**STAT 43000/STAT 53001 Applied Statistics Spring 2023**

**Lab -7**

1. The median per-capita income data aggregated at the county level for the state of Maine ( US Census ACS 2007 - 2011 dataset) are provided with this assignment. We would like to study the relationship between income and whether or not the county is on the coast.

   **a.** Import the data in R

```
> Q1=read.table(""C:\\Users\\Zhang\\Downloads\\Maine.txt",header=T)
Error: unexpected symbol in "Q1=read.table(""C"
> Q1=read.table("C:\\Users\\Zhang\\Downloads\\Maine.txt",header=T)
> head(Q1)
  SN Coast Income
1  1    no  23663
2  2    no  20659
3  3   yes  32277
4  4    no  21595
5  5   yes  27227
6  6    no  25023
>
```

   **b.** Does the income differ between coastal and non-coastal communities?

```
> y=(Coast=="yes")*1
> t.test(Income~Coast)

        Welch Two Sample t-test

data:  Income by Coast
t = -2.974, df = 9.1871, p-value = 0.01526
alternative hypothesis: true difference in me
95 percent confidence interval:
 -7836.701 -1077.299
sample estimates:
 mean in group no mean in group yes
         22252.12          26709.12
```

p-value=0.015<0.05, fail to reject h1.
true difference in means between group no and group yes is equal to 0

   **c.** Fit a simple logistic regression to model the likelihood of a county located in the coastal community.

```
> m=glm(y~Income,family="binomial")
> summary(m)

Call:
glm(formula = y ~ Income, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3578  -0.6948  -0.1863   0.5207   2.2137

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.222e+01  5.765e+00  -2.119   0.0341 *
Income       5.048e-04  2.401e-04   2.102   0.0355 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 14.807  on 14  degrees of freedom
AIC: 18.807

Number of Fisher Scoring iterations: 5
```

Π=[1+exp(12.22-0.0005*Income)]^(-1)

**d.** Calculate the Pseudo-R squared value.
R2=1−ResidualDeviance/NullDeviance=0.3324

**e.** Use nagelkerke(model) of the rcompanion library to calculate the various form of pseudo r-squred value

```
> nagelkerke(m)
$Models

Model: "glm, y ~ Income, binomial"
Null:  "glm, y ~ 1, binomial"

$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                             0.332428
Cox and Snell (ML)                   0.369248
Nagelkerke (Cragg and Uhler)         0.492331

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq   p.value
      -1     -3.6867 7.3735 0.0066192

$Number.of.observations

Model: 16
Null:  16
```

**f.** Use lrm function of rms package to compute another form of the pseudo $R^2$ called the Nagelkerke $R^2$.

```
> m2=lrm(y~Income)
> m2
Logistic Regression Model

lrm(formula = y ~ Income)

                      Model Likelihood     Discrimination     Rank Discrim.
                         Ratio Test            Indexes            Indexes
Obs            16    LR chi2      7.37    R2         0.492    C        0.828
 0              8    d.f.            1    R2(1,16)  0.329    Dxy      0.656
 1              8    Pr(> chi2) 0.0066    R2(1,12)  0.412    gamma    0.656
max |deriv| 0.4                           Brier      0.143    tau-a    0.350

          Coef     S.E.   Wald Z Pr(>|Z|)
Intercept -12.2176 5.7646 -2.12  0.0341
Income      0.0005 0.0002  2.10  0.0355
```

R2=0.492

2. The following simulated data provides the information regarding how many scholarships offers a high school baseball player receives based on their school division (“A”, “B”, or “C”) and their college entrance exam score (measured from 0 to 100).

data = data.frame(offers = c(rep(0, 50), rep(1, 30), rep(2, 10), rep(3, 7), rep(4, 3)), division = sample(c("A", "B", "C"), 100, replace = TRUE), exam = c(runif(50, 60, 80), runif(30, 65, 95), runif(20, 75, 95)))

**a.** Print first five observations

```
R Console

> data =  data.frame(offers = c(rep(0,
> head(data)
  offers division     exam
1      0        B 73.51972
2      0        B 65.95151
3      0        B 60.87850
4      0        C 64.24690
5      0        B 66.99052
6      0        C 67.74656
> head(data,5)
  offers division     exam
1      0        B 73.51972
2      0        B 65.95151
3      0        B 60.87850
4      0        C 64.24690
5      0        B 66.99052
> |
```

**b.** Provide the summary statistics of the variables offers , division and exam

```
> summary(data)
     offers        division              exam      
 Min.   :0.00   Length:100         Min.   :60.41  
 1st Qu.:0.00   Class :character   1st Qu.:67.01  
 Median :0.50   Mode  :character   Median :75.96  
 Mean   :0.83                      Mean   :76.04  
 3rd Qu.:1.00                      3rd Qu.:82.91  
 Max.   :4.00                      Max.   :93.90  
> |
```

**c.** Find the mean scores of the entrance exam by division and number of offers.

```
> aggregate(data$exam,by=list(data$offers,data$division),FUN=mean)
   Group.1 Group.2        x
1        0       A 68.72525
2        1       A 80.70206
3        2       A 86.21127
4        3       A 87.77707
5        4       A 77.96677
6        0       B 71.49162
7        1       B 79.67006
8        2       B 87.08834
9        3       B 78.66088
10       4       B 83.90929
11       0       C 69.27565
12       1       C 79.82872
13       2       C 83.47894
14       3       C 90.58236
15       4       C 82.88563
```

**d.** Fit a Poisson regression model for the data

```
> model=glm(offers~division+exam,family="poisson")
> model

Call:  glm(formula = offers ~ division + exam, family = "poisson")

Coefficients:
(Intercept)    divisionB    divisionC         exam  
   -6.59498     -0.19507     -0.19000      0.08202  

Degrees of Freedom: 99 Total (i.e. Null);  96 Residual
Null Deviance:      138.1 
Residual Deviance: 89.28        AIC: 214.2
> |
```

**e.** What does the coefficient for exam mean?

In this case, the coefficient for 'exam' is 0.08202. This means that for every one-unit increase in the exam score, the logged number of offers is expected to increase by 0.08202, assuming that the other predictor (division) is held constant.

**a.** f. What does the coefficient corresponding to DivisionB indicate?

The coefficient corresponding to DivisionB is -0.19507. it means holding the exam score constant, the logged number of offers for DivisionB is expected to be 0.19507 lower than the logged number of offers for DivisionA (the reference division).

**b.** What does the coefficient corresponding to DivisionC indicate?

The coefficient corresponding to DivisionC is -0.19000. This means that, holding the exam score constant, the logged number of offers for DivisionC is expected to be 0.19000 lower than the logged number of offers for DivisionA.

**c.** Does the data fits the Poisson regression? (You can check it by performing the chi-squared test on Residual deviance. )

```
> residual_deviance <- model$deviance
> residual_df <- model$df.residual
> p_value <- 1 - pchisq(residual_deviance, residual_df)
> p_value
[1] 0.6731236
>
```

0.67>0.05, so it fits well