Film Ratings Project

Logistic Regression

Brady Fisher

Introduction

For this project, I will be using data from the Film dataset in the Stat2Data package, with the goal of creating 3 models that will tell which explanatory variables are significant and are best to predict the level of "Good" for Films in the dataset.

Loading the Data

```
Film = read.csv("http://www.stat2.org/datasets/Film.csv")
```

First I have been tasked with creating an additional variable called "NorthAmerica" that is 1 if the origin of the movie is from the USA(Origin = 0), or Canada(Origin = 4), and 0 for a movie from anywhere else.

Adding the Variable

```
myData = data.frame(Film, NorthAmerica = ifelse((Film$Origin == 0 | Film$Origin == 4), 1,0))
head(myData)
```

```
##                          Title Year Time Cast Rating Description Origin
## 1            A_Ticklish_Affair 1963   89    5    2.0           7      0
## 2 Action_in_the_North_Atlantic 1943  127    7    3.0           9      0
## 3        And_the_Ship_Sails_On 1984  138    7    3.0          15      3
## 4                Autumn_Sonata 1978   97    5    3.0          11      5
## 5           Bachelor_Apartment 1931   77    6    2.5           7      0
## 6           Benson_Murder_Case 1930   69    8    2.5          10      0
##   Time_code Good NorthAmerica
## 1     short    0            1
## 2      long    1            1
## 3      long    1            0
## 4      long    1            0
## 5     short    0            1
## 6     short    0            1
```

Model 1: Chi-Square Test of Independence

Choose the Model

I want to first see if there is any relationship between the variables NorthAmerica and Good. If there is a relationship, I will use residuals or odds ratios to explain the association.

H_0: The NorthAmerica variable and the Good variable are independent, so there is no relationship between the two.

H_a: The NorthAmerica variable and the Good variable are dependent, so there is a relationship between the two.

Fit the Model

For a Chi-square Test of Independence I need to first construct a table of the observed cell counts and then the expected cell counts.

```
observedTable = table(myData$NorthAmerica, myData$Good)
observedTable
```

```
##
##      0  1
##   0 15  8
##   1 54 23
```

```
m1 = chisq.test(observedTable)
m1$expected
```

```
##
##         0     1
##   0 15.87  7.13
##   1 53.13 23.87
```

Assess the Model

Now to use the Chi-Square Test of Independence, I must check the three assumptions of: 1. Independent Random Sample 2. Large Sample Size

The first assumption of independent random sample is met because I assume the 100 films selected for the Film dataset represents a random sample of all films.

The second assumption of large sample size is met because I need the expected cell counts to be at least 5 and I saw the smallest was 7.13.

Use the Model

```
m1
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  observedTable
## X-squared = 0.036139, df = 1, p-value = 0.8492
```

From the chi-squared test of independence I see that I got a chi-squared value of 0.036139, which corresponds to a p-value of 0.8492, which is way greater than 0.05. This leads me to the conclusion of not rejecting the null hypothesis. Meaning there is NOT strong enough evidence to say NorthAmerica variable and the Good variable are associated.

Model 2: Simple Logistic Regression Model

Choose the Model

Now I want to create a Logistic Regression Model with Good as the response and NorthAmerica as the predictor, and test whether the slope for NorthAmerica is significantly different from 0.

H_0: The slope for NorthAmerica is 0.

H_a: The slope for NorthAmerica is not 0.

Fit the Model

```
m2 = glm(Good ~ NorthAmerica, family = binomial, data = myData)
```

Assess the Model

I need to check the assumptions of independent random sample and linearity for logistic regression. In this case, I again assume the data satisfies the independent random sample assumption. For the linearity assumption, since there are only 2 distinct values for the predictor NorthAmerica, this assumption is automatically satisfied, as I can always use a straight line to link 2 points.

Use the Model

```
summary(m2)
```

```
##
## Call:
## glm(formula = Good ~ NorthAmerica, family = binomial, data = myData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9246  -0.8424  -0.8424   1.4533   1.5546
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.6286     0.4378  -1.436    0.151
## NorthAmerica  -0.2249     0.5037  -0.447    0.655
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.82  on 99  degrees of freedom
## Residual deviance: 123.62  on 98  degrees of freedom
## AIC: 127.62
##
## Number of Fisher Scoring iterations: 4
```

From the summary I get a equation of logit(Good = 1 | NorthAmerica) = -0.6286 - 0.2249 * NorthAmerica. The slope for NorthAmerica has a point estimate $Beta_1$ = -.2249 and a p-value 0.655, which is greater than 0.05. So I fail to reject the null hypothesis meaning the slope is not significantly different than 0. This suggests the probability a film is "good" is not dependent of whether it was made in North America.

The results of model 1 and model 2 are consistent. They both suggest that the probability of a film being "good"" is not dependent on whether it is made in North America.

Model 3: Stepwise Selection Logistic Regression Model

Choose the Model

Now I want to include the variables Year, Time, Cast, and Description into the model besides the existing NorthAmerica in model 2. To do this I will perform a model selection to choose the predictors and interactions that give the best model based on AIC.

Fit the Model

In order to choose the best model I will start with the additive model and use a null model of only NorthAmerica as an explatory variable, and a full model of a 3 way interaction model. I will then use the step function to get the best model.

```
startingModel = glm(Good ~ NorthAmerica + Year + Time + Cast + Description,
                    data = myData, family = "binomial")
nullModel = glm(Good ~ NorthAmerica, data = myData, family = "binomial")
fullModel <- glm(Good ~ (NorthAmerica + Year + Time + Cast + Description)^3,
              data = myData, family = "binomial")
bestModel = step(startingModel, scope = list(upper = fullModel, lower = nullModel),
                 direction = "both", trace = 0)
summary(bestModel)
```

```
##
## Call:
## glm(formula = Good ~ NorthAmerica + Year + Time + Cast + Description +
##     Cast:Description + Time:Cast + NorthAmerica:Cast, family = "binomial",
##     data = myData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5210  -0.6637  -0.3279   0.5616   2.4592
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        99.31820   39.71542   2.501  0.01239 *
## NorthAmerica      -12.49716    4.53270  -2.757  0.00583 **
## Year               -0.05013    0.01961  -2.556  0.01059 *
## Time               -0.24773    0.09695  -2.555  0.01061 *
## Cast               -1.79198    1.46833  -1.220  0.22231
## Description         2.92979    0.93510   3.133  0.00173 **
## Cast:Description   -0.41520    0.13906  -2.986  0.00283 **
## Time:Cast           0.04980    0.01615   3.083  0.00205 **
## NorthAmerica:Cast   1.99373    0.73592   2.709  0.00674 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 123.820  on 99  degrees of freedom
## Residual deviance:  83.319  on 91  degrees of freedom
## AIC: 101.32
##
## Number of Fisher Scoring iterations: 6
```

Here I see my best model includes the variables NorthAmerica, Year, Time, Cast, and Description, while also including the interaction terms Cast:Description, Time:Cast and NorthAmerica:Cast.

Assess the Model

For this portion of the project I was told to accept that all assumptions were met.

Use the Model

Now I want to use the best model I found to predict the probability a film will be good given it was made in Europe in 1971, is 90 minutes long, has 5 cast members listed, and a description that is 10 lines long.

```
predict(bestModel, data.frame(NorthAmerica = 0, Year = 1971, Time = 90, Cast = 5, Description = 10))
```

```
##         1
## 0.2056993
```

```
predict(bestModel, data.frame(NorthAmerica = 0, Year = 1971, Time = 90, Cast = 5, Description = 10),
        type = "response")
```

```
##         1
## 0.5512443
```

Here I see the odds of the film being good is 0.2056993, and the probablitiy the film is good is 0.5512443 or about 55.12%.