Linear Statistical Analysis

Final Project (logistic Regression)

Name: Yukang Xu ID: 002462280

Ⅰ. Introduction

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

Ⅱ. Data Analysis

A. Model Selection

1. fit model by using all of variables *(Table: A.1)*
2. removing missing values
3. deleting insignificant variables
4. Using stepwise procedure with interactions and then removing the insignificant

After these four steps, we obtain the following model *(Table: A.2)*

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.328186   0.715449 -11.641  < 2e-16 ***
male             0.555279   0.109033   5.093 3.53e-07 ***
age              0.063515   0.006679   9.509  < 2e-16 ***
education       -0.047767   0.049395  -0.967  0.33353
currentSmoker    0.071601   0.156752   0.457  0.64783
cigsPerDay       0.017914   0.006238   2.872  0.00408 **
BPMeds           0.162496   0.234326   0.693  0.48802
prevalentStroke  0.693660   0.489569   1.417  0.15652
prevalentHyp     0.234208   0.138026   1.697  0.08973 .
diabetes         0.039167   0.315506   0.124  0.90120
totChol          0.002332   0.001127   2.070  0.03850 *
sysBP            0.015403   0.003808   4.044 5.24e-05 ***
diaBP           -0.004159   0.006438  -0.646  0.51831
BMI              0.006672   0.012758   0.523  0.60097
heartRate       -0.003246   0.004211  -0.771  0.44082
glucose          0.007127   0.002234   3.190  0.00142 **
---
```

(Table: A.1)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.834910   0.420465 -18.634  < 2e-16 ***
male         0.466471   0.101698   4.587 4.50e-06 ***
age          0.081085   0.005947  13.634  < 2e-16 ***
cigsPerDay   0.019626   0.004043   4.855 1.21e-06 ***
totChol      0.003319   0.001046   3.174   0.0015 **
glucose      0.008768   0.001642   5.340 9.30e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Table: A.2)

B. Interpreting the Model

Holding the other variables constant:

a. Having been investigated as a woman versus being a man, the log odds of ten-year CHD increase by 0.47.
b. Having been investigated as different age, the log odds of ten-year CHD increase by 0.081.
c. Having been investigated as usage of cigarettes, the log odds of ten-year CHD increase by 0.019.
d. Having been investigated as different totChol level, the log odds of ten-year CHD increase by 0.0033.
e. Having been investigated as different level of glucose , the log odds of ten-year CHD increase by 0.009.

For an easier interpretation, we can transform these values into odd's ratios:

```
 (Intercept)          male           age    cigsPerDay       totChol       glucose
0.0003956779 1.5943576161 1.0844628300 1.0198201987 1.0033245478 1.0088061058
```

Considering these estimates, we can say (while holding the other variables constant):

Having been investigated as a woman versus being a man, the odds of ten-year CHD increase by 1.59.

95% confidence intervals for the odds ratios are as follows:

```
                      OR       2.5 %       97.5 %
(Intercept) 0.0003956779 0.000171522 0.0008920894
male        1.5943576161 1.306463216 1.9466944201
age         1.0844628300 1.071988285 1.0972823881
cigsPerDay  1.0198201987 1.011743909 1.0279157098
totChol     1.0033245478 1.001260309 1.0053792545
glucose     1.0088061058 1.005606274 1.0121159075
```

As none of the intervals contain the value of one, we can see that there is a discrepancy between ten-year CHD between males and females. Most interesting, the odds of ten-year CHD for being a woman could be 1.95 times the odds of ten-year CHD of being a man!

C. Goodness of Fit (accuracy, Collinearity and Power)

The ANOVA table is created by adding the terms of the model sequentially.

```
          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      3816     3290.3
male       1   31.153      3815     3259.1 2.385e-08 ***
age        1  222.780      3814     3036.3 < 2.2e-16 ***
cigsPerDay 1   21.846      3813     3014.5 2.954e-06 ***
totChol    1   10.703      3812     3003.8  0.001069 **
glucose    1   29.542      3811     2974.3 5.471e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the residual deviance of the model decreases with each added predictor variable along with the fact that the p-values are significant, there is evidence that our fitted model is a good fit. Cooks distances for the data are created, yet none of them are significantly large. This indicates that there are no influential points.

We can also perform Wald Tests on each of the predictors to check and see if they are needed in the model.

```
> regTermTest(fit2,"totChol")
Wald test for totChol
 in glm(formula = TenYearCHD ~ male + age + cigsPerDay + totChol +
    glucose, family = "binomial", data = data)
F =  10.07717  on  1  and  3811  df: p= 0.0015131
> regTermTest(fit2,"glucose")
Wald test for glucose
 in glm(formula = TenYearCHD ~ male + age + cigsPerDay + totChol +
    glucose, family = "binomial", data = data)
F =  28.51355  on  1  and  3811  df: p= 9.8478e-08
> regTermTest(fit2,"male")
Wald test for male
 in glm(formula = TenYearCHD ~ male + age + cigsPerDay + totChol +
    glucose, family = "binomial", data = data)
F =  21.03905  on  1  and  3811  df: p= 4.6448e-06
> regTermTest(fit2,"age")
Wald test for age
 in glm(formula = TenYearCHD ~ male + age + cigsPerDay + totChol +
    glucose, family = "binomial", data = data)
F =  185.8864  on  1  and  3811  df: p= < 2.22e-16
>
```

Like the results before, these p-values indicate that each of the predictor variables are significant in predicting the odds that a customer will get CHD in ten years.

Lastly, we can use the Hosmer-Lemeshow Goodness of Fit Test to determine model adequacy.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit2$y, fitted(fit2)
X-squared = 12.318, df = 8, p-value = 0.1376
```

For the Hosmer-Lemeshow Test, significant p-values indicate that the model is not adequate for predicting ten-year CHD based on our variables. However, our p-value is .1376 so we can say that there is strong evidence that our model is a good fit.

After assessing the goodness of fit of the logistic model, we will check to see if there is any collinearity between the predictor variables. We will check this using variance inflation factors. If any are greater than 10, we will remove that variable from the model.

```
     male       age cigsPerDay    totChol    glucose
 1.173315   1.100699   1.228795   1.046793   1.008681
```

Since none of the VIF values are larger than 10, we can say that there is no collinearity between the predictor variables.

To assess the predictive power of the model, we use the McFadden R^2.

```
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.487125e+03 -1.645137e+03  3.160244e+02  9.604803e-02  7.945917e-02  1.375474e-01
```

A McFadden R^2 value between 0.2 and 0.4 is considered good. Therefore, since our McFadden R^2 is fairly small, we can say that the model selected is an excellent fit for predicting ten-year CHD.

D. Variable of Importance and Effect

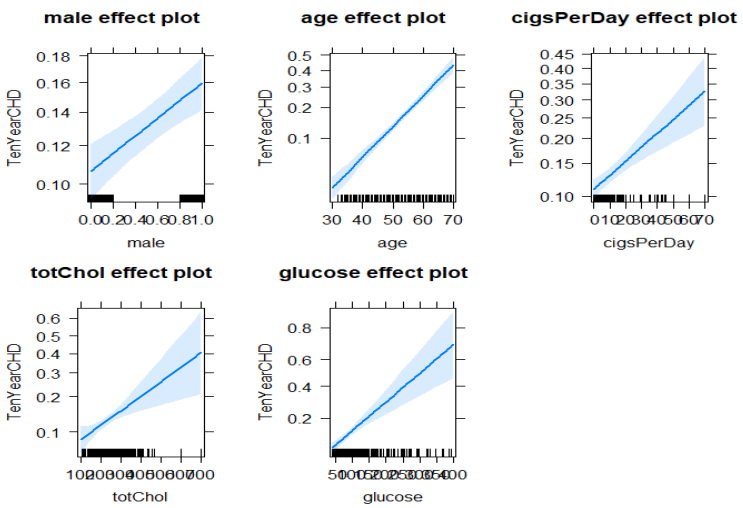
We can assess the importance of individual predictors in the model. Based on the all the sample:

```
glm variable importance

           Overall
age         100.00
glucose      20.70
cigsPerDay   16.06
male         13.50
totChol       0.00
```

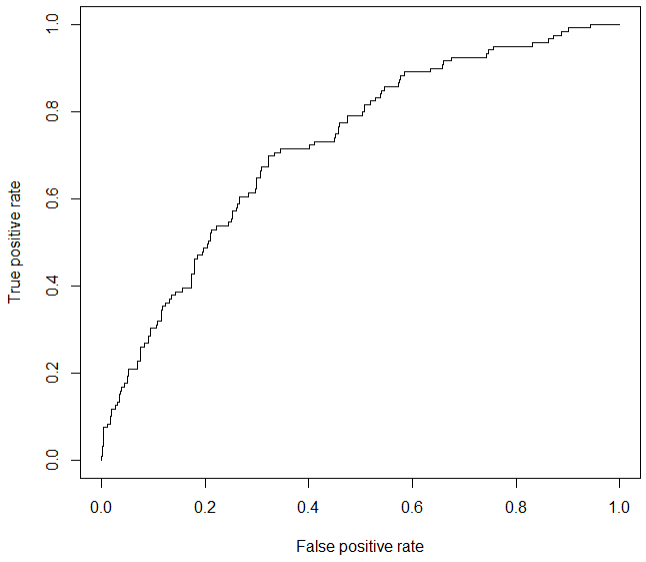
It appears that the age has the biggest impact on the probability of getting CHD for our clients

To determine the effect of the three individual predictor variables on the chances of getting CHD, let's make a plot to determine the effect each one has, individually, on ten-year CHD:



These plots are a little surprising. For example, if you consider being a male, your chances of get CHD in ten years is roughly 16% while it is 11% as a female.

E. Model Evaluation (Cross Validation and ROC Curve)



The area underneath this ROC curve is .7237. The curve is close to the left-hand border yet the top of the curve does not reach the y-value of 1 quickly. This indicates that the test is somewhat accurate. Since the area is .7237, the test does a good job of separating the customers who get CHD or not.

Using Cross Validation techniques on the model, we obtain the following results:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 643 116
         1   1   3

               Accuracy : 0.8467
                 95% CI : (0.8191, 0.8715)
    No Information Rate : 0.844
    P-Value [Acc > NIR] : 0.4449

                  Kappa : 0.039

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99845
            Specificity : 0.02521
         Pos Pred Value : 0.84717
         Neg Pred Value : 0.75000
             Prevalence : 0.84404
         Detection Rate : 0.84273
   Detection Prevalence : 0.99476
      Balanced Accuracy : 0.51183
```

III. Conclusion

The overall accuracy of the model to predict survival rate is .8467 with a sensitivity (the proportion who get CHD who were predicted to get CHD based on the model) is .998 yet the specificity (the proportion who will not get CHD who were predicted not to get CHD based on the model) was .025. This indicates that our model does a better job at correctly predicting the chances that someone get CHD than predicting the chances that someone will not get CHD.

R Code:

```
# Import data
data=read.csv("C:/Users/xuyuk/OneDrive - Georgia State University/Data import
/framingham.csv")

# Fit model
fit<-glm(TenYearCHD~.,data=data,family="binomial")
summary(fit)

# Feature selection
data=data[,-(3:4)]
data=data[,-(4:7)]
data=data[,-(5:8)]

# Remove missing value
data=data[complete.cases(data), ]
fit<-glm(TenYearCHD~.,data=data,family="binomial")
summary(fit)

# Model selection
full<-glm(TenYearCHD~male*age*cigsPerDay*totChol*glucose,data=data,family="bi
nomial")
null<-glm(TenYearCHD~1,data=data,family=binomial)
step(null,scope=list(lower=null,upper=full),direction="both")
fit1=glm(formula=TenYearCHD~age+cigsPerDay+glucose + male + totChol + age:tot
Chol + glucose:totChol, family = binomial, data = data)
summary(fit1)

# Remove insignificant interaction term
fit2=glm(TenYearCHD~male+age+cigsPerDay+totChol+glucose,data=data,family="bin
omial")
summary(fit2)

# Convert the coefficients to odds-ratios
exp(coef(fit2))

# Create a confidence interval of odds-ratios
exp(cbind(OR=coef(fit2),confint(fit2)))

# Anova Test to Determine Goodness of Fit
anova(fit2,test="Chisq")

# Cook's distance
cooks.distance<-cooks.distance(fit2)
which(cooks.distance>1)

# Wald Test to determine if predictors are significant
library(survey)
regTermTest(fit2,"male")
regTermTest(fit2,"age")
regTermTest(fit2,"CigsPerDay")
regTermTest(fit2,"totChol")
regTermTest(fit2,"glucose")

# Hoslem-Lemeshow Goodness of Fit Test
library(ResourceSelection)
hoslem.test(fit2$y,fitted(fit2),g=10)

# Looking at VIF for Collinearity
library(car)
vif(fit2)

# Determining the Pseudo-Rsq
```

```
library(pscl)
pR2(fit2)

# Plotting the effects of age, sex, and class to predict ten year CHD
library(effects)
plot(allEffects(fit2))

# Cross Validation to obtain accuracy of model
library(caret)
library(plyr)
ctrl<-trainControl(method="repeatedcv",number=10,savePredictions=TRUE)
mod_fit<-train(TenYearCHD~male+age+cigsPerDay+totChol+glucose,data=data,metho
d="glm",family="binomial",trControl=ctrl,tuneLength=5)
Train<-createDataPartition(data$TenYearCHD,p=0.8,list=FALSE)
training<-data[Train,]
testing<-data[-Train,]
y_testing=testing[,6]
x_testing=testing[,1:5]
prob <- predict(mod_fit, newdata=testing, type="raw")
results <- ifelse(prob > 0.5,1,0)
results=as.factor(results)
y_testing=as.factor(y_testing)
confusionMatrix(data=results,y_testing)

# Determining Variables of Importance
varImp(mod_fit)

# Graphing and finding the area underneath the ROC Curve:
library(ROCR)
p<-predict(fit2,newdata=subset(testing,select=c(1,2,3,4,5)),type="response")
pr<-prediction(p,testing$TenYearCHD)
prf<-performance(pr,measure="tpr",x.measure="fpr")
plot(prf)
auc<-performance(pr,measure="auc")
auc<-auc@y.values[[1]]
auc
```