STAT 43000/STAT 53001 Applied Statistics Spring 2023
Test 1-Part II- Solution

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.
b) Is there 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 a response variable.
d) Update the fitted model in (c) by incorporating a binary variable Sex
e) Predict the height of a male whose foot is 28.8 cm.
Solution:
a) We used R code below to import the data and print first 5 observations

```
> library(readxl)
> data=read_xlsx("G:\\Aryal\\STAT 43000\\Exams\\Foot data.xlsx")

> head(data,5)
# A tibble: 5 × 6
  Sex     Age `Foot length` `Shoe Print` `Shoe size` Height
  <chr> <dbl>         <dbl>        <dbl>       <dbl>  <dbl>
1 M        67          27.8         31.3          11   180.
2 M        47          25.7         29.7           9   175.
3 M        41          26.7         31.3          11   185.
4 M        42          25.9         31.8          10   178.
5 M        48          26.4         31.4          10   182.
```

b) Let μ_M and μ_F be the mean foot length of male and female respectively. We would like to test the following hypothesis

$$H_0 \ : \ \mu_M = \mu_F$$
$$H_a \ : \ \mu_M \neq \mu_F$$

Observe form R output below the p-value is much smaller than 0.05*. Therefore, we reject the null hypothesis and conclude that there is a significant difference in the foot length based on the gender.*

```
> t.test(Foot.length~Sex)

        Welch Two Sample t-test
```

```
data:  Foot.length by Sex
t = -8.1505, df = 35.92, p-value = 1.101e-09
alternative hypothesis: true difference in means between group F and group M is not equal
95 percent confidence interval:
 -3.903672 -2.347958
sample estimates:
mean in group F mean in group M
       24.19524        27.32105
```

c) Based on the R output below the fitted model is

$$\hat{Height} = 64.126 + 4.291 \times \textit{Foot length}$$

```
> model=lm(Height~Foot.length)
> summary(model)

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

Residuals:
     Min       1Q   Median       3Q      Max
-11.4565  -3.5664   0.8766   2.7702  10.0717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.126     11.485   5.583 2.12e-06 ***
Foot.length    4.291      0.446   9.623 9.83e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.506 on 38 degrees of freedom
Multiple R-squared:  0.709,     Adjusted R-squared:  0.7014
F-statistic:  92.6 on 1 and 38 DF,  p-value: 9.833e-12
```

d) Based on the R output below the updated model is

$$\begin{aligned} \textit{Female}: &\quad Height = 95.641 + 2.942\,\textit{Foot length} \\ \textit{Male}: &\quad Height = 102.237 + 2.942\,\textit{Foot length} \end{aligned}$$

```
  > newmodel=lm(Height~Foot.length+Sex)
> newmodel

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

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

e) Based on the computer output below it is predicated that a male with foot length 28.8 cm is 186.9669 cum tall.

```
 > predict(newmodel, data.frame(Foot.length=28.8, Sex="M"))
       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.
b) Fit a simple logistic regression model using percentage labeling index of the bone marrow leukemia cells (LI) as a predictor variable.
c) Calculate the odds ratio for LI.
d) Calculate the estimated odds of leukemia remission at LI=0.8 and LI=0.9. Now, calculate the odds ratio using the odds at LI= 0.9 and LI=0.8. How do you interpret this value?

Solution:
a) We use R code below to import the data and extract the variable names.

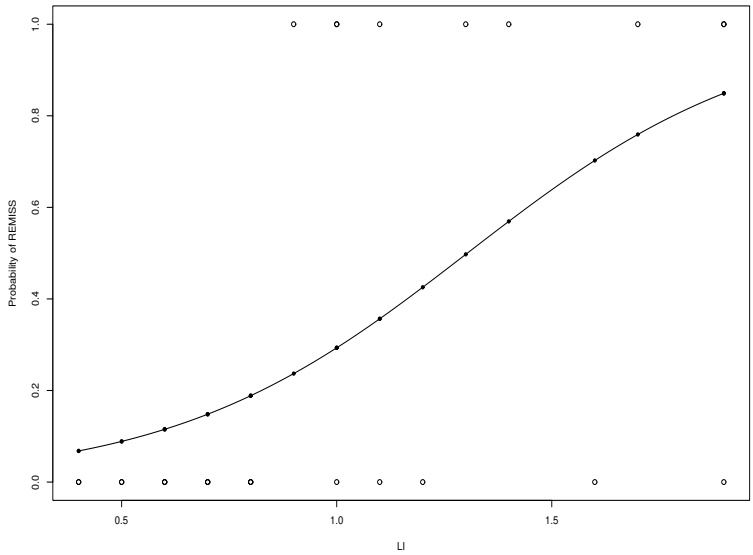
```
> data=read.table("C:\\Users\\aryal\\STAT 43000\\Exams\\leukemia Data.txt", header=T)
> head(data,5)
  REMISS CELL SMEAR INFIL  LI BLAST TEMP
1      1  0.8  0.83  0.66 1.9  1.10 1.00
2      1  0.9  0.36  0.32 1.4  0.74 0.99
3      0  0.8  0.88  0.70 0.8  0.18 0.98
4      0  1.0  0.87  0.87 0.7  1.05 0.99
5      1  0.9  0.75  0.68 1.3  0.52 0.98
> names(data)
[1] "REMISS" "CELL"   "SMEAR"  "INFIL"  "LI"     "BLAST"  "TEMP"
```

b) The fitted model is

$$\begin{aligned}\hat{\pi} &= \frac{e^{-3.777+2.897\times LI}}{1+e^{-3.777+2.897\times LI}} \\ &= [1+\exp(3.777-2.897\times LI)]^{-1}\end{aligned}$$

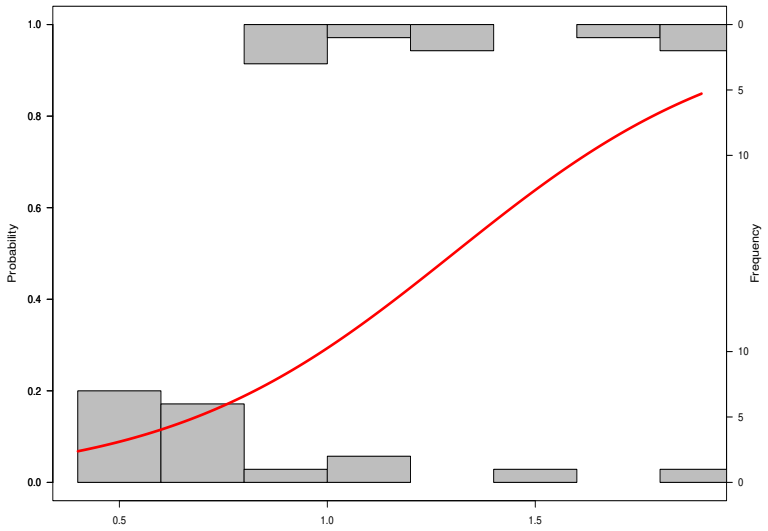
```
> plot(LI,REMISS,xlab="LI",ylab="Probability of REMISS")
> g=glm(REMISS~LI,family=binomial)
> curve(predict(g,data.frame(LI=x),type="resp"),add=TRUE)
> points(LI,fitted(g),pch=20)
> fit=glm(REMISS~LI,family=binomial)
```

```
> summary(fit)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9448  -0.6465  -0.4947   0.6571   1.6971

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.777      1.379  -2.740  0.00615 **
LI             2.897      1.187   2.441  0.01464 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 26.073  on 25  degrees of freedom
AIC: 30.073

Number of Fisher Scoring iterations: 4

> plot(LI,REMISS,xlab="LI",ylab="Probability of REMISS")
> curve(predict(fit,data.frame(LI=x),type="resp"),add=TRUE)
> points(LI,fitted(fit),pch=20)
```



One can use the R code below to produce the following graph

```
> library(popbio)
> logi.hist.plot(LI,REMISS,boxp=FALSE,type="hist",col="gray")
```



c) We know that the estimated parameter for LI is 2.897. so the odds ratio for LI is calculated as $\exp(2.897) = 18.1197$. *This means for every increase of 1 unit in LI, the estimated odds of leukemia remission are multiplied by 18.1197.*

d) We know that for a single predictor x *the odds of success are given by*

$$\frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 x).$$

Therefore, at LI=0.8 the estimated odds of leukemia remission is $\exp(-3.777 + 2.897 \times 0.8) = 0.232$ *and at LI=0.9 the estimated odds of leukemia remission is* $\exp(-3.777 + 2.897 \times 0.9) = 0.310$.
The odds ratio is $\frac{0.310}{0.232} = 1.336$, *which is the ratio of the odds of remission when LI=0.9 compared to the odds when L1=0.8.*
This means that for every 0.1 unit increase in LI, the estimated odds of remission is multiplied by 1.336. It should be noted that $\exp(2.897 \times 0.1) = 1.336$.

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`
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

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

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

```
> library(faraway)
> data(sexab)
> dim(sexab)
[1] 76  3
> head(sexab,5)
       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
```

It appears that there are 76 observations with three variables.

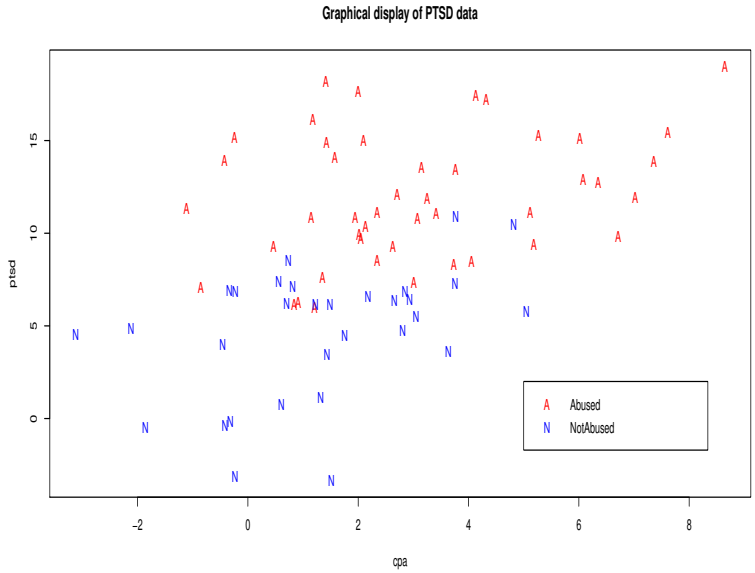
b) We can use R code below to display the data graphically

```
> library(faraway)
> data(sexab)
> attach(sexab)
> plot(ptsd~cpa,pch=as.character(csa), sexab,col=ifelse(csa=="Abused", "red", "blue"),
+ main="Graphical display of PTSD data")
> legend(5,2,levels(csa), pch=c("A","N"), col=c("red", "blue"))
```

We could also use "unclass" option as below to to replace the characters.

```
> library(faraway)
> data(sexab)
> data(sexab)
> attach(sexab)
```

```
> plot(ptsd~cpa,pch=unclass(csa), sexab, main="Graphical display of PTSD data")
> legend(6,2,levels(csa), pch=1:2)
```



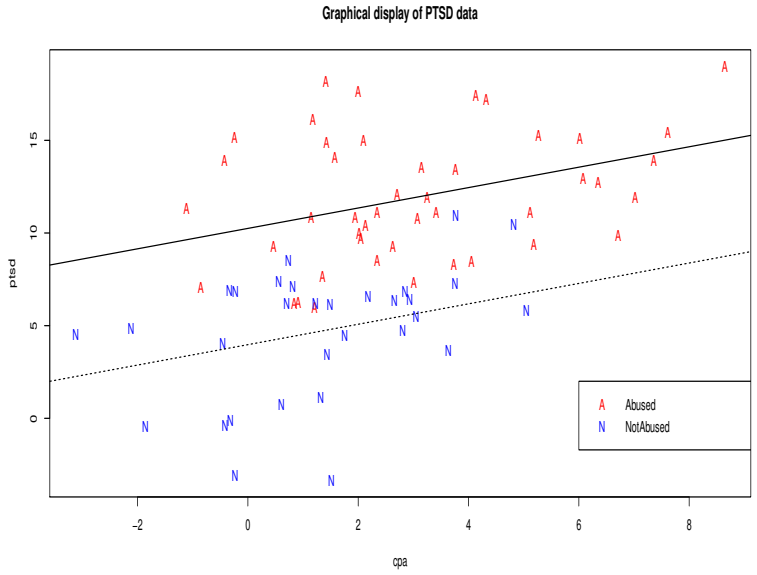
c) Using R code below the fitted linear regression model is

$$Abused: \quad ptsd = 10.2480 + 0.5506\, cpa$$
$$Not\ Abused: \quad ptsd = 3.9752 + 0.5506\, cpa$$

```
> model=lm(ptsd~cpa+factor(csa))
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.1567 -2.3643 -0.1533  2.1466  7.1417

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          10.2480     0.7187  14.260  < 2e-16 ***
cpa                   0.5506     0.1716   3.209  0.00198 **
factor(csa)NotAbused -6.2728     0.8219  -7.632 6.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.273 on 73 degrees of freedom
Multiple R-squared:  0.5786,    Adjusted R-squared:  0.5671
F-statistic: 50.12 on 2 and 73 DF,  p-value: 2.002e-14
```

We can display the model in the scatter plot using R code below

```
> plot(ptsd~cpa,pch=as.character(csa), sexab,col=ifelse(csa=="Abused", "red", "blue"),
+ main="Graphical display of PTSD data")
> legend(6,2,levels(csa), pch=c("A","N"), col=c("red", "blue"))
> abline(10.2480,0.5506)
> abline(10.2480-6.2728,0.5506, lty=2)
```



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).
b) Calculate the value of the Durbin-Watson test statistic. Do we have an evidence of autocorrelation at $\alpha = 0.05$.
c) Obtain estimates of the $\hat{\rho}$ based on the Cochrane-Orcutt procedure.
d) Obtain estimates of $\hat{\rho}$ based on the Hildreth-Lu procedure.

Solution:
a) We use R code below to import the data and estimate the parameters.

```
> data=read.table("C:\\aryal\\STAT 43000\\Exams\\Bus.txt", header=TRUE)
> t=data$t
> y=data$expenses
> x=data$miles
> model1=lm(y~x)
> model1

Call: lm(formula = y ~ x)
```

```
Coefficients: (Intercept)            x
                    0.6496       0.4467
```

Hence, the fitted model is $\hat{y} = 0.6496 + 0.4467x$. *Therefore,*

$$\textit{Total expenses (in £)} = 64960 + 0.4467 \times \textit{distance(in miles)}$$

b) We use R code test below to test the Durbin-Watson test

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho > 0 \end{aligned}$$

```
> library(lmtest)
> dwtest(model1)
        Durbin-Watson test

data:  model1 DW = 1.1603, p-value = 0.00365
alternative hypothesis: true autocorrelation is greater than 0
```

Note that the value of Durbin-Watson test statistic is 1.1603 with p-value 0.00365 which less than 0.05. So we reject the null hypothesis and conclude that there is an evidence of positive autocorrelation.
c) According to the Cochrane-Orcutt procedure we have

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=2}^{n} e_t^2}$$

We use R code below to obtain the estimate the value of ρ

```
> model=lm(y~x)
> et=resid(model)
> et=resid(model)[t]
> et1=resid(model)[t-1]
> s1=sum(et[-1]*et1)
> s2=sum(et1^2)
> phat=s1/s2
> phat
[1] 0.3672153
```

Hence, the estimated value of the autocorrelation coefficient is 0.3672153.
Or We could use R code below to find the value of the autocorrelation coefficient

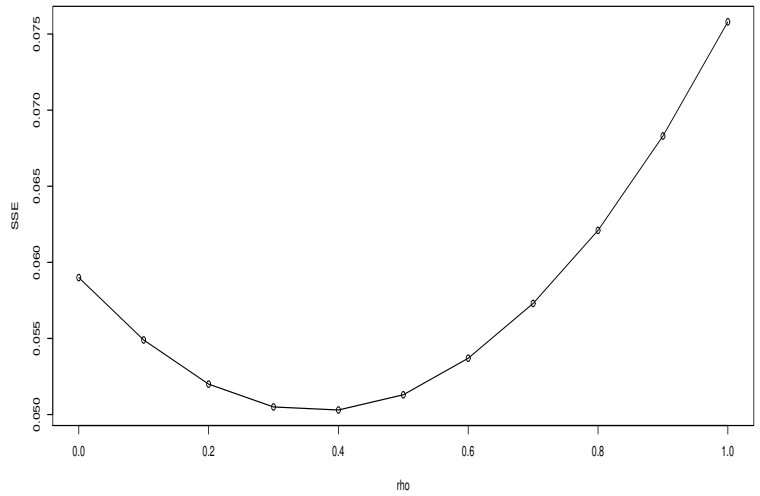
```
library(orcutt)
cochrane.orcutt(model)
Cochrane-orcutt estimation for first order autocorrelation
```

```
Call: lm(formula = y ~ x)
 number of interaction: 5
 rho 0.367587

Durbin-Watson statistic
(original):    1.16033 , p-value: 3.65e-03
(transformed): 2.41849 , p-value: 8.542e-01

 coefficients:
(Intercept)       miles
   0.645939    0.446578
```

d) According to the Hildreth-Lu we use the value of ρ that minimizes SSE of the transformed regression model. We use R code below for a sequence of ρ values

```
> rho<-seq(0,1, 0.1)
> hildreth.lu <- function(rho, model){
+ x <- model.matrix(model)[, -1]
+ y <- model.response(model.frame(model))
+ n <- length(y)
+ t <- 2:n
+ y <- y[t] - rho * y[t-1]
+ x <- x[t] - rho * x[t-1]
+ return(lm(y ~ x))}
> fit <- lm(y ~ x, data)
> tab <- data.frame('rho' = rho,
+ 'SSE' = sapply(rho, function(r) {deviance(hildreth.lu(r, fit))}))
> round(tab, 4)
   rho    SSE
1  0.0  0.0590
2  0.1  0.0549
3  0.2  0.0520
4  0.3  0.0505
5  0.4  0.0503
6  0.5  0.0513
7  0.6  0.0537
8  0.7  0.0573
9  0.8  0.0621
10 0.9  0.0683
11 1.0  0.0758
```

Note that the SSE is minimum at $\rho = 0.4$, therefore the estimated value of the autocorrelation coefficient is 0.4.



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.
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 PRESS indicate is likely to be the best for predicting new response observations?

Solution: We used R code below to fit a multiple linear regression model. The resulting model is $\hat{y} = 2.85473 - 0.29047x_1 + 0.20572x_2 + 0.45444x_3 - 0.59419x_4 + 0.00464x_5$.

```
> inverter=read.csv("C:\\aryal\\STAT 43000\\Exams\\inverter.csv")
> attach(inverter)
> head(inverter)
  x1 x2 x3 x4 x5     y
1  3  3  3  3  0 0.787
2  8 30  8  8  0 0.293
> model=lm(y~x1+x2+x3+x4+x5)
> model
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
```

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

b) Using the stepwise regression it can be observed that variable x_5 can be removed from our model

```
> library(MASS)
> step <- stepAIC(model, direction="both")
Start:  AIC=44.46
y ~ x1 + x2 + x3 + x4 + x5
       Df Sum of Sq     RSS    AIC
- x5    1     0.314  91.901 42.546
<none>               91.587 44.460
- x3    1    28.261 119.848 49.184
- x1    1    29.495 121.082 49.440
- x2    1    36.204 127.791 50.788
- x4    1    37.678 129.264 51.075
Step:  AIC=42.55
y ~ x1 + x2 + x3 + x4

       Df Sum of Sq     RSS    AIC
<none>               91.901 42.546
+ x5    1     0.314  91.587 44.460
- x3    1    28.388 120.289 47.276
- x1    1    29.406 121.307 47.486
- x2    1    38.393 130.294 49.273
- x4    1    42.879 134.780 50.119
```

The model is $\hat{y} = 3.1482 - 0.2900x_1 + 0.1992x_2 + 0.4554x_3 - 0.6092x_4$

```
> newmodel=lm(y~x1+x2+x3+x4)
> newmodel

Call:
lm(formula = y ~ x1 + x2 + x3 + x4)
Coefficients:
(Intercept)           x1           x2           x3           x4
     3.1482      -0.2900       0.1992       0.4554      -0.6092
```

c)

```
> library(MPV)
> PRESS(model)
[1] 252.695
> PRESS(newmodel)
[1] 238.2421
```

A model with smaller value of PRESS statistic is a preferred model. Since the PRESS statistic for new model (without the 5th variable) is lower than the original model we will choose the new model.