Homework: Ch 03

STAT 4510/7510

Due Wednesday, February 16, 11:59 pm

Instructions: Please list your name and student number clearly. In order to receive credit for a problem, your solution must show sufficient detail so that the grader can determine how you obtained your answer.

Submit a single pdf file for your final outcome. All `R` code should be included, as well as all output produced. Upload your work to the Canvas course site.

Problem 1

Read and complete 3.6 Lab: Linear Regression, found on pages 109 - 119.

- Read carefully the descriptions for each command from the textbook.
- Work through the textbook lab as written and execute the commands.
- Data files `Boston.csv` and `Carseats.csv` can be found on Canvas website under Module 3. You may want to read these data files with `stringAsFactors=TRUE`.
- Include your commands and output in your homework submission except the outcomes of `fix()` function.
- At page 114, when you use `vif()` function, part of the `car` package, you will see the format of an outcome that is slightly different from what you see from the textbook. You can simply refer to the GVIF values from the outcome to read the VIF values.

Problem 2

This question involves the use of multiple linear regression on the `Auto` data set. The data file `Auto.csv` can be found on Canvas under Module 1. We use the linear regression model with `mpg` as the response and all other variables except `name` as the predictors.

(a) The original data file includes some missing values recorded as the character `?`. To avoid problems caused by this missing values, we will simply remove the rows containing any missing value. Also the variable `origin` should be treated as a factor variable (`origin` of car: 1. American, 2. European, 3. Japanese). Read the data file by using

```
Auto <- read.csv("Auto.csv", na.strings ="?") # With the option, R recognizes ? as NA.
Auto <- na.omit(Auto) # Remove data rows including NA.
Auto$origin <- as.factor(Auto$origin) # Coerce the type of origin into factor
```

(b) Produce a scatterplot matrix which includes all of the variables in the data set except `name`. Which predictor variables do you think are shown as being associated with the response variable `mpg`?

(c) Compute the matrix of correlations between all the numerical variables using the function `cor()`. Exclude categorical variables `origin` and `name`.

1. Do you find outcomes being consistent with what you found from the scatterplot matrix? Explain.
2. Describe potential collinearity problems involved in this data.

(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 regarding following questions.

1. Is there a relationship between the predictors and the response?
2. Which predictors appear to have a statistically significant relationship to the response?
3. What does the coefficient for the `year` variable suggest?

(e) Use `constrast()` to learn how `R` created dummy variables for `origin`. Based on what you found, interpret the coefficients corresponding to the `origin2` and `origin3` predictors.

(f) Fit a simple linear regression with `mpg` as the response and `cylinders` as the predictor. Do you find a consistent result regarding the significance of `cylinders` with that obtained from the multiple linear regression fit? If not, provide possible explanation on the inconsistency.

(g) Let's call the model you obtain from the part (d) Model-1.

1. Use `vif()` function, part of the `car` package, to investigate collinearity problems involved in Model-1. (You can refer to the GVIF values from the outcome of `vif()` to read the VIF values.) The predictors with VIF values greater than 10 are usually treated as being problematic.

2. Remove `displacement` and `weight`, which have the most largest and the secondly largest VIF values, from the model. Refit the linear regression model (Model-2). Compare parameter estimates outcome of Model-2 with those of Model-1. Describe how standard errors and t-statistics of `cylinders`, `horsepower`, and `acceleration` are changed.

3. Re-investigate VIF for Model-2, and validate if collinearity problem is solved.

(h) Use the `plot()` function to produce diagnostic plots of the linear regression fit of Model-2. Comment on any problems you see with the fit.

1. Do the residual plots suggest any nonlinearity of the relationship between response and predictors?
2. Are there any unusually large outliers?
3. Does the leverage plot identify any observations with unusually high leverage? (For your information, traditionally we consider any observation for which the leverage statistic exceeds $2(p+1)/n$, twice the average, as high leverage points.)

(i) Try to fit linear regression models with an interaction effect between `horsepower` and `origin`. Does the interaction appear to be statistically significant?

(j) (7510 students only) Validate what you found from the part (i) by drawing a scatter plot with `horsepower` as x-axis and `mpg` as y-axis, grouped by levels of `origin`. Follow the description below.

1. Fit a simple linear regression model with `mpg` as the response and `horsepower` as the predictor only using data points with `origin` belongs to 1. Store the fitted model into the object named `lm.auto.origin1`. (Hint: *You can extract data points with* `origin==1` *by* `Auto[Auto$origin==1,]`.)

2. Repeat the previous simple linear regression model fit with data points with `origin` equal to 2 and 3, respectively. Store the fitted models into the objects named `lm.auto.origin2` and `lm.auto.origin3`.

3. Use `plot()` function to draw a scatter plot with x-axis `Auto$horsepower`, y-axis `Auto$mpg`, and the option `col=Auto$origin`. By using this option, data points with different levels of `origin` are displayed by different colors. (*You can also add legend in the plot if you want. For more information, see* https://r-charts.com/correlation/scatter-plot-group/.)

4. Add fitted lines of three simple linear regression models by using `abline()` function. You can also specify colors of these lines by using the option `col="colorName"`.

5. Describe what you found from the scatter plot, along with parameter estimates outcome from part (i).