110 學年度第二學期科學計算軟體作業十

姓名: 蕭合亭 學號: F64109527

1. 利用提供的 Titanic.csv 資料集，透過**向前法 Logistic regression** 探討鐵達尼號乘客之年齡、性別、艙等因子對存活狀態的影響。結果應包含下列內容：(答題提醒:請”**完整**”展示變數篩選**過程**、**結果**，且說明**選入那些因子**，並應**注意標註 p 值(sig.)**、**β(Beta, 估計值)**、**AIC 等主要判斷依據**，若未達到或錯誤皆會斟酌扣分)

   提示：建議將各變數以 as. factor()轉換為 factor 格式

   程式碼

```
mod.null <- glm(as.factor(Survived) ~ 1, family="binomial", data = dataset)
mod.full <- glm(as.factor(Survived) ~ ., family="binomial", data = dataset)
f.model = step(mod.null, scope =list(lower=mod.null, upper=mod.full), direction = "forward", trace = 1)
summary(f.model)
```

   結果

```
> mod.null <- glm(as.factor(Survived) ~ 1, family="binomial", data = dataset)
> mod.full <- glm(as.factor(Survived) ~ ., family="binomial", data = dataset)
> f.model = step(mod.null, scope =list(lower=mod.null, upper=mod.full), direction = "forward", trace = 1)
Start:  AIC=2771.46
as.factor(Survived) ~ 1

1
        Df Deviance    AIC
+ Sex    1   2335.0 2339.0
+ Class  3   2588.6 2596.6
+ Age    1   2749.9 2753.9
<none>      2769.5 2771.5

Step:  AIC=2338.99
as.factor(Survived) ~ Sex

2
        Df Deviance    AIC
+ Class  3   2228.9 2238.9
+ Age    1   2329.1 2335.1
<none>      2335.0 2339.0

Step:  AIC=2238.91
as.factor(Survived) ~ Sex + Class

3
       Df Deviance    AIC
+ Age   1   2210.1 2222.1
<none>     2228.9 2238.9

4
Step:  AIC=2222.06
as.factor(Survived) ~ Sex + Class + Age
```

```
> summary(f.model)

Call:
glm(formula = as.factor(Survived) ~ Sex + Class + Age, family = "binomial",
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0812  -0.7149  -0.6656   0.6858   2.1278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0438     0.1679  12.171  < 2e-16 ***
SexMale      -2.4201     0.1404 -17.236  < 2e-16 ***
Class2nd     -1.0181     0.1960  -5.194 2.05e-07 ***
Class3rd     -1.7778     0.1716 -10.362  < 2e-16 ***
ClassCrew    -0.8577     0.1573  -5.451 5.00e-08 ***
AgeChild      1.0615     0.2440   4.350 1.36e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2769.5  on 2200  degrees of freedom
Residual deviance: 2210.1  on 2195  degrees of freedom
AIC: 2222.1

Number of Fisher Scoring iterations: 4
```

(1)、 向前法變數篩選過程及結果，並說明選入的因子有哪些(50%)。

AIC 值越小，表示整體模型之合適度越佳。

添加任何變數都不會讓 AIC 值下降，以作為向前法所篩選出的最佳模型。

1. 目前模型 AIC 值以及模型變數為常數

   AIC 值=2771.46

   模型添加變數 Sex 以後，新的 AIC 值=2339.0，低於原本的 AIC 值則選入該變數，並進行下一輪的篩選。

2. 目前模型 AIC 值以及模型變數已添加 Sex

   AIC 值=2338.99

   模型添加變數 Class 以後，新的 AIC 值=2238.9，低於原本的 AIC 值則選入該變數，並進行下一輪的篩選。

3. 目前模型 AIC 值以及模型變數已添加 Sex、Class

   AIC 值=2238.91

   模型添加變數 Age 以後，新的 AIC 值=2222.1，低於原本的 AIC 值則選入該變數，結束篩選。

4. 目前模型 AIC 值以及模型變數已添加 Sex、Class、Age

   AIC 值=2222.06

因為都會下降，全部加至模型。

Summary：可見最後模型的 AIC 值=2222.1

(2)、 各個變數的 β 值及顯著性為何，請搭配計算結果說明(25%)？

看上圖 summary 部分，用紅色框框起來的部分為 Beta、p 值。

Sexmale :

β 值=-2.4201 ，拿 male 作為比較值，而得其比較係數。

p 值<2e-16 <0.05，達統計上的顯著性。

Class2nd :

β 值=-1.0181，拿 Class1st 作為比較值，而得其比較係數。

p 值<2e-16 <0.05，達統計上的顯著性。

Class3rd :

β 值=-1.7778，拿 Class1st 作為比較值，而得其比較係數。

p 值=2.05e-07 <0.05，達統計上的顯著性。

ClassCrew :

β 值=-0.8577，拿 Class1st 作為比較值，而得其比較係數。

p 值=5.00e-08 <0.05，達統計上的顯著性。

AgeChild :

β 值=1.0615，拿 Adult 作為比較值，而得其比較係數。

p 值=1.36e-05 <0.05，達統計上的顯著性。

(3)、 各個變數的勝算比為何(25%)？(結果需包含 95%信賴區間數值)

程式碼

```
exp(coef(f.model))
confint(f.model)
```

結果

```
> exp(coef(f.model))
 (Intercept)      SexMale     Class2nd     Class3rd    ClassCrew     AgeChild
 7.72017801   0.08891625   0.36128255   0.16901595   0.42414659   2.89082629
> confint(f.model)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept)   1.7206688  2.3791924
SexMale      -2.6993511 -2.1485860
Class2nd     -1.4052474 -0.6364156
Class3rd     -2.1175898 -1.4445910
ClassCrew    -1.1662816 -0.5490908
AgeChild      0.5835884  1.5413772
>
```

分析

SexMale 的勝算比=0.08891625，表示為男性相較於女性的勝算比。

Class2nd 的勝算比=0.36128255，表示第二艙對於第一船艙的勝算比。

Class3rd 的勝算比=0.16901595，表示第三艙對於第一船艙的勝算比。

ClassCrew 的勝算比=0.42414659，表示 Crew 對於第一船艙的勝算比。

AgeChild 的勝算比=2.890282629，表示小孩對於大人的勝算比。

上圖紅框框起來的為各變數提供模型參數的置信區間。