title: "Week10"

author: "Blandon Lee"

date: "2/16/2022"

output: pdf_document

Assignment Instructions

Fit a binary logistic regression model to the data set that predicts whether the patient survived or nor for one year (the Risk1Y variable) after the surgery. Use the glm() function to perform the logistic regression. See Generalized Linear Models for an example. Include a summary using the summary() function in your results.

A. According to the summary, which variables had the greatest effect on the survival rate?

B. To compute the accuracy of your model, use the dataset to predict the outcome variable. The percent of correct predictions is the accuracy of your model. What is the accuracy of your model?

```{r}

setwd('C:/Users/bland/Desktop')

library(readr)

library(Rcmdr)

library(readxl)

```

The data needed to be placed into a CSV and some revisions made. I updated the variables to make the data easier to interpret and work with. I also revised the Risk1Yr to display 0 for false and 1 for true.

```{r}

# Imports Data/Create DataFrame

thoraric_surgery <- read.csv('csv_result-ThoracicSurgery.csv')

View(thoraric_surgery)

head(thoraric_surgery)

```

```{r}

# Summary of Data

summary(thoraric_surgery)

```

The summary in this section you will that see the Number of Fisher Scoring iterations (25) and the Null deviance (3.9210e+02 on 469 degrees of freedom). This indicates that the response variable is predicted in the model that just possesses the intercept well.

According to the summary(Survive) the variables that have the greatest negative effect on survival rates are HaemoptysisBS, Diagnosis, and Smoking. However, FEV1, Asthma,TumourSize, MI_6mo, Type2DM, and Age.Surgery were shown to have the greatest positive effects.

```{r}

# Fit a binary logistic regression model to the data set that predicts whether or not the patient survived for one year.

thoraric_surgery$predict_survival <- with(thoraric_surgery, TumourSize >= 12 & Risk1Yr >= 1)

predict_survival

```

```{r}

# glm() function to perform the logistic regression

Survive <- glm(predict_survival ~ Diagnosis+FVC+FEV1+Performance+PBF+HaemoptysisBS+DyspnoeaBS+CoughBS+WeaknessBS+TumourSize+Type2DM+ MI_6mo+ArterialDisease+Smoking+Asthma+Age.Surgery+Risk1Yr, data = thoraric_surgery, family = binomial(link = 'logit'))

```

```{r}

# Summary of Regression

summary(Survive)

```

To see the number of deaths we can take 470 (number of lines) and * Risk1Yr (mean:0.1489) and we get 69.983 deaths. The predicted_survival number showed 69 dead (True). Then we can take the predict_survival and * by deaths and that gave the 0.9859537 accuracy for the model.

```{r}

# Compute the accuracy of your model
```

summary(thoraric_surgery)

```

```{r}

# Calculate the number of Risk1Yr that died within a year

deaths <- 470*0.1489

deaths

```

```{r}

# From the summary we have 69 deaths

PredictNumber <- 69

PredictNumber

```

```{r}

# Accuracy

accuracy <- PredictNumber/deaths

accuracy

```

A. Fit a logistic regression model to the binary-classifier-data.csv dataset

B. The dataset (found in binary-classifier-data.csv) contains three variables: label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.

What is the accuracy of the logistic regression classifier?

Keep this assignment handy, as you will be comparing your results from this week to next week.

```{r}

# Loading Data

library(readr)

binary_classifier_data <- read_csv("binary-classifier-data.csv")

View(binary_classifier_data)

```
```

I fit the regression model to see if the x variable is greater than 40 and the y variable is greater than 45. I was not sure what I was supposed to place there so I just chose those numbers. I then formulated the glm() and summarized regression. In the summary I got an AIC of 404.02 and a Null deviance: 1690.23 on 1497 degrees of freedom,

```{r}

# Fit a Logistic Regression Model

binary_classifier_data$FitRegression <- with(binary_classifier_data, x >= 40 & y >= 55)

```

```{r}

summary(FitRegression)

```

```{r}

# glm() function to fit a logistic regression model

regression <- glm(FitRegression ~ label + x + y, data = binary_classifier_data, family = binomial())

```

```{r}

# Summary

summary(regression)

```

Here I again showed the summary(binary_classifier_data) and then compared the labels mean to the true percentage. The results convey that the accuracy is low. This determination was made due to a label mean of 0.488 and the TrueLabelRegression being is 0.2516689.

```{r}

# Accuracy of Regression

summary(binary_classifier_data)

```

```{r}

## Compare the mean of labels to the percentage of true

LabelRegressionValues <- 377 +1121

LabelRegressionValues
```

```

```{r}

TrueLabelRegression <- 377/1498

TrueLabelRegression

```
```