ML 4375 – Intro to Machine Learning

Homework 2: Linear Regression

Objective: Learn how to use R to perform linear regression.

Turn in: an ".Rmd" file to eLearning
Important: Start the file with comments for your name, Homework 2; Have prominent headings for the steps like "# 1" so the TA can find your code.

Problem 1: Simple Linear Regression

1. Load library ISLR. Use the names() function and the summary() function to learn more about the Auto data set. Divide the data randomly into train and test sets, with 75% train. Use seed 1234 for reproducibility.
2. Use the lm() function to perform simple linear regression on the train data with mpg as the response and horsepower as the predictor. Use the summary() function to evaluate the results. Calculate the MSE.
3. (No code) Write in the white text area:
    a. Write the equation from the model , filling in the parameters w, b
    b. Is there a strong relationship between horsepower and mpg?
    c. Is it a positive or negative correlation?
    d. Comment on the RSE, R^2, and F-statistic
    e. Comment on the MSE
4. Plot train$mpg~train$horsepower and draw a blue abline(). Comment on how well the data fits the line. Predict mpg for horsepower of 98. Comment on the predicted value given the graph you created.
5. Test on the test data using the predict function. Find the correlation between the predicted values and the mpg values in the test data. Comment on the results. Calculate the mse on the test results. Compare this to the mse for the training data.
6. Plot the linear model in a 2x2 arrangement. Do you see evidence of non-linearity from the residuals?
7. Create a second linear model with log(mpg) predicted by horsepower. Compare the summary statistic R^2 of the two models.
8. Plot the function with an abline(). Comment on how well the line fits the data.
9. Predict on the test data using lm2. Find the correlation of the predictions and log() of test mpg. Compare this correlation with the correlation you got for lm1. Calculate the MSE for the test data on lm2 and compare to lm1.
10. Plot the second linear model in a 2x2 arrangement. How does it compare to the first set of graphs?

Problem 2: Multiple Linear Regression

1. Produce a scatterplot matrix which includes all the variables in the data set using the command “pairs(Auto)”. List any possible correlations that you observe, listing positive and negative correlations separately, with at least 3 in each category.
2. Display the matrix of correlations between the variables using function cor(). You should exclude the “name” variable since is it qualitative. Write the two strongest positive correlations and their values in the text area. Write the two strongest negative correlations and their values in the text area.
3. Convert the origin variable to a factor. Use the lm() function to perform multiple linear regression with mpg as the response and all other variables **except name** as predictors. Use the summary() function to print the results. Which predictors appear to have a statistically significant relationship to the response?
4. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Are there any leverage points? Display a row from the data set that seems to be a leverage point.
5. Use the * and + symbols to fit linear regression models with interaction effects, choosing whatever variables you think might get better results than your model in step 3 above. Compare the summaries of the two models, particularly R^2. Run anova() on the two models to see if your second model outperformed the previous one.

Element	Points
R script runs	20
Appropriate comments and white space	5
Steps (15)	5 points each