Framingham Heart Study

Josh Ye

Read data and install libraries

Introduction and Data

In this Project, we used data from the Framingham Heart Study, a long-term, ongoing cardiovascular cohor study of residents of Framingham, MA. The study began in 1948 with 5209 subjects. This dataset contains the original data, but with identifying patient information scrubbed. In the original study, individuals joined the study by accepting letters of invitation that were sent to a random sample of two of every three families, with members aged 30-59 years. Out of the 6507 original contacts, 4494 participants agreed to enter the study.

The dataset contains the following predictors:

- `male`: male or female, 1 for male, 0 for female (nomial)
- `age`: Age of the patient; (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- `currentSmoker`: whether or not the patient is a current smoker (Nominal)
- `cigsPerDay`: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- `BPmeds`: whether or not the patient was on blood pressure medication (Nominal)
- "prevalentStroke"': whether or not the patient had previously had a stroke (Nominal)
- `prevalentHyp`: whether or not the patient was hypertensive (Nominal)
- `diabetes`: whether or not the patient had diabetes (Nominal)
- `totChol`: total cholesterol level (Continuous)
- `sysBP`: systolic blood pressure (Continuous)
- `diaBP`: diastolic blood pressure (Continuous)
- `BMI`: Body Mass Index (Continuous)
- `heartRate`: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- `glucose`: glucose level (Continuous) For the purposes of the study, as well as our own analysis, the dependant variable is

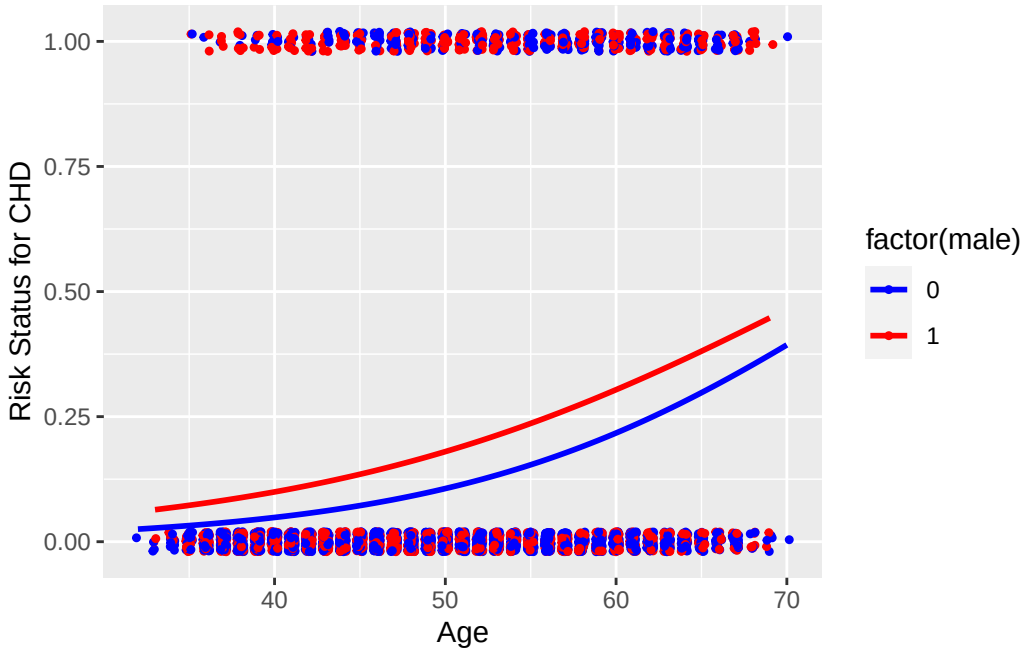
*`TenYearCHD`: 10 year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

Research Question

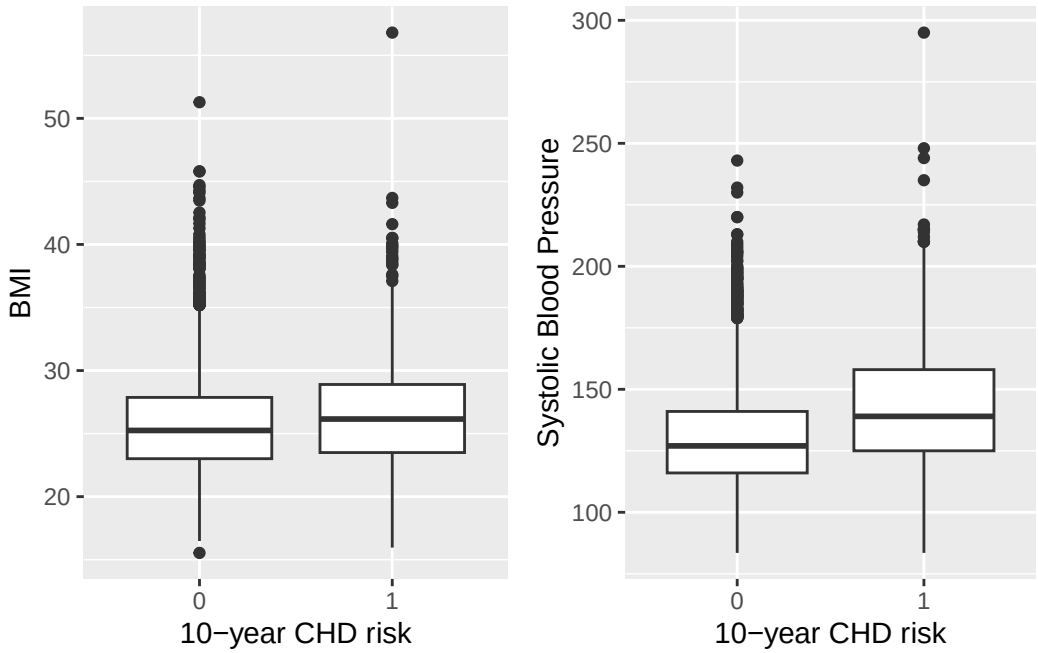
Our main research question will be to determine which factors are most important for predicting the Ten-Year risk level for Coronary Heart Disease based on the data provided in this dataset. We will investigate model selection and variable selection techniques, then we will thoroughly investigate the model that was produced using our chosen techniques.

Exploratory Data Analysis

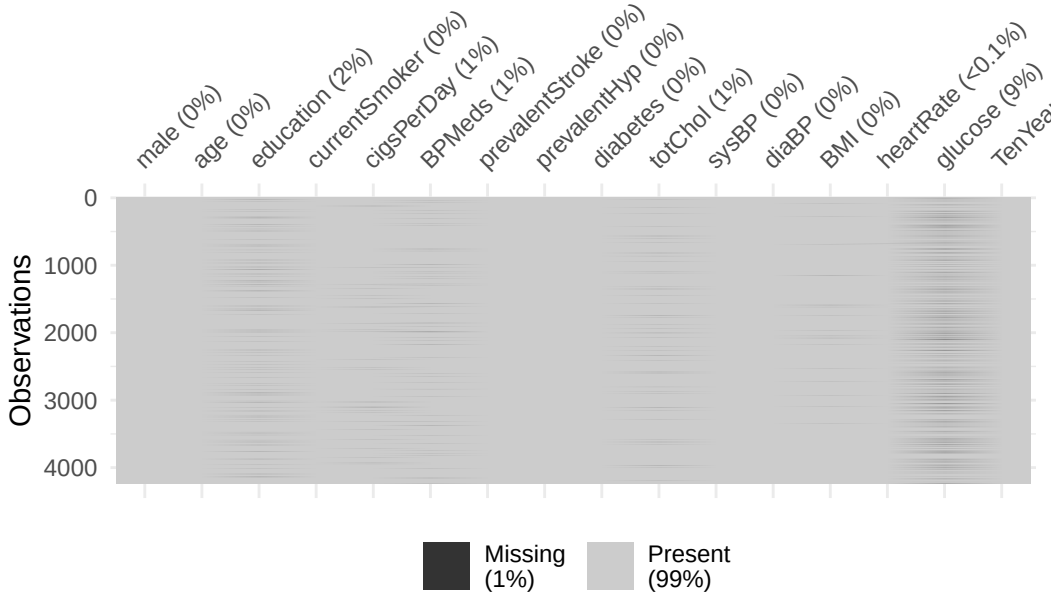
First, we take a look at the ages of individuals at risk for CHD and those not at risk. We notice that those who are at risk are older than those not at risk.



Additionally, we might assume that those who have a high BMI are generally more at risk for CHD. However, we see that there may not be such a strong statistical relationship. While the median BMI for the "at risk" category is marginally higher than the "no risk" category, there are a large large amount of outliers (on the high end) for the "no risk" group. On the other hand, the opposite is true for Systolic Blood Pressure, where both the median and outliers (on the high end) are higher for the "at risk" group than the "no risk" group.



Furthermore, we investigate missingness in our model.



Based on the graph above, we can see that the `glucose` predictor had the highest missing data rate. While this could possibly imply that there is some issue specific to measuring glucose

levels which causes the data to be missing, this is hard to determine conclusively. Therefore, although it is possible that the data is MNAR, we will *not* impute data and simply continue our analysis by removing the missing data.

No additional data cleaning (besides removing rows with missing data) was performed, since the column names were sufficient and the data types were ok.

Methodology

Since our dependent variable, `TenYearCHD` is binary, this is a binary classification problem and a *logistic regression* model is best suited for this type of analysis. Furthermore, we have a few highly correlated variables, namely `diaBP` and `sysBP` ($r = 0.787$), and `cigsPerDay` and `currentSmoker` ($r = 0.774$).

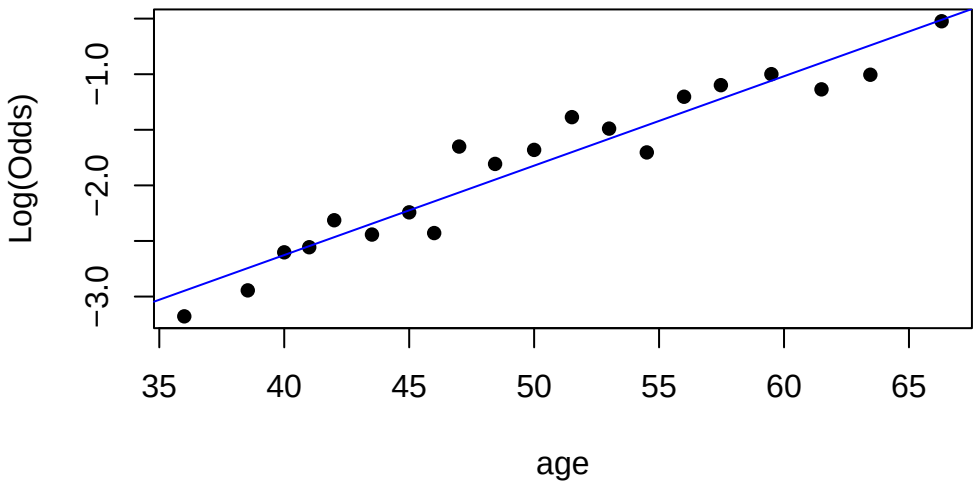
The method that we utilized for variable selection was stepwise backward AIC. There were two reasons why this method was chosen. Firstly, the LASSO variable selection process, with a lambda produced by k-fold cross validation, did not produce meaningfully better results, especially because the cross-validation process produced an extremely low λ value. Secondly, by inspecting the results of backward AIC satisfactorily removing predictors that had little statistical significance in the model. Additionally, the backward AIC results did not contain any pairs of highly correlated predictors.

The results of our backward AIC selection process is the following model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.739521355	0.522563322	-16.724330	8.715099e-63
male	0.553152392	0.107037102	5.167857	2.367939e-07
age	0.065337127	0.006443762	10.139594	3.686274e-24
cigsPerDay	0.019574222	0.004181536	4.681108	2.853283e-06
prevalentStroke	0.751411972	0.483562132	1.553910	1.202059e-01
prevalentHyp	0.226230801	0.135098080	1.674567	9.401920e-02
totChol	0.002247965	0.001122375	2.002865	4.519175e-02
sysBP	0.014219087	0.002857258	4.976479	6.475119e-07
glucose	0.007314476	0.001672662	4.372954	1.225766e-05

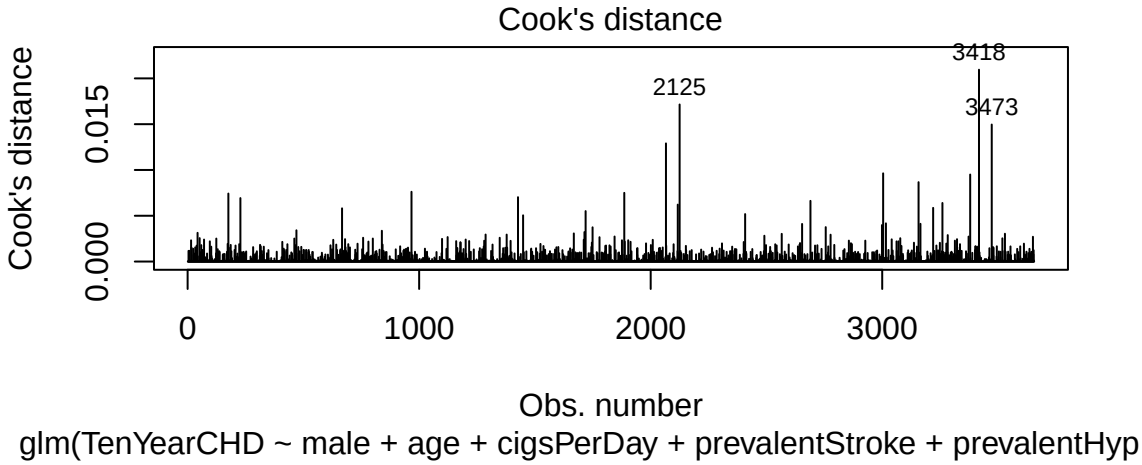
As we can see, while this refined model contains only eight predictors, most of them are statistically significant, while most of the statistically *insignificant* predictors have been dropped.

Now we verify that our data satisfies the assumptions for the Logistic Regression Model. Logistic Regression models must satisfy two assumptions. The first is Linearity. We use an empirical logit plot to demonstrate this - if there are a roughly equal number of points on both sides of the line (for the continuous predictors), as they are below, then we consider this requirement satisfied:



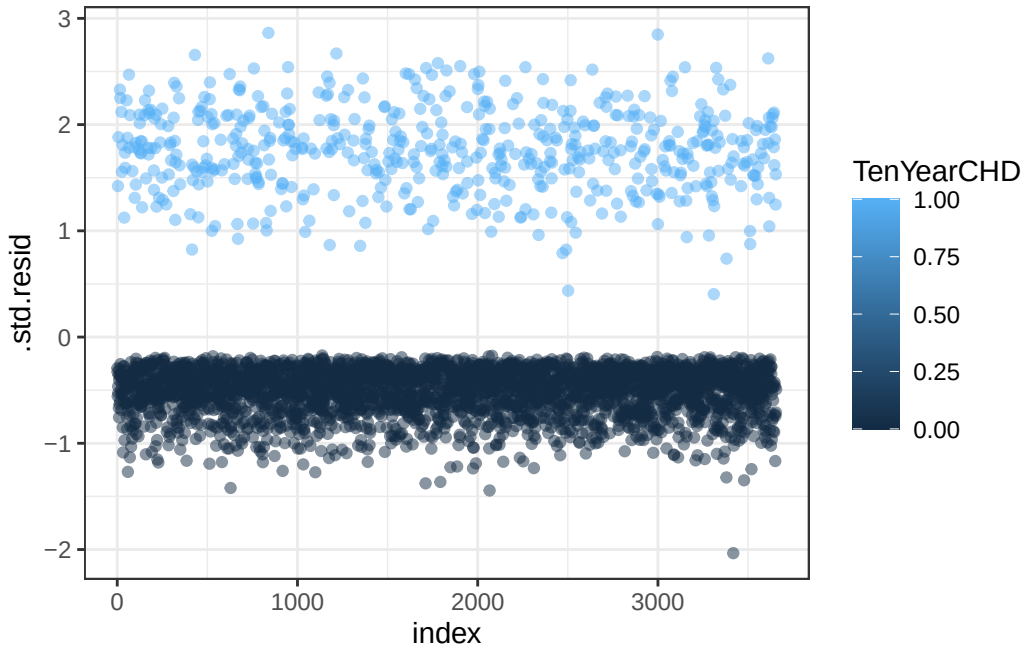
Furthermore, we must check the independence assumption. While all samples were taken within the town of Framingham, MA, every individual did not live in the same community; instead, letters were sent out randomly inviting famlies to participate in the study. Since the sampling was relatively random, for our purposes we may say that the independence assumption is satisfied.

Now, we search for influential values by Cook's distance.



We note that there are 3 high Cook's Distance values which we want to investigate:

However, after plotting the standard residuals, we conclude that there are **no influential values** as none of the high Cook's Distance values had standard residuals greater than `abs(3)`.



Furthermore, we can notice here that individuals who are *not* at risk for Coronary Heart Disease in a ten year span have standard residuals that are clustered close together. On average, there is about a 1 standard residual gap between the two groups.

Results

Now, we perform some tests on the tests on how well we predicted. Since this is a logistic regression model, in order to interpret our model effectively, we convert the log-odds to an odds ratio. First, we use a threshold of 0.5, where a model prediction of 0.5 or above implies that the individual in question is predicted to be at risk for CHD within 10 years. In context of our investigation, the Binary Classifiers that we care most about are True Positives, True Negatives, False Positives, and False Negatives. However, we would especially like to minimize false negatives, that is, individuals who *truly* are at risk of developing CHD in the next 10 years but are missed by the model. Thus, for a threshold value of 0.5 we see that there were 512 false negatives, which is a lot.

```
   0    1
```

```
  At Risk       18   45
  Not At Risk 3081  512
```

This implies that we should lower our threshold in order to have a more accurate assessment, even if our false positive rate goes up a bit. If we try evaluate our model with a threshold value of 0.1 for the Ten year CHD risk, then we see

```
                0    1
  At Risk     1629  469
  Not At Risk 1470   88
```

that the number of false negatives drops down significantly.

Finally, we conclude that our model predicts fairly well, as the area under the ROC curve is `0.737`.

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.737
```

We can interpret some of our slope coefficients and answer our research question. Since our model is in log-odds, we must exponentiation the slopes in order to get the odds ratio. For instance, exponentiating the coefficient for `male`, we note that holding all other variables constant, males have $\exp(0.553) \equiv 1.738$ times greater odds than women for acquiring Coronary Heart Disease over a 10 year span. Furthermore, in addition to age, the other categorical variables, `prevalentStroke` and `prevalentHyp` (even though they are statistically insignificant at the $\alpha = 0.05$ significance level) have a very large effect on the log-odds of being at risk for Coronary Heart Disease over a 10 year span. For instance, while every additional cigarette that is smoked per day increases your odds of being at risk for Coronary Heart Disease by about 1.02 times, having had a stroke in the past increases your odds of being at risk for CHD 2.12 times!

Discussion

From our model, the being at risk for acquiring Coronary Heart Disease over a ten year period is most strongly statistically associated with (based on p-value) being male, being older, smoking cigarettes daily, having high cholesterol, having high Systolic Blood Pressure, and having high blood sugar (glucose). Although we have showed that this model has relatively strong predictive power, there are still some possible limitations of our analysis. For instance,

the coefficients were selected using backward stepwise AIC selection. Since this is a greedy algorithm, and not exhaustive, it is possible that there are better sets of predictors that were missed. Furthermore, there are also possible concerns retarding the source of data itself. Because the data was taken from one city in Massachussetts, it is possible that the lifestyle and demographic of the inidivudals of that Framington are not representative of Americans or humans as a whole, but are only locally representative of that area.

There are many possibilities for future work. Firstly, more predictors could be included. These could include other lifestyle factors, such as exercise duration, average weekly vegetable consumption, etc. that could give more insight into potential factors that could actually be associated with a lowered risk of developing Heart Disease. Secondly, the study could be expanded to different regions of the united states, to see if cultural factors, weather, or other factors play any significant role in the development of heart disease.

References and Links

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5156338/
2. https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression/download?datasetVersionNumber=1