330 Assignmnt 2

Amir Ghobadi

2024-03-30

Question 1

A)

```
# adding in and naming the data frame
Area.df<- read.csv("data.Areas_A2.csv")



# GLM model
model <- glm(Y ~ previous.eruption + UrbanRural + offs_time +
earthquakes.year + total.land.area , data = Area.df, family = poisson)

# summary of the modek
exp(coef(model))

##                (Intercept)          previous.eruption
##                  3.1029801                  0.2329841
## UrbanRuralRural settlement            UrbanRuralUrban
##                  1.3113147                  0.5589173
##                  offs_time           earthquakes.year
##                  2.7399924                  1.0498360
##            total.land.area
##                  1.0157011

summary(model)

##
## Call:
## glm(formula = Y ~ previous.eruption + UrbanRural + offs_time +
##     earthquakes.year + total.land.area, family = poisson, data = Area.df)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.1323630  0.0114001   99.33   <2e-16 ***
## previous.eruption         -1.4567852  0.0070800 -205.76   <2e-16 ***
## UrbanRuralRural settlement  0.2710302  0.0132009   20.53   <2e-16 ***
## UrbanRuralUrban           -0.5817537  0.0080039  -72.68   <2e-16 ***
## offs_time                  1.0079551  0.0035531  283.68   <2e-16 ***
## earthquakes.year           0.0486339  0.0028080   17.32   <2e-16 ***
## total.land.area            0.0155791  0.0006856   22.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 200101  on 18738  degrees of freedom
## Residual deviance:  23589  on 18732  degrees of freedom
## AIC: 90137
##
## Number of Fisher Scoring iterations: 5
```

```
# interactions one and two
fit1 <- glm(Y~ earthquakes.year  * total.land.area, data=Area.df, family =
poisson )

fit2 <- glm(Y~ earthquakes.year  * previous.eruption , data=Area.df, family =
poisson )


# model summeries
exp(coef(fit1))
```

```
##                      (Intercept)                 earthquakes.year 
##                        10.302900                         1.030076 
##                  total.land.area earthquakes.year:total.land.area 
##                         1.035427                         1.006672 
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = Y ~ earthquakes.year * total.land.area, family = poisson,
##     data = Area.df)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                      2.3324255  0.0034436  677.33   <2e-16 ***
## earthquakes.year                 0.0296323  0.0028524   10.39   <2e-16 ***
## total.land.area                  0.0348142  0.0007705   45.18   <2e-16 ***
## earthquakes.year:total.land.area 0.0066502  0.0006109   10.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 200101  on 18738  degrees of freedom
## Residual deviance: 195274  on 18735  degrees of freedom
## AIC: 261816
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(fit2))
```

```
##                        (Intercept)                   earthquakes.year
##                         13.7197637                          1.0956858
##                  previous.eruption earthquakes.year:previous.eruption
##                          0.3655698                          0.5495752

summary(fit2)

##
## Call:
## glm(formula = Y ~ earthquakes.year * previous.eruption, family = poisson,
##     data = Area.df)
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         2.618837   0.003597  727.98   <2e-16
***
## earthquakes.year                    0.091380   0.002910   31.40   <2e-16
***
## previous.eruption                  -1.006298   0.009297 -108.24   <2e-16
***
## earthquakes.year:previous.eruption -0.598610   0.010141  -59.03   <2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 200101  on 18738  degrees of freedom
## Residual deviance: 126388  on 18735  degrees of freedom
## AIC: 192930
##
## Number of Fisher Scoring iterations: 5

# chi square test
chi_sq_test = pchisq(23589, 18732)

print(chi_sq_test)

## [1] 1
```

b)

- In the model above I chose to fit and compare previous erruptions, urban rural, off time, earthquake per year and total land area variables. the reason I chose these variables was because they all had significant p values and seemed relevant enough to effect the number eruptions during the observed period.
- based on the summary of my fitted model above, we can see that the log count of observed value of eruptions is at approximately 3.1 when other variables are not being active (on zero).

- We estimate that for every unit increase of previous eruptions, the expected log count of observed value of eruptions will have a decrease of approximately 0.23 times

- we estimate that for every unit increase of urban rural settlement the expected log count of observed value of eruptions will increase by 1.31 times, this is not the same with urban rural urban variable which causes the expected log count of observed value of eruptions to go down by 0.558 times.

- we estimate that for every year of off time unit increase during the observation period, the expected log count of observed value of eruptions will increase by approximately 2.74 times. This means the longer the observation period, the higher the number of eruptions

- we estimate that for every unit increase of earthquakes per year, our expected log count of observed value of eruptions will increase by approximately 1.05 times.

- we estimate for every each square km increase of total land area, our expected log count of observed value of eruptions will increase by approximately 1.015 times.

Interactions:

- for this section I fitted three models that each individually effected the observed number of eruptions in an observed period of time and their interactions together and whether their interaction effected observed number of eruptions too, this was ensured by looking at the p-value ensuring that all the individual variables and interactions are significant.

- In the first model there is a summary of how earthquakes per year and total land area and their interactions effects the number of eruptions. The baseline log count of eruptions is at 10.30 and when the earthquakes per year and total land area variables interact the there is log count increase of 1.006 in the log count of the eruptions.

- there may have been more interactions between the variables as well but I struggled to find more as the interaction effects on the eruption number were no statically significant enough.

To evaluate the goodness of fit and whether our model is adequate we will need to run a chi squared test. This is done by checking the number of evaluation of $\mu \geq 5$m in which in this case there are well more than 5 of them and lastly check for the p value which in this case is 1, meaning that there is no evidence against the hypothesis that our model is incorrect making our model the right fit.

Question 2:

A)

```r
# Loading and naming the data frame
health.df <- read.csv("HealthFacilities.csv")
```

```
# creating a changing the variables with muatate function
health.corrected <- health.df %>% mutate(Y = as.integer((incidents /
Total.beds) > 0.05),
    Type_recoded = case_when(Type == 1 ~ 1,
                             Type == 2 ~ 1,
                             Type == 3 ~ 2,
                             Type == 4 ~ 2),
                      Size = case_when(d1 <= 2  ~ 'Small',
                                       d1 <= 4  ~ 'Medium',
                                       d1 <= 6 ~ 'Large'),
                      Score = factor(case_when(d2 <= 3  ~ 'Low',
                                     d2 <= 6  ~ 'Medium',
                                     d2 <= 9 ~ 'High'))) %>%
                      mutate(Type_recoded = factor(Type_recoded, levels =
c('1', '2')),
                             Size = factor(Size, levels = c('Small',
'Medium', 'Large')),
                             Score = factor(Score, levels = c('Low',
'Medium', 'High')))



summary(health.corrected [,c("Size")])

##  Small Medium  Large
##     73     87     72

levels(health.corrected$Size)

## [1] "Small"  "Medium" "Large"
```

B)

```
model.null <- glm( Y ~ 1, family=binomial , data=health.corrected)

model2 <- glm( Y ~ City+ Certified+Type_recoded+Size , family=binomial ,
data=health.corrected)

summary(model2)

##
## Call:
## glm(formula = Y ~ City + Certified + Type_recoded + Size, family =
binomial,
##     data = health.corrected)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.62910    0.39672   1.586  0.11279
## CityCityB       1.49543    0.81490   1.835  0.06649 .
## CityCityC     -1.50306    0.37049  -4.057 4.97e-05 ***
## CertifiedYes   1.02138    0.33144   3.082  0.00206 **
## Type_recoded2  0.05444    0.31620   0.172  0.86332
## SizeMedium    -0.89935    0.37120  -2.423  0.01540 *
## SizeLarge     -1.78062    0.41260  -4.316 1.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 319.13  on 231  degrees of freedom
## Residual deviance: 254.51  on 225  degrees of freedom
## AIC: 268.51
##
## Number of Fisher Scoring iterations: 4
```

```
hltest(model2)
```

```
##
##    The Hosmer-Lemeshow goodness-of-fit test
##
##  Group Size Observed  Expected
##      1   27        4  2.436875
##      2   26        4  5.060065
##      3   26        5  6.772338
##      4   19        7  5.908059
##      5   13        8  5.626993
##      6   26        9 11.602620
##      7   26       15 13.692343
##      8   24       14 16.067837
##      9   30       23 23.353082
##     10   15       15 13.479787
##
##          Statistic =  7.90149
## degrees of freedom =  8
##            p-value =  0.44315
```

```
model3 <- glm( Y ~ City+Certified +Type_recoded +  Size+Score,
family=binomial , data=health.corrected)

summary(model3)
```

```
##
## Call:
## glm(formula = Y ~ City + Certified + Type_recoded + Size + Score,
##     family = binomial, data = health.corrected)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.644287   0.500930  -1.286 0.198380
## CityCityB      1.497812   0.955118   1.568 0.116836
## CityCityC     -2.344860   0.482136  -4.863 1.15e-06 ***
## CertifiedYes   1.323790   0.397496   3.330 0.000867 ***
## Type_recoded2  0.004874   0.369015   0.013 0.989462    
## SizeMedium    -1.135121   0.428544  -2.649 0.008078 ** 
## SizeLarge     -2.637761   0.537267  -4.910 9.13e-07 ***
## ScoreMedium    1.442612   0.479411   3.009 0.002620 ** 
## ScoreHigh      3.286205   0.521362   6.303 2.92e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 319.13  on 231  degrees of freedom
## Residual deviance: 197.67  on 223  degrees of freedom
## AIC: 215.67
## 
## Number of Fisher Scoring iterations: 5
```

```
hltest(model3)
```

```
## 
##    The Hosmer-Lemeshow goodness-of-fit test
## 
##  Group Size Observed   Expected
##      1   25        0  0.3278014
##      2   24        5  1.5587682
##      3   25        5  3.5608615
##      4   27        6  7.3271219
##      5   23        5  9.1276845
##      6   26        9 13.5899092
##      7   23       19 15.8247648
##      8   23       20 18.6989648
##      9   23       22 21.2001384
##     10   13       13 12.7839853
## 
##          Statistic =  18.93928 
## degrees of freedom =  8 
##            p-value =  0.015188
```

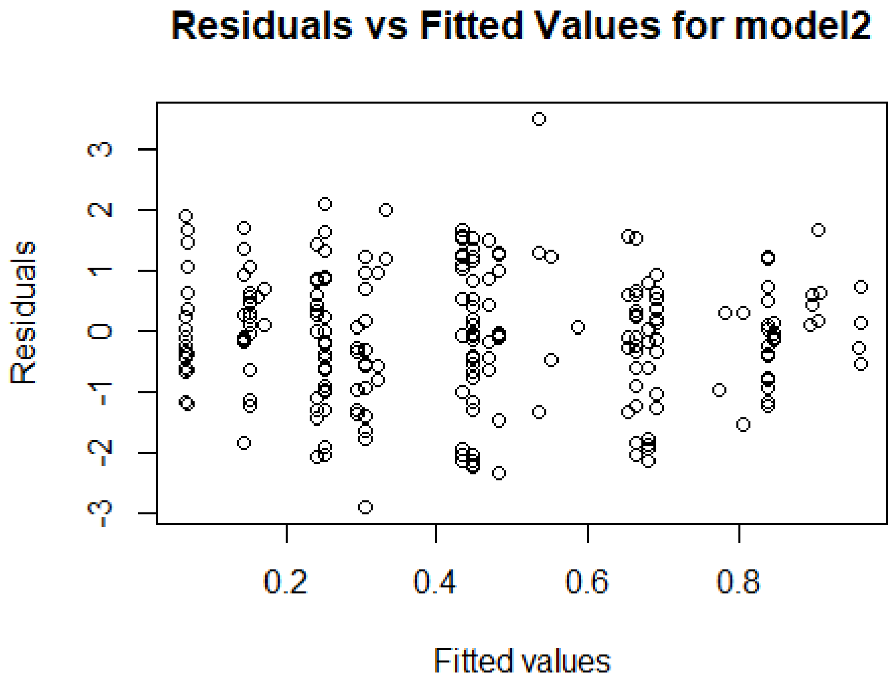
Hosmer-Lemeshow was used to fit a goodness of fit, Hosmer-Lemeshow compares the models by checking if there is a significant difference between the expected and observed values, this is ensured by looking at the p-values, In this case model 2 has a higher p-value (p = 0.44) compared to the model 3 (p = 0.015), indicating there is not a significant difference between the expected and observed values confirming the model is adequate unlike model 3 that has has a smaller p value showing there is more differences between expected and observed values.

c)

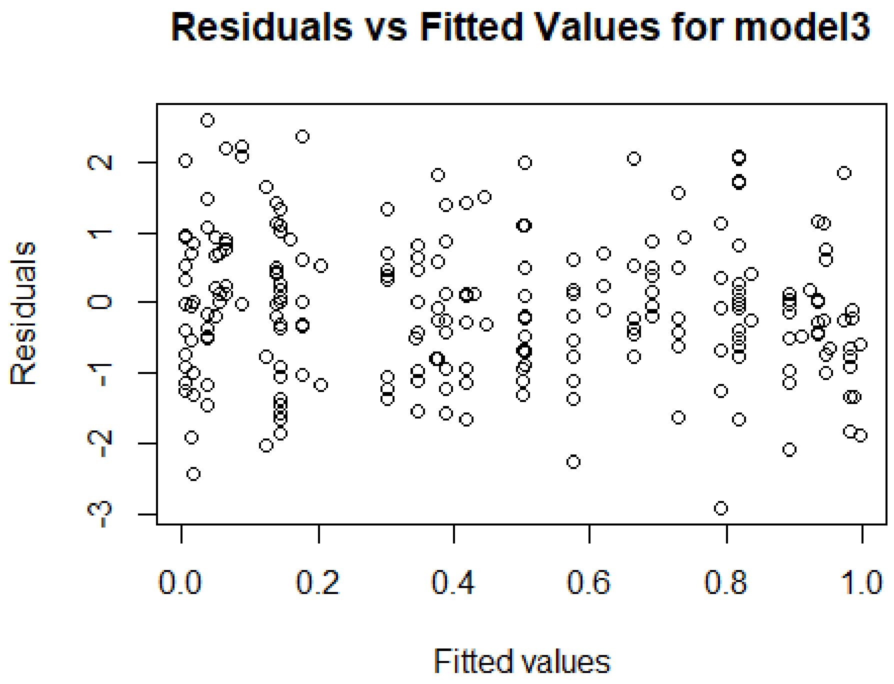
We can determine the deviance and the maximized log likelihood of a model by mainly looking at how the model is plotted and the variables in it without needing to do calculations. In this case, model 3 has the most variables. More variables usually means a decrease in deviance and an increase in the maximized likelihood of the model as there more parameters, making the fit better, which is what maximized log likelihood is measuring. So in this case, model 3 has the smallest deviance and highest maximized log likelihood compared to the more simpler models, assuming the added variables contain useful information that will help the model to better fit. The null model, which is the simplest model, has the smallest maximized log likelihood (could be because it is the simplest model and has the least variables) and the highest deviance as it lacks sufficient variables, information, and values. It is also important to check what the new variables add to the model, as more variables may make the data too cluttered and not lead to a better fit. In this case, model 3 is potentially better because the added value (Score) variable which adds more information to the data, making it a fit better.

D)

```
plot(fitted(model2), qresiduals(model2),
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs Fitted Values for model2")
```



```
plot(fitted(model3), qresiduals(model3),
     xlab = "Fitted values", ylab = "Residuals",
     main = "Residuals vs Fitted Values for model3")
```



The plots above show the deviance residuals against the fitted values. In both models the residuals are mainly spread around and close to 0 meaning the fit is adequate and there is no signs of lack of fit. Both models predictions and outcomes on the plots and fitted values match well enough to claim adequacy. The residuals in model 3 are more spread and variable in model 3 compared to model 2 covering more fitted values on the x axis with no trends and patterns, while in model 2 the values are less variable and there are no obvious trends and patterns as well.