10.2 Exercise

Harlan Wittlieff

11/7/2021

1. Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset

i. Fit a binary logistic regression model to the data set that predicts whether or not the patient survived 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.

```r
# Update PRE14 to numerical values based on size
thoracic_df$PRE14 <- gsub("[a-zA-z ]", "", thoracic_df$PRE14)
thoracic_df$PRE14 <- as.numeric(thoracic_df$PRE14)

mymodel <- glm(Risk1Yr ~ PRE9 + PRE11 + PRE14, data = thoracic_df, family = 'binomial')
summary(mymodel)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE9 + PRE11 + PRE14, family = "binomial",
##     data = thoracic_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2074  -0.5365  -0.5365  -0.4001   2.2649
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.2942     2.0539  -4.525 6.03e-06 ***
## PRE9T         0.9409     0.4284   2.196 0.028070 *
## PRE11T        0.6979     0.3226   2.163 0.030541 *
## PRE14         0.6190     0.1712   3.616 0.000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 374.58  on 466  degrees of freedom
## AIC: 382.58
##
## Number of Fisher Scoring iterations: 4
```

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

From the variables I selected, PRE9 has the greatest effect size of .94, followed by PRE11, and PRE14

iii. 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?

```
res <- predict(mymodel, thoracic_df, type="response")
confmatrix <- table(Actual_Value=thoracic_df$Risk1Yr, Predicted_Value = res >0.5)
confmatrix
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            F   397    3
##            T    69    1
```

```
accuracy <- (confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)
```

The accuracy of my model is 84.7%.

1. Fit a Logistic Regression Model to Binary Classifier Dataset

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

```
mymodel_binary <- glm(label ~ x + y, data = binary_df, family = 'binomial')
summary(mymodel)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE9 + PRE11 + PRE14, family = "binomial",
##     data = thoracic_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2074  -0.5365  -0.5365  -0.4001   2.2649
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -9.2942     2.0539  -4.525 6.03e-06 ***
## PRE9T         0.9409     0.4284   2.196 0.028070 *
## PRE11T        0.6979     0.3226   2.163 0.030541 *
## PRE14         0.6190     0.1712   3.616 0.000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 374.58  on 466  degrees of freedom
## AIC: 382.58
##
## Number of Fisher Scoring iterations: 4
```

b. What is the accuracy of the logistic regression classifier?

```
res_binary <- predict(mymodel_binary, binary_df, type="response")
confmatrix_binary <- table(Actual_Value=binary_df$label, Predicted_Value = res_binary >0.5)
confmatrix_binary
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   429  338
##            1   286  445
```

```
accuracy_binary <- (confmatrix_binary[[1,1]] + confmatrix_binary[[2,2]]) / sum(confmatrix_binary)
accuracy_binary
```

```
## [1] 0.5834446
```

The accuracy is 58.3%.