LINEAR STATISTICAL ANALYSIS

Coronary Heart Disease (CHD) Prediction

(Logistic Regression)

Ruixie Fang

002422276

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 project intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression. We will construct a model using all of the predictor variables and using a stepwise procedure. We will determine the goodness of fit of that model using a chi-square procedure and then consider the McFadden Pseudo R^2 to estimate the predictive power. Next, I will plot the estimated effects of the predictor variables to determine what they individually contribute to the model. Lastly, we will measure the accuracy of the model create a ROC curve to compare the rates of false positive predictions with false negative predictions.

Maximum Likelihood Estimation

To obtain the model of interest, we need to find the values of the coefficients that solve, we have three variables in final model:

$$\max_{\beta_0,\beta_1,\beta_2,\beta_3} \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_i x_i}}{1+e^{\beta_0+\beta_i x_i}} \right)^{y_i} \left(\frac{1}{1+e^{\beta_0+\beta_i x_i}} \right)^{1-y_i}$$

We cannot solve this equation by hand. As a result, we use statistical software to obtain the values of the coefficients.

Model Selection

1. Using the all the predictor variables to determine the model, we come up with the following model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.159948   0.462477 -19.806  < 2e-16 ***
male         0.549104   0.103896   5.285 1.26e-07 ***
age          0.066848   0.006250  10.696  < 2e-16 ***
cigsPerDay   0.019902   0.004068   4.892 1.00e-06 ***
totChol      0.002456   0.001060   2.317   0.0205 *
sysBP        0.016960   0.002095   8.096 5.70e-16 ***
glucose      0.007614   0.001647   4.622 3.80e-06 ***
```

All of the predictor variables are significant in this model which tells us that there is evidence to suggest that all of the slopes are different from zero.

2. Using stepwise procedure with interactions and then removing the insignificant terms, we obtain the following model:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.921451   0.381097 -20.786  < 2e-16 ***
age          0.070450   0.006190  11.382  < 2e-16 ***
sysBP        0.017081   0.002042   8.366  < 2e-16 ***
cigsPerDay   0.026523   0.003796   6.988  2.8e-12 ***
```

Interpreting the Model

Holding the other variables constant:

a. For each unit the age grows, the log odds of having CHD increase by 0.07045.
b. For each unit the systolic blood pressure grows, the log odds of having CHD increase by 0.017081.
c. For the number of cigarettes that the person smoked on average in one day grows, the log odds of having CHD increase by 0.026523.

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

```
  (Intercept)           age         sysBP    cigsPerDay
0.0003628753 1.0729905076 1.0172280643 1.0268778769
```

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

a. For each unit the age grows, the odds of having CHD increase by 1.0729905076.
b. For each unit the systolic blood pressure grows, the odds of having CHD increase by 1.0172280643.
c. For the number of cigarettes that the person smoked on average in one day grows, the odds of having CHD increase by 1.0268778769.

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

```
                          OR        2.5 %       97.5 %
(Intercept) 0.0003628753 0.0001702499 0.0007587762
age         1.0729905076 1.0601246172 1.0861702229
sysBP       1.0172280643 1.0131741601 1.0213202152
cigsPerDay  1.0268778769 1.0192368164 1.0345258301
```

Since the confidence intervals are related to the p-values, and none of the confidence interval includes 0, we can conclude that all coefficients are statistically significant.

Goodness of Fit

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

```
           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        3814     3289.6
age         1  215.229      3813     3074.4 < 2.2e-16 ***
sysBP       1   65.431      3812     3008.9 6.020e-16 ***
cigsPerDay  1   46.959      3811     2962.0 7.249e-12 ***
---
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.

```
> #cook'sdistance
> cooks.distance=cooks.distance(fit3)
> which(cooks.distance>1)
named integer(0)
```

From the results of Cooks distances, none of them are significantly large. It indicates that there are no influential points.

Next, we perform Wald Tests on each of the predictors to check if they are needed in the model.

```
> library(survey)
> regTermTest(fit3,"age")
Wald test for age
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  129.5522  on  1  and  3811  df: p= < 2.22e-16
> regTermTest(fit3,"sysBP")
Wald test for sysBP
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  69.9826  on  1  and  3811  df: p= < 2.22e-16
> regTermTest(fit3,"cigsPerDay")
Wald test for cigsPerDay
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  48.82556  on  1  and  3811  df: p= 3.2876e-12
```

Like the results before, these p-values indicate that each of the predictor variables are significant in predicting the odds that a people have a CHD.

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

```
> library("ResourceSelection")
ResourceSelection 0.3-5        2019-07-22
> hoslem.test(fit3$y,fitted(fit3),g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit3$y, fitted(fit3)
X-squared = 6.1653, df = 8, p-value = 0.6287
```

For the Hosmer-Lemeshow Test, significant p-values indicate that the model is not adequate. However, our p-value is .6287, we can say that there is strong evidence that our model is a good fit.

Collinearity

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.

```
> library(car)
> vif(fit3)
       age      sysBP cigsPerDay
  1.193809   1.128126   1.089852
```

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

Power

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

```
> library(pscl)
> pR2(fit3)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.480992e+03 -1.644801e+03  3.276181e+02  9.959204e-02  8.229227e-02  1.424230e-01
```

A McFadden R^2 value between 0.2 and 0.4 is considered good. Although our McFadden R^2 is 9.959204e-02, after comparing with the results of other two models (shown below), we can conclude that this model is better than others. And it's a good fit for predicting CHD.

```
> pR2(fit1)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.454173e+03 -1.644801e+03  3.812561e+02  1.158974e-01  9.510473e-02  1.645974e-01
> pR2(fit2)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.449468e+03 -1.644801e+03  3.906646e+02  1.187574e-01  9.733362e-02  1.684550e-01
```

Cross Validation

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

```
Predicted    0.0540  0.332  0.275
cvpred       0.0506  0.328  0.264
 [ reached getOption("max.print") -- omitted 2 rows ]

Sum of squares = 50    Mean square = 0.13    n = 381

Overall (Sum over all 381 folds)
  ms
0.12
```

The value of 0.12 is low, and it represents a good accuracy result.

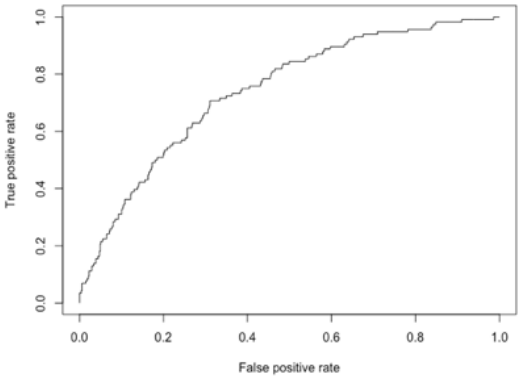
Variable of Importance

We can assess the importance of individual predictors in the model:

```
glm variable importance

            Overall
age           100.0
sysBP          31.4
cigsPerDay      0.0
```

It appears that the age has the biggest impact on the probability of having CHD. It isn't surprising to see that the overall importance of the number of cigarettes that the person smoked on average in one day is 0, because the p -value of third factor is the largest one among all the variables.

ROC Curve



The area underneath this ROC curve is .777. 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 .777, the test does a good job of separating the people with risk of CHD from all samples and making predictions using the chosen model.

Conclusions

- Age
- systolic blood pressure
- the number of cigarettes that the person smoked on average in one day)

Based on this model, we can pinpoint these three as high-risk factors to help high-risk patients make decisions on lifestyle changes and also help other people to live in a healthy way, decreasing/avoiding the potential risk of having CHD.

R-Code

```
> #read data and drop missing values
> heart<-read.csv("/Users/balloon_n/Documents/F/study/class/Linear
Stat/project/framingham1.csv")
> hrt <- na.omit(heart)
```

Full Model

```
> fit1<-glm(TenYearCHD~.,data=hrt,family="binomial")
> summary(fit1)

Call:
glm(formula = TenYearCHD ~ ., family = "binomial", data = hrt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0405  -0.6009  -0.4340  -0.2854   2.8770

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.159948   0.462477 -19.806  < 2e-16 ***
male         0.549104   0.103896   5.285 1.26e-07 ***
age          0.066848   0.006250  10.696  < 2e-16 ***
cigsPerDay   0.019902   0.004068   4.892 1.00e-06 ***
totChol      0.002456   0.001060   2.317   0.0205 *
sysBP        0.016960   0.002095   8.096 5.70e-16 ***
glucose      0.007614   0.001647   4.622 3.80e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3289.6  on 3814  degrees of freedom
Residual deviance: 2908.3  on 3808  degrees of freedom
AIC: 2922.3

Number of Fisher Scoring iterations: 5
```

Stepwise Model Selection (including interaction terms)

```
>full=glm(TenYearCHD~male*age*cigsPerDay*totChol*sysBP*glucose,data=hrt,family="binomial"
)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> null=glm(TenYearCHD~1,data=hrt,family="binomial")
> step(null,scope=list(lower=null,upper=full),direction="both")
Start:  AIC=3291.6
TenYearCHD ~ 1

             Df Deviance    AIC
+ age         1   3074.4 3078.4
+ sysBP       1   3119.4 3123.4
+ glucose     1   3242.3 3246.3
+ totChol     1   3258.1 3262.1
+ male        1   3258.4 3262.4
+ cigsPerDay  1   3279.3 3283.3
<none>            3289.6 3291.6

Step:  AIC=3078.37
```

```
TenYearCHD ~ age

             Df Deviance    AIC
+ sysBP       1   3008.9 3014.9
+ cigsPerDay  1   3032.2 3038.2
+ male        1   3036.1 3042.1
+ glucose     1   3046.3 3052.3
+ totChol     1   3068.5 3074.5
<none>            3074.4 3078.4
- age         1   3289.6 3291.6

Step:  AIC=3014.94
TenYearCHD ~ age + sysBP

             Df Deviance    AIC
+ male        1   2957.7 2965.7
+ cigsPerDay  1   2962.0 2970.0
+ glucose     1   2988.5 2996.5
+ age:sysBP   1   3006.5 3014.5
+ totChol     1   3006.7 3014.7
<none>            3008.9 3014.9
- sysBP       1   3074.4 3078.4
- age         1   3119.4 3123.4

Step:  AIC=2965.69
TenYearCHD ~ age + sysBP + male

             Df Deviance    AIC
+ cigsPerDay  1   2935.5 2945.5
+ glucose     1   2937.8 2947.8
+ totChol     1   2951.5 2961.5
+ male:sysBP  1   2955.4 2965.4
<none>            2957.7 2965.7
+ age:sysBP   1   2957.0 2967.0
+ male:age    1   2957.6 2967.6
- male        1   3008.9 3014.9
- sysBP       1   3036.1 3042.1
- age         1   3074.2 3080.2

Step:  AIC=2945.46
TenYearCHD ~ age + sysBP + male + cigsPerDay

                   Df Deviance    AIC
+ glucose           1   2913.6 2925.6
+ totChol           1   2930.0 2942.0
+ male:sysBP        1   2933.1 2945.1
<none>                  2935.5 2945.5
+ age:sysBP         1   2934.8 2946.8
+ male:cigsPerDay   1   2935.1 2947.1
+ age:cigsPerDay    1   2935.3 2947.3
+ cigsPerDay:sysBP  1   2935.4 2947.4
+ male:age          1   2935.4 2947.4
- cigsPerDay        1   2957.7 2965.7
- male              1   2962.0 2970.0
- sysBP             1   3014.2 3022.2
- age               1   3067.6 3075.6

Step:  AIC=2925.64
TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose
```

```
                     Df Deviance    AIC
+ totChol             1   2908.3 2922.3
+ male:sysBP          1   2910.8 2924.8
<none>                    2913.6 2925.6
+ age:sysBP           1   2913.0 2927.0
+ male:glucose        1   2913.2 2927.2
+ male:cigsPerDay     1   2913.3 2927.3
+ cigsPerDay:glucose  1   2913.3 2927.3
+ age:cigsPerDay      1   2913.5 2927.5
+ cigsPerDay:sysBP    1   2913.5 2927.5
+ male:age            1   2913.6 2927.6
+ sysBP:glucose       1   2913.6 2927.6
+ age:glucose         1   2913.6 2927.6
- glucose             1   2935.5 2945.5
- cigsPerDay          1   2937.8 2947.8
- male                1   2939.0 2949.0
- sysBP               1   2983.9 2993.9
- age                 1   3040.3 3050.3

Step:  AIC=2922.35
TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose + totChol

                     Df Deviance    AIC
+ totChol:glucose     1   2904.8 2920.8
+ male:sysBP          1   2905.3 2921.3
+ age:totChol         1   2905.6 2921.6
<none>                    2908.3 2922.3
+ male:totChol        1   2906.4 2922.4
+ male:cigsPerDay     1   2907.8 2923.8
+ totChol:sysBP       1   2907.9 2923.9
+ male:glucose        1   2907.9 2923.9
+ age:sysBP           1   2908.0 2924.0
+ cigsPerDay:glucose  1   2908.0 2924.0
+ age:cigsPerDay      1   2908.1 2924.1
+ cigsPerDay:sysBP    1   2908.2 2924.2
+ cigsPerDay:totChol  1   2908.2 2924.2
+ sysBP:glucose       1   2908.3 2924.3
+ male:age            1   2908.3 2924.3
+ age:glucose         1   2908.3 2924.3
- totChol             1   2913.6 2925.6
- glucose             1   2930.0 2942.0
- cigsPerDay          1   2931.8 2943.8
- male                1   2936.4 2948.4
- sysBP               1   2974.0 2986.0
- age                 1   3027.5 3039.5

Step:  AIC=2920.79
TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose + totChol +
    glucose:totChol

                     Df Deviance    AIC
+ male:sysBP          1   2901.5 2919.5
+ age:totChol         1   2901.9 2919.9
+ male:totChol        1   2902.6 2920.6
<none>                    2904.8 2920.8
+ totChol:sysBP       1   2903.8 2921.8
+ male:cigsPerDay     1   2904.2 2922.2
+ male:glucose        1   2904.3 2922.3
+ cigsPerDay:glucose  1   2904.3 2922.3
- glucose:totChol     1   2908.3 2922.3
```

```
+ age:sysBP           1   2904.3 2922.3
+ age:cigsPerDay      1   2904.5 2922.5
+ cigsPerDay:totChol  1   2904.5 2922.5
+ cigsPerDay:sysBP    1   2904.7 2922.7
+ sysBP:glucose       1   2904.7 2922.7
+ age:glucose         1   2904.7 2922.7
+ male:age            1   2904.8 2922.8
- cigsPerDay          1   2928.4 2942.4
- male                1   2932.5 2946.5
- sysBP               1   2970.9 2984.9
- age                 1   3024.6 3038.6

Step:  AIC=2919.48
TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose + totChol +
    glucose:totChol + sysBP:male

                     Df Deviance    AIC
+ age:totChol         1   2898.9 2918.9
<none>                    2901.5 2919.5
+ male:totChol        1   2899.8 2919.8
- sysBP:male          1   2904.8 2920.8
+ totChol:sysBP       1   2901.0 2921.0
+ cigsPerDay:glucose  1   2901.1 2921.1
+ male:cigsPerDay     1   2901.1 2921.1
+ male:glucose        1   2901.2 2921.2
+ age:sysBP           1   2901.3 2921.3
+ male:age            1   2901.3 2921.3
+ cigsPerDay:totChol  1   2901.3 2921.3
- glucose:totChol     1   2905.3 2921.3
+ age:cigsPerDay      1   2901.3 2921.3
+ sysBP:glucose       1   2901.4 2921.4
+ cigsPerDay:sysBP    1   2901.4 2921.4
+ age:glucose         1   2901.4 2921.4
- cigsPerDay          1   2925.2 2941.2
- age                 1   3023.0 3039.0

Step:  AIC=2918.94
TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose + totChol +
    glucose:totChol + sysBP:male + age:totChol

                     Df Deviance    AIC
<none>                    2898.9 2918.9
- age:totChol         1   2901.5 2919.5
- sysBP:male          1   2901.9 2919.9
+ male:totChol        1   2897.9 2919.9
+ male:age            1   2898.6 2920.6
+ cigsPerDay:glucose  1   2898.6 2920.6
+ male:glucose        1   2898.6 2920.6
+ male:cigsPerDay     1   2898.7 2920.7
+ age:cigsPerDay      1   2898.8 2920.8
+ sysBP:glucose       1   2898.8 2920.8
+ cigsPerDay:sysBP    1   2898.9 2920.9
+ age:sysBP           1   2898.9 2920.9
+ totChol:sysBP       1   2898.9 2920.9
+ age:glucose         1   2898.9 2920.9
- glucose:totChol     1   2902.9 2920.9
+ cigsPerDay:totChol  1   2898.9 2920.9
- cigsPerDay          1   2922.2 2940.2

Call:  glm(formula = TenYearCHD ~ age + sysBP + male + cigsPerDay +
```

```
    glucose + totChol + glucose:totChol + sysBP:male + age:totChol,
    family = "binomial", data = hrt)

Coefficients:
    (Intercept)              age            sysBP             male       cigsPerDay
     -9.8747285        0.1182846        0.0140338       -0.4661567        0.0199314
        glucose          totChol  glucose:totChol       sysBP:male      age:totChol
     -0.0103276        0.0072706        0.0000737        0.0070279       -0.0002120

Degrees of Freedom: 3814 Total (i.e. Null);  3805 Residual
Null Deviance:     3290
Residual Deviance: 2899       AIC: 2919
```

Review model significant

```
> fit2=glm(TenYearCHD ~ age + sysBP + male + cigsPerDay + glucose + totChol +
glucose:totChol + sysBP:male + age:totChol,
+          family = "binomial", data = hrt)
> summary(fit2)

Call:
glm(formula = TenYearCHD ~ age + sysBP + male + cigsPerDay +
    glucose + totChol + glucose:totChol + sysBP:male + age:totChol,
    family = "binomial", data = hrt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7127  -0.5999  -0.4306  -0.2856   2.8940

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.875e+00  1.914e+00  -5.160 2.47e-07 ***
age              1.183e-01  3.239e-02   3.652  0.00026 ***
sysBP            1.403e-02  2.700e-03   5.198 2.01e-07 ***
male            -4.662e-01  5.793e-01  -0.805  0.42098
cigsPerDay       1.993e-02  4.090e-03   4.873 1.10e-06 ***
glucose         -1.033e-02  9.742e-03  -1.060  0.28907
totChol          7.271e-03  7.834e-03   0.928  0.35337
glucose:totChol  7.370e-05  3.975e-05   1.854  0.06374 .
sysBP:male       7.028e-03  4.086e-03   1.720  0.08546 .
age:totChol     -2.120e-04  1.331e-04  -1.593  0.11117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3289.6  on 3814  degrees of freedom
Residual deviance: 2898.9  on 3805  degrees of freedom
AIC: 2918.9

Number of Fisher Scoring iterations: 5
```

Remove the insignificant interaction term

```
> fit3=glm(TenYearCHD ~ age + sysBP + cigsPerDay,family = "binomial", data = hrt)
> summary(fit3)

Call:
glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4437  -0.6077  -0.4395  -0.3126   2.7238

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.921451   0.381097 -20.786  < 2e-16 ***
age          0.070450   0.006190  11.382  < 2e-16 ***
sysBP        0.017081   0.002042   8.366  < 2e-16 ***
cigsPerDay   0.026523   0.003796   6.988  2.8e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3289.6  on 3814  degrees of freedom
Residual deviance: 2962.0  on 3811  degrees of freedom
AIC: 2970

Number of Fisher Scoring iterations: 5
```

Interpretation of Model

```
> #To convert the coefficients to odds-ratios:
> exp(coef(fit3))
 (Intercept)          age        sysBP   cigsPerDay
0.0003628753 1.0729905076 1.0172280643 1.0268778769
> #odds ratio CI
> exp(cbind(OR =coef(fit3),confint(fit3)))
Waiting for profiling to be done...
                      OR        2.5 %       97.5 %
(Intercept) 0.0003628753 0.0001702499 0.0007587762
age         1.0729905076 1.0601246172 1.0861702229
sysBP       1.0172280643 1.0131741601 1.0213202152
cigsPerDay  1.0268778769 1.0192368164 1.0345258301
```

Goodness of Fit

```
> #ANOVA test to determine GOF
> anova(fit3,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: TenYearCHD

Terms added sequentially (first to last)


           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        3814     3289.6
age         1  215.229      3813     3074.4 < 2.2e-16 ***
sysBP       1   65.431      3812     3008.9 6.020e-16 ***
cigsPerDay  1   46.959      3811     2962.0 7.249e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #cook'sdistance
> cooks.distance=cooks.distance(fit3)
```

```
> which(cooks.distance>1)
named integer(0)
```

Wald Test to determine if predictors are significant

```
> library(survey)
> regTermTest(fit3,"age")
Wald test for age
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  129.5522  on  1  and  3811  df: p= < 2.22e-16
> regTermTest(fit3,"sysBP")
Wald test for sysBP
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  69.9826  on  1  and  3811  df: p= < 2.22e-16
> regTermTest(fit3,"cigsPerDay")
Wald test for cigsPerDay
 in glm(formula = TenYearCHD ~ age + sysBP + cigsPerDay, family = "binomial",
    data = hrt)
F =  48.82556  on  1  and  3811  df: p= 3.2876e-12
```

Hosmer-Lemeshow Goodness of Fit Test

```
> library("ResourceSelection")
ResourceSelection 0.3-5        2019-07-22
> hoslem.test(fit3$y,fitted(fit3),g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit3$y, fitted(fit3)
X-squared = 6.1653, df = 8, p-value = 0.6287
```

Collinearity

```
> library(car)
> vif(fit3)
       age      sysBP cigsPerDay
  1.193809   1.128126   1.089852
```

Power (Determining the Pseudo-Rsq)

```
> library(pscl)
> pR2(fit3)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.480992e+03 -1.644801e+03  3.276181e+02  9.959204e-02  8.229227e-02  1.424230e-01
> pR2(fit1)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.454173e+03 -1.644801e+03  3.812561e+02  1.158974e-01  9.510473e-02  1.645974e-01
> pR2(fit2)
          llh       llhNull            G2      McFadden          r2ML          r2CU
-1.449468e+03 -1.644801e+03  3.906646e+02  1.187574e-01  9.733362e-02  1.684550e-01
```

Cross Validation

```
> library("DAAG")
> form=lm(TenYearCHD ~ age + sysBP + cigsPerDay,data=hrt)
> Cv.fit3<-CVlm(data=hrt,form.lm=form,m=10)
[ … ]
```

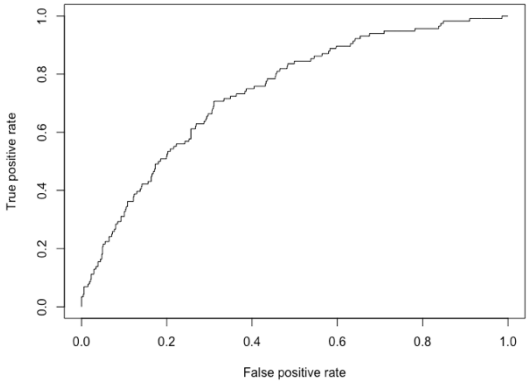
```
fold 10
Observations in test set: 381
                3     25    34       36     58     80     83    101   116
Predicted    0.164  0.188 0.240  0.00184  0.155  0.242  0.168  0.168 0.115
cvpred       0.163  0.186 0.241 -0.00335  0.151  0.237  0.166  0.167 0.114
                122    123     129   139    145   149    158       172    176
Predicted    0.0539  0.202  0.0750 0.205  0.197 0.221  0.165  0.003741  0.357
cvpred       0.0536  0.208  0.0743 0.199  0.196 0.224  0.160 -0.000772  0.356
                179     182    226   228    238     253    257     287     292
Predicted    0.0165  0.0838  0.178 0.293  0.149 -0.0210  0.104 -0.0394  0.0710
cvpred       0.0134  0.0805  0.175 0.286  0.149 -0.0233  0.092 -0.0414  0.0704
                293    306    338   341    381     404     417    418      421
Predicted    0.207 0.0409  0.190 0.260  0.146  0.0132  0.0973  0.127 -0.01149
cvpred       0.209 0.0380  0.187 0.256  0.147  0.0106  0.0983  0.121 -0.00878
                436    465     501   524    539    547   554     563    566
Predicted    0.282  0.179 0.00392 0.181  0.198  0.144 0.261  0.0272  0.151
cvpred       0.276  0.173 0.00368 0.181  0.202  0.145 0.253  0.0221  0.152
                569     595      597    601    606   611     624     637     642
Predicted   -0.0647  0.0310  0.00247  0.103  0.305 0.275  0.0691  0.0311  0.0429
cvpred      -0.0684  0.0268 -0.00197  0.104  0.306 0.272  0.0710  0.0301  0.0397
                664     669    671    688     702    703    724    734    744
Predicted   0.355  0.0262  0.267  0.118 -0.0102  0.174  0.187  0.239  0.121
cvpred      0.352  0.0244  0.268  0.117 -0.0123  0.173  0.182  0.243  0.118
                749     750    751   788    796   805      821    867   868
Predicted    0.0887 -0.0191  0.420 0.115  0.364 0.103  0.01326  0.241 0.226
cvpred       0.0824 -0.0191  0.414 0.114  0.359 0.100  0.00739  0.247 0.231
                880     882   887   900    913    933    935   949     951     963
Predicted    0.108  0.0564 0.355 0.293  0.313  0.380  0.284 0.409  0.0906 -0.0704
cvpred       0.106  0.0559 0.350 0.293  0.313  0.371  0.288 0.400  0.0881 -0.0719
                965    968    969      972   1011   1032   1040   1054    1060
Predicted    0.0433  0.193  0.174  0.01262  0.106  0.195  0.113  0.261 -0.0229
cvpred       0.0413  0.196  0.172  0.00758  0.107  0.197  0.113  0.254 -0.0243
                1072  1079   1082   1095    1103  1111   1138   1167  1178  1194
Predicted   0.231  0.16  0.235  0.165  0.0965 0.244  0.151  0.198 0.228 0.231
cvpred      0.235  0.16  0.234  0.163  0.0929 0.242  0.148  0.196 0.223 0.233
                1195   1205    1224    1243   1255   1259    1268   1279     1282
Predicted    0.359  0.110  0.0758  0.0538  0.126  0.244  0.0849  0.221 -0.00704
cvpred       0.352  0.106  0.0732  0.0519  0.125  0.245  0.0819  0.223 -0.00855
                1306   1309   1325   1335  1341   1343  1359   1426   1428  1431
Predicted    0.212  0.347  0.221  0.132 0.115  0.238 0.213  0.116  0.214 0.177
cvpred       0.214  0.344  0.223  0.130 0.119  0.240 0.217  0.116  0.214 0.178
                1467   1475   1487   1491   1495   1526    1548   1558   1560  1561
Predicted    0.149  0.164  0.184  0.305  0.181  0.198  0.0773  0.299  0.213  0.29
cvpred       0.152  0.157  0.181  0.307  0.182  0.187  0.0716  0.293  0.213  0.28
                1569    1585  1590  1609   1621  1625   1633   1668   1681    1694
Predicted    0.163  0.0599 0.295 0.257  0.325 0.321  0.136  0.220  0.256  0.0754
cvpred       0.165  0.0591 0.283 0.244  0.315 0.323  0.137  0.218  0.256  0.0706
                1700    1726    1754  1761    1767   1781   1819    1828    1830
Predicted    0.207  0.0887  0.0305 0.280  0.0145  0.141  0.263  0.0183  0.0519
cvpred       0.204  0.0903  0.0297 0.272  0.0102  0.139  0.266  0.0137  0.0509
                1835   1848   1859   1869   1872   1895   1898   1908  1923   1924
Predicted    0.118  0.197  0.179  0.145  0.185  0.113  0.195  0.172 0.288 0.0453
cvpred       0.122  0.200  0.181  0.146  0.185  0.115  0.200  0.171 0.289 0.0430
                1926     1930   1934  1940    1945   1972   1977    1989   1991
Predicted    0.214 -0.00514  0.104 0.269  0.0513  0.218  0.166  0.0691  0.103
cvpred       0.221 -0.00755  0.104 0.269  0.0448  0.209  0.165  0.0685  0.100
                1994    2006  2026    2041   2048    2055  2064    2074    2123
Predicted    0.125 -0.0156 0.295  0.0469  0.107  0.0817 0.152  0.0446 -0.0406
cvpred       0.122 -0.0165 0.288  0.0493  0.110  0.0782 0.153  0.0425 -0.0425
                2133    2136  2137     2140     2153   2164   2172    2177  2180
Predicted    0.397  0.0451 0.312 -0.00541  0.01151  0.184  0.191  0.0989 0.417
cvpred       0.387  0.0410 0.304 -0.00476  0.00977  0.187  0.195  0.1014 0.406
                2192  2207   2211   2227   2253    2261   2283    2284    2289
Predicted    0.165 0.268  0.139  0.229  0.152  0.0216  0.273  0.0672  0.0817
cvpred       0.168 0.259  0.138  0.228  0.147  0.0229  0.265  0.0679  0.0829
                2306    2330    2364     2367   2380   2383   2386    2398   2404
Predicted    0.338  0.0840  0.0298  0.00184  0.147  0.172  0.214  0.0722  0.285
cvpred       0.337  0.0818  0.0283 -0.00177  0.137  0.169  0.213  0.0669  0.286
                2407  2408   2434    2436   2439  2442   2459  2470   2473  2476
Predicted    0.222 0.162  0.198  0.0650  0.203 0.312  0.192 0.183  0.218 0.255
cvpred       0.224 0.162  0.199  0.0607  0.203 0.309  0.193 0.181  0.216 0.250
                2478    2490   2491   2494   2500   2502   2519    2534   2541
Predicted    0.0770  0.0481  0.273  0.131 0.0998  0.243  0.283  0.0748  0.177
cvpred       0.0744  0.0457  0.273  0.125 0.1000  0.242  0.282  0.0740  0.181
                2544   2550    2561    2564   2566   2596   2600   2608   2612
Predicted    0.142  0.330  0.0808  0.0373  0.207  0.302  0.193  0.188  0.173
cvpred       0.137  0.327  0.0785  0.0348  0.201  0.300  0.192  0.188  0.173
                2618  2626   2632   2633   2634  2638   2653   2660   2663    2667
Predicted    0.164 0.262  0.252  0.076 0.0437 0.237  0.165  0.175  0.186 -0.0138
cvpred       0.160 0.260  0.253  0.072 0.0408 0.242  0.163  0.171  0.181 -0.0130
                2693   2737    2745   2751    2752   2759   2768  2794  2802
Predicted    0.0333  0.204  0.0808 0.0786  0.0707  0.176  0.198 0.208 0.271
cvpred       0.0309  0.202  0.0812 0.0744  0.0684  0.174  0.204 0.204 0.259
                2803   2824   2830   2836   2844    2875   2892    2914   2922
Predicted    0.308  0.135  0.123  0.185  0.167  0.1005  0.116  0.0316  0.318
cvpred       0.305  0.136  0.120  0.185  0.164  0.0989  0.114  0.0286  0.320
                2936   2938   2943   2958   2967   2969    2976    2977   2996
Predicted    0.327  0.149  0.129  0.140  0.219  0.164  0.0777  0.0612  0.212
```

```
cvpred      0.320  0.146  0.132  0.137  0.215  0.168  0.0742  0.0571  0.214
             3013   3024   3033   3059   3070   3110    3121   3134    3139
Predicted    0.321  0.157  0.133  0.173  0.271  0.0349  0.0958  0.246  0.0616
cvpred       0.318  0.155  0.129  0.167  0.272  0.0309  0.0963  0.248  0.0609
             3149   3151   3161   3217   3225   3238   3244    3250  3255
Predicted   0.205  0.176  0.228  0.111  0.0500  0.0910  0.120  0.0348 0.176
cvpred      0.199  0.171  0.223  0.101  0.0467  0.0907  0.119  0.0293 0.180
               3260   3266   3268   3301   3315   3322   3329    3337  3339
Predicted   -0.0400  0.117  0.150  0.197  0.0830  0.223  0.109  0.0659 0.227
cvpred      -0.0427  0.119  0.147  0.196  0.0809  0.222  0.106  0.0672 0.224
              3345   3350  3372   3390   3402   3404   3424    3452    3474
Predicted    0.238  0.114 0.156  0.136  0.243  0.168  0.193  0.0754  0.1020
cvpred       0.235  0.114 0.152  0.139  0.243  0.168  0.188  0.0769  0.0957
               3496  3518   3563   3565   3566   3578   3590    3597    3627
Predicted    0.0621 0.340  0.138 0.0202  0.176  0.227  0.402  0.0693  0.0443
cvpred       0.0631 0.335  0.142 0.0163  0.170  0.222  0.397  0.0666  0.0374
             3628   3632   3635    3638    3666   3675  3681   3683   3688
Predicted   0.371  0.179  0.215  0.0919  0.0741  0.371 0.237  0.209  0.107
cvpred      0.373  0.172  0.208  0.0877  0.0742  0.369 0.231  0.212  0.103
               3692    3693    3699   3712   3719   3724   3725   3734   3736
Predicted    0.0380  0.0824  0.0437  0.152  0.109  0.233  0.120  0.134  0.169
cvpred       0.0362  0.0827  0.0360  0.148  0.107  0.230  0.111  0.139  0.168
              3740   3757    3763   3764    3773    3777   3779    3786  3811
Predicted    0.305  0.209  0.0497  0.294 -0.00686 -0.0273  0.185 -0.0140 0.143
cvpred       0.304  0.203  0.0473  0.290 -0.00725 -0.0308  0.182 -0.0175 0.141
             3834    3842   3851  3867  3869   3875   3883    3922   3934   3956
Predicted   0.264  0.0912  0.168 0.151 0.139 0.0211  0.301  0.0849  0.105  0.248
cvpred      0.268  0.0927  0.162 0.146 0.138 0.0197  0.308  0.0851  0.101  0.246
             3963  3974   3976    3986    3993    3996   3998   4012    4013
Predicted   0.333 0.262  0.192 -0.0672  0.0395  0.0662  0.144  0.135 -0.0145
cvpred      0.332 0.262  0.192 -0.0691  0.0409  0.0644  0.141  0.135 -0.0139
              4021   4024    4028    4032   4035    4039    4048    4078    4084
Predicted    0.106  0.221  0.0405  0.0565  0.104 -0.0362  0.0491  0.0272  0.0519
cvpred       0.109  0.217  0.0323  0.0518  0.101 -0.0376  0.0466  0.0253  0.0509
              4100   4115    4119   4121   4140   4163    4170   4173  4194
Predicted    0.193  0.209  0.0335  0.160  0.129  0.121  0.0142  0.251 0.395
cvpred       0.196  0.213  0.0265  0.159  0.129  0.119  0.0145  0.248 0.387
               4196   4198   4229
Predicted    0.0540  0.332  0.275
cvpred       0.0506  0.328  0.264
 [ reached getOption("max.print") -- omitted 2 rows ]

Sum of squares = 50    Mean square = 0.13    n = 381
```

```
Overall (Sum over all 381 folds)
  ms
0.12
```

Variable of Importance

```
> fit4=train(factor(TenYearCHD) ~ age + sysBP +
cigsPerDay,data=hrt,method="glm",family="binomial")
> varImp(fit4)
glm variable importance

           Overall
age          100.0
sysBP         31.4
cigsPerDay     0.0
```



ROC Curve

```
> library(ROCR)
> Train=createDataPartition(hrt$TenYearCHD,p=0.8,list=FALSE)
> training=hrt[Train,]
> testing=hrt[-Train,]
> p=predict(fit3,newdata=subset(testing,select=c(2,3,5)),type="response")
> pr=prediction(p,testing$TenYearCHD)
> prf=performance(pr,measure="tpr",x.measure="fpr")
> plot(prf)
> auc=performance(pr,measure="auc")
> auc=auc@y.values[[1]]
> auc
[1] 0.777
```