Assignment 7.1

Anna Harvey

July 19, 2020

## 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.

logregmod <- glm(Risk1Yr ~ PRE30 + PRE17 + PRE14 + PRE9, data = surgery, family = binomial)  
  
summary(logregmod)

##   
## Call:  
## glm(formula = Risk1Yr ~ PRE30 + PRE17 + PRE14 + PRE9, family = binomial,   
## data = surgery)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4318 -0.5496 -0.4601 -0.3614 2.4980   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.0748 0.4732 -6.498 8.14e-11 \*\*\*  
## PRE30T 0.8821 0.4362 2.022 0.04316 \*   
## PRE17T 1.0239 0.4174 2.453 0.01418 \*   
## PRE14OC12 0.3790 0.3090 1.226 0.22004   
## PRE14OC13 1.2999 0.5735 2.267 0.02341 \*   
## PRE14OC14 1.7493 0.5625 3.110 0.00187 \*\*   
## PRE9T 1.0384 0.4434 2.342 0.01919 \*   
## ---  
## 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: 368.78 on 463 degrees of freedom  
## AIC: 382.78  
##   
## Number of Fisher Scoring iterations: 5

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

The variable PRE14OC14 which indicates the size of the original tumor as being the largest size is the most significant of the included variables. The other significant variables (in order from most to least significant) are: Type 2 Diabetes, Dyspnoea before surgery, Size 13 tumors (2nd largest), and Smoking.

## 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?

#Split the data into training and testing sets  
split <- sample.split(surgery, SplitRatio = 0.8)  
train <- subset(surgery, split == TRUE)  
test <- subset(surgery, split == FALSE)  
  
#Create the GLM with the training set  
mod <- glm(Risk1Yr ~ PRE30 + PRE17 + PRE14 + PRE9, data = train, family = binomial)  
  
#Run the data through the model  
res <- predict(mod, test, type = "response")  
res <- predict(mod, train, type = "response")  
  
#Validate model with confusion matrix  
confmatrix <- table(Actual\_value = train$Risk1Yr, Predicted\_value = res > 0.5)  
confmatrix

## Predicted\_value  
## Actual\_value FALSE TRUE  
## F 308 2  
## T 50 0

#Calculate accuracy  
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.8555556

The accuracy of the model is 83.7%