7.1 Assignment

Scott Breitbach

7/13/2020

## Assignment 7.1: Fit a Logistic Regression Model to the Thoracic Surgery Binary Dataset

surgery\_df <- read.arff("completed/Week7/ThoraricSurgery.arff")

names(surgery\_df) <- c("Diagnosis", "FVC", "Vol\_Exhaled", "Perform\_Zb", "Pain\_B4\_Srg", "Haem\_B4\_Srg", "Dysp\_B4\_Srg", "Cough\_B4\_Srg", "Weak\_B4\_Srg", "Tumor\_Size", "Diabetes", "Heart\_Atk", "Pr\_Art\_Dis", "Smoking", "Asthma", "Age", "Died\_1Yr")

surgery\_df$Pain\_B4\_Srg <- relevel(surgery\_df$Pain\_B4\_Srg, "F")  
surgery\_df$Haem\_B4\_Srg <- relevel(surgery\_df$Haem\_B4\_Srg, "F")  
surgery\_df$Dysp\_B4\_Srg <- relevel(surgery\_df$Dysp\_B4\_Srg, "F")  
surgery\_df$Cough\_B4\_Srg <- relevel(surgery\_df$Cough\_B4\_Srg, "F")  
surgery\_df$Weak\_B4\_Srg <- relevel(surgery\_df$Weak\_B4\_Srg, "F")  
surgery\_df$Diabetes <- relevel(surgery\_df$Diabetes, "F")  
surgery\_df$Heart\_Atk <- relevel(surgery\_df$Heart\_Atk, "F")  
surgery\_df$Pr\_Art\_Dis <- relevel(surgery\_df$Pr\_Art\_Dis, "F")  
surgery\_df$Smoking <- relevel(surgery\_df$Smoking, "F")  
surgery\_df$Asthma <- relevel(surgery\_df$Asthma, "F")  
surgery\_df$Died\_1Yr <- relevel(surgery\_df$Died\_1Yr, "F")

### 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.

survival\_glm <- glm(Died\_1Yr ~ Diagnosis + FVC + Vol\_Exhaled + Perform\_Zb + Pain\_B4\_Srg + Haem\_B4\_Srg + Dysp\_B4\_Srg + Cough\_B4\_Srg + Weak\_B4\_Srg + Tumor\_Size + Diabetes + Heart\_Atk + Pr\_Art\_Dis + Smoking + Asthma + Age, data = surgery\_df, family = binomial())  
summary(survival\_glm)

##   
## Call:  
## glm(formula = Died\_1Yr ~ Diagnosis + FVC + Vol\_Exhaled + Perform\_Zb +   
## Pain\_B4\_Srg + Haem\_B4\_Srg + Dysp\_B4\_Srg + Cough\_B4\_Srg +   
## Weak\_B4\_Srg + Tumor\_Size + Diabetes + Heart\_Atk + Pr\_Art\_Dis +   
## Smoking + Asthma + Age, 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   
## DiagnosisDGN2 1.474e+01 2.400e+03 0.006 0.99510   
## DiagnosisDGN3 1.418e+01 2.400e+03 0.006 0.99528   
## DiagnosisDGN4 1.461e+01 2.400e+03 0.006 0.99514   
## DiagnosisDGN5 1.638e+01 2.400e+03 0.007 0.99455   
## DiagnosisDGN6 4.089e-01 2.673e+03 0.000 0.99988   
## DiagnosisDGN8 1.803e+01 2.400e+03 0.008 0.99400   
## FVC -2.272e-01 1.849e-01 -1.229 0.21909   
## Vol\_Exhaled -3.030e-02 1.786e-02 -1.697 0.08971 .   
## Perform\_ZbPRZ1 -4.427e-01 5.199e-01 -0.852 0.39448   
## Perform\_ZbPRZ2 -2.937e-01 7.907e-01 -0.371 0.71030   
## Pain\_B4\_SrgT 7.153e-01 5.556e-01 1.288 0.19788   
## Haem\_B4\_SrgT 1.743e-01 3.892e-01 0.448 0.65419   
## Dysp\_B4\_SrgT 1.368e+00 4.868e-01 2.811 0.00494 \*\*  
## Cough\_B4\_SrgT 5.770e-01 4.826e-01 1.196 0.23185   
## Weak\_B4\_SrgT 5.162e-01 3.965e-01 1.302 0.19295   
## Tumor\_SizeOC12 4.394e-01 3.301e-01 1.331 0.18318   
## Tumor\_SizeOC13 1.179e+00 6.165e-01 1.913 0.05580 .   
## Tumor\_SizeOC14 1.653e+00 6.094e-01 2.713 0.00668 \*\*  
## DiabetesT 9.266e-01 4.445e-01 2.085 0.03709 \*   
## Heart\_AtkT -1.466e+01 1.654e+03 -0.009 0.99293   
## Pr\_Art\_DisT -9.789e-02 1.003e+00 -0.098 0.92227   
## SmokingT 1.084e+00 4.990e-01 2.172 0.02984 \*   
## AsthmaT -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

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

Both Dyspnoea before surgery and a tumor size score of OC14 (the largest size in this study) were the highest predictors of death, with z-values of 2.81 and 2.71, respectively. Both were significant with p values < 0.01.

The next best predictors of death were whether the individual was a smoker and whether they had diabetes, with z-values of 2.17 and 2.09 (both at significance < 0.05).

If we want to look strictly at variables associated with surviving, then “Volume that has been exhaled at the end of the first second of forced expiration” had the greatest negative correlation with the death outcome at a z-value of -1.70 (significance of 0.1)

# Odds Ratio  
exp(survival\_glm$coefficients)

## (Intercept) DiagnosisDGN2 DiagnosisDGN3 DiagnosisDGN4 DiagnosisDGN5   
## 6.481697e-08 2.511211e+06 1.440574e+06 2.209615e+06 1.301120e+07   
## DiagnosisDGN6 DiagnosisDGN8 FVC Vol\_Exhaled Perform\_ZbPRZ1   
## 1.505091e+00 6.785355e+07 7.967257e-01 9.701510e-01 6.422903e-01   
## Perform\_ZbPRZ2 Pain\_B4\_SrgT Haem\_B4\_SrgT Dysp\_B4\_SrgT Cough\_B4\_SrgT   
## 7.454996e-01 2.044884e+00 1.190456e+00 3.928338e+00 1.780613e+00   
## Weak\_B4\_SrgT Tumor\_SizeOC12 Tumor\_SizeOC13 Tumor\_SizeOC14 DiabetesT   
## 1.675616e+00 1.551720e+00 3.251796e+00 5.222483e+00 2.525890e+00   
## Heart\_AtkT Pr\_Art\_DisT SmokingT AsthmaT Age   
## 4.317676e-07 9.067446e-01 2.956473e+00 8.455364e-07 9.905394e-01

# Confidence Interval  
exp(confint(survival\_glm))

## 2.5 % 97.5 %  
## (Intercept) NA 1.861968e+203  
## DiagnosisDGN2 1.717218e-206 NA  
## DiagnosisDGN3 8.098202e-207 NA  
## DiagnosisDGN4 1.675824e-206 NA  
## DiagnosisDGN5 1.264378e-205 NA  
## DiagnosisDGN6 2.066233e-24 2.939776e+20  
## DiagnosisDGN8 5.691816e-171 NA  
## FVC 5.499148e-01 1.138007e+00  
## Vol\_Exhaled 9.264310e-01 9.993543e-01  
## Perform\_ZbPRZ1 2.300552e-01 1.783025e+00  
## Perform\_ZbPRZ2 1.540289e-01 3.470770e+00  
## Pain\_B4\_SrgT 6.558696e-01 5.928649e+00  
## Haem\_B4\_SrgT 5.383928e-01 2.497049e+00  
## Dysp\_B4\_SrgT 1.466379e+00 1.007288e+01  
## Cough\_B4\_SrgT 7.094170e-01 4.740878e+00  
## Weak\_B4\_SrgT 7.532542e-01 3.596887e+00  
## Tumor\_SizeOC12 8.231331e-01 3.022655e+00  
## Tumor\_SizeOC13 9.225453e-01 1.064690e+01  
## Tumor\_SizeOC14 1.540476e+00 1.723680e+01  
## DiabetesT 1.017658e+00 5.900292e+00  
## Heart\_AtkT NA 1.949037e+106  
## Pr\_Art\_DisT 9.525986e-02 5.459928e+00  
## SmokingT 1.197920e+00 8.705307e+00  
## AsthmaT NA 8.570374e+105  
## Age 9.561182e-01 1.026545e+00

Dsyp\_B4\_Srg has an odds ratio of 3.928 indicating that an individual with Dyspnoea has 3.9 times greater risk of dying within the year than someone without. Additionally the odds ratios for Tumor\_SizeOC14, Diabetes, and Smoking were 5.22, 2.53, and 2.95 respectively. On the flip side, Vol\_Exhaled had an odds ratio of only 0.097. None of their confidence intervals cross zero.

### 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?

# "Run the data through the model"  
res <- predict(survival\_glm, surgery\_df, type = "response")  
# "Validate the model - confusion matrix"  
confmatrix <- table(Actual\_Value=surgery\_df$Died\_1Yr, Predicted\_Value = res > 0.5)  
confmatrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 390 10  
## T 67 3

"Accuracy"

## [1] "Accuracy"

(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.8361702

The model appears to be 83.6% accurate.