Victor Agboli

STAT 8561

Logistic Regression

Predicting the Odds of developing Coronary Heart Disease (CHD) in the next ten years.

Introduction

The heart is a fist-sized organ that pumps blood throughout the human body. It's the primary organ of the human circulatory system. It pumps blood in the human body for 70 beats per minute, and it does so using arteries. When these arteries are blocked, it poses a serious problem. Coronary heart disease (CHD) is a type of heart disease where the arteries of the heart cannot deliver enough oxygen-rich blood to the heart. It is the leading cause of death in the United States. About 18.2 million American adults have coronary artery disease, making it the most common type of heart disease in the United States, according to the Centers for Disease Control and Prevention [1].

In this project, I want to determine the factors that cause a patient to develop coronary heart disease and predict the odds of patients developing coronary heart disease in ten years using the logistic regression model. I will construct the model using the predictor variables which are; gender, age, prevalent stroke, prevalent hypertension, diabetes, and current smoker, and select the best using the stepwise procedure. I will also determine the model goodness of fit using the chi-square goodness of fit procedure, Cook's distance for influential points, and check the predictive power of the model using the McFadden Pseudo R^2. Next, I will plot the estimated effects of the predictor variables to determine what they individually contribute to the model. Lastly, I will measure the accuracy of the model concerning unknown variables and finally create a ROC curve to compare the rates of false positive predictions with false negative predictions and give recommendations to avoid the risk of developing CHD.

Logistic Function

The logistic function is given as:

$$p = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^{6} \beta_i x_i}}$$

The log of odds is given as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{6} \beta_i x_i$$

Maximum Likelihood Estimation

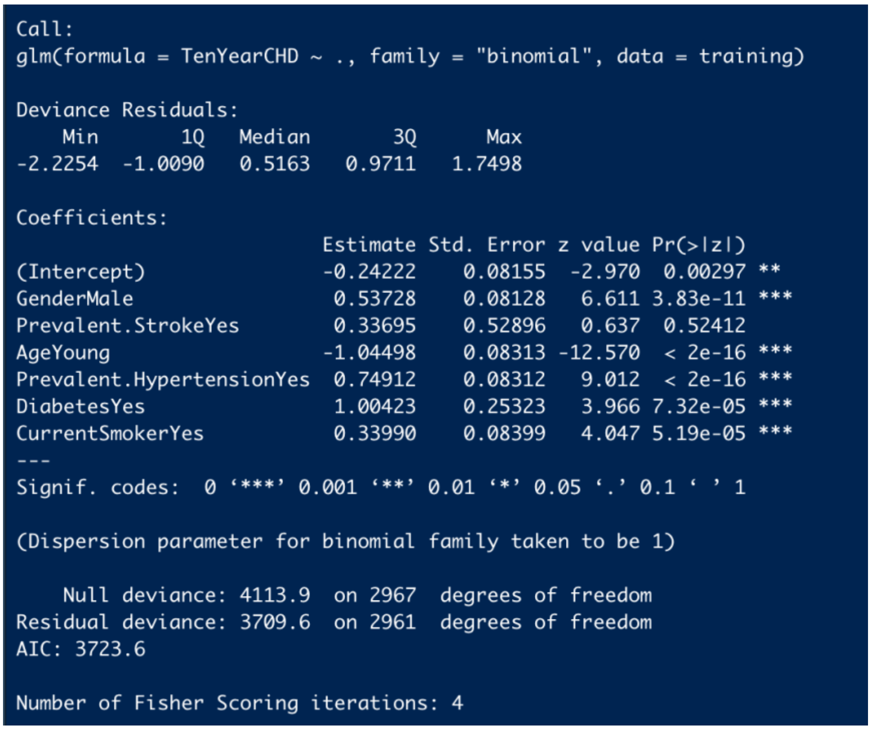
To obtain the model of interest, I need to find the values of the coefficients that solve:

$$\max_{\beta_0, \beta_{1,}\beta_2, \beta_3, \beta_4, \beta_5, \beta_6} \prod_{i=1}^{n} (\frac{e^{\beta_0 + \beta_i x_i}}{1 + e^{\beta_0 + \beta_i x_i}})^{y_i} (\frac{1}{1 + e^{\beta_0 + \beta_i x_i}})^{1-y_i}$$

But this equation can't be solved by hand. So, I will use the R statistical software to obtain the values of the coefficients.

Model Selection

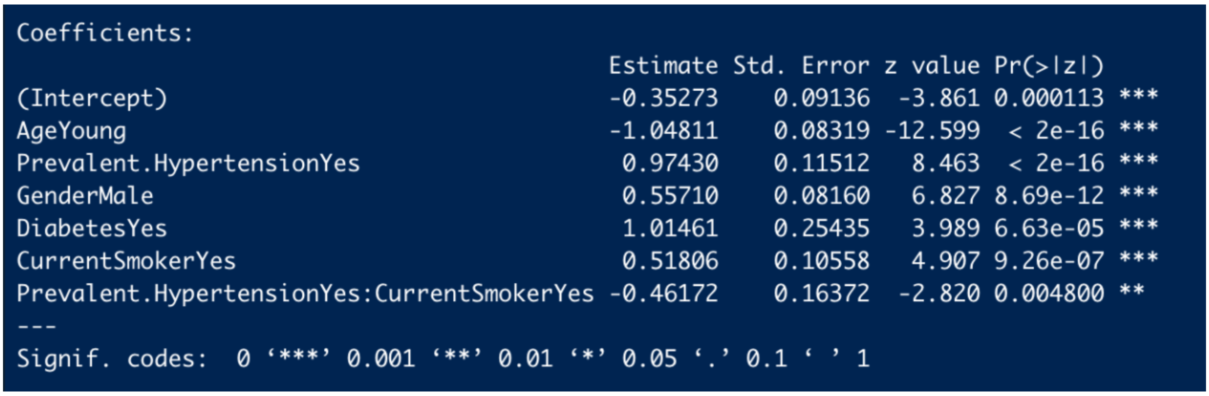
1. Using all the predictor variables to determine the model, I came up with the following model.



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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2254  -1.0090   0.5163   0.9711   1.7498

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.24222    0.08155  -2.970  0.00297 **
GenderMale                 0.53728    0.08128   6.611 3.83e-11 ***
Prevalent.StrokeYes        0.33695    0.52896   0.637  0.52412
AgeYoung                  -1.04498    0.08313 -12.570  < 2e-16 ***
Prevalent.HypertensionYes  0.74912    0.08312   9.012  < 2e-16 ***
DiabetesYes                1.00423    0.25323   3.966 7.32e-05 ***
CurrentSmokerYes           0.33990    0.08399   4.047 5.19e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4113.9  on 2967  degrees of freedom
Residual deviance: 3709.6  on 2961  degrees of freedom
AIC: 3723.6

Number of Fisher Scoring iterations: 4
```

From the output, all but one of the predictor variables are significant. This means the prevalent stroke variable is not needed in our model because it’s not significant.

2. Using a stepwise procedure with interactions and removing the insignificant terms, I obtained this model:



```
Coefficients:
                                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                                -0.35273    0.09136  -3.861 0.000113 ***
AgeYoung                                   -1.04811    0.08319 -12.599  < 2e-16 ***
Prevalent.HypertensionYes                   0.97430    0.11512   8.463  < 2e-16 ***
GenderMale                                  0.55710    0.08160   6.827 8.69e-12 ***
DiabetesYes                                 1.01461    0.25435   3.989 6.63e-05 ***
CurrentSmokerYes                            0.51806    0.10558   4.907 9.26e-07 ***
Prevalent.HypertensionYes:CurrentSmokerYes -0.46172    0.16372  -2.820 0.004800 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The new model is:

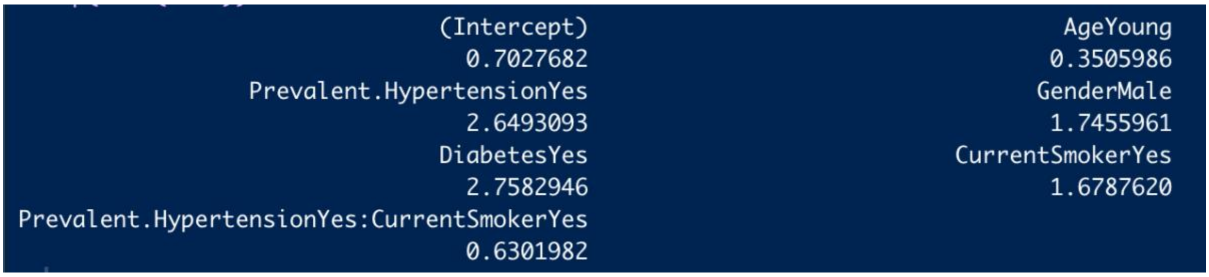
$$\begin{aligned} TenYearCHD = {} & -0.35273 - 1.04811 \times AgeYoung \\ & + 0.97430 \times Prevalent.HypertensionYes + 0.55710 \times GenderMale \\ & + 1.01461 \times DiabetesYes + 0.51806 \times CurrentSmokerYes \\ & - 0.46172 \times Prevalent.HypertensionYes{:}CurrentSmokerYes \end{aligned}$$

Interpreting the model

Holding the other variables constant:

a. Having been a young person (less than 65 years) versus being old, the log odds of developing coronary heart disease decrease by -1.04811.
b. Having been a male versus a female, the log odds of developing coronary heart disease increase by 0.55710.
c. Having been hypertensive, the log odds of developing coronary heart disease increase by 0.97430.
d. Having been a smoker, the log odds of developing coronary heart disease increase by 0.51806.
e. Having been diabetic, the log odds of developing coronary heart disease increase by 1.01461.

For easier interpretation, I’m transforming these values into odd’s ratios (OR):

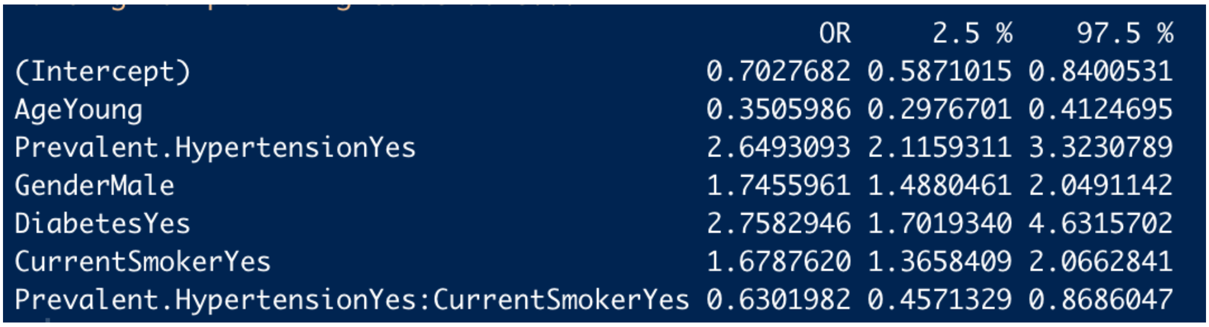


```
                               (Intercept)                                   AgeYoung
                                 0.7027682                                  0.3505986
                 Prevalent.HypertensionYes                                 GenderMale
                                 2.6493093                                  1.7455961
                               DiabetesYes                           CurrentSmokerYes
                                 2.7582946                                  1.6787620
Prevalent.HypertensionYes:CurrentSmokerYes
                                 0.6301982
```

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

a. Having been a young person (less than 65 years) versus being old, the odds of developing coronary heart disease decrease by 0.3506.
b. Having been hypertensive, the odds of developing coronary heart disease increase by 2.6493.
c. Having been a male versus a female, the odds of developing coronary heart disease increase by 1.7456.
d. Having been a smoker, the odds of developing coronary heart disease increase by 1.6788.
e. Having been diabetic, the odds of developing coronary heart disease increase by 2.7583.
f. Surprisingly, having been hypertensive and a smoker, the odds of developing coronary heart disease decrease by 0.6302.

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

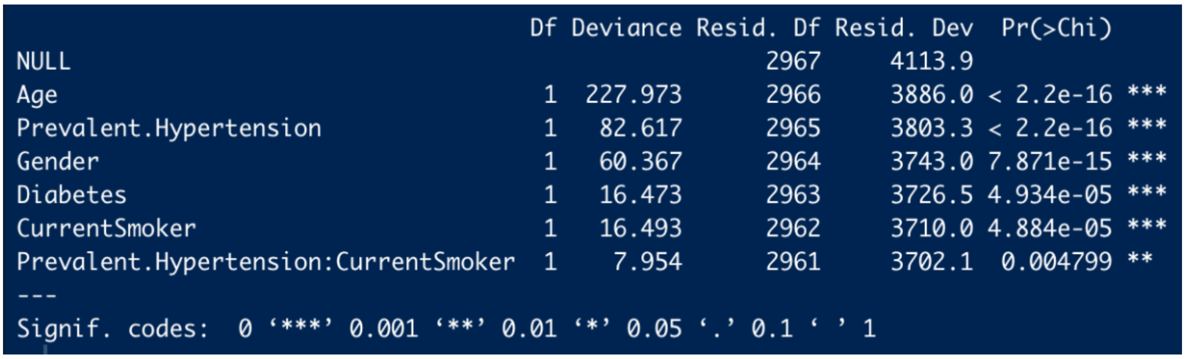


```
                                                  OR     2.5 %    97.5 %
(Intercept)                                0.7027682 0.5871015 0.8400531
AgeYoung                                   0.3505986 0.2976701 0.4124695
Prevalent.HypertensionYes                  2.6493093 2.1159311 3.3230789
GenderMale                                 1.7455961 1.4880461 2.0491142
DiabetesYes                                2.7582946 1.7019340 4.6315702
CurrentSmokerYes                           1.6787620 1.3658409 2.0662841
Prevalent.HypertensionYes:CurrentSmokerYes 0.6301982 0.4571329 0.8686047
```

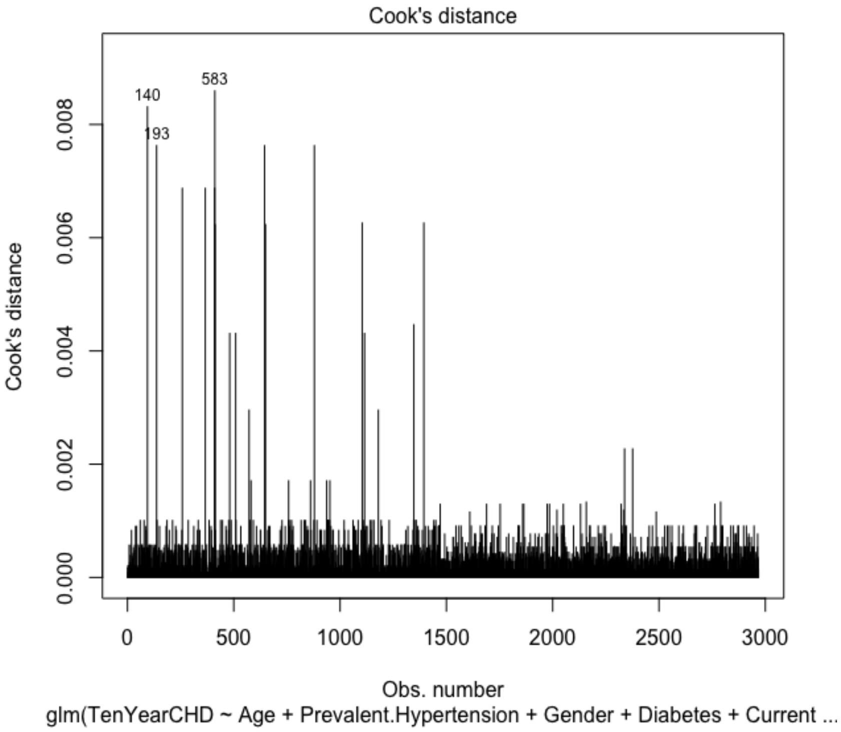
It seems obvious to me that none of the intervals have the value 1. It's interesting to note that someone with diabetes has 4.6316 times higher risk of developing coronary heart disease than someone without the condition.

Goodness of Fit

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

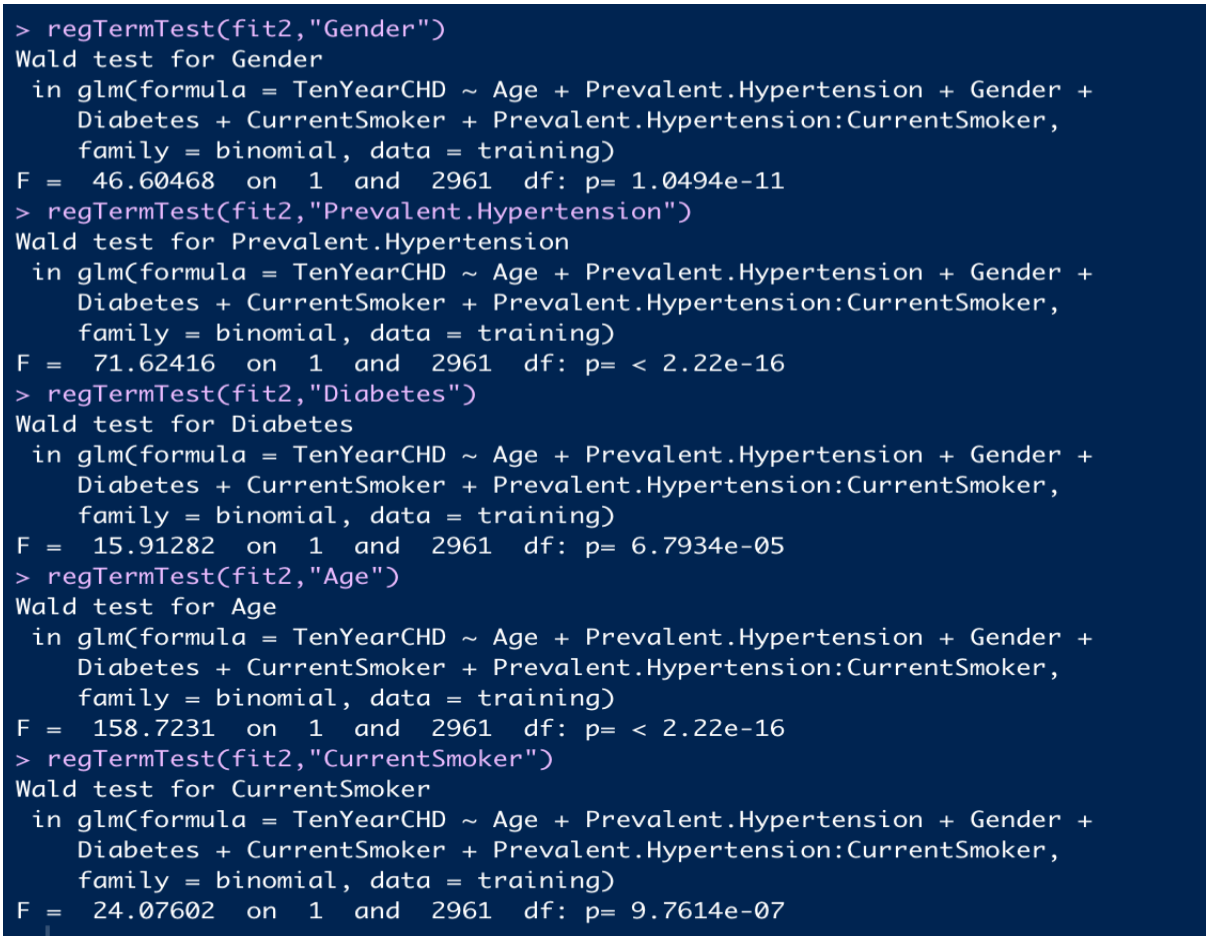


```
                                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                               2967     4113.9              
Age                                 1  227.973      2966     3886.0 < 2.2e-16 ***
Prevalent.Hypertension              1   82.617      2965     3803.3 < 2.2e-16 ***
Gender                              1   60.367      2964     3743.0 7.871e-15 ***
Diabetes                            1   16.473      2963     3726.5 4.934e-05 ***
CurrentSmoker                       1   16.493      2962     3710.0 4.884e-05 ***
Prevalent.Hypertension:CurrentSmoker 1    7.954      2961     3702.1  0.004799 ** 
---
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's distance plot for the data was created, yet none of them are significantly large (greater than 1). This indicates that there are no influential points.

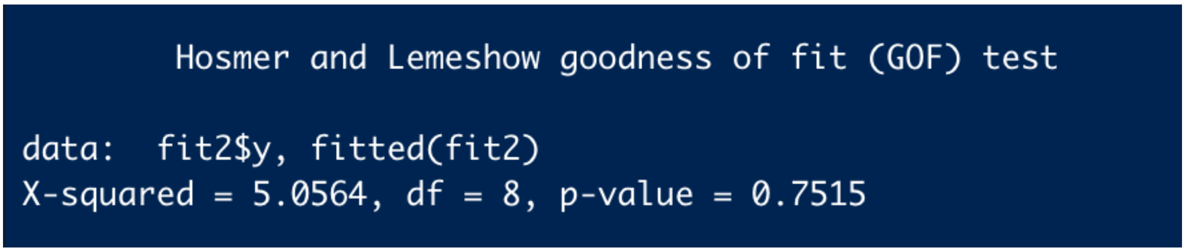
I can also perform Wald Tests on each of the predictors to check and see if they are needed in the model.



```
> regTermTest(fit2,"Gender")
Wald test for Gender
 in glm(formula = TenYearCHD ~ Age + Prevalent.Hypertension + Gender +
    Diabetes + CurrentSmoker + Prevalent.Hypertension:CurrentSmoker,
    family = binomial, data = training)
F =  46.60468  on  1  and  2961  df: p= 1.0494e-11
> regTermTest(fit2,"Prevalent.Hypertension")
Wald test for Prevalent.Hypertension
 in glm(formula = TenYearCHD ~ Age + Prevalent.Hypertension + Gender +
    Diabetes + CurrentSmoker + Prevalent.Hypertension:CurrentSmoker,
    family = binomial, data = training)
F =  71.62416  on  1  and  2961  df: p= < 2.22e-16
> regTermTest(fit2,"Diabetes")
Wald test for Diabetes
 in glm(formula = TenYearCHD ~ Age + Prevalent.Hypertension + Gender +
    Diabetes + CurrentSmoker + Prevalent.Hypertension:CurrentSmoker,
    family = binomial, data = training)
F =  15.91282  on  1  and  2961  df: p= 6.7934e-05
> regTermTest(fit2,"Age")
Wald test for Age
 in glm(formula = TenYearCHD ~ Age + Prevalent.Hypertension + Gender +
    Diabetes + CurrentSmoker + Prevalent.Hypertension:CurrentSmoker,
    family = binomial, data = training)
F =  158.7231  on  1  and  2961  df: p= < 2.22e-16
> regTermTest(fit2,"CurrentSmoker")
Wald test for CurrentSmoker
 in glm(formula = TenYearCHD ~ Age + Prevalent.Hypertension + Gender +
    Diabetes + CurrentSmoker + Prevalent.Hypertension:CurrentSmoker,
    family = binomial, data = training)
F =  24.07602  on  1  and  2961  df: p= 9.7614e-07
```

Like the results before, these p-values indicate that each of the predictor variables is significant in predicting the odds that a person develops coronary heart disease in ten years.

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

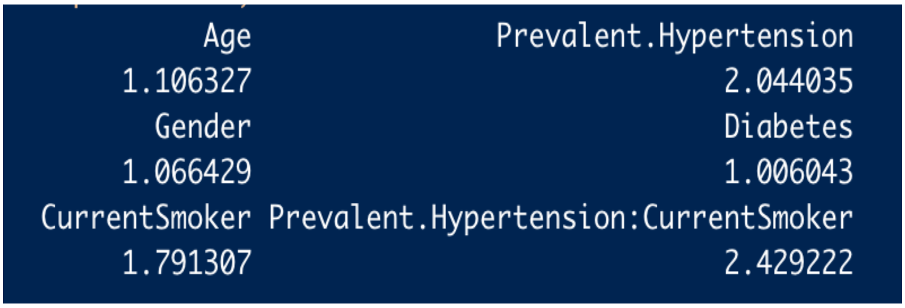


```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  fit2$y, fitted(fit2)
X-squared = 5.0564, df = 8, p-value = 0.7515
```

For the Hosmer-Lemeshow Test, significant p-values indicate that the model is not adequate for predicting the odds of developing coronary heart disease based on gender, age, hypertension, diabetes, and smoking. However, the p-value is 0.7515 so I can say that there is strong evidence that the model is a good fit.

Collinearity

After assessing the goodness of fit of the logistic model, I 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.

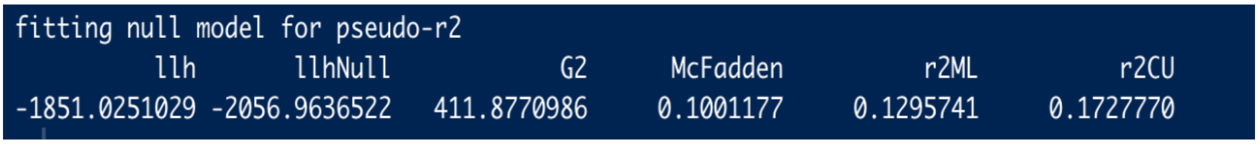


```
          Age             Prevalent.Hypertension
     1.106327                           2.044035
       Gender                          Diabetes
     1.066429                           1.006043
 CurrentSmoker Prevalent.Hypertension:CurrentSmoker
     1.791307                           2.429222
```

Since none of the VIF values are larger 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.

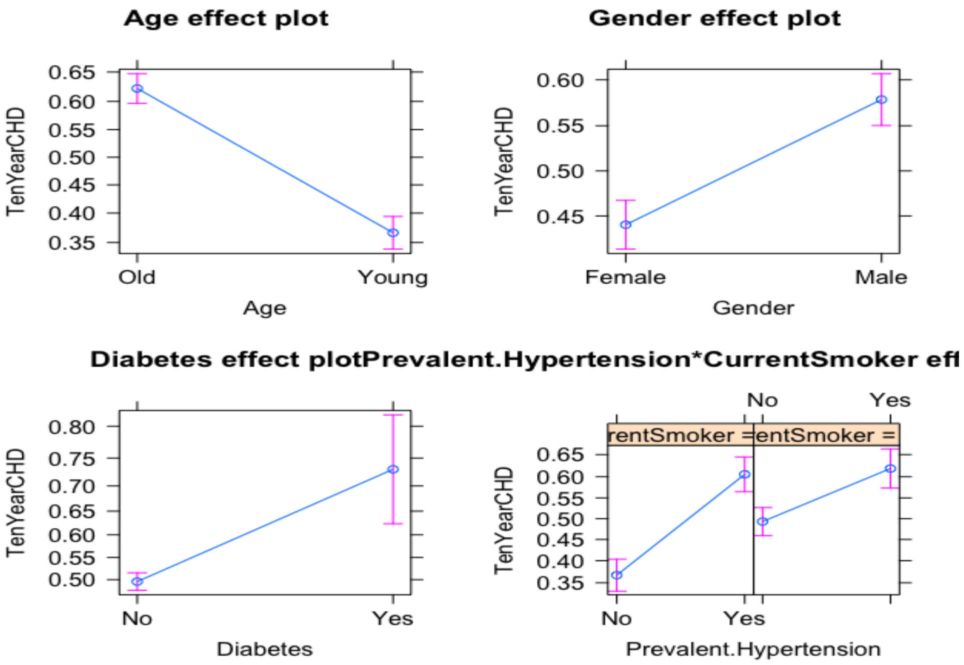


```
fitting null model for pseudo-r2
          llh       llhNull           G2      McFadden          r2ML          r2CU
-1851.0251029 -2056.9636522  411.8770986     0.1001177     0.1295741     0.1727770
```

A McFadden R^2 value between 0.2 and 0.4 is considered good. Therefore, since our McFadden R^2 is 0.1001 we can say that the model selected is a fairly good fit for predicting the odds of developing coronary heart disease.

Effect Size

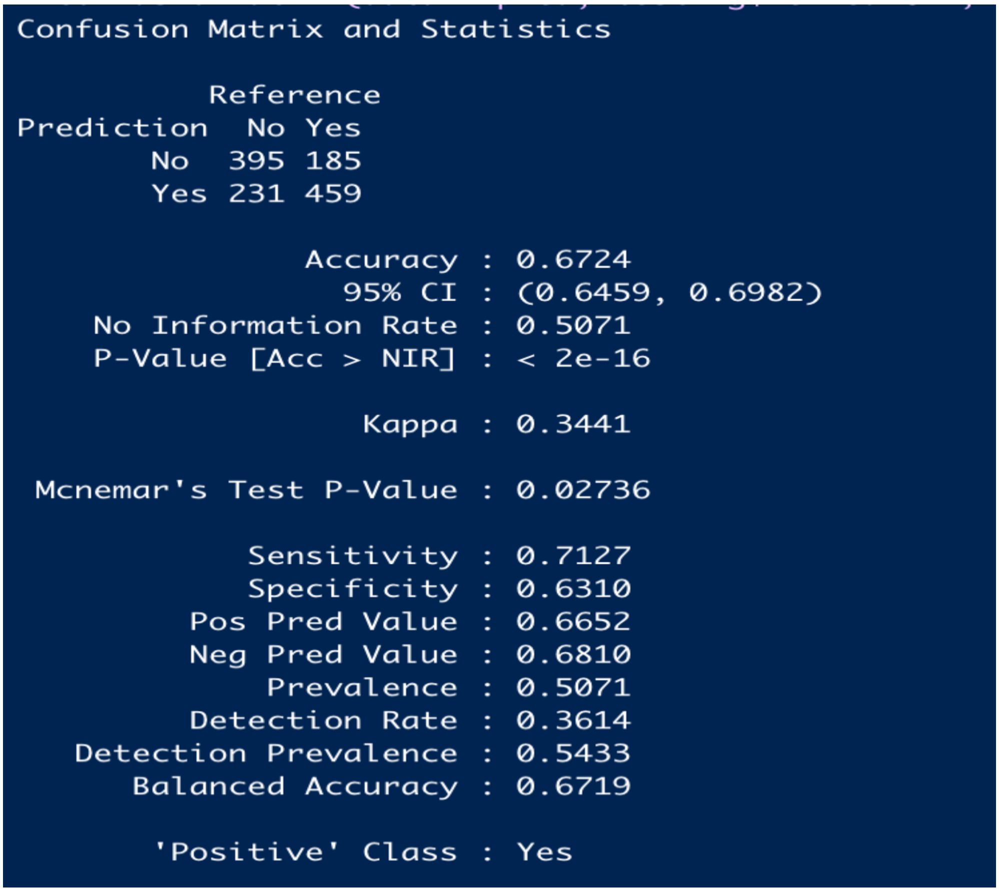
To determine the effect of the significant five individual predictor variables on the odds of developing coronary heart disease, let’s make a plot to determine the effect each one has, individually:



If a person is old (above 65 years), the chances of developing coronary heart disease in ten years are 30% more than a young one. This is corroborated by the National Institute of Aging [2]. Also, if a person is male, he's almost 20% more likely to develop coronary heart disease than a female in ten years. This is corroborated by Weidner G. (2000) [3]. A diabetic person is also 20% more likely to develop coronary heart disease than a non-diabetic in ten years, corroborated by a John Hopkins Medicine report [4] and a hypertensive smoker is almost 10% more likely to develop coronary heart disease than a non-hypertensive smoker in ten years. Surprisingly, a hypertensive smoker has almost the same likelihood of developing coronary heart disease as a hypertensive non-smoker.

Cross Validation

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

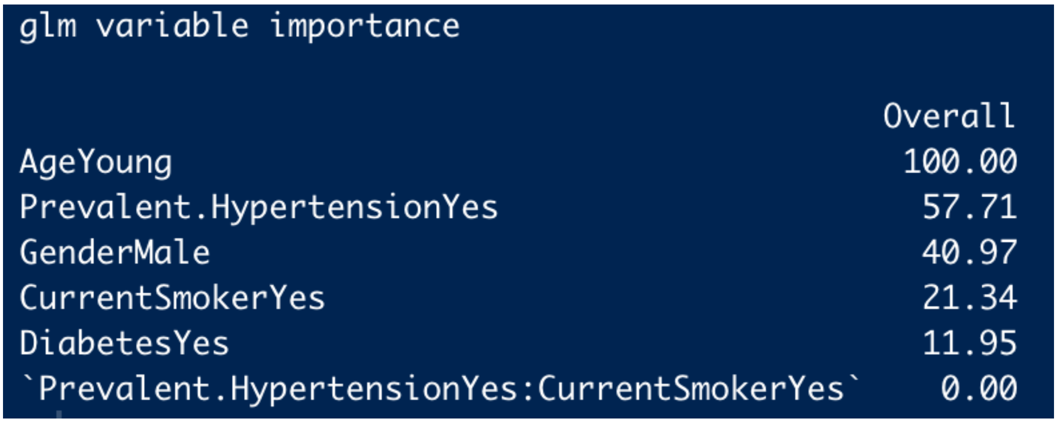


```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  395 185
       Yes 231 459

               Accuracy : 0.6724
                 95% CI : (0.6459, 0.6982)
    No Information Rate : 0.5071
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.3441

 Mcnemar's Test P-Value : 0.02736

            Sensitivity : 0.7127
            Specificity : 0.6310
         Pos Pred Value : 0.6652
         Neg Pred Value : 0.6810
             Prevalence : 0.5071
         Detection Rate : 0.3614
   Detection Prevalence : 0.5433
      Balanced Accuracy : 0.6719

       'Positive' Class : Yes
```

The overall accuracy of the model to predict survival rate is 0.6724 with a sensitivity (the proportion who developed coronary heart disease who were predicted to have had it based on the model) is 0.7127 yet the specificity (the proportion who did not develop coronary heart disease who were predicted not to have had it based on the model) was 0.6310. This indicates that our model does a better job at correctly predicting the chances that someone developed coronary heart disease than predicting the chances that someone did not. Despite not-so-high accuracy, the sensitivity of 0.7127 is high enough, which means our model can distinguish 71% of the time, persons who later developed coronary heart disease and those who didn't.

Variable of Importance

I can assess the importance of individual predictors in the model.

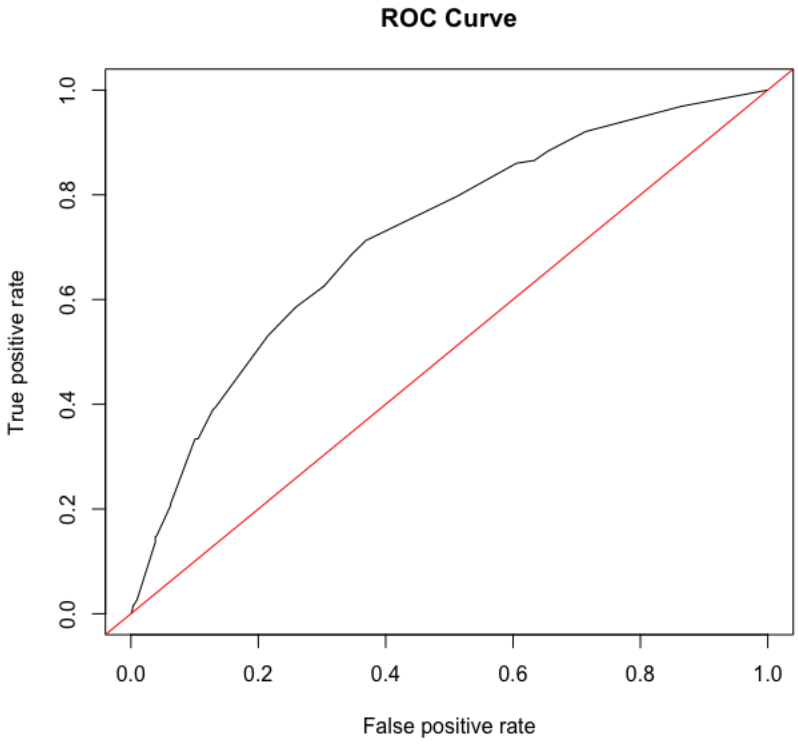


```
glm variable importance

                                              Overall
AgeYoung                                       100.00
Prevalent.HypertensionYes                       57.71
GenderMale                                      40.97
CurrentSmokerYes                                21.34
DiabetesYes                                     11.95
`Prevalent.HypertensionYes:CurrentSmokerYes`     0.00
```

It appears that age has the biggest impact on the probability of developing coronary heart disease in ten years. This is closely followed by prevalent hypertension, gender, smoking, and diabetes variables. It isn't surprising to see that the overall importance of prevalent hypertension and smoking is 0 since that interaction term has the highest p-value.

Receiver Operating Characteristics (ROC) Curve

The ROC curve is shown below:



The area underneath this ROC curve is 0.7194. 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 0.7194, the test does a good job of separating the persons that developed coronary heart disease in ten years from those that didn't, making predictions using the chosen model.

Conclusion

The World Health Organization estimates that 12 million people die each year as a result of heart disease. Cardiovascular diseases account for half of all fatalities in the United States and other developed countries. The early diagnosis of cardiovascular disease can help high-risk individuals make lifestyle adjustments, which can lessen consequences. This research revealed that age, prevalent hypertension, gender, smoking, and diabetes all have a significant impact on assessing the chance of having coronary heart disease in the next ten years. Smoking because both nicotine and carbon monoxide (from the smoke) put a strain on the heart by making it work faster, increasing the risk of blood clots which could lead to coronary heart disease. Hypertension because it puts a strain on the heart, increasing the odds of developing CHD. Diabetes because over time, high blood sugar can damage blood vessels and the nerves that control the heart leading to coronary heart disease. Lastly, aging is associated with a progressive decline in numerous physiological processes, leading to an increased risk of health complications and disease. By delivering oxygenated blood to all tissues in the body, the health of the cardiovascular system is vital for the health of every tissue and the longevity of the organism as a whole. Aging has a remarkable effect on the heart and arterial system, leading to an increase in coronary disease [5]

Recommendation

To summarize, to reduce the risk of coronary heart disease, one must prioritize one's health. Eat a good and balanced diet to keep your blood sugar in check and avoid diabetes; exercise to maintain a healthy weight and reduce your chances of getting hypertensive; and stop smoking and drinking alcohol. If you are already hypertensive or diabetic, take your medications seriously.

References

1. What Is Coronary Heart Disease? (2022, March 24). NHLBI, NIH. https://www.nhlbi.nih.gov/health/coronary-heart-disease
2. Heart Health and Aging. (n.d.). National Institute on Aging. https://www.nia.nih.gov/health/heart-health-and-aging
3. Weidner G. (2000). Why do men get more heart disease than women? An international perspective. Journal of American college health: J of ACH, 48(6), 291–294. https://doi.org/10.1080/07448480009596270
4. Diabetes and Heart Disease. (2019, November 19). Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-heart-disease
5. North, B. J., & Sinclair, D. A. (2012). The Intersection Between Aging and Cardiovascular Disease. Circulation Research, 110(8), 1097–1108. https://doi.org/10.1161/circresaha.111.246876