Multiple Linear Regression

Due: Tuesday 09/20 (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) Fit a regression model to the data for the three predictor variables. State the estimated regression function.

   (b) Calculate the coefficient of multiple determination R^2.

   (c) Test whether there is a regression relation, using $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2 and β_2?

   (d) Test whether X_3 can be dropped from the regression model given that X_1, X_2 are retained. Use level of significance 0.025. State the alternatives, decision rule and conclusion.

   (e) Test whether both X_2 and X_3 can be dropped from the regression model, given that X_1 is retained. Use level of significance 0.025. State the alternatives, decision rule and conclusion.

2. In the *punting* data from the *faraway* library we find *average distance* and *hang times* of 10 punts of an American football as related to various measures of leg strength for 13 volunteers.

   (a) Fit a regression model with *Distance* as the response, and the *right* and *left* strengths and flexibilities as predictors. Which predictors are significant at the 5% level?

   (b) Use an F -test to determine whether collectively these four predictors are significant at the 5% level.

   (c) Relative to the model in (a), test whether the right and left strength have the same effect.

   (d) Fit a model to test the hypothesis that it is total leg strength defined by adding the right and left leg strengths that is sufficient to predict the response, in comparison to using individual left and right strengths.

   (e) Relative to the model in (a), test whether the right and left leg flexibilities have the same effect.

   (f) Fit a model with *Hang* as the response, and the same four predictors. Can we make a test to compare this model to that used in (a)? Explain.

Part II: Homework Questions – to be submitted

1. Derive a formula relating R^2 and the F -test for the regression. You can find the formula in the lecture slides. Here you need to derive it

2. 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) Fit a regression model to the data for the four predictor variables mentioned above. State the estimated regression function.

   (b) Prepare partial scatter plots for all 5 variables under consideration. What do you observe?

   (c) Compute the correlation matrix for all the 5 variables. What are your observations?

   (d) Calculate the coefficient of multiple determination R^2. Interpret your results.

   (e) Test whether there is a linear regression relation, using $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What does your test *imply* about β_1, β_2, β_3 and β_4?

   (f) Test whether X_3 can be dropped from the regression model given that X_1, X_2 are retained. Use level of significance 0.025. State the alternatives, decision rule and conclusion.

   (g) Test whether both X_2 and X_3 can be dropped from the regression model, given that X_1 is retained. Use level of significance 0.025. State the alternatives, decision rule and conclusion.