Project II - Angina

Erin Melcon(Instructor)
Dandi Peng 915553480, Yuhan Ning 915486450

3/1/2019

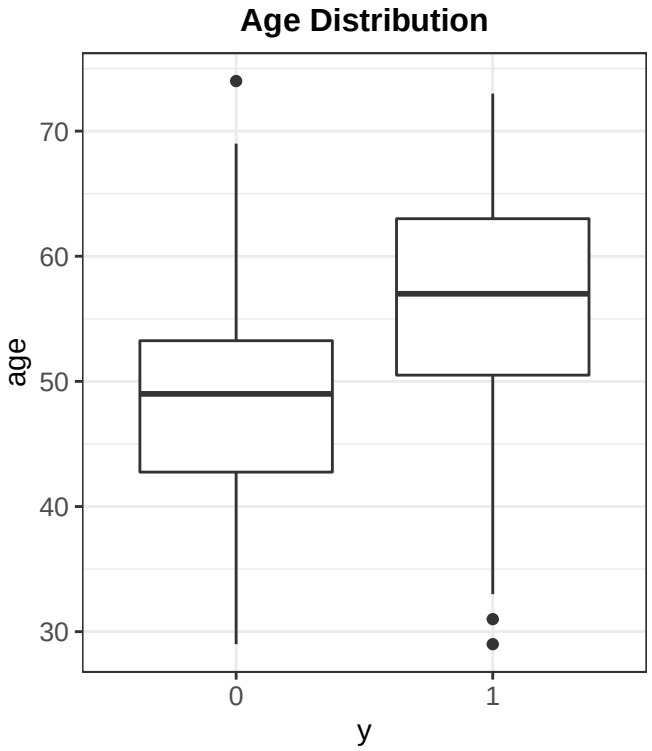
I. Summary

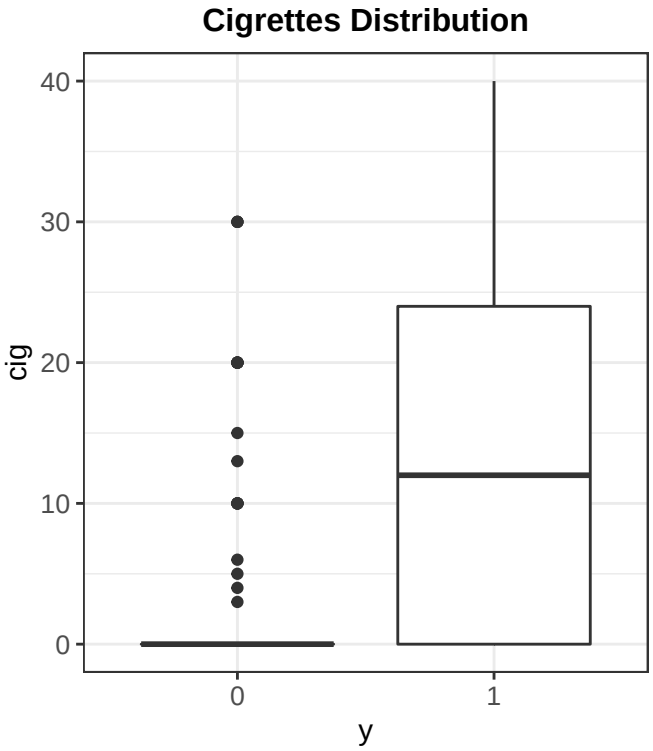
We choose Problem II since we are interested in how the subject's health status and habits related to angina. In this project, we aim to build a model where Y = angina status and predict the probability of angina for a 50 year old who has never smoked, with history of hypertension, angina, and stroke.

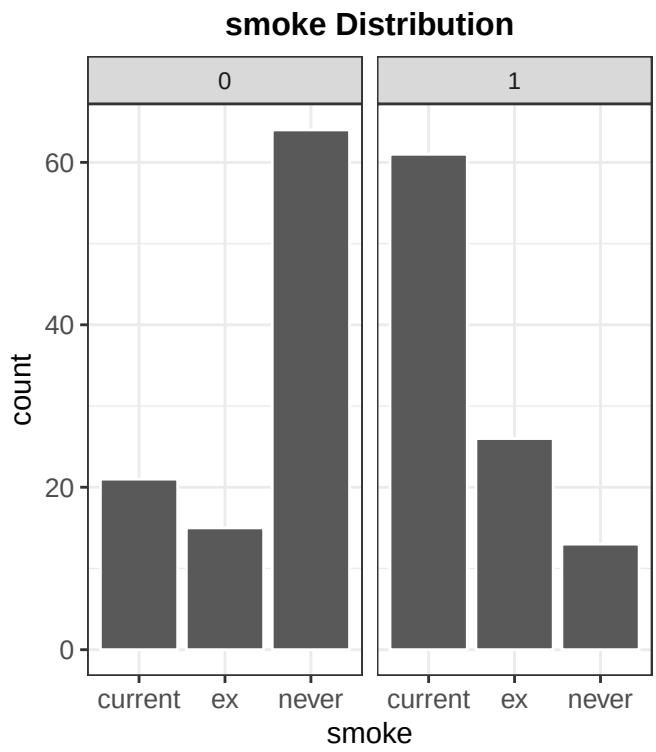
First we take a glance at the dataset, there are 200 samples in total. The table of counts for each categorical explanatory variables and five number for each numerical explanatory variables are shown below.

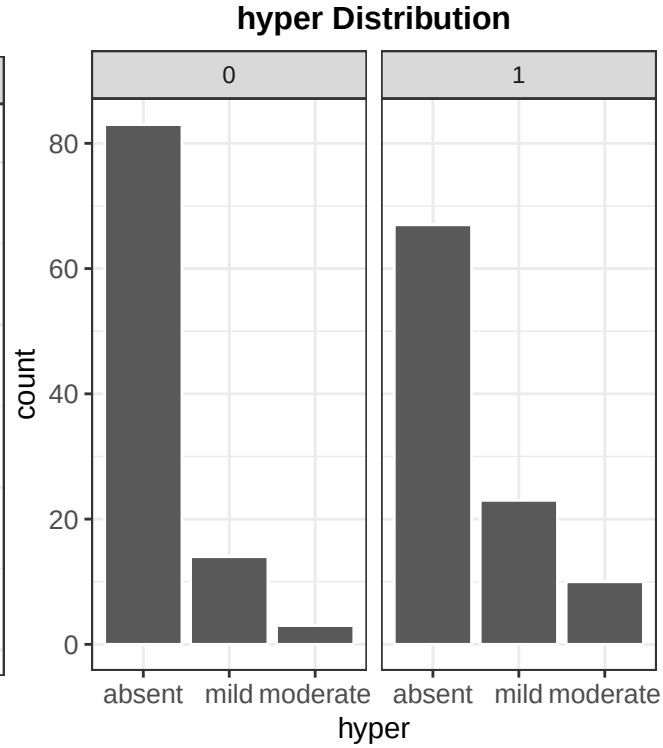
Variable	Summary	Value	Variable	Summary	Value
	Min.	29.00		current	82
	1st Qu.	46.00	smoke	ex	41
age	Median	53.00		never	77
	Mean	52.45		absent	150
	3rd Qu.	59.00	hyper	mild	37
	Max.	74.00		moderate	13
	Min.	0.000	myofam	no	135
	1st Qu.	0.000		yes	65
cig	Median	0.000	strokefam	no	188
	Mean	8.205		yes	12
	3rd Qu.	16.250	diabetes	no	191
	Max.	40.000		yes	9
	Min.	0.0			
	1st Qu.	0.0			
y	Median	0.5			
	Mean	0.5			
	3rd Qu.	1.0			
	Max.	1.0			

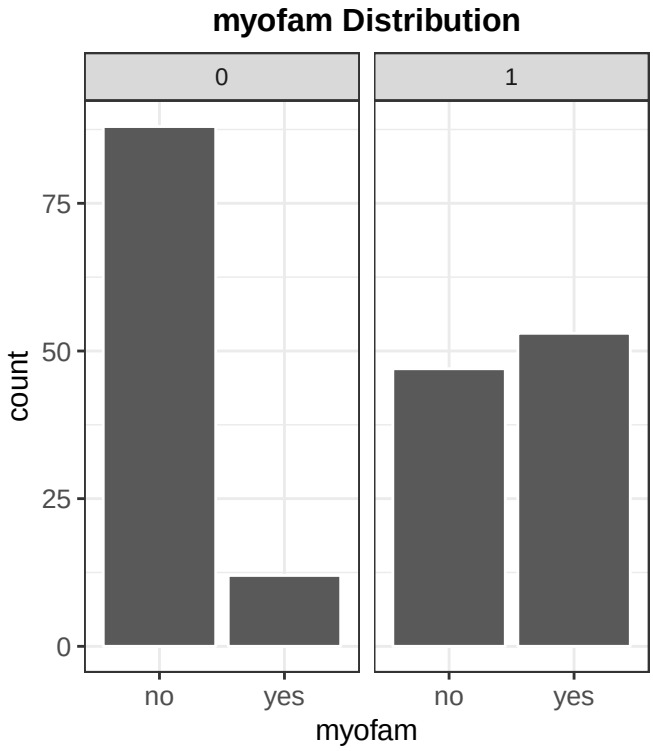
Besides, explore the relationship between our explanatory variables and response variable using ggplot.

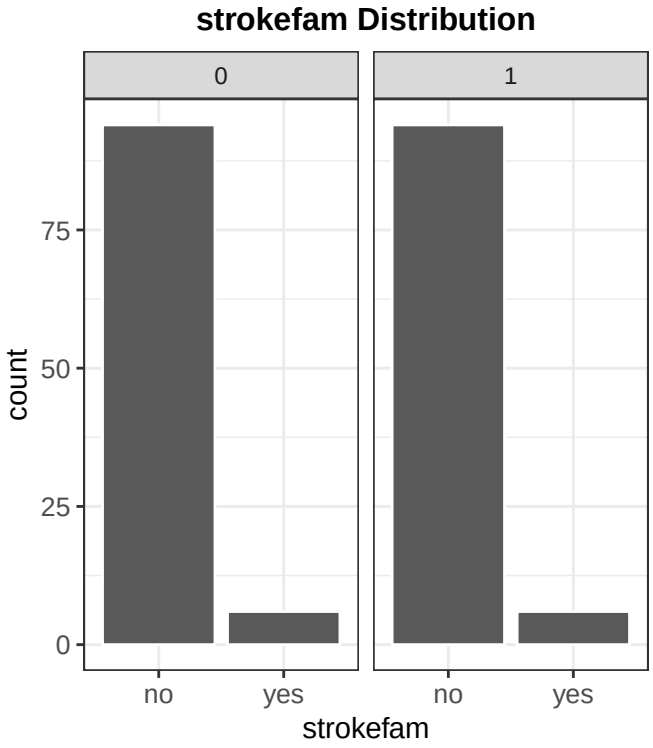

Age Distribution
age
70
60
50
40
30
0
1
y

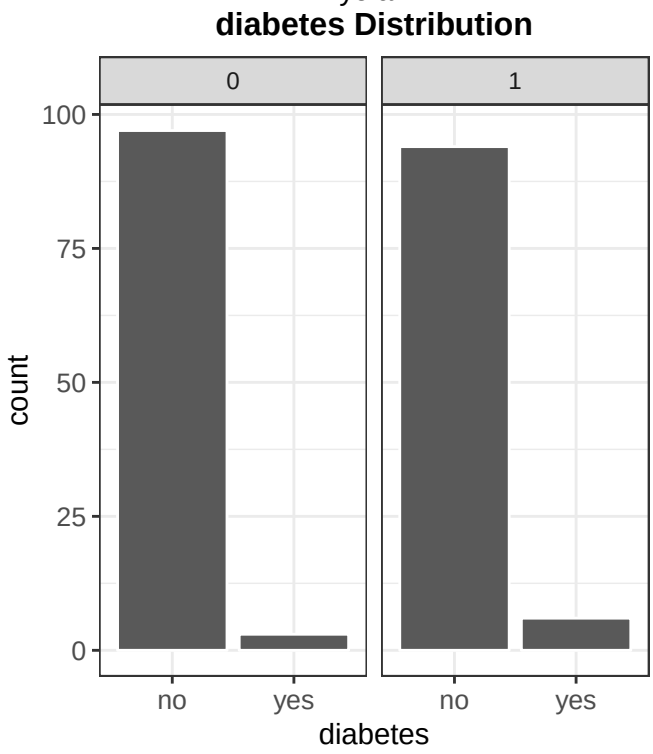

Cigrettes Distribution
cig
40
30
20
10
0
0
1
y


smoke Distribution
0
1
count
60
40
20
0
current
ex
never
current
ex
never
smoke


hyper Distribution
0
1
count
80
60
40
20
0
absent
mild
moderate
absent
mild
moderate
hyper

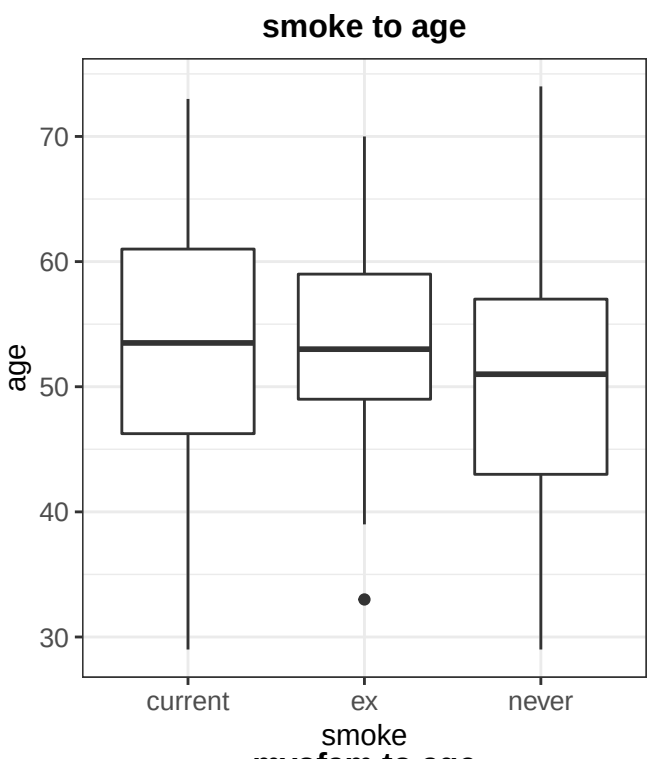


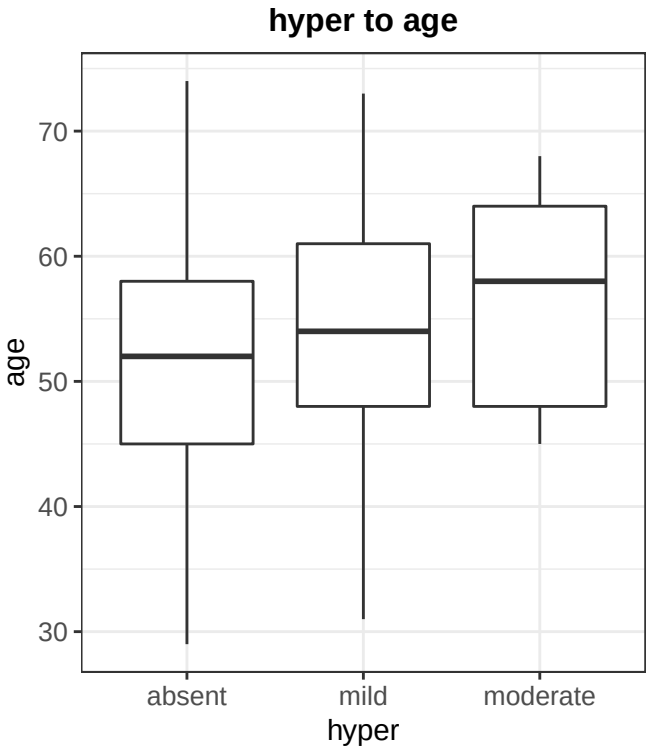


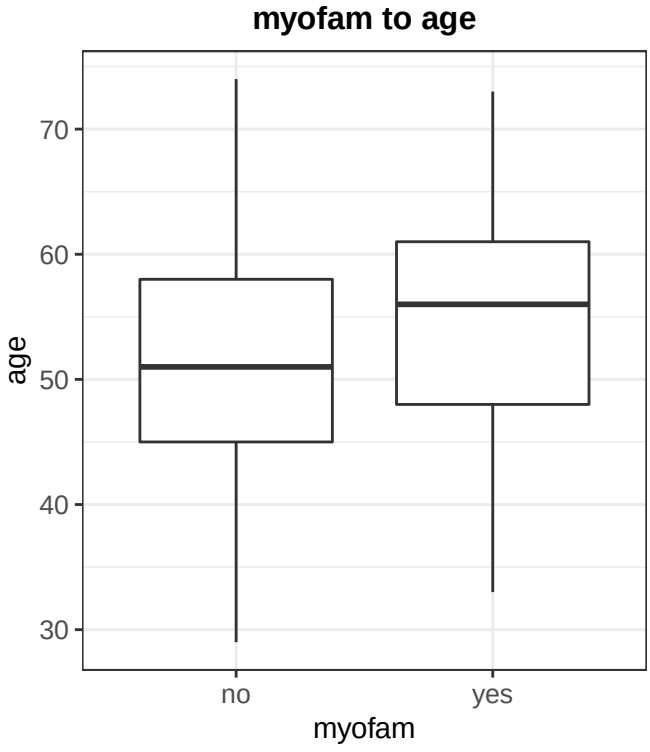


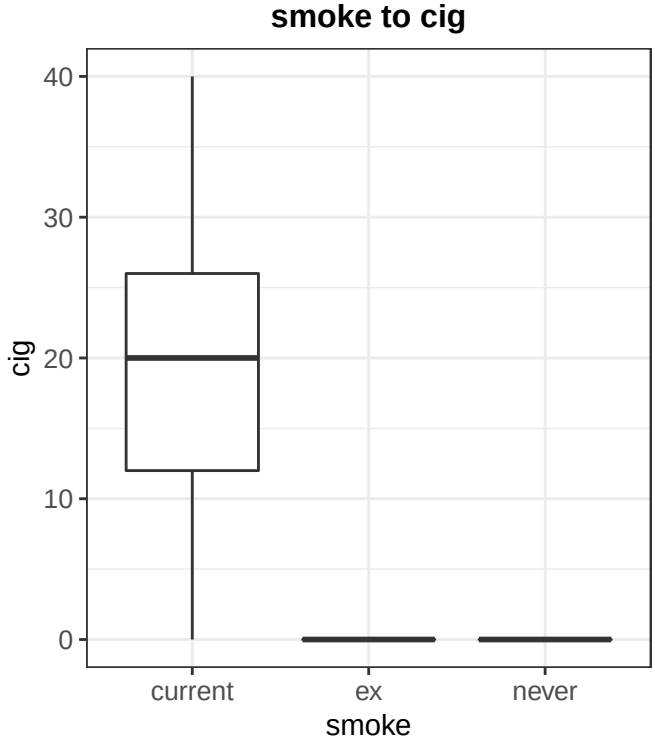
Based on these plots, we see that for 'y vs. strokefam' and 'y vs. diabetes', there is no significant distribution difference for Y=1 and Y=0. This may suggest variables 'strokefam' and 'diabetes' have no effect on predicting angina status.

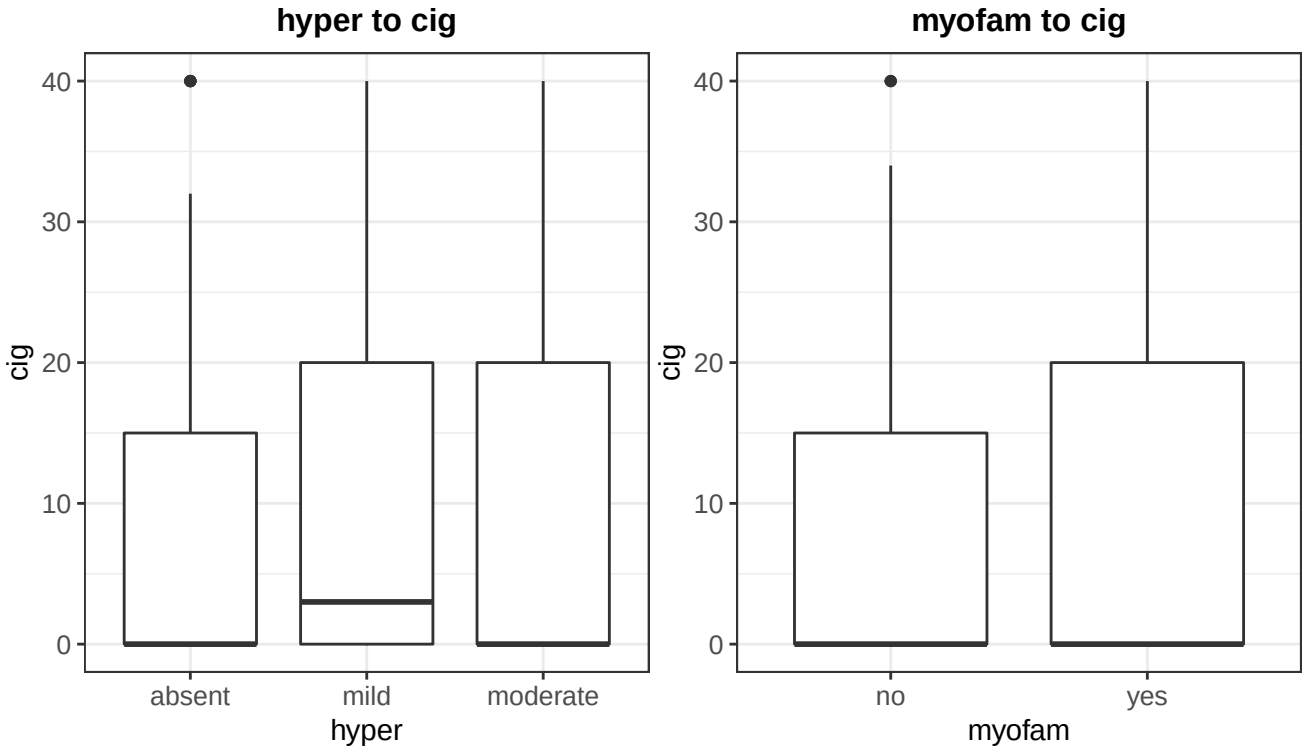
Also, plot among explanatory variables to check dependence of either two variables:


smoke to age
age
70
60
50
40
30
current
ex
never
smoke


hyper to age
age
70
60
50
40
30
absent
mild
moderate
hyper


myofam to age
age
70
60
50
40
30
no
yes
myofam

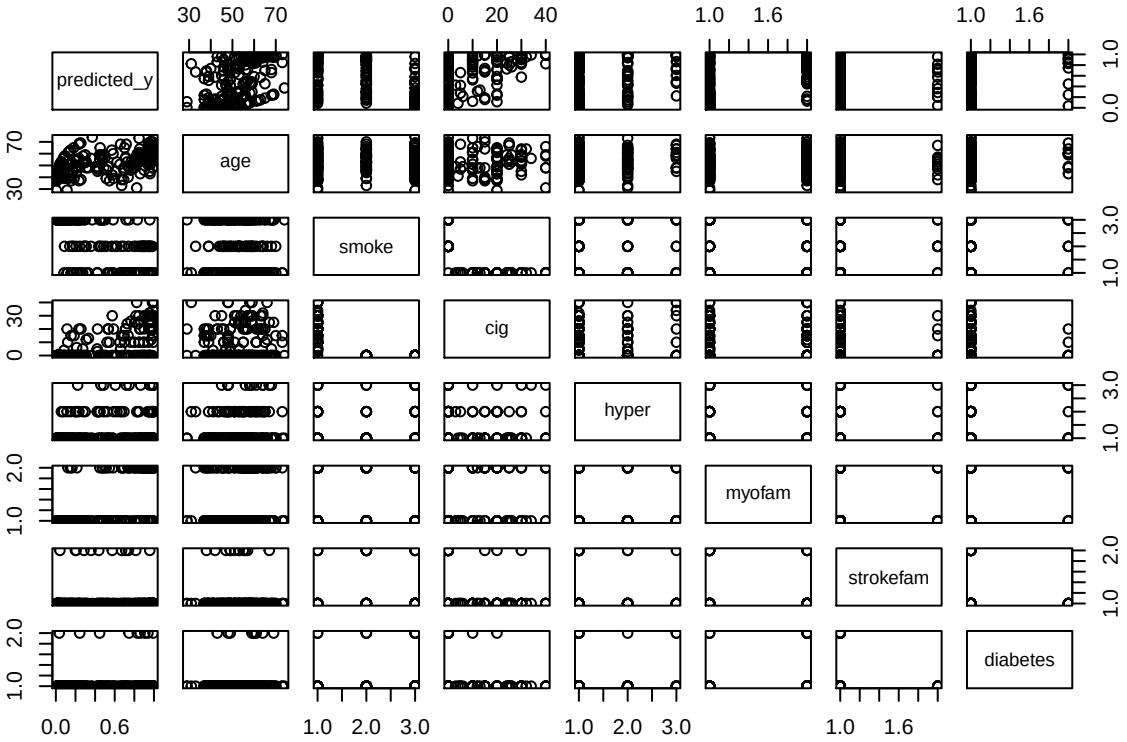

smoke to cig
cig
40
30
20
10
0
current
ex
never
smoke



From the above plots for continuous variables to categorical variables, the interactions between 'myofam' and 'age', between 'hyper' to 'age' are initially suggested to be checked in later modeling.

Since Y = angina is a binomial variable, we build a logistic regression with all explanatory variables and find their relationship. Using this model to predict probability of Y = 1 and plot matrix plot to see the relationship between them.

```
## $coefficients
##                 Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)   -7.2764153 1.57474921 -4.6206820 3.824806e-06
## age            0.1091375 0.02589784  4.2141556 2.507145e-05
## smokeex        0.6820394 0.85538120  0.7973514 4.252470e-01
## smokenever    -1.3262547 0.84063608 -1.5776800 1.146391e-01
## cig            0.1031558 0.03937200  2.6200295 8.792216e-03
## hypermild      1.3132511 0.54805819  2.3961892 1.656654e-02
## hypermoderate  2.1130279 0.88903793  2.3767579 1.746555e-02
## myofamyes      2.3921758 0.50604951  4.7271577 2.276844e-06
## strokefamyes  -0.1004183 0.75102140 -0.1337089 8.936327e-01
## diabetesyes   -0.0605750 1.02289107 -0.0592194 9.527774e-01
```

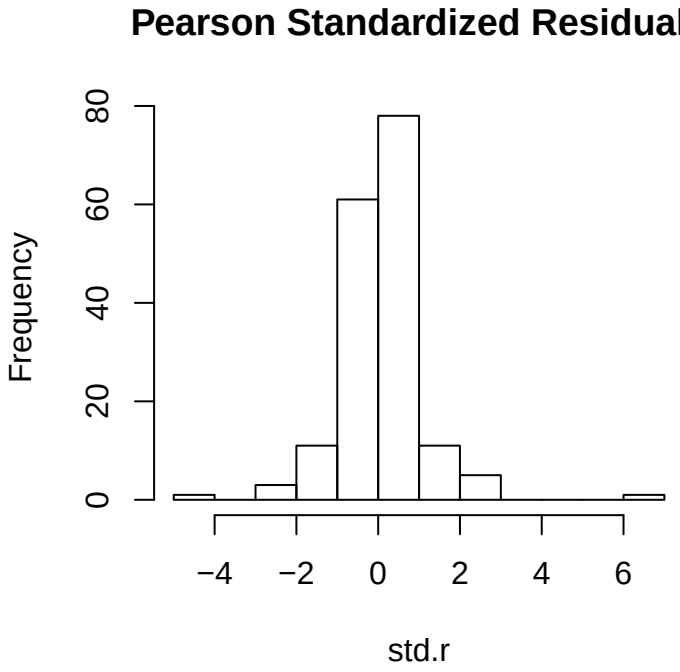


Based on the plot, we see no strong multicollinearity among X variables and the model summary suggest that 'smoke', 'strokefam' and 'diabetes' variables may be insignificant and we will further check in the following steps.

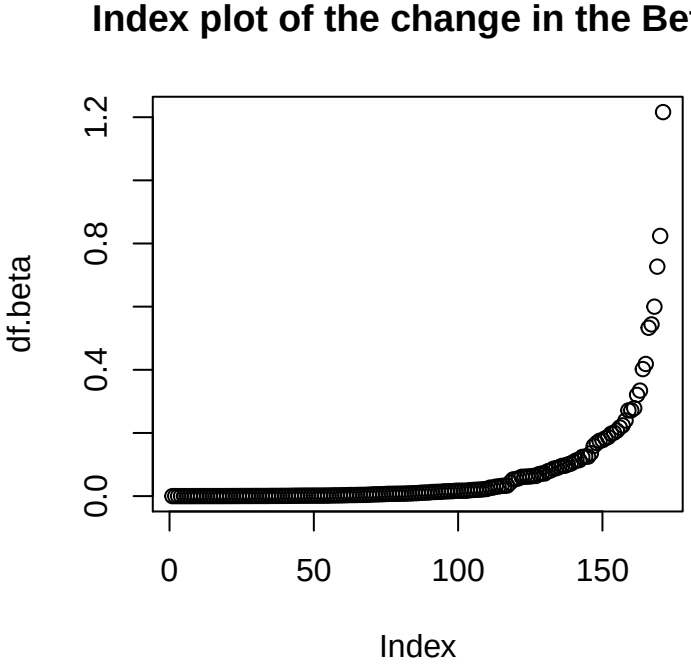
II. Data Preparation

Continually, use the full model to consider outliers and influential points. Set cutoffs of pearson standardized residual as 3, change in the betas as 0.5 and change in the pearson test-statistic as 10.

The result plots are shown below.

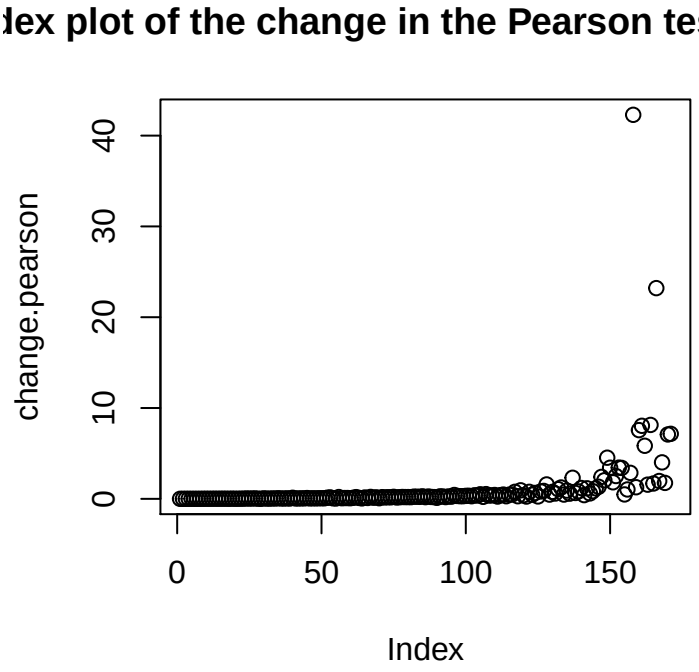



```
##     (Intercept) age smokeex smokenever cig hypermild hypermoderate
## 158           1  38       0          1   0         0             0
## 166           1  45       0          0  30         0             0
##     myofamyes strokefamyes diabetesyes y       sPr
## 158         0            0          0 1  6.503159
## 166         1            0          0 0 -4.817092
```




```
##     (Intercept) age smokeex smokenever cig hypermild hypermoderate
## 166           1  45       0          0  30         0             0
## 167           1  45       1          0   0         0             1
```

```
## 168           1  48       0           1   0         0              1
## 169           1  59       1           0   0         0              0
## 170           1  58       1           0   0         0              1
## 171           1  43       0           0  20         0              0
##     myofamyes strokefamyes diabetesyes y     dBhat
## 166         1            0           0 0 0.5331769
## 167         0            0           0 0 0.5438178
## 168         0            0           0 1 0.5999283
## 169         0            0           1 1 0.7266942
## 170         0            0           0 0 0.8239320
## 171         1            0           1 0 1.2161134
```




```
##     (Intercept) age smokeex smokenever cig hypermild hypermoderate
## 158           1  38       0           1   0         0              0
## 166           1  45       0           0  30         0              0
##     myofamyes strokefamyes diabetesyes y   dChisq
## 158         0            0           0 1 42.29108
## 166         1            0           0 0 23.20438
```

Based on above analysis, index 158 and 166 are considered as outliers and index 166 also represents as influential point. Since outliers do not help describe what happens 'in average', we drop index 158 and 166.

III. Model Selection/Analysis

III-1 Preliminary Model Selection

Now we will do model selection (backward/forward selection, forward/backward selection and all subset selection method) using 'AIC' criteria since our goal is to predict.

```
## Loading required package: leaps

## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

## y ~ age + smoke + cig + hyper + myofam

## y ~ smoke + myofam + age + cig + hyper

## $coefficients
##                  Estimate Std. Error    z value     Pr(>|z|)
## (Intercept)   -7.1835405 1.57131543 -4.5716731 4.838453e-06
## age            0.1069600 0.02579783  4.1460841 3.382097e-05
## smokeex        0.7647024 0.84366357  0.9064068 3.647206e-01
## smokenever    -1.3204902 0.83826440 -1.5752670 1.151948e-01
## cig            0.1028389 0.03908515  2.6311493 8.509664e-03
## hypermild      1.2937617 0.54918777  2.3557729 1.848421e-02
## hypermoderate  2.0952788 0.88857345  2.3580254 1.837244e-02
## myofamyes      2.4718877 0.51260788  4.8221805 1.419974e-06
```

Therefore, all methods agree and select $y \sim age + smoke + cig + hyper + myofam$ as our 'best' model.

The estimated logistic regression function is: $log\frac{\hat{\pi}}{1-\hat{\pi}} = -7.1835 + 0.1070X_1 + 0.7647X_{2,ex} - 1.3205X_{2,never} + 0.1028X_3 + 1.2938X_{4,mild} + 2.0953X_{4,moderate} + 2.4719X_{5,yes}$, where X_1, X_2, X_3, X_4, X_5 represent age, smoke, cig, hyper and myofam respectively.

The only concern is that the significancy of smoke variable is not satisfied. The following two hypothesis tests are applied to manage this situation.

```
##                                      LL p   n      AIC      BIC
## y~age+myofam+hyper+cig       -78.63790 6 198 169.2758 189.0054
## y~smoke+age+myofam+hyper+cig -71.38502 8 198 158.7700 185.0762
```

LR Test-1

$$H_0 : \beta_{2,ex} = 0,\ \beta_{2,nv} = 0$$

$$H_a :\ at\ least\ one\ of\ \beta_{2,i} \neq 0$$

Based on the above output, the test statistics is $G^2 = -2(L_0 - L_1) = -2(-78.63790 - (-71.38502)) = 14.50576$, and the d.f. $= 8 - 6 = 2$.

The corresponding p value $= P(\chi_2^2 > G^2) = 0.000708$, which is less than any $\alpha's$, therefore, we reject the null hypothesis and cannot drop the smoke variable.

Wald Test

$$H_0 : \beta_{2,ex} = 0, \quad H_a : \beta_{2,ex} \neq 0$$

The Wald test statistics is $\frac{\hat{\beta_{2,ex}} - 0}{SE(\hat{\beta_{2,ex}})} = \frac{0.7647}{0.8437} = 0.9064$, and its corresponding p value is $P(Z^2 > 0.786) = 0.36472$, which is large than any $\alpha's$, therefore, we fail to reject the null hypothesis and can drop $\beta_{2,ex}$.

Combined the above two tests, we can conclude that smoke varaible should be contained, but there is no significant difference between "ex" and "current" smoking status and we can merge them to be one level - 'some history smoking' vs. the rest 'never smoked'.

Repeat the procedure and fit the new model:

```
## Morgan-Tatar search since family is non-gaussian.

## Note: factors present with more than 2 levels.

## $coefficients
##                   Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)   -8.58386128 1.57694016 -5.443365 5.228328e-08
## age            0.10822693 0.02574789  4.203332 2.630147e-05
## smokesmoked    1.90060283 0.53860763  3.528734 4.175533e-04
## cig            0.07656995 0.02426222  3.155934 1.599851e-03
## hypermild      1.26735475 0.54602880  2.321040 2.028468e-02
## hypermoderate  2.15362579 0.88139215  2.443437 1.454813e-02
## myofamyes      2.49375445 0.51138214  4.876499 1.079853e-06
```

Now our best model is selected as $y \sim age + smoke + cig + hyper + myofam$ and the estimated logistic regression function is: $log\frac{\hat{\pi}}{1-\hat{\pi}} = -8.58386 + 0.10823X_1 + 1.90060X_{2,smoked} + 0.07657X_3 + 1.26735X_{4,mild} + 2.15363X_{4,moderate} + 2.49375X_{5,yes}$, where X_1, X_2, X_3, X_4, X_5 represent age, smoke, cig, hyper and myofam respectively.

III-2 Check Interaction Terms

Next, we need to check whether interaction terms between these variables are needed.

Still use LR Test and check all possible 10 interaction terms. H_0 is model with no interaction is a better fit. Set $\alpha = 0.01$.

LR Test-2

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                                              LL p   n     aic     bic
## y~age+smoke+cig+hyper+myofam             -71.803 7 198 157.606 180.624
## y~age+smoke+cig+hyper+myofam+age*smoke   -71.357 8 198 158.715 185.021
## y~age+smoke+cig+hyper+myofam+age*cig     -71.781 8 198 159.562 185.868
```

```
## y~age+smoke+cig+hyper+myofam+age*hyper    -70.834 9 198 159.669 189.263
## y~age+smoke+cig+hyper+myofam+age*myofam   -69.391 8 198 154.782 181.088
## y~age+smoke+cig+hyper+myofam+smoke*cig    -71.803 8 198 159.606 185.912
## y~age+smoke+cig+hyper+myofam+smoke*hyper  -68.962 9 198 155.923 185.518
## y~age+smoke+cig+hyper+myofam+smoke*myofam -71.792 8 198 159.584 185.890
## y~age+smoke+cig+hyper+myofam+cig*hyper    -71.419 9 198 160.838 190.433
## y~age+smoke+cig+hyper+myofam+cig*myofam   -69.840 8 198 155.681 181.987
## y~age+smoke+cig+hyper+myofam+hyper*myofam -70.583 9 198 159.166 188.760
```

```
## p-value for model: 0.344935901804226p-value for model: 0.833853612615098p-value for m
```

All models with interaction have p-values larger than 0.01, so we fail to reject H_0 and conclude that model with no interactions is a better fit. Till now, we keep the last best model and it is our final model.

III-3 Confidence Interval

We calculate the 95% Wald confidence intervals for each beta:

```
##                lower.bounds2 upper.bounds2
## (Intercept)    -11.67460720    -5.4931154
## age              0.05776199     0.1586919
## smokesmoked      0.84495127     2.9562544
## cig              0.02901688     0.1241230
## hypermild        0.19715798     2.3375515
## hypermoderate    0.42612893     3.8811227
## myofamyes        1.49146387     3.4960450
```

Also, we calculate the 95% profile likelihood CI:

```
## Waiting for profiling to be done...
```

```
##                     2.5 %     97.5 %
## (Intercept)   -11.94673218 -5.7163011
## age             0.06019072  0.1619037
## smokesmoked     0.87721583  3.0046977
## cig             0.03154052  0.1274660
## hypermild       0.21729553  2.3723775
## hypermoderate   0.49428577  4.0178872
## myofamyes       1.54113512  3.5606339
```

IV. Interpretation

In part IV, we get the estimated logistic regression function is: $log\frac{\hat{\pi}}{1-\hat{\pi}} = -8.58386 + 0.10823X_1 + 1.90060X_{2,smoked} + 0.07657X_3 + 1.26735X_{4,mild} + 2.15363X_{4,moderate} + 2.49375X_{5,yes}$

Now, we interprete the coefficients based on model and their 95% profile CI as the following:

$exp(0.10823)$: When the age increases by 1 year, the estimated odds of having angina is $exp(0.10823) = 1.1143$ (between $exp(0.0602)$ and $exo(0.1619)$) times for what they were, holding other variables constant.

$exp(1.90060)$: The estimated odds of having angina is for subject who smoked is $exp(1.90060) = 6.689907$ (between $exp(0.8772)$ and $exp(3.0047)$) times for subject who never smoked, holding other variables constant.

$exp(0.07637)$: When the subject smokes one more cigarette per day, the estimated odds of having angina is $exp(0.07637) = 1.079362$ (between $exp(0.0315)$ and $exp(0.1275)$) times for what they were, holding other variables constant.

$exp(1.26735)$: The estimated odds of having angina is for subject who has mild hypertension history in family is $exp(1.26735) = 3.551429$ (between $exp(0.2173)$ and $exp(2.3724)$) times for subject who has no hypertension history, holding other variables constant.

$exp(2.15363)$: The estimated odds of having angina is for subject who has moderate hypertension history in family is $exp(2.15363) = 8.616078$ (between $exp(0.4943)$ and $exp(4.0179)$) times for subject who has no hypertension history, holding other variables constant.

$exp(2.47375)$: The estimated odds of having angina is for subject who has myocardial infarction history in family is $exp(2.47375) = 11.86686$ (between $exp(1.5411)$ and $exp(3.5606)$) times for subject who has no history, holding other variables constant.

From the confidence intervals, since they are all strictly above 0, we are 95% confident that the estimated odds of having angina increases with the age, number of cigarette per day, history of hypertension in family and history of myocardial infarction in family.

V. Prediction

Set our cutoff as 0.5, now we measure the sensitivity, specificity and error rate based on error matrix:

```
##      predicted
## truth  0  1
##     0 82 17
##     1 17 82
```

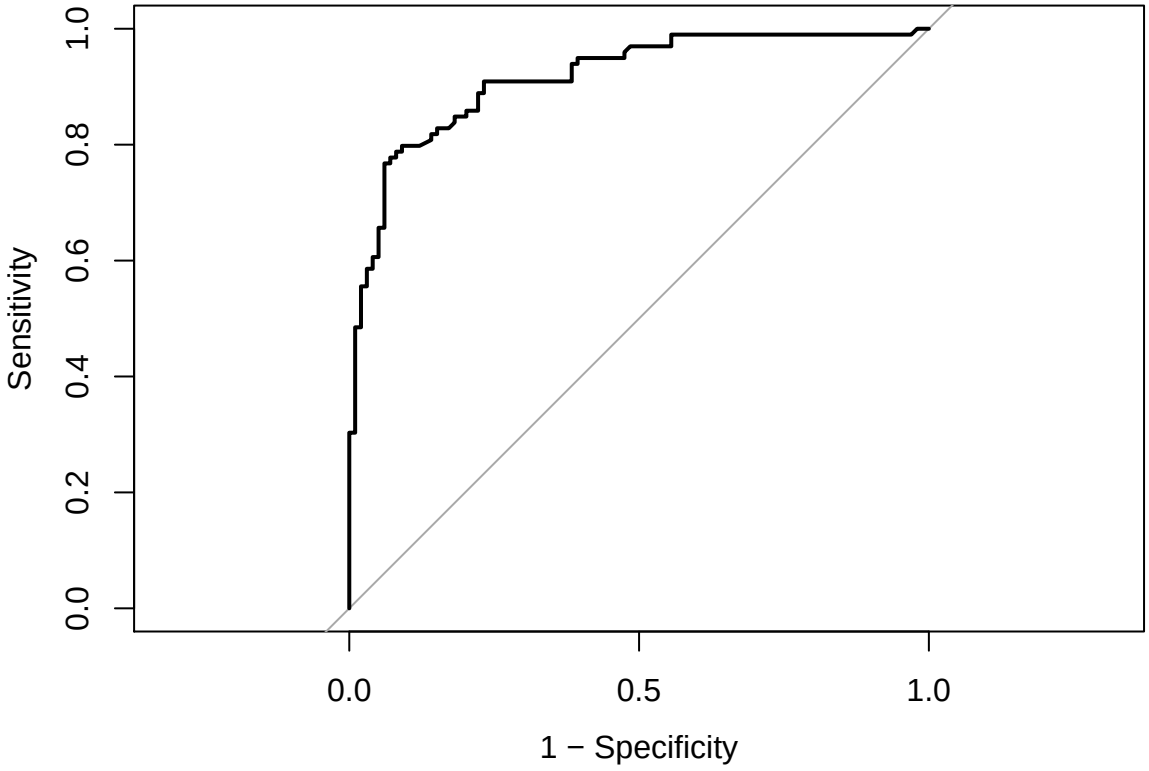
```
## Sensitivity Specificity  Error-Rate
##   0.8282828   0.8282828   0.1717172
```

Also plot the ROC and calculate AUC to assess our 'best' model:

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```




```
## Area under the curve: 0.9156
```

```
## 95% CI: 0.8766-0.9545 (DeLong)
```

This indicates that our fit does a reasonable job, since AUC is above 0.8.

Using the 'best' model we choose, the probability of angina for a 50 year old who has never smoked, with history of angina, and stroke (and no other history of medical issues) is 0.1295256. Since it is less than cutoff, the subject is predicted to have no angina.

VI. Conclusion

```
## $coefficients
##                  Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)   -8.58386128 1.57694016 -5.443365 5.228328e-08
## age            0.10822693 0.02574789  4.203332 2.630147e-05
## smokesmoked    1.90060283 0.53860763  3.528734 4.175533e-04
## cig            0.07656995 0.02426222  3.155934 1.599851e-03
## hypermild      1.26735475 0.54602880  2.321040 2.028468e-02
## hypermoderate  2.15362579 0.88139215  2.443437 1.454813e-02
## myofamyes      2.49375445 0.51138214  4.876499 1.079853e-06
```

From summary of our model, we see that age and myofam have extremely small p-values and they have large impact on our model: $y \sim age + smoke + cig + hyper + myofam$. The odds of having angina increase with the increase of age, number of cigarettes per day, and the existence of family history on myocardial infarction.

R Appendix

```
knitr::opts_chunk$set(echo = TRUE)
data = read.csv('angina.csv')
n = nrow(data)
#summary(data)
data$y = as.factor(data$y)
library(ggplot2)
ggplot(data=data)+geom_boxplot(aes(x=y,y=age))+theme_bw()+
  ggtitle('Age Distribution')+
  theme(axis.text=element_text(size=10),
          plot.title = element_text(size = 12, hjust=0.5, face = 'bold'),
          legend.position="none")
ggplot(data=data)+geom_boxplot(aes(x=y,y=cig))+theme_bw()+
  ggtitle('Cigrettes Distribution')+
  theme(axis.text=element_text(size=10),
          plot.title = element_text(size = 12, hjust=0.5, face = 'bold'),
          legend.position="none")

res_images <- function(x.var,df){
  ggplot(df,aes_string(x = x.var))+
  geom_bar(aes(y = ..count.., group=y), color = "white",
    position = position_dodge(0.9))+
  facet_wrap(~y)+theme_bw()+ggtitle(paste(x.var, 'Distribution'))+
  theme(axis.text=element_text(size=10),
          plot.title = element_text(size = 12, hjust=0.5, face = 'bold'),
          legend.position="none")
}

for (i in c(colnames(data)[c(2,4,5,6,7)])){
  print(res_images(i,data))
}
prd_images <- function(x.var,y.var,df){
  ggplot(df,aes_string(x = x.var, y =y.var))+geom_boxplot()+
    theme_bw()+ggtitle(paste(x.var, 'to', y.var))+
  theme(axis.text=element_text(size=10),
          plot.title = element_text(size = 12, hjust=0.5, face = 'bold'),
          legend.position="none")
}

# age
for (i in c(colnames(data)[c(2,4,5)])){
  print(prd_images(i,'age', data))
}
```

```r
# cig
for (i in c(colnames(data)[c(2,4,5)])){
  print(prd_images(i,'cig', data))
}
full.logit = glm(y ~ age + smoke + cig + hyper + myofam + strokefam + diabetes,
                 data = data, family = binomial)
summary(full.logit)[12]
predicted_y = predict(full.logit, newdata = data[,1:7], type = 'response')
plot(data.frame(predicted_y, data[,1:7]))
library(LogisticDx)
good.stuff = as.data.frame(dx(full.logit))
pear.r = good.stuff$Pr
std.r = good.stuff$sPr
df.beta = good.stuff$dBhat
change.pearson = good.stuff$dChisq
cutoff.std = 3.0
cutoff.beta = 0.5
cutoff.pearson = 10

hist(std.r, main='Pearson Standardized Residuals')
good.stuff[abs(std.r) > cutoff.std, c(1:(length(full.logit$coefficients)+1),
                                      which(names(good.stuff) == 'sPr'))]

plot(df.beta, main = 'Index plot of the change in the Betas')
good.stuff[df.beta > cutoff.beta, c(1:(length(full.logit$coefficients)+1),
                                    which(names(good.stuff) == 'dBhat'))]

plot(change.pearson, main = 'Index plot of the change in the Pearson test-statistic')
good.stuff[change.pearson > cutoff.pearson, c(1:(length(full.logit$coefficients)+1),
                                              which(names(good.stuff) == 'dChisq'))]
newdata = data[-c(158,166),]
library(bestglm)
best.subset.aic = bestglm(Xy = newdata, family = binomial(link=logit),
                          IC = "AIC",method = "exhaustive")
empty.model = glm(y~1, data = newdata, family = binomial(link = logit))
full.model = glm(y~., data = newdata, family = binomial(link = logit))
best.fb.aic = step(empty.model, scope = list(lower = empty.model, upper = full.model),
                   direction = 'both', criterion = 'AIC', trace = F)
best.bf.aic = step(full.model, scope = list(lower = empty.model, upper = full.model),
                   direction = 'both', criterion = 'AIC', trace = F)
best.bf.aic$formula
best.fb.aic$formula
summary(best.subset.aic$BestModel)[12]
# III. Hypothesis Test Based on the p value
```

```r
## a) X_2: smoke (can be dropped or not)
All.Criteria = function(the.model){
  p = length(the.model$coefficients)
  n = length(the.model$residuals)
  the.LL = logLik(the.model)
  the.BIC =  -2*the.LL + log(n)*p
  the.AIC =  -2*the.LL + 2*p
  the.results = c(the.LL,p,n,the.AIC,the.BIC)
  names(the.results) = c("LL","p","n","AIC","BIC")
  return(the.results)
}

Models = c("y~age+myofam+hyper+cig", "y~smoke+age+myofam+hyper+cig")
model.crit = t(sapply(Models,function(M){
  current.model = glm(M,data = newdata,family = binomial(link = logit))
  All.Criteria(current.model)
}))
model.crit
smoke_2 = as.factor(ifelse(newdata$smoke == 'ex' |
                             newdata$smoke == 'current','smoked','never'))
newdata['smoke'] = smoke_2
best.subset.aic2 = bestglm(Xy = newdata, family = binomial(link=logit),
                           IC = "AIC",method = "exhaustive")
summary(best.subset.aic2$BestModel)[12]
all.models = c('y~age+smoke+cig+hyper+myofam',
               'y~age+smoke+cig+hyper+myofam+age*smoke',
               'y~age+smoke+cig+hyper+myofam+age*cig',
               'y~age+smoke+cig+hyper+myofam+age*hyper',
               'y~age+smoke+cig+hyper+myofam+age*myofam',
               'y~age+smoke+cig+hyper+myofam+smoke*cig',
               'y~age+smoke+cig+hyper+myofam+smoke*hyper',
               'y~age+smoke+cig+hyper+myofam+smoke*myofam',
               'y~age+smoke+cig+hyper+myofam+cig*hyper',
               'y~age+smoke+cig+hyper+myofam+cig*myofam',
               'y~age+smoke+cig+hyper+myofam+hyper*myofam')
all.criteria = function(the.model){
  p = length(the.model$coefficients)
  n = length(the.model$residuals)
  the.ll = logLik(the.model)
  the.bic = -2*the.ll+log(n)*p
  the.aic = -2*the.ll+2*p
  the.results = c(the.ll, p, n, the.aic, the.bic)
  names(the.results) = c('LL','p','n','aic','bic')
  return(the.results)
```

```
}
all.model.crit = t(sapply(all.models, function(M){
  current.model = glm(M,data = newdata,
                      family = binomial(link = logit))
  all.criteria(current.model)
}))
m = round(all.model.crit,3)
m
for (i in 2:11){
  sta = -2*(m[1,1]-m[i,1])
  p_val = pchisq(sta, m[i,2]-m[1,2],lower.tail = F)
  cat(c('p-value for model:',p_val))
}
estimates2 = summary(best.subset.aic2$BestModel)$coefficients[,1]
se2 = summary(best.subset.aic2$BestModel)$coefficients[,2]
z_sta = qnorm(1-0.05/2)
upper.bounds2 = estimates2 + z_sta * se2
lower.bounds2 = estimates2 - z_sta * se2
wald.ci2 = cbind(lower.bounds2, upper.bounds2)
wald.ci2
confint(best.subset.aic2$BestModel)
truth = newdata$y
predicted = ifelse(fitted(best.subset.aic2$BestModel)>0.5, 1, 0)
my.table = table(truth, predicted)
sens = sum(predicted == 1 & truth == 1)/sum(truth == 1)
spec = sum(predicted == 0 & truth == 0)/sum(truth == 0)
error = sum(predicted != truth)/length(predicted)
results = c(sens, spec, error)
names(results) = c('Sensitivity','Specificity','Error-Rate')
my.table
results
library(pROC)
my.auc = auc(truth, fitted(best.subset.aic2$BestModel),
             plot = T, legacy.axes = T)
my.auc
auc.CI = ci(my.auc, level = 1-0.05)
auc.CI
pre = predict(best.subset.aic2$BestModel,
              newdata = data.frame(age = 50, smoke = 'never', cig = 0,
                                   hyper = 'mild', myofam = 'no'), type = 'response')
summary(best.subset.aic2$BestModel)[12]
```