# Week 6 – Logistic, Multinomial and Polynomial Regression

By: Zach Adair

Regis University

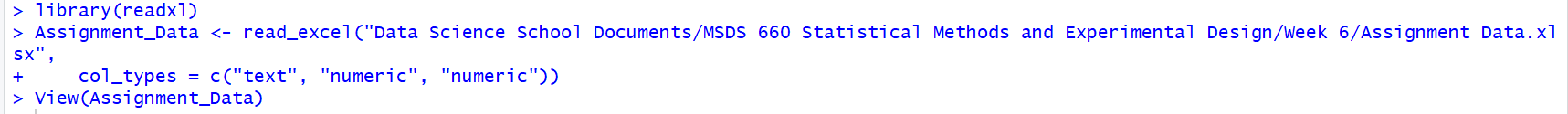
MSDS – 660 Statistical Methods and Experimental Design

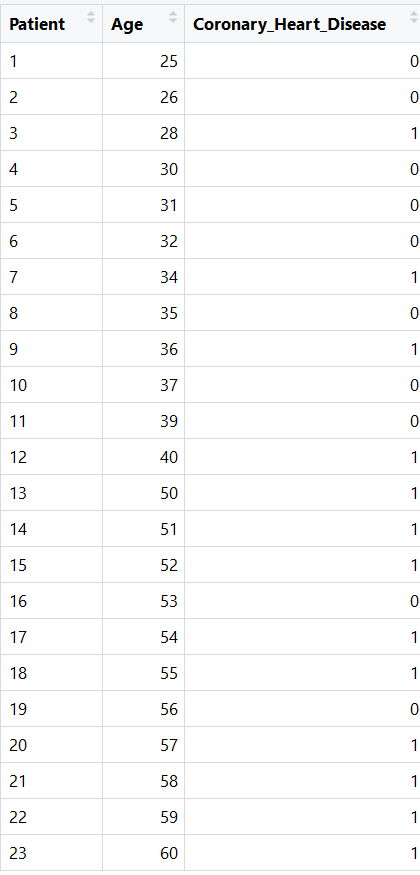
## Introduction

This week’s assignment involves studying the relationship between age and the presence/absence of coronary heart disease. Throughout the assignment I will be using a mix of regression techniques to analyze the data given a significance level of 5% and finally construct a predictor model to help show what is the potential someone may have coronary heart disease.

## Load the Data

The first thing we must always do before any analysis is get our data and load it into the application we are using. Since we are using RStudio again I will load the data into RStudio. After importing the data from Microsoft Excel and into RStudio our result is so:





Now that my data is loaded I can begin my analysis.

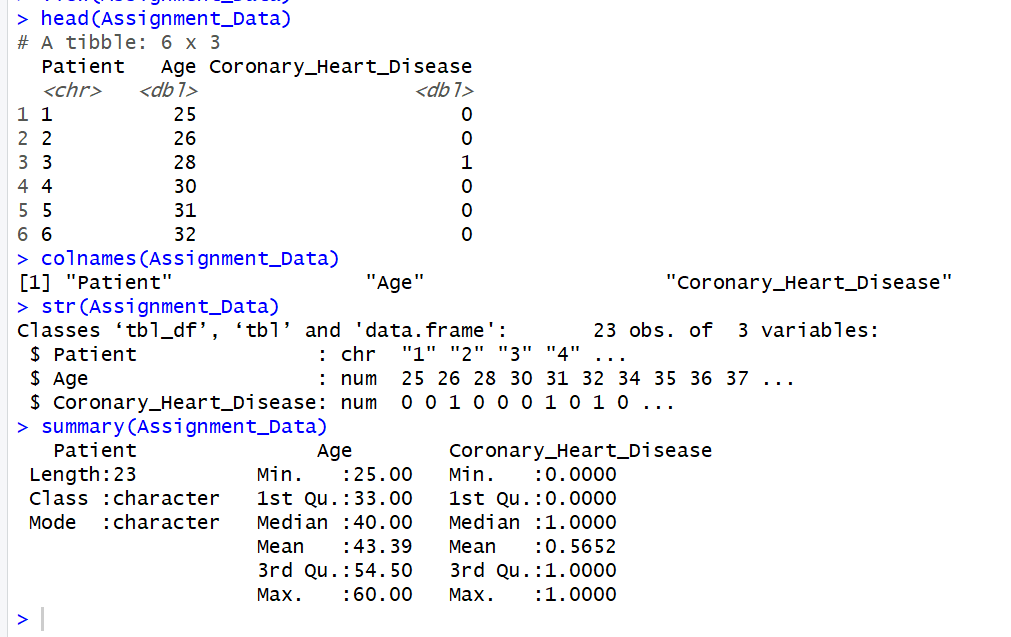
## State Hypothesis and other Assumptions

Null Hypothesis: The coefficient of age has no effect on if the patient has Coronary Heart Disease.

Alternative Hypothesis: The coefficient of age has an effect on if the patient has Coronary Heart Disease.

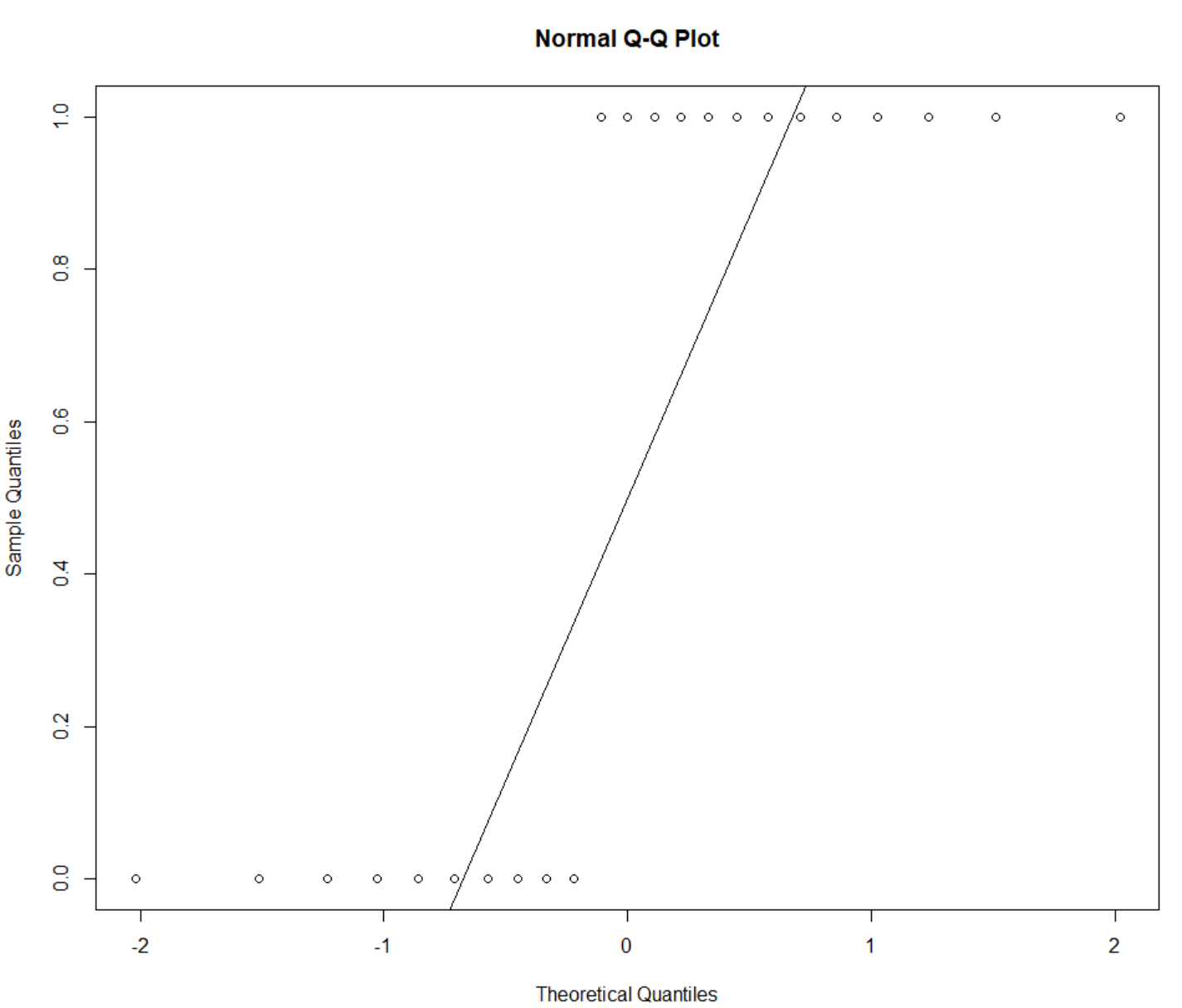
## Analyzing the Data

We will start with getting comfortable with the data through some summary analysis.

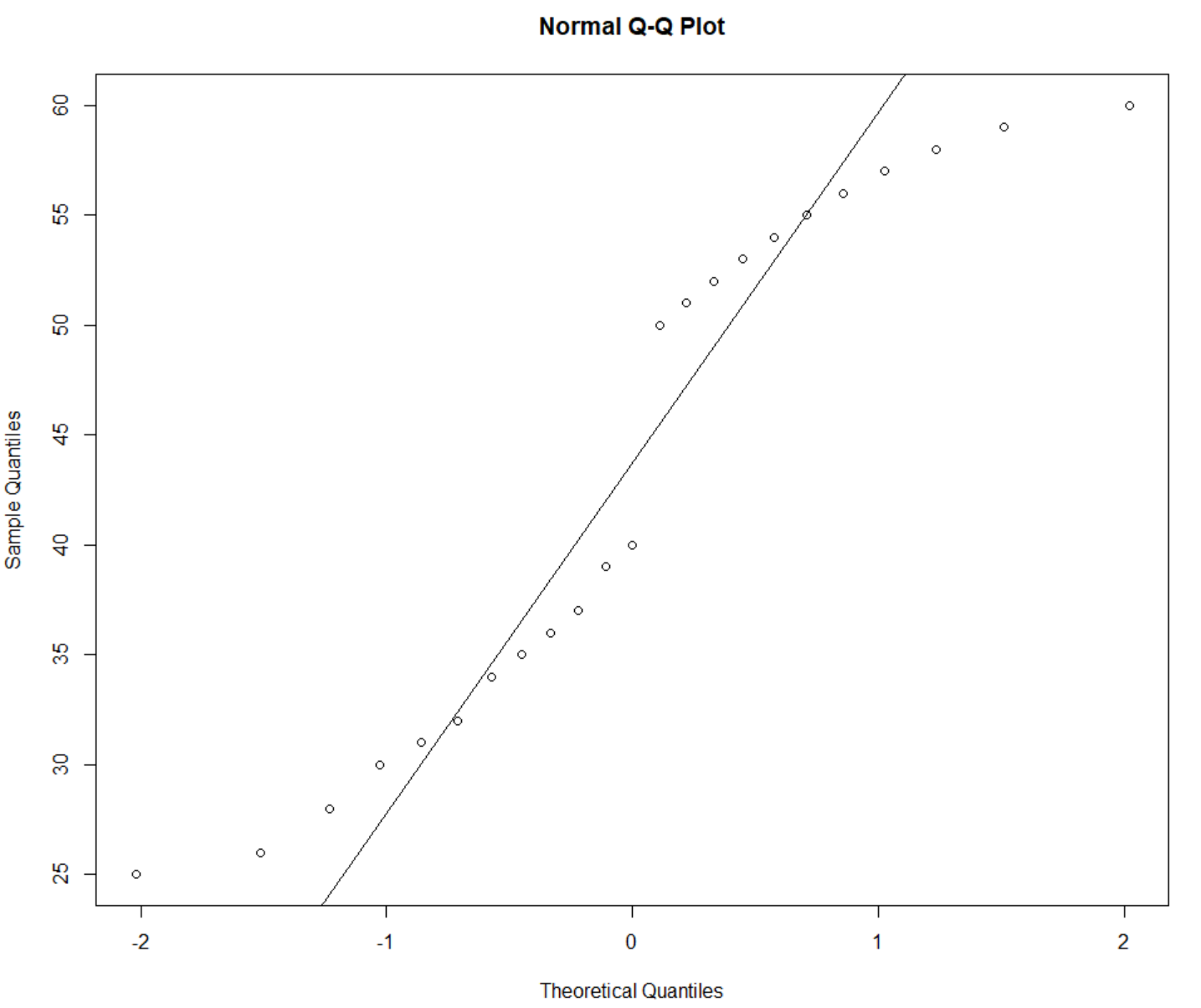


After going through summary statistics to learn about the data involved in this analysis it is pretty difficult to make any real assumptions about the data. It appears based on the mean of Coronary Heart Disease that there might be a couple more patients diagnosed positively than not. The only other bit of information about the data that I can see is that the average age of the patients is in the early 40s. Next I’m going to plot Coronary Heart Disease and age.

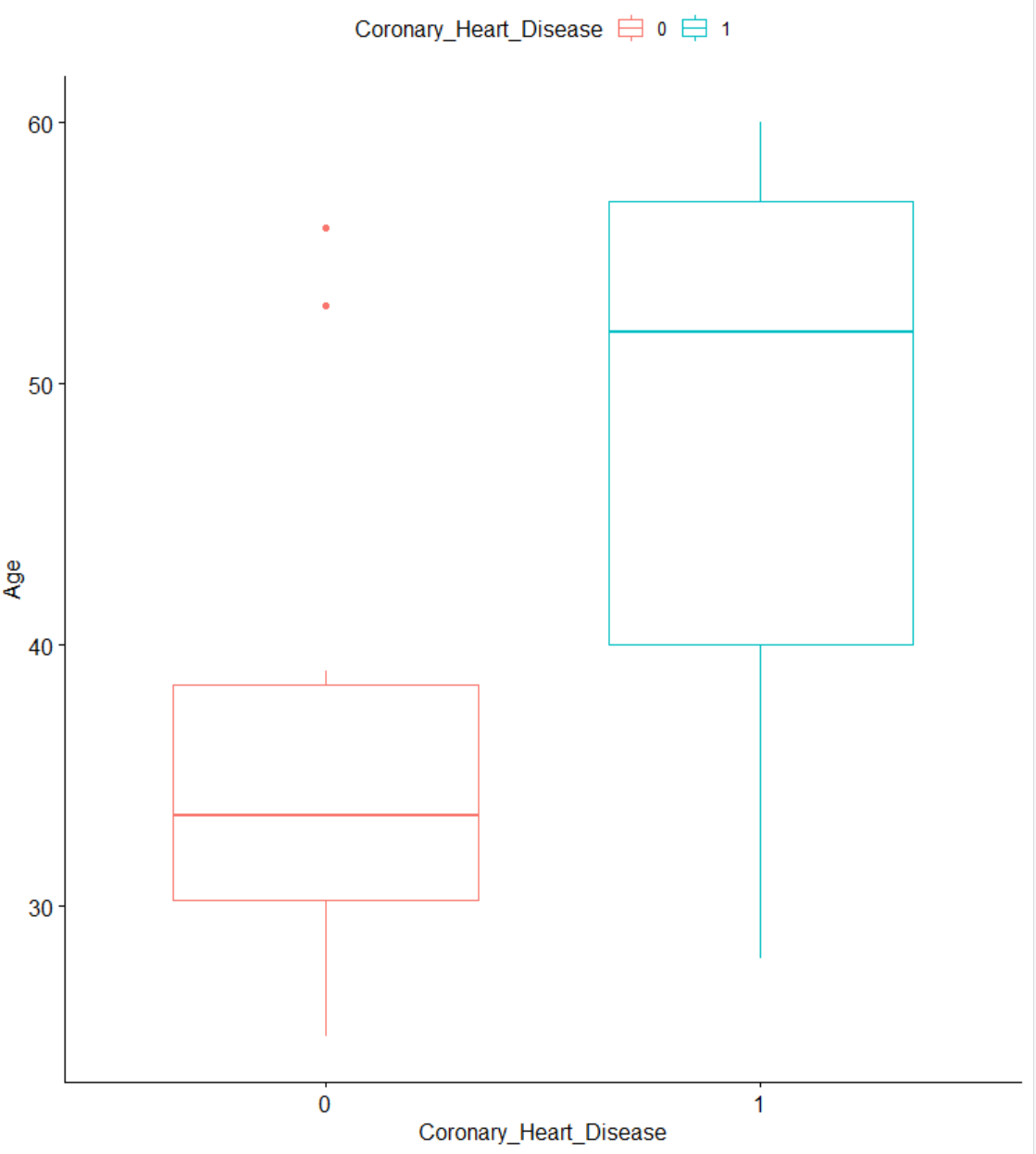


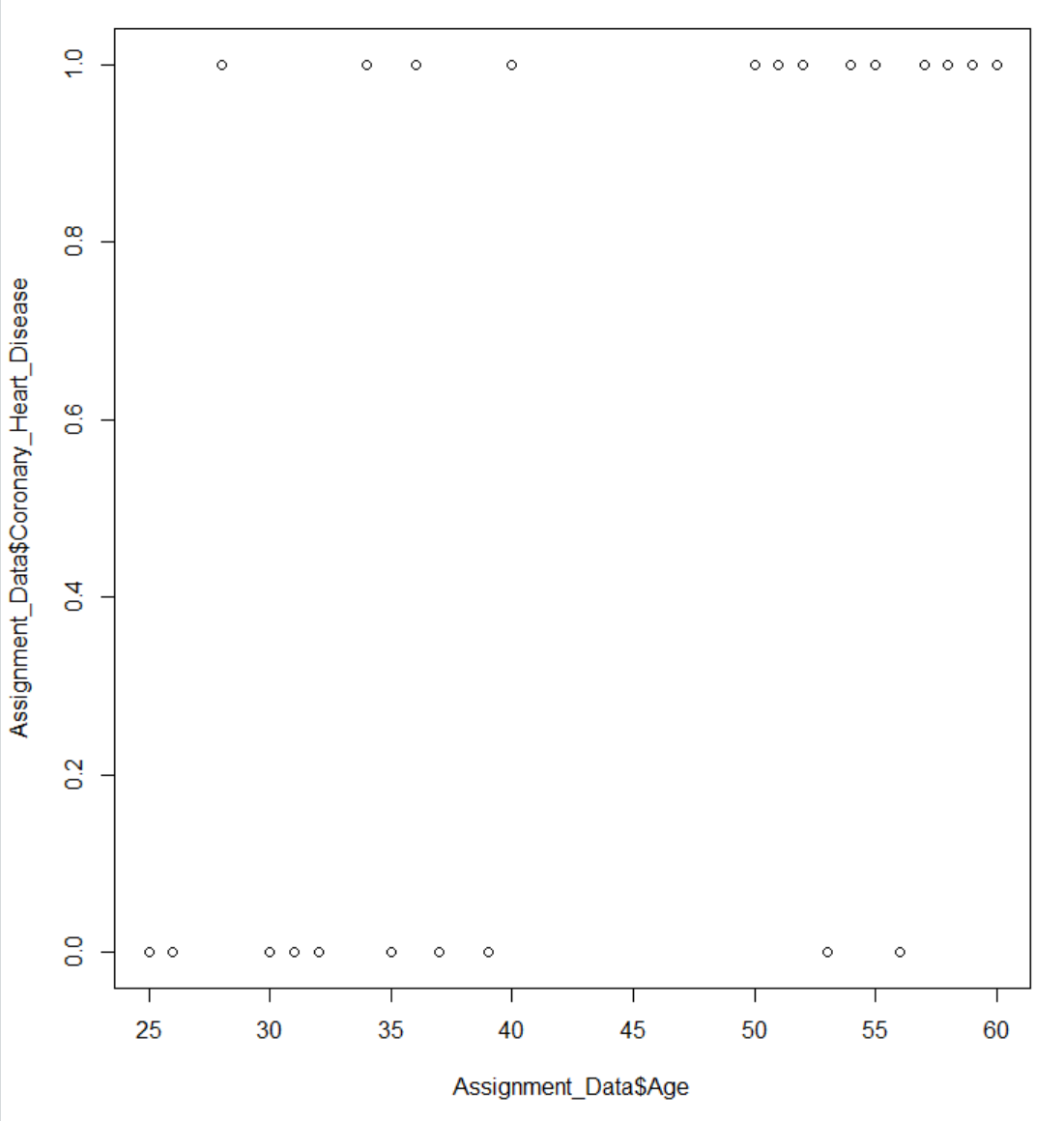






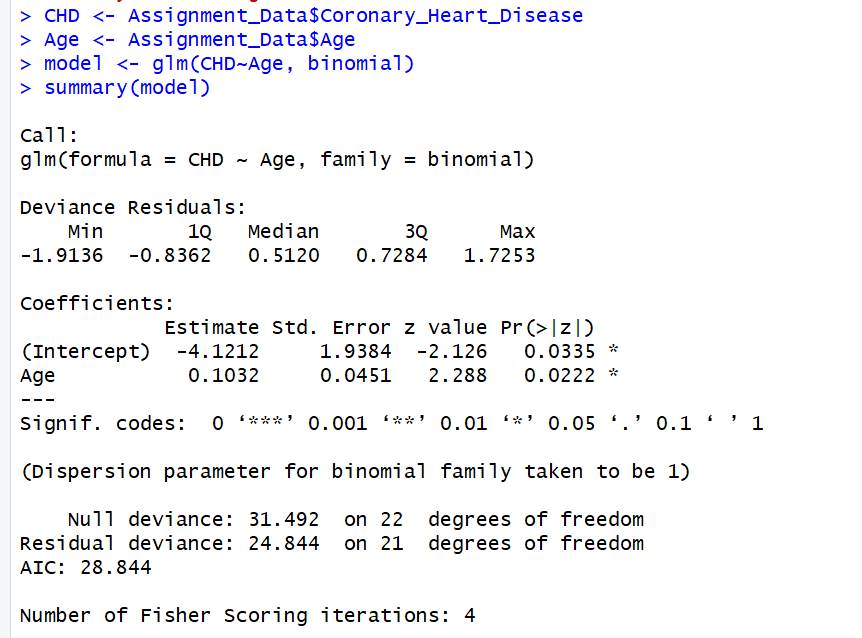






From the plots above I can make some assumptions about the data and how it is distributed. From what I’m seeing, it appears that there does seem to be some sort of correlation between the higher the age, the more likely of a patient having Coronary Heart Disease. We have records where that isn’t true but it appears a majority of cases where he patient is over the age of 40 ended up having Coronary Heart Disease.

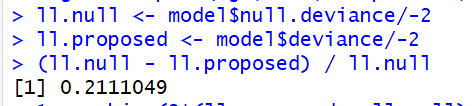
After looking at the data through simple plots I can now try find more evidence of my assumption that I’m starting to grow which is that age is a coefficient that effects if the patient has Coronary Heart Disease.



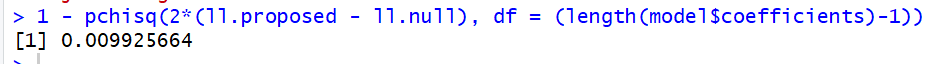
Before creating my model and running a summary of it, I assigned the fields of Age and Coronary Heart Disease to shorten the code in the model.

From the result, the p-value of Age was less than 0.05 so in this situation we would conclude that we should reject the null hypothesis which stated that the coefficient of age didn’t have any effect on the patient’s chances of having Coronary Heart Disease.

After diagnosing from our logistic model that the null hypothesis should be rejected because age is a coefficient that effects Coronary Heart Disease, I would like to calculate the McFadden’s Pseudo R^2 value.

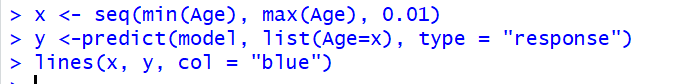


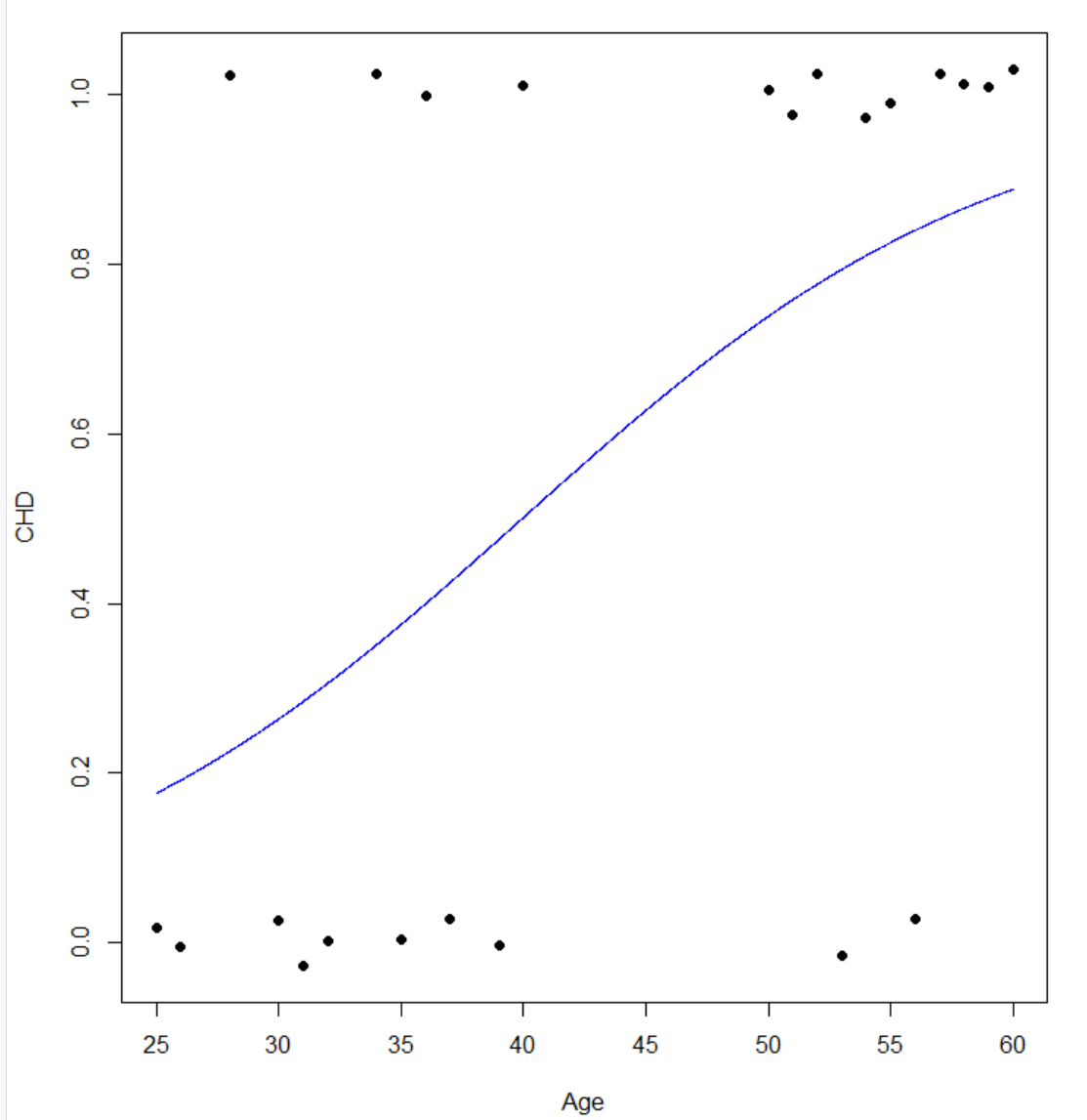
Setting up to calculate McFadden’s Pseudo R^2, we first pull the log-likelihood of the null model out of the logistic model, next we pull our log-likelihood from a proposed model of the data. We use those to form a calculation of the Pseudo R^2, our total is about 0.211 which is the overall effect size, next I will use the p-values for that R^2 to calculate a Chi-square distribution.



From the result of the chi-square model our p-value is under 0.05, thus our R^2 value is validated.

The next thing we would do is try and create a predictive model of the data and try to predict the probability of a random age and the potential of Coronary Heart Disease.

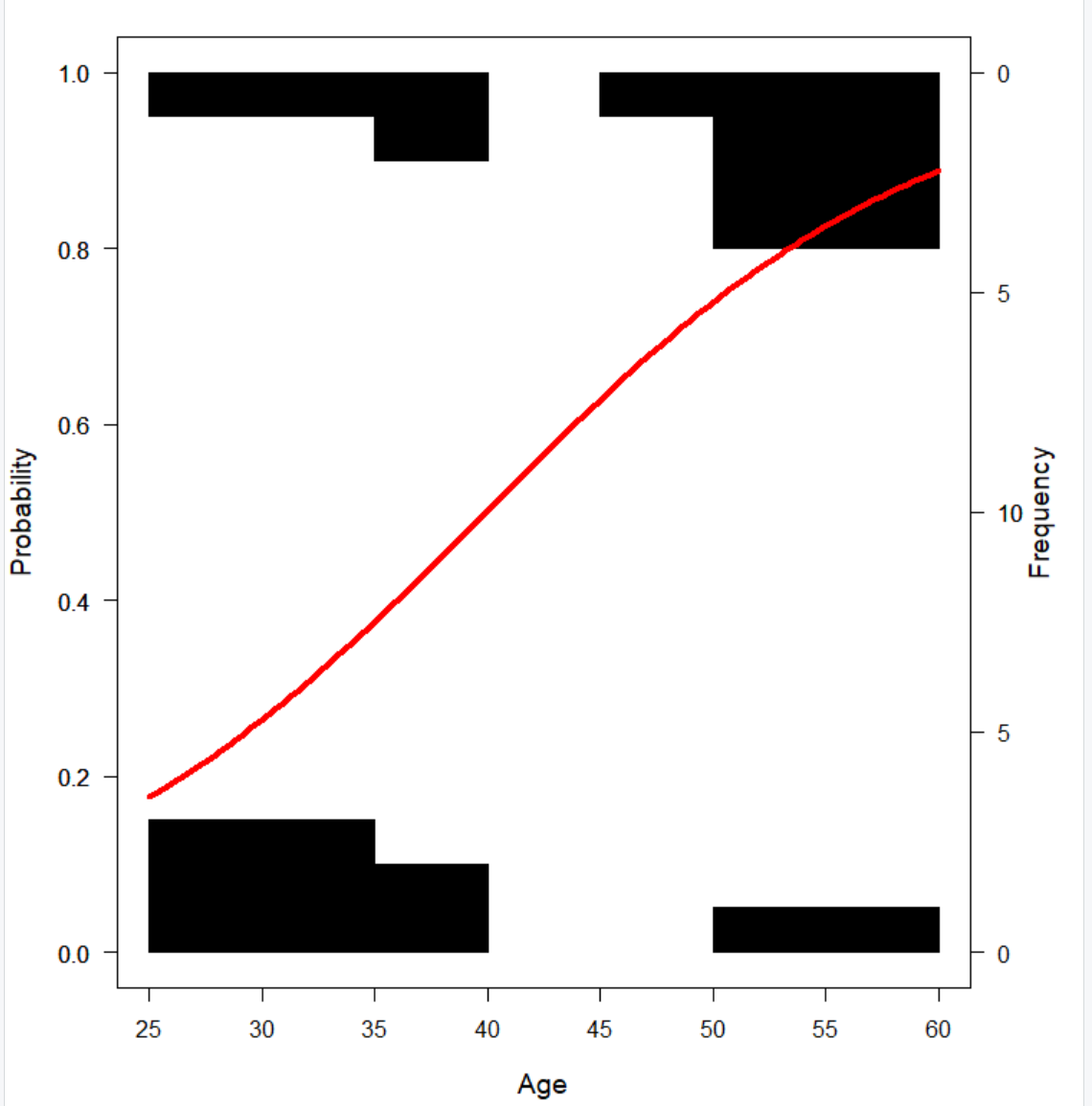




Above I created an x value to use and y is to predict that x value along our data and then a plot of that predictor in a line along our actual data points. From our curve what we are seeing is what is generated from our predictor model as the age of the patient increases, so does the chance that the patient has Coronary Heart Disease.

Now, I’m going to look at our predictor along with the distribution of the data to see the look of our data and how it is distributed as well as how that predictor line looks when plotted with that.





From our plot, even though we are only looking at 23 total patients we can distinguish right away that our highest number of patients come from either the low end of the age spectrum or the high end, and when you look at our predictor model we see a predictor that coincides with our actual distribution of our data.

## Final Conclusions and Concerns

From the logistic analysis ran on the data it was pretty simple to conclude that the coefficient of age has an effect on if the patient has Coronary Heart Disease. With what we saw originally from the simple statistics as well as the original plotting of the data it appeared to skew to that result, but after running a logistic model and backing that up with our predictor model it is evident.

The concern I do have about the data and conclusions I would like to be able to see this data with more data points and other variables as well. I think its definitive that the age and Coronary Heart Disease have a relationship, but I would like to see if there are other factors which are stronger, like weight or how active the patient is and if those kind of factors are bigger contributors.