A Statistical Analysis of Cardiovascular Health Data

MOTIVATION

Heart disease is the most common cause of death in the United States. (Center for Disease Control and Prevention, 2021) This aspect of mortality in the United States has been true for nearly an entire century this point in time and sees no signs of changing. (Center for Disease Control and Prevention, 2007). Thus, it should be stated that heart disease is a generally notable concern within health economics and biostatistics. Additionally, there are categorical variables that are noted to impact cardiovascular health metrics (e.g; cholesterol levels), such as physiological sex. This paper is therefore meant to explore some of the relationships with cardiovascular health data; namely, whether there notable differences between people with heart disease or no heart disease and/or people categorized “male” or “female” with regards to cholesterol, maximum heart rates, and resting blood pressure.

DATA DESCRIPTION

The data I have pulled for this project was pulled from Kaggle, which may not be the best primary academic source. (fedesoriano, 2021) However, it is a well-documented data set that is compiled from five separate datasets that are all available on the University of California - Irvine Machine Learning Repository, which makes the data convenient for use and allows for the analysis of five separate heart disease datasets simultaneously. (University of California - Irvine, 2019) That said, the combination of the five datasets should not cause any major concerns, and the data have not been transformed.

DATA EXPLORATION

Using the following SAS snippet, I imported the above mentioned .csv file into SAS:

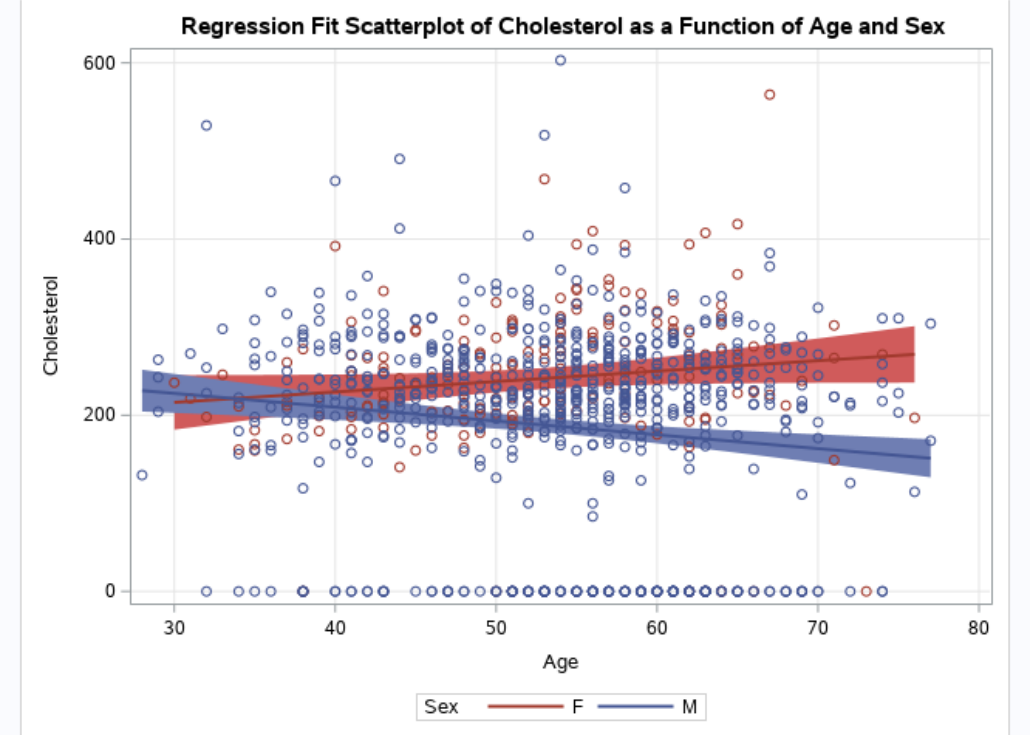
```
FILENAME CSV "/home/u49665201/sasuser.v94/STA3064/heart.csv" TERMSTR=CRLF;

/** Import the CSV file.  **/


PROC IMPORT DATAFILE=CSV
                    OUT=Heart
                    DBMS=CSV
                    REPLACE;
RUN;
```

STA 3064 PROJECT B

Some of the more interesting scatterplots were generated using the following segments of code:

I. The most interesting (and unexpected) scatterplot shows that cholesterol seems to increase with age for people categorized as female, and that cholesterol seems to decrease with age for people categorized as male. However, it can be noted that people categorized as male who are younger tend to have higher cholesterol than the people who are categorized as female in the same age bracket, up until the mid-30s.

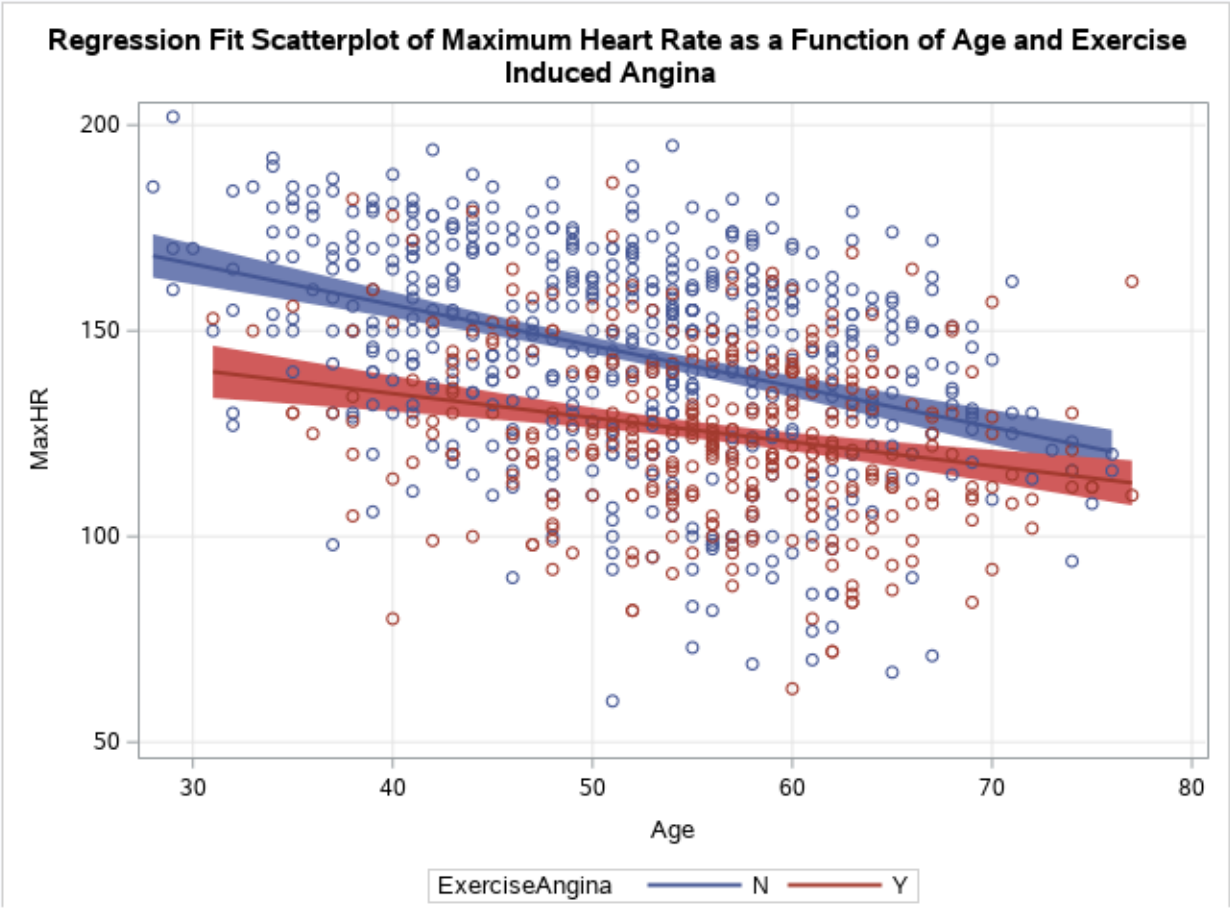
```
proc sgplot data=heart;
 reg x=age y=cholesterol
 /group=sex
 CLM alpha=0.05;
 xaxis grid;
 yaxis grid;
 title 'Regression Fit Scatterplot of Cholesterol as a Function of Age and Sex';
run;
```



STA 3064 PROJECT B

II. This is somewhat expected, but maximum heart rates tend to go down for everyone as they age. There does seem to be a difference between which whether someone has exercise induced angina, but they start to converge as age goes up, making for a less dramatic difference with elderly populations.

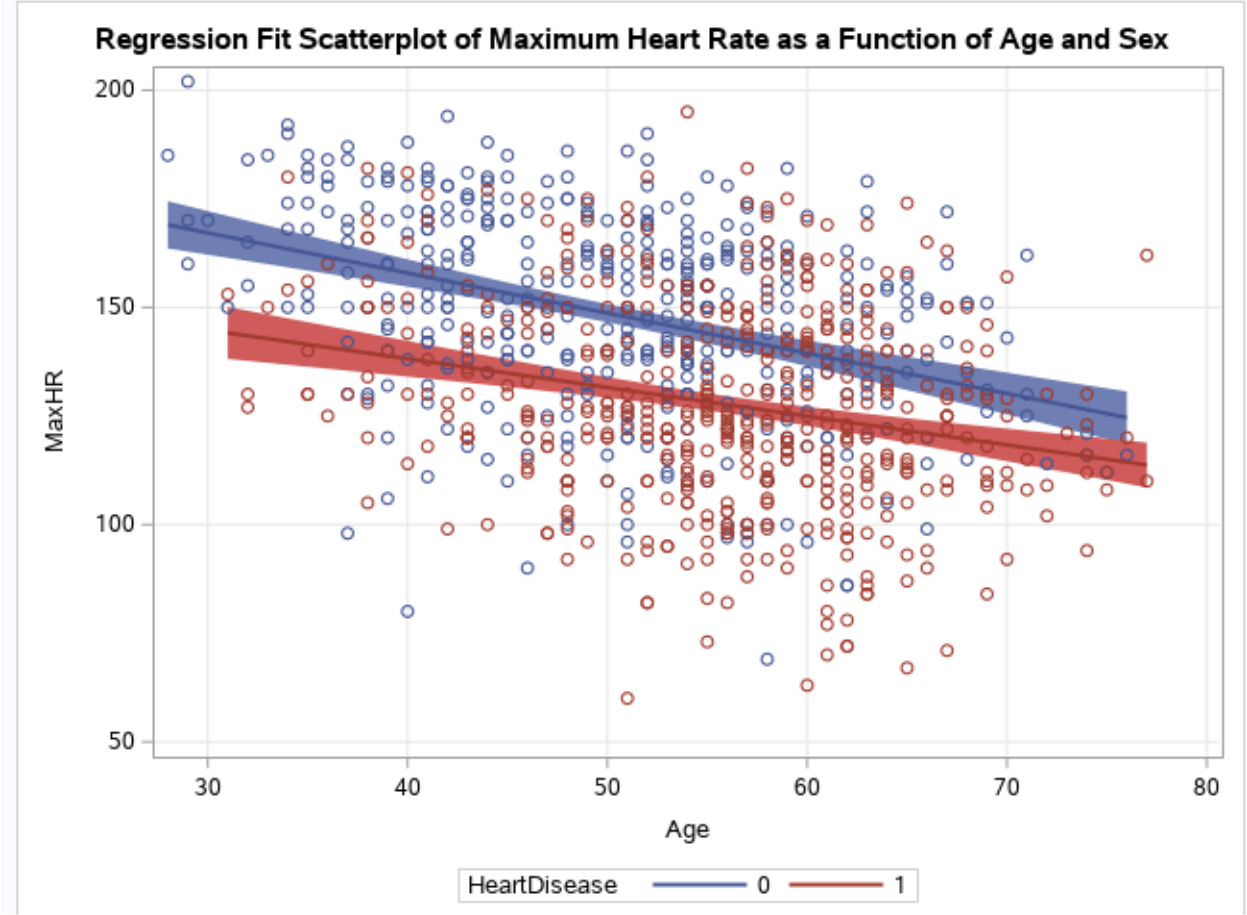
```
proc sgplot data=heart;
 reg x=age y=MaxHR
 /group=ExerciseAngina
 CLM alpha=0.05;
 xaxis grid;
 yaxis grid;
 title 'Regression Fit Scatterplot of Maximum Heart Rate as a Function of Age and Exercise Induced Angina';
run;
```



STA 3064 PROJECT B

I. This seems to be a relatively obvious statement to make, but there is a clear difference between the maximum heart rate for people assigned male and people assigned female at birth. Both groups have decreasing maximum heart rates with age.

```
proc sgplot data=heart;
 reg x=age y=MaxHR
 /group=HeartDisease
 CLM alpha=0.05;
 xaxis grid;
 yaxis grid;
 title 'Regression Fit Scatterplot of Maximum Heart Rate as a Function of Age and Sex';
run;
```



ANOVA

The most promising model in terms of the normality and linearity of the residuals seems to be MaxHR= 146.5446 -19.3425(if HeartDisease is present) + 4.06062 (if Sex='F') + ε. While there is another model within my code that does show a difference in cholesterol levels between the two defined physiological sex groups, as well as the presence of heart disease, that model had problems with non-normal residuals, which made it a less viable model.

I would also like to note that I did individual Tukey tests on the group "Heart Disease" and "Sex", and it appears that there is a more notable difference between the Heart Disease/No Heart Disease groups, as opposed to the "Male"/"Female" groups, even if Tukey does pick up on both of them. (The difference between Heart Disease and No Heart Disease appears is 148.15-127.66-20.49, whereas the difference between "Male" and "Female" is 146.14-134.33=11.81.)

It can also be noted that the F test statistic is 141.93 with a p-value of <0.0001, so we can clearly reject the null hypothesis that there are no differences between these categories.

```
proc glm data=Heart plots=diagnostics;
class HeartDisease(ref='0') Sex(ref='M');
model MaxHR = HeartDisease Sex/solution ss3;
means HeartDisease/tukey;
run;

proc glm data=Heart plots=diagnostics;
class HeartDisease(ref='0') Sex(ref='M');
model MaxHR = HeartDisease Sex/solution ss3;
means Sex/tukey;
run;
```

STA 3064 PROJECT B

The GLM Procedure

Class Level Information		
Class	Levels	Values
HeartDisease	2	1 0
Sex	2	F M

Number of Observations Read	918
Number of Observations Used	918

The GLM Procedure

Dependent Variable: MaxHR

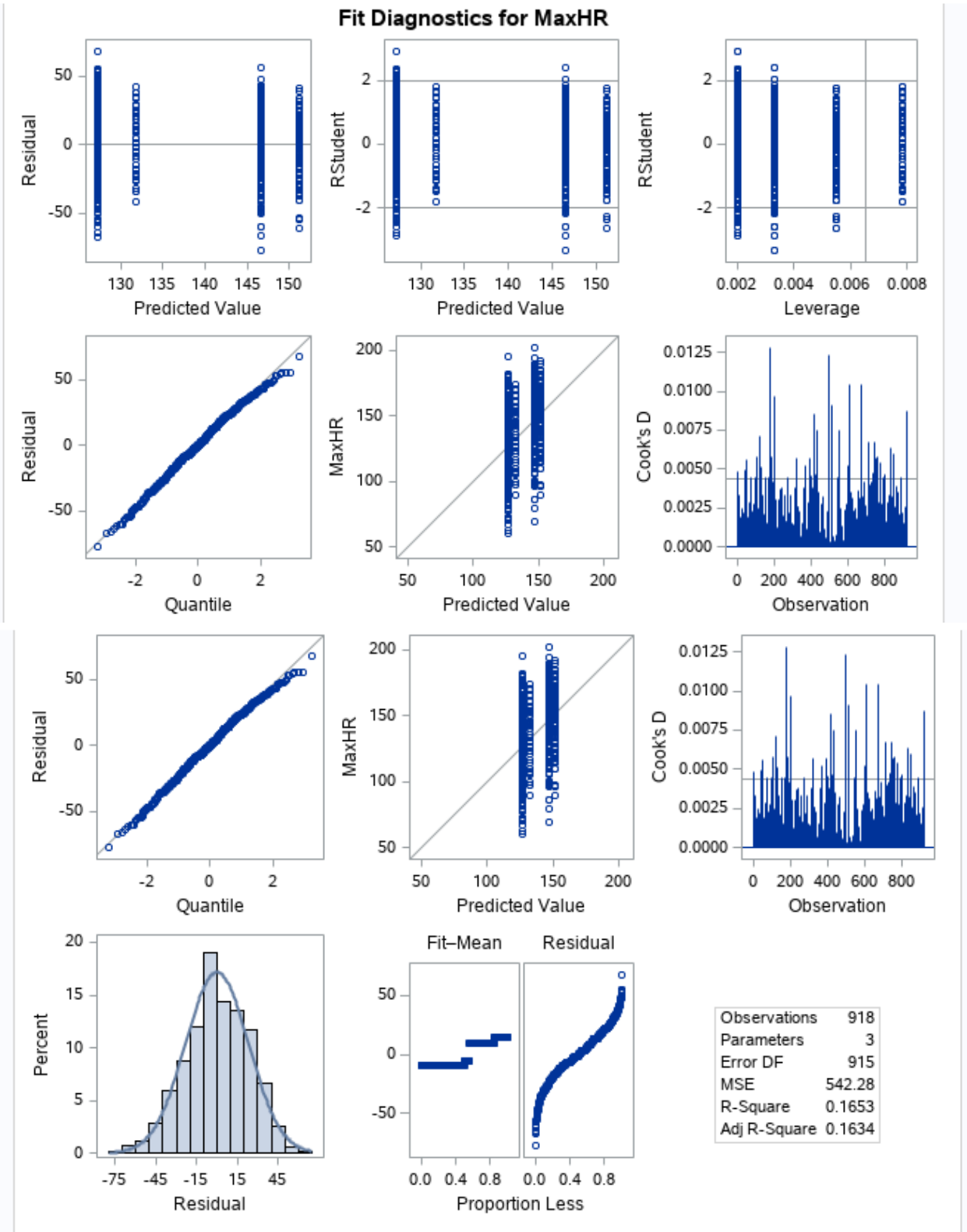
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	98240.6478	49120.3239	90.58	<.0001
Error	915	496184.9916	542.2787		
Corrected Total	917	594425.6394			

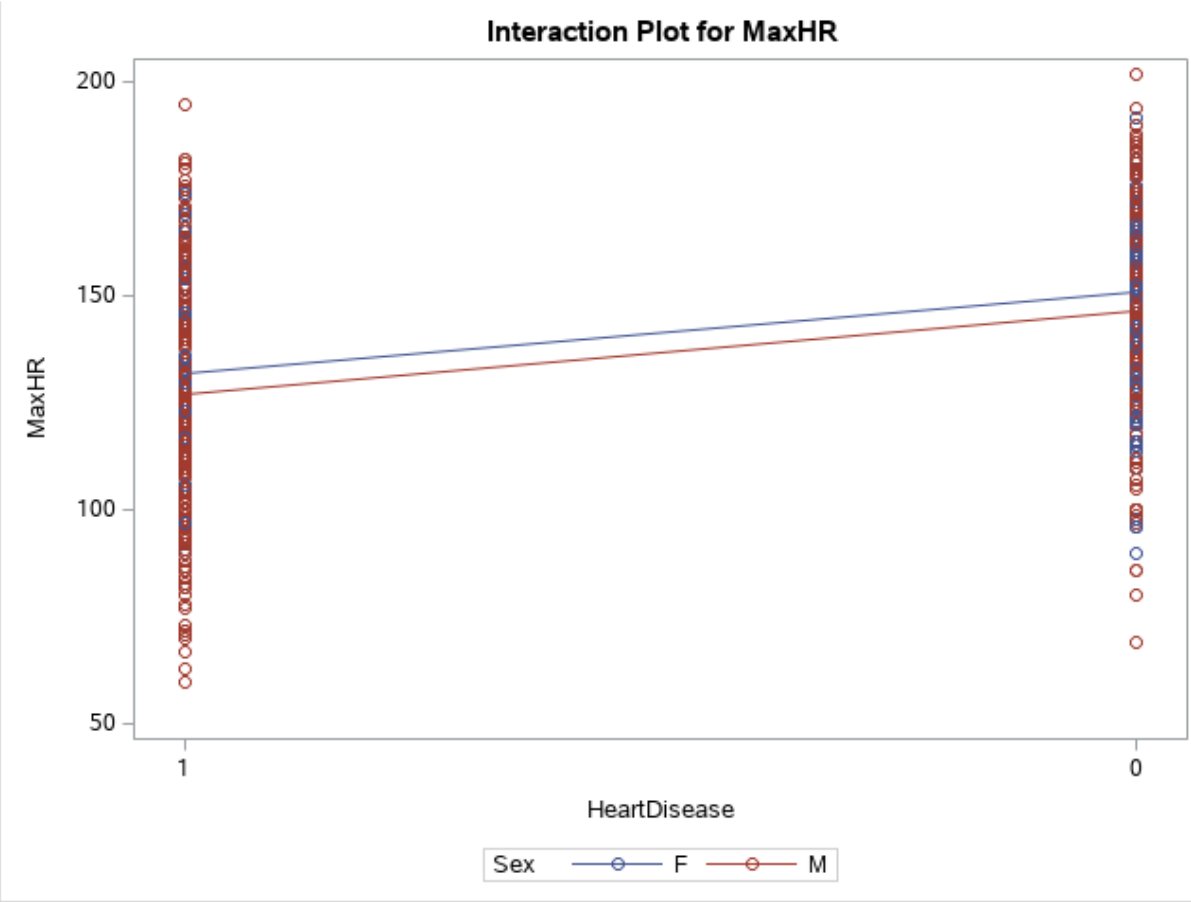
R-Square	Coeff Var	Root MSE	MaxHR Mean
0.165270	17.02141	23.28688	136.8094

Source	DF	Type III SS	Mean Square	F Value	Pr > F
HeartDisease	1	76965.40913	76965.40913	141.93	<.0001
Sex	1	2932.34736	2932.34736	5.41	0.0203

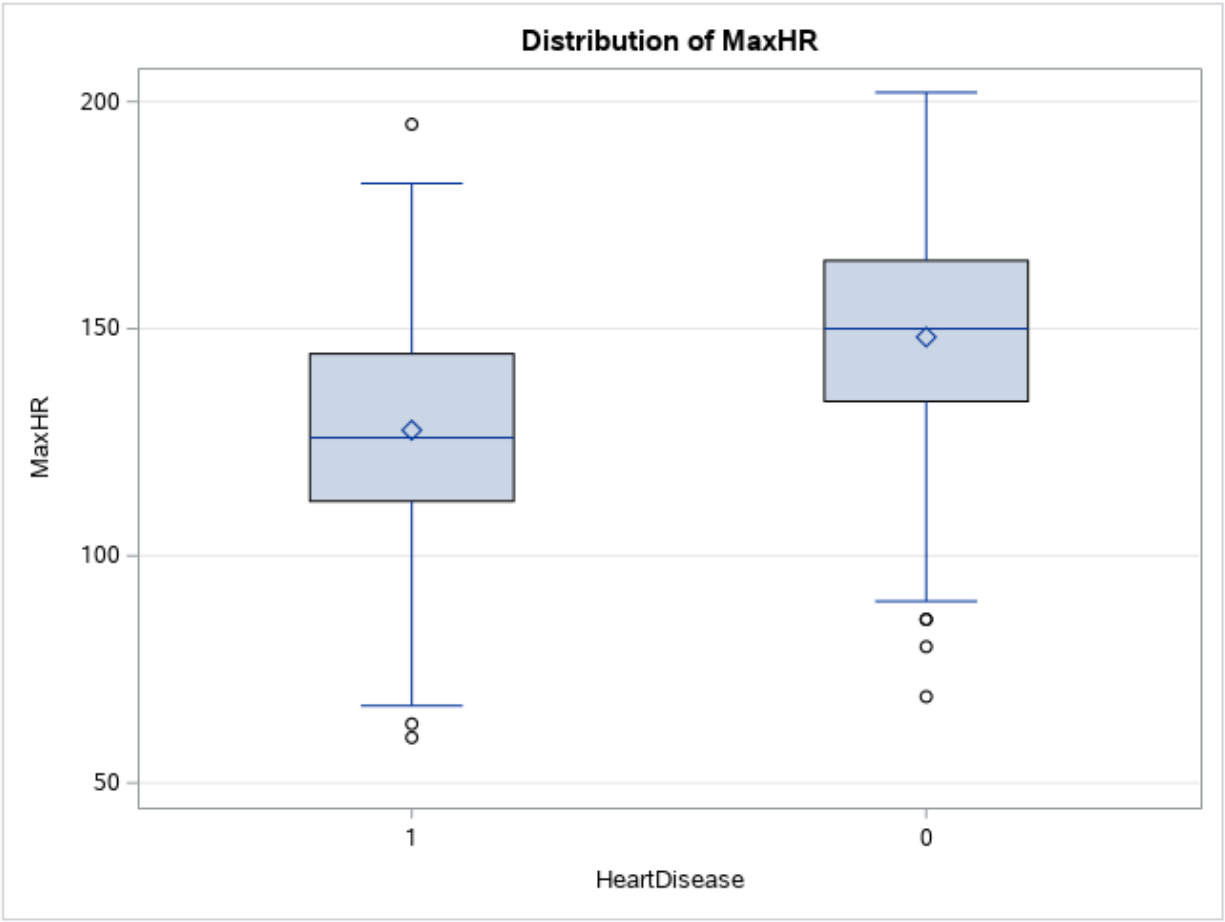
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	146.5446447	B	1.34162224	109.23	<.0001
HeartDisease 1	-19.3425053	B	1.62358937	-11.91	<.0001
HeartDisease 0	0.0000000	B	.	.	.
Sex F	4.6062634	B	1.98085335	2.33	0.0203
Sex M	0.0000000	B	.	.	.

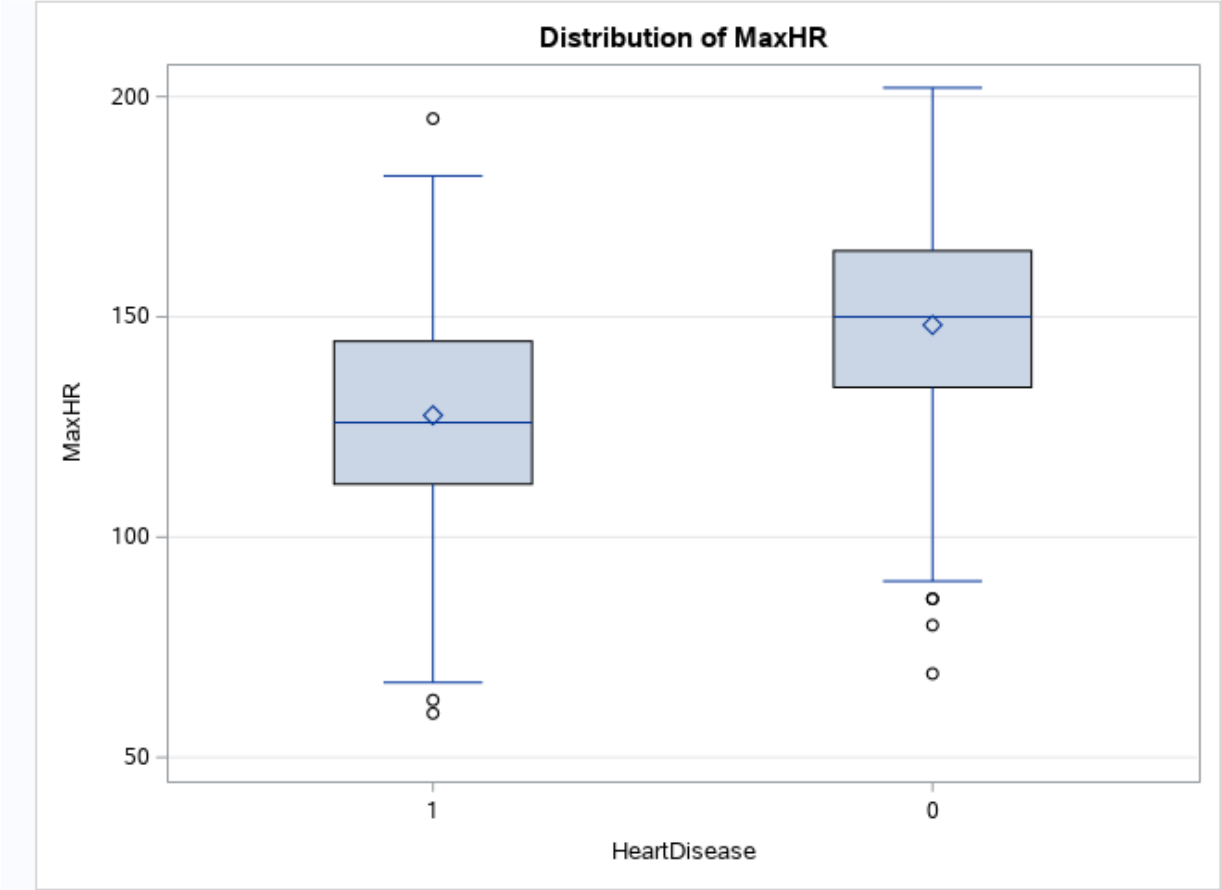
STA 3064 PROJECT B

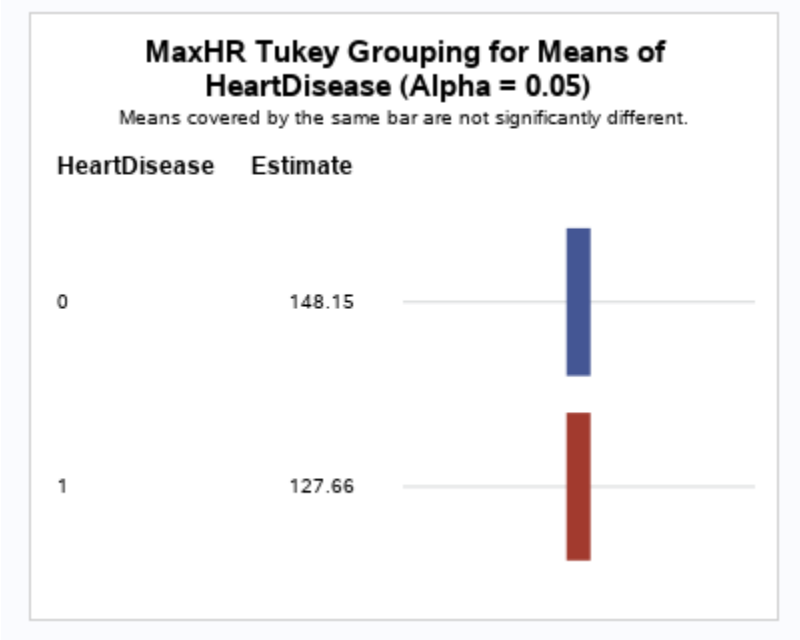


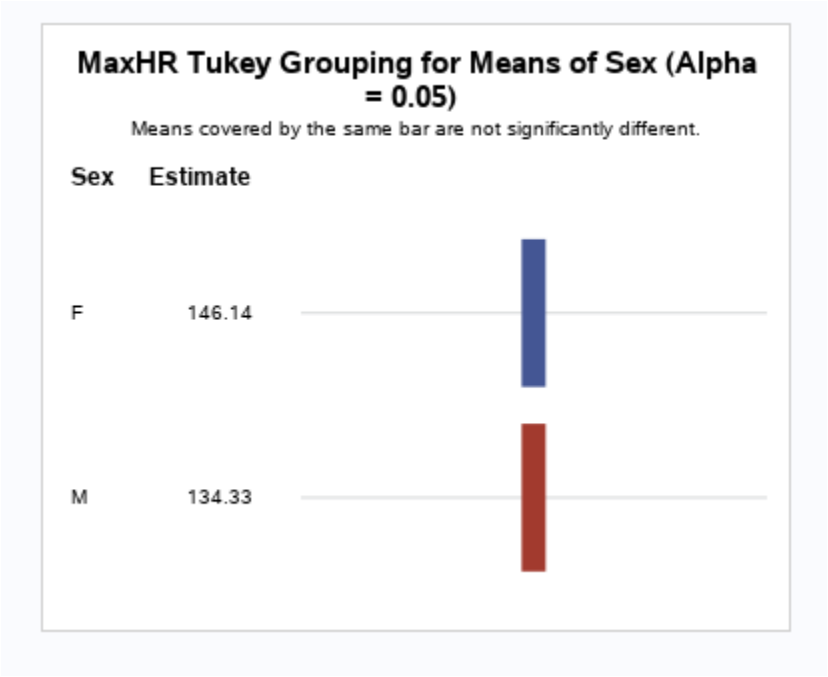

Interaction Plot for MaxHR
200
150
100
50
MaxHR
1
0
HeartDisease
Sex
F
M

STA 3064 PROJECT B




Distribution of MaxHR
200
150
100
50
MaxHR
1
0
HeartDisease


MaxHR Tukey Grouping for Means of HeartDisease (Alpha = 0.05)
Means covered by the same bar are not significantly different.
HeartDisease
Estimate
0
148.15
1
127.66



ANCOVA

Using the same model above, but doing ANCOVA on it with the following code, we can verify that the above model reasonable shows a difference between populations with heart disease and without heart disease, as well as patients marked as “male” or “female” within the population. It can be noted that there is a somewhat larger p-value for the population marked “female” (at 0.0057), as well as a somewhat lower F-test statistic in comparison to the other variables and their F-test statistics (7.68). However, it can still be deemed significant.

The model generated can be defined as MaxHR = 186.5986746 + -0.7966085(Age) + 5.2241415(Female) - 14.9272456(Heart Disease) + ε

```
proc glm data=heart plots=diagnostics;
class sex (ref='M') HeartDisease(ref='0');
model MaxHR = Age Sex HeartDisease/solution ss3;
run;
```

STA 3064 PROJECT B

The GLM Procedure

Class Level Information		
Class	Levels	Values
Sex	2	F M
HeartDisease	2	1 0

Number of Observations Read	918
Number of Observations Used	918

The GLM Procedure

Dependent Variable: MaxHR

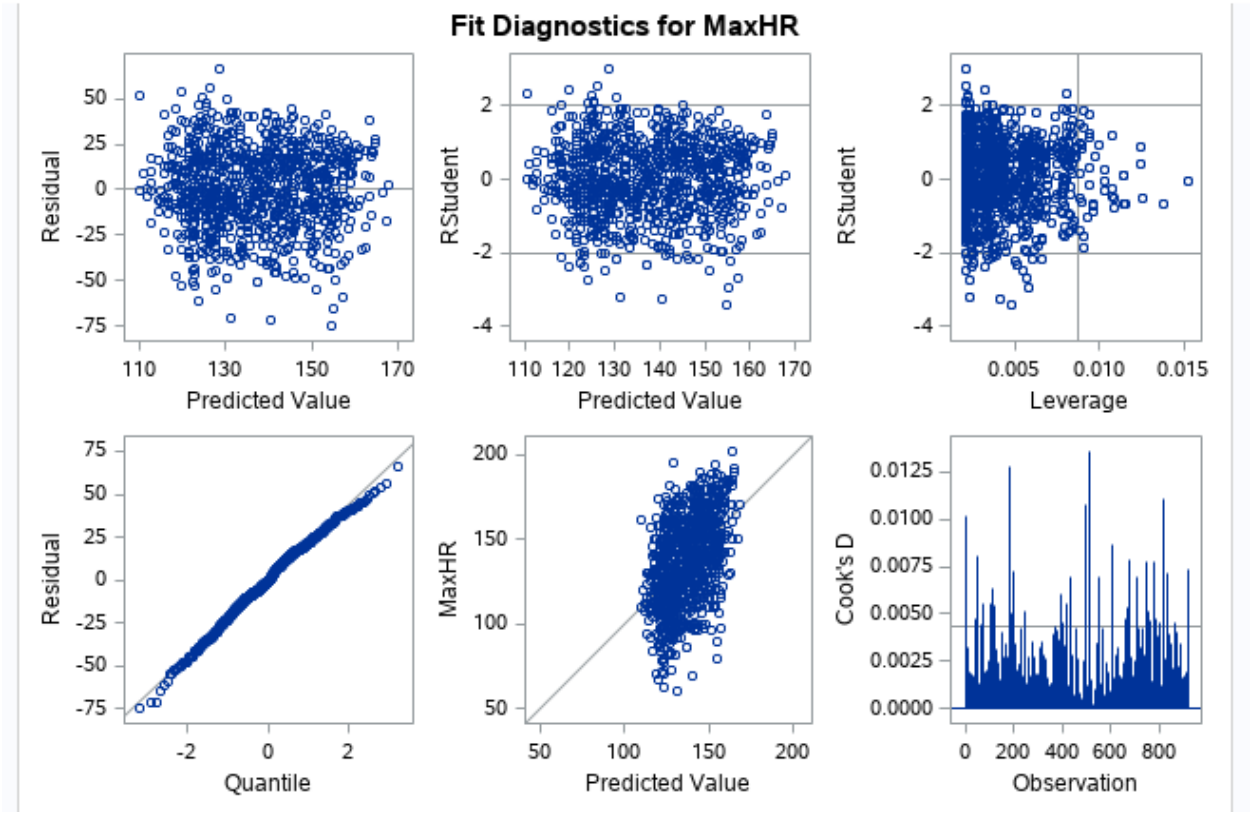
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	145844.7912	48614.9304	99.05	<.0001
Error	914	448580.8482	490.7887		
Corrected Total	917	594425.6394			

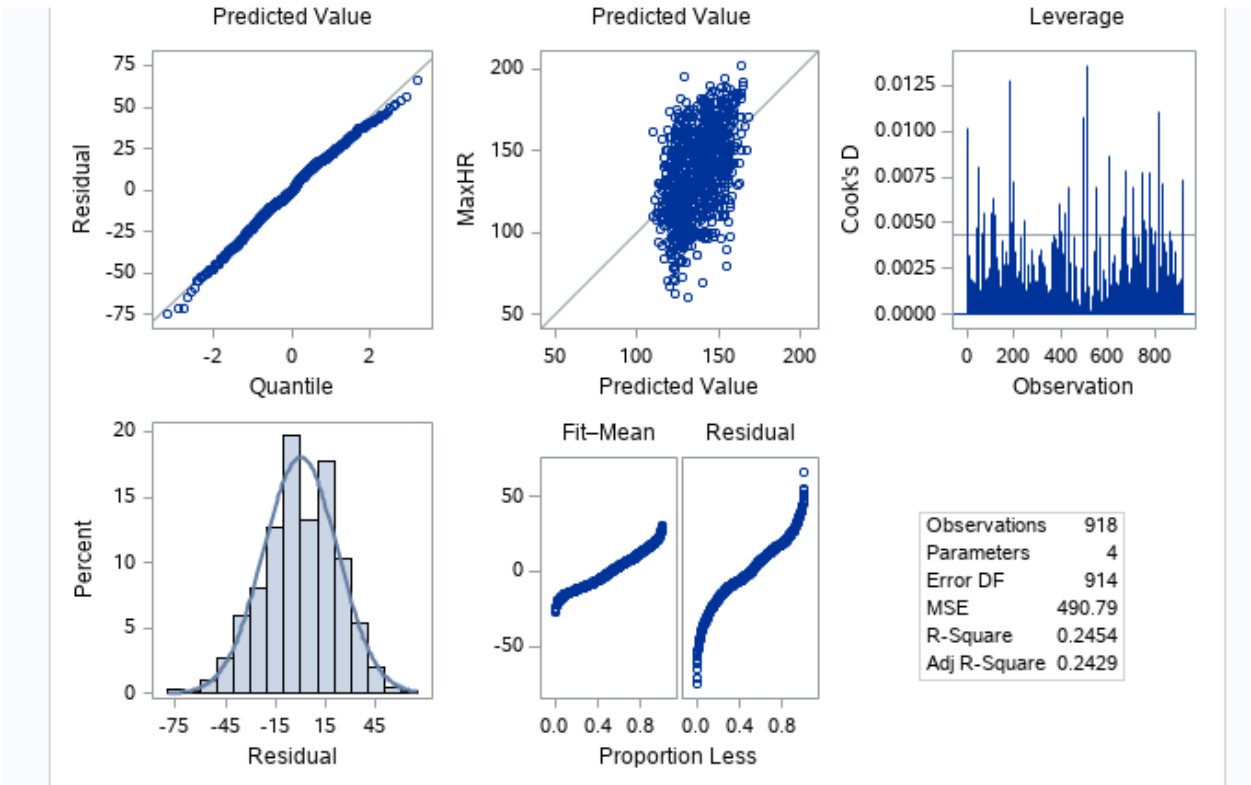
R-Square	Coeff Var	Root MSE	MaxHR Mean
0.245354	16.19315	22.15375	136.8094

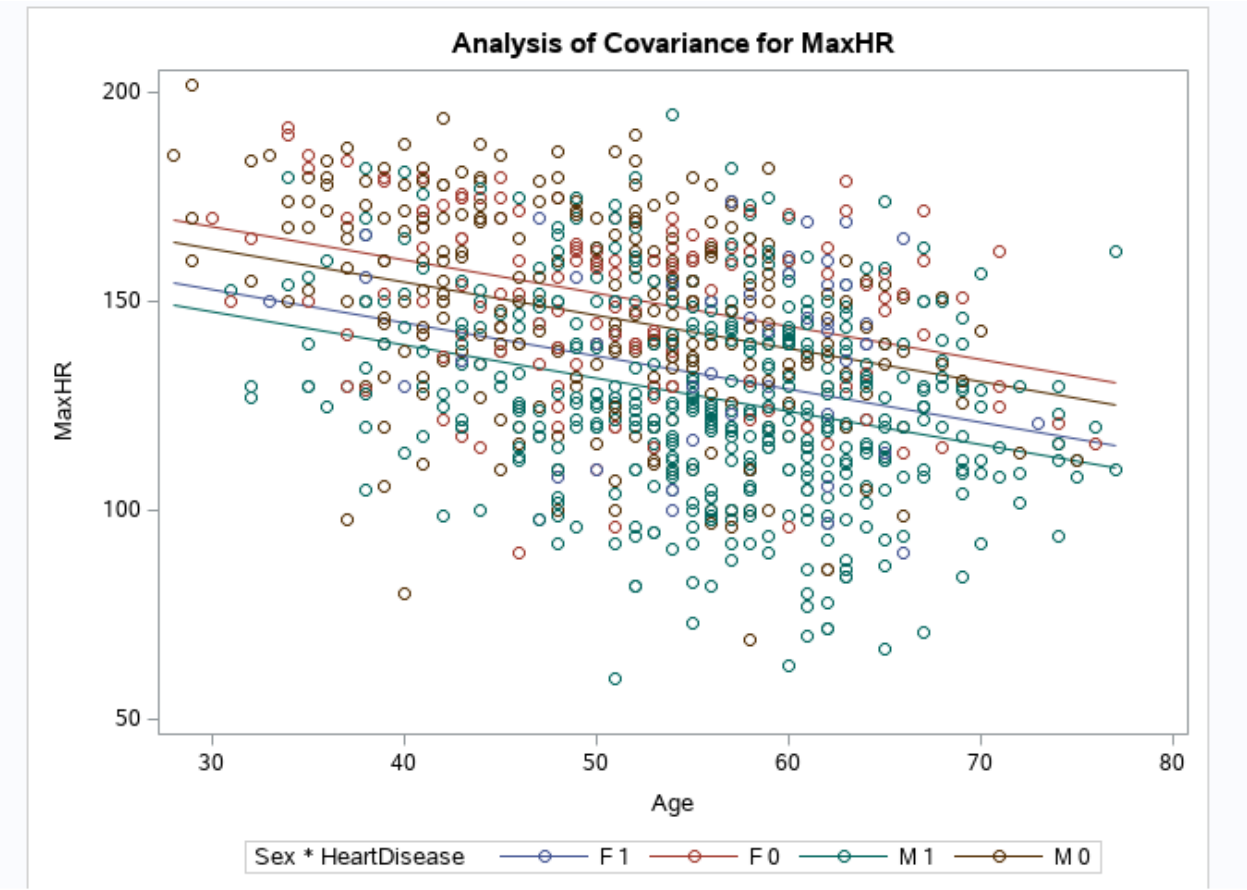
Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	47604.14339	47604.14339	97.00	<.0001
Sex	1	3767.61620	3767.61620	7.68	0.0057
HeartDisease	1	42276.84287	42276.84287	86.14	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	186.5986746	B	4.26254598	43.78	<.0001
Age	-0.7966085		0.08088535	-9.85	<.0001
Sex F	5.2241415	B	1.88551012	2.77	0.0057
Sex M	0.0000000	B	.	.	.
HeartDisease 1	-14.9272456	B	1.60833188	-9.28	<.0001
HeartDisease 0	0.0000000	B	.	.	.

STA 3064 PROJECT B







CONCLUSION

First and foremostly, it can be noted that the categorical variables analyzed in this sample appear to have the most impact on maximum heart rate, followed by cholesterol, with little to no difference in resting blood pressure. Secondarily, these measures of cardiovascular health appear to have less of a difference between binary physiological sexes as they do between whether someone has a cardiovascular disease or not. That said, none of the R-square values for the analyses went above .3, so further exporation of the data would be necessary to come up with a more accurate model. However, using ANOVA and ANCOVA on this data still leads to interesting insights.

REFERENCES

Center for Disease Control and Prevention. (2007, February 16). *CDC Mortality and Morbidity Weekly Report*. Retrieved from cdc.gov: https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5606a2.htm

Center for Disease Control and Prevention. (2021, October 19). *Center for Disease Control and Prevention National Center for Health Statistics*. Retrieved from cdc.gov: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved October 2021 from https://www.kaggle.com/fedesoriano/heart-failure-prediction.