Homework 2: SDGB 7840

Instructor: Prof. Nagaraja

Due: 2/28

Submit two files through Blackboard: (a) .Rmd R Markdown file with answers and code and (b) Word document of knitted R Markdown file. Your file should be named as follows: "HW[X]-[Full Name]-[Class Time]" and include those details in the body of your file.

Complete your work individually and comment your code for full credit. For an example of how to format your homework see the files posted with Lecture 1 on Blackboard. **Show all of your code in the knitted Word document.**

1. Read the posted article, "Bordeaux wine vintage quality and weather," by Ashenfelter, Ashmore, and LaLonde (CHANCE, 1995). Three regression models are considered in this article. Answer the following questions:

   (a) What is a wine "vintage"?

   (b) What is the response variable for the three models described in this paper?

   Now, download the data in "wine.txt". This is some of the data the authors used to fit their models. The columns are: vintage (VINT), log of average vintage price relative to 1961 (LPRICE2), rainfall in the months preceding the vintage in mL (WRAIN), average temperature over the growing season in °C (DEGREES), rainfall in September and August in mL (HRAIN), and age of wine in years (TIME_SV).

   Note: the average temperature in September is not available in our data set so we cannot fit the third regression model from the paper.

   (c) Which values of LPRICE2 are missing and, according to the article, why have they been omitted?

   (d) Make a scatterplot matrix of the variables (explanatory and response) included in the models. Describe what you see.

(e) Fit the two regression models from the paper. Which is the best regression model? Justify your answer and include relevant output (let $\alpha = 0.05$). Did you choose the same model as the authors?

(f) What is the sample size for your models?

(g) Write out the regression equation of the model you chose in part (e). Remember to include the units of measurement. Interpret the partial slopes and the y-intercept. Does the y-intercept have a practical interpretation?

(h) Make a table with the following statistics for both models: SSE, RMSE, PRESS, and $\text{RMSE}_{jackknife}$. Compare the relevant statistics. Based on this information, would you change your answer to part (e)? Justify your answers.

(i) Could we use these regression models to predict quality for wines produced in 2005? Justify your answer.

2. 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.

(b) Create a scatterplot matrix of all the quantitative variables. 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.

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

(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 the model and 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 the model and provide the relevant output.

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