STAT 149 Framingham Heart Study Course Project

Jordan Turley

May 6, 2020

1 Introduction

We are tasked with modeling the probability of being at a 10-year risk of coronary heart disease. We are using data from the Framingham Heart Study, a long term ongoing study of heart health of the residents of Framingham, MA.

There are two important aspects of modeling data like this. The first is inference. If we can create an interpretable model, this will help to see what causes an individual to be at risk for heart disease. This would allow us to see, for example, how much smoking or high blood pressure puts an individual at risk.

The second aspect is prediction. If we are able to create an accurate model to predict a user's risk of heart disease, this can be used to save valuable time in a doctor's day. A model would allow for an individual to have a few quick tests that could be done by a nurse, like heart rate and blood pressure, and for the individual to give a few self-reported pieces of information like their education level and number of cigarettes smoked per day. This is compared to having a full examination done by a doctor. It may be more useful for the doctor's time to be dedicated to more high risk individuals than for an individual that the model only says has a 5% chance of being at risk for heart disease.

We will use methods involving generalized linear models to model the probability of being at risk for heart disease as a function of several other predictor variables.

2 Data

The data is obtained from the Framingham Heart Study. Below, we will take an in-depth look at the variables that we are given, check the multicollinearity of the variables, and look at the relationship between each predictor and the response variable through several plots.

2.1 Dataset

The dataset contains 4,238 residents of Framingham, MA that were involved in the Framingham Heart Study. The response variable is `CHD_Risk`, which is a "yes" or "no" indicating whether the individual is at a 10-year risk of coronary heart disease. We are given several quantitative and categorical variables that are listed in Figure 1. We will use these variables to create a model that will predict the probability of an individual being at a 10-year risk of coronary heart disease.

2.2 Preprocessing

The first issue we see with this dataset is that there is some missing data. In Figure 2, we see the percentage of each variable that is missing.

While there is missing data, we see that no variable has too significant of a number of rows missing its data. The highest percentage of missing data is `glucose`, but it is less than ten percent. Because of this, it wouldn't be too big of a deal to simply drop the rows that contain missing data. However, there could be some reasoning behind the missing data that could help predict a probability of risk of heart disease.

We use a simple method to deal with missing data. For quantitative variables, we impute the mean and create a new column indicating whether the original value was missing. For categorical variables we simply create a new factor for a missing value. This allows us to use the fact that data was missing as part of our model rather than just dropping rows with missing data.

Next, we check the collinearity of the variables in this dataset after fixing the missing values. If there is high multicollinearity between two or more of the variables, this could cause problems like inflated coefficient or standard error estimates, insignificant Wald tests, or illogical signs of variables. We calculate the generalized variance inflation factor of each variable. We are given the square root of the GVIF, and as a rule of thumb, we compare this to $\sqrt{10} \approx 3.162$. The GVIF for each variable is shown in Figure 3.

The highest GVIF is for systolic blood pressure and is 1.937 which is well below our threshold of $\sqrt{10}$. We do not have evidence of collinearity.

Finally, to visually see our variables and their relationship with the response, we plot each variable versus the 10-year risk response in Figures 4 and 5. We can clearly see some relationships between the variables. For example, the average age is noticeably higher for individuals who are at a 10 year risk of heart disease than those who are not. Average systolic blood pressure is noticeably higher for individuals at risk than those at not as well. We see that our dataset is very imbalanced, as there are about seven times as many individuals who are not at risk than those who are. However, we see significant relationships for sex, blood pressure medication, hypertension, and diabetes. It was surprising that there was not a larger relationship for the smoking variables, but we will see if these are statistically significant in the modeling section.

Before we start modeling the data, we split the dataset into 80% training and 20% test sets.

3 Modeling

We would like to model the 10-year risk of coronary heart disease, `CHD_Risk`, as a function of the other variables we have. To do this, we will use a logistic regression model, a generalized additive model, and a classification tree model. We will compare the logistic regression and generalized additive model using likelihood ratio tests, and we will compare all three using the test set.

3.1 Logistic Regression

We start by simply doing a logistic regression of `CHD_Risk` on all of the variables we are given. This yields the model shown in Figure 6. We see several variables that are significnat, as well as several that are not significant. We will later see which of these can be removed through an analysis of deviance.

First, we would like to see if there are any interactions that are significant between the categorical variables or between a categorical variable and a quantitative variable. There seems to be validity to some interactions, like sex for example. For example, it seems like it could be more dangerous to smoke or have high blood pressure for a male than a female, or vice versa. This would be a good place to consult with a doctor to see which interactions would make medical sense; however, we can simply test all interaction terms between a categorical variable and another variable with a likelihood ratio test.

When we perform this test between the model with no interactions and the model with all interactions, we get a deviance of 191.98 on 167 degrees of freedom, with a p-value of $p = 0.09$. This is a gray area of statistical significance. The test is significant at the 10% level, but not at the 5% level. Depending on the significance level we want, we can either include or exclude the interaction terms. However, this actually tells us a lot about the interaction terms. With a significance of 9%, we see that some of the interaction terms may be significant and contribute to our model, but in practical terms, these are likely not contributing to our model in a way that is significantly better than our base model with no interactions. There may be medical reasons to include some of the interaction terms, but in terms of statistical significance and practical use, there do not seem to be any useful interaction terms, so we will continue using the base model with no interaction terms.

Next, we conduct an analysis of deviance to reduce and simplify the model and find the set of predictors that will still give us a thorough and accurate model. We use backwards selection to reduce the full model. The analysis of deviance is shown in Figure 7. Because we have so many variables, I listed the variables that I was removing rather than the variables I was including.

After the analysis of deviance, we are left with the reduced model in Figure 8. The significant predictor variables are `age`, `cigsPerDay`, `totChol`, `sysBP`, `glucose`, `sexmale`, and `PrevStroke`. We likely could have removed some of the other predictor variables, like `diaBP` and `smoker`, but I decided to leave these variable because they were paired with another variable, and it is no extra effort to include these variables. For example, if this was used in the future and you were taking someone's blood pressure to plug into the model, you get both the systolic and diastolic numbers in the blood pressure reading, so it's no more trouble to plug in both. The smoker variable is in a way redundant with the number of cigarettes smoked per day, so it probably could have been removed, but I decided to leave it. Finally, we see in the last likelihood ratio test that none of the missing data variables are significant. We could have gone either way for these variables. I ultimately decided to leave them in the model since removing them seemed like we were only executing half of the solution we implemented for dealing with missing data. Also, the missing data is minimal, so these coefficients will be multiplied by zero most of the time.

Next, we would like to see if we lost anything significant by reducing the model from the full model. We use one final likelihood ratio test to test the full model versus the reduced model. We get a deviance of 8.2 on 10 degrees of freedom, with a p-value of $p = 0.6093$. Since this test is insignificant, we see that we do not have significant evidence that the full model is doing more than the reduced model, so we would prefer to use the reduced model.

Finally, we can also look at the ratio of the deviance to the degrees of freedom for the reduced model to evaluate fit. We have a deviance of 2602.5 on 3376 degrees of freedom, yielding a ratio of 0.7708. A value close to one indicates a good fit, and our ratio is close to one, which indicates that this model is fitting the data well. A value of less than one may indicate overfitting, but we are not significantly lower than one to the point we would worry about overfitting, so it seems like our model is fitting this data well.

3.2 Generalized Additive Model

Next, we would like to fit a generalized additive model to the data to see if there are some nonlinearities that are not captured by the simple logistic regression model. We fit two generalized additive models. We fit one on the full dataset with smoothers for the continuous variables and regular categorical variables. We also fit the same model on the reduced set of predictor variables found in the previous section using an analysis of deviance. We will compare these two models, and we will also compare these models to the logistic regression models from the previous section.

We first fit the full GAM model. If we compare this model to the full logistic regression model using a likelihood ratio test, we get a deviance of 21.873 and a p-value of $p = 0.0009$. This test is very significant and indicates that there is evidence that the full GAM model is capturing aspects that our regular logistic regression model is not.

Next, we fit a reduced GAM model on the reduced predictor set. We compare this model to the full GAM model using a likelihood ratio test and get a deviance of -11.019 and a p-value of $p = 0.4941$. This test is not significant, indicating that there is not significant evidence to use the full model over the reduced model, so we will continue with the reduced model.

Finally, we compare the reduced GAM model to the reduced logistic regression model. We get a deviance of -19.054 and a p-value of $p = 0.6235$. This indicates that there is not significant evidence to use the smoother model over the regular logistic regression model. The smoothed variable plots are included in Figure 9.

It is peculiar that, in terms of likelihood ratio tests, we prefer the full GAM over the full logistic regression model, but we prefer the reduced GAM over the full GAM, and the reduced logistic regression model over the reduced GAM. One possible explanation for this is that there was a nonlinear variable that was discarded because it was not significant in the logistic regression model, but it would have been significant in the GAM.

An analysis of deviance of the GAM would reveal if this is the case and would be interesting to look at. However, as we will see with the predictions and accuracy on the test set that, even if a variable that is statistically significant in the GAM was discarded, it does not make a practical difference.

3.3 Classification Tree

Finally, we fit a classification tree to the data. First, we fit the tree on the full dataset with no regularization. Next, we use cross-validation to select the optimal value of `cp`. Finally, we prune the full tree to find the reduced tree.

The full and pruned trees are shown in Figure 10, and the cross-validation results are shown in Figure 11. We see that the original unpruned tree is much too broad and deep. This tree is likely overfit and will not perform well on new data. To determine the value of `cp`, we look at the output and plot we get using cross-validation, and determine that the best value is 0.0044. After pruning, we get a much more reasonable tree. The variables at the top two levels are `age` and `sysBP`, which seems to indicate that these two variables are important, among others.

3.4 Model Comparison

To compare the three models, we make predictions on the held-out test set. We look at the number of yes/no's predicted for each model as well as the overall accuracy. These results are shown in Figure 12. The training set is about 84% no/16% yes, so we will compare our accuracy to the accuracy we could achieve by predicting all no.

All of the models give similar predictions and accuracy numbers. The worst model is the full tree model, since the model is very overfit to the training data. The other models all give very similar accuracy and yes/no counts. All of our models overpredict the number of no's and underpredict the number of yes's, but we are able to predict better than the naive way of just predicting no.

Occam's razor says that the simpler solution is the better one. Using this, the statistical significance tests we performed in the previous sections, and the accuracy numbers we observe, I would suggest that we use the reduced logistic regression. The likelihood ratio tests we performed between the logistic regression and GAM models showed that the GAM was not statistically signifcantly better than the logistic regression, and we see from the accuracy numbers that the predictions between the logistic regression and the GAM are almost identical. The tree model seems to perform worse than the logistic regression and the GAM, so this model should not be used. The final decision is between the full and reduced logistic regression model. We did a likelihood ratio test and saw that there was not evidence that we should use the full model, and we see that the accuracy is only marginally better for the full model compared to the reduced model. The small increase in accuracy could simply be random. Because of this, I would recommend that the reduced logistic regression model be used.

3.5 Model Diagnostics and Interpretation

We would like to interpret the reduced logistic regression model and see what makes an individual be at risk for heart disease. The model summary is shown in Figure 8.

The two largest coefficient estimates (ignoring the na variables which are not significant) are `sexmale = 0.491786` and `PrevStrokeYes = 0.982953`. We see that being male and/or previously having a stroke gives you a much higher probability to be at risk for heart disease. For example, previously having a stroke causes an individual's log-odds of being at risk of heart disease to be 0.983 higher than an individual that did not have a stroke. Unfortunately, these are two attributes that a person cannot change, but it is important for an individual to know that men and stroke victims are inherently at a higher risk than women and non-stroke victims. It is possible that having a stroke is influenced by some other factor or factors that are not included in this dataset, or maybe having a stroke causes damage to your body that can lead to heart disease. Consulting with a medical doctor could explain this more.

We look at the sign of the coefficients and see that most of them are as expected. The coefficients for `age`, `cigsPerDay`, `totChol`, `sysBP`, `glucose`, `sexmale`, and `PrevStrokeYes` are all statistically significant and positive, indicating that as these varaibles increase, so does the probability of being at risk for heart disease. The coefficient for `diaBP` is negative, which does not make sense, but this variable is not significant. The coefficient for `smokerSmoker` is insignificant as well, but the sign is positive which we would expect. For example, as an individual's age increases by one year, the log-odds of being at risk increases by 0.063.

The type of individual that would have the highest probability of being at risk is a person that is older, smokes a lot every day, has high cholesterol, high blood pressure, high glucose, is male, and has previously had a stroke. An individual that would have a low probability of being at risk is the opposite: a young person that doesn't smoke, low cholesterol, low blood pressure, low glucose, female, and has never had a stroke.

Finally, we see that several variables do not matter as much as we may think by comparing the full and reduced model. We saw that the reduced model was preferred to the full model, indicating that we can exclude variables `education`, `BMI`, `heartRate`, `OnBPMeds`, `Hyp`, and `Diab`. Education could be a proxy for socioeconomic status, but we see it is insignificant. BMI and heart rate can indicate overall health, but these are insignificant. Taking blood pressure medication and having hyptertension are somewhat redundant with the blood pressure numbers we have, so they do not matter, and having diabetes does not matter either. This does not mean that these variables are unimportant for overall health, but in terms of being at risk for heart disease, factors like age, cholesterol, and sex are more important in a statistical and practical sense.

4 Conclusion

This is an interesting dataset to study and allows us to use several different methods of modeling. We can clearly see the importance of studying this dataset and predicting the probability of being at risk. Seeing a doctor is expensive for the individual and time-consuming for a doctor. The current pandemic has shown us that doctors' hours are valuable, and even more in a time like now where treating a COVID-19 patient may be more urgent than determining whether or not a person is at risk for heart disease. This does not mean that heart disease is unimportant, but if a doctor is needed elsewhere, an accurate model would be useful for individuals worried about their heart health.

We used several methods of modeling. We used a basic logistic regression, a generalized additive model, and a classification tree to model the probability of an individual being at risk for heart disease. We determine that several variables are insignificant, both in terms of statistical significance and predictions in practice. We determined that the best model is a reduced logistic regression modeling the probability as a function of age, daily cigarettes smoked, total cholesterol, systolic and diastolic blood pressure, glucose, sex, smoker/nonsmoker, and previously having a stroke.

When we interpret the model, we do not discover anything groundbreaking, but we do confirm intuitions around heart health with statistical evidence. All of the significant variables followed the sign we would expect; for example, the probability of being at risk increases as age or cholesterol increases. It is more surprising which variables we exclude. We see that the variables education, BMI, heart rate, blood pressure medicine yes/no, hypertension, and diabetes are not statistically significant. Something like BMI is still important relating to an individual's overall health, but with respect to heart disease, it is not as important as one might think.

This study yields a lot of future work. It would be interesting to talk to a medical doctor that is an expert in heart health to review this study. We found that interactions between the variables may or may not have been statistically significant, but a doctor could tell us if two variables have a medical reason to be interacted. A doctor could also tell us if a variable should be included in the model that we excluded that may be important. It would also be interesting to further analyze the previous stroke variable and see if strokes are caused by some excluded variable that would be significant to our model. Finally, more data is always helpful in statistics. The data in this study was limited to the individuals in the Framingham Heart Study, which are residents of Framingham, MA. It would be interesting to look at data from other parts of the country or other parts of the world. Just in the United States, life is different enough on the East and West coast, or in the North and South that location might be a significant variable. This data also may be older, so more recent data may give a different analysis.

The inference value of this model is the most valuable part of this study. This reinforces our precognitions of heart health. We have strong statistical evidence that certain variables affect risk of heart disease, which is useful for an individual to know. Even if an individual cannot change a variable, like their age or sex, an individual can use this information for their own benefit. An individual knowing that they are at increased risk could encourage them to make other changes in their life, like exercising more or eating healthier foods, which could be the difference between the individual developing heart disease or not.

5 Figures

Quantitative	**Categorical**
Age	Education*
Average Cigarettes per Day	Sex
Total Cholesterol Level (mg/dL)	Smoker/Nonsmoker
Systolic Blood Pressure (mmHg)	On Blood Pressure Medication (yes/no)
Diastolic Blood Pressure (mmHg)	Had a Stroke (yes/no)
Body Mass Index (kg/m^2)	Has Hypertension (yes/no)
Heart Rate (beats per minute)	Has Diabetes (yes/no)
Glucose (mg/dL)	

Figure 1: Quantitative and Categorical Variables
*Some High School, High School/GED, Some College, College, or Higher

Variable	**% Missing (Count)**	**Variable**	**% Missing (Count)**
Age	0% (0)	Education	2.478% (105)
Average Cigarettes per Day	0.684% (29)	Sex	0% (0)
Total Cholesterol Level	1.179% (50)	Smoker/Nonsmoker	0% (0)
Systolic Blood Pressure	0% (0)	On Blood Pressure Medication	1.251% (53)
Diastolic Blood Pressure	0% (0)	Had a Stroke	0% (0)
Body Mass Index	0.448% (19)	Has Hypertension	0% (0)
Heart Rate	0.024% (1)	Has Diabetes	0% (0)
Glucose	9.155% (388)		

Figure 2: Percent of Missing Data for each Column

Variable	**GVIF$\hat{(}$1/(2*Df))**	**Variable**	**GVIF$\hat{(}$1/(2*Df))**
Age	1.189	Education	1.015
Average Cigarettes per Day	1.635	Sex	1.109
Total Cholesterol Level	1.053	Smoker/Nonsmoker	1.602
Systolic Blood Pressure	1.937	On Blood Pressure Medication	1.026
Diastolic Blood Pressure	1.726	Had a Stroke	1.014
Body Mass Index	1.115	Has Hypertension	1.434
Heart Rate	1.048	Has Diabetes	1.262
Glucose	1.269	`cigsPerDay.na`	1.010
		`totChol.na`	1.042
		`BMI.na`	1.007
		`heartRate.na`	1.003
		`glucose.na`	1.044

Figure 3: Generalized Variance Inflation Factor of each Variable

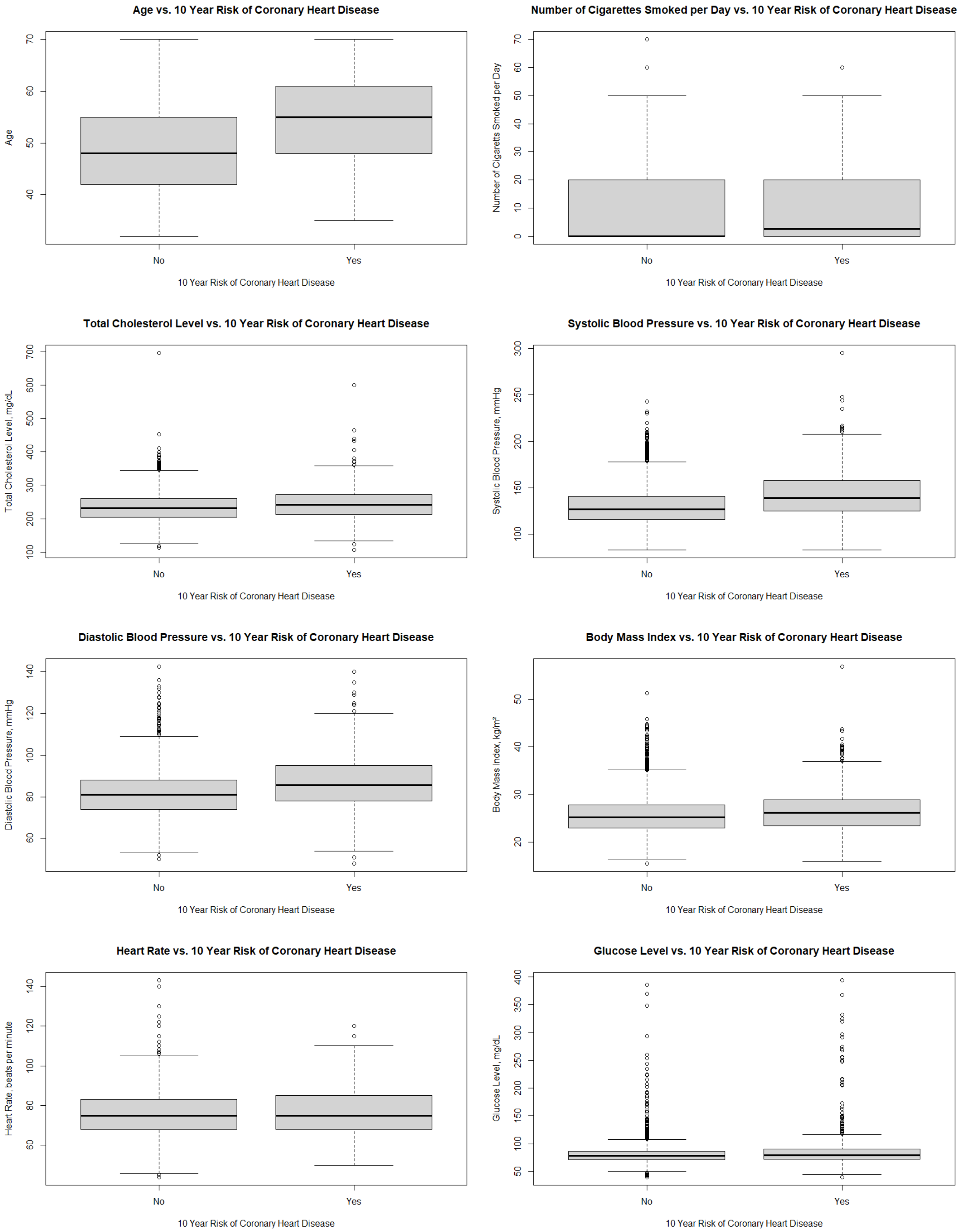


Figure 4: Quantitative Variable Plots with 10 Year Risk of Coronary Heart Disease

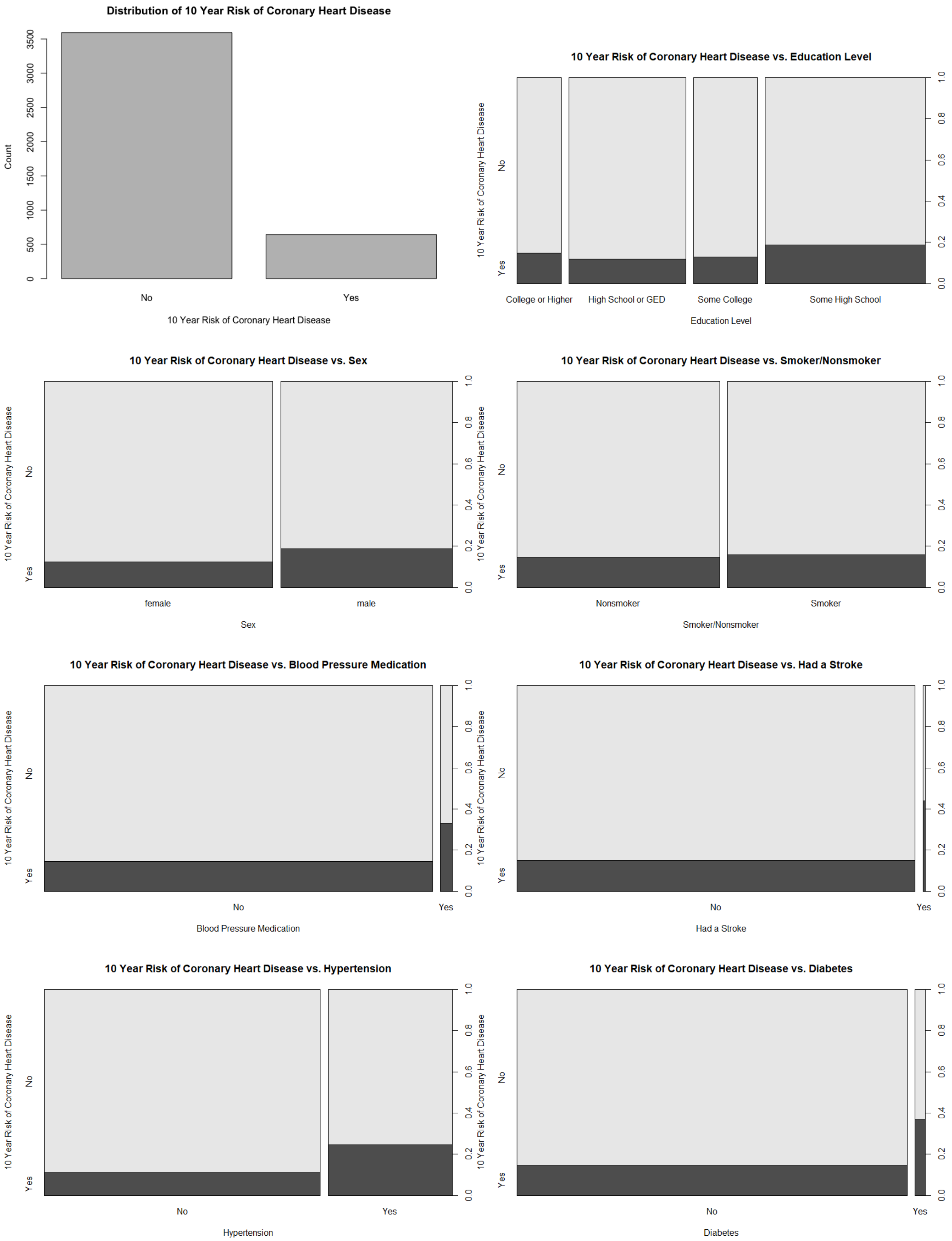


Figure 5: Categorical Variable Plots with 10 Year Risk of Coronary Heart Disease

```
Call:
glm(formula = CHD_Risk ~ ., family = binomial, data = chd.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0209  -0.6004  -0.4311  -0.2901   2.8301

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -7.905780   0.818989  -9.653  < 2e-16 ***
age                           0.060743   0.006969   8.716  < 2e-16 ***
educationHigh School or GED  -0.271775   0.182432  -1.490 0.136294
educationNA                  -0.263868   0.339788  -0.777 0.437414
educationSome College        -0.077201   0.199245  -0.387 0.698408
educationSome High School    -0.039896   0.168736  -0.236 0.813092
cigsPerDay                    0.017427   0.006396   2.724 0.006441 **
totChol                       0.002273   0.001126   2.018 0.043592 *
sysBP                         0.014774   0.004041   3.656 0.000256 ***
diaBP                        -0.003234   0.006653  -0.486 0.626909
BMI                           0.005305   0.013028   0.407 0.683837
heartRate                    -0.002367   0.004332  -0.546 0.584836
glucose                       0.007658   0.002337   3.276 0.001051 **
sexmale                       0.467272   0.112618   4.149 3.34e-05 ***
smokerSmoker                  0.156424   0.160247   0.976 0.328994
OnBPMedsNo                   -0.261541   0.388319  -0.674 0.500616
OnBPMedsYes                  -0.078204   0.450756  -0.173 0.862261
PrevStrokeYes                 0.895243   0.478384   1.871 0.061291 .
HypYes                        0.207406   0.143830   1.442 0.149298
DiabYes                       0.053471   0.317394   0.168 0.866215
cigsPerDay.na                -1.112669   1.043642  -1.066 0.286360
totChol.na                    0.441503   0.429226   1.029 0.303667
BMI.na                        0.669953   0.760095   0.881 0.378098
heartRate.na                 12.264550 324.743771   0.038 0.969874
glucose.na                   -0.147145   0.197515  -0.745 0.456282
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2936.5  on 3390  degrees of freedom
Residual deviance: 2594.3  on 3366  degrees of freedom
AIC: 2644.3

Number of Fisher Scoring iterations: 11
```

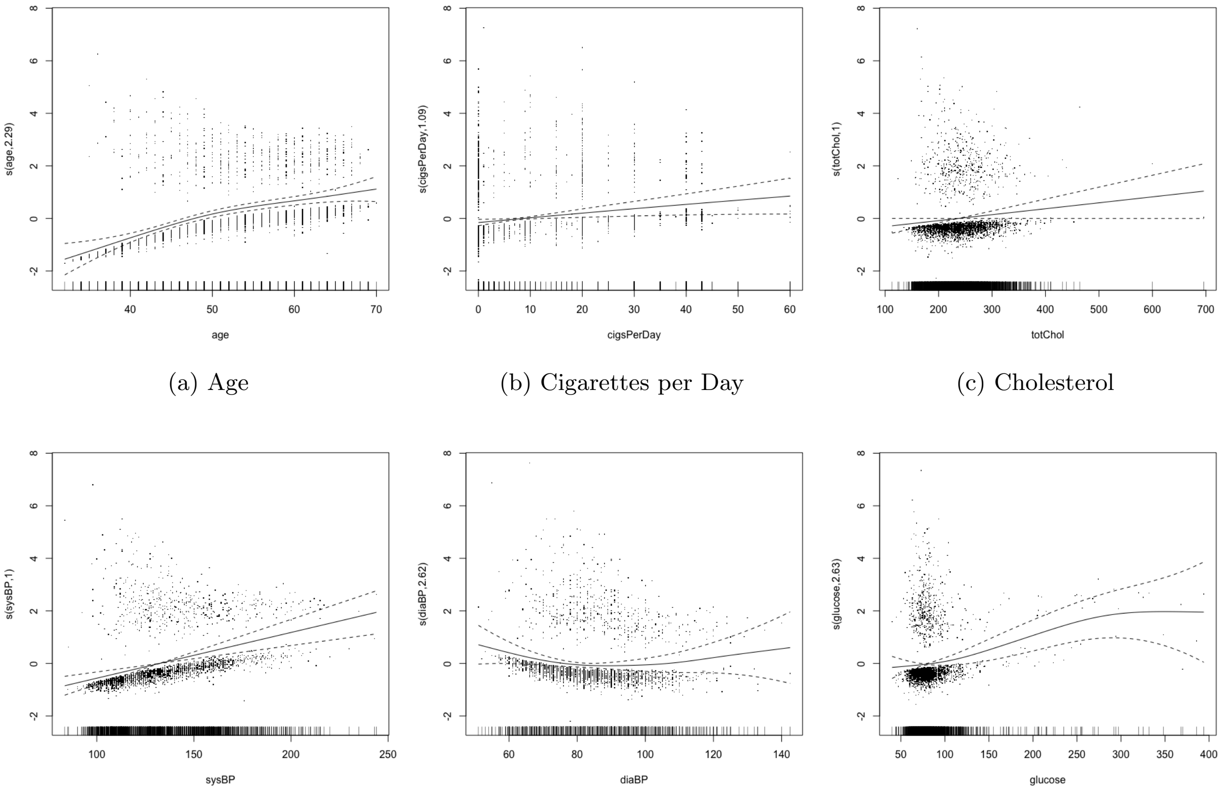
Figure 6: Full Logistic Regression Model

Model 1	**Model 2**	**Deviance**	**p-Value**
Full	Full - education	-4.1807	0.3821
Full - education	Full - education - cigsPerDay - smoker	-25.333	3.155e-06
Full - education	Full - education - sysBP - diaBP	-18.788	8.322e-05
Full - education	Full - education - BMI - heartRate - OnBPMeds	-1.6064	0.8076
Full - education - BMI - heartRate - OnBPMeds	Full - education - BMI - heartRate - OnBPMeds - Hyp - Diab	-2.4129	0.2993
Full - education - BMI - heartRate - OnBPMeds - Hyp - Diab	Full - education - BMI - heartRate - OnBPMeds - Hyp - Diab - na variables	-4.9434	0.4228

Figure 7: Analysis of Deviance for Logistic Regression Model

```
Call:
glm(formula = CHD_Risk ~ . - education - BMI - heartRate - OnBPMeds -
    Hyp - Diab, family = binomial, data = chd.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0501  -0.6023  -0.4379  -0.2926   2.8099

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.869818   0.549216 -16.150  < 2e-16 ***
age             0.063375   0.006812   9.304  < 2e-16 ***
cigsPerDay      0.017420   0.006388   2.727  0.00639 **
totChol         0.002239   0.001120   1.999  0.04558 *
sysBP           0.017472   0.003634   4.808 1.53e-06 ***
diaBP          -0.001627   0.006453  -0.252  0.80098
glucose         0.007860   0.001732   4.539 5.66e-06 ***
sexmale         0.491786   0.110431   4.453 8.46e-06 ***
smokerSmoker    0.134118   0.159044   0.843  0.39907
PrevStrokeYes   0.982953   0.468328   2.099  0.03583 *
cigsPerDay.na  -1.132774   1.042487  -1.087  0.27721
totChol.na      0.464424   0.428589   1.084  0.27854
BMI.na          0.608957   0.764368   0.797  0.42564
heartRate.na   12.342846 324.743761   0.038  0.96968
glucose.na     -0.160422   0.196723  -0.815  0.41480
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2936.5  on 3390  degrees of freedom
Residual deviance: 2602.5  on 3376  degrees of freedom
AIC: 2632.5

Number of Fisher Scoring iterations: 11
```

Figure 8: Reduced Logistic Regression Model



(a) Age

(b) Cigarettes per Day

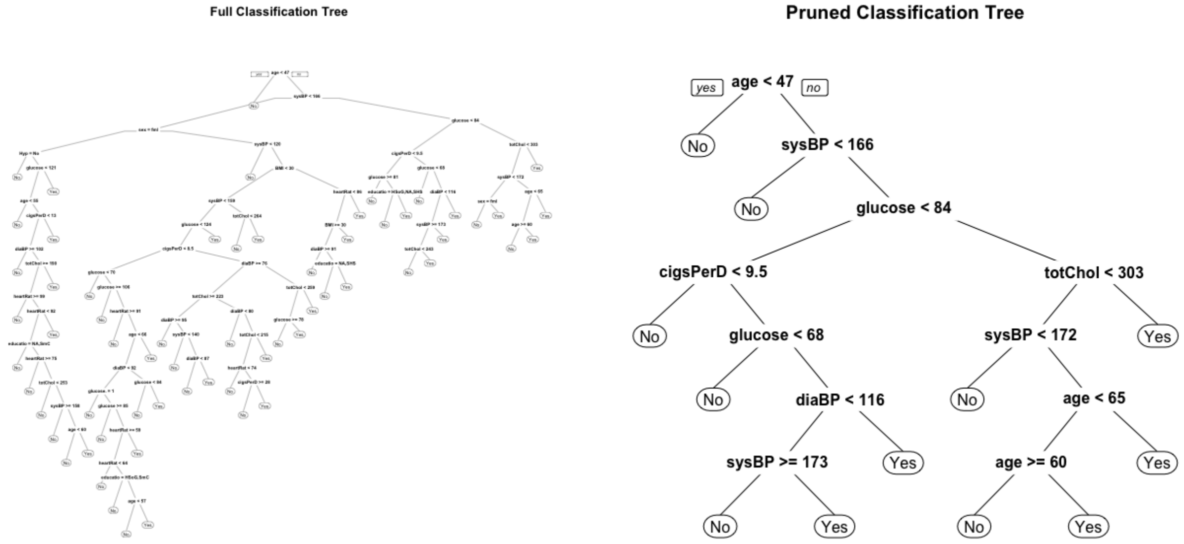
(c) Cholesterol

(d) Systolic Blood Pressure

(e) Diastolic Blood Pressure

(f) Glucose

Figure 9: Smoothed Variable Plots for Reduced GAM



(a) Full Classification Tree

(b) Pruned Classification Tree

Figure 10: Full and Pruned Classification Trees

```
         CP nsplit rel error xerror     xstd
1 0.0051985      0   1.00000 1.0000 0.039943
2 0.0037807     11   0.92439 1.0095 0.040096
3 0.0031506     26   0.86200 1.0548 0.040815
4 0.0028355     29   0.85255 1.0718 0.041078
5 0.0018904     35   0.83554 1.1021 0.041535
6 0.0012602     56   0.79017 1.1569 0.042335
7 0.0010000     62   0.78261 1.1796 0.042656
```

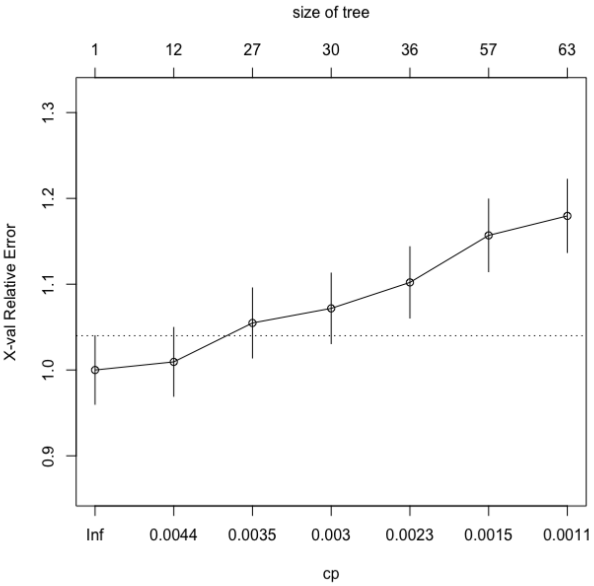


Figure 11: Cross-Validation Results for Pruning Classification Tree

	No	**Yes**	**Accuracy**
Observed	732	115	
Full Logistic Regression	832	15	0.8701
Reduced Logistic Regression	832	15	0.8678
Full GAM	832	15	0.8678
Reduced GAM	832	15	0.8678
Full Tree	763	84	0.8217
Reduced Tree	829	18	0.8548

Figure 12: Comparison of Models on Test Set