STAT 40001/STAT 50001 **Statistical Computing** **Fall 2024**

Homework 3-Solution

Name:

Instruction: Please submit your R code along with a brief write-up of the solutions (do not submit raw output containing ERRORs). Some of the questions below can be answered with very little or no programming. However, write code that outputs the final answer and does not require any additional paper calculations.

Q.N. 1) The R data frame "HairEyeColor" contains classifications of 592 students by gender, hair color and eye color.
a) Is hair color independent of eye color for men?
b) Is hair color independent of eye color for women?
We use R code below to extract the data

```
> data("HairEyeColor")
> menData = HairEyeColor[,,1]
> womenData = HairEyeColor[,,2]
```

Alternately, we can read the data as below

```
> data("HairEyeColor")
> Data=data.frame(HairEyeColor)
> head(Data,3)
   Hair   Eye  Sex Freq
1 Black Brown Male   32
2 Brown Brown Male   53
3   Red Brown Male   10
```

a) In order to investigate whether hair color is independent of eye color for men we test the following hypotheses

$$\boldsymbol{H_0} \ : \ \textit{Hair colors and eye colors are independent}$$
$$\boldsymbol{H_a} \ : \ \textit{Hair colors and eye colors are not independent}$$

We use R code below:

```
> chisq.test(menData)

        Pearson's Chi-squared test

data:  menData
X-squared = 41.28, df = 9, p-value = 4.447e-06

Warning message:
In chisq.test(menData) : Chi-squared approximation may be incorrect
> chisq.test(menData,simulate.p.value = T)
 Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  menData
X-squared = 41.28, df = NA, p-value = 0.0004998
```

The p-value (0.0004998) is much smaller than 0.05 so the null hypothesis is rejected and we conclude that the eye colors and hair colors are not independent.

b) In order to investigate whether hair color is independent of eye color for women we test the following hypotheses

$$\boldsymbol{H_0} \ : \ \textit{Hair colors and eye colors are independent}$$
$$\boldsymbol{H_a} \ : \ \textit{Hair colors and eye colors are not independent}$$

We use R code below:

```
> chisq.test(womenData)

        Pearson's Chi-squared test

data:  womenData
X-squared = 106.66, df = 9, p-value < 2.2e-16

Warning message:
In chisq.test(womenData) : Chi-squared approximation may be incorrect
> chisq.test(womenData,simulate.p.value = T)

        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)

data:  womenData
X-squared = 106.66, df = NA, p-value = 0.0004998
```

The p-value (0.0004998) is much smaller than 0.05 so the null hypothesis is rejected and we conclude that the eye colors and hair colors are not independent.

Q.N. 2)A clinical dietician wants to compare two different diets, A and B, for diabetic patients. She hypothesizes that diet A (Group 1) will be better than diet B (Group 2), in terms of lower blood glucose. She plans to get a random sample of diabetic patients and randomly assign them to one of the two diets. At the end of the experiment, which lasts 6 weeks, a fasting blood glucose test will be performed on each patient. She also expects that the average difference in blood glucose measure between the two group will be about 10 mg/dl. Furthermore, she also assumes the standard deviation of blood glucose distribution for diet A to be 15 and the standard deviation for diet B to be 17. How many subjects are needed in each group assuming equal sized groups? (Please use $\boldsymbol{\alpha = 0.05}$ and Power=0.8).

Solution: Since what really matters is the difference, instead of means for each group, we can enter a mean of 10 for Group 1 and 0 for the mean of Group 2, so that the difference in means will be 10. Next, we need to specify the pooled standard deviation, which is the square root of the average of the two standard deviations squared (i.e., variances). In this case, it is $\sqrt{\frac{15^2+17^2}{2}} = \mathbf{16.03.}$ *Therefore,*

```
> library(pwr)
> pwr.t.test(d=(10-0)/16.03,power=.8,sig.level=.05,type="two.sample", alt="greater")

     Two-sample t test power calculation

              n = 32.47164
              d = 0.6238303
      sig.level = 0.05
          power = 0.8
    alternative = greater

NOTE: n is number in *each* group
```

Based on the R output, it is required that there are at lest 33 individuals in each group.

Q.N. 3) Data on the average SAT scores for US states in 1982 and possible associated factors are provided with this assignment. The variables are listed below
State: US state
SAT: state averages of the total SAT (verbal + quantitative) scores
Takers: the percentage of all eligible students (high school seniors) in the state who took the exam.
Income: the median income of families of test–takers (in hundreds of dollars)
Years: average number of years that the test–takers had formal studies
Public: the percentage of the test–takers who attended public secondary schools
Expend: the total state expenditure on secondary schools (in hundreds of dollars per student)
Rank: the median percentile ranking of the test–takers within their secondary school classes

a) Import the dataset in R and determine its dimension.
b) Identify the state with the highest total SAT score and the lowest total SAT score.
c) Create a correlation matrix for this data set.
d) Display the correlation matrix using a correlogram. Please customize the graph using different customization options as discussed in the class.

Solution:
a) We used the R code below to import the data and determine its dimension. It appears that there are 50 observations with 8 variables

```
> library(readxl)
> Q3=read_excel("C:\\Users\\aryalg\\statesatscores.xls")
> head(Q3)
# A tibble: 6 × 8
  State         SAT Takers Income Years Public Expend  Rank
  <chr>       <dbl>  <dbl>  <dbl> <dbl>  <dbl>  <dbl> <dbl>
1 Iowa         1088      3    326  16.8   87.8   25.6  89.7
2 SouthDakota  1075      2    264  16.1   86.2   20.0  90.6
3 NorthDakota  1068      3    317  16.6   88.3   20.6  89.8
4 Kansas       1045      5    338  16.3   83.9   27.1  86.3
5 Nebraska     1045      5    293  17.2   83.6   21.0  88.5
6 Montana      1033      8    263  15.9   93.7   29.5  86.4
> dim(Q3)
[1] 50  8
```

b) Based on the R output below, Iowa and South Carolina have been identified as the states with the highest and lowest SAT scores, respectively.

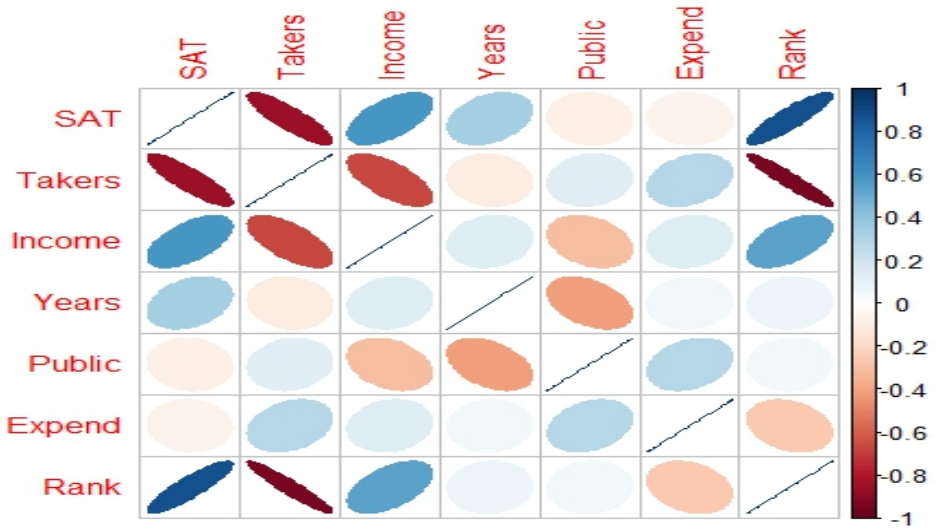
```
> attach(Q3)
> Q3$State[which.max(SAT)]
[1] "Iowa"
> Q3$State[which.min(SAT)]
[1] "SouthCarolina"
```

c) The correlation matrix is provided below.

```
> cor(Q3[,-1])
               SAT     Takers     Income       Years      Public      Expend        Rank
SAT     1.00000000 -0.8578100  0.5844666  0.33096873 -0.08035685 -0.06287764  0.87990911
Takers -0.85780996  1.0000000 -0.6619351 -0.10154338  0.12355623  0.28363042 -0.94283310
Income  0.58446657 -0.6619351  1.0000000  0.13476223 -0.30656698  0.13151940  0.53269990
Years   0.33096873 -0.1015434  0.1347622  1.00000000 -0.41711828  0.05982859  0.07022351
Public -0.08035685  0.1235562 -0.3065670 -0.41711828  1.00000000  0.28459119  0.05062356
Expend -0.06287764  0.2836304  0.1315194  0.05982859  0.28459119  1.00000000 -0.26496898
Rank    0.87990911 -0.9428331  0.5326999  0.07022351  0.05062356 -0.26496898  1.00000000
```

d) We used “corrplot” function available in “corrplot” library to create the correlogram

```
> library(corrplot)
> corrplot(cor(Q3[,-1]), method="ellipse")
```

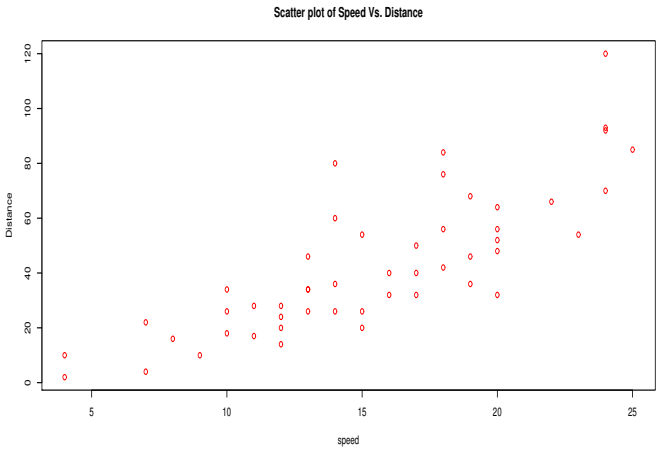


Q.N. 4) The “cars” data set is one of the R-installed data sets and is available in the base package. The data set contains 50 observations of speed(mph) and dist(stopping distance in feet).

a) Display the data using scatter plot.

Solution: We will read and plot the scatter plot of the data using R code below:

```
> data=cars
> attach(cars)
> names(cars)
[1] "speed" "dist"
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
```



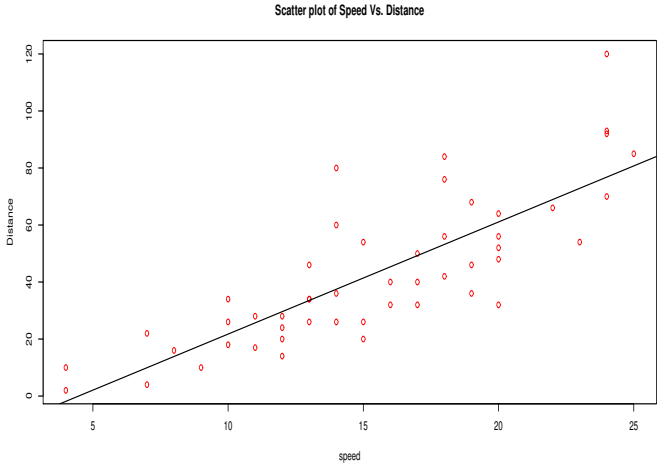
b) Fit a simple regression model using speed as a predictor variable.
Solution: Use R code below to estimate the parameters

```
> model=lm(dist~speed)
> model
Call:
lm(formula = dist ~ speed)
Coefficients:
(Intercept)        speed
    -17.579        3.932
```

Therefore, the fitted model is $\boldsymbol{distance = -17.579 + 3.932 \times speed}$

c) Add the fitted line to the scatter plot.
Solution: We can add the fitted line to the scatter plot using the code below:

```
> plot(speed,dist, main= "Scatter plot of Speed Vs. Distance", ylab="Distance", col=2)
> model=lm(dist~speed)
> abline(model)
```

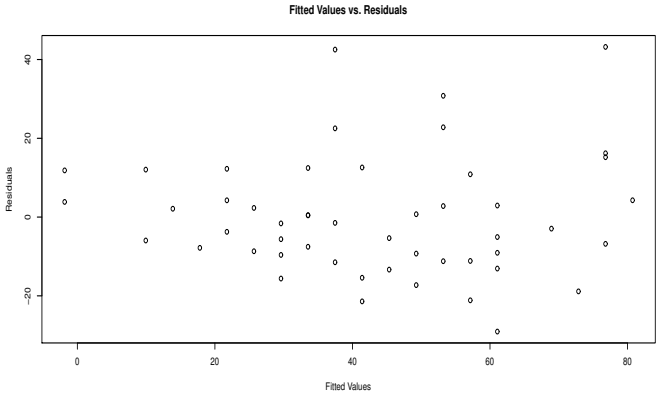


d) Calculate the residuals and fitted values and print only first five observations of the residuals and fitted values.
Solution: The fitted value and residuals of the model are calculated using the R code below:

```
> model=lm(dist~speed)
> fitted=fitted(model)
> residuals=resid(model)
> head(fitted,5)
        1         2         3         4         5
-1.849460 -1.849460  9.947766  9.947766 13.880175
> head(residuals,5)
        1         2         3         4         5
 3.849460 11.849460 -5.947766 12.052234  2.119825
```

e) Create a scatter plot of the residuals and fitted values.
Solution: R code below is used to create the scatter plot of the fitted value and residuals.

```
> model=lm(dist~speed)
> fitted=fitted(model)
> residuals=resid(model)
> plot(fitted,residuals,xlab="Fitted Values",ylab="Residuals",main="Fitted Values vs. Residuals")
```



f) Assuming that no intercept model is appropriate fit a simple linear regression model.
Solution: No-intercept model can be fitted using the R code below:

```
> model1=lm(dist~-1+speed)
> model1
Call:
lm(formula = dist ~ -1 + speed)
Coefficients:
speed
2.909
```

Hence, the fitted model is $\boldsymbol{Distance = 2.909 \times Speed}$
g) Calculate and compare the coefficient of determination for both the with intercept and no-intercept models.
Solution: In order to calculate the coefficient of determination we use the R code below:

```
> summary(model)
> summary(model1)
```

Note that we have the following values for the coefficient of determination:

Model	$\boldsymbol{R^2}$	Adjusted$\boldsymbol{R^2}$
Intercept Model	0.6511	0.6438
No-intercept	0.8963	0.8942

h) Using your fitted model predict the stopping distance for a car with an speed of 21 mph.

Solution: We can use R code below to predict the stopping distance for a car with an speed of 21 mph

```
> model=lm(dist~speed)
> model1=lm(dist~-1+speed)
> xval=as.data.frame(21)
> colnames(xval)="speed"
> predict(model,xval)
       1
65.00149
> predict(model1,xval)
       1
61.09178
```

Note that the intercept model predicts a stopping distance of 65.00149 feet and the no-intercept model predicts a stopping distance of 61.09178 feet.

Q.N. 5) Suppose that a fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The study is conducted in a suburb of a major metropolitan area. The data collected were the distance in miles between the nearest fire station and the fire and the amount of damage to the house (in thousands of dollars).

Distance	Damage
3.4	26.2
1.8	17.8
4.6	31.3
2.3	23.1
3.1	27.5
5.5	36.0
0.7	14.1
3.0	22.3
2.6	19.6
4.3	31.3
2.1	24.0
1.1	17.3
6.1	43.2
4.8	36.2
3.8	26.5

a) Fit a simple linear regression model and analyze the residual plots.
b) What is the expected Damage if the fire station is 4 miles away?
c) Use the Box-Cox transformation to choose an appropriate value of $\boldsymbol{\lambda}$ to improve the model.
d) Fit a simple linear regression model after transformation.
e) Compare and contrast models in (a) and (d).

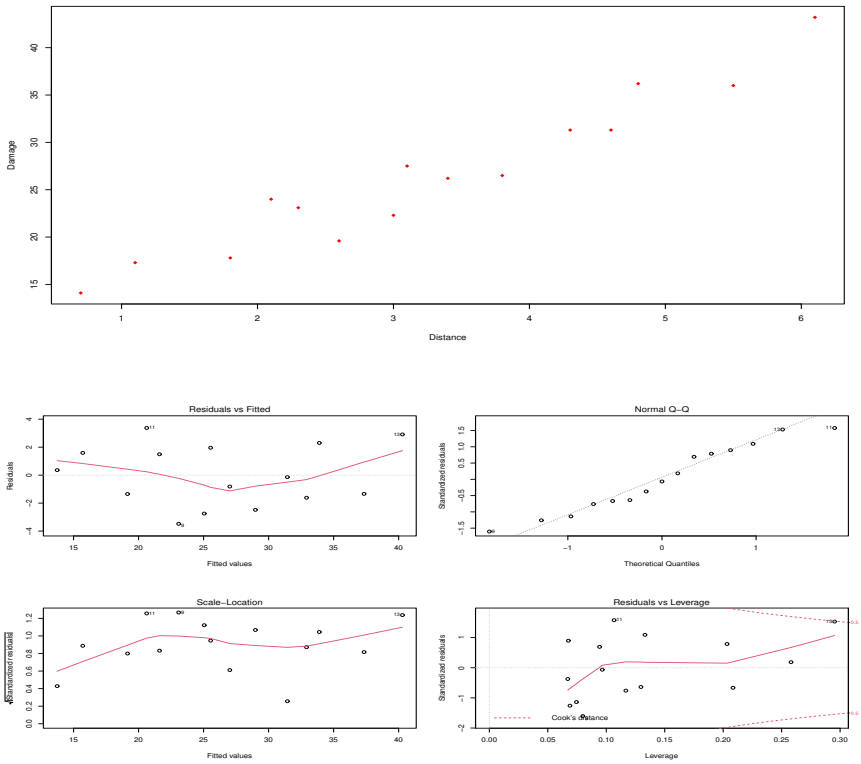
Solution: We use R code below to develop the model

```
> Distance
 [1] 3.4 1.8 4.6 2.3 3.1 5.5 0.7 3.0 2.6 4.3 2.1 1.1 6.1 4.8 3.8
> Damage
 [1] 26.2 17.8 31.3 23.1 27.5 36.0 14.1 22.3 19.6 31.3 24.0 17.3 43.2 36.2 26.5
> plot(Distance, Damage, col="red", pch=18)
> model=lm(Damage~Distance)
> model

Call:
lm(formula = Damage ~ Distance)

Coefficients:
(Intercept)     Distance
     10.300        4.917
```

The fitted model is

$$\boldsymbol{Damage = 10.300 + 4.917 \times Distance}$$

The residual plot shows a structured pattern.

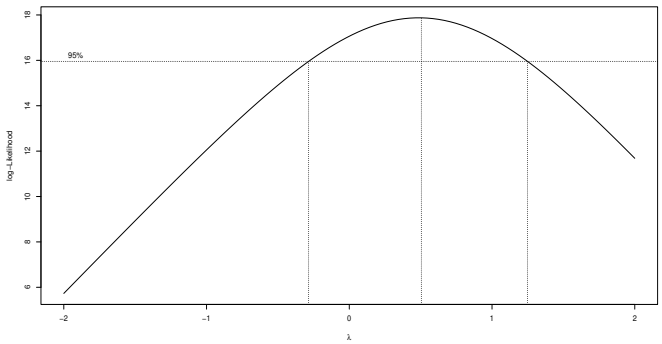


b) Based on the R output below the expected damage is **$29,666**.

```
> predict(model, data.frame(Distance=4))
      1
29.9666
```

c) Based on the R output below $\boldsymbol{\lambda = 0.5}$.

```
> library(MASS)
> b=boxcox(model)
```



d) We will transform the data and fit a linear regression model

```
> y=Damage^0.5
> newmodel=lm(y~Distance)
> newmodel
Call:
lm(formula = y ~ Distance)
Coefficients:
(Intercept)     Distance
     3.5183       0.4777
```

The fitted model is

$$\begin{aligned}(\boldsymbol{Damage})^{\mathbf{0.5}} &= \mathbf{3.5183 + 0.4777 \times \boldsymbol{Distance}} \\ \boldsymbol{Damage} &= \mathbf{(3.5183 + 0.4777 \times \boldsymbol{Distance})^2}\end{aligned}$$

e) It appears that the coefficient of determination is slightly increase. Moreover, the residual plots shows a structurless behaviours.

```
> summary(model)$r.square
[1] 0.9265571
> summary(newmodel)$r.square
[1] 0.9315393
```