Assignment 9.2

Kimberly Adams

08/7/2022

Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset

For this problem, you will be working with the thoracic surgery data set from the University of California Irvine machine learning repository. This dataset contains information on life expectancy in lung cancer patients after surgery.

The underlying thoracic surgery data is in ARFF format. This is a text-based format with information on each of the attributes. You can load this data using a package such as foreign or by cutting and pasting the data section into a CSV file.

0 = False 1 = True

Variables Definitions (for reference):

1. DGN: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
2. PRE4: Forced vital capacity - FVC (numeric)
3. PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
4. PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)
5. PRE7: Pain before surgery (T,F)
6. PRE8: Haemoptysis before surgery (T,F)
7. PRE9: Dyspnoea before surgery (T,F)
8. PRE10: Cough before surgery (T,F)
9. PRE11: Weakness before surgery (T,F)
10. PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
11. PRE17: Type 2 DM - diabetes mellitus (T,F)
12. PRE19: MI up to 6 months (T,F)
13. PRE25: PAD - peripheral arterial diseases (T,F)
14. PRE30: Smoking (T,F)
15. PRE32: Asthma (T,F)
16. AGE: Age at surgery (numeric)
17. Risk1Yr: 1 year survival period - (T)rue value if died (T,F)

Assignment Instructions:

1. 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.

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE30 + PRE7, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7798  -0.5849  -0.5849  -0.4001   2.2651
##
## Coefficients:
##             Estimate Std. Error z value      Pr(>|z|)
## (Intercept)  -2.4852     0.4029  -6.168 0.00000000069 ***
## PRE30         0.8062     0.4209   1.916        0.0554 .
## PRE7          0.6442     0.4566   1.411        0.1583
## ---
## 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: 389.89  on 467  degrees of freedom
## AIC: 395.89
##
## Number of Fisher Scoring iterations: 5
```

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

Of the variables tried, PRE30 (Smoking) had the highest correlation coefficient of 0.81 followed by PRE7 (Pain before surgery, 0.64). Smoking is significant at a 90% confidence level.

3. 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?

```
##             Predicted_value
## Actual_Value FALSE
##            0   400
##            1    70
```

According to the confusion matrix, the model predicts that everyone would die, but there are 70 cases in which the patient actually lived. This means that the accuracy is $400/470 = 85\%$. Although this value is high, I personally have some significant doubts on this model due to the fact that it did not predict even 1 survival case. More study is needed.

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

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?

```
##
```

```
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = BinClass.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 0.0000207 ***
## ---
## 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
```

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

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

The results show that this model is only about 58% accurate.