Assignment 1

Abhishek Alate * Ahmed Aqdam Tariq * Laura Le

```
library(pscl)

## Warning: package 'pscl' was built under R version 3.6.2

library(SDMTools)

## Warning: package 'SDMTools' was built under R version 3.6.2

library(heplots)

## Warning: package 'heplots' was built under R version 3.6.2

## Warning: package 'car' was built under R version 3.6.2

library(pROC)

## Warning: package 'pROC' was built under R version 3.6.2
```

Problem: FlixIT Inc. purchases unlimited licenses to movie content that is then streamed on-demand to FlixIt subscribers. Subscribers pay a flat monthly fee, and are provided with unlimited access to FlixIt content. Last year, FlixIt implemented a "Recruit A Friend"(RAF) initiative. Under this initiative, any current FlixIt subscriber who recruits someone who purchases an annual FlixIt subscription is given a one-month rebate. FlixIt now wishes to determine the characteristics of subscribers who have participated in this initiative. Data collected from a random sample of FlixIt subscribers (contained in the file FlixIt.dat, which includes a header record) include age of the subscriber (Age: integer), region of the country in which the subscriber resides (Region:1=north, 2=south, 3=east, 4=west), and whether or not the subscriber participated in the RAF initiative (Partic: 0=no, 1=yes). Based on these data, and using a Logistic Regression framework, use R, to complete the following questions. Use the alpha level of 0.05.

Data setup.

```
flixitframe <- read.table("F:/GWU/Courses/Spring 2020/1. Statistics for
Analytics II/Assigments/Assignment 1/FlixIt.dat", header = TRUE)
flixitframe$Region <- as.factor(flixitframe$Region)
```

---

1. Can we be reasonably certain that Age predicts Partic? Explain.

```
# Running regression model with age independent variable with the Partic as
dependent variable.
flixitframe.age.logit <- glm(Partic ~ Age, data = flixitframe, family =
```

```
"binomial")
summary(flixitframe.age.logit)

##
## Call:
## glm(formula = Partic ~ Age, family = "binomial", data = flixitframe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5162  -0.7765  -0.4960   0.7455   2.2824
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.80213    1.43953  -5.420 5.96e-08 ***
## Age          0.16482    0.03298   4.997 5.82e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.92  on 149  degrees of freedom
## Residual deviance: 149.88  on 148  degrees of freedom
## AIC: 153.88
##
## Number of Fisher Scoring iterations: 4
```

Given extremely small p-value, at 5% signicance level, we reject the null hypothesis and conclude that age is signicant.

---

2. Can we be reasonably certain that Region predicts Partic? Explain.

```
# Running regression model with Re independent variable with the Partic as
dependent variable.
flixitframe.region.logit <- glm(Partic ~ Region, data = flixitframe, family =
"binomial")
summary(flixitframe.region.logit)

##
## Call:
## glm(formula = Partic ~ Region, family = "binomial", data = flixitframe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0074  -0.5168  -0.5168   0.5350   2.0393
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -18.57    1581.97  -0.012    0.991
## Region2        16.62    1581.97   0.011    0.992
## Region3        18.42    1581.97   0.012    0.991
```

```
## Region4        20.44    1581.97   0.013    0.990
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.92  on 149  degrees of freedom
## Residual deviance: 134.57  on 146  degrees of freedom
## AIC: 142.57
##
## Number of Fisher Scoring iterations: 17
```

Given large p-value, at 5% signicance level, we fail to reject the null hypothesis and conclude that region is NOT signicant on it's own.

---

3. Can we be reasonably certain that Age predicts Partic after controlling for Region? Explain.

```
flixitall.logit <- glm(Partic ~ Age+Region, data = flixitframe, family =
"binomial")
summary(flixitall.logit)
```

```
##
## Call:
## glm(formula = Partic ~ Age + Region, family = "binomial", data =
flixitframe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9783  -0.5738  -0.4542   0.5242   2.2015
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -20.31823 1568.82457  -0.013    0.990
## Age            0.05545    0.04340   1.278    0.201
## Region2       16.21317 1568.82401   0.010    0.992
## Region3       17.74620 1568.82403   0.011    0.991
## Region4       19.23919 1568.82436   0.012    0.990
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 184.92  on 149  degrees of freedom
## Residual deviance: 132.90  on 145  degrees of freedom
## AIC: 142.9
##
## Number of Fisher Scoring iterations: 17
```

Given large p-value, at 5% signicance level, we fail to reject the null hypothesis and conclude that Age and Region are not significant.

---

4. Can we be reasonably certain that Region predicts Partic after controlling for Age? Explain.

Similarly, given large p-value, at 5% signicance level, we fail to reject the null hypothesis and conclude that Age and Region are not significant.

---

5. What is your evaluation of the model fit in terms of McFadden's score?

```
pR2(flixitframe.age.logit)
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -74.9377756 -92.4611364  35.0467216   0.1895214   0.2083571   0.2940699
```

```
pR2(flixitframe.region.logit)
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -67.2850713 -92.4611364  50.3521301   0.2722881   0.2851488   0.4024518
```

```
pR2(flixitall.logit)
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -66.4516907 -92.4611364  52.0188914   0.2813014   0.2930481   0.4136006
```

When Age is the only predictor, 18% of the variation in the model is explained. When Region is the only predictor, 27% of the variation in the model is explained. When Age and region are predictors, 28% of the variation in the model is explained.

---

6. Using a threshold value of 0.5, create the confusion matrix, and find the total correct classification rate.

```
flixitframe["PredVal"] <- predict(flixitframe.age.logit,
list(Age=flixitframe$Age), type="link")
flixitframe["PredProb"] <- predict(flixitframe.age.logit,
list(Age=flixitframe$Age), type="response")
flixitframe["PredBin"] <- (flixitframe$PredProb>0.5)+0
confusion <- t(confusion.matrix(flixitframe$Partic, flixitframe$PredBin))
confusion <- addmargins(confusion)
confusion
```

```
##      pred
## obs     0  1 Sum
##   0    95  9 104
##   1    28 18  46
##   Sum 123 27 150
```

(TN+TP)/(TN+FN+TP+FP) = 95+18/150 = 0.7533333, the total correct classification rate is 0.75.

---

7. When predicting Partic using Age, what value does the AUROC take and how would you interpret this value?

```
ROC.curve <- roc(Partic ~ Age, data = flixitframe)
```

```
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

ROC.curve

##
## Call:
## roc.formula(formula = Partic ~ Age, data = flixitframe)
##
## Data: Age in 104 controls (Partic 0) < 46 cases (Partic 1).
## Area under the curve: 0.7744
```

The area under the curve is 0.7744 but it is below 0.8 which indicates that the model does not do a great job in discriminating between the two categories of the outcome variable.

---

8. Find the odds that a 35 year old subscriber from the East will be a RAF participant.

```
exp(-20.31823+0.05545*35+17.74620)

## [1] 0.5319105
```

The odds are 0.5319105.

---

9. Find the probability that a 35 year old subscriber from the east will be a RAF participant.

```
exp(-20.31823+0.05545*35+17.74620)/(1+exp(-20.31823+0.05545*35+17.74620))

## [1] 0.3472204
```

The probability is 0.3472204.

---

10. Find the best estimate of the coefficient associated with AGE in the full model and interpret its meaning.

```
exp(0.05545)

## [1] 1.057016
```

Age coefficient is 0.05545, it's exponent is 1.057016. This implies odds of participation are multiplied by 1.057016 for each unit increase in the Age.

---

11. If you were asked to provide the best estimate of the correlation between Age and Region what would you say?

```
ANOVA <- lm(Age~Region, data = flixitframe)
summary(ANOVA)

##
## Call:
## lm(formula = Age ~ Region, data = flixitframe)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6481  -3.4844   0.3426   3.5156  11.3519
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   31.176      1.209  25.779  < 2e-16 ***
## Region2        7.308      1.361   5.371 3.02e-07 ***
## Region3       12.472      1.387   8.994 1.15e-15 ***
## Region4       22.490      1.766  12.732  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.986 on 146 degrees of freedom
## Multiple R-squared:  0.571,  Adjusted R-squared:  0.5621
## F-statistic: 64.76 on 3 and 146 DF,  p-value: < 2.2e-16
```

```
etasq(ANOVA, anova=TRUE, partial=FALSE)
```

```
## Anova Table (Type II tests)
##
## Response: Age
##             eta^2 Sum Sq  Df F value    Pr(>F)
## Region    0.57095 4830.7   3  64.763 < 2.2e-16 ***
## Residuals         3630.1 146
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

eta-square value in synonymous to the R-squared value and it represents the strength of relationship which is 57%
