Modeling Risk Associated with Colectomy Procedures for Anastomotic Leaking

Jacob Kramp

8/8/2021

Introduction

We have been asked by a hospital that has conducted several colectomy procedures to analyze several covariates collected about each patient and the covariate's relationship to the risk associated to anastomotic leaking. We have been given a datset containing 179 cases from which we will attempt to model the risk using a generalized linear model. We will also assess the assumptions made while using the model and whther or not there may be a better fitting model available.

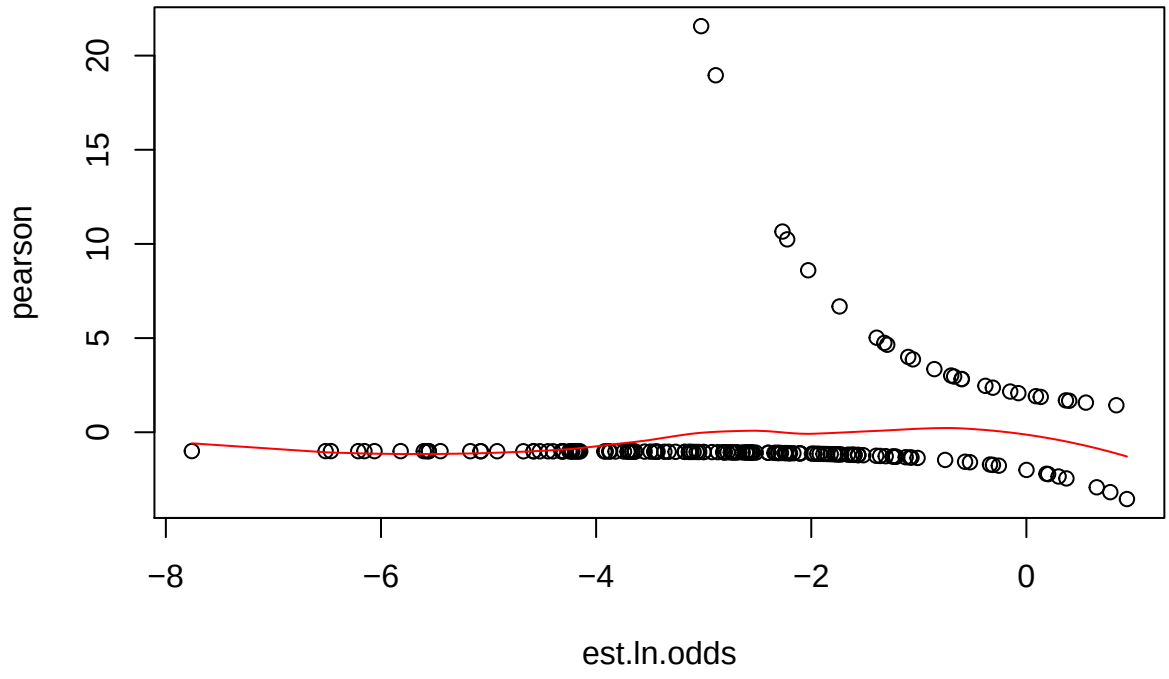
Analyzing the Model with Respect to BMI of Patient

In order to reduce noise in our model created by very small changes in our covariate, BMI, we will discretize the BMI by rounding the BMI either up or down. We will continue to treat BMI as a numeric predictor.Using the glm() function in R, The results of our model are printed below:

```
##
## Call:
## glm(formula = Anastamotic.Leak ~ ., family = "binomial", data = dff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5913  -0.5080  -0.3020  -0.1448   2.4783
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -19.23298   18.51826  -1.039 0.298993
## GenderMale         0.55285    0.71975   0.768 0.442422
## Height..in.        0.18157    0.26803   0.677 0.498128
## Weight..lbs.      -0.02013    0.04124  -0.488 0.625393
## BMI                0.22677    0.27366   0.829 0.407284
## Age                0.08175    0.02668   3.064 0.002181 **
## RaceW             -0.24339    0.52123  -0.467 0.640537
## Tobacco            1.09906    0.58740   1.871 0.061338 .
## DM                -0.78865    0.64587  -1.221 0.222064
```

```
## CAD.PAD           -0.66341    0.73598  -0.901 0.367374
## Cancer             0.72275    0.61989   1.166 0.243644
## Albumin..g.dL.    -1.31428    0.36967  -3.555 0.000378 ***
## Operative.Length   7.04466    3.60607   1.954 0.050754 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 110.95  on 166  degrees of freedom
## AIC: 136.95
##
## Number of Fisher Scoring iterations: 6
```

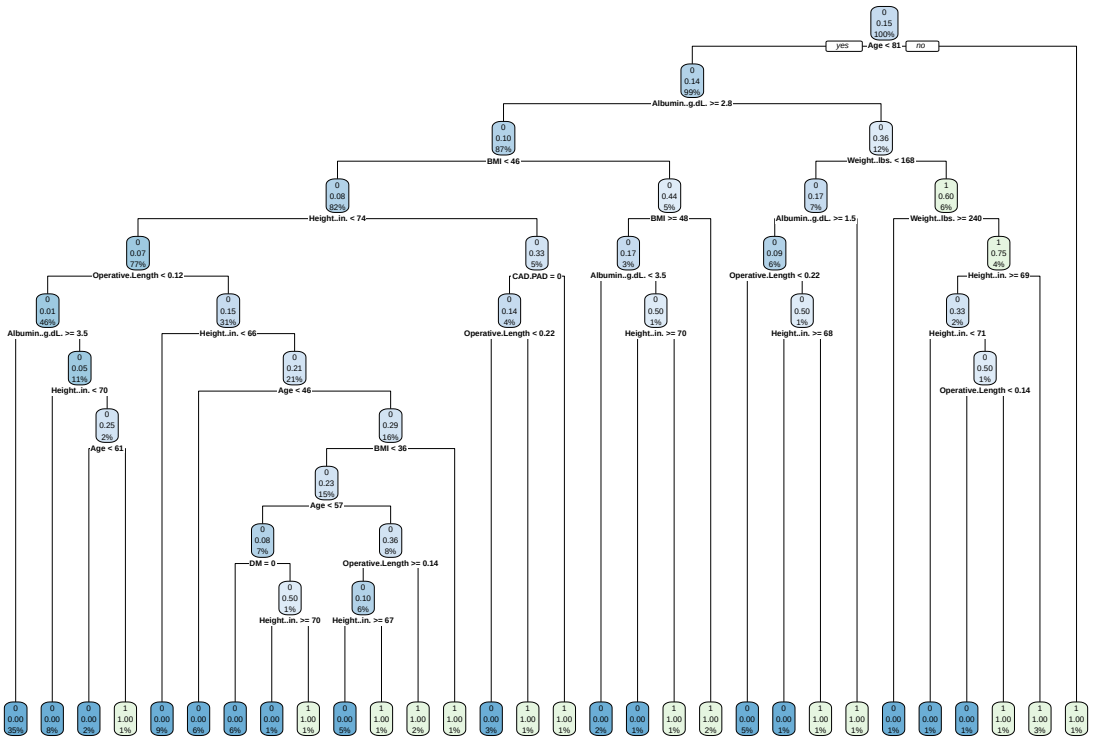
According to our model,a one unit increase in BMI will lead to an estimated +1.25 modification to the odds of having an anastomic leak.In other words, a 1 unit increase of BMI, is an estimated 25.45% change in odds of an anastomic leak. Now, before we continue, we should assess the validity of our model to be sure that our results are trustworthy.



The red line shown in the graph would ideally be shaped as a contant line at Pearson = 0. The behavior of the red line indicates that this model may not fit our data very well when our covariate values get much larger. There may be a more competitive model available.

Would a Tree Model Make a More Accurate Prediction?

The use of a tree model (the only other classification model covered thus far) should be evaluated for a possibly better performance in this situation. The results of the tree model constructed can be visualized below.



To assess which model is performing best, we can calculate the log likelihood of each model using a cross-validated set of p-value estimates. The cross-validated method that we'll use here is the leave one out method. We'll show the first five rows of our computed p-values and cross-validated p-values to show some of the process.

y	p	cv.p
0	0.0148346	0.0151855
0	0.0020025	0.0020099
0	0.1198983	0.1294721
0	0.0014790	0.0014825
0	0.0196423	0.0199162
0	0.4363138	0.4964412

Log-likelihood of GLM	CV Log-likelihood of GLM
-55.47313	-73.33

y	p.0	p.1	cv.p
0	0.99999	1e-05	0.1509434
0	0.99999	1e-05	0.1142857
0	0.99999	1e-05	0.1635220
0	0.99999	1e-05	0.1572327
0	0.99999	1e-05	0.1509434
0	0.99999	1e-05	0.1572327

Log-likelihood of Tree	CV Log-likelihood of Tree
-0.00179	-78.94713

The tree model is obviously overfitting the data according to the log likelihood before cross validating. A log likelihood of -0.00179 compared to the cross validated log likelihood of -78.9471271 is evidence of that. Still, the GLM model seems to perform better according to the cross validated log likelihood. We will continue to use that model as a result.

Variable Selection

Now we'll consider every relevant covariate given to us in our dataset.

```
## The following objects are masked from dff (pos = 3):
##
##     Age, Albumin..g.dL., Anastamotic.Leak, BMI, CAD.PAD, Cancer, DM,
##     Gender, Height..in., Operative.Length, Race, Tobacco, Weight..lbs.
```

```
##
## Call:
## glm(formula = Anastamotic.Leak ~ ., family = "binomial", data = dff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5905  -0.5082  -0.3035  -0.1452   2.4774
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -19.10961   18.62296  -1.026 0.304829
## GenderMale         0.54382    0.73178   0.743 0.457388
## Height..in.        0.17989    0.26936   0.668 0.504217
## Weight..lbs.      -0.01981    0.04154  -0.477 0.633456
## BMI                0.22505    0.27498   0.818 0.413116
## Age                0.08160    0.02675   3.050 0.002288 **
## RaceW             -0.23284    0.54400  -0.428 0.668641
## Racewhite         -0.30198    1.01597  -0.297 0.766286
## Tobacco            1.09960    0.58760   1.871 0.061299 .
## DM                -0.78409    0.64932  -1.208 0.227220
## CAD.PAD           -0.66189    0.73640  -0.899 0.368750
```

```
## Cancer             0.72680    0.62281   1.167 0.243220
## Albumin..g.dL.    -1.31788    0.37375  -3.526 0.000422 ***
## Operative.Length   7.03521    3.61206   1.948 0.051451 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 110.94  on 165  degrees of freedom
## AIC: 138.94
##
## Number of Fisher Scoring iterations: 6
```

Interestingly enough, in the presence of the other covariates, BMI is not statistically significant. This does not directly mean that BMI is not an important predictor. In fact, we should use an objective measure of fit to identify which predictors are important for the performance of the model and which of them are not. We will use AIC to determine which predictors to keep and which to toss out of our dataset. The step() function in R will remove a predictor one at a time and evaluate the new model's AIC to determine if the model will fit better after removing that variable. It will do this process repeatedly until the AIC is best when not removing any covariates. The results are shown below.

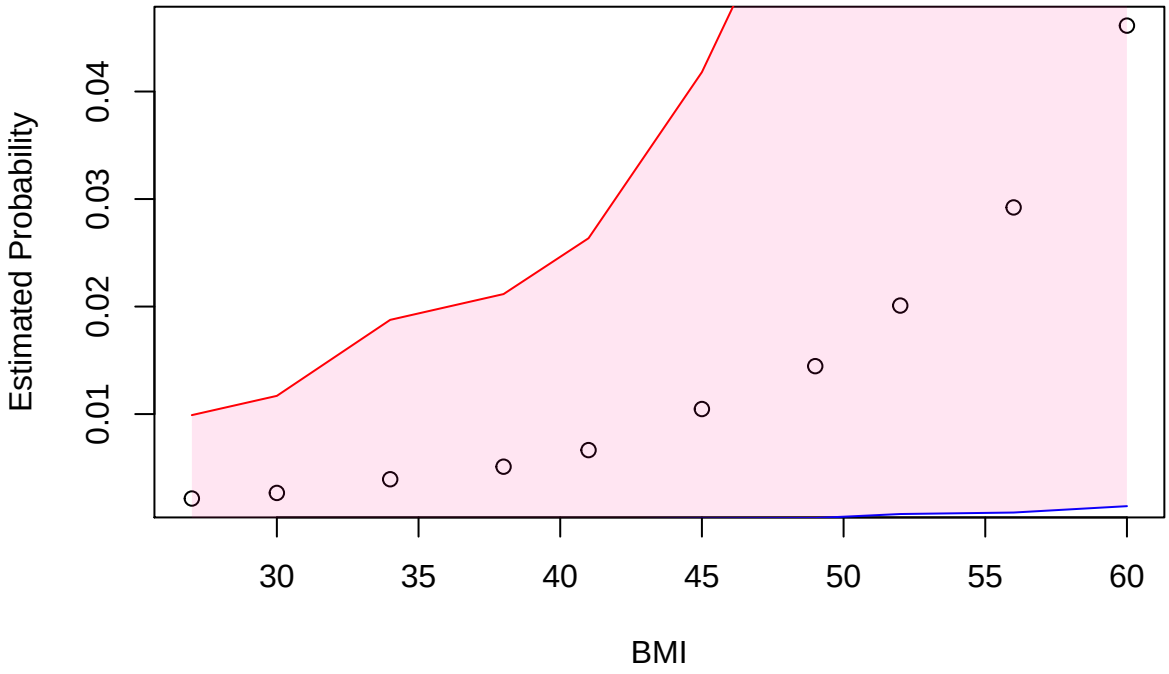
```
##
## Call:
## glm(formula = Anastamotic.Leak ~ Gender + BMI + Age + Tobacco +
##     DM + Albumin..g.dL. + Operative.Length, family = "binomial",
##     data = dff)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7600  -0.5141  -0.3220  -0.1457   2.4042
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -6.30447    1.95529  -3.224 0.001263 **
## GenderMale        0.77174    0.49998   1.544 0.122699
## BMI               0.08909    0.03041   2.930 0.003389 **
## Age               0.08276    0.02350   3.522 0.000428 ***
## Tobacco           0.78186    0.53601   1.459 0.144655
## DM               -0.86507    0.63472  -1.363 0.172911
## Albumin..g.dL.   -1.38995    0.35815  -3.881 0.000104 ***
## Operative.Length  7.71134    3.46733   2.224 0.026148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 148.35  on 178  degrees of freedom
## Residual deviance: 113.93  on 171  degrees of freedom
## AIC: 129.93
##
## Number of Fisher Scoring iterations: 6
```

In conclusion, the most important predictors that were given to us appear to be Gender, BMI, Age, whether

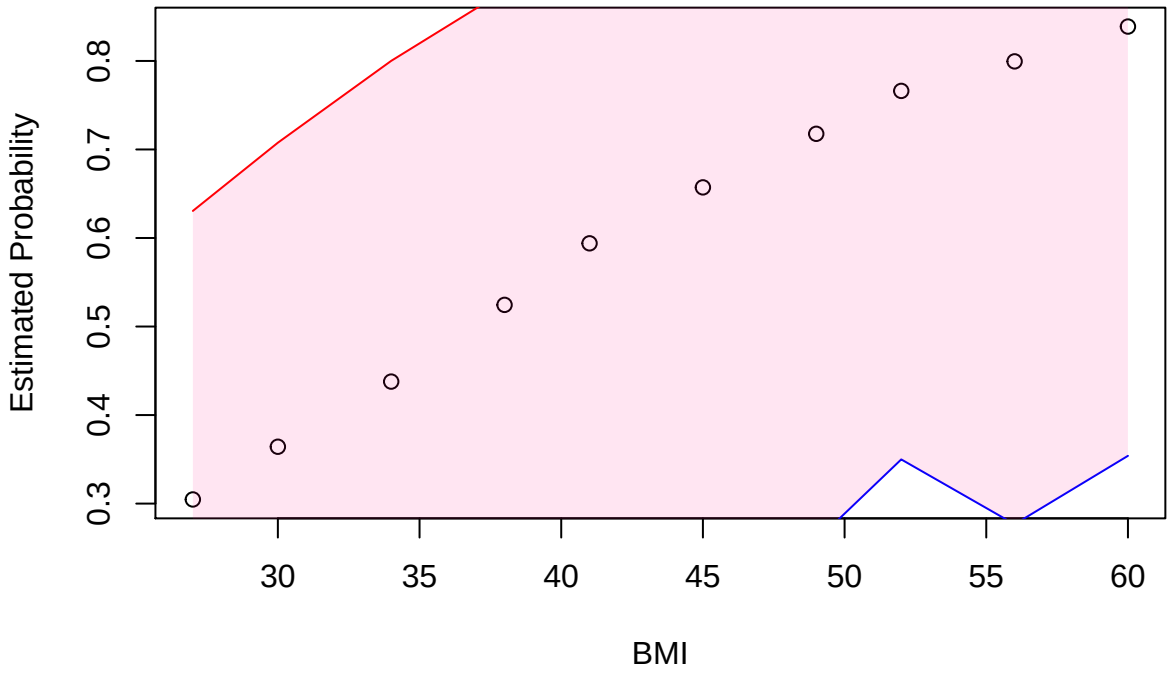
the subject uses Tobacco or not, whether the subject has Diabetes or not, the Albumin levels, and the length of operation.

Case Studies and a Visual of Risk associated with Obesity

Our first visual shows the increasing probability of a leak for Arizona Robbins; a 35 year old white female who doesn't use tobacco, doesn't have diabetes, doesn't have CAD or PAD, doesn't have cancer, has a post-operative albumin level of 4.2 and whose operation length took 90 minutes.



The second is for Richard Webber, a 62 African American male who uses tobacco and has diabetes and whom had an albumin level of 2.8 following a 210 minute operation.



Although the male subject is more likely to have a leak after surgery, both visualizations show the increase rik caused by obesity are large. This is especially true if you have other underlying factors that significantly increase yor risk.

Index (Code)

```
library(tidyverse)
library(data.table)
library(rpart)
library(caret)
library(knitr)
library(car)
df <- data.frame(fread('colon2017.csv'))
set.seed(2021)

# Multivariate Case
dff <- df[,c("Gender","Height..in.","Weight..lbs.",'BMI','Age','Race',"Tobacco",'DM',"CAD.PAD",
             'Cancer',"Albumin..g.dL." ,"Operative.Length","Anastamotic.Leak")]
#Adjust spelling of white
dff$Race[grep('W',dff$Race)] <- 'W'
dff$Race[grep('w',dff$Race)] <- 'W'
```

```r
attach(dff)
# GLM Model
model.glm <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = dff)
summary(model.glm)

y <- dff$Anastamotic.Leak
pearson = (y - model.glm$fit)/(model.glm$fit*(1-model.glm$fit))
est.ln.odds = log(model.glm$fit/(1-model.glm$fit))
loess1 = loess(pearson~est.ln.odds)
plot(est.ln.odds,pearson)
lines(est.ln.odds[order(est.ln.odds)],loess1$fit[order(est.ln.odds)],col = 'red')

#Tree
model.tree <- rpart(Anastamotic.Leak~.,data = dff,method = 'class',control=rpart.control(minsplit=2, cp=
rpart.plot::rpart.plot(model.tree)

#Cross validate by computing log likelihood of glm and tree model
results = data.frame(y = df$Anastamotic.Leak,p = predict.glm(model.glm,newdata=dff,type="response"),cv.p
for(i in 1:nrow(dff)){
  #Leave one out
  traind <- dff[-i,]
  testd <- dff[i,]
  glm.t <- glm(Anastamotic.Leak~.,data = traind,family = 'binomial')
  results[i,'cv.p'] <- predict.glm(glm.t,newdata=testd,type="response")
}
LL.glm = sum(results$y*log(results$p) + (1-results$y)*log(1-results$p))
CV.LL.glm = sum(results$y*log(results$cv.p) + (1-results$y)*log(1-results$cv.p))

knitr::kable(head(results))
lldf <- data.frame(ll.glm = LL.glm,cv.ll.glm = CV.LL.glm)
colnames(lldf) <- c('Log-likelihood of GLM','CV Log-likelihood of GLM')
knitr::kable(lldf)

#Same thing for tree model
results = data.frame(y = df$Anastamotic.Leak,p = predict(model.tree,newdata=dff,type="prob"),cv.p=0)
for(i in 1:nrow(dff)){
  #leave 10% out
  t <- sample(1:nrow(dff),20)
  traind <- dff[-t,]
  testd <- dff[t,]
  tree.t <- rpart(Anastamotic.Leak~.,data = traind,method = 'class',control=rpart.control(minsplit=2, cp
  pred <- predict(tree.t,newdata=testd,type="prob")
  results[t,'cv.p'] <- pred[,'1']
}
#Make small adjustment to be able to calculate log likelihood
results[,-1][results[,-1]==0] <- .00001
results[,-1][results[,-1]==1] <- .99999

LL.tree = sum(results$y*log(results$p.1) + (1-results$y)*log(1-results$p.1))
CV.LL.tree = sum(results$y*log(results$cv.p) + (1-results$y)*log(1-results$cv.p))

knitr::kable(head(results))
lldf <- data.frame(ll.tree = LL.tree,cv.ll.tree = CV.LL.tree)
```

```r
colnames(lldf) <- c('Log-likelihood of Tree','CV Log-likelihood of Tree')
knitr::kable(lldf)

# Multivariate Case
dff <- df[,c("Gender","Height..in.","Weight..lbs.",'BMI','Age','Race',"Tobacco",'DM',"CAD.PAD",
             'Cancer',"Albumin..g.dL." ,"Operative.Length","Anastamotic.Leak")]
#Adjust spelling of white
dff$Race[grep('W',dff$Race)] <- 'W'
attach(dff)
# GLM Model
model.glm <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = dff)
summary(model.glm)

s <- step(model.glm)
summary(s)

#Bootstrap  predictions and their confidence intervals
dff <- df[,c("Gender",'BMI','Age',"Tobacco",'DM',"Albumin..g.dL." ,"Operative.Length","Anastamotic.Leak

AR <- data.frame(
  Gender = as.character(rep('Female',10)),
  BMI = as.integer(seq(27,60,length.out = 10)),
  Tobacco = as.integer(rep(0,10)),
  DM = as.integer(rep(0,10)),
  Albumin..g.dL. = as.numeric(rep(4.2,10)),
  Operative.Length = as.numeric(rep(90 *0.000694444,10)),
  Age = as.integer(rep(35,10))
)

#Train model with original data
model.glm <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = dff)
predictions = data.frame(bmi = numeric(),pstar = numeric(),lower = numeric(),upper = numeric())
for(w in 1:nrow(AR)){
  phat.star = data.frame(bmi = numeric(),phat = numeric())
  for(i in 1:1000){
    #Get size of possible sample
    n <- nrow(dff)
    #Take a sample of size in with independence
    BS.x = dff[sample(1:n,n,replace = T),]
    #Plug BS.x into model for BS.phat
    log.lik <- predict.glm(model.glm,BS.x)
    phat <- 1/(1+exp(-log.lik))
    #Bootstrap y values
    BS.y <- sapply(1:length(phat),function(p){
      sample(c(1,0),1,prob = c(phat[p],(1-phat[p])))
    })
    new.data = data.frame(Anastamotic.Leak=BS.y,BS.x[,-grep('Anastamotic.Leak',colnames(BS.x))])
    #Fit new model
    new.model <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = new.data)
    #Make the desired prediction
    phat <- predict.glm(new.model,AR[w,],type='response')
    phat.star <- rbind(phat.star,data.frame(bmi = AR[w,'BMI'],phat=phat))
  }
```

```
  lower = quantile(phat.star[,'phat'],.025)
  upper = quantile(phat.star[,'phat'],.975)
  pstar = mean(phat.star[,'phat'])
  bmi = unique(phat.star[,'bmi'])
  predictions = rbind(predictions,data.frame(bmi=bmi,pstar=pstar,lower = lower,upper=upper))
}
plot(predictions$bmi,predictions$pstar,ylab = 'Estimated Probability',xlab = 'BMI')
lines(x = predictions$bmi,y = predictions$lower,col = 'blue')
lines(x = predictions$bmi,y = predictions$upper,col = 'red')
polygon(c(predictions$bmi, rev(predictions$bmi)), c(predictions$upper, rev(predictions$lower)),
        border = NA,col=rgb(1, 0, .5,0.1))


#Bootstrap  predictions and their confidence intervals

AR <- data.frame(
  Gender = as.character(rep('Male',10)),
  BMI = as.integer(seq(27,60,length.out = 10)),
  Tobacco = as.integer(rep(1,10)),
  DM = as.integer(rep(1,10)),
  Albumin..g.dL. = as.numeric(rep(2.8,10)),
  Operative.Length = as.numeric(rep(210 *0.000694444,10)),
  Age = as.integer(rep(62,10))
)

#Train model with original data
model.glm <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = dff)
predictions = data.frame(bmi = numeric(),pstar = numeric(),lower = numeric(),upper = numeric())
for(w in 1:nrow(AR)){
  phat.star = data.frame(bmi = numeric(),phat = numeric())
  for(i in 1:1000){
    #Get size of possible sample
    n <- nrow(dff)
    #Take a sample of size in with independence
    BS.x = dff[sample(1:n,n,replace = T),]
    #Plug BS.x into model for BS.phat
    log.lik <- predict.glm(model.glm,BS.x)
    phat <- 1/(1+exp(-log.lik))
    #Bootstrap y values
    BS.y <- sapply(1:length(phat),function(p){
      sample(c(1,0),1,prob = c(phat[p],(1-phat[p])))
    })
    new.data = data.frame(Anastamotic.Leak=BS.y,BS.x[,-grep('Anastamotic.Leak',colnames(BS.x))])
    #Fit new model
    new.model <- glm(Anastamotic.Leak ~ .,family = 'binomial',data = new.data)
    #Make the desired prediction
    phat <- predict.glm(new.model,AR[w,],type='response')
    phat.star <- rbind(phat.star,data.frame(bmi = AR[w,'BMI'],phat=phat))
  }
  lower = quantile(phat.star[,'phat'],.025)
  upper = quantile(phat.star[,'phat'],.975)
  pstar = mean(phat.star[,'phat'])
  bmi = unique(phat.star[,'bmi'])
```

```
  predictions = rbind(predictions,data.frame(bmi=bmi,pstar=pstar,lower = lower,upper=upper))
}
plot(predictions$bmi,predictions$pstar,ylab = 'Estimated Probability',xlab = 'BMI')
lines(x = predictions$bmi,y = predictions$lower,col = 'blue')
lines(x = predictions$bmi,y = predictions$upper,col = 'red')
polygon(c(predictions$bmi, rev(predictions$bmi)), c(predictions$upper, rev(predictions$lower)),
        border = NA,col=rgb(1, 0, .5,0.1))
```