RMarkdown Assignment # 13 - Week 08

Pushkar Chougule

Oct 22nd 2020

## Note - Remove example code and comments before submitting assignment. Producing a professional R Markdown document is the goal.

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

##Read the ThoracicSurgery data and use summary() on the dataframe

library(foreign)  
  
surgery\_df <- read.arff('ThoraricSurgery.arff')  
  
str(surgery\_df)

## 'data.frame': 470 obs. of 17 variables:  
## $ DGN : Factor w/ 7 levels "DGN1","DGN2",..: 2 3 3 3 3 3 3 2 3 3 ...  
## $ PRE4 : num 2.88 3.4 2.76 3.68 2.44 2.48 4.36 3.19 3.16 2.32 ...  
## $ PRE5 : num 2.16 1.88 2.08 3.04 0.96 1.88 3.28 2.5 2.64 2.16 ...  
## $ PRE6 : Factor w/ 3 levels "PRZ0","PRZ1",..: 2 1 2 1 3 2 2 2 3 2 ...  
## $ PRE7 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 1 1 1 1 ...  
## $ PRE8 : Factor w/ 2 levels "F","T": 1 1 1 1 2 1 1 1 1 1 ...  
## $ PRE9 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 1 1 1 1 ...  
## $ PRE10 : Factor w/ 2 levels "F","T": 2 1 2 1 2 2 2 2 2 2 ...  
## $ PRE11 : Factor w/ 2 levels "F","T": 2 1 1 1 2 1 1 1 2 1 ...  
## $ PRE14 : Factor w/ 4 levels "OC11","OC12",..: 4 2 1 1 1 1 2 1 1 1 ...  
## $ PRE17 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 2 1 1 1 ...  
## $ PRE19 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 1 1 1 1 ...  
## $ PRE25 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 1 2 1 1 ...  
## $ PRE30 : Factor w/ 2 levels "F","T": 2 2 2 1 2 1 2 2 2 2 ...  
## $ PRE32 : Factor w/ 2 levels "F","T": 1 1 1 1 1 1 1 1 1 1 ...  
## $ AGE : num 60 51 59 54 73 51 59 66 68 54 ...  
## $ Risk1Yr: Factor w/ 2 levels "F","T": 1 1 1 1 2 1 2 2 1 1 ...

summary(surgery\_df)

## DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9   
## DGN1: 1 Min. :1.440 Min. : 0.960 PRZ0:130 F:439 F:402 F:439   
## DGN2: 52 1st Qu.:2.600 1st Qu.: 1.960 PRZ1:313 T: 31 T: 68 T: 31   
## DGN3:349 Median :3.160 Median : 2.400 PRZ2: 27   
## DGN4: 47 Mean :3.282 Mean : 4.569   
## DGN5: 15 3rd Qu.:3.808 3rd Qu.: 3.080   
## DGN6: 4 Max. :6.300 Max. :86.300   
## DGN8: 2   
## PRE10 PRE11 PRE14 PRE17 PRE19 PRE25 PRE30 PRE32   
## F:147 F:392 OC11:177 F:435 F:468 F:462 F: 84 F:468   
## T:323 T: 78 OC12:257 T: 35 T: 2 T: 8 T:386 T: 2   
## OC13: 19   
## OC14: 17   
##   
##   
##   
## AGE Risk1Yr  
## Min. :21.00 F:400   
## 1st Qu.:57.00 T: 70   
## Median :62.00   
## Mean :62.53   
## 3rd Qu.:69.00   
## Max. :87.00   
##

##split data into test and train, create models, summary() of model

**a. 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.**

library(caTools)  
  
# split the dataset into two for training and testing purpose  
  
split <- sample.split(surgery\_df, SplitRatio = 0.8)  
split

## [1] TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  
## [13] TRUE TRUE FALSE TRUE FALSE

train <- subset(surgery\_df, split == "TRUE")  
test <- subset(surgery\_df, split == "FALSE")  
  
#log\_reg\_all <- glm(Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + #PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE, data = train, family = #binomial(link = "logit"))  
  
log\_reg\_all <- glm(Risk1Yr ~ ., data = surgery\_df, family = "binomial")  
  
summary(log\_reg\_all)

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = "binomial", data = surgery\_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   
## PRE7T 7.153e-01 5.556e-01 1.288 0.19788   
## PRE8T 1.743e-01 3.892e-01 0.448 0.65419   
## PRE9T 1.368e+00 4.868e-01 2.811 0.00494 \*\*  
## PRE10T 5.770e-01 4.826e-01 1.196 0.23185   
## PRE11T 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 \*\*  
## PRE17T 9.266e-01 4.445e-01 2.085 0.03709 \*   
## PRE19T -1.466e+01 1.654e+03 -0.009 0.99293   
## PRE25T -9.789e-02 1.003e+00 -0.098 0.92227   
## PRE30T 1.084e+00 4.990e-01 2.172 0.02984 \*   
## PRE32T -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

exp(log\_reg\_all$coefficients)

## (Intercept) DGNDGN2 DGNDGN3 DGNDGN4 DGNDGN5 DGNDGN6   
## 6.481697e-08 2.511211e+06 1.440574e+06 2.209615e+06 1.301120e+07 1.505091e+00   
## DGNDGN8 PRE4 PRE5 PRE6PRZ1 PRE6PRZ2 PRE7T   
## 6.785355e+07 7.967257e-01 9.701510e-01 6.422903e-01 7.454996e-01 2.044884e+00   
## PRE8T PRE9T PRE10T PRE11T PRE14OC12 PRE14OC13   
## 1.190456e+00 3.928338e+00 1.780613e+00 1.675616e+00 1.551720e+00 3.251796e+00   
## PRE14OC14 PRE17T PRE19T PRE25T PRE30T PRE32T   
## 5.222483e+00 2.525890e+00 4.317676e-07 9.067446e-01 2.956473e+00 8.455364e-07   
## AGE   
## 9.905394e-01

exp(confint(log\_reg\_all))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) NA 1.861968e+203  
## DGNDGN2 1.717218e-206 NA  
## DGNDGN3 8.098202e-207 NA  
## DGNDGN4 1.675824e-206 NA  
## DGNDGN5 1.264378e-205 NA  
## DGNDGN6 2.066233e-24 2.939776e+20  
## DGNDGN8 5.691816e-171 NA  
## PRE4 5.499148e-01 1.138007e+00  
## PRE5 9.264310e-01 9.993543e-01  
## PRE6PRZ1 2.300552e-01 1.783025e+00  
## PRE6PRZ2 1.540289e-01 3.470770e+00  
## PRE7T 6.558696e-01 5.928649e+00  
## PRE8T 5.383928e-01 2.497049e+00  
## PRE9T 1.466379e+00 1.007288e+01  
## PRE10T 7.094170e-01 4.740878e+00  
## PRE11T 7.532542e-01 3.596887e+00  
## PRE14OC12 8.231331e-01 3.022655e+00  
## PRE14OC13 9.225453e-01 1.064690e+01  
## PRE14OC14 1.540476e+00 1.723680e+01  
## PRE17T 1.017658e+00 5.900292e+00  
## PRE19T NA 1.949037e+106  
## PRE25T 9.525986e-02 5.459928e+00  
## PRE30T 1.197920e+00 8.705307e+00  
## PRE32T NA 8.570374e+105  
## AGE 9.561182e-01 1.026545e+00

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

According to summary(), following variables have Pr(>|z|) value less than or closer to 0.05

1. PRE9T (Dyspnoea before surgery)

2. PRE14OC14 (PRE14 is the column and OC14 is the largest size of the original tumor)

3. PRE17T (Type 2 DM - diabetes mellitus)

4. PRE30T (Smoking)

Couple other variables below have Pr(>|z|) value just above 0.05 but below 0.10, which can also be considered as the predictors, if needed, to improve model further

1. PRE14OC13 (PRE14 is the column and OC13 is the large size of the original tumor)

2. PRE5 (Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric))

Odds ratio for above mentioned variables are greater than 1 as well, confirming that as predictors change, the odds of outcome changing increase, as per (Field, Miles, and Field 2012).

**c. 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?**

#Calculating accuracy for model with all variables and evaluate against the test data  
  
result <- predict(log\_reg\_all, test, type="response")  
  
confusion\_matrix <- table(Actual\_Value=test$Risk1Yr, Predicted\_Value = result > 0.5)  
confusion\_matrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 97 1  
## T 10 1

#Accuracy calculation based on confusion matrix  
accuracy = (confusion\_matrix[[1,1]] + confusion\_matrix[[2,2]])/sum(confusion\_matrix) \* 100  
accuracy

## [1] 89.90826

The above results indicate, based on the test dataset (subset of original dataset), that the model accuracy is 89.91%. There are some False negative outcomes from the model. But it can be improved with additional data points or by using more relevant predictor variables or the combination of the two.

## References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. <https://books.google.com/books?id=wd2K2zC3swIC>.