HOMEWORK 4 MATH 484-564, REGRESSION

DUE SEPTEMBER 27TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) You've already learned how to import and analyze CSV data sets in R. R also includes built-in data sets that you can use. In this task, we'll explore one of these data sets, called mtcars, for analysis.

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

Note

Henderson and Velleman (1981) comment in a footnote to Table 1: 'Hocking [original transcriber]'s noncrucial coding of the Mazda's rotary engine as a straight six-cylinder engine and the Porsche's flat engine as a V engine, as well as the inclusion of the diesel Mercedes 240D, have been retained to enable direct comparisons to be made with previous analyses.'

Above, you can find information about the mtcars data set. We will focus on specific variables rather than all the available ones. Using the piping function % > % , you'll select the variables mpg, cyl, disp, hp, and wt, and save them to my_data. The code for doing that using R Markdown is shown below. Note: You might need to install the necessary library if you haven't done so.

````
## Question 1


```{r}
library("tidyverse")

?mtcars
my_data<-mtcars %>%
  select(mpg, cyl, disp, hp, wt)

model1<-lm(mpg~cyl+disp+hp+wt, data=my_data)
summary(model1)

```
````

Run a regression analysis with mpg as the dependent variable and cyl, disp, hp, and wt as independent variables. Then, include the summary of the regression in your R Markdown.

After obtaining the summary from the regression model (model1), please answer the following questions. Write your answer to this question in the R Markdown document and not on a separate PDF file.

(a) If we choose to regress mpg using only three predictors instead of the four from model 1 (cyl, disp, hp, and wt), which variable would you remove from the analysis, and why?

(b) Rerun the regression using the three variables from Part (a). Label this model as model2 and print the summary. Based on the result, if we have to drop one more predictor from the regression model, which predictor will that be? Why?

(c) Using your answer from (b), create a regression model named model3 with just two predictors. In this model, are all the p values for the predictors considered 'significant'? Print the summary of model3.

(d) Notice that even with only two predictors in (c), the R^2 for model3 is still above 0.8. Give one possible reasons why this is the case.

Remark: The process of eliminating predictors outlined above is known as variable selection.

(2) (6 points) We are once again exploring a built-in dataset from the ISLR2 library. Run the code below in your R Markdown, and you will see the following description of the data. Note the screen shot below only shows part of the description.

```{r}
library("ISLR2")

?Carseats
```

Carseats {ISLR2}

Sales of Child Car Seats

Description

A simulated data set containing sales of child car seats at 400 different stores.

Usage

```
Carseats
```

Format

A data frame with 400 observations on the following 11 variables.

`Sales`

Unit sales (in thousands) at each location

`CompPrice`

Price charged by competitor at each location

`Income`

Community income level (in thousands of dollars)

`Advertising`

Local advertising budget for company at each location (in thousands of dollars)

`Population`

Population size in region (in thousands)

`Price`

Price company charges for car seats at each site

(a) Fit a multiple regression model to predict Sales using Advertising and Price.

(b) Based on your answer in Part (2a), for which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

(c) Would you predict the sales of car seats simply based on the two predictors Advertising and Price? Explain your answer.

3. (6 points)A person's muscle mass tends to decline with age. To investigate this trend in women, a nutritionist randomly selected 15 women from each decade, starting at age 40 and continuing through age 79. The data is stored in a TXT file named MMass. In this file, the first column represents muscle mass, while the second column represents age. Use the following code to read the file, transform the variables, and create a plot of the data (x, y).

```r
# Load the txt file and named it as data1
data1<- read.table("MMass.txt", header = FALSE)

# Showing the first 6 data points
head(data1)

# Rename variables
x<- data1$V2
y<- data1$V1

# Plotting data
plot(x,y)
```

(a) Use the lm() function to compute the regression model. Then, add the regression line to the previous (x, y) plot.

(b) Obtain a point estimate of the average change in the muscle mass for women differing in age by one year.

(c) Obtain a point estimate of σ^2, where σ^2 is the constant variance of ϵ in the following model

$$Muscle\ mass = \beta_0 + \beta_1 Age + \epsilon.$$

(d) Does the linear regression model seem to provide a good fit for the data? Does your plot support the expectation that muscle mass decreases with age? Use the information from the summary function to support your argument.

4. (6 points) For a simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, n$$

show that R^2 is equal to r^2, where r is the correlation between x and y given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

5. (6 points) Show that when moving from a simple linear regression model with one predictor to a multiple linear regression model with two predictors, the R^2 (coefficient of determination) will either increase or stay the same. Explain your reasoning using the definition of R^2 and the effect of adding a predictor on the Residual Sum of Squares (RSS).