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

Homework 4-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 data set *Cars93* provided in the library MASS contains data on cars sold in the United States in the year 1993.

a) How many variables are included in the data set?
Solution: Based on the R code below it appears that there are 27 variables included in the dataset.

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> dim(Cars93)
[1] 93 27
```

b) Fit a regression model for `MPG.city` using the numerical variables `EngineSize`, `Weight`, `Passengers`, and `Price`.
Solution: We use R code below to estimate the model parameters

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> model=lm(MPG.city~EngineSize+Weight+Passengers+Price)
> model

Call:
lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)

Coefficients:
(Intercept)   EngineSize       Weight   Passengers        Price
  46.389413     0.196119    -0.008207     0.269622    -0.035804
```

Hence, the desired regression model is

$$\boldsymbol{MPG.city = 46.3894 + 0.1961 \times EngineSize - 0.0082 \times Weight + 0.2696 \times Passengers - 0.0358 \times Price}$$

c) Which variables are marked as statistically significant by the marginal t-test in model(b)?

Solution: We use R code below to test the significance of the model parameters

```
> library(MASS)
> data(Cars93)
> attach(Cars93)
> model=lm(MPG.city~EngineSize+Weight+Passengers+Price)
> summary(model)
Call:
lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-6.1207 -1.9098  0.0522  1.1294 13.9580

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.389413   2.097516  22.116  < 2e-16 ***
EngineSize   0.196119   0.588880   0.333    0.740
Weight      -0.008207   0.001343  -6.111 2.63e-08 ***
Passengers   0.269622   0.424951   0.634    0.527
Price       -0.035804   0.049179  -0.728    0.469
---

Residual standard error: 3.06 on 88 degrees of freedom
Multiple R-squared: 0.7165,     Adjusted R-squared: 0.7036
F-statistic: 55.59 on 4 and 88 DF,  p-value: < 2.2e-16
```

It appears that the Weight is only significance variable to determine the City MPG.

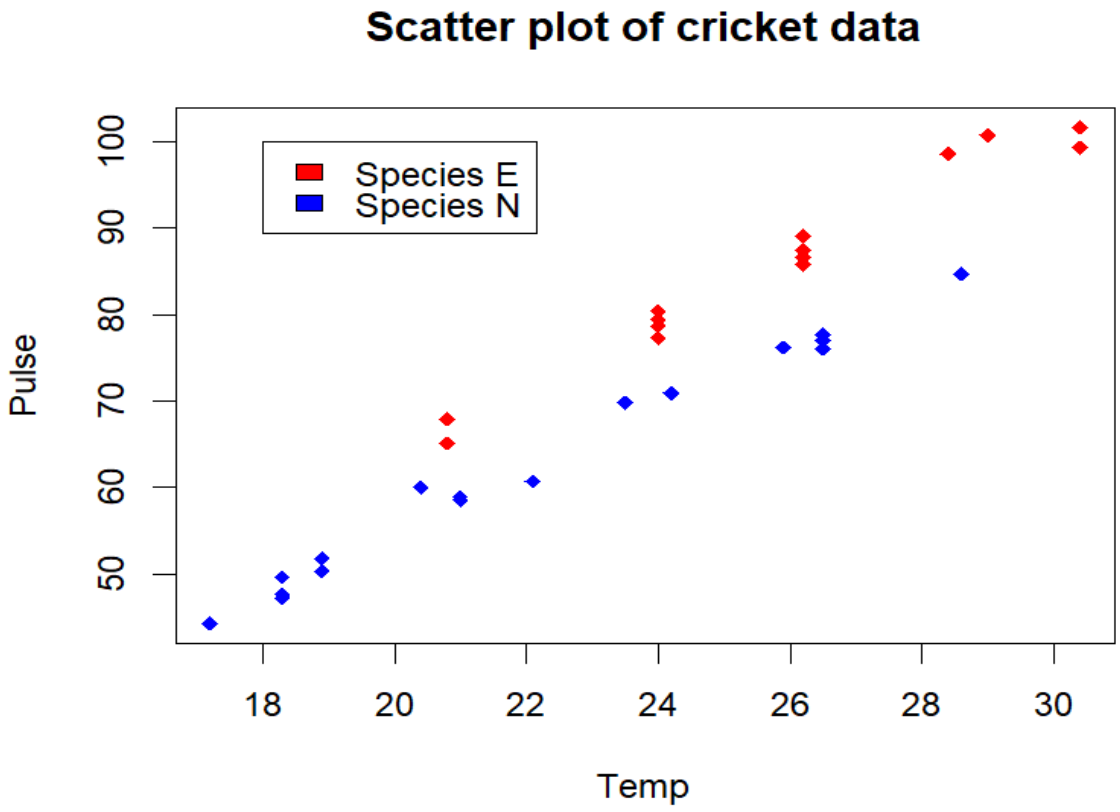
Q.N. 2) Walker (1962) studied the mating songs of male tree crickets. Each wingstroke by a cricket produces a pulse of song, and females may use the number of pulses per second to identify males of the correct species. Walker (1962) wanted to know whether the chirps of the crickets Oecanthus exclamationis (E) and Oecanthus niveus (N) had different pulse rates. He measured the pulse rate of the crickets at a variety of temperatures, and the data are provided in the Brightspace with this assignment.
a) Import the dataset in R.
b) Display the pulse rate of the crickets at different temperature using a scatterplot. Please be sure to use different color or point characters to distinguish the pulse rate of each species.
c) Fit a multiple linear regression model using the temperature and species as a predictor variables and Pulse rate as a response variable.
d) Display the fitted models for both species in the scatterplot.
e) Estimate the pulse rate of a Oecanthus niveus at 26.5 degree Celsius. Also construct a 95% confidence interval for your predicted value.

Solution:
a) We used R code below to import the data in R. Note that there are 31 observations with three variables.

```
> data=read.csv("C:\\Users\\aryalg\\cricket.csv")
> attach(data)
> head(data)
  Temp Pulse Species
1 20.8  67.9       E
2 20.8  65.1       E
3 24.0  77.3       E
4 24.0  78.7       E
5 24.0  79.4       E
6 24.0  80.4       E
> dim(Q1)
[1] 31  3
```

b) We used R code below to create the graph

```
> plot(Temp,Pulse,col=ifelse(Species=="E","red","blue"),pch=18,main="Scatter plot of cricket data")
> legend(18,100, legend=c("Species E", "Species N"), fill=c("red", "blue"))
```



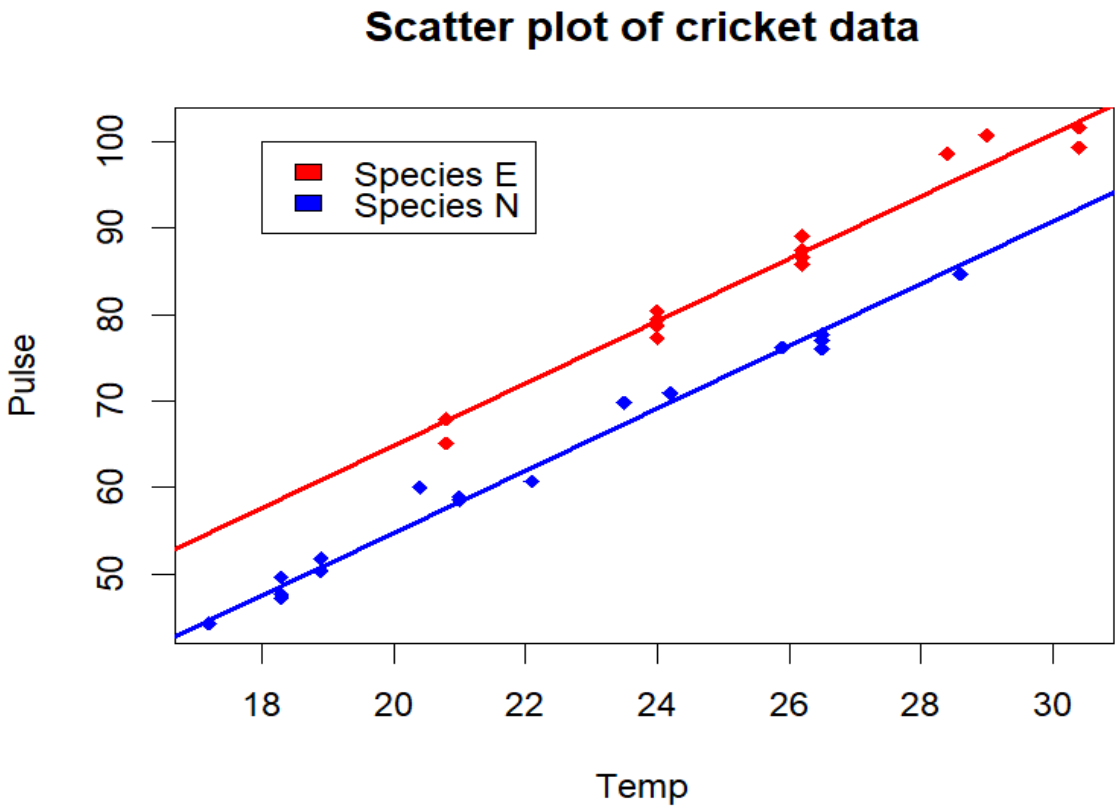
c) The fitted model is given by

$$\hat{Pulse} = \begin{cases} -7.211 + 3.603 \times Temp, & \textit{for Species E,} \\ -17.276 + 3.603 \times Temp & \textit{for Species N} \end{cases}$$

```
> model=lm(Pulse  ~Temp+Species, data=data)
> model

Call:
lm(formula = Pulse ~ Temp + Species, data = data)

Coefficients:
(Intercept)         Temp     SpeciesN
     -7.211        3.603      -10.065
```

d) We used R code below to create and diplay the fitted models

```
> plot(Temp,Pulse,col=ifelse(Species=="E","red","blue"),pch=18,main="Scatter plot of cricket data")
> legend(18,100,legend=c("Species E", "Species N"), fill=c("red", "blue"))
> abline(-7.211,3.603, col="red", lwd=2)
> abline(-17.276, 3.603, col="blue", lwd=2)
```



e) Based on the R output below, the pulse rate of a Oecanthus niveus at 26.5 degree Celsius is **78.19675** with 95% confidence interval **(76.95245, 79.44106).**

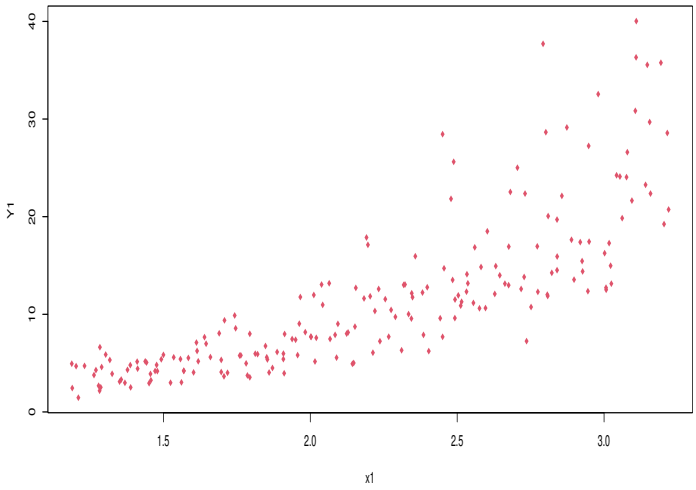
```
> predict(model, data.frame(Temp=26.5, Species="N"))
       1
78.19675
> predict(model, data.frame(Temp=26.5, Species="N"), interval="conf")
       fit      lwr      upr
1 78.19675 76.95245 79.44106
>
```

Q.N. 3) The dataset "SimDataST" in "PASWR" package in R contains data frame with 200 observations on the following 5 numeric variables: $\boldsymbol{Y_1, Y_2, x_1, x_2}$ and $\boldsymbol{x_3}$.
a) Import the data into R and draw a scatter plot using variables $\boldsymbol{x_1}$ and $\boldsymbol{Y_1}$.
b) Fit a simple linear regression model using $\boldsymbol{Y_1}$ as response variable and $\boldsymbol{x_1}$ as regressor variable and assess the residual plots of the model.
c) Determine the value of $\boldsymbol{\lambda}$ using Box-Cox transformation to improve the model.
d) Fit a simple linear regression model after transformation.
e) Compare the results in (b) and (d). Was the transformation worth it?

Solution:
a) We use R code below to access the data and display the scatterplot

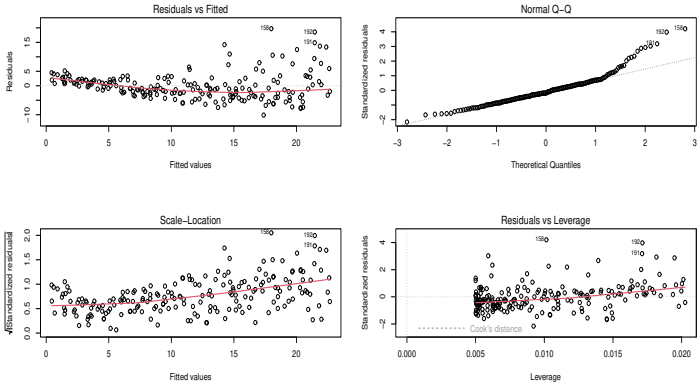
```
> library(PASWR)
> data(SimDataST)
> head(SimDataST,3)
        Y1        Y2       x1       x2       x3
```

```
1 4.941516 0.6605367 1.188179 1.204445 1.454922
2 2.445700 0.6864355 1.189769 1.211087 1.470046
3 4.688932 0.9097890 1.202898 1.233897 1.535170
> attach(SimDataST)
> plot(x1,Y1, col=2, pch=18)
```



b) Based on the R code below, the fitted model is $\hat{\boldsymbol{Y}}_{\mathbf{1}} = -\mathbf{12.57} + \mathbf{10.95}\boldsymbol{x}_{\mathbf{1}}$.

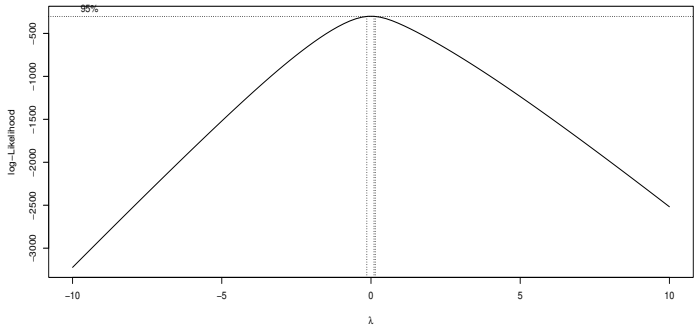
```
> model=lm(Y1~x1)
> model
Call:
lm(formula = Y1 ~ x1)
Coefficients:
(Intercept)           x1
     -12.57        10.95
> par(mfrow=c(2,2))
> plot(model)
```



The residual plot shows a clear pattern in the residuals.

c) Based on the R output below we have $\boldsymbol{\lambda = 0}$. This means a log tranformation will be appropriate.

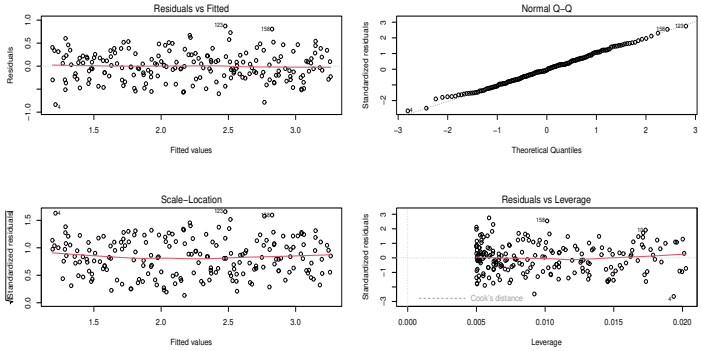
```
> library(MASS)
> boxcox(model, lambda = seq(-10,10,1))
```



d) We use R code below to fit a new model

```
 > y=log(Y1)
> newmodel=lm(y~x1)
> newmodel

Call:
lm(formula = y ~ x1)

Coefficients:
(Intercept)           x1
   -0.01783      1.01800
```

e) It can be observed that after transformation the residuals are structure less and the the coefficient of determination is higher. Therefore, the transformation was worth it.



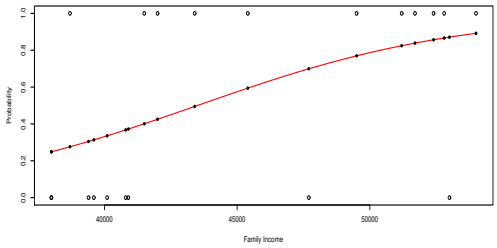

```
 > summary(model)$r.squared
[1] 0.6504271
> summary(newmodel)$r.squared
[1] 0.7792601
```

Q.N. 4) A study was conducted attempting to relate home ownership to family income. Twenty households were selected and family income (x) was estimated along with information on home ownership (y = 1 indicates yes and y = 0 indicates no). The data are provided in Brightspace with this assignment.
a) Fit a simple logistic regression model for the subject data and state the logistic regression model.
b) Display the fitted model in the scatterplot.
c) What is and estimated probability that a family with an income of $45, 000 owns a house?

Solution: We imported the data in R and used R code below to perform all the tasks. The fitted model is

$$\hat{\pi} = [1 + exp(8.7395 - 0.00002 * x)]^{-1}$$

```
> model=glm(y~x,family=binomial)
> summary(model)
Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0232  -0.8766   0.5072   0.7980   1.6046

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.7395139  4.4394326  -1.969   0.0490 *
x            0.0002009  0.0001006   1.998   0.0458 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on 19  degrees of freedom
Residual deviance: 22.435  on 18  degrees of freedom
AIC: 26.435

> plot(x,y, xlab="Family Income", ylab="Probability")
> curve(predict(model,data.frame(x=x),type="resp"),add=TRUE, col=2, lwd=2)
> points(x,fitted(model),pch=20)
```



b) Based on the R output below, the probability that a family with an income of $45, 000 owns a house is 0.575.

```
> predict(model,data.frame(x=45000), type="resp")
        1
0.5747456
```

Q.N. 5) According to the web site `http://www.keepkidshealthy.com`, risk factors associated with premature births include smoking and maternal malnutrition. A birth is considered premature if the gestation period is less than 37 full weeks. Also note that the body mass index(BMI) can be used as a measure of malnutrition. Do you find this to be true with the data in babies provided in the `UsingR` package?
Tasks to perform:
a) Extract the variables of interest: gestation, smoking status, mother's height and weight, and birth weight of the babies.
b) Clean the data set as there are some missing values coded as 9, 99, or 999.
c) Calculate the BMI of mothers.
d) Create indicator variable(1 for premature and 0 for not premature) babies.
e) Fit a logistic regression model with `smoke` and `BMI` as a predictor variable and `premature` as response variable.

Solution: We use R code below to extract the variables of interest

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> dim(data)
[1] 1236    5
> head(data,5)
  gestation smoke wt1 ht  wt
1       284     0 100 62 120
2       282     0 135 64 113
3       279     1 115 64 128
4       999     3 190 69 123
5       282     1 125 67 108
```

We clean the data set using the following R code:

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> dim(Clean)
[1] 1175    5
```

We calculate the BMI of mothers using R code below

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> attach(Clean)
> BMI=wt1/(ht)^2*703
> BMI[1:10]
 [1] 18.28824 23.17017 19.73755 19.57563 17.00806 32.55307 23.29467 22.86030
 [9] 21.94858 18.24394
```

We create an indicator variable premature using the R code below.

```
> library(UsingR)
> data(babies)
> data=subset(babies,select=c("gestation","smoke","wt1","ht","wt"))
> Clean=subset(data, gestation !=999&smoke!=9 & wt1!=999 & ht!=99 & wt!=999)
> dim(Clean)
[1] 1175    5
```

```
> preemie=as.numeric(Clean$gestation<7*37)
> table(preemie)
preemie
   0    1
1079   96
```

We can now model the variable preemie by the levels of smoke and the variable BMI.

```
> model=glm(preemie~factor(Clean$smoke)+BMI, family=binomial)
> summary(model)
Call:
glm(formula = preemie ~ factor(Clean$smoke) + BMI, family = binomial)
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.42458    0.71159  -4.813 1.49e-06 ***
factor(Clean$smoke)1  0.19353    0.23569   0.821    0.412
factor(Clean$smoke)2  0.31370    0.38896   0.806    0.420
factor(Clean$smoke)3  0.10114    0.40499   0.250    0.803
BMI                   0.04023    0.03050   1.319    0.187
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 664.83  on 1174  degrees of freedom
Residual deviance: 662.34  on 1170  degrees of freedom
AIC: 672.34
Number of Fisher Scoring iterations: 5
```

Note that none of the variables are flagged as significant. This indicates that the model with no effect is, perhaps, preferred. In order to check which model is preferred by AIC we use R code below

```
> library(MASS)
> stepAIC(model)
Start:  AIC=672.34
preemie ~ factor(Clean$smoke) + BMI
                      Df Deviance    AIC
- factor(Clean$smoke)  3   663.35 667.35
- BMI                  1   663.98 671.98
<none>                     662.34 672.34
Step:  AIC=667.35
preemie ~ BMI
       Df Deviance    AIC
- BMI   1   664.83 666.83
<none>      663.35 667.35

Step:  AIC=666.83
preemie ~ 1
Call:  glm(formula = preemie ~ 1, family = binomial)
Coefficients:
(Intercept)
     -2.419
Degrees of Freedom: 1174 Total (i.e. Null);  1174 Residual
Null Deviance:      664.8
Residual Deviance: 664.8        AIC: 666.8
```

Since, the model of constant mean is chosen, this data indicate that neither smoking status nor BMI are the risk factors for premature babies.