University of California, Davis

Take Home Project 2:
Problem 2: Building a Model where Y = Angina status.

Ana Boeriu & Victoria Gribben
STA 138
Dr. Erin Melcon
March 1, 2019

I. Introduction

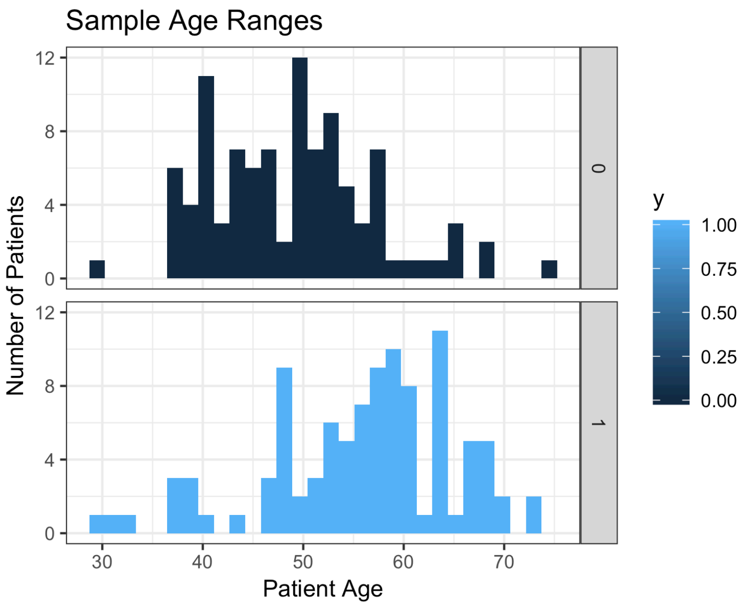
The goal of this project is to build the "best" correct model that can help medical doctors accurately predict angina in the general population, using only the most significant variables. It is important that the model works well with as few predictive variables as possible because angina is a type of chest pain resulting from reduced blood flow to the heart muscle, which is an emergency. During medical emergencies, doctors' ability to gather patient information and access to patient medical history is limited.

We will then use the best model to calculate confidence intervals and predictions that give us insight as to which variable has a significant effect on angina. These results may be useful to anyone in the general public specifically medical doctors and researchers who can provide more information to patients about managing this condition. The approach we are taking is using logistic regression because our response variable and most of our predictor variables are categorical. We are regressing the categorical response variable (Y) angina against the explanatory variables (X) of age, smoking status, cigarettes, hypertension, myocardial infarction, stroke, and diabetes in order to choose our best model.

II. Data Summary

There are seven predictor variables in the data set: age, smoking status, number of cigarettes smoked per day, family history of hypertension, family history of heart attack, family history of stroke, and family history of diabetes.

A. Age:



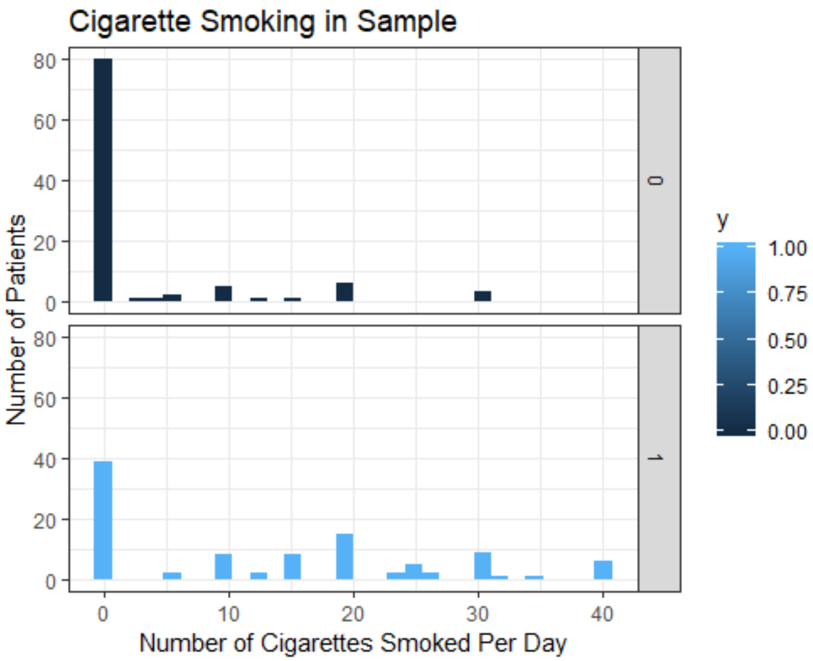
Very few patients younger than 35 years old experienced angina. The majority of patients who experienced angina were at least 50 years old.

B. Smoking Status:

Smoking Status	No Angina	Angina
Current	21	61
Ex	15	26
Never	64	13
Total	100	100

Smoking status appears to have an effect on the likelihood on a patient reporting chest pain: 73.5% of current smokers sampled reported angina, while only 16.9% of patients who had never smoked reported angina.

C. Number of Cigarettes per day



Most patients in the sample did not smoke. Smokers tended to smoke in half-pack increments, with the most common amount being 20 (1 pack per day).

D. Family History: Hypertension

Family History: Hypertension	No Angina	Angina

Absent	83	67
Mild	14	23
Moderate	3	10

The majority of subjects did not have a family history of hypertension. Patients with a family history of hypertension were more likely to also report chest pain.

E. Family History: Heart Attack (Myocardial Infarction)

Family History: Heart Attack	No Angina	Angina
No	88	47
Yes	12	53

The majority of subjects did not have a family history of heart attack. Patients with a family history of heart attack were more likely to also report chest pain.

F. Family History: Stroke

Family History: Stroke	No Angina	Angina
No	94	94
Yes	6	6

Since the patient distribution is the same whether or not the patient had a family history of stroke, we are confident this predictor is not useful.

G. Family History: Diabetes

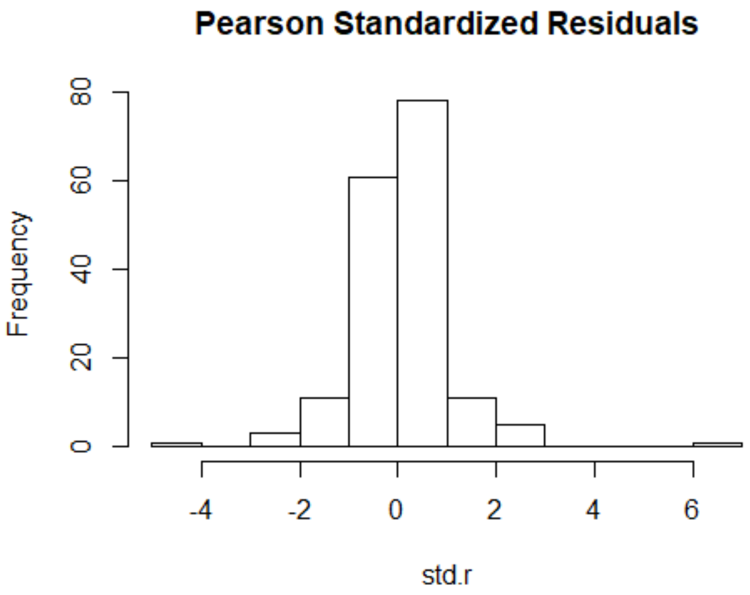
Family History: Diabetes	No Angina	Angina
No	97	94
Yes	3	6

The patient distributions for this predictor variable are very similar and the difference may not be statistically significant.

III. Data Preparation

In this section we will graph the data with all of its predictors and look for any anomalies in the data. We will discuss any potential outliers and whether we will remove the observation or not.

A. Pearson Standardized Residuals

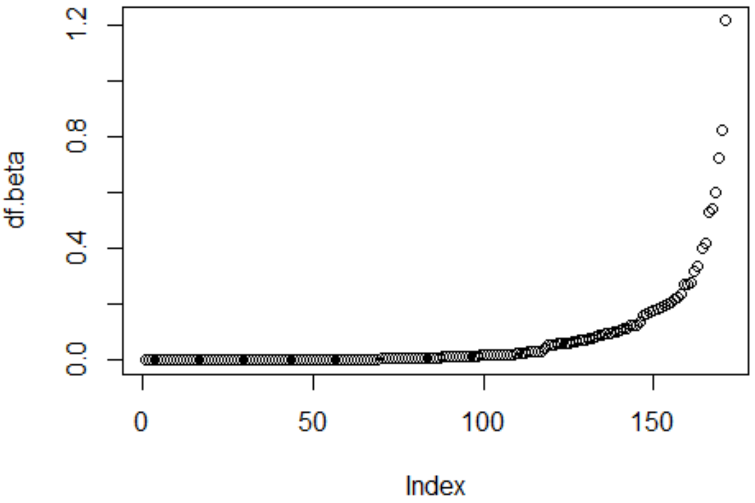


The standardized residual plot shows outliers in the data: there are values of $|r_{ij}| > 3$.

Standardized Residual Outlier Points								
Row	Age	Smoke	Cig	Hyper	Myofam	Strokefam	Diabetes	Angina
158	46	Ex	0	Absent	Yes	No	No	No
166	64	Ex	0	Absent	Yes	No	No	Yes

While these points are outliers, they are not obvious errors, so we compare them to the influential points.

B. DF Beta



Based on the plot of df.beta values, a reasonable cutoff appears to be around 0.3.
With this cutoff, the only overlap between outliers and influential points is patient 166. Since this outlier has proven to be influential, and since this observation is 0.5% of the data, we will remove it before proceeding.

IV. Model Selection and Analysis

In this section, we are using stepwise regression for all subsets and a specific subset selection process to see which combination of explanatory variables will produce the best model based off of the criteria we have chosen to use which is BIC. There are many variables in our model and we must make sure the ones in our model are significant. We looked at the model chosen when using all subsets.

A. Model Criteria

Because our goal is to have a correct model, we are choosing only explanatory variables which have a significant relationship with angina. This model may be smaller than ones which have the goal of prediction. For the purposes of this class AIC or BIC are often used as model selection criteria as they penalize large models. AIC may overfit correct models and BIC penalizes large models even more, so we chose to focus on BIC as we imagine if somebody is trying to see if a person has angina they will really need the result to be as correct as possible. We choose the model that lowers BIC the most.

B. All Subsets

When there are a fewer number of predictor variables all possible models and the corresponding model criteria can be calculated. This was the model with the lowest BIC when looking at all possible models.

Best Model by BIC
$Y \sim X_{age} + X_{smoke} + X_{cig} + X_{hyper} + X_{myofam}$

Best Model Fit by BIC
$Y \sim -7.2446 + 0.1086X_{age} + 0.6486X_{smoke,ex} - 1.3332X_{smoke,never} + 0.1025X_{cig} + 1.3121X_{hyper,mild} + 2.1151X_{hyper,moderate} + 2.3796X_{myofam,yes}$

With our new model, we again evaluate the data's influential points and outliers before proceeding. We find two outliers, patients 154 and 161, whose absolute standardized residual values are larger than three, and six influential points, patients 158-163, whose dfbeta values are larger than 0.3. We once again remove the point that is both an influential point and an outlier, resulting in a total loss of 1% of our original data. Without this outlier, our model fit is slightly different.

Best No-Interactions Model Fit (Outlier Removed)
$Y \sim -7.2067 + 0.1081X_{age} + 0.6357X_{smoke,ex} - 1.3440X_{smoke,never} + 0.1013X_{cig} + 1.3121X_{hyper,mild} + 2.116X_{hyper,moderate} + 2.3791X_{myofam,yes}$

C. Interaction Terms

From the above model we know which X's are significant. Thus, the next step is to fit the models with interaction terms and compare their model selection criteria to our chosen no-interactions model.

Models with Interaction Terms:		
Interaction Term	AIC	BIC
None	161.1164	187.4225*
$X_{age} * X_{smoke}$	163.6735	196.5562
$X_{age} * X_{cig}$	163.0854	192.6798
$X_{age} * X_{hyper}$	163.2556	196.1383
$X_{age} * X_{myofam}$	158.2063*	187.8007
$X_{smoke} * X_{cig}$	161.1164	187.4225
$X_{smoke} * X_{hyper}$	159.6817*	199.1409
$X_{smoke} * X_{myofam}$	163.0246	195.9073
$X_{cig} * X_{hyper}$	164.4534	197.3361

$X_{cig} * X_{myofam}$	160.0889	189.6843
$X_{hyper} * X_{myofam}$	162.581	195.4637

The best predictor model may include interaction terms, but according to our BIC model selection term, the “most correct” model does not benefit from interaction terms. Since adding one interaction term is not helpful, we will not test adding two or more interaction terms.

D. Final Model

Final Estimated Logistic Regression Function
$Y \sim - 7.2067 + 0.1081X_{age} + 0.6357X_{smoke,ex} - 1.3440X_{smoke,never} + 0.1013X_{cig} + 1.3121X_{hyper,mild} + 2.116X_{hyper,moderate} + 2.3791X_{myofam,yes}$

Our final model uses age, smoking status, average number of cigarettes smoked per day, and family histories of hypertension and heart attack (myofam).

Final Estimated Odds Function
$ODDS_{ANGINA} = exp(-7.2067 + 0.1081X_{age} + 0.6357X_{smoke,ex} - 1.3440X_{smoke,never} + 0.1013X_{cig} + 1.3121X_{hyper,mild} + 2.116X_{hyper,moderate} + 2.3791X_{myofam,yes})$

Age, being an ex-smoker, cigarette use, family history of hypertension, and especially family history of heart attack increase the odds that a patient will have angina. The patient having never smoked reduces the odds.

E. Corrected Confidence Intervals (g=8)

95% Corrected Confidence Intervals for Exponentiated β_i's			
	Lower Bound	Exp(β) Value	Upper Bound
Intercept	0.0000	0.022	0.0398
Age	1.043	1.1142	1.2033
Smoke - Ex	0.1970	1.8883	20.7170
Smoke - Never	0.0258	0.2608	2.7422
Cig	1.0037	1.1066	1.2512
Hyper - Mild	0.8626	3.7140	17.8247

Hyper - Moderate	0.8217	8.2979	125.4575
Myofam - Yes	2.9859	10.7952	47.8070

V. Interpretation

A. Model

Age increases the odds of angina by 1.1142 times with every increase in year, all other predictors held constant. Being an ex-smoker increases the odds by 1.8883. Having never smoked reduces the odds by a factor of 0.2608. Each additional cigarette smoked per day, on average, increases the odds of angina by 1.1066, other predictors held constant. A family history of mild hypertension increases the odds of angina by 3.7140, and a family history of moderate hypertension increases those odds by 8.2979. Lastly, a family history of heart attack (myocardial infarction) increases the odds of angina by 10.7952 times.

B. Confidence Intervals

1. **Intercept**
   Since there is a measurement in the final model (ie age cannot be zero) it is not meaningful to interpret the intercept.
2. **Age - exp(β_1)**
   We are overall 95% certain that when age increases by one year the true odds of having angina is multiplied by between 1.043 and 1.2033 times the previous odds, all other predictive values held constant.
3. **Cig**
   We are overall 95% certain that when the average number of cigarettes smoked per day increases by 1 cigarette, the true odds that a patient has angina are multiplied by between 1.0037 and 1.2512.
4. **Myofam - Yes**
   We are 95% confident that the true odds of having angina for a patient who has a family history of myocardial infarction is between 2.9859 and 47.8070 times that of a patient who does not have a history of myocardial infarction.

The model also selected smoking status and family history of hypertension as being important predictor variables, but our confidence intervals suggest that if we did the same model selection process on many samples, these predictor variables would not always be chosen.

VI. Prediction

For this part we will be predicting whether or not a specific patient has chest pain, and then report on measures of prediction.

A. Pi - hat

Our model predicts that the odds of angina for a 50 year old who has never smoked, with a mild history of hypertension, history of stroke, and no other history of medical issues is 0.1380. The probability that this patient has angina is about 0.1213. We would predict that this patient does not have angina.

B. Proportion of Reduction in Error

Our PRE (Proportion of Reduction in Error) value was 0.5431, which means that our model achieves a 54.31% reduction in error compared to using the MLE (Maximum Likelihood Estimator), which is y-bar.

C. Error Matrices for Cutoff Values

Comparing $\pi_0 = 0.50$, the MLE, to the median of our model fitted values, 0.4770:

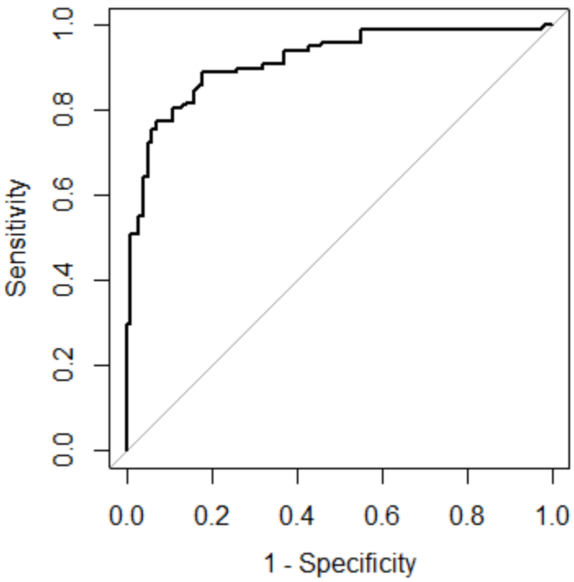
$\pi_0 = 0.50$			$\pi_0 = 0.4770$		
	Predicted = 0	Predicted = 1		Predicted = 0	Predicted = 1
Truth = 0	84	16	Truth = 0	84	16
Truth = 1	18	80	Truth = 1	15	83

Sensitivity	Specificity	Error Rate	Sensitivity	Specificity	Error Rate
0.8163	0.84	0.1717	0.8469	0.84	0.1566

We chose to test the median of the model fitted values as a cutoff value because we know that exactly 100 patients in the sample have angina and exactly 100 do not. Since using the median of the model fitted values results in better sensitivity and error rate with no loss to specificity, we would suggest that this is a better cutoff value than 0.50.

However, in order to test overall model fit with many different cutoff values, we generated the ROC plot:

D. ROC Plot and AUC



This plot shows that this model fits the data well since we have high values for sensitivity at low values for 1-specificity, as well as a large area under the curve (AUC).

95% Confidence Interval for AUC		
Lower Bound	Measured AUC	Upper Bound
0.8766	0.9158	0.955

We are 95% confident that the true AUC for our model is between 0.8766 and 0.955, which indicates that our model is a very good to excellent fit.

VII. Conclusion

We found that the most important predictors were the predictors chosen by the model, which are age, smoking status, average cigarette consumption per day, and family histories of hypertension and heart attack. For a patient, the most important predictor was actually his or her age, since $1.1142^{50} = 222.9245$ times the base odds of angina.

VIII. R Appendix

II. Data Summary

```
data = read.csv("angina.csv")
str(data)
```

```
attach(data)
#cig and age are both not going to have mu ij > 5 for all, either
table(data$diabetes)  #only 9 patients had diabetes, so this is probably not going to be very helpful.
table(data$smoke)  #reasonable distribution here.
table(data$strokefam)  #again, 12 is a little low, but maybe it's useful.
table(data$hyper)  #same
table(data$y)   #so we know how our sample was chosen, 100 of each.
table(data$myofam)  #reasonable dist
hist(cig,breaks=30)
ha = data[which(y==1),]
nh = data[which(y==0),]
hist(ha$cig)
hist(nh$cig)
#fit the full model and see if anything looks odd.
full.model = glm(y~.,data = data,family=binomial(link=logit))
summary(full.model)
```

A. <u>Age</u>

```
ggplot(data,aes(age,fill=y))+geom_histogram()+xlab("Patient Age")+ylab("Number of Patients")+ggtitle("Sample Age Ranges")+theme_bw()+facet_grid(y~.,)+scale_y_continuous(breaks = seq(0, 12, by= 4))+ggsave("SAMPLE age range.png",height=4,width=5)
```

B. Smoking Status

```
table(data$smoke,data$y)
```

C. Number of Cigarettes Per Day

Note: according to Google results, a pack of cigarettes is usually 20 cigarettes in the US. So our results peaking in breaks of 10 is probably not surprising, nor is the second largest result being 1 pack.

```
library(ggplot2)
dev.off()
ggplot(data,aes(cig,fill=y))+geom_histogram()+xlab("Number of Cigarettes Smoked Per Day")+ylab("Number of Patients")+ggtitle("Cigarette Smoking in Sample")+theme_bw()+facet_grid(y~.,)
```

D. Family History: Hypertension

```
table(data$hyper,data$y)  #mu ij less than 5
```

E. Family History: Heart Attack (Myocardial Infarction)

```
table(data$myofam,data$y)
```

F. Family History: Stroke

```
table(data$strokefam,data$y)
```

G. Family History: Diabetes

```
table(data$diabetes,data$y)
#note mu ij less than 5
```

<u>III. Data Preparation</u>

A. <u>Pearson's Residuals</u>

```
library(LogisticDx)
good.stuff = dx(full.model)
good.stuff = as.data.frame(good.stuff)
pear.r = good.stuff$Pr #Pearsons Residuals
std.r = good.stuff$sPr #Standardized residuals (Pearson)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
change.pearson = good.stuff$dChisq
#Change in pearson X^2 for each observation
hist(std.r, main = "Pearson Standardized Residuals")
which(abs(std.r)>3)
data[which(abs(std.r)>3),]
good.stuff = dx(our.model)
good.stuff = as.data.frame(good.stuff) #Convert to dataframe because of annoying things
```

```
pear.r = good.stuff$Pr #Pearsons Residuals
std.r = good.stuff$sPr #Standardized residuals (Pearson)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
hist(std.r, main = "Pearson Standardized Residuals")
#There are unusual points in the data.
which(abs(std.r)>3)
data[which(abs(std.r)>3),]
```

B. DF Beta

```
hist(df.beta)
which(df.beta > 0.3)
data[which(df.beta>0.3),]
```

- ended up removing 166

```
data = data[-166,]
```

IV. Model Selection and Analysis

A. Model Criteria

```
empty.model = glm(y~1, data = data, family=binomial(link=logit))
full.model = glm(y~.,data = data,family=binomial(link=logit))
BIC.select.F = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction = "both",criterion = "BIC",trace=FALSE)
BIC.select.F$formula
BIC.select.B = step(full.model,scope = list(lower = empty.model, upper = full.model),direction = "both",criterion = "BIC",trace=FALSE)
BIC.select.B$formula
AIC.select.F = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction = "both",criterion = "AIC",trace=FALSE)
AIC.select.F$formula
AIC.select = step(full.model,scope = list(lower = empty.model, upper = full.model),direction = "both",criterion = "AIC",trace=FALSE)
AIC.select$formula
```

B. All Subsets

```
y = data$y
data = data[,-1]
data$y=y
library(bestglm)
best.subset.BIC = bestglm(Xy = data, family = binomial(link=logit),IC = "BIC",method = "exhaustive")
best.subset.BIC
best.subset.AIC = bestglm(Xy = data, family = binomial(link=logit),IC = "AIC",method = "exhaustive")
best.subset.AIC
our.model = glm(AIC.select$formula,data=data,family=binomial(link=logit))
Our.model
```

C. Interaction Terms

```
model1 = glm(y ~ age + smoke + cig + hyper + myofam + age*smoke,data=data,family=binomial(link=logit))
BIC(model1)
BIC(our.model)
#not an improvement.
model2 = glm(y ~ age + smoke + cig + hyper + myofam + age*cig,data=data,family=binomial(link=logit))
BIC(model2)
BIC(our.model)
#not an improvement
model3 = glm(y ~ age + smoke + cig + hyper + myofam + age*hyper,data=data,family=binomial(link=logit))
BIC(model3)
BIC(our.model)
#Not an improvement
```

```
model4 = glm(y ~ age + smoke + cig + hyper + myofam + age*myofam,data=data,family=binomial(link=logit))
BIC(model4)
BIC(our.model)
#no change.
model5 = glm(y ~ age + smoke + cig + hyper + myofam + smoke*cig,data=data,family=binomial(link=logit))
BIC(model5)
BIC(our.model)
#no change.
model6 = glm(y ~ age + smoke + cig + hyper + myofam + smoke*hyper,data=data,family=binomial(link=logit))
BIC(model6)
BIC(our.model)
#no improvement.
model7 = glm(y ~ age + smoke + cig + hyper + myofam + smoke*myofam,data=data,family=binomial(link=logit))
BIC(model7)
BIC(our.model)
#no improvement
model8 = glm(y ~ age + smoke + cig + hyper + myofam + cig*hyper,data=data,family=binomial(link=logit))
BIC(model8)
BIC(our.model)
#no improvement
model9 = glm(y ~ age + smoke + cig + hyper + myofam + cig*myofam,data=data,family=binomial(link=logit))
BIC(model9)
BIC(our.model)
#no improvement
model10 = glm(y ~ age + smoke + cig + hyper + myofam + hyper*myofam,data=data,family=binomial(link=logit))
BIC(model10)
BIC(our.model)
```

- no improvement.
- So as far as model correctness goes, there doesn't seem to be a reason to add any interaction terms.

D. **<u>Final Model</u>**

```
our.model = glm(AIC.select$formula,data=data,family=binomial(link=logit))
our.model
```

E. **<u>Corrected Confidence Intervals (g=8)</u>**

```
alpha = 0.05/8
conf = confint(our.model, level = 1-alpha)
exp(conf)
```

VI. Prediction

A. **<u>Pi-hat</u>**

```
predict(our.model,newdata = data.frame(age=50,smoke="never",cig=0,hyper="mild",myofam="no"),type="response")
```

B. **<u>Proportion of Reduction of Error</u>**

```
r = cor(our.model$y, our.model$fitted.values)
r
prop.red = 1- sum((our.model$y - our.model$fitted.values)^2)/sum((our.model$y - mean(our.model$y))^2)
prop.red
```

C. **<u>Error Matrices</u>**

```
pi.0 = 0.50
truth = data$y #The true values of y
predicted = ifelse(fitted(our.model)>pi.0,1,0)
#The predicted values of y based on pi.0
my.table = table(truth,predicted)
sens = sum(predicted == 1 & truth == 1 )/sum(truth == 1)
spec = sum(predicted == 0 & truth == 0 )/sum(truth == 0)
```

```
error = sum(predicted !=
truth)/length(predicted)
results = c(sens,spec,error)
names(results) =
c("Sensitivity","Specificity","Error-
Rate")
results
my.table
median(our.model$fitted.values)
pi.0 = 0.477
truth = data$y #The true values of y
predicted =
ifelse(fitted(our.model)>pi.0,1,0)
#The predicted values of y based on
pi.0
my.table = table(truth,predicted)
sens = sum(predicted == 1 & truth
== 1 )/sum(truth == 1)
spec = sum(predicted == 0 & truth
== 0 )/sum(truth == 0)
error = sum(predicted !=
truth)/length(predicted)
results = c(sens,spec,error)
names(results) =
c("Sensitivity","Specificity","Error-
Rate")
results
My.table
```

D. **<u>ROC and AUC plots</u>**

```
library(pROC)
my.auc =
auc(our.model$y,fitted(our.model),pl
ot = TRUE,legacy.axes = TRUE)
My.auc
auc.CI = ci(my.auc,level = 1-0.05)
#for a 95% confidence interval.
auc.CI
```