STAT 43000/STAT 53001 Applied Statistics Spring 2023 Homework 3

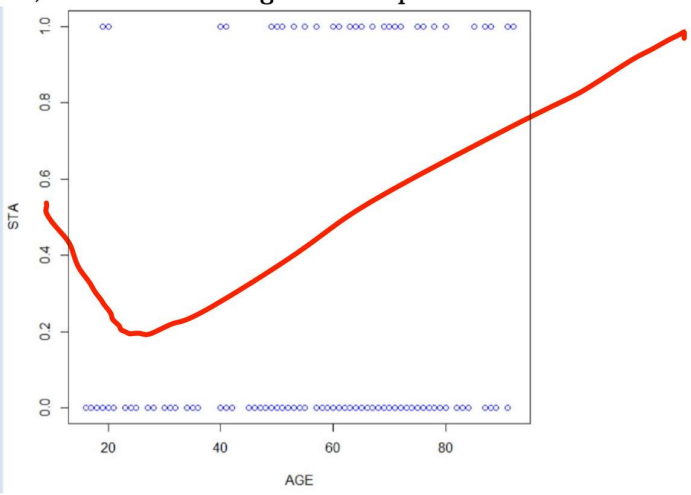
Due Date : April 3, 2023

Name:
PUID:

Q.N. 1) The **ICU** dataset provided in the Brightspace consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. A number of publications have appeared which have focused on various aspects of this problem.

a)Import the data and display vital status (STA) versus AGE using a scatter plot.

```
> ql=read.csv("C:\\Users\\Zhang\\Desktop\\courses\\ing\\Applied Statistics\\hmw$
>
> head(ql,3)
  ID STA AGE SEX RACE SER CAN CRN INF CPR SYS HRA PRE TYP FRA PO2 PH PCO BIC
1 552   0  16   0    1   1   0   0   0   0 100 140   0   1   1   0  0   0   0
2 102   0  16   1    1   0   0   0   0   0 104 111   0   1   0   0  0   0   0
3 837   0  17   1    3   0   0   0   0   0 130 140   0   1   0   0  0   0   0
  CRE LOC
1   0   0
2   0   0
3   0   0
> attach(ql)
> plot(STA~AGE,main="Scatter plot of STA with respect to Age",col="blue")
>
```



b) Fit a simple logistic regression model.

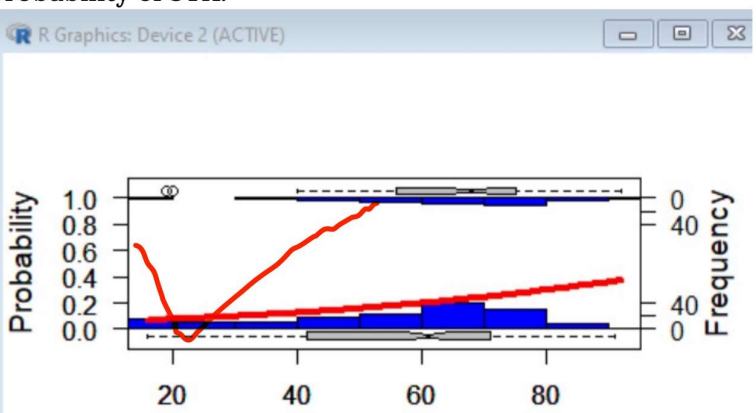
```
> modelql=glm(STA~AGE,family="binomial")
> summary(modelql)

Call:
glm(formula = STA ~ AGE, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9536  -0.7391  -0.6145  -0.3905   2.2854

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
AGE          0.02754    0.01056   2.607  0.00913 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.31  on 198  degrees of freedom
AIC: 196.31

Number of Fisher Scoring iterations: 4
```

c) Write down the equation for the logistic regression model of STA on AGE.

Π=[1+exp(-12.040+4.024*wt)]^-1

d) Plot the logistic curve to display the probability of STA.

```
> install.packages("popbio")
Installing package into 'C:/Users/Zhan
(as 'lib' is unspecified)
--- Please select a CRAN mirror for us
trying URL 'https://cloud.r-project.or
Content type 'application/zip' length
downloaded 295 KB

package 'popbio' successfully unpacked

The downloaded binary packages are in
        C:\Users\Zhang\AppData\Local\T
> library(popbio)
> logi.hist.plot(AGE,STA,type="hist")
>
```



e) Using the fitted model determine the STA probability of an individual of AGE 60.

```
> predict(modelql,data.frame(AGE=60),type="response")
         1
0.1968726
>
```

Q.N. 2) Suppose you are investigating allegations of sex discrimination in the hiring practices of a particular firm. An equal-rights group claims that females are less likely to be hired than male within the same background, experiences and other qualifications. The data collected on 28 former applicants provided in the Brightspace (hiring data) will be used to fit the model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, where

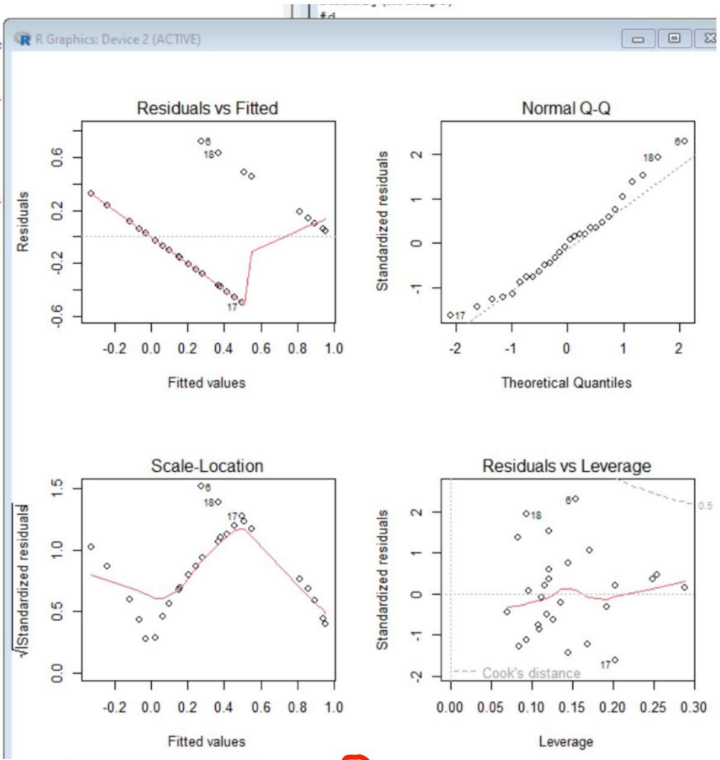
$$y = \begin{cases} 1 & \text{if hired} \\ & \text{f not} \end{cases} \qquad 0 \quad \text{i}$$

x_1 = Years of higher education (4, 6 or 8) x_2 = Years of experience

$$x_3 = \begin{cases} 1 & \text{if male applicant } 0 \quad \text{if} \quad \text{female applicant} \end{cases}$$

a) Fit a multiple linear regression model. Does this model seem appropriate? Justify your answer.

```
> file.choose()
[1] "C:\\Users\\Zhang\\Desktop\\courses\\ing\\Applied Statistic
> head(q2,3)
Error in head(q2, 3) : object 'q2' not found
> q2=read.table("C:\\Users\\Zhang\\Desktop\\courses\\ing\\Appli
> head(q2,3)
  V1 V2 V3 V4
1  y x1 x2 x3
2  0  6  2  0
3  0  4  0  1
> q2=read.table("C:\\Users\\Zhang\\Desktop\\courses\\ing\\Appli
> head(q2,3)
  y x1 x2 x3
1 0  6  2  0
2 0  4  0  1
3 1  6  6  1
> attach(q2)
> modelq2=lm(y~x1+x2+x3)
> par(mfrow=(c(2,2)))
> plot(modelq2)
> summary(modelq2)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49572 -0.24532 -0.00046  0.15100  0.72310

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.84436    0.28128  -3.002 0.006179 **
x1           0.10692    0.04230   2.528 0.018482 *
x2           0.08863    0.02069   4.285 0.000256 ***
x3           0.48473    0.13814   3.509 0.001802 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3433 on 24 degrees of freedom
Multiple R-squared:  0.5369,    Adjusted R-squared:  0.479
F-statistic: 9.273 on 3 and 24 DF,  p-value: 0.0002969

> |
```



THE residual standard error is not big but x1 isn't quite meaningful

b) Find the maximum likelihood estimates of β_0,β_1,β_2 and β_3 to fit a logistic regression model

$$E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{[1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)]}$$

```
> modelq2log=glm(y~x1+x2+x3, family="binomial")
> summary(modelq2log)

Call:
glm(formula = y ~ x1 + x2 + x3, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4380  -0.4573  -0.1009   0.1294   2.1804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.2483     6.0805  -2.343   0.0191 *
x1            1.1549     0.6023   1.917   0.0552 .
x2            0.9098     0.4293   2.119   0.0341 *
x3            5.6037     2.6028   2.153   0.0313 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.165  on 27  degrees of freedom
Residual deviance: 14.735  on 24  degrees of freedom
AIC: 22.735

Number of Fisher Scoring iterations: 7
```

B0=14.2483
B1=1.15
B2=-0.9
B3=-5.6

c) Calculate the 95% confidence interval of the parameters.

```
> confint(modelq2log,level=0.95)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -31.3101488 -5.628821
x1            0.2049156  2.744315
x2            0.3302805  2.136563
x3            1.8033770 12.674158
>
```

d) Conduct a test of model adequacy. Use α = 0.05.

```
> qchisq(0.05,24,lower.tail=FALSE)
[1] 36.41503
>
```

e) Is there sufficient evidence to indicate that gender is an important predictor of hiring status? Test using α = 0.05.

```
> cor.test(x3,y, alternative = "two.side", method = "pearson",conf.level = 0.95)

        Pearson's product-moment correlation

data:  x3 and y
t = 1.4831, df = 26, p-value = 0.1501
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1046963  0.5908066
sample estimates:
      cor
0.2792896
```

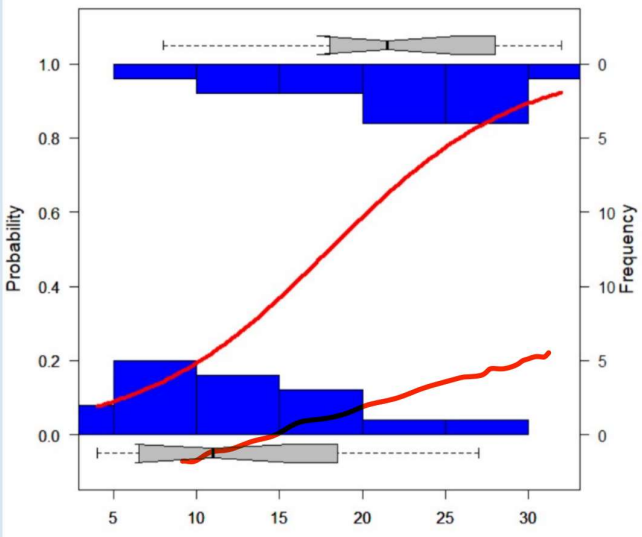
Cor=0.279 under 95%confidence interval is low, gender may not be an important predictor of hiring status

Q.N. 3) The following data resulted from a study commissioned by a large management consulting company to investigate the relationship between amount of job experience (months) for a junior consultant and the likelihood of the consultant being able to perform a certain complex task.

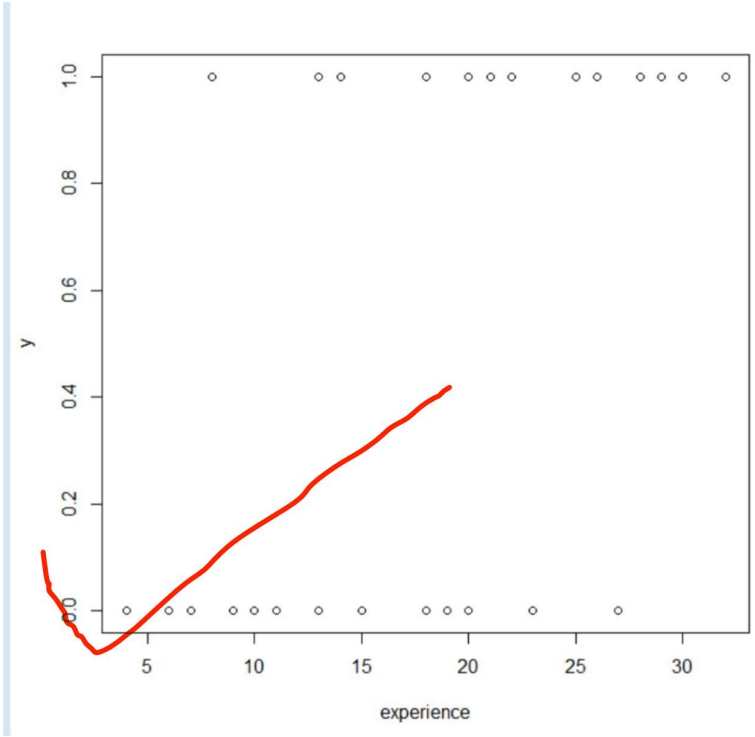
Success: 8, 13, 14, 18, 20, 21, 21, 22, 25, 26, 28, 29, 30, 32 Failure:
4, 4, 6, 6, 7, 9, 10, 11, 11, 13, 15, 18, 19, 20, 23, 27

a) Develop a simple logistic regression model

```
-1.8828  -0.7095  -0.3987   0.8525   1.9689

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.20140    1.23399  -2.594  0.00948 **
experience   0.17732    0.06562   2.702  0.00689 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.455  on 29  degrees of freedom
Residual deviance: 30.770  on 28  degrees of freedom
AIC: 34.77

Number of Fisher Scoring iterations: 4

> install.packages("popbio")
Installing package into 'C:/Users/Zhang/AppData/Local/R/win-library/4.2'
(as 'lib' is unspecified)
trying URL 'https://cloud.r-project.org/bin/windows/contrib/4.2/popbio_2.7.zip'
Content type 'application/zip' length 302228 bytes (295 KB)
downloaded 295 KB

package 'popbio' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Zhang\AppData\Local\Temp\RtmpslQbda\downloaded_packages
> library(popbio)
> logi.hist.plot(experience,y,type="hist")
Warning messages:
1: In (function (z, notch = FALSE, width = NULL, varwidth = FALSE,  :
  some notches went outside hinges ('box'): maybe set notch=FALSE
2: In (function (z, notch = FALSE, width = NULL, varwidth = FALSE,  :
  some notches went outside hinges ('box'): maybe set notch=FALSE
>
```



b) Display the data and the fitted model.



c) Construct a 95% confidence interval for β_1.

```
> confint(modelq3,level=0.95)
Waiting for profiling to be done...
                    2.5 %     97.5 %
(Intercept) -6.07751046 -1.082857
experience   0.06430993  0.329305
>
```

d) What is the estimated probability of task performance if a consultant has 2 years of experience.

```
> predict(modelq3,data.frame(experience=2),type="response")
        1
0.7415817
>
```

e) Estimate the emotional job experience for which 90% of the consultants will be able to performa certain complex task.

```
> predict(modelq3,data.frame(y=.9),type="response")
         1          2          3          4          5          6          7          8
0.14394995 0.28981572 0.32762151 0.49757622 0.58538962 0.62767496 0.62767496 0.66809288
         9         10         11         12         13         14         15         16
0.77408492 0.80358279 0.85364407 0.87443887 0.89265061 0.92220856 0.07641171 0.07641171
        17         18         19         20         21         22         23         24
0.10550539 0.10550539 0.12344757 0.16720786 0.19337468 0.22254217 0.22254217 0.28981572
        25         26         27         28         29         30
0.36780486 0.49757622 0.54180756 0.58538962 0.70617709 0.83007460
Warning message:
'newdata' had 1 row but variables found have 30 rows
>
```

Q.N. 4) An experiment was conducted on the effect of toxicity on the number of offspring produced by the aquatic animal C. dubia (water flea). The variables in the data are:

animal -- ID for each C. dubia (water flea) tested offspring -- count of young produced

conc -- concentration of pollutant, 5 levels (micro grams/L)

(a) Fit a Poisson regression to the data and Provide the summary output.

```
> modelq4=glm(offspring~conc,family="poisson")
> summary(modelq4)

Call:
glm(formula = offspring ~ conc, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0940  -0.8667  -0.4584   1.0255   2.4064

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.9964989  0.0588449  50.922  < 2e-16 ***
conc        -0.0028093  0.0003598  -7.806 5.89e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 141.37  on 49  degrees of freedom
Residual deviance:  78.80  on 48  degrees of freedom
AIC: 297.1

Number of Fisher Scoring iterations: 4
```

model equation?

(b) Is there evidence that increasing pollutant decreases the mean offspring count?
THE ratio for conc<0 and it's significant, which means increasing pollutant decreases the mean offspring count

(c) Provide the estimated mean number of offspring for concentration level 20, 80, 235.

```
> predict(modelq4,data.frame(conc=c(20,80,235)),interval="pred",type="response")
       1        2        3
18.92213 15.98789 10.34537
>
```

Q.N. 5) A partially filled computer output of ANOVA is shown below. Fill in the blanks. You may give a range for the p-value.

Source	DF	SS	MS	F value	P value
Model	?	?	22.75	?	?
Error	15	186	?		
Total	19	277			

Source	DF	SS	MS	F-value	P-value
Model	4	91		1.835	0.1746
Error			12.4		
Total					