STAT 43000/STAT 53001 Applied Statistics Spring 2023

Test 1-Part II

Due Date: March 8, 2023
Time: 2:00 PM (CST)

Name:
PUID:

Q.N. 1) A data set containing data on 40 foot and height measurements of human is provided in the Brightspace (**Foot measurement**). This data is from "Estimation of Stature from Foot Length: Applications in Forensic Science.

a) Import the data in R and print the first 5 observations.

```
> data=read.csv("C:\\Users\\Zhang\\Downloads\\Test 1 Part II attached files Mar 8, 2023 1232 PM\\Foot data.xlsx")
Error in file(file, "rt") : cannot open the connection
In addition: Warning message:
In file(file, "rt") :
  cannot open file 'C:\Users\Zhang\Downloads\Test 1 Part II attached files Mar 8, 2023 1232 PM\Foot data.xlsx': No s
> file.choose()
[1] "C:\\Users\\Zhang\\Downloads\\Test 1 Part II attached files Mar 8, 2023 1232 PM\\Foot data.csv"
> data=read.csv("C:\\Users\\Zhang\\Downloads\\Test 1 Part II attached files Mar 8, 2023 1232 PM\\Foot data.csv")
> head(data)
  Sex Age Foot.length Shoe.Print Shoe.size Height
1   M  67        27.8       31.3        11  180.3
2   M  47        25.7       29.7         9  175.3
3   M  41        26.7       31.3        11  184.8
4   M  42        25.9       31.8        10  177.8
5   M  48        26.4       31.4        10  182.3
6   M  34        29.2       31.9        13  185.4
> head(data,5)
  Sex Age Foot.length Shoe.Print Shoe.size Height
1   M  67        27.8       31.3        11  180.3
2   M  47        25.7       29.7         9  175.3
3   M  41        26.7       31.3        11  184.8
4   M  42        25.9       31.8        10  177.8
5   M  48        26.4       31.4        10  182.3
>
```

b) Is there a significant difference in the foot length of male and female?

```
> male_data <- subset(data, Sex== "M")$Foot.length
> female_data <- subset(data, Sex== "F")$Foot.length
> t.test(male_data, female_data)

        Welch Two Sample t-test

data:  male_data and female_data
t = 8.1505, df = 35.92, p-value = 1.101e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.347958 3.903672
sample estimates:
mean of x mean of y
 27.32105  24.19524
```

p value 1.101e-09 << 0.05, so there is a significant difference in the foot length of male and female

c) Fit a simple linear regression model using Foot length as a predictor variable and height as aresponse variable.

```
>ml=lm(Height~Foot.length)
Error in l = lm(Height ~ Foot.length) :
  invalid (do_set) left-hand side to assignment
> ml <- lm(Height ~ Foot.length)
> ml

Call:
lm(formula = Height ~ Foot.length)

Coefficients:
(Intercept)  Foot.length
     64.126        4.291

>
```

Height=4.291*Foot.length+64.126

d) Update the fitted model in (c) by incorporating a binary variable Sex

```
> Sex=(Sex=="M")*1
Error: unexpected input in "Sex=(Sex=="
> Sex=(Sex=="M")*1
> m2 <- lm(Height ~ Foot.length+Sex)
> m2

Call:
lm(formula = Height ~ Foot.length + Sex)

Coefficients:
(Intercept)  Foot.length          Sex
     95.641        2.942        6.596
```

Set M=1 and F=0
So for male, Height=2.942 *Foot.length+95.641+6.596
And for female, Height=2. 942 *Foot.length+95.641

e) Predict the height of a male whose foot is 28.8 cm.

```
> predict(m2,data.frame(Foot.length=28.8,Sex=1))
       1
186.9669
>
```

Q.N. 2) The **leukemia data** provided in the Brightspace provides the information of 27 patients. The response variable of whether leukemia remission occurred (REMISS), which is given by a 1. a) Import the data in R and print the variables.

```
> file.choose()
[1] "C:\\Users\\Zhang\\Downloads\\Test1\\LD.csv"
> data=read.csv("C:\\Users\\Zhang\\Downloads\\Test1\\LD.csv")
> head(data)
  REMISS CELL SMEAR INFIL  LI BLAST TEMP
1      1  0.8  0.83  0.66 1.9  1.10 1.00
2      1  0.9  0.36  0.32 1.4  0.74 0.99
3      0  0.8  0.88  0.70 0.8  0.18 0.98
4      0  1.0  0.87  0.87 0.7  1.05 0.99
5      1  0.9  0.75  0.68 1.3  0.52 0.98
6      0  1.0  0.65  0.65 0.6  0.52 0.98
> names(data)
[1] "REMISS" "CELL"   "SMEAR"  "INFIL"  "LI"     "BLAST"  "TEMP"
>
```

b) Fit a simple logistic regression model using percentage labeling index of the bone marrow leukemiacells (LI) as a predictor variable.

```
> model <- glm(REMISS ~ LI, data = data, family = binomial)
> model

Call:  glm(formula = REMISS ~ LI, family = binomial, data = data)

Coefficients:
(Intercept)           LI
     -3.777        2.897

Degrees of Freedom: 26 Total (i.e. Null);  25 Residual
Null Deviance:      34.37
Residual Deviance: 26.07        AIC: 30.07
```

Π= =[1+exp(3.777-2.897LI)]

c) Calculate the odds ratio for LI.

```
> odds_ratio = exp(coef(model)[2])
> odds_ratio
      LI
18.12449
>
```

d) Calculate the estimated odds of leukemia remission at LI=0.8 and LI=0.9. Now, calculate theodds ratio using the odds at LI= 0.9 and LI=0.8. How do you interpret this value?

```
> odds_08 <- exp(predict(model, data.frame(LI = 0.8)))
> odds_09 <- exp(predict(model, data.frame(LI = 0.9)))
> odds_08
        1
0.2323921
> odds_09
        1
0.3104903
> odds_ra=odds_09/odds_08
> odds_ra
        1
1.336062
>
```

A value greater than 1 suggests that higher values of LI are associated with higher odds of leukemia

Q.N. 3) A data set sexab available in faraway package is related to a study of the effects of childhood sexual abuse on adult females reported by Rodriguez et al. (1997).

a) Install the library faraway and access the data sexab

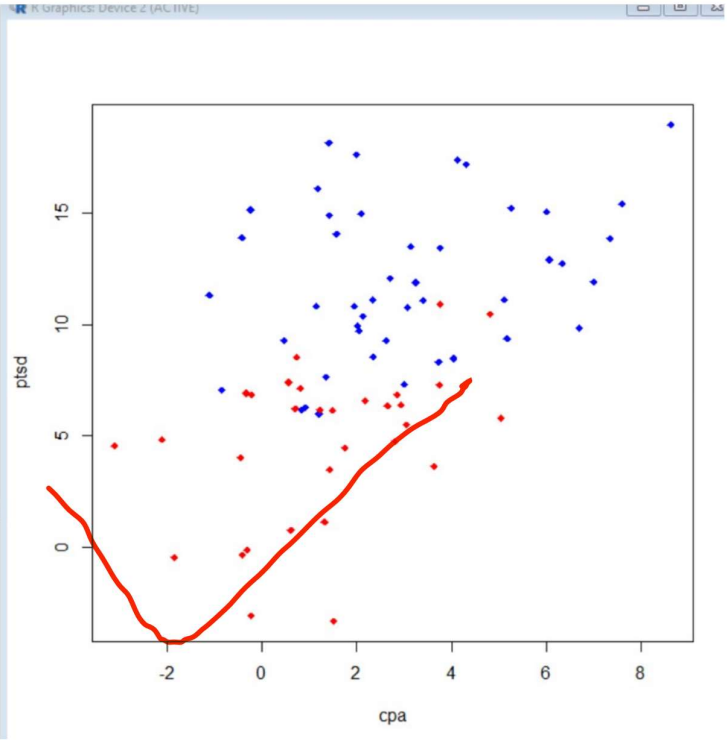
```
> install.packages("faraway")
Installing package into 'C:/Users/Zhang/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cloud.r-project.org/bin/windows/contrib/4.2/faraway_1.0.8.zip'
Content type 'application/zip' length 772008 bytes (753 KB)
downloaded 753 KB

package 'faraway' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Zhang\AppData\Local\Temp\RtmpMh9Pcz\downloaded_packages
> library(faraway)
> data(sexab)
```

```
> sexab
         cpa      ptsd       csa
1    2.04786   9.71365    Abused
2    0.83895   6.16933    Abused
3   -0.24139  15.15926    Abused
4   -1.11461  11.31277    Abused
5    2.01468   9.95384    Abused
6    6.71131   9.83884    Abused
7    1.20814   5.98491    Abused
8    2.34284  11.11053    Abused
9    0.91188   6.25528    Abused
10  -0.85308   7.04109    Abused
```

b) Note that the data include the variables:

cpa-Childhood physical abuse on standard scale *csa*-
Childhood sexual abuse - abused or not abused *ptsd*-
Post-traumatic stress disorder on standard scale
Display the data graphically using different colors to csa variable: abused and not abused

```
58  0.73195  8.53706 NotAbused
59 -0.41639 -0.37549 NotAbused
60  2.80928  4.75267 NotAbused
61  2.93373  6.39374 NotAbused
62 -0.22780  6.83305 NotAbused
63  4.82039 10.46992 NotAbused
64  1.32165  1.11152 NotAbused
65  0.56215  7.40117 NotAbused
66  1.22299  6.17468 NotAbused
67  3.04951  5.49228 NotAbused
68  3.76859 10.91447 NotAbused
69 -2.11876  4.83856 NotAbused
70  3.63574  3.62940 NotAbused
71 -0.31402 -0.14339 NotAbused
72  2.17626  6.57281 NotAbused
73 -0.23208 -3.11622 NotAbused
74 -1.85753 -0.46996 NotAbused
75  2.85253  6.84304 NotAbused
76  0.81138  7.12918 NotAbused
> attach(sexab)
Error: unexpected input in "attach("
> attach(sexab)
> csa=(csa=="Abused")*1
Error: unexpected input in "csa=(csa=="
> csa=(csa=="Abused")*1
> plot(cpa,ptsd,col=ifelse(csa=="0","red","blue"),pch=18)
>
```



c) Fit a linear regression model by choosing ptsd a response variable and using other variables as Predictors

```
> ml=lm(ptsd~cpa+csa)
> ml

Call:
lm(formula = ptsd ~ cpa + csa)

Coefficients:
(Intercept)          cpa          csa
     3.9753       0.5506       6.2728
```

Ptsd=0.5506*cpa+3.9753
(not abused)

Ptsd=0.5506*cpa+3.9753
+6.2728(abused)

Q.N. 4) An economic study followed a British bus company for n = 33 time periods, recording y= Total Expenses (adjusted for inflation in 100,000s of pounds) and x=car miles(in millions). The data are available in the Brightspace (**Bus**)

a) Fit a simple linear regression model relating Total Expenses (y) to car miles (x).

```
> data=read.csv("C:\\Users\\Zhang\\Downloads\\Test1\\bus.csv")
> attach(data)
> ml=lm(expenses~miles)
> ml

Call:
lm(formula = expenses ~ miles)

Coefficients:
(Intercept)        miles
     0.6496       0.4467

> |
```

Expenses=0.4467*miles+0.6496

b) Calculate the value of the Durbin-Watson test statistic. Do we have an evidence of auto correlation at α = 0.05.

```
> library(lmtest)
Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric

> dwtest(ml)

        Durbin-Watson test

data:  ml
DW = 1.1603, p-value = 0.00365
alternative hypothesis: true autocorrelation is greater than 0
```

Pvalue is much small than 0.05

c) Obtain estimates of the $\hat{\rho}$ based on the Cochrane-Orcutt procedure.

```
> library(orcutt)
> orc=cochrane.orcutt(ml)
> orc$rho
[1] 0.3675868
>
```

d) Obtain estimates of $\hat{\rho}$ based on the Hildreth-Lu procedure.

Q.N. 5) The transient points of an electronic inverter data are provided in the Brightspace as **inverter**. The variables under study are

y: Transient point (volts) of PMOS-NMOS inverters
X1: Width of the NMOS device
X2: Length of the NMOS device
X3: Width of the PMOS device
X4: Length of the PMOS device
X5: Temperature (^oC)

a) Fit a multiple linear regression model for this data.

```
> ml=lm(y ~ x1 + x2 + x3 + x4 + x5, data = data)
> ml

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5
    2.85473     -0.29047      0.20572      0.45444     -0.59419      0.00464

>
```

b) Use stepwise regression criteria to find an appropriate regression model for these data .

c) Calculate the PRESS statistics for both models in (a) and (b). Which model would PRESSindicate is likely to be the best for predicting new response observations?