HW4

2025-02-16

Homework 4: Multiple Linear Regression

Setup 1

This question involves the use of multiple linear regression on the Auto data set. The data set is provided in a CSV file. Check the UCI repository for the details of the data set, link: https://archive.ics.uci.edu/ml/datasets/Auto+MPG.

Question set 1

a. Load data in R and see the summary statistics. Mention if you need any pre-processing of the data. Hints: missing data is coded as a '?' symbol. Origin is a categorical variable. The name of a car should not be used to model mpg. Finally, you may delete or impute missing values (see question c).

```r
setwd("/Users/atanugiri/OneDrive - University of Texas at El Paso/Class Documents/Data Mining/Homework/a
```

b. Produce a pair plot that includes all relevant variables in the data set. You may use the 'ggpairs' function from the GGally package.

c. Compute the matrix of correlations between the variables. You will need to exclude the name. Also, do not use the origin, as it is a categorical variable. However, the origin should be used for modeling in the next question. Hints: you may use the cor() function from base R. The corrplot package gives a nice visualization of the correlation matrix.

d. Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

i) Is there a relationship between the predictors and the response?

ii) Which predictors appear to have a statistically significant relationship to the response?

iii) What does the coefficient for the year variable suggest?

e. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

f. Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant? Formula $y \sim .^2$ is used to include all interaction terms.

g. Try a few different transformations of the variables, such as $log(X)$, $\sqrt{X}$, X^2. Note that X^2 transformation needs $I()$ in the formula: $y \sim I(X^2)$. For $\sqrt{(X)}$ and X^2, you can simply use $log(X)$ and $\sqrt{X}$, respectively. Comment on your findings. You may also consider transforming the response variable. The goal is to be familiar with some variable transformations, although they may not be the optimum ones.

Setup 2

This question should be answered using the Carseats data set. The data set is provided in a CSV file.

Question set 2

a. Fit a multiple regression model to predict Sales using Price, Urban, and ShelveLoc.

b. Provide an interpretation of each coefficient in the model. Be careful – some of the variables in the model are qualitative!

c. Write out the model in equation form, being careful to handle the qualitative variables properly.

d. Add the interaction between Urban and Price in the model. Interpret the fitted coefficients.

e. For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

f. Now fit a multiple linear model for Sales using all variables provided in the data set (intercept and main effects only). Comment on the model fitting.

g. Fit a smaller model that only uses the predictors for which there is evidence of association with the outcome. We will discuss variable selection in a later chapter, but for this question, select variables with significant p-values.

h. How well do the models in (f) and (g) fit the data? You may use anova() function to compare to models.