Multiple Linear Regression

Due: Monday 09/27 (11.00PM)
Submission: On Gradescope

The Homework contains two parts:
Part I consists of practice problems that you can work on to practice; you do not need to submit these. Some of these will be discussed during Thursday's office hours. Part II consists of the problems that you have to submit. Use R and R Markdown as necessary and submit your solutions as a PDF or HTML file.

Part I: Practice Questions

1. **Grocery Retailer**
   A large national grocery retailer tracks productivity and cost of its facilities closely. The data in the `grocery.txt` file were obtained from a single distribution center for a one year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is called in one of the week has a holy day and zero otherwise (X_3), and the total labor hours (Y).

   (a) Identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State decision rule and conclusion.

   (b) Obtain the diagonal elements of the hat matrix. Identify any high leverage points. If any, are they good or bad?

   (c) Cases 16, 22, 43, and 48 appear to be outlying X observations, and cases 10, 32, 38, and 40 appear to be outlying Y observations. Obtain the Cook's distance values for each of these cases to assess their influence. What do you conclude?

2. Use the `teengamb` data from the *faraway* library to fit a model with gamble as the response and the other variables as predictors.

   (a) Predict the amount that a male with average (given this data) status, income and verbal score would gamble along with an appropriate 95% confidence interval.

   (b) Repeat the prediction for a male with maximal values (for this data) of status income and verbal score. Which confidence interval is wider and why? Is this result expected?

   (c) Fit a model with $\sqrt{gamble}$ as a response but with the same predictors. Now predict the response and give a 95% prediction interval for the individual in (a). Take care to give your answer in the original units of the response.

   (d) Repeat the prediction for the model in (c) for a female with status= 20, income = 1, verbal=10. Comment on the credibility of the result.

   (e) Using the model with gamble as the response, check for large leverage points, outliers, and influential points.

Part II: Homework Questions – to be submitted

The `whitewines.csv` data set contains information related to white variants of the Portuguese "Vinho Verde" wine. Specifically, we have recorded the following information:

(a) `fixed acidity`, (b) `volatile acidity`, (c) `citric acid` , (d) `residual sugar`, (e) `chlorides` , (f) `free sulfur dioxide`, (g) `total sulfur dioxide`, (h) `density`, (i) `pH`, (j) `sulphates`, (k) `alcohol`, (l) `quality` (score between 0 and 10)

In this homework, our goal is to explain the relationship between `alcohol level` (dependent variable) and `residual sugar`, `pH`, `density` and `fixed acidity`.

(a) Identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = .05$. State decision rule and conclusion.

(b) Obtain the diagonal elements of the hat matrix and identify any high leverage points. If any, are they good or bad?

(c) Use Cook's distance to investigate whether there are any high influential points. What do you conclude?

(d) Calculate Cook's distance D_i for each case and prepare an index plot. Are any cases influential according to this measure?

(e) Predict the amount of alcohol of a white wine with residual.sugar = 1.7 , pH = 3, density = 1, fixed.acidity = 6.3 with an appropriate 95% confidence interval.

(f) Predict the amount of alcohol of a white wine with residual.sugar = 67 , pH = 4, density = 1.1, fixed.acidity = 15 with an appropriate 95% prediction interval.

(g) Construct a 95% confidence region for the slope coefficients of pH and density. What do you conclude about the statistical significance of β_{pH} and $\beta_{density}$?

(h) Regress `alcohol` against `fixed acidity` and construct a 95% simultaneous confidence band for the fitted regression line.

(i) Plot the raw data corresponding to question (h), fitted regression line, 95% point-wise confidence intervals and 95% confidence band calculated in (h). What do you observe?