Framingham Heart Study

Josh Ye

Read data and install libraries

Project Introduction and Predictors

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:

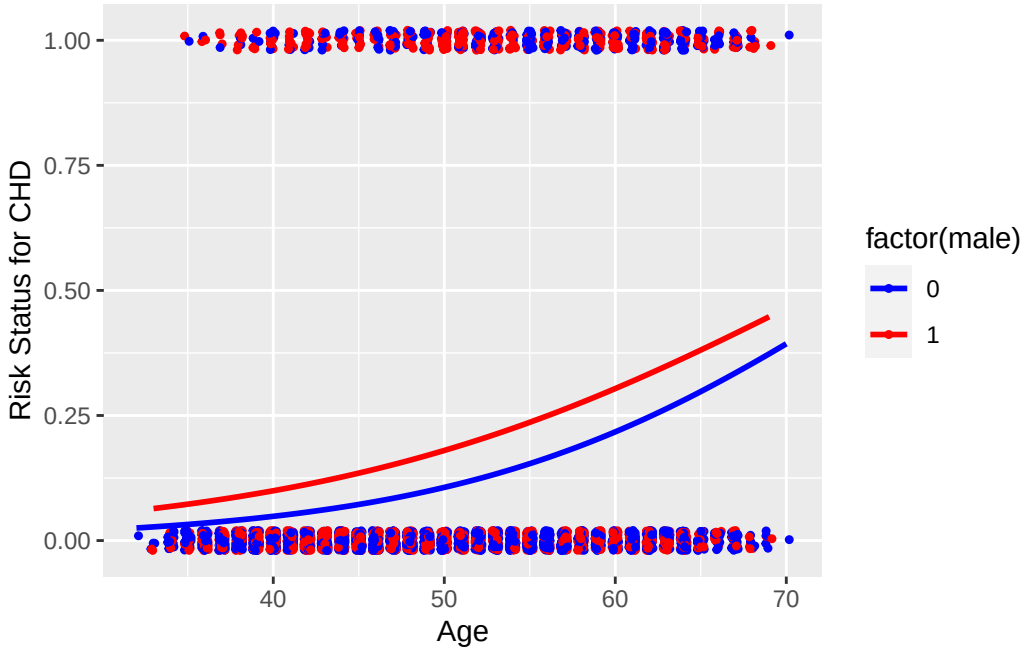
- `male`: male or female, 1 for male, 0 for female (nomial)
- `age`: Age of the patient; (Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- `currentSmoker`: whether or not the patient is a current smoker (Nominal)
- `cigsPerDay`: the number 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

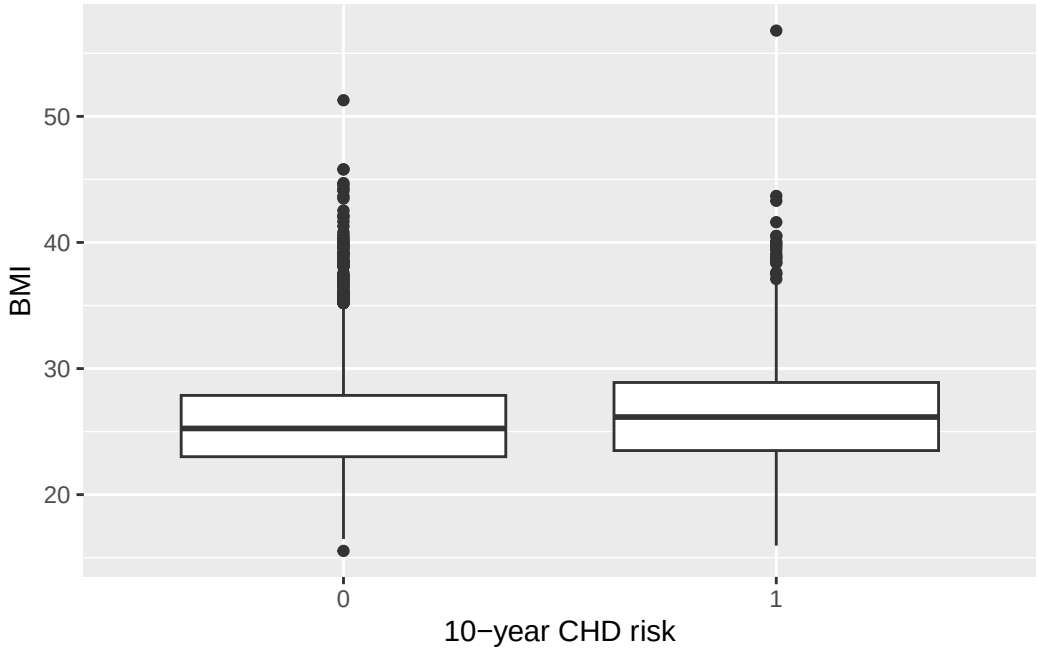
*`TenYearCHD`: 10 year risk of coronary heart disease CHD (binary: "1", means "Yes", "0" means "No")

Exploratory Data Analysis

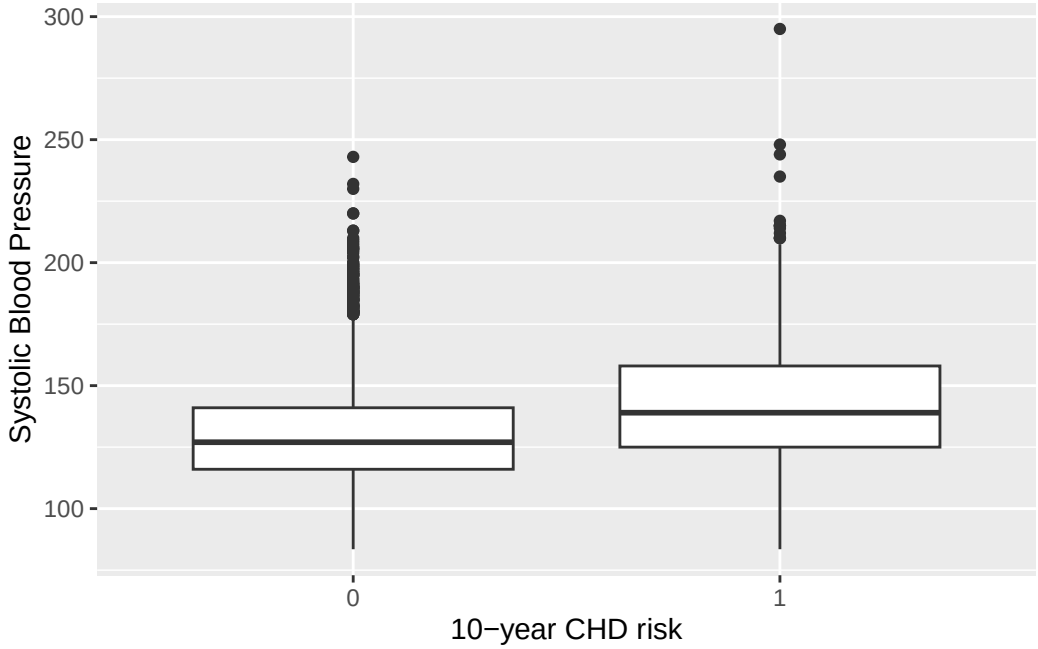
First, we take a look at the ages of individuals at risk for CHD and those not at risk. We notice that those who are at risk are older than those not at risk.



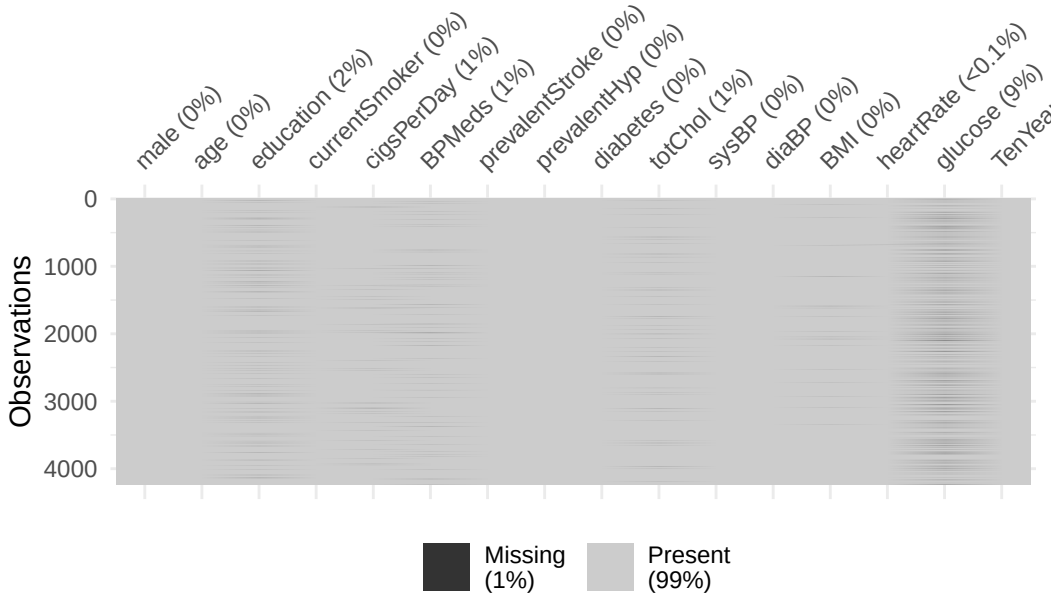
Additionally, we might assume that those who have a high BMI are generally more at risk for CHD. However, we see that there may not be such a strong statistical relationship. While the median BMI for the "at risk" category is marginally higher than the "no risk" category, there are a large large amount 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.

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.

Variable Selection and Assumption Testing

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

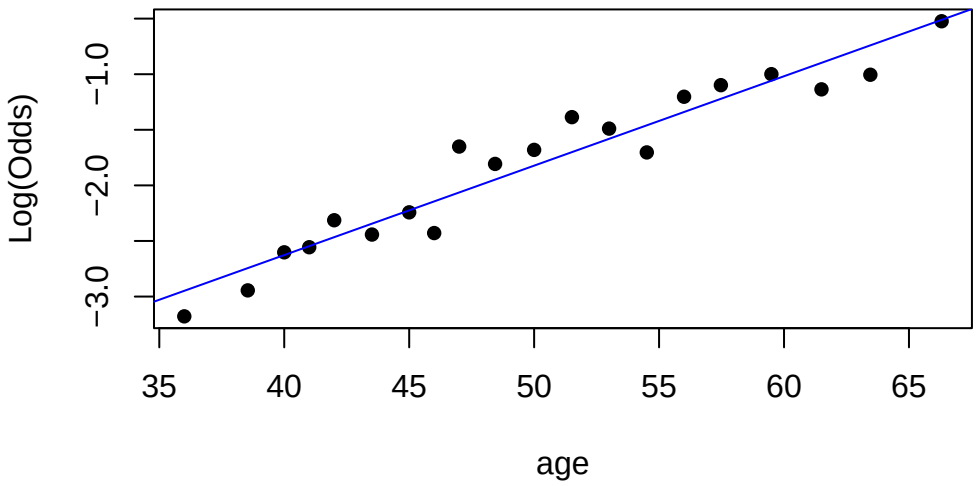
a lambda produced by k-fold cross validation, did not produce meaningfully better results, especially because the cross-validation process produced an extremely low λ value. Secondly, by inspecting the results of backward AIC satisfactorily removing predictors that had little statistical significance in the model. Additionally, the backward AIC results did not contain any pairs of highly correlated predictors.

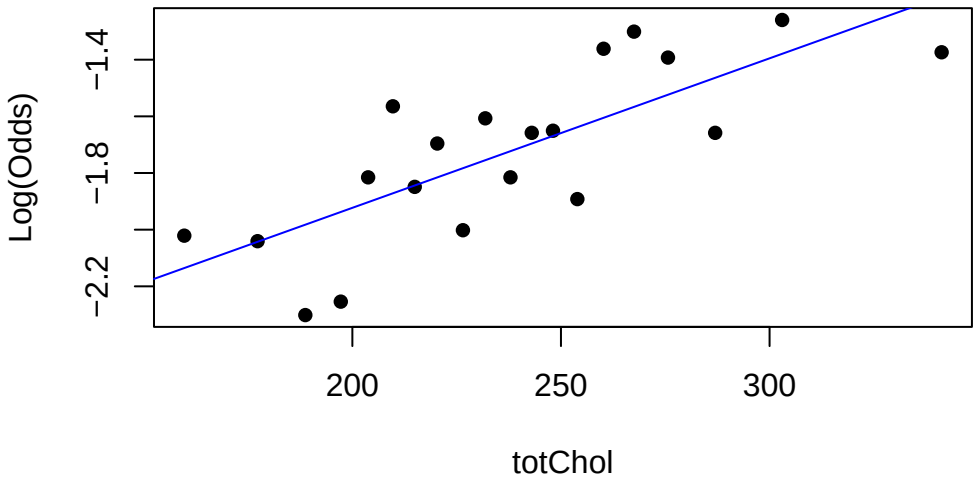
The results of our backward AIC selection process is the following model:

	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:

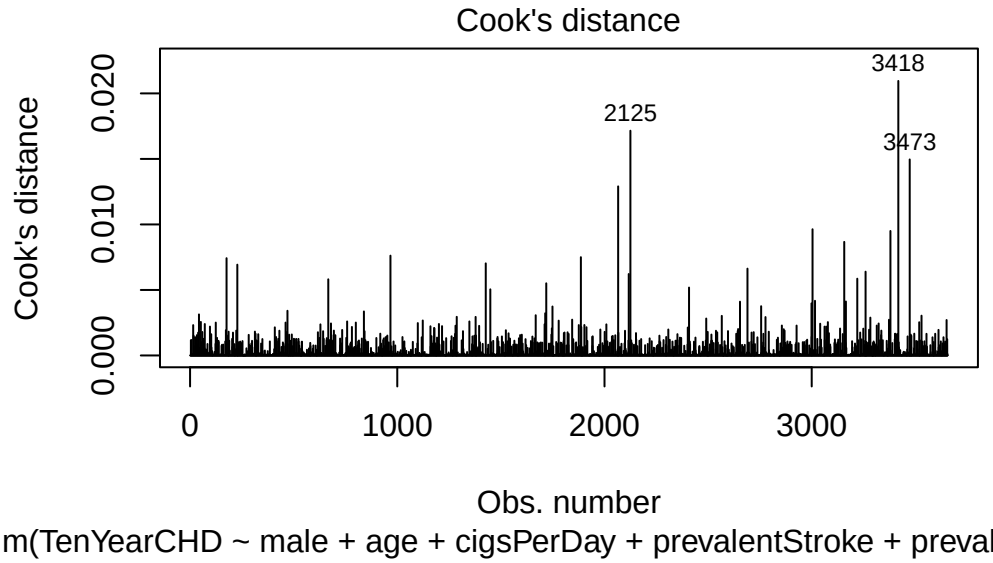




Furthermore, we must check the independence assumption. While all samples were taken within the town of Framingham, MA, every individual did not live in the same community; instead, letters were sent out randomly inviting famlies to participate in the study. Since the sampling was relatively random,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

assumption is satisfied.

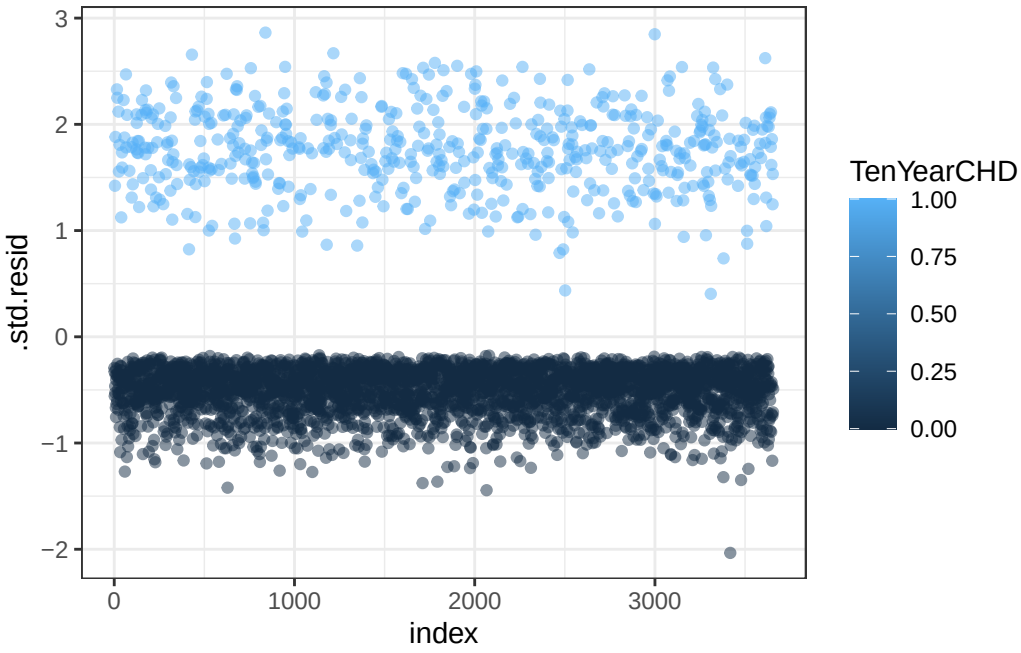
Now, we search for influential values by Cook's distance.



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

```
# A tibble: 3 x 16
  TenYearCHD  male   age cigsPer~1 preva~2 preva~3 totChol sysBP glucose .fitted
       <dbl> <dbl> <dbl>     <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>
1          1     1    52         0       1       0     202   136      67  -1.16
2          0     1    64         0       0       1     195   176     370   1.87
3          1     0    52         5       1       1     205   159      83  -0.938
# ... with 6 more variables: .resid <dbl>, .std.resid <dbl>, .hat <dbl>,
#   .sigma <dbl>, .cooksd <dbl>, index <int>, and abbreviated variable names
#   1: cigsPerDay, 2: prevalentStroke, 3: prevalentHyp
```

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)`.



Key Results and Findings

Now, we perform some tests on the tests on how well we predicted.