Home Work # 4. AMS 380

Name:__________________________________SBU ID:__________________________________

Dear all, the homework is due on Tuesday, Sep 21, 2021, at 11:59 PM. Please submit your homework to the Blackboard in a pdf or word (.doc) document. Rmarkdown is highly recommended.

Please include (1) R code; (2) Output from R; (3) Answers to all the questions asked.

1. Please write up the entire R code necessary to answer the following questions. Please refer to the following website for the R procedures:
   http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

   First, you need to install the car package and load the data:

   ```
   install.packages("car")
   library(carData)
   # Load the data
   data("Salaries", package = "carData")
   # Inspect the data
   Salaries
   ```

(a) Please fit the general linear model with the response variable being 'salary' and a single predictor being 'sex'. Please write down your regression model and point out which gender group is the baseline group. Is gender a significant predictor of salary based on your analysis? Please report the p-value.

(b) Now you will note that the default baseline group used in part (a) for the categorical variable 'sex' is the "female" group. Can you rerun part (a) with the "male" group as the baseline?

(c) The categorical variable 'rank' in the Salaries data has three levels: "AsstProf", "AssocProf" and "Prof". Now, please fit the general linear model with the response variable being 'salary' and the predictors being 'yrs.service', 'rank', 'discipline' and 'sex'. Please use the Anova() function [in car package] to show the p-values of each variable. Which variables are significant at the significance level of $\alpha = 0.05$? Please use the summary() function to write down the entire regression equation.

(d) Report the coefficient of determination for the model in part (c) above – does this statistic indicate a good linear model fit?

(e) What assumptions are necessary for the regression in part (c) above? Please test these assumptions.

2. Now we learn how to do variable selection using the best subset method, and the stepwise variable selection method. We shall use the built-in R dataset 'swiss'.
   Please write up the entire R code necessary to answer the following questions.
   Please refer to the following websites for the R procedures:
   http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/
   http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/154-stepwise-regression-essentials-in-r/

First, please install the following packages, and load the data:

```
install.packages("tidyverse")
install.packages("caret")
install.packages("leaps")
install.packages("MASS")
library(tidyverse)
library(caret)
library(leaps)
library(MASS)

# Load the data
data("swiss")
# Inspect the data
sample_n(swiss, 3)
```

(a) We shall use ‘Fertility’ as the response variables, and there are a total of 5 repressors to choose from. First, please use the R function regsubsets() [leaps package] for best-subset variable selection to identify different best models of different sizes ranging from 1 to 5.

(b) Please use the 5-fold cross-validation to select the best overall model from all 5 best subset models identified in part (a) above. Please write down the equation of this best overall model.

(c) Please use the R function stepAIC() [MASS package] to identify the best model using the stepwise variable selection method. Please write down the equation of this best overall model.

3. Scientists wish to analyze the effect of fertilizer type on crop yield. The dataset ‘crop.data.csv’ tabulates crop yields from 3 different fertilizers.
Please write up the entire R code necessary to answer the following questions.
You may refer to the following website for the R procedures:
http://www.sthda.com/english/wiki/one-way-anova-test-in-r
http://www.sthda.com/english/wiki/compare-multiple-sample-variances-in-r

(a) Please draw side-by-side box plots to visually compare the yields from the three fertilizers.

(b) Test at $\alpha = 0.05$ whether the three fertilizers are equally effective. What assumptions are necessary? Please test these assumptions.

(c) At the familywise error rate of $\alpha = 0.05$, please perform pairwise comparison of the three fertilizers using the Tukey HSD test.

(d) Please compare fertilizers 2 and 3 using the usual pooled-variance t-test at the significance level $\alpha = 0.05$. What assumptions are necessary? Please test these assumptions.