ISyE 312 Data Management and Analysis Project

Investigating the Relationship Between Various Health Factors and Resting Blood Pressure

Group 12: Sanjana Parikh, Manavv Shah, Akshaya Thiru, Nandan V

Table of Contents

1. Introduction

The dataset we selected is from the University of California at Irvine's Machine Learning repository and provides medical information of patients in Cleveland, which helps determine how likely the patient is to get heart disease. The file was retrieved from Kaggle and contains age, sex, resting blood pressure, maximum heart rate, and nine more variables. For over 300 individuals.

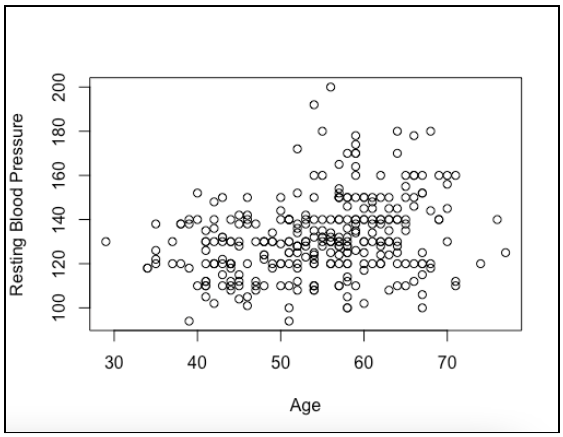
After examining the data, we noted that the aim of this project is to investigate the relationship between various health factors and resting blood pressure. To successfully investigate the same, we decided to split the process in the following manner:

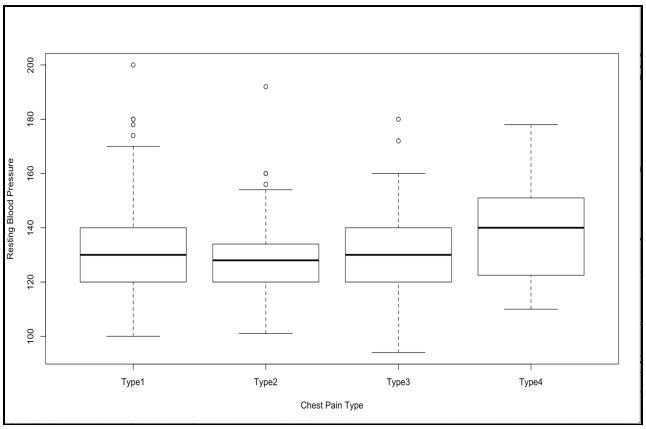
- Find a relationship between independent variables and the dependent variable (resting blood pressure)
- Run a simple regression model to confirm assumptions (expand)
- Run a multiple regression as we got a low R2 value for simple linear regression
- Use variance inflation factors to check for multicollinearity
- Evaluate Model adequacy
    - Check if the errors have zero mean and constant variance
    - Calculate Cook's distance to check for influential factors
    - Confirm the normality assumption
    - Studying for an interaction term and evaluating an alternative model, respectively

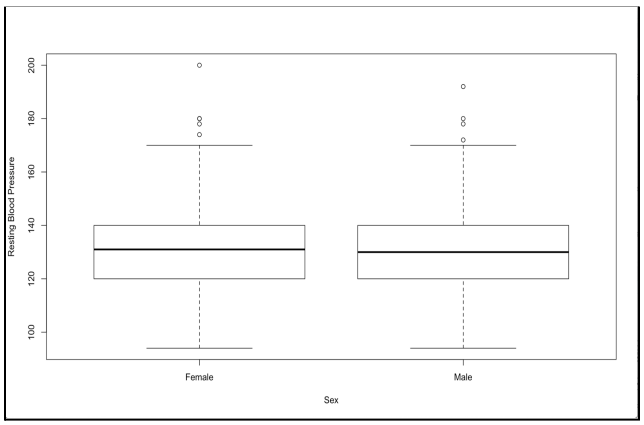
2. Data Analysis and Exploring the Dataset

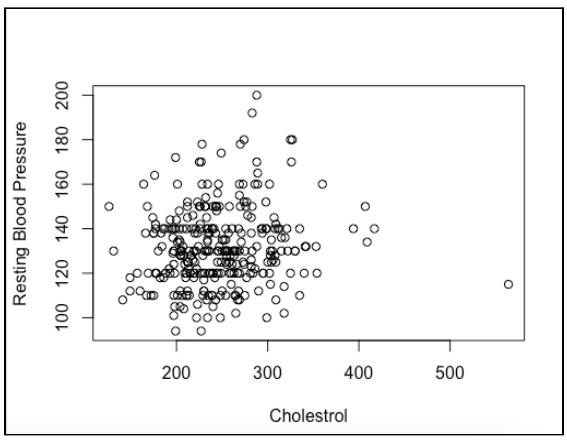
Our dataset holds continuous, binary and categorical data and so before proceeding with a simple linear regression it is essential to explore how these independent variables relate to the dependent variable.

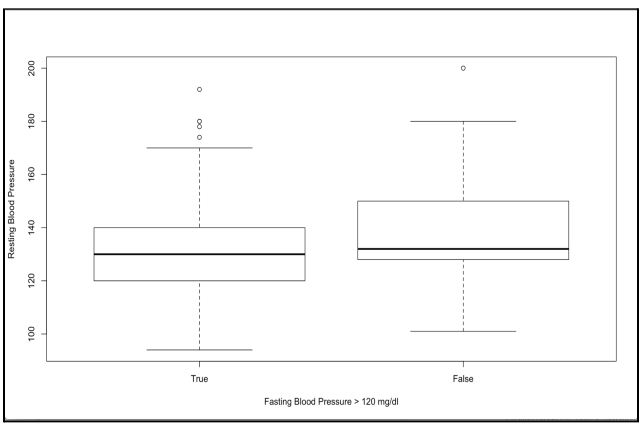
Relationships between our continuous and categorical variables with the dependent variable are shown in the plots below:











After exploring the data, we can see that relations exist between the independent and dependent variables. However, we need to analyze further how well the data is captured to move forward with an adequate model.

We conduct a simple linear regression of resting blood pressure on age to understand the summary statistics.

3. Simple Linear Regression

The linear regression model we ran is:

$$Resting\ blood\ pressure\ =\ \beta 0\ +\ \beta 1\ \times\ age\ in\ years$$

Call:

lm(formula = heart$trestbps ~ heart$age)

Residuals:

Min	1Q	Median	3Q	Max
-38.439	-11.499	-1.044	10.192	67.495

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	102.2961	5.8906	17.366	< 2e-16	***
heart$age	0.5394	0.1069	5.048	7.76e-07	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.87 on 301 degrees of freedom

Multiple R-squared: 0.07804, Adjusted R-squared: 0.07497

F-statistic: 25.48 on 1 and 301 DF, p-value: 7.762e-07

The summary statistics give us a better insight of how the variation in y is captured by the variation in x. Using the R-squared value, we interpret that 7.89% of the variation in the resting blood pressure data is captured by variation in the age data. We can see from the summary of

simple_model that the p-value is less than 0.001. Therefore, we can conclude that our model is statistically significant.

Additionally, from the coefficients in the summary, we can deduce the regression equation. The coefficient of age is 0.5394 - this means that for every increase in age (in years) the resting blood pressure increases by 0.5394. The regression equation for our model is:

$$Resting\ blood\ pressure\ =\ 102.2961\ +\ (0.5394\ \times\ age\ in\ years)$$

The visualization of our simple regression model of resting blood pressure on age can be seen below.

As only 7.89% is explained, this suggests that a multiple regression would work better, and this will also reduce the possibility of omitted variable bias in the simple model.

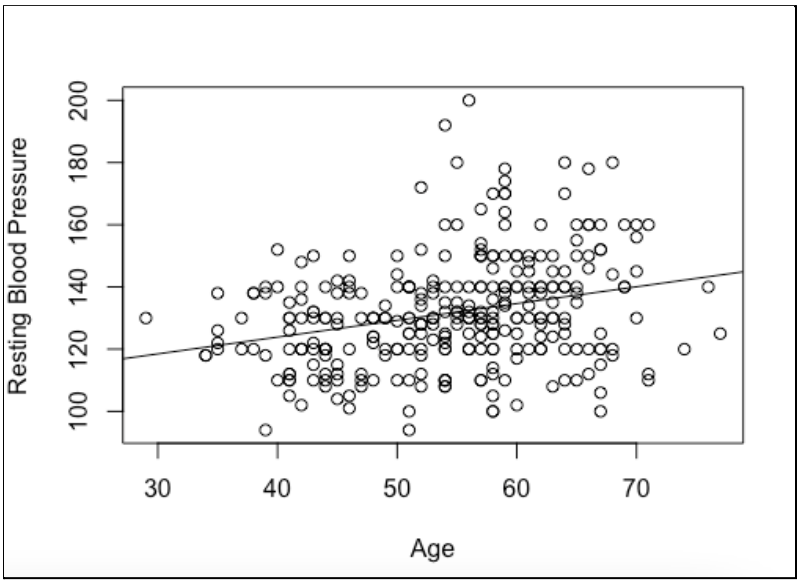


Figure 2 :Simple linear regression plot of resting blood pressure on age

By plotting the regression line, we can confirm that this is a poor regression as the residual values appear to be large with several data points lying far from the regression line.

4. Multiple Linear Regression

As simple regression plot and our low R-Squared value suggested that we need more variables to explain the data, a multiple regression model was run to regress the resting blood pressure on age, cholesterol, maximum heart rate, sex, chest pain, and fasting blood pressure. The explanation of variation increased to 10.89%, which is still relatively low, but this is understandable as there is a significant amount of variation between our sample units (human beings). The model should not be dismissed due to the low r-squared value. It can still possess a significant trend that helps us understand the relationship between our independent and dependent variables. According we moved to further analyze our model with a hypothesis test as follows:

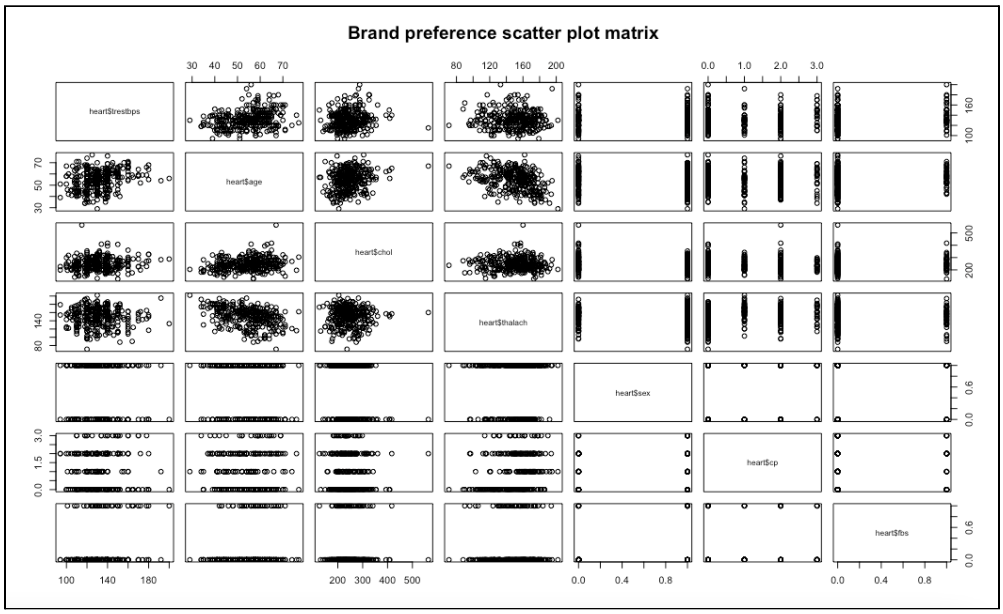
Null and alternative hypothesis :

$H0: \beta i = 0 \; for \; all \; i = age, cholestrol, max \; heart \; rate, \; sex, \; chestpain, fasting \; blood \; pressure$

$H1: \quad At \; Least \; one \; \beta i \neq 0$

Out of the thirteen explanatory variables in the dataset, we ran our regression on six independent variables. We were able to run a stepwise regression to confirm our model selection and choose which independent variables to include in the model using the stepAIC( ) function. Using a significance level of 5%, we can see that our p-value of 5.702e-06 is less than the risk level 0.05, and so the null hypothesis can be rejected. We can conclude that our model is jointly significant.
Even though a multiple regression model does explain the variation of y better. Multicollinearity could impact the usefulness of our selected model. Adding more variables to our model could lead to larger variances and less precise estimates, damaging the results and use of the model. This will impact the efficiency of our model as large variances, and covariances of our estimators will increase the chances of failing to reject the null hypothesis that the variables are not significant.

To



investigate if there is any linear correlation between our independent variables, we can use a scatter plot matrix.

Figure 3: Scatter plot matrix

Using the scatter plot matrix, we cannot conclude that there are any strong linear relationships between the independent variables. To ensure that we can move past the issue of

multicollinearity, we conduct an additional multicollinearity diagnostic using variance inflation factors and obtain the following result:

heart$age	heart$thalach	heart$sex	heart$cp	heart$fbs	heart$oldpeak	heart$restecg	heart$thal
1.262978	1.404194	1.080637	1.134815	1.037310	1.191049	1.026664	1.111619

Variation inflation factors greater than 5 to 10 are considered significant, but as the VIFs are lower than this range and close to 1, we can say that multicollinearity is not a concern. Even though the goodness of fit value is higher, this is no indication that the model is adequate as adding variables to our model will lead to this regardless. We need to proceed and check for model adequacy.

4.1 Checking assumptions for our model

We need to ensure that our model is adequate and that the linear regression model's assumptions are met. Violations of the assumptions could lead to an unstable model, and so this is an essential step in ensuring that we are using the correct model.

The first assumption, the relationship between the explanatory variables and the response is linear, can be satisfied as we inferred this when exploring our data in the first step. Our plots suggested that some linear relationship exists between the independent and dependent variables. The other assumptions will be discussed below.

4.2 Checking if the errors have zero mean and constant variance

The plot below looks at the relationship between the residuals generated by the model and the fitted value. The residuals appear to be evenly distributed around the zero line. This is a satisfactory residual plot as there are no signs of a funnel, double bow, or non-linear pattern. With this plot, we can assume that the errors have zero mean, constant variance, and there are no significant outliers in our data. We can also see that our residuals are uncorrelated and show no patterns (non-linear, funnel etc)

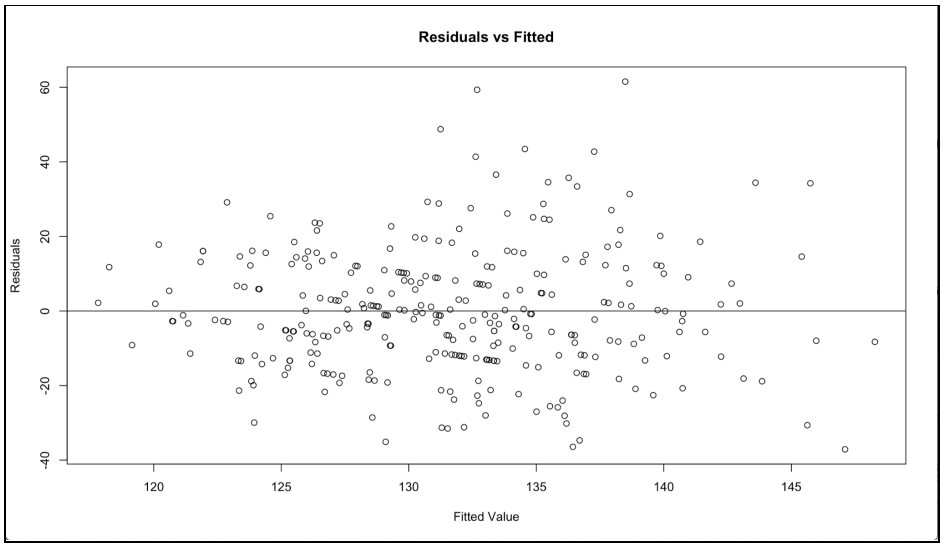


Figure 4 : Residual plot

To further test our model for any influential factors, we calculated the Cook's distance for each observation. This can be seen below:

	dfb.1_	dfb.hrt.g	dfb.hrt.ch	dfb.hrt.t	dfb.hrt.s	dfb.hrt.cp	dfb.hrt.f	dffit	cov.r	cook.d
1	-0.003571	0.005576	-1.26e-03	-9.22e-04	0.005061	1.29e-02	0.014674	0.023912	1.063	8.20e-05
2	0.000552	-0.030945	1.34e-02	1.62e-02	0.016292	1.64e-02	-0.007495	0.058917	1.044	4.97e-04
3	0.021219	-0.023020	-1.80e-02	8.76e-03	-0.035450	-6.70e-03	-0.002769	0.053607	1.044	4.12e-04
4	0.056372	-0.039840	1.08e-02	-6.59e-02	-0.033491	1.42e-02	0.022040	-0.086664	1.026	1.07e-03
5	0.047770	-0.006138	-8.66e-02	-4.16e-02	0.055372	5.32e-02	0.015031	-0.145882	1.033	3.04e-03
6	-0.004175	0.024979	-4.02e-02	1.77e-02	0.017762	-3.62e-02	-0.014505	0.070729	1.028	7.16e-04
7	-0.001076	-0.000333	1.53e-02	1.60e-03	-0.030538	6.70e-04	-0.008646	0.046026	1.032	3.03e-04
8	0.004592	0.016764	-1.37e-02	-1.40e-02	-0.016141	2.62e-03	0.007151	-0.041271	1.033	2.44e-04
9	0.013660	-0.018677	-9.29e-02	3.51e-02	0.055924	7.67e-02	0.293779	0.370143	0.942	1.93e-02
10	-0.053436	0.078744	-9.91e-02	8.04e-02	0.038920	3.89e-02	-0.040631	0.166650	1.017	3.96e-03
11	-0.020122	0.013329	-5.68e-03	3.06e-02	0.024094	-3.85e-02	-0.013425	0.063452	1.025	5.76e-04
12	0.006171	-0.005631	2.81e-03	-5.76e-03	-0.006245	5.63e-03	-0.001404	0.011395	1.047	1.86e-05
13	-0.002452	-0.000833	2.35e-03	3.28e-03	0.003310	-5.01e-04	-0.001677	0.007056	1.034	7.14e-06
14	0.038541	-0.113160	5.57e-02	2.59e-02	-0.073547	-1.93e-01	0.074896	-0.264148	0.993	9.92e-03
15	-0.004132	-0.000355	1.36e-02	-3.32e-04	-0.035242	4.30e-02	0.059149	0.092056	1.062	1.21e-03
16	-0.025224	0.012955	2.07e-02	4.78e-03	0.050165	-2.84e-02	0.011851	-0.072005	1.033	7.42e-04
17	0.065324	-0.021629	-8.96e-02	-4.29e-02	0.057000	-4.99e-02	0.030104	-0.160066	1.025	3.66e-03
18	0.052754	0.028700	-2.89e-02	-8.75e-02	-0.067184	1.16e-01	-0.032130	0.170437	1.054	4.15e-03
19	-0.004436	-0.072278	2.17e-02	7.37e-02	0.054700	-1.02e-01	-0.018820	0.193008	0.984	5.30e-03
20	0.000128	-0.000325	1.16e-04	-2.65e-05	0.000236	-3.66e-04	0.000147	-0.000619	1.060	5.48e-08
21	-0.009862	0.010033	-3.84e-03	1.13e-02	0.006886	-1.10e-02	-0.004392	0.019739	1.038	5.58e-05
22	-0.002433	-0.007043	1.22e-03	8.54e-03	0.008608	9.17e-03	-0.005249	0.024959	1.037	8.93e-05
23	-0.007007	-0.039699	-1.06e-02	6.61e-02	0.030999	-7.25e-02	-0.011328	0.133384	1.020	2.54e-03
24	0.001541	0.008787	1.27e-03	-1.92e-02	0.019913	3.18e-02	0.069853	0.092636	1.046	1.23e-03
25	0.025464	-0.056006	-2.06e-02	1.18e-02	0.032839	9.69e-02	-0.025686	0.153619	1.029	3.37e-03

Figure 5 : Cook's Distance values

The figure above shows the first 25 values (complete data is given in the appendix). We can see that there are no values where the Cook's distance is greater than 1. Therefore, we can conclude that there are no influential factors in our data.

4.3 Checking that the errors are normally distributed

To check the assumption that the errors are normally distributed, we can evaluate if the residuals are close to the diagonal line on a qq plot. Using the qq plot below, we can see that the residuals are close to being normally distributed as they lie close to the diagonal line.

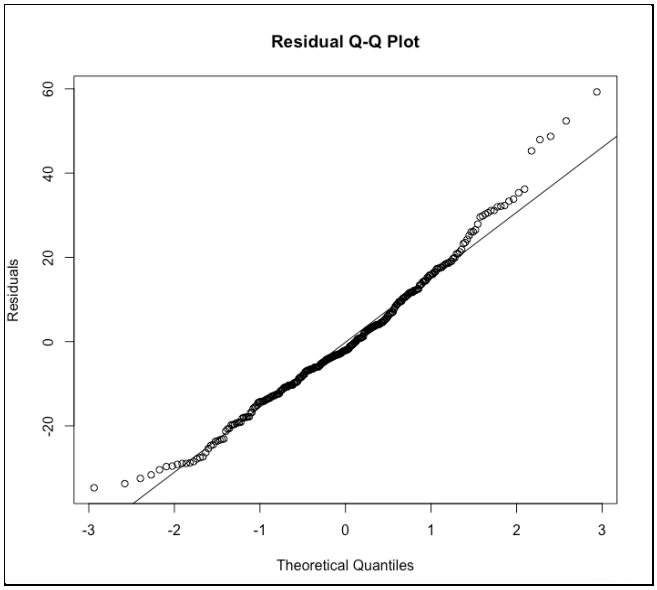


Figure 6 : QQ plot to check for the normality assumption

Fortunately, we can meet the assumptions of a linear regression model. So we can state that the linear model we use is adequate in exploring the relationship between the resting blood pressure and other health determinant factors. A final check that we could use is conducting a formal test for lack of fit for the regression model.

4.4 Evaluating an alternative model with interaction terms

Some models include the presence of interactive terms, and enabling this interaction between our dependent variables can lead to more precise models. The interactive term we chose to investigate is cholesterol and age, as older men and women have been linked with higher levels of cholesterol.

```
> interaction_model =
lm(heart$trestbps~heart$age+heart$chol+heart$thalach+heart$sex+heart$cp+heart$fbs+heart$chol*heart$age)
```

We can conduct a lack of fit test using our newly proposed model, "interaction model," using an ANOVA test. Our null hypothesis argues that both models have no significant difference, so our original model should be sufficient.

Analysis of Variance Table

Model 1: heart$trestbps ~ heart$age + heart$chol + heart$thalach + heart$sex +
heart$cp + heart$fbs

Model 2: heart$trestbps ~ heart$age + heart$chol + heart$thalach + heart$sex +
heart$cp + heart$fbs + heart$chol * heart$age

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	296	82771			
2	295	82346	1	425.13	1.523 0.2181

As the p-value is greater than the significance level 0.05, we fail to reject the null hypothesis that the two models have no significant difference.

We can confirm this using a residual plot for the new model and compare this to our previous plots.

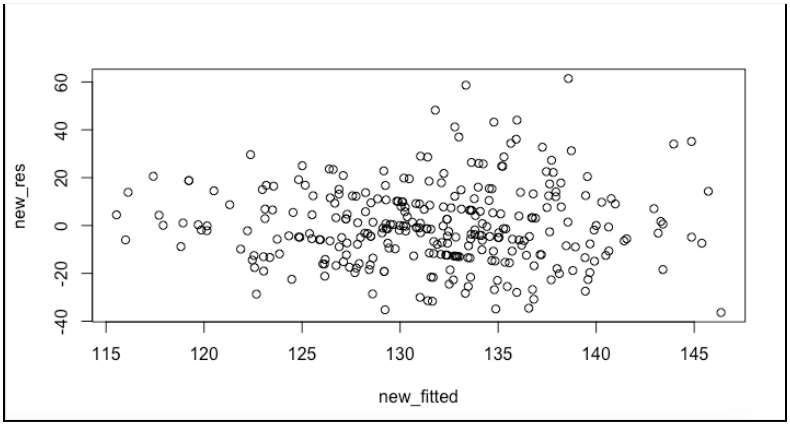


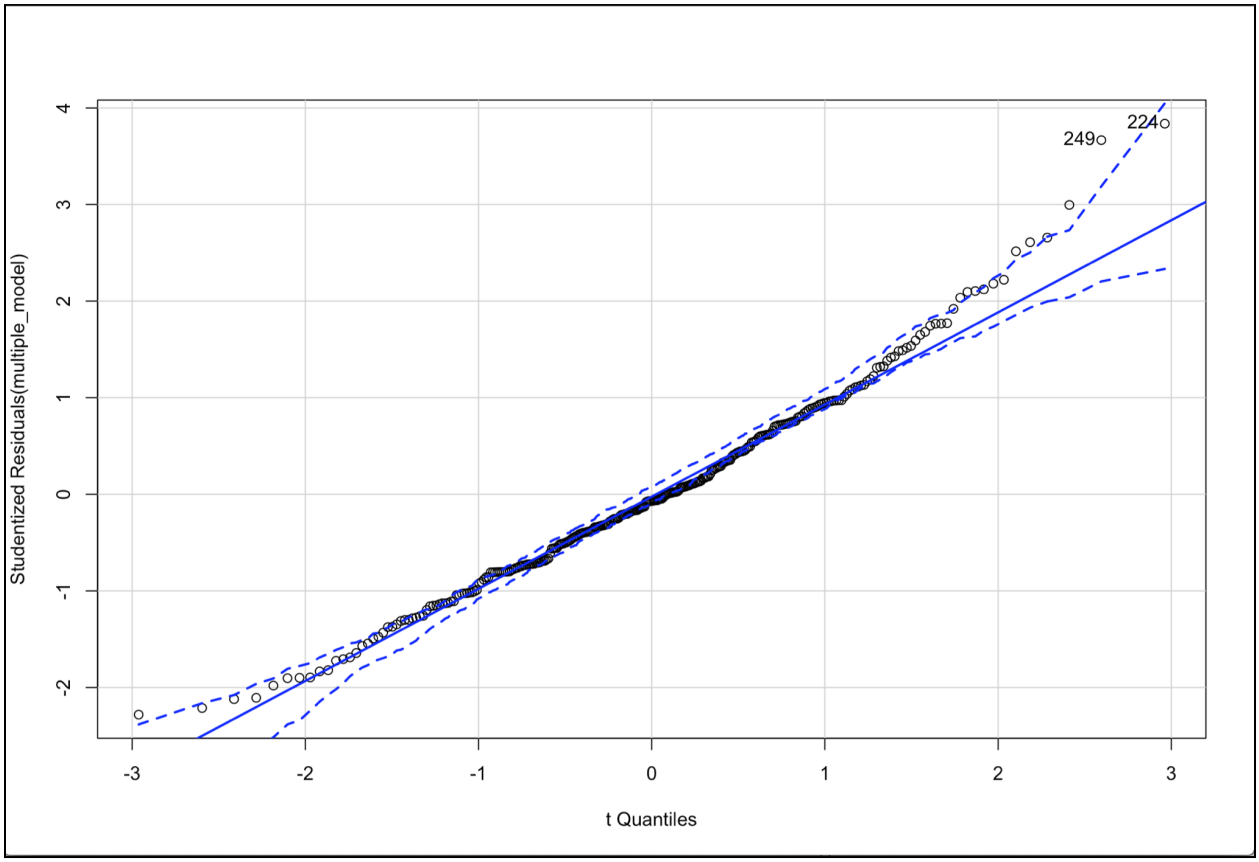
Figure 7 : Residual plot for alternate model

The residual plot for the alternate model does not appear to provide a significant advantage, and so there might be no benefit of adding this interactive term. The interactive term could also lead

to increased multicollinearity, and so using the plot above, we can safely say that our original model is sufficient.

Further Analysis and Recommendations

5.1 Enhanced Regression



The graph above shows us the theoretical quantile comparison for variables and studentized residuals from linear models. The two dotted lines act as comparison lines to check if our data goes above or below the limits. As we can see, almost all our data points are between the limits - also called a 'confidence envelope.'

5.2 Recommendations for future analysis : Least Absolute Deviation

Our group also decided to explore how our model can further be improved in other situations. There might be a case where an increased sample is used with more data points, leading to the possibility of outliers existing in our data.

Dealing with an outlier can be done using two methods :

1. Removing the outlier from the dataset
2. Find a more efficient way to estimate the parameters of the regression model

Removing an outlier from the dataset could lead to different statistical interpretations if the outlier significantly influences the model. However, we also acknowledge that sometimes outliers do provide valuable information and cannot be discarded. To further test whether our model is correct in terms of statistical significance, we decided to use the least absolute deviation method, which is less sensitive to outliers. Using the least-squares method, we attempt to find estimates for our parameters by minimizing the squared residuals. Outliers tend to have larger residuals as they are farther away from the regression line. This leads to a more considerable impact when squaring their residuals, and so estimates will be influenced. Using the LAD model ensures that residuals of outliers do not heavily influence our estimation as it minimizes the absolute value of the residuals instead of the squared residuals.

We ran this regression to observe if having a different way to estimate will lead to drastically different results when the outlier is included.

Call:

```
rq(formula = heart$trestbps ~ heart$age + heart$chol + heart$thalach +
    heart$sex + heart$cp + heart$fbs)
```

Coefficients:

```
 (Intercept)     heart$age    heart$chol heart$thalach     heart$sex      heart$cp
 84.25480880    0.55161315    0.01019235    0.09663576   -2.54494597    0.30247529
    heart$fbs
  1.75316116
```

Using the LAD regression model will be more efficient to reduce the effect of the outlier in the case that both models lead to different conclusions about joint significance. So accordingly, in the case that there is an outlier, we can proceed with this regression model.

5. Conclusion

After completing our report, we can conclude that our simple linear regression model is not adequate as our R-squared value is 7.89%. This suggests that only 7.89% of the variation in the resting blood pressure data is captured by variation in the age data. This can also be seen in Figure 2; as we plot the regression, many residual values lie far from the regression line. Therefore, it is more suitable to use a multiple regression model.

Our multiple regression model used more variables such as sex, cholesterol, fasting blood pressure, maximum heart rate, and chest pain type. This resulted in getting a p-value of 5.702e-06, which is less than the alpha level 0.05, and so the null hypothesis can be rejected. We can conclude that our model is jointly significant.

Moving forward, we conducted tests to confirm certain assumptions (i.e., normality, multicollinearity) and some analysis to make sure our data did not include any outlier or our regression was not influenced by any variables. We achieved this with the help of Cook's Distance, VIF (Variance Inflation Factors), Residual Plots, and QQ - Plots. From the results of these analyses, we confirmed that there were no outliers, and all our assumptions were met.

To improve our model further, we evaluated an alternative model using interaction terms to check if we could get a more accurate result. As seen in Section 4.4, the alternative model did not provide any significant advantage, and therefore it is advisable to stick to our original model.

Appendix

1. "Cholesterol Levels: What You Need to Know." *MedlinePlus*, U.S. National Library of Medicine, 2 Oct. 2020, medlineplus.gov/cholesterollevelswhatyouneedtoknow.html#:~:text=Things%20outside%20of%20your%20control,men%20of%20the%20same%20age.
2. "Lec-3-Diagnostics." *www2.Stat.duke.edu*.
3. *Real Statistics Using Excel*, www.real-statistics.com/multiple-regression/lad-regression/.
4. robk@statmethods.net, Robert Kabacoff -. "Regression Diagnostics." *Quick-R: Regression Diagnostics*, www.statmethods.net/stats/rdiagnostics.html.