Heart Disease Analysis Project
By: Yolanda Lewis

Introduction

Cardiovascular diseases (CVDs) are a group of disorders that affect the heart and blood vessels. These diseases can include coronary heart disease (CHD), congenital heart disease, and pulmonary embolism. CVDs are known to be one of the most common causes of death globally. People at high risk for CVDs often exhibit behavioral and health risk factors such as smoking, unhealthy diet, obesity, physical inactivity, hypertension, high cholesterol, and diabetes. According to the National Center for Biotechnology Information (NCBI), hypertension is one of the highest risk factors for many if not all CVDs. Identifying people at a high risk of CVDs can help to prolong life and ensure people receive the necessary medical treatment.

In this report, I will focus on how the presence of hypertension testing improves the predictability of total cholesterol levels. More specifically, I will use multiple regression with second order models to try to improve the overall accuracy of the first order model presented in milestone 7 for predicting total cholesterol. While it does not appear that the variables in the data set along with hypertension test results are good predictors of total cholesterol, building a second order model may improve the accuracy of the model.

Data

This specific data set, acquired from Kaggle, includes 3,612 residents with 13 variables. Three variables including sex, currentSmoker, and education were previously removed from the dataset as these variables will be used by other members of the group project to explore the dataset The dataset contains 12 explanatory variables which include age, cigPerDay, BPMeds, prevalentStroke, prevalentHyp, sysBP, diaBP, BMI, glucose, diabetes, and heartRate. Total cholesterol will be used as the response variable in this report. Outliers from the total cholesterol variable were removed so that the datapoints considered as outliers do not influence the analysis due to extremely high or low values.

Model Building Process

Based on my previous milestone, I found that certain variables were not useful in the model for predicting cholesterol. The goal of this report is to test whether the presence of hypertension test results increase the predictability of total cholesterol levels. The variables previously removed included glucose, prevalentStroke, diabetes, and TenYearCHD. These variables were removed because the individual t-tests did not show that the estimations for the variables would be useful in the model. See the final model below that was presented in milestone 7.

Final First Order Model

```
> MyDataset_model5 <- lm(totChol ~ Age + cigsPerDay + BPMeds + prevalentHyp +
+                          sysBP + diaBP + BMI + heartRate,  data = totChol_outliers_removed)
> summary(MyDataset_model5)

Call:
lm(formula = totChol ~ Age + cigsPerDay + BPMeds + prevalentHyp +
    sysBP + diaBP + BMI + heartRate, data = totChol_outliers_removed)

Residuals:
    Min      1Q  Median      3Q     Max
-133.72  -27.02   -1.93   25.24  119.17

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.49989    8.11570  14.478  < 2e-16 ***
Age           1.15819    0.08490  13.642  < 2e-16 ***
cigsPerDay    0.10098    0.05548   1.820 0.068842 .
BPMeds        6.77255    4.01611   1.686 0.091816 .
prevalentHyp -2.37886    2.00319  -1.188 0.235094
sysBP         0.07944    0.05631   1.411 0.158355
diaBP         0.25279    0.09275   2.725 0.006453 **
BMI           0.58553    0.17402   3.365 0.000774 ***
heartRate     0.17686    0.05554   3.185 0.001462 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.94 on 3603 degrees of freedom
Multiple R-squared:  0.09672,   Adjusted R-squared:  0.09471
F-statistic: 48.22 on 8 and 3603 DF,  p-value: < 2.2e-16

> vif(MyDataset_model5)
      Age cigsPerDay      BPMeds prevalentHyp       sysBP       diaBP         BMI
 1.257948    1.046363    1.094301    2.038597    3.643794    2.937001    1.190227
   heartRate
 1.049646
```

The final first order regression model uses total cholesterol as the response variable and contains 8 explanatory variables which include age, cigPerDay, BPMeds, prevalentHyp, sysBP, diaBP, BMI, and heartRate. The F-test in this model shows a p-value of 2.2e-16, since the p-value is low, I would reject the null hypothesis and accept the alternative, that at least one of the betas is not equal to zero. The adjusted R^2 explains that 9.5% of the variability in total cholesterol is explained by the model. Even though the t-tests for prevalent hypertension and systolic BP were not good, they were left in the model because the goal to determine how hypertension test results affects the predictability of total cholesterol levels for the patients in the dataset. Removing systolic blood pressure from the model worsened the t-test for hypertension and did not improve the overall model.

To build a second order model for this dataset, I first performed a residual analysis with the variables that were included in the final first order model to check the residual plots against the independent variables and predicted values. The residual analysis is done to look for trends(or patterns) within the data, changes in variability, and to determine data points that maybe outside two standard deviations from the regression line.

To observe the normality of the data, I looked at a histogram and QQ plot of the dataset residuals. The histogram in figure 1 shows the distribution of the residuals from a linear model of total cholesterol. The distribution is fairly- normal with a few outliers. The QQ Plot in figure 2 shows that the residuals are normal and lay roughly on the diagonal line.

Figure 1

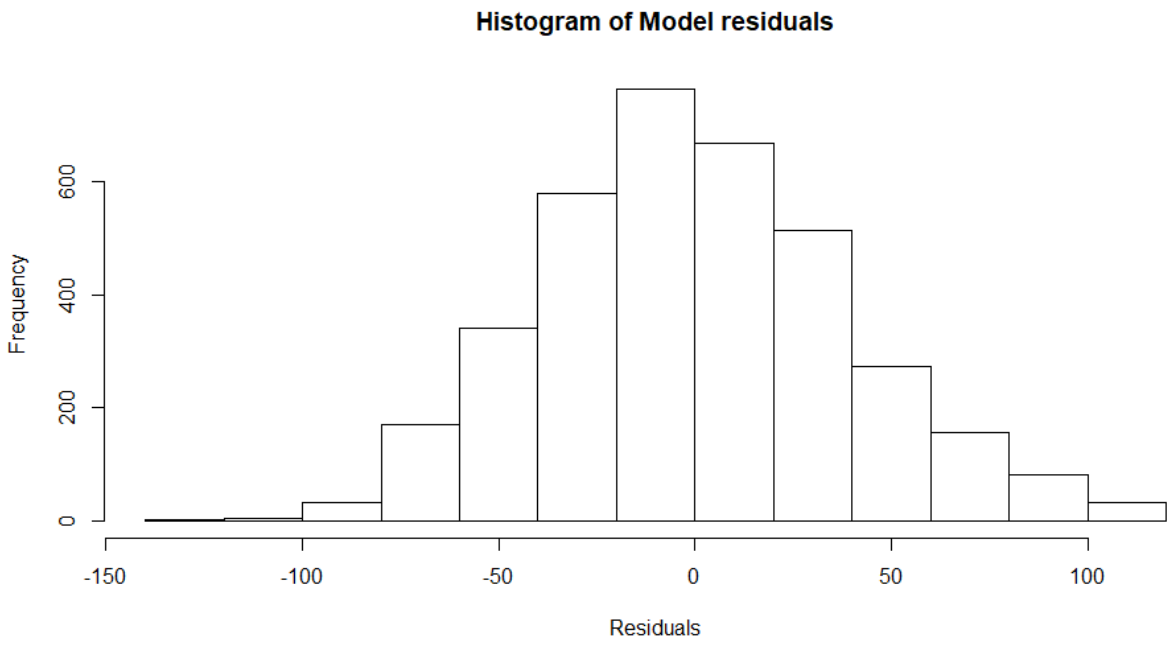
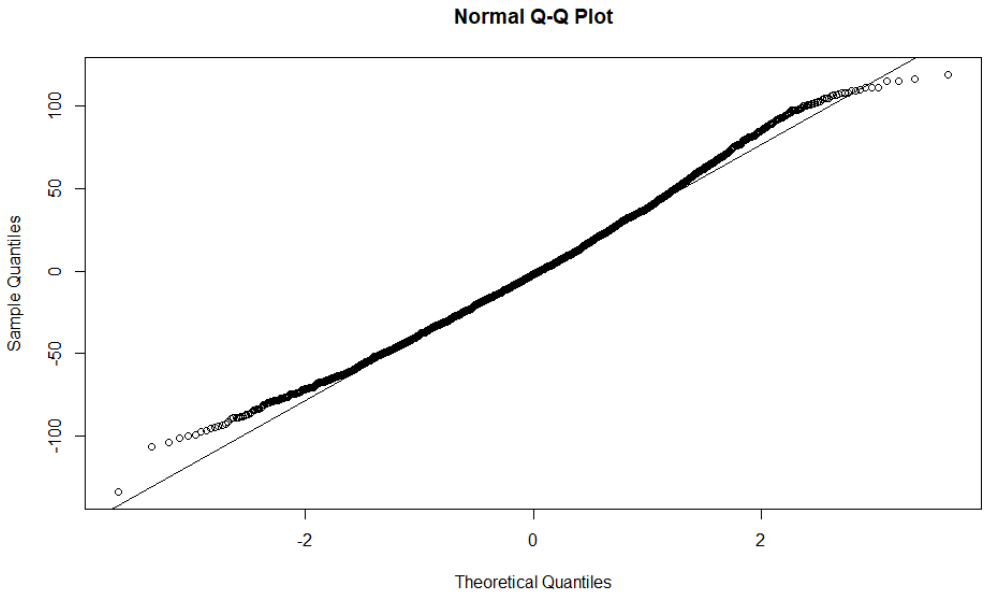


Figure 2

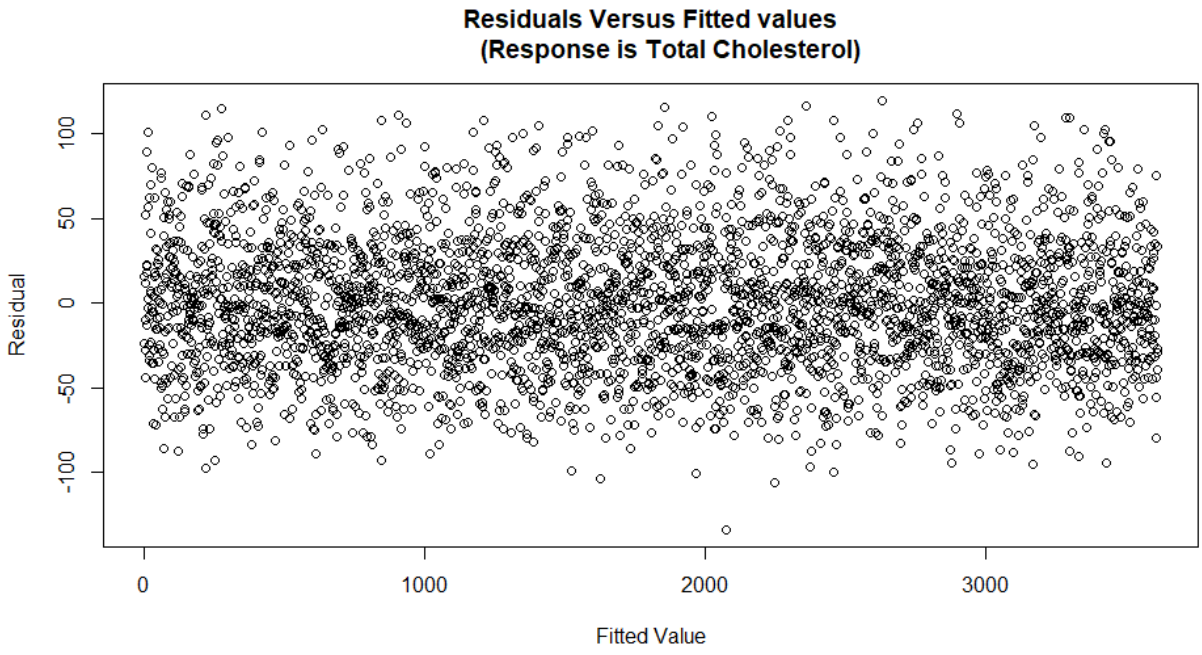


The sum of the residuals was also observed to determine if they have a mean of zero. The sum confirmed that the residuals meet this assumption See the sum of the residuals below:

```
> # check the sum of the residuals
> sum(MyDataset_model5$residuals)
[1] 5.4e-12
```

To help with building a second order model I plotted the fitted values versus the predicted values of the model to determine if I had a healthy residual plot. I then plotted the residuals of the independent variables and dependent variable to assist with determining any second order or interaction terms that could be added to the model to increase the adjusted R^2. The variables in the model include Age, cigsPerDay, BPMeds, prevalent hypertension, sysBP, diaBP, BMI, and heart rate. See the plots below:

Figure 3




```
> MyDataset_model5 <- lm(totChol ~ Age + cigsPerDay + BPMeds + prevalentHyp +
+                          sysBP + diaBP + BMI + heartRate,  data = totChol_outliers_removed)

> durbinWatsonTest(MyDataset_model5)
 lag Autocorrelation D-W Statistic p-value
   1          0.0119          1.98   0.516
 Alternative hypothesis: rho != 0
>
```

Figure 3 displays the residuals versus the fitted values of the model. All the residuals in this report have been standardized used the rstandard function in R. The plot showed no obvious pattern or no big change in the variance within the data, so the plot appeared to be healthy . The Durbin-Watson Test showed there maybe a slight independence among the residuals in the model. Next, I plotted residuals against the independent variables in the model.

Figure 4

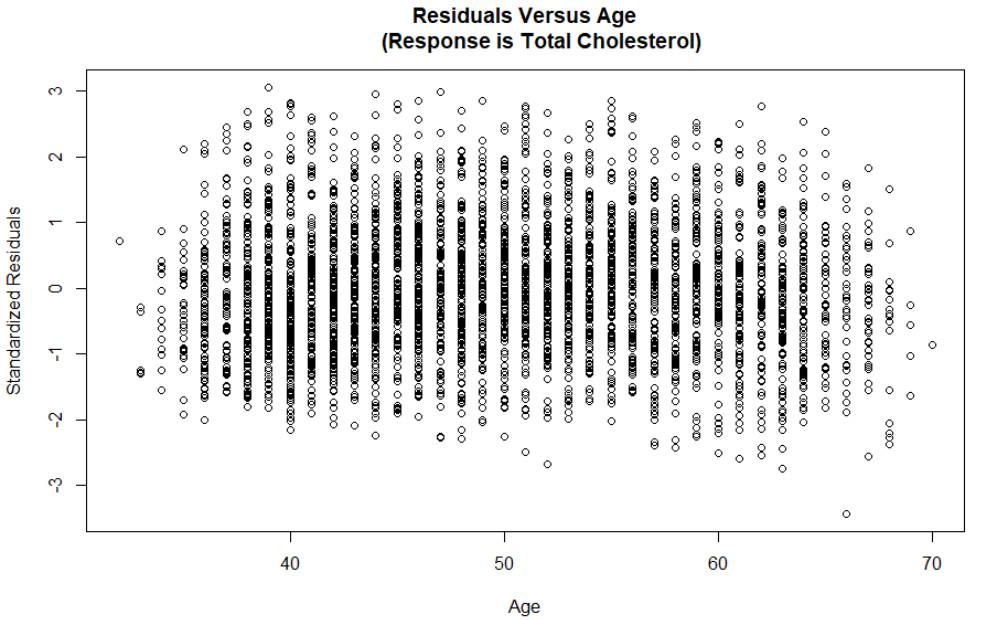


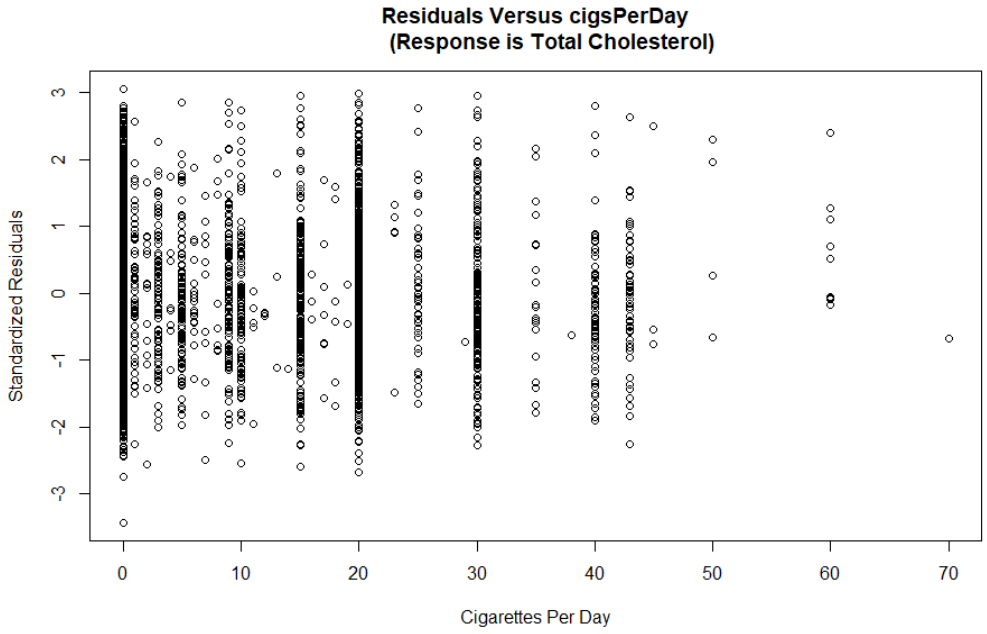
Figure 4 appears to display somewhat of a binomial pattern, but the variable could not be transformed using a stabilizing transformation as the results produced a result of NaN.

To test the dependency of the age variable, I performed a Durbin-Watson Test. The Durbin Watson test provides information regarding the independence of the residuals. See the output below:

```
> durbinWatsonTest(Resid_model1.MydataSet)
 lag Autocorrelation D-W Statistic p-value
   1         0.00449          1.99   0.806
 Alternative hypothesis: rho != 0
> |
```

According to the results the variable may not be independent, there may be some correlation among age and total cholesterol levels.

Figure 5



The scatterplot in figure 5 appears to be normal with no observable pattern. It does not appear that a second order term would be useful for cigsPerDay.

Figure 6

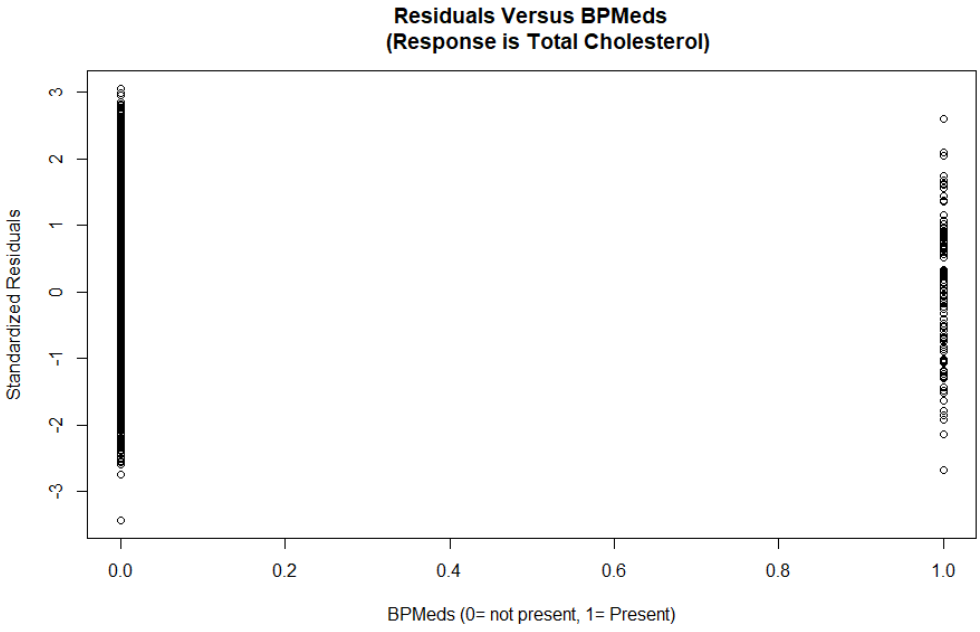
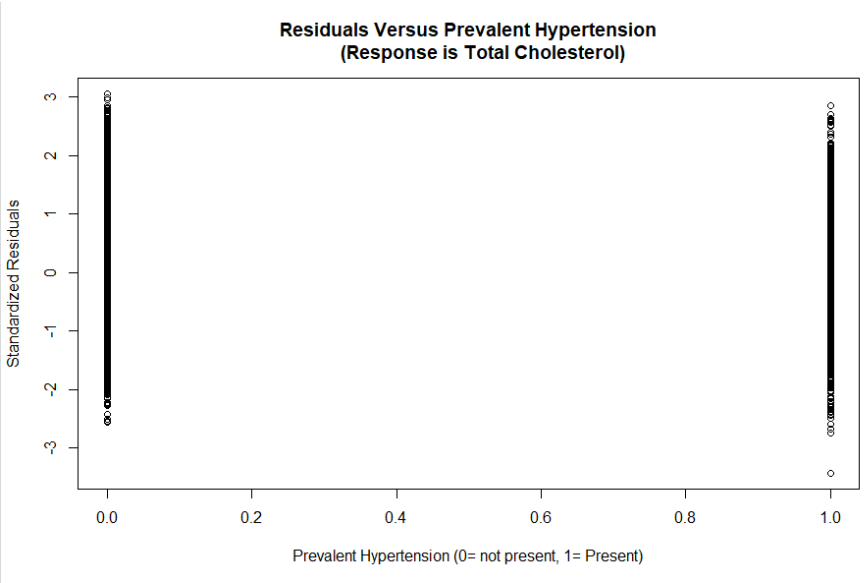
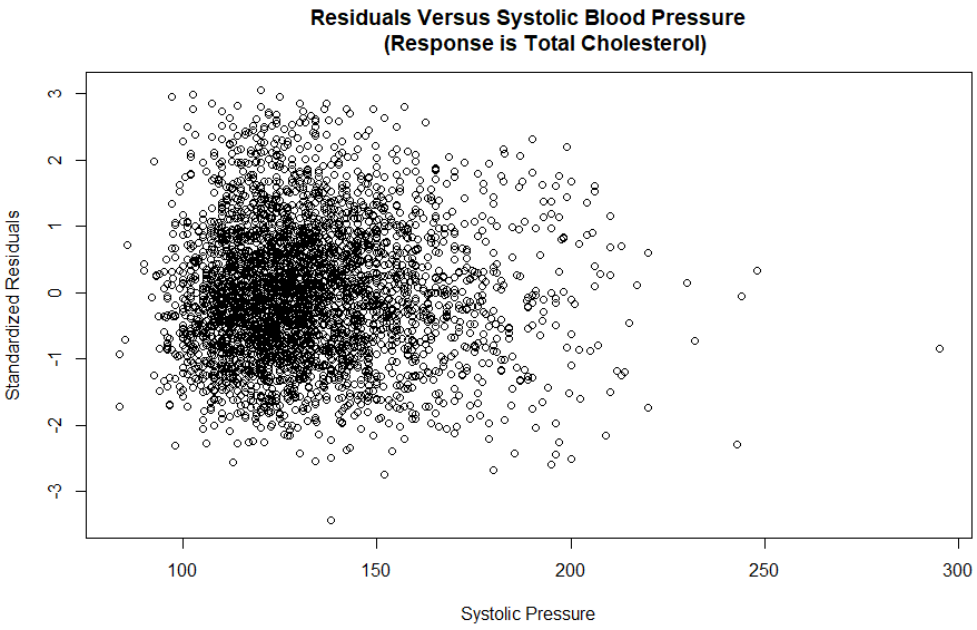


Figure 7



Figures 6 and 7 are binary variables and contain no visible pattern

Figure 8



In Figure 8 the scatterplot for the residuals versus systolic blood pressure shows many of the datapoints grouped together between 100 and 150. There is no visible pattern that would suggest the use of a second order or an interaction term.

Figure 9

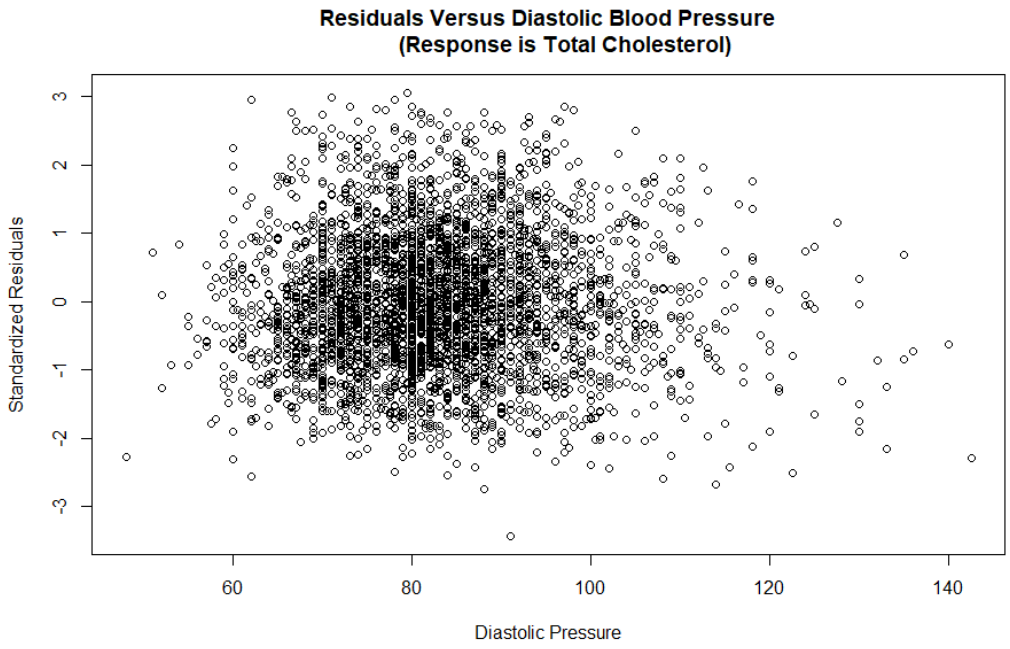


Figure 9 displays the residual plot of diastolic blood pressure. Many of the values appear to be grouped between 60-100. There does not appear to be a visible pattern or trend, but diastolic may be useful in creating an interaction term.

Figure 10

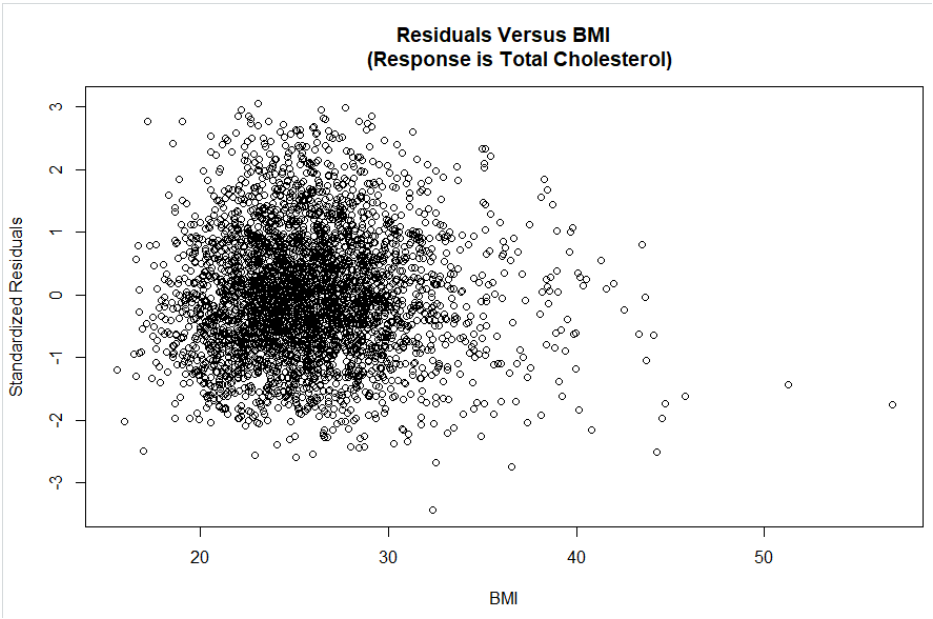


Figure 10 displays the residual plot of BMI. Many of the values appear to be grouped between 20-30. There does not appear to be a visible pattern or change in variability, however BMI may be useful with creating an interaction term.

Figure 11

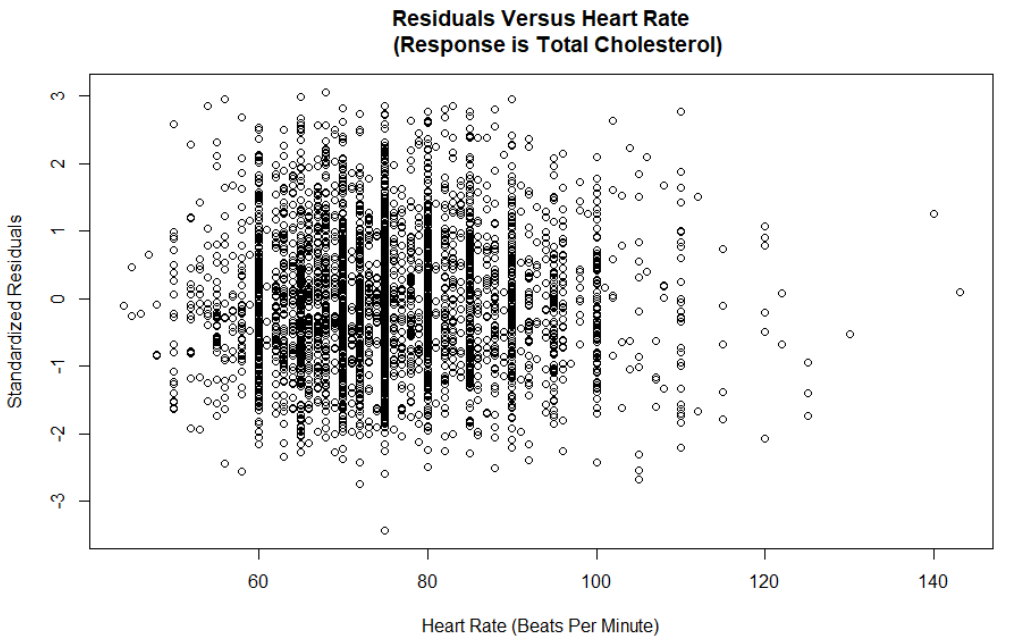


Figure 11 displays the residual plot of BMI. There does not appear to be a noticeable pattern that would suggest the use of a second order term.

There did not appear to be any plots that indicated the need for a second order term to be used in the model. The residual analysis showed that the residuals conformed to the assumptions about the errors.

Several interaction terms were tried to improve the overall adjusted R^2 of the model. The interaction terms that proved most useful included:

- Diastolic Blood Pressure and BMI
- Diastolic Blood pressure and Age
- BMI and Age

These interaction terms may have been useful because a patient body mass index may have an impact on Blood pressure levels. A patients age might be deterministic of average blood pressure levels and BMI.

The first order model contained cigsPerDay. The variable was removed because it did not seem to be useful to the second order model. The t-test for cigsPerDay in figure 12 was 0.12882, so I would not use the estimation in the model. See the Regression model below:

Figure 12

```
> MyDataset_model9 <- lm(totChol ~ Age + BPMeds +cigsPerDay + prevalentHyp +
+                          + sysBP + diaBP + BMI + heartRate
+                        + diaBP_Age + BMI_Age + diaBP_BMI, data = totChol_outliers_removed)
> summary(MyDataset_model9)

Call:
lm(formula = totChol ~ Age + BPMeds + cigsPerDay + prevalentHyp +
    +sysBP + diaBP + BMI + heartRate + diaBP_Age + BMI_Age +
    diaBP_BMI, data = totChol_outliers_removed)

Residuals:
    Min      1Q  Median      3Q     Max
-122.21  -26.40   -2.08   25.06  122.74

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.59e+02   3.56e+01   -4.46  8.6e-06 ***
Age          4.88e+00   6.21e-01    7.87  4.8e-15 ***
BPMeds       9.32e+00   4.00e+00    2.33  0.01974 *
cigsPerDay   8.37e-02   5.51e-02    1.52  0.12882
prevalentHyp -2.58e+00   1.99e+00   -1.30  0.19388
sysBP        1.35e-01   5.64e-02    2.40  0.01648 *
diaBP        2.46e+00   4.36e-01    5.63  1.9e-08 ***
BMI          7.44e+00   1.26e+00    5.89  4.3e-09 ***
heartRate    1.74e-01   5.51e-02    3.15  0.00164 **
diaBP_Age    -2.45e-02   6.92e-03   -3.54  0.00041 ***
BMI_Age      -6.91e-02   1.98e-02   -3.48  0.00051 ***
diaBP_BMI    -3.95e-02   1.07e-02   -3.69  0.00023 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.6 on 3600 degrees of freedom
Multiple R-squared:  0.113,     Adjusted R-squared:  0.11
F-statistic: 41.5 on 11 and 3600 DF,  p-value: <2e-16

>
```

Final Second Order Model

```
> MyDataset_model9 <- lm(totChol ~ Age + BPMeds + prevalentHyp +
+                          + sysBP + diaBP + BMI + heartRate
+                        + diaBP_Age + BMI_Age + diaBP_BMI, data = totChol_outliers_removed)
> summary(MyDataset_model9)

Call:
lm(formula = totChol ~ Age + BPMeds + prevalentHyp + +sysBP +
    diaBP + BMI + heartRate + diaBP_Age + BMI_Age + diaBP_BMI,
    data = totChol_outliers_removed)

Residuals:
    Min      1Q  Median      3Q     Max
-122.41  -26.36   -2.13   24.80  124.25

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.59e+02   3.56e+01   -4.46  8.5e-06 ***
Age          4.90e+00   6.21e-01    7.90  3.7e-15 ***
BPMeds       9.25e+00   4.00e+00    2.31  0.02069 *
prevalentHyp -2.55e+00   1.99e+00   -1.28  0.19888
sysBP        1.34e-01   5.65e-02    2.38  0.01746 *
diaBP        2.46e+00   4.36e-01    5.64  1.8e-08 ***
BMI          7.48e+00   1.26e+00    5.92  3.6e-09 ***
heartRate    1.79e-01   5.50e-02    3.26  0.00113 **
diaBP_Age    -2.46e-02   6.92e-03   -3.56  0.00038 ***
BMI_Age      -7.03e-02   1.98e-02   -3.54  0.00040 ***
diaBP_BMI    -3.94e-02   1.07e-02   -3.68  0.00024 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.6 on 3601 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.11
F-statistic: 45.4 on 10 and 3601 DF,  p-value: <2e-16
```

Final Model Assessment

The final second order regression model uses total cholesterol as the response variable and contains 10 explanatory variables which include age, BPMeds, prevalentHyp, sysBP, diaBP, BMI, and heartrate. Interaction terms with Diastolic Blood Pressure and BMI, Diastolic Blood Pressure and Age, and BMI and Age were added to the model. The F-test in this model shows a p-value of 2.2e-16, since the p-value is low, I would reject the null hypothesis and accept the alternative, that at least one of the betas is not equal to zero. The adjusted R^2 explains that 11% of the variability in total cholesterol is explained by the model. . As seen above in figures 12, cigsPerDay was removed from the model.

CigsPerDay was removed from the model because the t-test had a P value of 0.12882, which indicated the estimation for this variable would not be useful in the model.

The variables selected to be used in the model included the following:

- Age
- BPMeds
- prevalentHyp
- sysBP
- diaBP
- BMI
- heartrate
- diaBP_Age
- BMI_Age
- diaBP_BMI

Even though the t-test for prevalent hypertension is not good, it is left in the model because the goal to determine how hypertension test results affects the predictability of total cholesterol levels for the patients in the dataset. It also appeared that with the addition of the second order terms the t-test P-values of sysBP and prevalent hypertension decreased. The p-value of prevalent hypertension decreased from 0.235094 to 0.19888, and the p-value for sysBP decreased from 0.158355 to 0.01746. The value for sysBP was not very useful to the first order model, but with the addition of second order terms the p-value appeared to improve.

First order model

prevalentHyp	-2.37886	2.00319	-1.188	0.235094
sysBP	0.07944	0.05631	1.411	0.158355

Second order model

prevalentHyp	-2.55e+00	1.99e+00	-1.28	0.19888
sysBP	1.34e-01	5.65e-02	2.38	0.01746 *

The T-Test for age is good with a p-value 3.7e-15. Based on the T-test for Age , I would reject the null hypothesis and accept the alternative, I would use the estimation for Age in the model. The T-Test for cigsPerDay is not the best with a p-value of 0.068842. The T-Test for BPMeds is good with a p-value of 0.02069. Based on the T-test for BPMeds , I would reject the null hypothesis and accept the alternative, I would use the estimation for BPMeds in the model. The T-Test for prevalentHyp is not good with a p-value of 0.19888. Based on the T-test for prevalentHyp, I would fail to reject the null hypothesis, I would not use the estimation for prevalentHyp in the model. The regression coefficient for hypertension has a

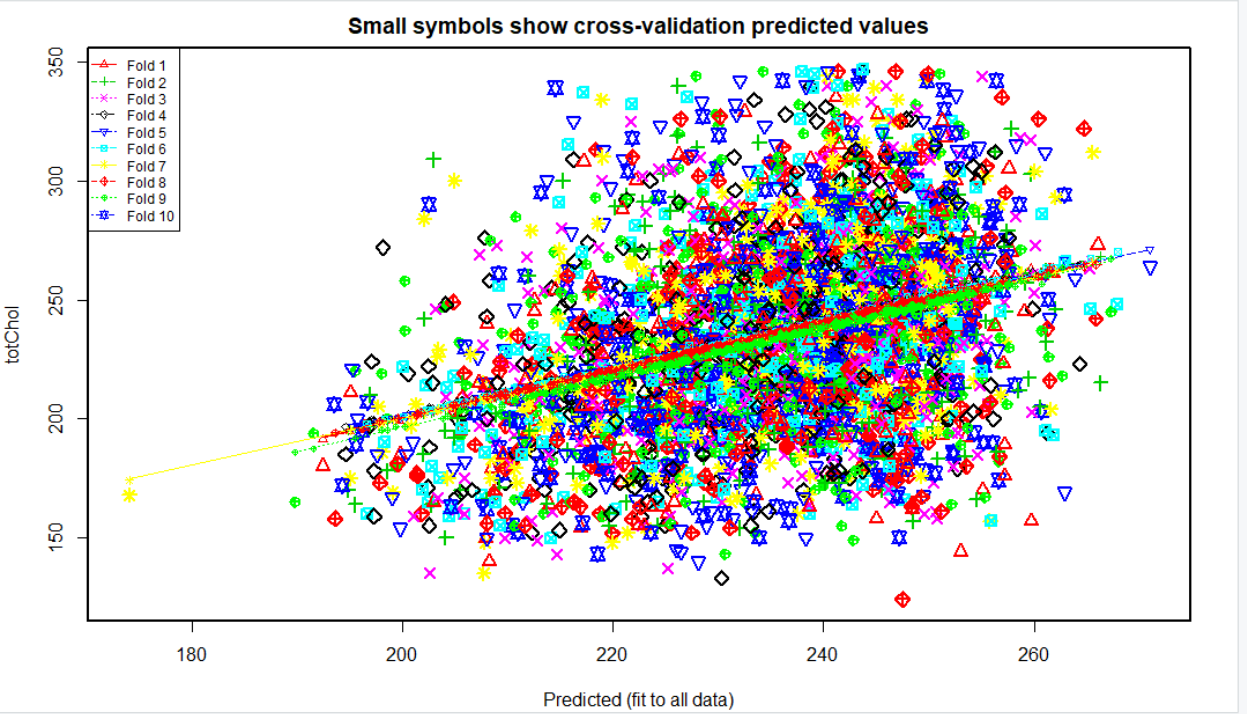
negative estimated value which might suggest that hypertension is negatively correlated with total cholesterol.

The T-Test for sysBP is good with a p-value of 0.01746. Based on the T-test for sysBP, I would reject the null hypothesis and accept the alternative, I would use the estimation for sysBP in the model. The T-Test for diaBP is good with a p-value 1.8 e-08. Based on the T-test for diaBP , I would reject the null hypothesis and accept the alternative, I would use the estimation for diaBP in the model. The T-Test for BMI is good with a p-value 3.6e-09. Based on the T-test for BMI , I would reject the null hypothesis and accept the alternative, I would use the estimation for BMI in the model. The T-Test for heartRate is good with a p-value of 0.00113. Based on the T-test for heartRate , I would reject the null hypothesis and accept the alternative, I would use the estimation for heartRate in the model.

The T-Test for diaBP_Age is good with a p-value of 0.00038. Based on the T-test for diaBP_Age , I would reject the null hypothesis and accept the alternative, I would use the estimation for diaBP_Age in the model. The T-Test for BMI_Age is good with a p-value of 0.00040. Based on the T-test for BMI_Age , I would reject the null hypothesis and accept the alternative, I would use the estimation for BMI_Age in the model. The T-Test for diaBP_BMI is good with a p-value of 0.00024. Based on the T-test for diaBP_BMI , I would reject the null hypothesis and accept the alternative, I would use the estimation for diaBP_BMI in the model.

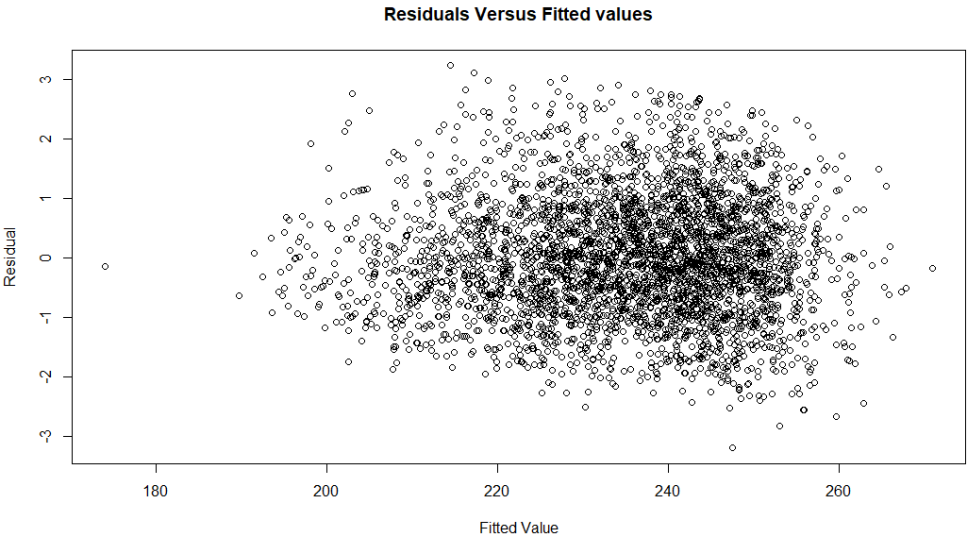
I performed 10-fold cross validation (90/10 split) on the final model to determine if the model was overfitting or biased. Roughly 361 observations were used in each fold as the test set to perform training and testing on the model. The result showed an overall mean square error of 82311 which is quite high. This might indicate that the model is not performing very well at predicting total cholesterol. See the output and plot below:

```
> TenFold_cv<- cv.lm(data= MyDataset_model9_data, form.lm = formula(totChol ~ .), m=10)
Analysis of Variance Table

Response: totChol
              Df  Sum Sq Mean Sq F value  Pr(>F)
Age            1  447299  447299   299.5 < 2e-16 ***
prevalentHyp   1   32458   32458    21.7 3.2e-06 ***
sysBP          1   44341   44341    29.7 5.4e-08 ***
diaBP          1   19335   19335    12.9 0.00032 ***
BMI            1   16567   16567    11.1 0.00088 ***
heartRate      1   15864   15864    10.6 0.00113 **
diaBP_Age      1   48268   48268    32.3 1.4e-08 ***
BMI_Age        1   25220   25220    16.9 4.1e-05 ***
diaBP_BMI      1   19987   19987    13.4 0.00026 ***
Residuals   3602 5379689    1494
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Overall (Sum over all 361 folds)
   ms
83211
```



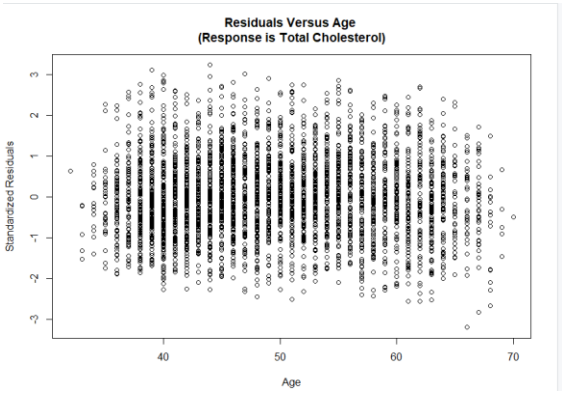
I also plotted the residuals of the independent and predicted values of the final model:

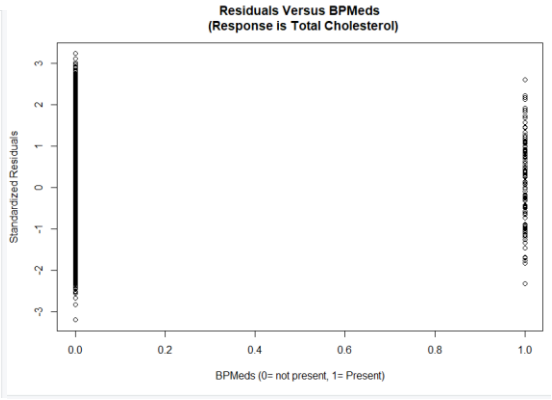
Residuals Versus fitted values of the final model:

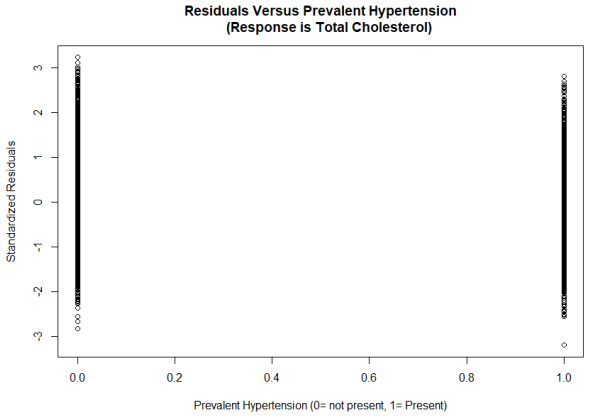


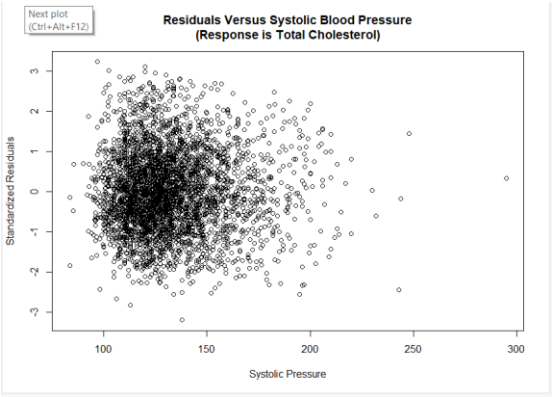
The residual plot of the predicted values did not appear to have an observable pattern and appeared to be normal or homoscedastic. The values however appear to be bunched together between 220 and 260.

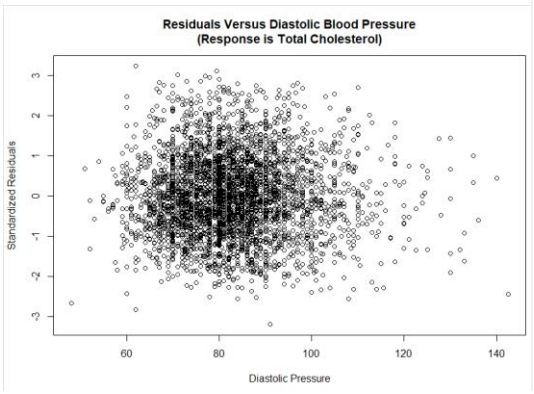
Plots for residuals versus the independent variables

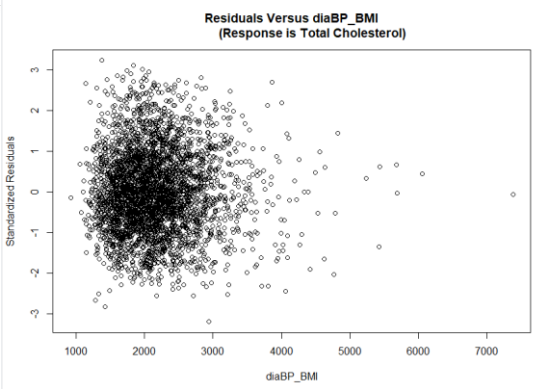


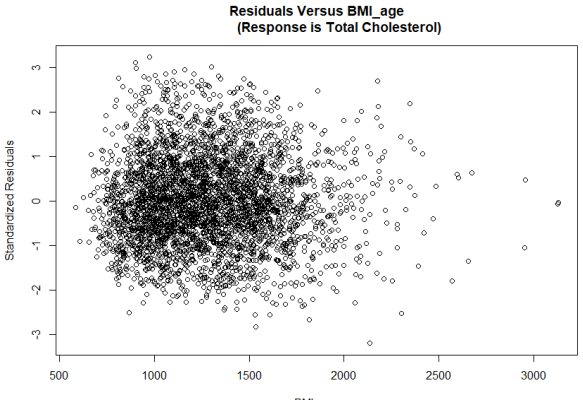


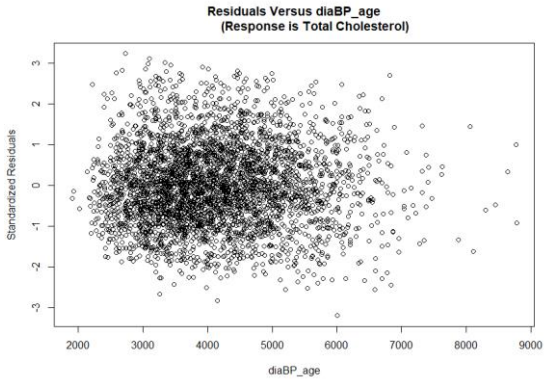


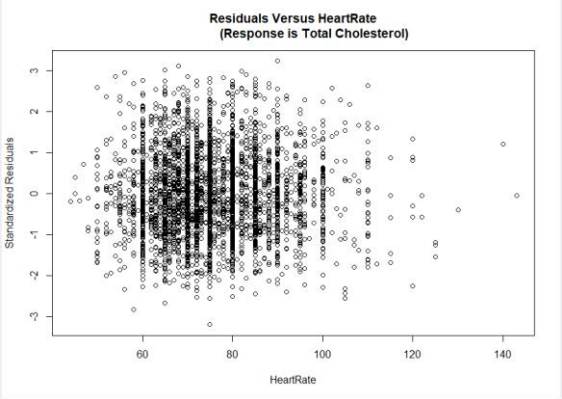












The residual plots of the independent variables did not show any observable patterns and were almost identical to the residual plots observed for the first order model shown earlier in the report. The age variable appeared to be similar to a binomial pattern, but the variable could not be transformed as described earlier. The plots containing systolic blood pressure, diastolic blood pressure and heart rate appeared to have data values that were grouped together. Some of the standardized residuals appeared to be over 2 but the values are most likely outliers within the dataset.

Summary

In summary it appears that the inclusion of hypertension testing along with the other explanatory variables including interaction terms improved the predictability of total cholesterol. However, the inclusion of hypertension testing(hypertension present or not present) in the dataset does not offer additional predictability of total cholesterol. Moreover, removing the hypertension feature appears yield a better overall model. It appears that presence of hypertension may not be a good indicator of a patients' total cholesterol levels. Age appeared to be the most useful when trying to predict cholesterol levels.

Patients sex, smoking habits and education were removed from the dataset to perform this analysis. It could be possible that the inclusion of these additional features might improve the ability to predict patient total cholesterol levels. More factors such as physical activity, diet, previous family history, and

health care coverage may be useful to include in the patient records to help with predicting cholesterol levels. Overall, it appears that hypertension testing results are not useful in determining a patient's cholesterol levels, rather a patients age, diastolic blood pressure, BMI, and heartRate. appear to be most predictive of total cholesterol.