assignment_week10_kanaparthiVenkata

Venkata Kanaparthi

5/22/2021

```
## Warning: package 'ggm' was built under R version 4.0.5

## Warning: package 'plyr' was built under R version 4.0.5

## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following objects are masked from 'package:pastecs':
##
##     first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'magrittr'

## The following object is masked from 'package:pastecs':
##
##     extract

## Warning: package 'purrr' was built under R version 4.0.5

##
## Attaching package: 'purrr'
## The following object is masked from 'package:magrittr':
##
##     set_names

## The following object is masked from 'package:plyr':
##
##     compact

## Warning: package 'stringr' was built under R version 4.0.5

## Warning: package 'QuantPsyc' was built under R version 4.0.5

## Loading required package: boot

## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':
##
##     norm

## Warning: package 'caTools' was built under R version 4.0.5

## Warning: package 'survival' was built under R version 4.0.5

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml
```

Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset

a. 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.

Set the working directory to the root of your DSC 520 directory

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 ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +
##     PRE32 + AGE, family = binomial(), data = thoracicData_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
## DGNDGN2      1.474e+01  2.400e+03   0.006  0.99510
## DGNDGN3      1.418e+01  2.400e+03   0.006  0.99528
## DGNDGN4      1.461e+01  2.400e+03   0.006  0.99514
## DGNDGN5      1.638e+01  2.400e+03   0.007  0.99455
## DGNDGN6      4.089e-01  2.673e+03   0.000  0.99988
## 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
```

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

Ans : PRE14 and PRE9 has the most significant impact.

```
##             Predicted_value
## Actual_value FALSE TRUE
##        FALSE   390   10
##        TRUE     67    3
```

```
## [1] "Accuracy of the model :  83.62 %"
```

Fit a Logistic Regression Model

1.) Fit a logistic regression model to the binary-classifier-data.csv dataset 2.) 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. a.) What is the accuracy of the logistic regression classifier? b.) Keep this assignment handy, as you will be comparing your results from this week to next week.

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial(), data = binaryClassifier_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
```

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

```
## [1] "Accuracy of the model :  58.34 %"
```