- 1 = if person has health insurance coverage; 0 if not
- commune (like a district)

(a) Clean your data. Randomly assign your data into training and test sets. Half of the data should be in the training set and half should be in the test set. How many observations are in the training set? Is there any missing data?

(b) Create a table of summary statistics (mean, median, standard deviation, minimum, and maximum) and make scatterplots for the pairs of numerical variables. If you need to transform any data because of potential linearity issues, do that now. Also, if you see any interaction terms which may be useful while doing your exploratory data analysis, you can add them in when you do questions 2c, 2d, and 2e. (Note: since the response variable has already been transformed, do not transform it again.)

(c) Use **forward selection** to choose the "best" model. Explain how you decided which model was "best."

(d) Use **backward elimination** to choose the "best" model. Explain how you decided which model was "best."

(e) Use the **sequential replacement** (or called forward and backward stepwise) method to choose the "best" model (`seqrep` in the `regsubsets()` function). Explain how you decided which model was "best."

(f) Compute the test set RMSE for the models you chose in 2c, 2d, and 2e. Compare them to the RMSE values from the training set. Which model performs best? Justify your answer.

(g) Check your regression assumptions (including collinearity/multicollinearity) for the model you chose in question 2f. Are the assumptions satisfied? Justify your answer.

(h) Interpret the partial slopes and y-intercept of the model you chose in question 2f.