thoracic_logistic_exercise_PhillipsEmily

Emily Phillips

8/3/2021

Number 1

Reading in the Thoracic Surgey CSV file

```
##   ï..DGN PRE4 PRE5 PRE6  PRE7  PRE8  PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25
## 1   DGN2 2.88 2.16 PRZ1 FALSE FALSE FALSE  TRUE  TRUE  OC14 FALSE FALSE FALSE
## 2   DGN3 3.40 1.88 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC12 FALSE FALSE FALSE
## 3   DGN3 2.76 2.08 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
## 4   DGN3 3.68 3.04 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC11 FALSE FALSE FALSE
## 5   DGN3 2.44 0.96 PRZ2 FALSE  TRUE FALSE  TRUE  TRUE  OC11 FALSE FALSE FALSE
## 6   DGN3 2.48 1.88 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
##   PRE30 PRE32 AGE Risk1Yr
## 1  TRUE FALSE  60   FALSE
## 2  TRUE FALSE  51   FALSE
## 3  TRUE FALSE  59   FALSE
## 4 FALSE FALSE  54   FALSE
## 5  TRUE FALSE  73    TRUE
## 6 FALSE FALSE  51   FALSE
```

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.

```
#FALSE will be taken as the intial baseline which is good because this represents that the individual d
#one-year survival period
#Therefore, our model coefficients will reflect the probability of surviving rather than the probabilit

thoracicModel.1 <- glm(Risk1Yr ~ ï..DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE
summary(thoracicModel.1)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ ï..DGN + PRE4 + PRE5 + PRE6 + PRE7 +
##     PRE8 + PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 +
##     PRE30 + PRE32 + AGE, family = binomial(), data = thoracic_df)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.6084  -0.5439  -0.4199  -0.2762   2.4929
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.655e+01  2.400e+03  -0.007  0.99450
## ï..DGNDGN2   1.474e+01  2.400e+03   0.006  0.99510
## ï..DGNDGN3   1.418e+01  2.400e+03   0.006  0.99528
## ï..DGNDGN4   1.461e+01  2.400e+03   0.006  0.99514
## ï..DGNDGN5   1.638e+01  2.400e+03   0.007  0.99455
## ï..DGNDGN6   4.089e-01  2.673e+03   0.000  0.99988
## ï..DGNDGN8   1.803e+01  2.400e+03   0.008  0.99400
## PRE4        -2.272e-01  1.849e-01  -1.229  0.21909
## PRE5        -3.030e-02  1.786e-02  -1.697  0.08971 .
## PRE6PRZ1    -4.427e-01  5.199e-01  -0.852  0.39448
## PRE6PRZ2    -2.937e-01  7.907e-01  -0.371  0.71030
## PRE7TRUE     7.153e-01  5.556e-01   1.288  0.19788
## PRE8TRUE     1.743e-01  3.892e-01   0.448  0.65419
## PRE9TRUE     1.368e+00  4.868e-01   2.811  0.00494 **
## PRE10TRUE    5.770e-01  4.826e-01   1.196  0.23185
## PRE11TRUE    5.162e-01  3.965e-01   1.302  0.19295
## PRE14OC12    4.394e-01  3.301e-01   1.331  0.18318
## PRE14OC13    1.179e+00  6.165e-01   1.913  0.05580 .
## PRE14OC14    1.653e+00  6.094e-01   2.713  0.00668 **
## PRE17TRUE    9.266e-01  4.445e-01   2.085  0.03709 *
## PRE19TRUE   -1.466e+01  1.654e+03  -0.009  0.99293
## PRE25TRUE   -9.789e-02  1.003e+00  -0.098  0.92227
## PRE30TRUE    1.084e+00  4.990e-01   2.172  0.02984 *
## PRE32TRUE   -1.398e+01  1.645e+03  -0.008  0.99322
## AGE         -9.506e-03  1.810e-02  -0.525  0.59944
## ---
## 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: 341.19  on 445  degrees of freedom
## AIC: 391.19
##
## Number of Fisher Scoring iterations: 15
```

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

The variables that had the greatest effect on the survival rate can be found by assessing whether their z-statistic was significant at less than p = 0.05. For this model, those variables are the following: PRE9TRUE, PRE14OC14, PRE17TRUE, PRE30TRUE.

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?

```
#Find probability for each observation
thoracic_df$model_prob <- predict(thoracicModel.1, thoracic_df, type = "response")

#Transform the probabilities into success & failures, 1s & 0s
#risk_binary turns "TRUE" into 1 and "FALSE" into 0
thoracic_df <- thoracic_df  %>% mutate(model_pred = 1*(model_prob > .53) + 0,
                                       risk_binary = 1*(Risk1Yr == "TRUE") + 0)

#Calculate the accuracy of the model, compare model_pred and risk_binary
thoracic_df <- thoracic_df %>% mutate(accurate = 1*(model_pred == risk_binary))
sum(thoracic_df$accurate)/nrow(thoracic_df)
```

```
## [1] 0.8361702
```

The accuracy of our model is 83.6%, which is the percent of correct predictions that came from our model for the Risk1Yr outcome variable.

Number 2

The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.

```
##   label        x        y
## 1     0 70.88469 83.17702
## 2     0 74.97176 87.92922
## 3     0 73.78333 92.20325
## 4     0 66.40747 81.10617
## 5     0 69.07399 84.53739
## 6     0 72.23616 86.38403
```

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

```
binaryModel.1 <- glm(label ~ x + y, data = binary_df,family = binomial())
summary(binaryModel.1)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial(), data = binary_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

What is the accuracy of the logistic regression classifier?

```
#Find probability for each observation
binary_df$model_prob <- predict(binaryModel.1, binary_df, type = "response")

#Transform the probabilities into success & failures, 1s & 0s
#risk_binary turns "TRUE" into 1 and "FALSE" into 0
binary_df <- binary_df  %>% mutate(model_pred = 1*(model_prob > .53) + 0,
                                   label_binary = 1*(label == 1) + 0)

#Calculate the accuracy of the model, compare model_pred and risk_binary
binary_df <- binary_df %>% mutate(accurate = 1*(model_pred == label_binary))
sum(binary_df$accurate)/nrow(binary_df)
```

```
## [1] 0.4706275
```

The accuracy of the logistic regression classifier is 47.1%.