DSC_520_week9_Assignment00

Reenie Christudass

2022-08-15

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)

PRE4: Forced vital capacity - FVC (numeric)

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

PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)

PRE7: Pain before surgery (T,F)

PRE8: Haemoptysis before surgery (T,F)

PRE9: Dyspnoea before surgery (T,F)

PRE10: Cough before surgery (T,F)

PRE11: Weakness before surgery (T,F)

PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)

PRE17: Type 2 DM - diabetes mellitus (T,F)

PRE19: MI up to 6 months (T,F)

PRE25: PAD - peripheral arterial diseases (T,F)

PRE30: Smoking (T,F)

PRE32: Asthma (T,F)

AGE: Age at surgery (numeric)

Risk1Y: 1 year survival period - (T)rue value if died (T,F)

Load Libraries

```r
if(!require('foreign')) {
  install.packages('foreign')
  library('foreign')
}
```

```
## Loading required package: foreign
```

```r
if(!require('tidyr')) {
  install.packages('tidyr')
  library('tidyr')
}
```

```
## Loading required package: tidyr
```

```
## Warning: package 'tidyr' was built under R version 4.2.1
```

```
install.packages("MASS", repos="http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/chris/AppData/Local/R/win-library/4.2'
## (as 'lib' is unspecified)
```

```
## package 'MASS' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\chris\AppData\Local\Temp\RtmpmkPE9C\downloaded_packages
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.1
```

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/chris/dsc520/data")
```

```
## Load the `data/r4ds/heights.csv` to
df <- read.arff("C:/Users/chris/dsc520/data/ThoraricSurgery.arff")
head(df)
```

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

```
data_new <- sapply(df, unclass)           # Convert categorical variables
head(data_new)
```

```
##      DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25
## [1,]   2 2.88 2.16    2    1    1    1     2     2     4     1     1     1
## [2,]   3 3.40 1.88    1    1    1    1     1     1     2     1     1     1
## [3,]   3 2.76 2.08    2    1    1    1     2     1     1     1     1     1
## [4,]   3 3.68 3.04    1    1    1    1     1     1     1     1     1     1
## [5,]   3 2.44 0.96    3    1    2    1     2     2     1     1     1     1
## [6,]   3 2.48 1.88    2    1    1    1     2     1     1     1     1     1
##      PRE30 PRE32 AGE Risk1Yr
## [1,]     2     1  60       1
## [2,]     2     1  51       1
## [3,]     2     1  59       1
## [4,]     1     1  54       1
## [5,]     2     1  73       2
## [6,]     1     1  51       1
```

```
# convert the matrix into dataframe
newdata=as.data.frame(data_new)
head(newdata)
```

```
##   DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25 PRE30
## 1   2 2.88 2.16    2    1    1    1     2     2     4     1     1     1     2
## 2   3 3.40 1.88    1    1    1    1     1     1     2     1     1     1     2
## 3   3 2.76 2.08    2    1    1    1     2     1     1     1     1     1     2
## 4   3 3.68 3.04    1    1    1    1     1     1     1     1     1     1     1
## 5   3 2.44 0.96    3    1    2    1     2     2     1     1     1     1     2
## 6   3 2.48 1.88    2    1    1    1     2     1     1     1     1     1     1
##   PRE32 AGE Risk1Yr
## 1     1  60       1
## 2     1  51       1
## 3     1  59       1
## 4     1  54       1
## 5     1  73       2
## 6     1  51       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.

newdata2 <-newdata[,c("DGN","PRE4","PRE5","PRE6","PRE7","PRE8","PRE9","PRE11","PRE14","PRE17","PRE19","
                      ,"PRE32","AGE","Risk1Yr")]

riskmodel<-glm(as.factor(Risk1Yr)~DGN+PRE4+PRE5+PRE6+PRE7+PRE8+PRE9+PRE11+PRE14+PRE17+PRE19+PRE25+PRE30+
                family=binomial,data=newdata2)
summary(riskmodel)
```

```
##
## Call:
## glm(formula = as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE6 +
##     PRE7 + PRE8 + PRE9 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 +
##     PRE30 + PRE32 + AGE, family = binomial, data = newdata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5778  -0.5689  -0.4405  -0.3213   2.4665
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    18.14865 1391.66427   0.013 0.989595
## DGN             0.46286    0.19017   2.434 0.014938 *
## PRE4          -0.18753    0.17465  -1.074 0.282923
## PRE5          -0.02177    0.01673  -1.302 0.192990
## PRE6          -0.01923    0.30876  -0.062 0.950352
## PRE7           0.45697    0.51184   0.893 0.371973
## PRE8           0.33550    0.37456   0.896 0.370399
## PRE9           1.27502    0.47405   2.690 0.007153 **
## PRE11          0.63815    0.37492   1.702 0.088741 .
## PRE14          0.68003    0.18320   3.712 0.000206 ***
## PRE17          0.85492    0.43012   1.988 0.046850 *
## PRE19        -13.82120  984.05000  -0.014 0.988794
## PRE25          0.11986    0.92301   0.130 0.896683
## PRE30          0.91929    0.45608   2.016 0.043837 *
## PRE32        -13.20624  984.05778  -0.013 0.989293
## AGE           -0.01012    0.01697  -0.596 0.551017
## ---
## 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: 355.50  on 454  degrees of freedom
## AIC: 387.5
##
## Number of Fisher Scoring iterations: 14
```

```
##VARIABLE SELECTION
riskmodel_new <- stepAIC(riskmodel)
```

```
## Start:  AIC=387.5
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 +
##     AGE
##
##         Df Deviance    AIC
## - PRE6   1   355.50 385.50
## - PRE25  1   355.52 385.52
## - AGE    1   355.86 385.86
## - PRE32  1   355.88 385.88
## - PRE19  1   356.17 386.17
## - PRE7   1   356.26 386.26
## - PRE8   1   356.28 386.28
## - PRE4   1   356.67 386.67
## <none>       355.50 387.50
## - PRE5   1   357.81 387.81
## - PRE11  1   358.31 388.31
## - PRE17  1   359.13 389.13
## - PRE30  1   360.23 390.23
## - DGN    1   360.99 390.99
## - PRE9   1   362.05 392.05
## - PRE14  1   369.21 399.21
##
## Step:  AIC=385.5
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 +
##     PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE
```

```
##
##         Df Deviance    AIC
## - PRE25  1   355.52 383.52
## - AGE    1   355.87 383.87
## - PRE32  1   355.89 383.89
## - PRE19  1   356.18 384.18
## - PRE7   1   356.27 384.27
## - PRE8   1   356.28 384.28
## - PRE4   1   356.67 384.67
## <none>       355.50 385.50
## - PRE5   1   357.87 385.87
## - PRE11  1   358.75 386.75
## - PRE17  1   359.14 387.14
## - PRE30  1   360.34 388.34
## - DGN    1   360.99 388.99
## - PRE9   1   362.18 390.18
## - PRE14  1   369.34 397.34
##
## Step:  AIC=383.52
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 +
##     PRE11 + PRE14 + PRE17 + PRE19 + PRE30 + PRE32 + AGE
##
##         Df Deviance    AIC
## - AGE    1   355.89 381.89
## - PRE32  1   355.90 381.90
## - PRE19  1   356.19 382.19
## - PRE7   1   356.27 382.27
## - PRE8   1   356.35 382.35
## - PRE4   1   356.70 382.70
## <none>       355.52 383.52
## - PRE5   1   357.91 383.91
## - PRE11  1   358.77 384.77
## - PRE17  1   359.20 385.20
## - PRE30  1   360.43 386.43
## - DGN    1   361.03 387.03
## - PRE9   1   362.37 388.37
## - PRE14  1   369.35 395.35
##
## Step:  AIC=381.89
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 +
##     PRE11 + PRE14 + PRE17 + PRE19 + PRE30 + PRE32
##
##         Df Deviance    AIC
## - PRE32  1   356.26 380.26
## - PRE19  1   356.52 380.52
## - PRE7   1   356.62 380.62
## - PRE8   1   356.67 380.67
## - PRE4   1   356.81 380.81
## <none>       355.89 381.89
## - PRE5   1   358.13 382.13
## - PRE11  1   358.83 382.83
## - PRE17  1   359.45 383.45
## - PRE30  1   360.69 384.69
## - DGN    1   361.09 385.09
## - PRE9   1   362.51 386.51
## - PRE14  1   369.69 393.69
##
## Step:  AIC=380.26
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 +
##     PRE11 + PRE14 + PRE17 + PRE19 + PRE30
##
##         Df Deviance    AIC
## - PRE19  1   356.89 378.89
## - PRE7   1   357.00 379.00
## - PRE8   1   357.06 379.06
## - PRE4   1   357.14 379.14
## <none>       356.26 380.26
## - PRE5   1   358.49 380.49
## - PRE11  1   359.25 381.25
## - PRE17  1   359.86 381.86
## - PRE30  1   361.11 383.11
## - DGN    1   361.48 383.48
## - PRE9   1   362.92 384.92
## - PRE14  1   370.08 392.08
##
## Step:  AIC=378.89
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 +
##     PRE11 + PRE14 + PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE7   1   357.63 377.63
## - PRE8   1   357.73 377.73
## - PRE4   1   357.76 377.76
## <none>       356.89 378.89
## - PRE5   1   359.13 379.13
## - PRE11  1   359.74 379.74
## - PRE17  1   360.56 380.56
## - PRE30  1   361.72 381.72
## - DGN    1   362.14 382.14
## - PRE9   1   363.57 383.57
## - PRE14  1   370.75 390.75
##
## Step:  AIC=377.63
## as.factor(Risk1Yr) ~ DGN + PRE4 + PRE5 + PRE8 + PRE9 + PRE11 +
##     PRE14 + PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE4   1   358.45 376.45
## - PRE8   1   358.90 376.90
## - PRE5   1   359.54 377.54
## <none>       357.63 377.63
## - PRE11  1   360.25 378.25
## - PRE17  1   361.42 379.42
## - PRE30  1   362.26 380.26
## - DGN    1   363.13 381.13
## - PRE9   1   364.39 382.39
## - PRE14  1   371.99 389.99
##
```

```
## Step:  AIC=376.45
## as.factor(Risk1Yr) ~ DGN + PRE5 + PRE8 + PRE9 + PRE11 + PRE14 +
##     PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE8   1   359.93 375.93
## <none>       358.45 376.45
## - PRE5   1   360.46 376.46
## - PRE11  1   361.35 377.35
## - PRE17  1   362.75 378.75
## - PRE30  1   363.16 379.16
## - DGN    1   363.65 379.65
## - PRE9   1   365.05 381.05
## - PRE14  1   372.51 388.51
##
## Step:  AIC=375.93
## as.factor(Risk1Yr) ~ DGN + PRE5 + PRE9 + PRE11 + PRE14 + PRE17 +
##     PRE30
##
##         Df Deviance    AIC
## - PRE5   1   361.65 375.65
## <none>       359.93 375.93
## - PRE11  1   363.03 377.03
## - PRE17  1   364.36 378.36
## - PRE30  1   364.42 378.42
## - DGN    1   364.75 378.75
## - PRE9   1   367.27 381.27
## - PRE14  1   373.99 387.99
##
## Step:  AIC=375.65
## as.factor(Risk1Yr) ~ DGN + PRE9 + PRE11 + PRE14 + PRE17 + PRE30
##
##         Df Deviance    AIC
## <none>       361.65 375.65
## - PRE11  1   365.08 377.08
## - PRE17  1   366.15 378.15
## - DGN    1   366.42 378.42
## - PRE30  1   366.63 378.63
## - PRE9   1   367.94 379.94
## - PRE14  1   375.79 387.79
```

```
##CONCLUSION:
##At the very last step stepAIC has produced the optimal set of features {DGN + PRE9 + PRE11 + PRE14 +
##    PRE17 + PRE30}. stepAIC also removes the Multicollinearity.
```

```
summary(riskmodel_new)
```

```
##
## Call:
## glm(formula = as.factor(Risk1Yr) ~ DGN + PRE9 + PRE11 + PRE14 +
##     PRE17 + PRE30, family = binomial, data = newdata2)
##
## Deviance Residuals:
```

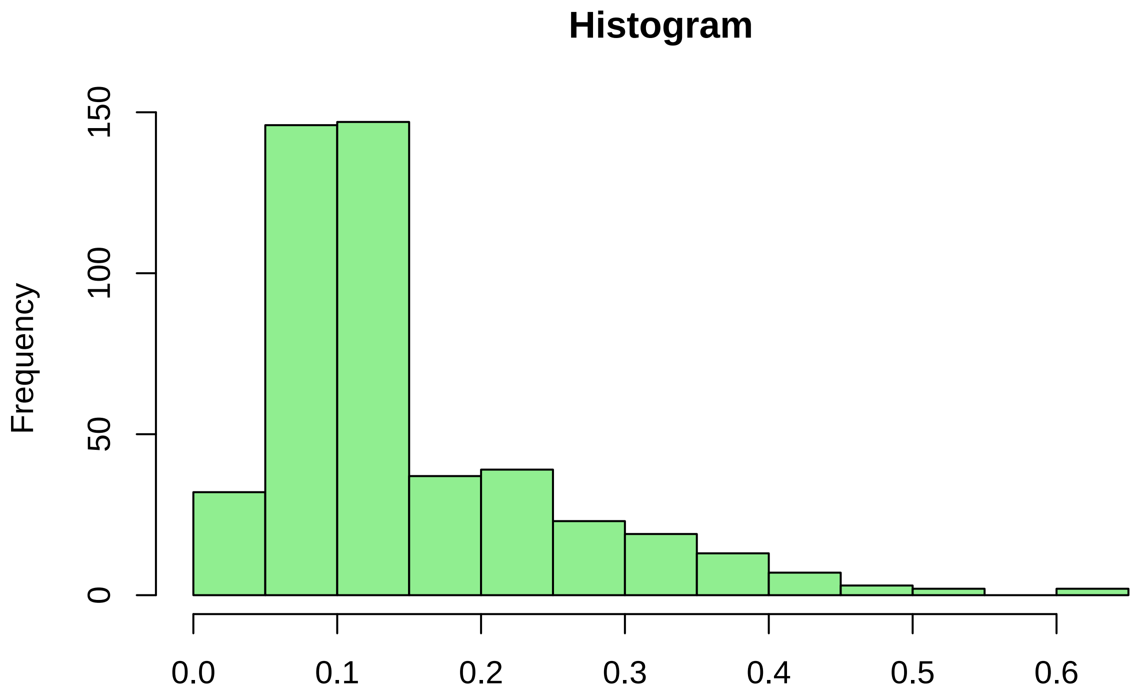
```
##     Min       1Q   Median       3Q      Max  
## -1.3552  -0.5313  -0.4369  -0.3434   2.4622  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -9.0559     1.5356  -5.897  3.7e-09 ***
## DGN           0.4146     0.1828   2.268 0.023317 *  
## PRE9          1.1762     0.4411   2.666 0.007668 ** 
## PRE11         0.6251     0.3287   1.901 0.057240 .  
## PRE14         0.6808     0.1795   3.793 0.000149 ***
## PRE17         0.9338     0.4193   2.227 0.025954 *  
## PRE30         0.9145     0.4448   2.056 0.039772 *  
## ---
## 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: 361.65  on 463  degrees of freedom
## AIC: 375.65
## 
## Number of Fisher Scoring iterations: 5
```

```
##Analysis of the outcome

summary(newdata2$fitted.values)
```

```
## Length  Class   Mode 
##      0   NULL   NULL
```

```
hist(riskmodel_new$fitted.values,main = " Histogram ",xlab = "", col = 'light green')
```



```
newdata2$Predict <- ifelse(riskmodel_new$fitted.values >0.5,"Survived","Not Survive")
head(newdata2)
```

```
##   DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9 PRE11 PRE14 PRE17 PRE19 PRE25 PRE30 PRE32
## 1   2 2.88 2.16    2    1    1    1     2     4     1     1     1     2     1
## 2   3 3.40 1.88    1    1    1    1     1     2     1     1     1     2     1
## 3   3 2.76 2.08    2    1    1    1     1     1     1     1     1     2     1
## 4   3 3.68 3.04    1    1    1    1     1     1     1     1     1     1     1
## 5   3 2.44 0.96    3    1    2    1     2     1     1     1     1     2     1
## 6   3 2.48 1.88    2    1    1    1     1     1     1     1     1     1     1
##   AGE Risk1Yr     Predict
## 1  60       1 Not Survive
## 2  51       1 Not Survive
## 3  59       1 Not Survive
## 4  54       1 Not Survive
## 5  73       2 Not Survive
## 6  51       1 Not Survive
```

```
##Model Performance Evaluation
riskmodel$aic
```

```
## [1] 387.5008
```

```
riskmodel_new$aic
```

```
## [1] 375.6534
```

```
## CONCLUSION : A model with minimum AIC value is preferred.The above shows the AIC of the original mode
```

```
##Confusion Matrix
mytable <- table(newdata2$Risk1Yr,newdata2$Predict)
mytable
```

```
##
##     Not Survive Survived
##   1         397        3
##   2          69        1
```

```
efficiency <- sum(diag(mytable))/sum(mytable)
efficiency
```

```
## [1] 0.8468085
```

```
## CONCLUSION: The accuracy of our model is 84.7%
```