DSC 423/323 Spring 2021
Heart Attack Risk Prediction
Teammates: Ximan Liu, Nair Varum, Borad Dhruv Kantilal, Ingilela Reuben, Chloe Tian

Abstract

We are working on a data analysis project for predicting heart disease. The project starts with raw data in the form of a .csv file and then changes it into Data Analysis. This project will use SAS code to analyze data for Heart Disease Prediction by following data science pipelines and data analytics. One of the leading causes of morbidity and mortality among the world's population is heart disease. One of the most important topics in the domain of clinical data analysis is the prediction of cardiovascular disease. This massive volume of data must be mined and filtered in order for us to be able to extract relevant information. The goal of this research is to compare alternative algorithms for predicting cardiac disease.

Introduction

Heart disease is the biggest cause of death in the world today. The development of computer algorithms that can forecast the existence of heart disease is expected to considerably reduce heart disease-related deaths, while early identification could lead to large cost savings in health care. Traditional statistical methods rely on a small number of variables gathered through experiments conducted under controlled settings to make judgments. Cardiovascular disease is a term used to describe problems involving restricted or obstructed blood vessels, which can result in a heart attack, chest pain (angina), or stroke. Other types of cardiac illness include those that impact the muscle, valves, or rhythm of your heart. We have data that classifies people as having heart disease or not based on certain characteristics. Machine Learning algorithms can determine whether a patient has cardiac disease based on a huge number of frequently complex characteristics acquired from a range of medical data banks. We'll aim to use this information to build a model that can predict whether or not a patient has this ailment.

CHD is the most common type of heart disease, claiming the lives of about 370,000 people each year. Every year, around 805,000 Americans suffer from a heart attack. 605,000 of them are first-time heart attacks, while 200,000 occur in patients who have already had a heart attack.[1] About one-fifth of all heart attacks are silent, meaning that the damage has already been done but the victim is unaware of it.

Methodology

DV(Y): OUTPUT (output is binary)
IV (Numeric variables): There are 14 variables in total which are, SEX(Gender of the person), CP(Chest Pain type chest pain type), TRTBPS(resting blood pressure (in mm Hg)), CHOL(cholesterol in mg/dl fetched via BMI sensor) FBS((fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)), RESTECG(resting electrocardiographic results), THALACHH(maximum heart rate achieved), EXNG(exercise induced angina (1 = yes; 0 = no)), OLDPEAK(Previous peak), SLP(Slope), CAA(number of major vessels), THALL(thal rate β-thalassemia cardiomyopathy) AGE(Age of the person)

Step 1: Check for missing values in all variables. No missing values were found.

Step 2: Handle variables by making into dummy variables SEX (MALE=1, FEMALE=0), EXNG= Exercise Induced Angina (1 = yes; 0 = no), FBS (Fasting Blood Sugar > 120 mg/dl) (1 = True, 0 = False), OUTPUT (0 = low chance of heart attack, 1 = high chances of heart attack), CP= Chest Pain Type (1= typical angina, 2= atypical angina, 3= non-anginal pain, 4= asymptomatic).

Step 3: Explore and separate the variables as dependent variables and independent variables. Quantitative or qualitative is a significant indicator to categorize the data.

Step 4: For our dataset, to investigate quantitative variables in univariate analysis, we could use histograms and boxplots. As we all know, the distribution of the variables and the trend can help us consider the data well. Considering graphically depicting groups of numerical data the quartiles, boxplot is also a good analysis strategy for our exploration.

Step 5: Run a full model and do Before analysis and summarize the report, we need to check out extreme multicollinearity with VIF values. If VIF is greater than 10, it means the regression coefficient is not precise.

Step 6: Remove non-significant variables, outliers and influential points by p-values and studentized residual plots .

Step 7: Select 75% of data for training set (227 values) and 25% test set (75 values)

Step 8: Use stepwise, forward and backward model selection for training set.

Step 7: Select 75% of data for training set (227 values) and 25% test set (75 values)

Step 8: Use stepwise, forward and backward model selection for training set.

Step 9: Select the best model by comparing performance between test and train set.

Rahman, Rashik. “Heart Attack Analysis & Prediction Dataset.” *kaggle*, 2021, https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

Analysis, Results and Findings

Explanatory analysis

The Univariate Procedure distribution of age: Mean is 54.36634, std is 9.082101, skewness is -0.2024634. Which is skewed to the left, median is 55 and mode is 58, the min is 29, Q1 is 42, Q3 is 61, the maximum is 77. Figure 2.

The Univariate Procedure distribution of trtbps: Mean is 131.6238, std is 17.53814, skewness is 0.71376844 which is skewed to the right, median is 130, mode is 120, the min is 94, Q1 is 110, Q3 is 140, the maximum is 200.Figure 3.

The Univariate Procedure distribution of chol: Mean is 246.264, std is 51.83075, skewness is 1.14340082 which is skewed to the right, median is 240, mode is 197, and the min is 126, Q1 is 188, Q3 is 275, the maximum is 564. Figure 4.

The Univariate Procedure distribution of thalachh: The mean is 149.646865, std is 22.9051611, skewness is -0.5374097 which skewed to the left, the median is 153, the mode is 162, the min is 71, the Q1 is 116, the Q3 os 166 the maximum is 202. Figure 5.

The Univariate Procedure distribution of oldpeak: The mean is 1.03960396, the std is 1.16107502, the skewness is 1.26971993 which is skewed to the right, the median is 0.800000, the mode is 0, and the min is 0, and Q1 is 0, Q3 is 1.6, the maximum is 6.2. Figure 6.

The boxplot vs IDVS distribution of trtbps by output:

It is seen that people with lower resting blood pressure have higher chances of getting heart attack, average value for those is 128.09, while people with lower chance of heart attack have an average of 138.16. There were some people who had resting blood pressure of 200 and still had a low chance of getting a heart attack but there were exceptions where blood pressure of 170 had more chances. In the article,“What Happens to Blood Pressure During a Heart Attack?” It explains that “Blood pressure is not an accurate predictor of a heart attack. Sometimes a heart attack can cause an increase or decrease in blood pressure, but having a change in blood pressure reading doesn’t always mean it’s heart-related.” [4] From our findings however we find that resting blood pressure(TRTBPS) is an influential variable in determining risk of heart attack. But from a medical standpoint however, a blood pressure of 138.16 would put one in the prehypertension class which puts one at risk for a heart attack. Figure 7.

The boxplot vs IDVS distribution of chol by output:

People having more chances of heart attack have an average cholesterol level of 257.33 but there are some cases when cholesterol level hits 564. Whereas, people with less chance of heart attack have an average cholesterol level of 237.97. Also people with a low cholesterol level of 126 had more chances of heart attack while the other people had a minimum cholesterol level of 164. “With high cholesterol, you can develop fatty deposits in your blood vessels. Eventually, these deposits grow, making it difficult for enough blood to flow through your arteries. Sometimes, those deposits can break suddenly and form a clot that causes a heart attack or stroke.”[3] which is one of the main points we are trying to portray with the findings in our dataset. This further validates that higher levels of cholesterol does in fact lead to a higher chance of a heart attack. Figure 8.

The boxplot vs IDVS distribution of age by output:

Older people have less chance of heart attack than younger people, according to the mean of older people 57.219 which is 0, and the younger people 53.907 which is 1. Figure 9.

The boxplot vs IDVS distribution of sex by output:

Sex (gender of person) 1 = male; 0 = female. The data seems that females have higher risks of getting heart attacks, because the mean of male is 0.6744 and female is 0.7813. And the standard deviation shows that the male is higher than the female. Statistically, older people have fewer heart attacks than younger people, and then Q1 and Q3 versus older people are also correlated with lower heart attacks than younger people. But older people note that older people are more likely to have heart disease at age 71 than they are at age 69. So overall, the younger you are, the more likely you are to have a heart attack, but the older you are, the more likely you are to have a heart attack. Figure 10.

The boxplot vs IDVS distribution of cp by output:

cp: chest pain type

(1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)

People with a higher chance of having a heart attack are more likely to have (1) typical angina and (2) atypical angina, and the mean is 1.3758 which means the appearance of an uncomfortable situation. Vice versa, people with a lower chance of having a heart attack have the mean value of 0.4783 which means they barely suffer from angina. Besides, people with a higher chance of having a heart attack have a median value of 2, which means atypical angina is a somewhat average common symptom. On the other hand, people with a lower chance of having a heart attack have the median value of 0. Figure 11.

The boxplot vs IDVS distribution of fbs by output:

fbs (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

People with a slightly higher chance of not having a low fasting blood sugar than having a high fasting blood sugar, when fbs = 0 the mean is 0.1875 and fbs = 1 the mean is 0.1628. Also according to the data show that the range q3 and q1 datamax and datamin have the exactly the same value. Which means the fasting blood sugar does not cause particularly large heart attacks. Figure 12.

The boxplot vs IDVS distribution of restecg by output:

restecg: resting electrocardiogram results

(0 = normal, 1 = ST-T wave anomaly, 2 = Show probable or definite left ventricular hypertrophy according to ESTES criteria)

People with a higher chance of having a heart attack have the median value of 1 and the mean value of 0.5939. Vice versa, people with a lower chance of having a heart attack have the median value of 0 and the mean value of 0.4493. It shows the problem that people with a higher chance of having a heart attack are more likely to have ST-T wave anomaly but people with a lower chance of having a heart attack tend to be normal on resting electrocardiogram results. Figure 13.

The boxplot vs IDVS distribution of thalachh by output:

The maximum heart rate achieved for people with more chances of heart attack is 187 and minimum of 105 with an average of 159.35. People with lower chances of heart attack had a maximum heart rate of 177 and minimum of 71 with an average of 139.16. Figure 14.

The boxplot vs IDVS distribution of exng by output:

exng (exercise included angina) (1 = yes; 0 = no). The numbers for exercise excluding angina were much higher than those for exercise including angina, when exercise is not included the mean is 0.4083 and the exercise included the mean is 0.093. The value of the datamin and datamax is exactly the same: the max is 1 and the min is 0. The q3 when exercise included is 0 and exercise not included is 1. Figure 15.

The boxplot vs IDVS distribution of oldpeak by output:

The data shows that if the average old peak is higher there is less chance of getting heart attack. The maximum old peak for people having less chance of heart attack is 6.2 with an average of 1.52. And for people with a higher chance of heart attack has a maximum old peak of 3.5 with an average of 0.75. Figure 16.

The boxplot vs IDVS distribution of caa by output:

With the higher average value of caa (number of major vessels) there is less chance of heart attack with an average of 1.09 and for people with more chance of heart attack have an average old peak of 0.34. Figure 17.

Collinearity Check

With the correlation matrix we check all variables. There was no multicollinearity among the variables. It was verified by checking that all VIF values are less than 10 which concluded that. Figure 18.

Studentized plots

The residual plots Sex VS residual the plot :are not random because the points only located along 0 and 1. So can concluded that it's not a linear, constant variance and independent. Figure 20.
The residual plots Cp VS residual the plot: are not random because the points are only located along 0, 1,2 and 3. So can concluded that it's not a linear, constant variance and independent. Figure 21.
The residual plots trtbps VS residual the plot:are not random because the points are not randomed around the zero line.So can conclude that it's not a linear, constant variance and independent. Figure 22.
The residual plots thalachh VS residual the plot: are not random because the points are not randomed around the zero line.So can conclude that it's not a linear, constant variance and independent. Figure 23.
The residual plots exng VS residual the plot: are not random because the points only located along 0 and 1. So can concluded that it's not a linear, constant variance and independent. Figure 24.
The residual plots oldpeak VS residual the plot: are not random because the points are not randomed around the zero line. The most points around the left-hand side, and like a funnel shape. So can conclude that it's not a linear, constant variance and independent. Figure 25.
The residual plots caa VS residual the plot: are not random because the points only located along 0,1,2,3 and 4. So can concluded that it's not a linear, constant variance and independent. Figure 26.
The residual plots thall VS residual the plot: are not random because the points only located along 0, 1, 2 and 3. So can concluded that it's not a linear, constant variance and independent. Figure 27.

Linearity and normality look satisfied by looking at the probability plot. With adj-R2 of 0.49 and RMSE of 0.355. Figure 29.

Independence

The residual plots and scatter plot can see that all plots show points having a pattern funnel shape, because it's spread increases and it's not spread constant so it's not independent.

Linearity
According to the boxplot diagram the relationship between variables and CDF of studentized residuals and the normal cumulative distribution is almost a straight line. There is a strong positive correlation between residuals vs variables. So, by proving that it's linearity.
It can be proven that there is a strong positive correlation coefficient. So, by proving that it's linearity.

Constant variance
According to the diagram and the residuals vs predicted values because the plot is not randomly scattered around the zero line so it's not constant variance.

Normality
According to the diagram, I can get the points close to the line and the pattern of the spread shows a 45-degree line and it's almost a straight line so it's normal.

Outliers and Influential points

Based on the Cook's D metric. There were 13 influential points in total. But only one was removed as it was significant. The other influential points were kept in the dataset as in healthcare, sometimes there are unique cases appearances. We did not want to change to much of dataset as it would be against the ethics of data science and to avoid discrimination. Figure 30

Can your model be improved?

More data would be needed for this set in order to draw a better conclusion, perhaps having more determining factors would help as well. With the current data we are satisfied that the cholesterol level and risk of heart attack have a positive correlation, but other determining factors such as age and blood pressure don't make sense. I think we should also have more gender variants like transgender, non-binary, etc. to not show gender discrimination. Also, a lot of the data used contradicts the general consensus in the medical field meaning that the author may have mixed up the dummy variables included in the set.

Model Equation.

The variables that predict the response variable are thalach, cp, caa, sex, slp, exng, thall, trtbps. Figure 31.

Output = 0.6552+0.0033 (thalach)+0.1157 (cp)-0.1138(caa)-0.1884 (sex)+0.13323 (slp)- 0.1568 (exng) - 0.1323 (thall) - 0.0026 (trtbps)

Dummy Vars: cp → 1 - typical angina, 2 - atypical angina, 3 - non-anginal pain, 4 - asymptomatic
Sex → 1 - Male , 0 - Female
Exng → 1 - Yes, 0 - No
Thall → 1 - fixed defect, 2 - normal, 3 - reversible defect

Prediction 1: A male with maximum heart rate of 162, atypical angina chest pain with 2 major vessels, slope of 0 with presence of exercise included angina, normal thal rate and resting blood pressure of 131.

The prediction resulted in a value of 0.5454. Since there was no transformation used there is no need to transform this value.

Final Model and Predictions

Using the selected regression model to examine the relationship and associations amongst the variables in this study, it can be seen that there is a high correlation between the independent variables cp, thall, and exng when in comparison to our response variable output. CP (Chest Pain type) has a positively moderate correlation when it comes to comparing it to the chance at a heart attack. A similar observation can be made about Thall (thal rate β-thalassemia cardiomyopathy) as well. The strongest predictors for the response variable would be CP (Chest Pain type), thalachh (maximum heart rate achieved), exng (exercise induced angina), and oldpeak.

Prediction 1: A male with maximum heart rate of 175, atypical angina chest pain with 1 major vessel, slope of 1 with presence of exercise included angina, normal thall rate and resting blood pressure of 120.

The prediction resulted in a value of 0.44603

Prediction 2: A male with maximum heart rate of 152, atypical angina chest pain with 2 major vessels, slope of 0 with presence of exercise included angina, normal thall rate and resting blood pressure of 89.

The prediction resulted in a value of 0.1607

Train and Test Model

I split the data into a test and train set in order to test the performance of the model. After splitting the dataset in 75:25 ratio, the training set consists of 227 observations while the test set contains 75 observations. We used the train set on our model, using a seed value of 47274 and samprate of 0.75. The model resulted in a R-square value of 0.5179 and adjusted-R2 value of 0.5003. P-value associated with f-value if less than 0.05, meaning that there is at least one significant variable in the model to predict the output. Figure 32

According to the proposal we also see that people who have a higher chance of heart attack have an average of 52. And chest pain of type 2 (atypical angina). Figure 33, 34

Sensitivity: 79% correctly +ve labeled by our program to all who have a high chance of heart attack.

Accuracy: how many did we correctly label? 82%

Specificity: 84% are correctly -ve labeled by the program to all who are healthy in reality.

Precision: 87% of the correctly +ve labeled by our program to all +ve labeled.

Future Work

In our data set we found that all of the variables are influential in determining the risk of heart disease however, some of the results were not lining up with the general consensus. According to 1h, we found that people who were younger were more at risk for a heart attack. There were quite a few discrepancies found in the data that did not make sense such as younger age and according to 1f, The higher resting blood pressure would determine a higher risk of heart attack. A few articles from the Mayo Clinic and Everyday Health explained how some of the data results contradict real world findings. We also removed the influential points to see if the conclusion from the data would change, however the results stayed the same. The general consensus from our group is that we would need more data, and more determining factors in the data set to draw a better conclusion from than the one we are using now. From what we can determine from the data we can conclude that an older man with high blood pressure, low cholesterol, low blood sugar, with no typical angina or atypical angina, an average low heart rate, and who exercises regularly will have a low risk of a heart attack according to 1.

In addition, we found through reading the literature that in addition to physically measuring the body level and indicating the state of the body, psychological conditions may also play an important role in the prediction of heart disease, and this is a very potential and challenging Research direction, because it is not like blood pressure or blood sugar that can be quantified obviously.

Generally speaking, the psychological condition is related to the perception of heart disease. According to the cited research, the psychologically optimistic people will underestimate their heart health problems more, and the psychologically pessimistic people will overestimate the risk of heart health problems more. [5] But this situation is also a double-edged sword. People who are pessimistic are more likely to have poor health. Furthermore, the optimism of dealing with the problem of heart disease is greatly affected by age, and age is also one of the important factors affecting heart disease.

Research References

[1] Virani, S.S. *et al.* (2020) Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association. *Circulation* 141, e139–e596

[2] Cherney, Kristeen. "What Happens to Blood Pressure During a Heart Attack?" *Healthline*, Healthline Media, 27 Mar. 2017, https://www.healthline.com/health/blood-pressure-changes-during-heart-attack.

[3] Mayo Clinic. "High Cholesterol - Symptoms and Causes - Mayo Clinic." *Mayo Clinic*, 13July2019,

https://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/symptoms-causes/syc-20350800#:~:text=Your%20body%20needs%20cholesterol%20to,to%20flow%20through%20your%20arteries.

[4] Orenstein, Beth. "Understanding the Stages of Hypertension | Everyday Health." *EverydayHealth.Com*, EverydayHealth, 19 Apr. 2009, https://www.everydayhealth.com/hypertension/understanding-the-stages-of-hypertension.aspx.

[5] N E Avis, K W Smith, and J B McKinlay. Cambridge Research Center, American Institutes for Research, MA. "Accuracy of perceptions of heart attack risk: what influences perceptions and can they be changed?", American Journal of Public Health 79, no. 12 (December 1, 1989): pp. 1608-1612.
https://doi.org/10.2105/AJPH.79.12.1608

Appendix Section

Figure 1: Scatterplot Matrix

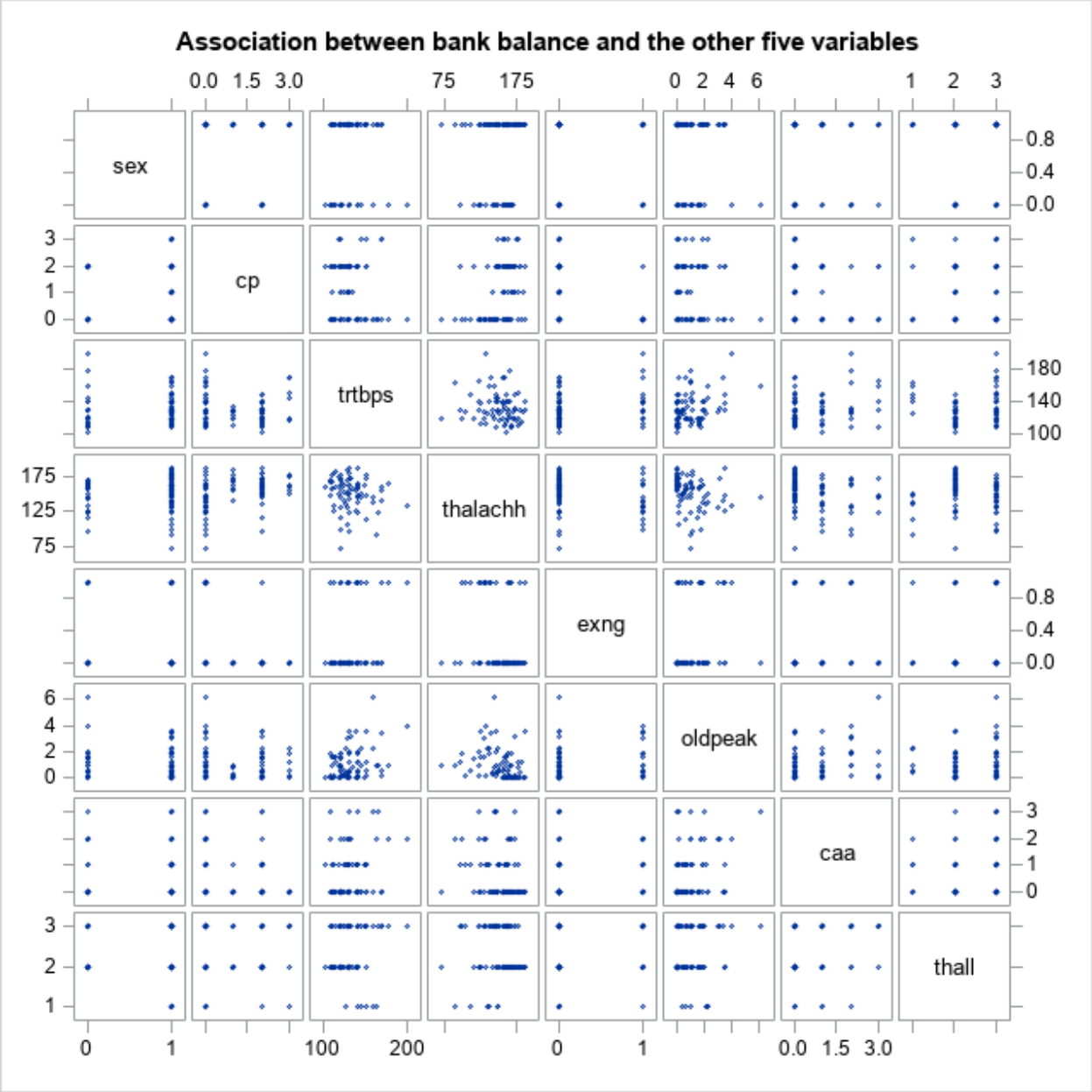


Figure 2

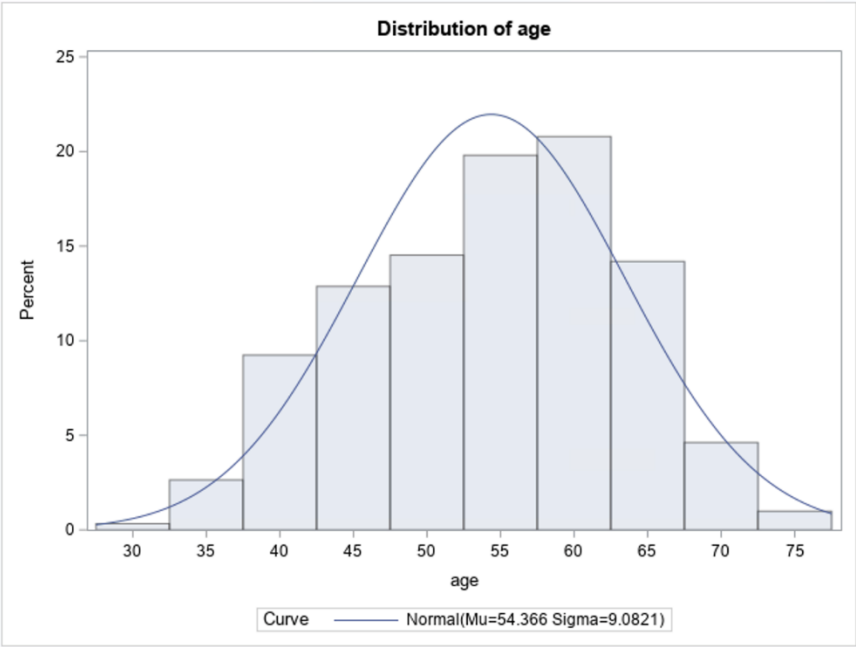


Figure 3

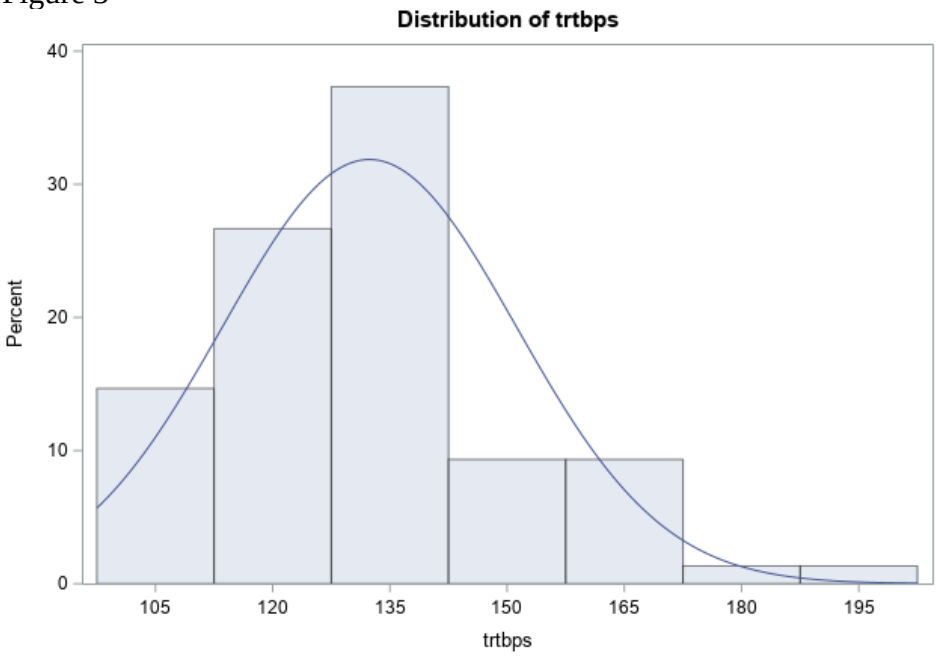


Figure 4

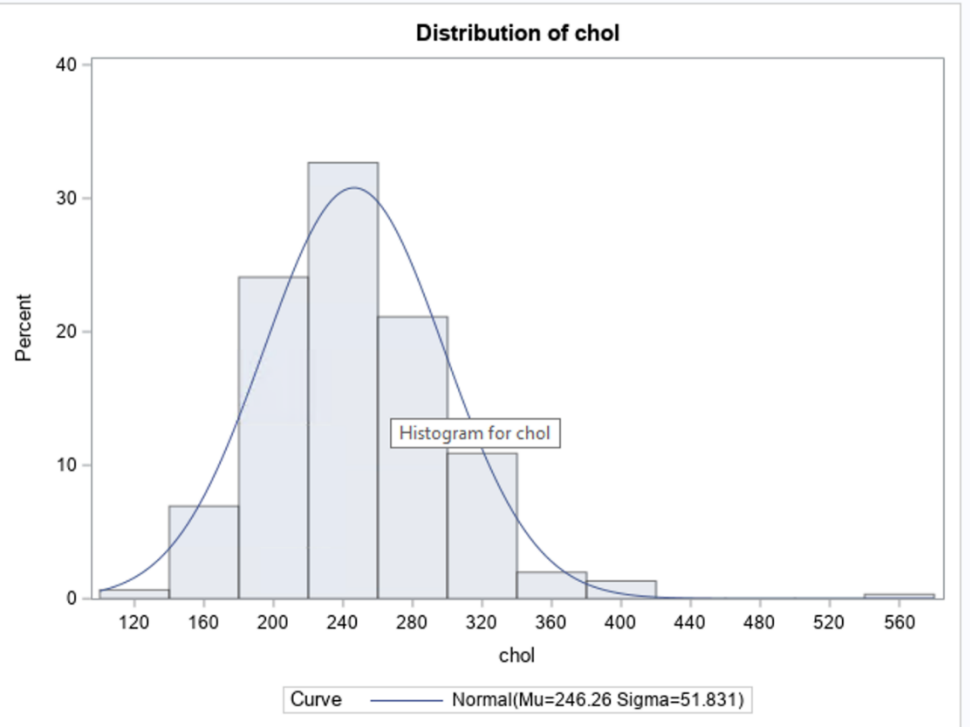


Figure 5

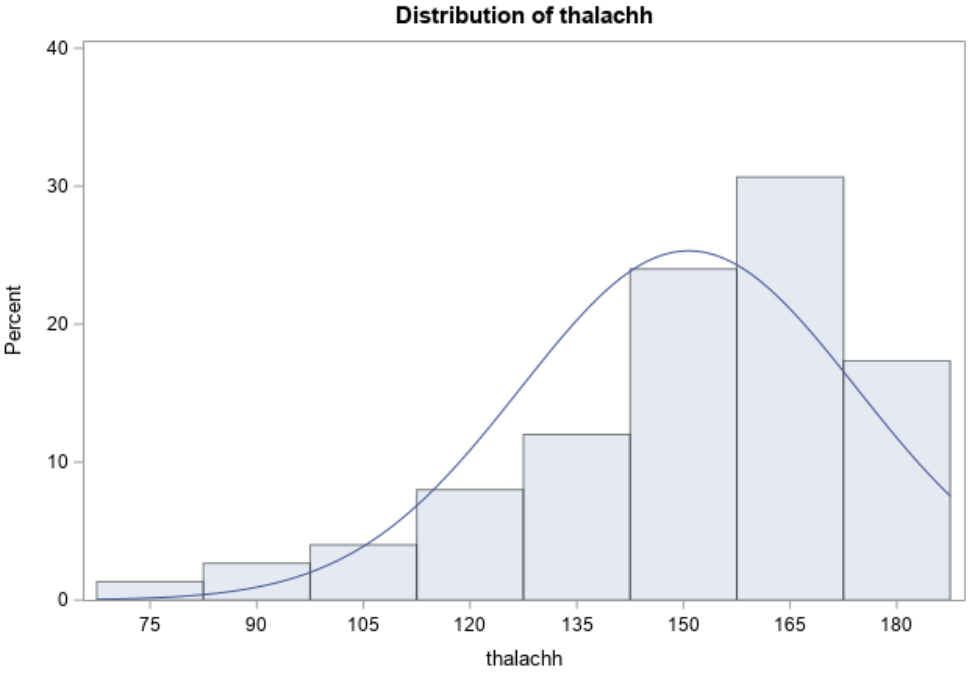


Figure 6

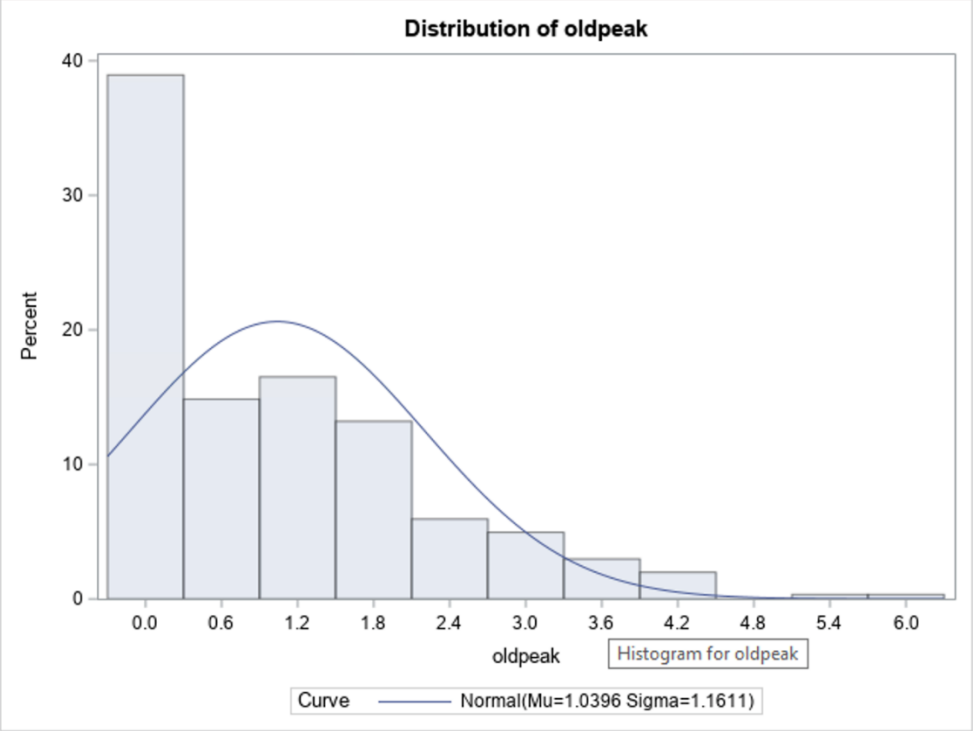


Figure 7

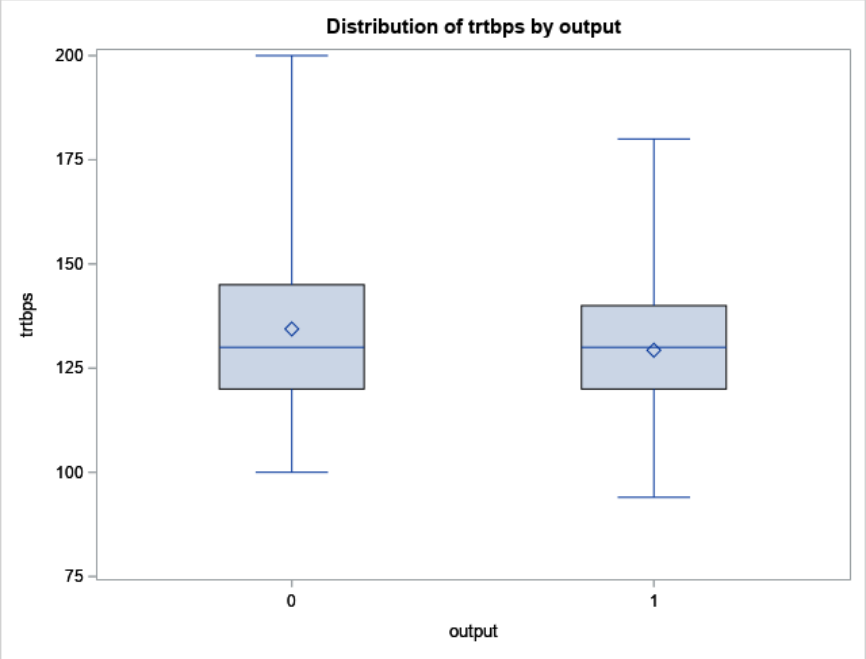


Figure 8

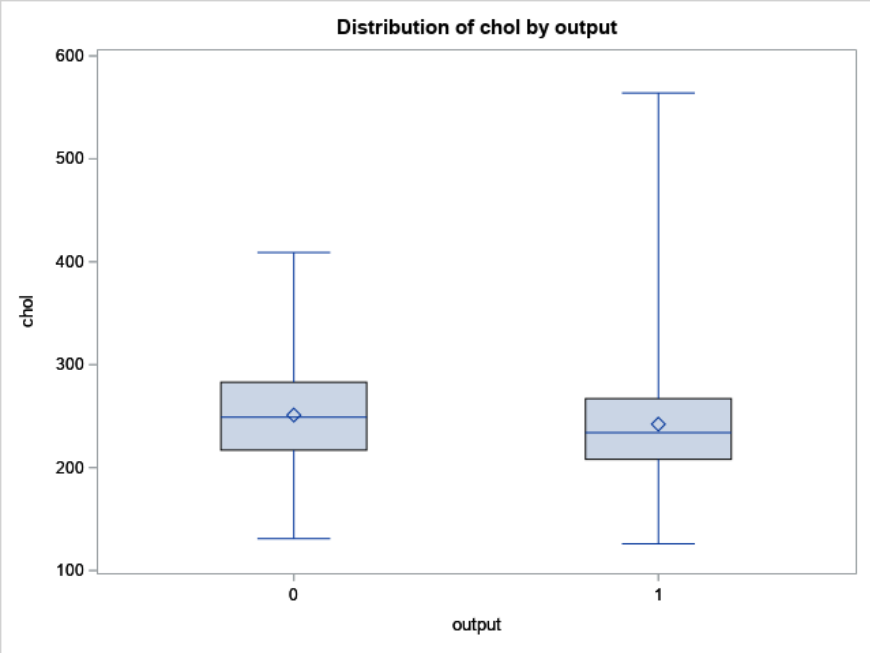


Figure 9

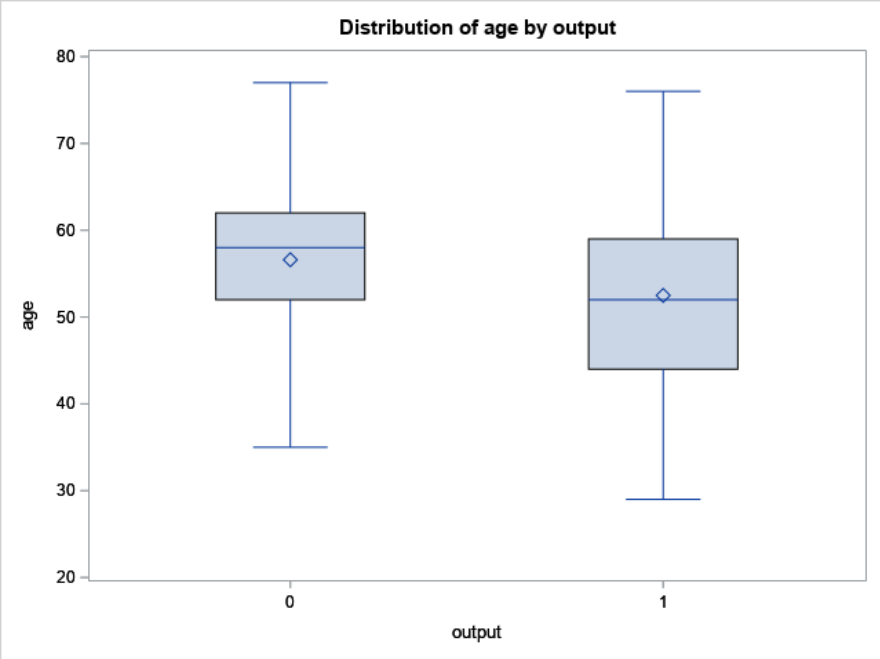


Figure 10

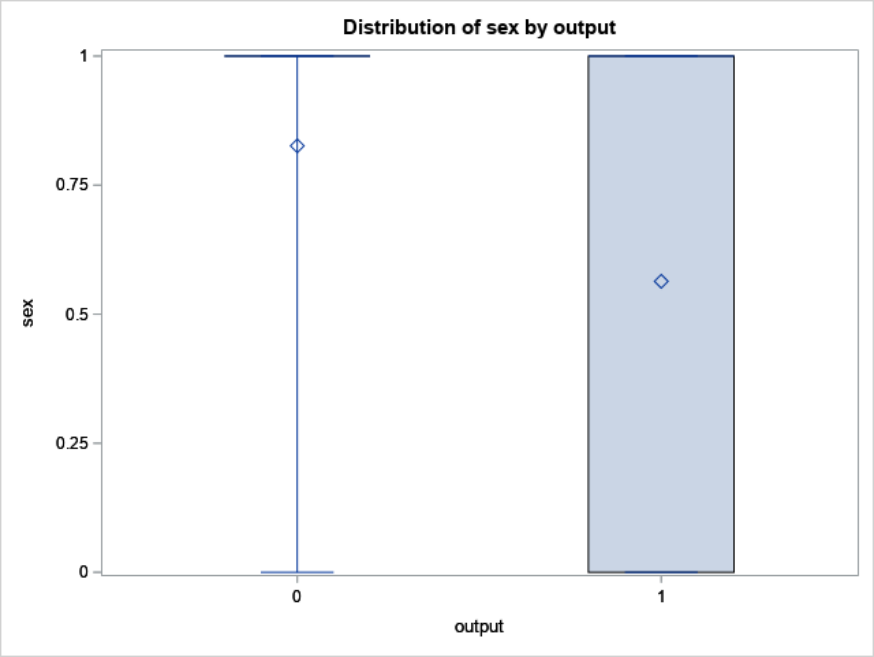


Figure 11

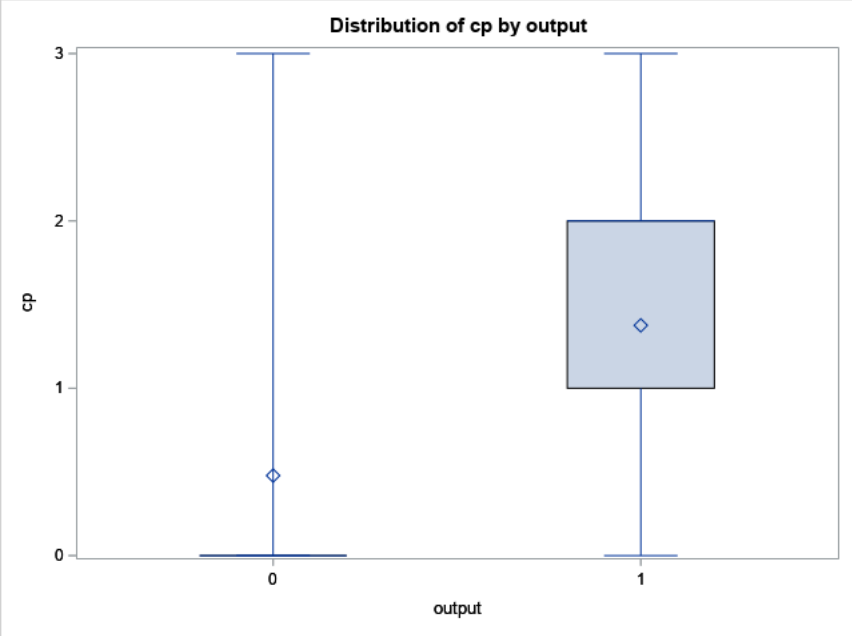


Figure 12

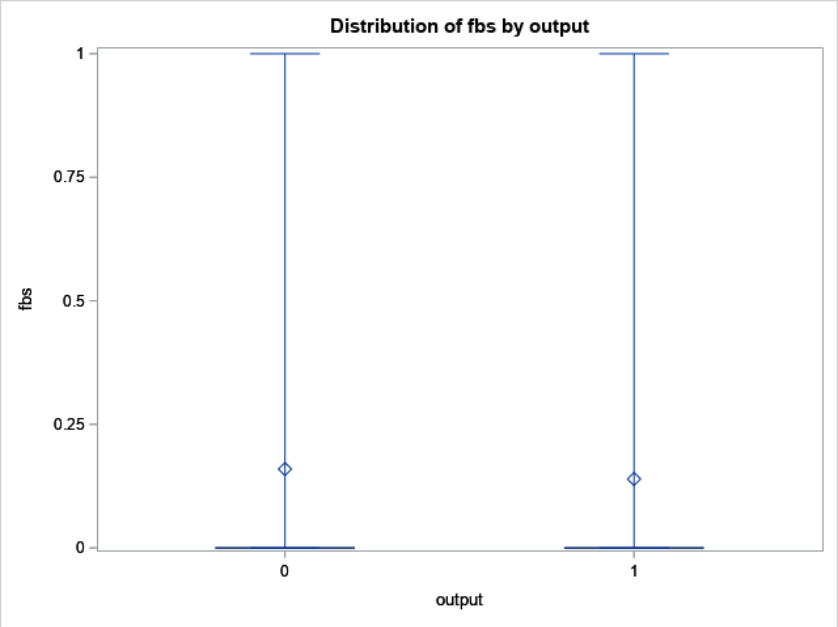


Figure 13

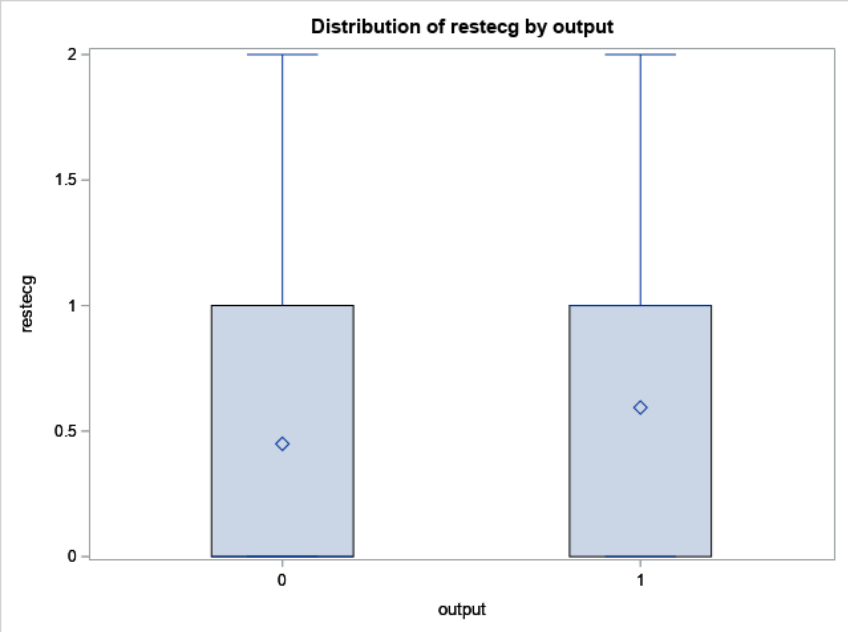


Figure 14

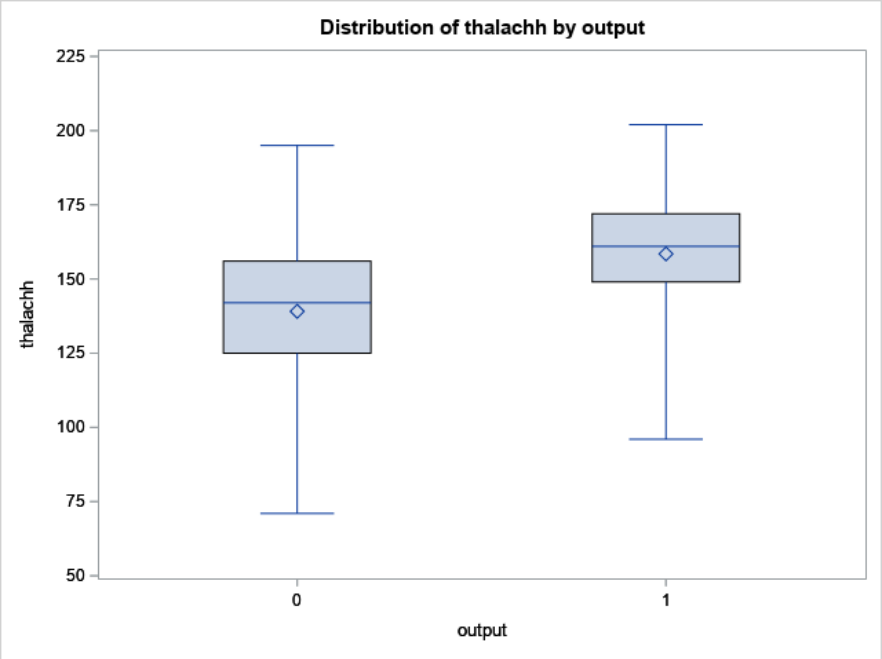


Figure 15

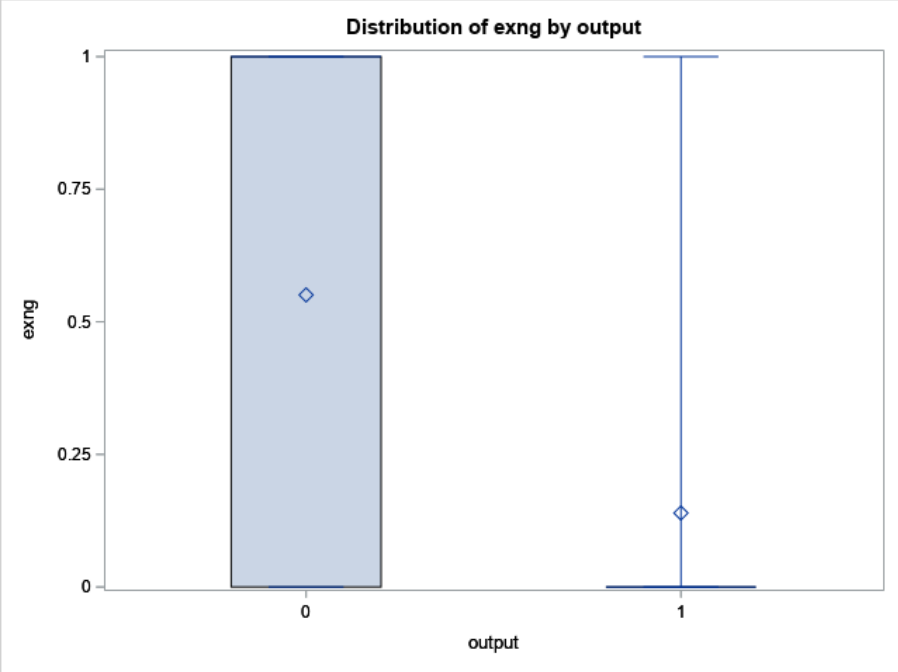


Figure 16

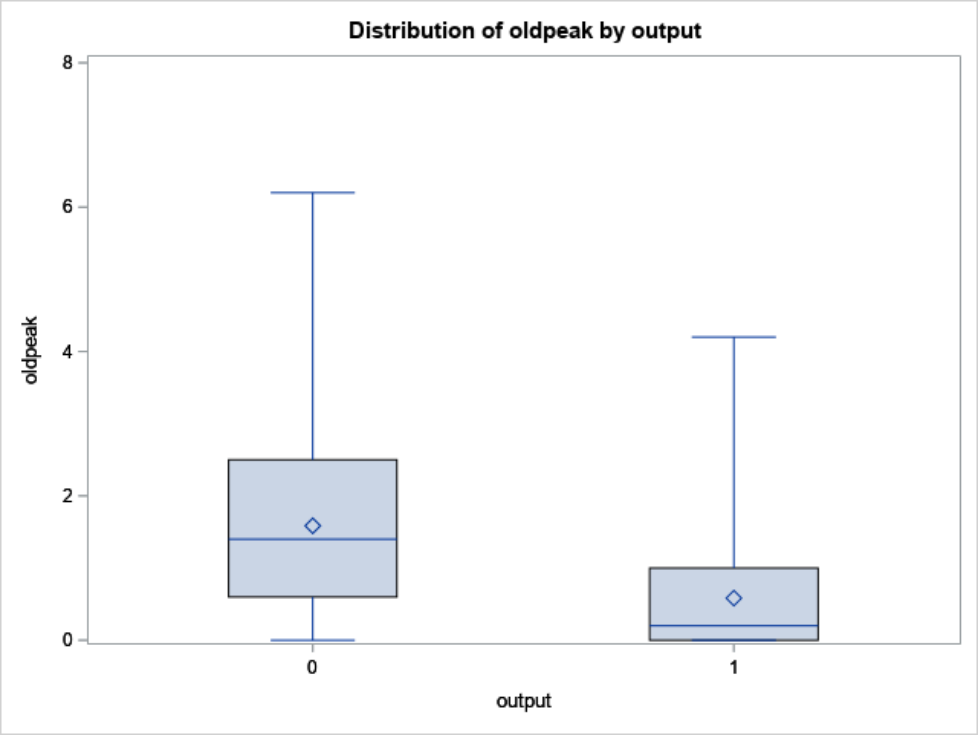


Figure 17

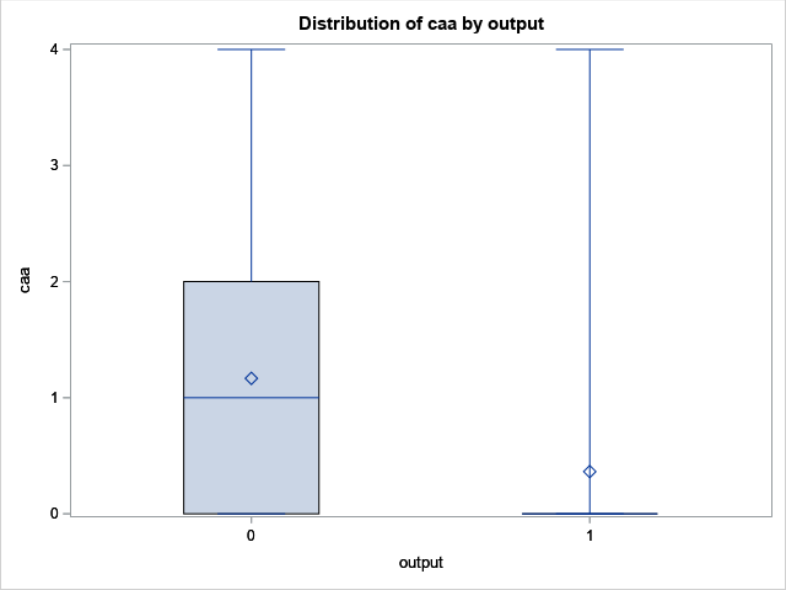


Figure 18

Estimated Correlation Matrix														
Parameter	Intercept	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	age
Intercept	1.0000	-0.2288	0.0783	-0.3128	-0.2003	0.0602	-0.1493	-0.5976	-0.1830	-0.1617	-0.0536	-0.0351	-0.1655	-0.5617
sex	-0.2288	1.0000	-0.1700	0.1265	0.3351	-0.0859	0.0724	-0.1433	0.0881	-0.0741	-0.1364	0.0429	-0.1032	0.1120
cp	0.0783	-0.1700	1.0000	-0.1348	0.0132	-0.1690	0.0918	-0.0577	0.1978	-0.1904	0.1050	-0.0970	-0.0963	-0.0291
trtbps	-0.3128	0.1265	-0.1348	1.0000	0.0315	-0.0944	-0.0008	-0.1815	-0.0468	-0.0493	-0.0636	0.0716	0.0272	-0.2271
chol	-0.2003	0.3351	0.0132	0.0315	1.0000	-0.0076	0.1661	-0.1900	0.0243	-0.0698	-0.0416	0.1148	-0.0931	-0.1766
fbs	0.0602	-0.0859	-0.1690	-0.0944	-0.0076	1.0000	0.0185	-0.0502	-0.0737	0.0894	0.1143	-0.1086	0.1672	-0.0937
restecg	-0.1493	0.0724	0.0918	-0.0008	0.1661	0.0185	1.0000	0.0139	-0.0593	-0.0998	-0.0565	-0.0566	-0.1037	0.0865
thalachh	-0.5976	-0.1433	-0.0577	-0.1815	-0.1900	-0.0502	0.0139	1.0000	0.1638	0.0828	-0.1541	0.0089	-0.0601	0.3986
exng	-0.1830	0.0881	0.1978	-0.0468	0.0243	-0.0737	-0.0593	0.1638	1.0000	-0.0598	0.0333	0.0897	-0.0098	0.0555
oldpeak	-0.1617	-0.0741	-0.1904	-0.0493	-0.0698	0.0894	-0.0998	0.0828	-0.0598	1.0000	0.4519	-0.0449	0.0601	0.0372
slp	-0.0536	-0.1364	0.1050	-0.0636	-0.0416	0.1143	-0.0565	-0.1541	0.0333	0.4519	1.0000	-0.2121	-0.0129	-0.0103
caa	-0.0351	0.0429	-0.0970	0.0716	0.1148	-0.1086	-0.0566	0.0089	0.0897	-0.0449	-0.2121	1.0000	0.0415	-0.1178
thall	-0.1655	-0.1032	-0.0963	0.0272	-0.0931	0.1672	-0.1037	-0.0601	-0.0098	0.0601	-0.0129	0.0415	1.0000	-0.0335
age	-0.5617	0.1120	-0.0291	-0.2271	-0.1766	-0.0937	0.0865	0.3986	0.0555	0.0372	-0.0103	-0.1178	-0.0335	1.0000

Figure 19

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.65519	0.23461	2.79	0.0056
thalachh	1	0.00337	0.00105	3.19	0.0016
cp	1	0.11574	0.02238	5.17	<.0001
caa	1	-0.11379	0.02116	-5.38	<.0001
sex	1	-0.18846	0.04589	-4.11	<.0001
slp	1	0.13323	0.03678	3.62	0.0003
exng	1	-0.15689	0.05129	-3.06	0.0024
thall	1	-0.13299	0.03536	-3.76	0.0002
trtbps	1	-0.00261	0.00120	-2.17	0.0307

Figure 20

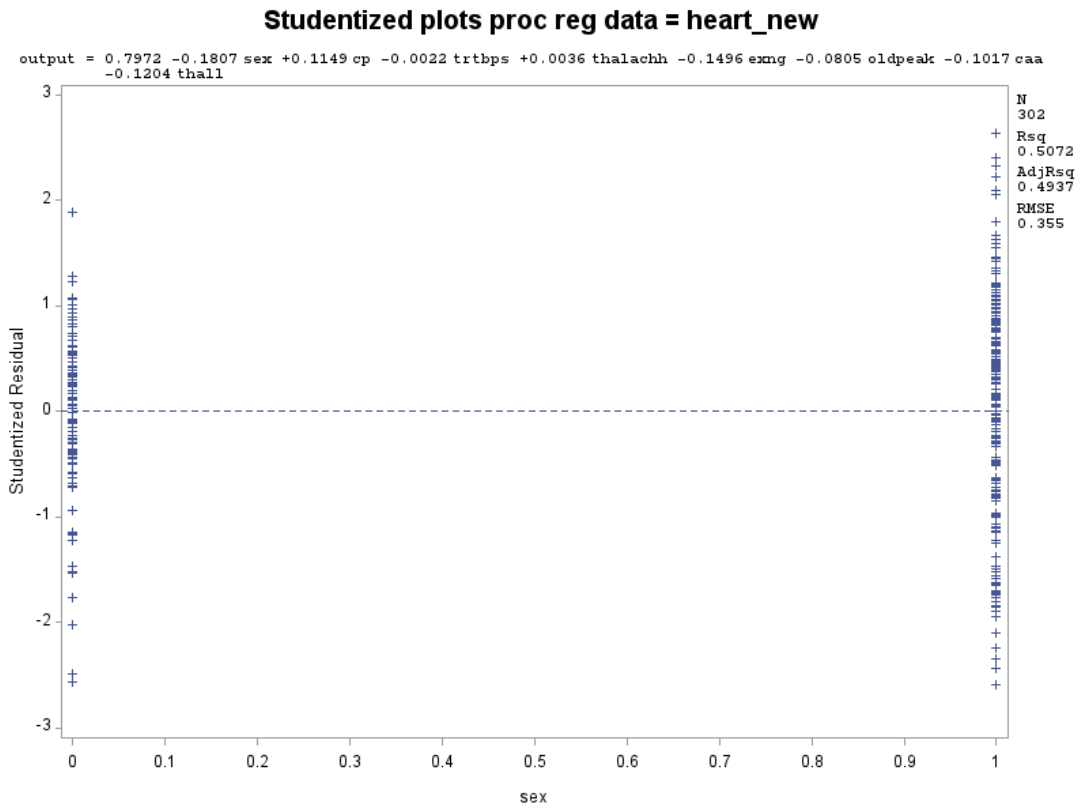

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
3
2
1
0
-1
-2
-3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
sex

Figure 21

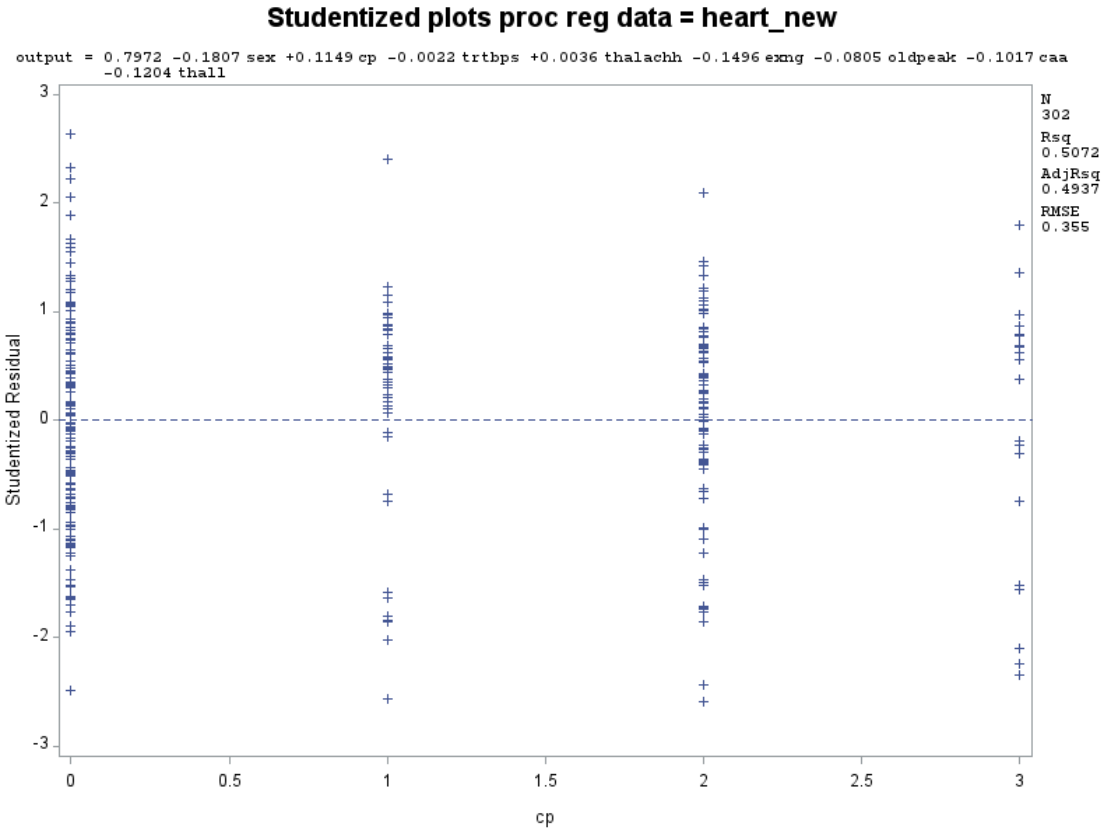

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
cp
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3

Figure 22

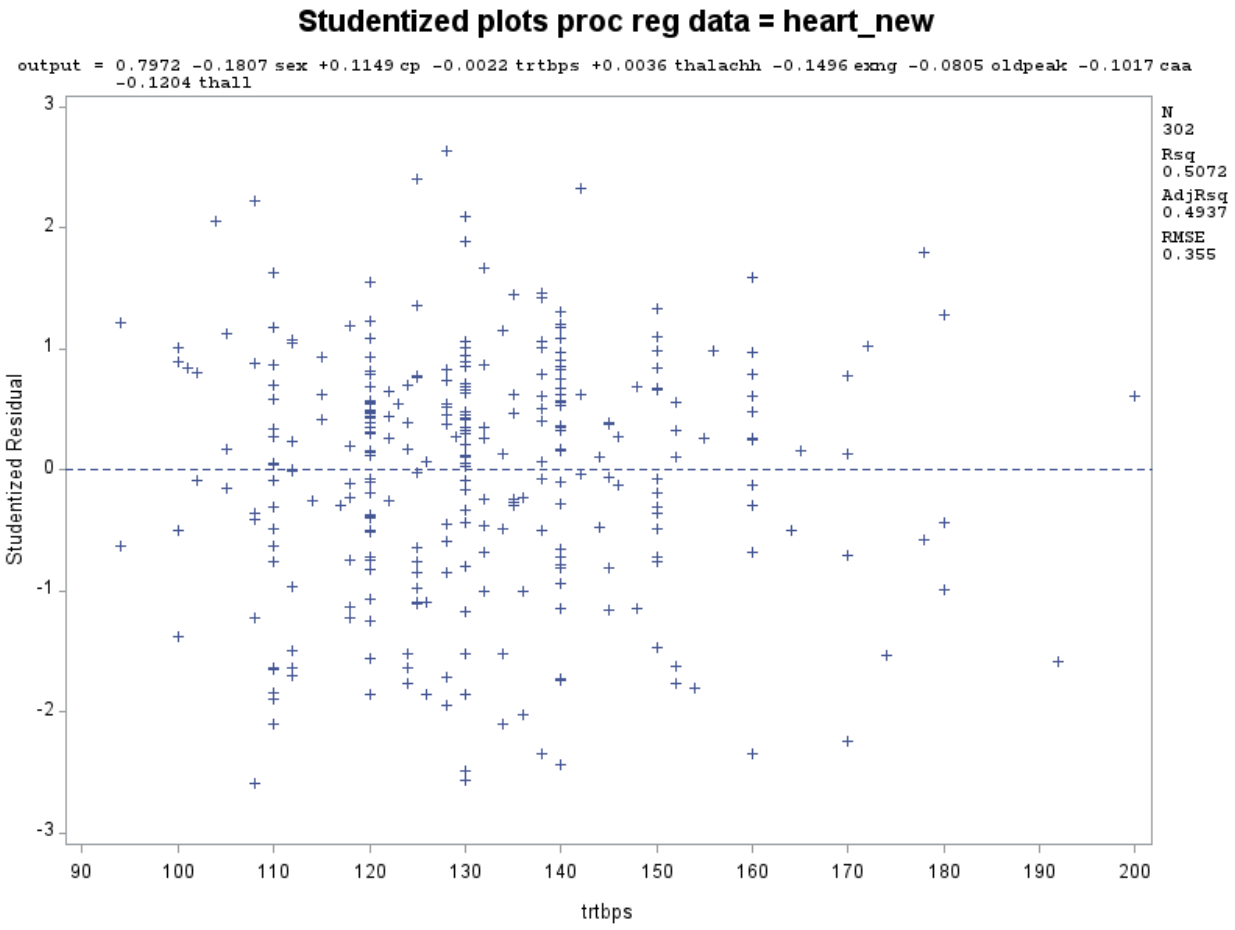

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
trtbps
3
2
1
0
-1
-2
-3
90
100
110
120
130
140
150
160
170
180
190
200

Figure 23

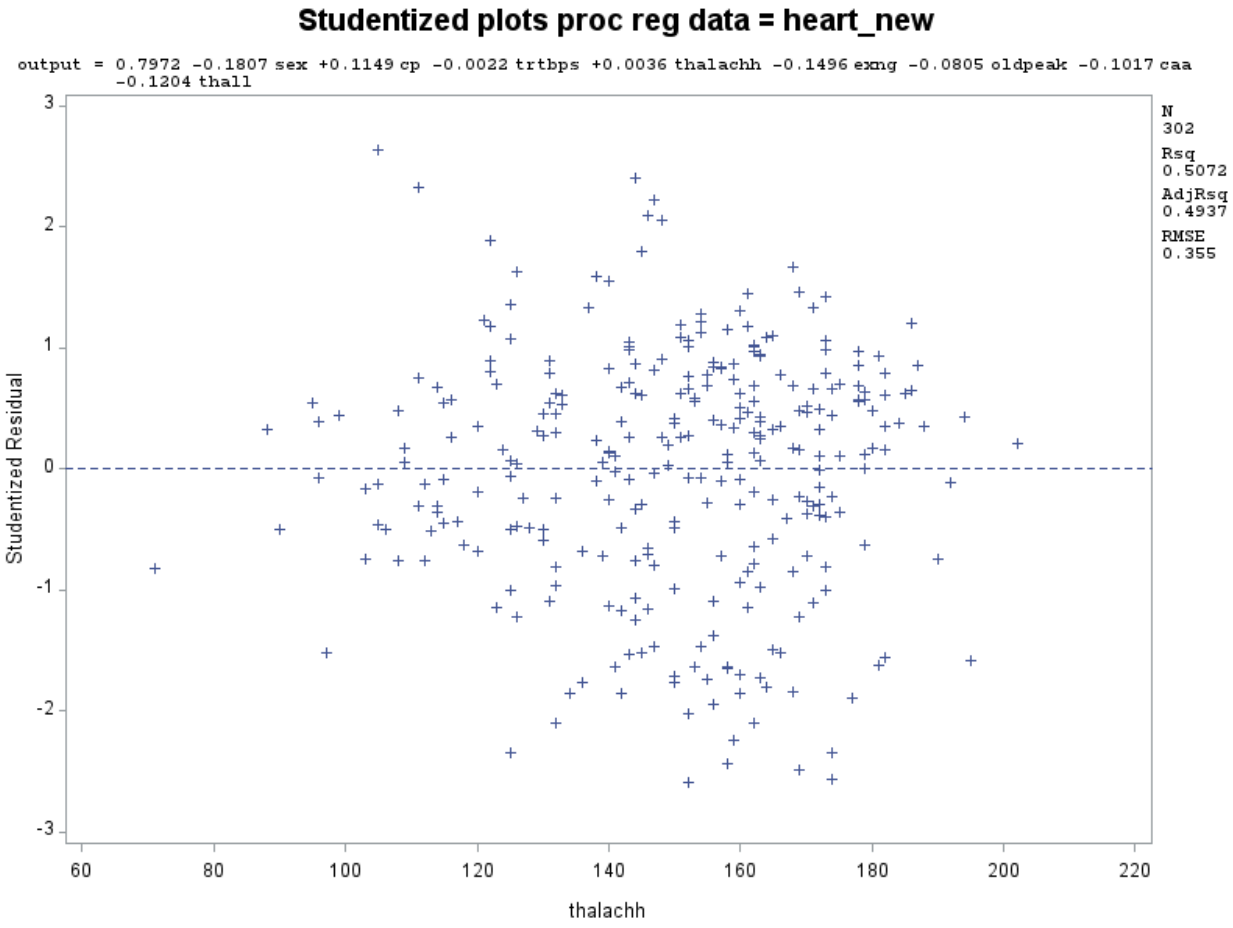

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
thalachh
3
2
1
0
-1
-2
-3
60
80
100
120
140
160
180
200
220

Figure 24

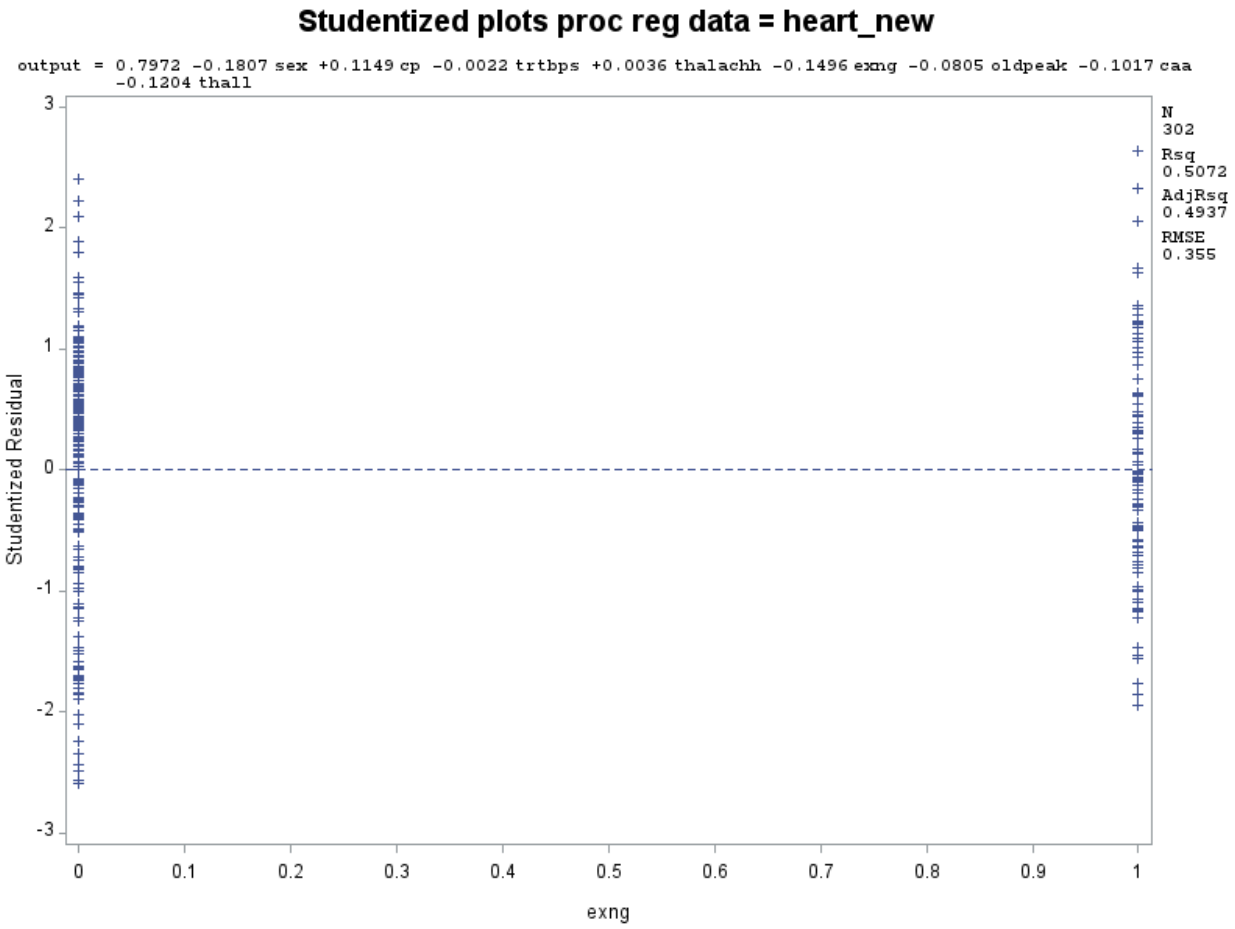

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
3
2
1
0
-1
-2
-3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
exng

Figure 25

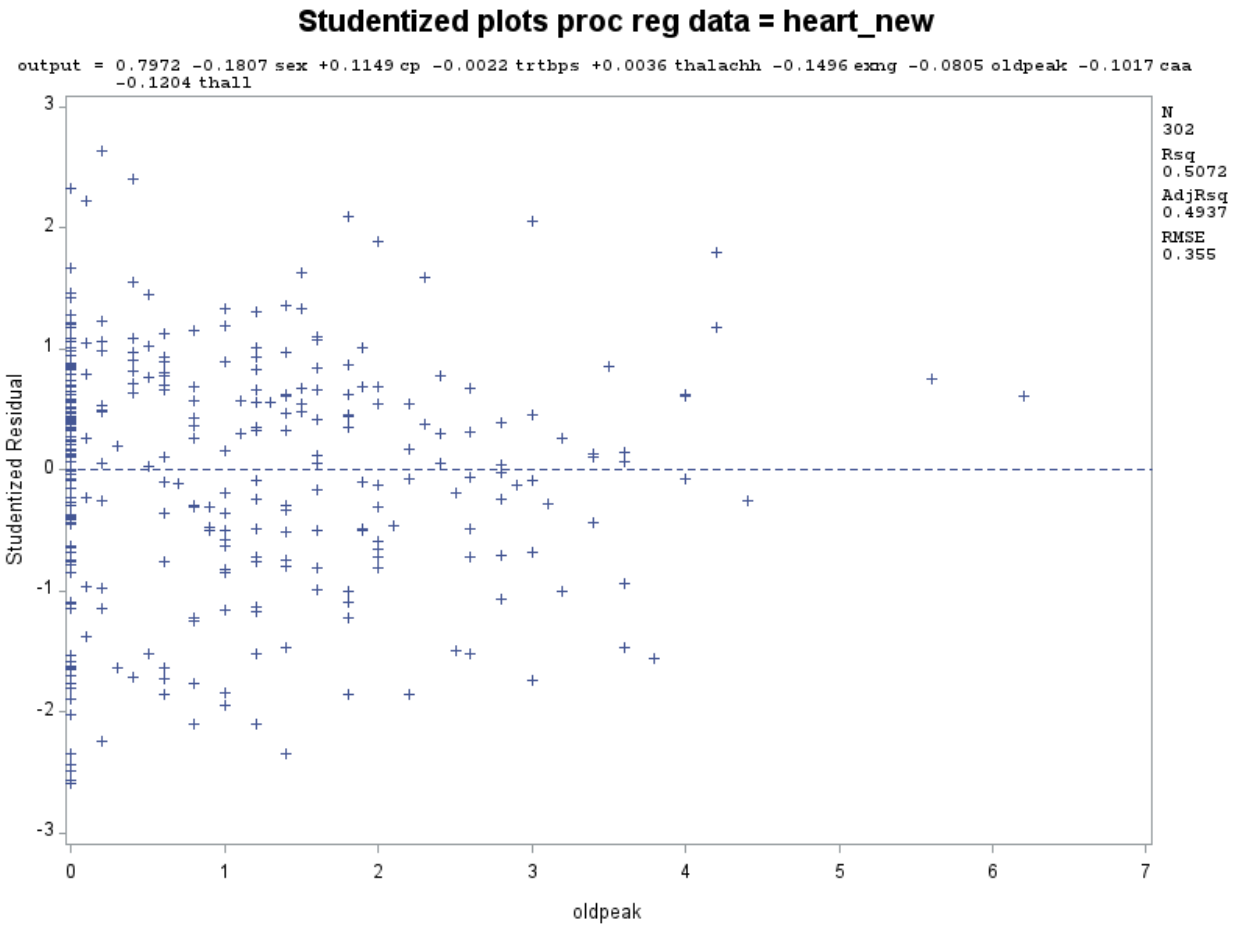

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
oldpeak
3
2
1
0
-1
-2
-3
0
1
2
3
4
5
6
7

Figure 26

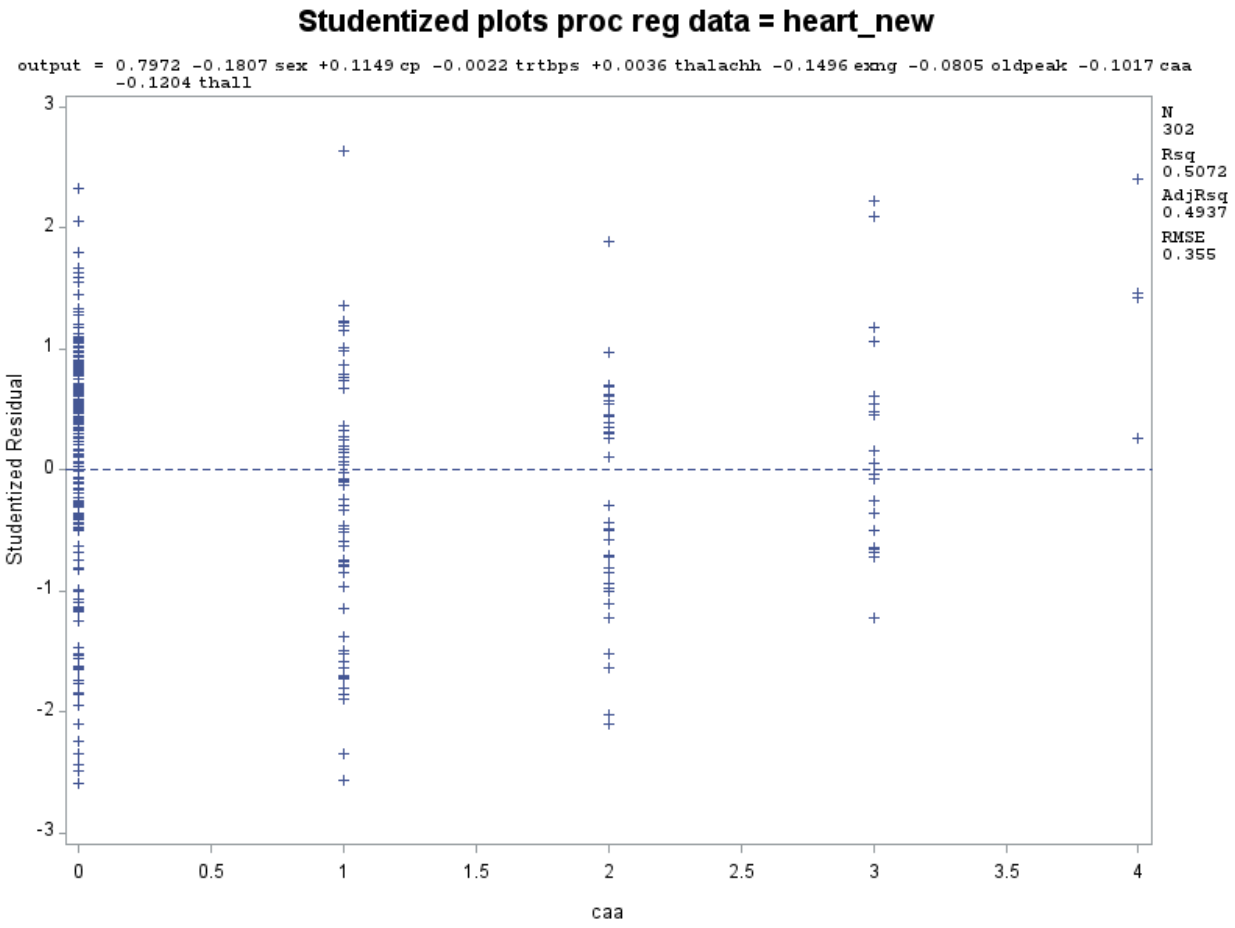

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
caa
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3
3.5
4

Figure 27

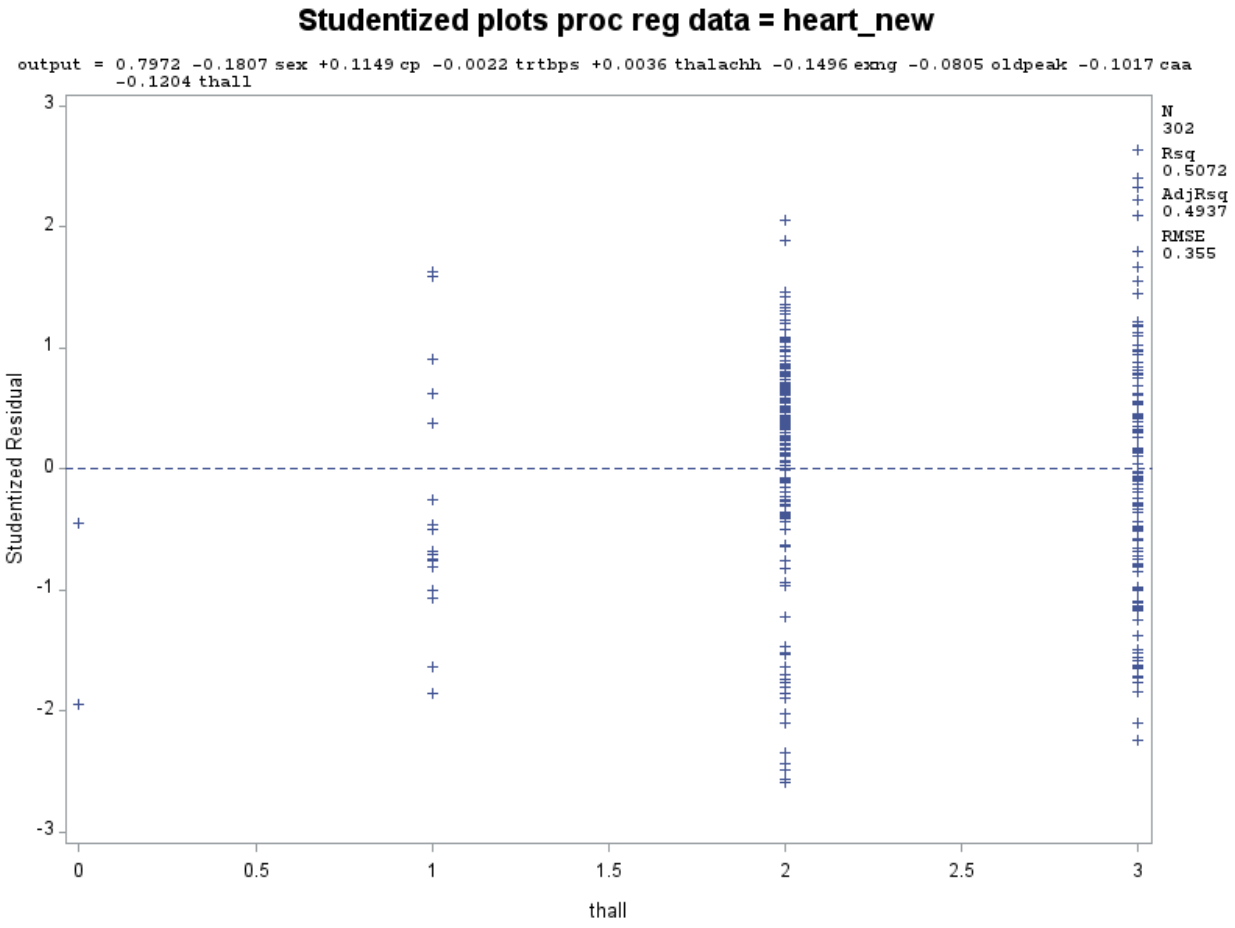

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
thall
3
2
1
0
-1
-2
-3
0
0.5
1
1.5
2
2.5
3

Figure 28

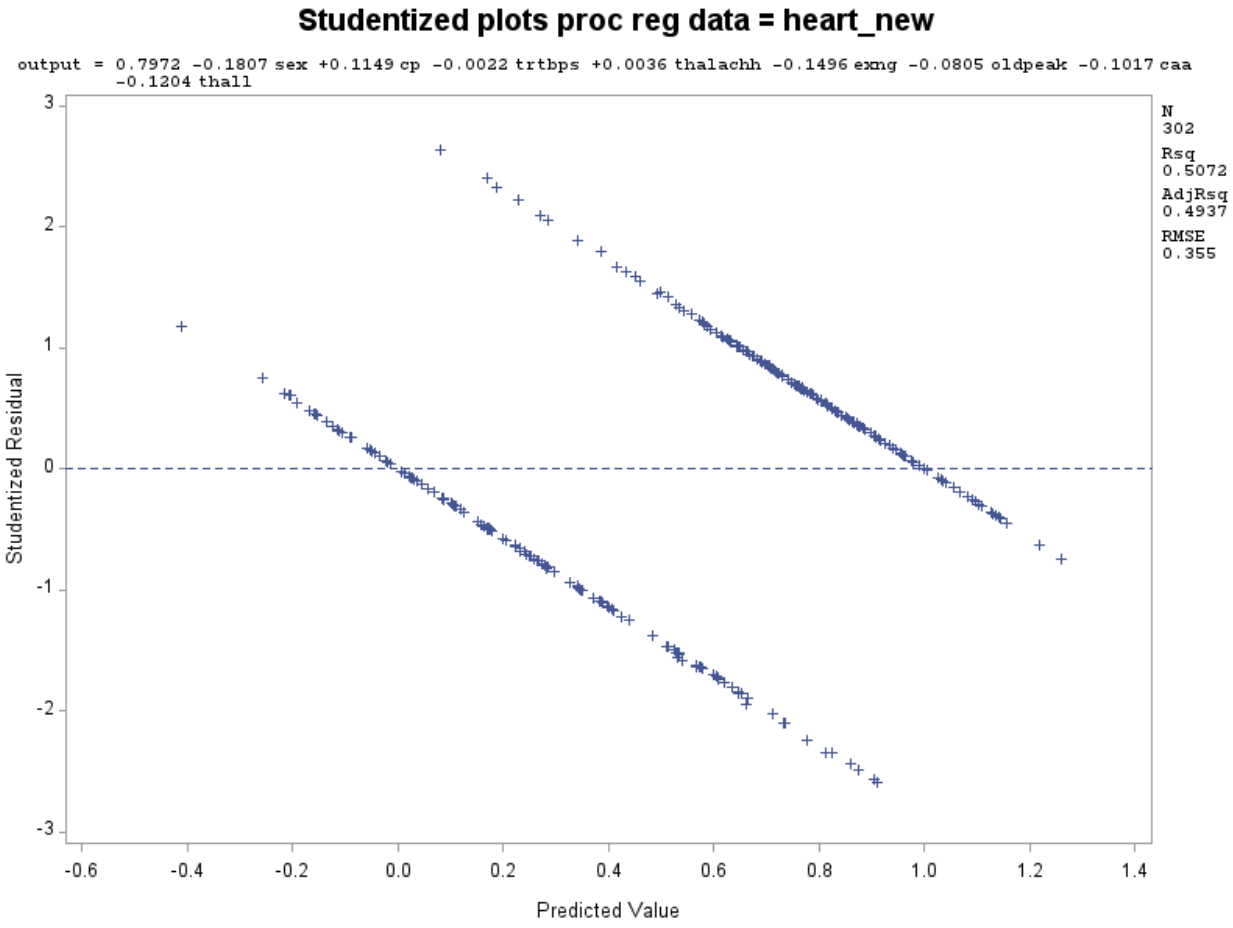

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Studentized Residual
Predicted Value
3
2
1
0
-1
-2
-3
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

Figure 29

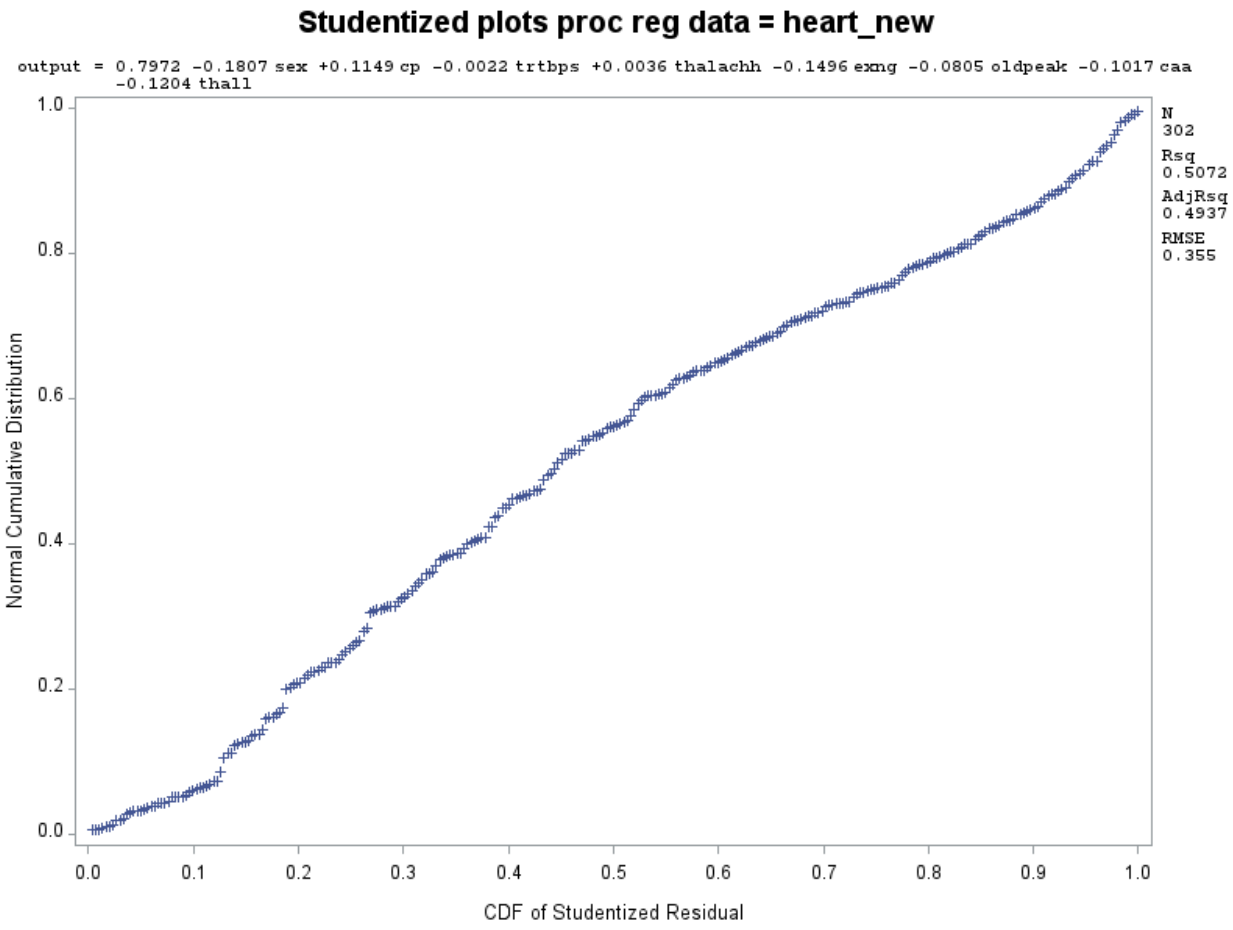
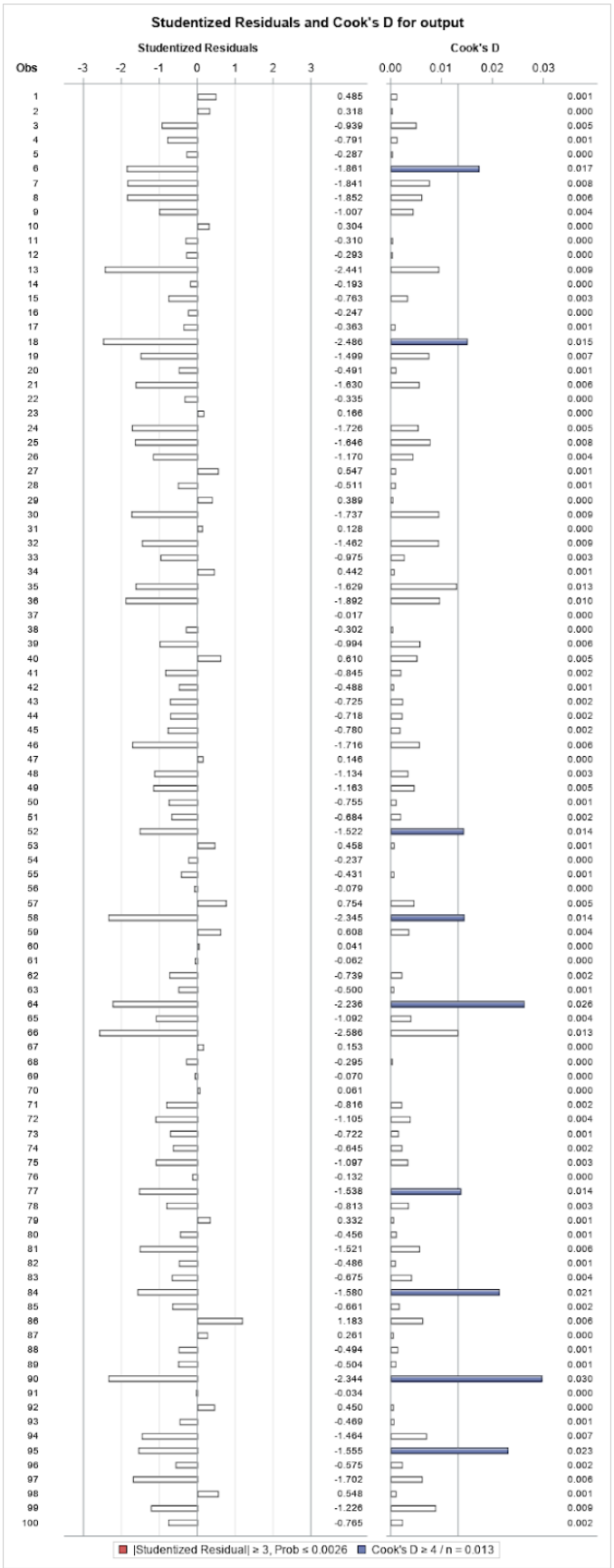
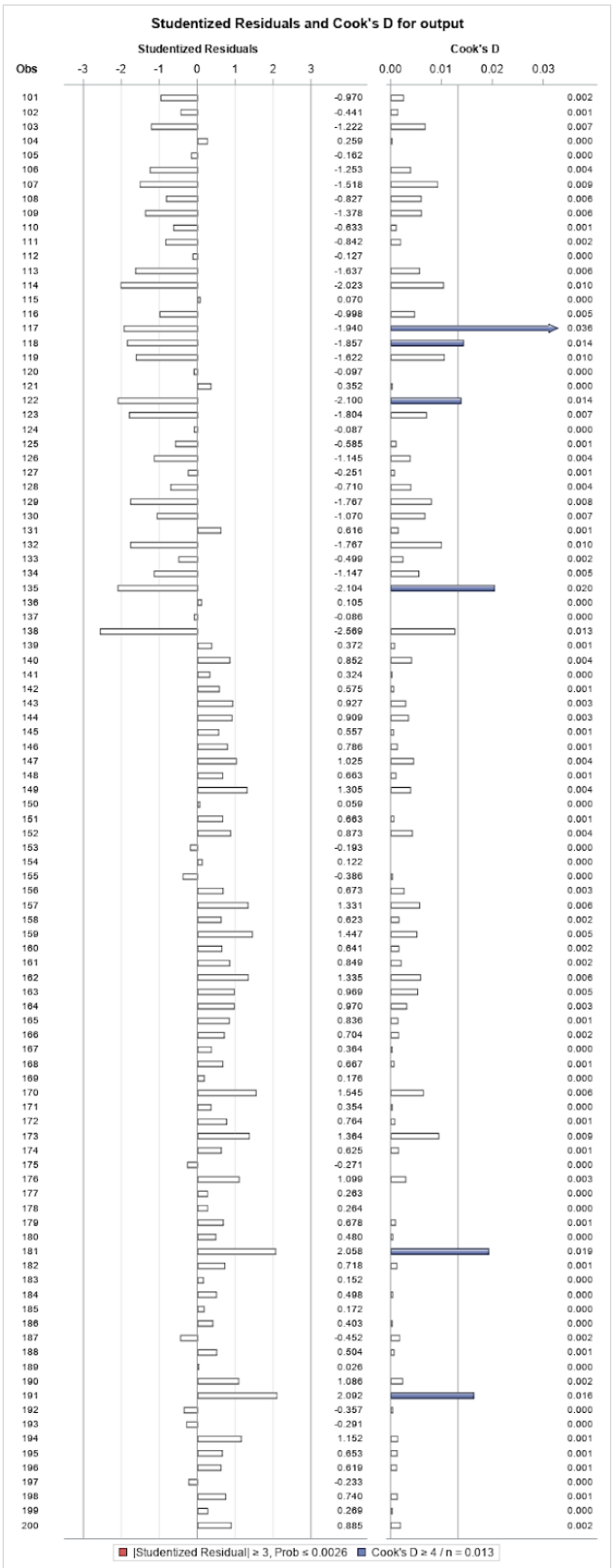
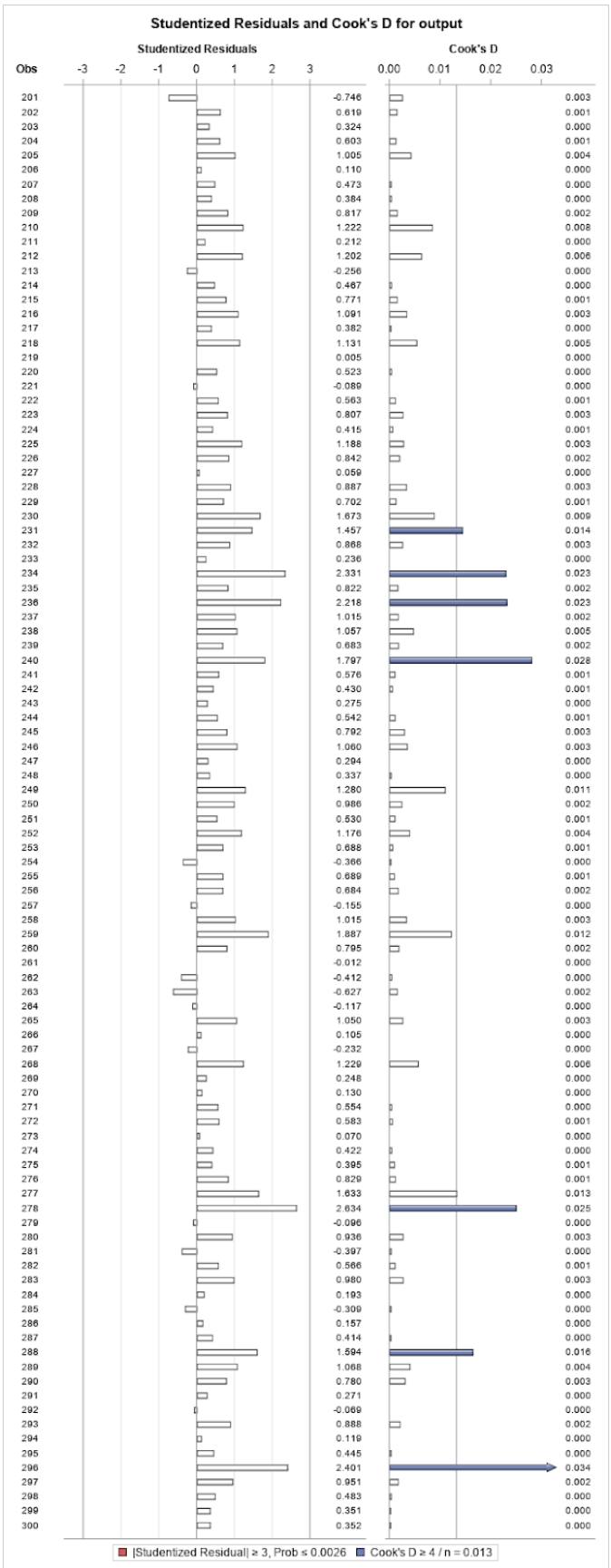

Studentized plots proc reg data = heart_new
output = 0.7972 -0.1807 sex +0.1149 cp -0.0022 trtbps +0.0036 thalachh -0.1496 exng -0.0805 oldpeak -0.1017 caa -0.1204 thall
N
302
Rsq
0.5072
AdjRsq
0.4937
RMSE
0.355
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure 30


Studentized Residuals and Cook's D for output
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.00 0.01 0.02 0.03
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.013


Studentized Residuals and Cook's D for output
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.00 0.01 0.02 0.03
101 -0.970 0.002
102 -0.441 0.001
103 -1.222 0.007
104 0.259 0.000
105 -0.162 0.000
106 -1.253 0.004
107 -1.518 0.009
108 -0.827 0.006
109 -1.378 0.006
110 -0.633 0.001
111 -0.842 0.002
112 -0.127 0.000
113 -1.637 0.006
114 -2.023 0.010
115 0.070 0.000
116 -0.998 0.005
117 -1.940 0.036
118 -1.857 0.014
119 -1.622 0.010
120 -0.097 0.000
121 0.352 0.000
122 -2.100 0.014
123 -1.804 0.007
124 -0.087 0.000
125 -0.585 0.001
126 -1.145 0.004
127 -0.251 0.001
128 -0.710 0.004
129 -1.767 0.008
130 -1.070 0.007
131 0.616 0.001
132 -1.767 0.010
133 -0.499 0.002
134 -1.147 0.005
135 -2.104 0.020
136 0.105 0.000
137 -0.086 0.000
138 -2.589 0.013
139 0.372 0.001
140 0.852 0.004
141 0.324 0.000
142 0.575 0.001
143 0.927 0.003
144 0.909 0.003
145 0.557 0.001
146 0.786 0.001
147 1.025 0.004
148 0.663 0.001
149 1.305 0.004
150 0.059 0.000
151 0.663 0.001
152 0.873 0.004
153 -0.193 0.000
154 0.122 0.000
155 -0.386 0.000
156 0.673 0.003
157 1.331 0.006
158 0.623 0.002
159 1.447 0.005
160 0.641 0.002
161 0.849 0.002
162 1.335 0.006
163 0.969 0.005
164 0.970 0.003
165 0.836 0.001
166 0.704 0.002
167 0.364 0.000
168 0.667 0.001
169 0.176 0.000
170 1.545 0.006
171 0.354 0.000
172 0.764 0.001
173 1.364 0.009
174 0.625 0.001
175 -0.271 0.000
176 1.099 0.003
177 0.263 0.000
178 0.264 0.000
179 0.678 0.001
180 0.480 0.000
181 2.058 0.019
182 0.718 0.001
183 0.152 0.000
184 0.498 0.000
185 0.172 0.000
186 0.403 0.000
187 -0.452 0.002
188 0.504 0.001
189 0.026 0.000
190 1.086 0.002
191 2.092 0.016
192 -0.357 0.000
193 -0.291 0.000
194 1.152 0.001
195 0.653 0.001
196 0.619 0.001
197 -0.233 0.000
198 0.740 0.001
199 0.269 0.000
200 0.885 0.002
|Studentized Residual| ≥ 3, Prob ≤ 0.0026
Cook's D ≥ 4 / n = 0.013



Obs	Studentized Residuals	Cook's D
201	-0.746	0.003
202	0.619	0.001
203	0.324	0.000
204	0.603	0.001
205	1.005	0.004
206	0.110	0.000
207	0.473	0.000
208	0.384	0.000
209	0.817	0.002
210	1.222	0.008
211	0.212	0.000
212	1.202	0.006
213	-0.256	0.000
214	0.467	0.000
215	0.771	0.001
216	1.091	0.003
217	0.382	0.000
218	1.131	0.005
219	0.005	0.000
220	0.523	0.000
221	-0.089	0.000
222	0.563	0.001
223	0.807	0.003
224	0.415	0.001
225	1.188	0.003
226	0.842	0.002
227	0.059	0.000
228	0.887	0.003
229	0.702	0.001
230	1.673	0.009
231	1.457	0.014
232	0.868	0.003
233	0.236	0.000
234	2.331	0.023
235	0.822	0.002
236	2.218	0.023
237	1.015	0.002
238	1.057	0.005
239	0.683	0.002
240	1.797	0.028
241	0.576	0.001
242	0.430	0.001
243	0.275	0.000
244	0.542	0.001
245	0.792	0.003
246	1.060	0.003
247	0.294	0.000
248	0.337	0.000
249	1.280	0.011
250	0.986	0.002
251	0.530	0.001
252	1.176	0.004
253	0.688	0.001
254	-0.366	0.000
255	0.689	0.001
256	0.684	0.002
257	-0.155	0.000
258	1.015	0.003
259	1.887	0.012
260	0.795	0.002
261	-0.012	0.000
262	-0.412	0.000
263	-0.627	0.002
264	-0.117	0.000
265	1.050	0.003
266	0.105	0.000
267	-0.232	0.000
268	1.229	0.006
269	0.248	0.000
270	0.130	0.000
271	0.554	0.000
272	0.583	0.001
273	0.070	0.000
274	0.422	0.000
275	0.395	0.001
276	0.829	0.001
277	1.633	0.013
278	2.634	0.025
279	-0.096	0.000
280	0.936	0.003
281	-0.397	0.000
282	0.566	0.001
283	0.980	0.003
284	0.193	0.000
285	-0.309	0.000
286	0.157	0.000
287	0.414	0.000
288	1.594	0.016
289	1.068	0.004
290	0.780	0.003
291	0.271	0.000
292	-0.069	0.000
293	0.888	0.002
294	0.119	0.000
295	0.445	0.000
296	2.401	0.034
297	0.951	0.002
298	0.483	0.000
299	0.351	0.000
300	0.352	0.000

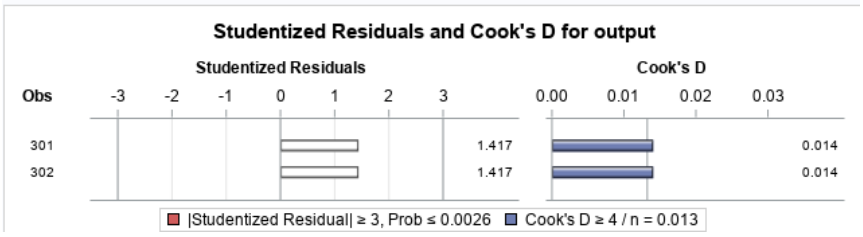


Figure 31

Model 1

The REG Procedure
Model: MODEL1
Dependent Variable: train_y

Number of Observations Read	302
Number of Observations Used	227
Number of Observations with Missing Values	75

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	29.26487	3.65811	29.28	<.0001
Error	218	27.23733	0.12494		
Corrected Total	226	56.50220			

Root MSE	0.35347	R-Square	0.5179
Dependent Mean	0.53304	Adj R-Sq	0.5003
Coeff Var	66.31237		

The FREQ Procedure

Selection Indicator				
Selected	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	75	24.83	75	24.83
1	227	75.17	302	100.00

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.69738	0.27271	2.56	0.0112
thalachh	1	0.00292	0.00123	2.37	0.0187
cp	1	0.13119	0.02555	5.14	<.0001
caa	1	-0.11751	0.02340	-5.02	<.0001
sex	1	-0.20051	0.05359	-3.74	0.0002
slp	1	0.13933	0.04320	3.23	0.0015
exng	1	-0.15735	0.05819	-2.70	0.0074
thall	1	-0.09688	0.04133	-2.34	0.0200
trtbps	1	-0.00318	0.00140	-2.26	0.0246

Figure 32

The FREQ Procedure

Frequency	Table of output by pred_y		
	pred_y		
output	0	1	Total
0	27	5	32
1	9	34	43
Total	36	39	75

Figure 33

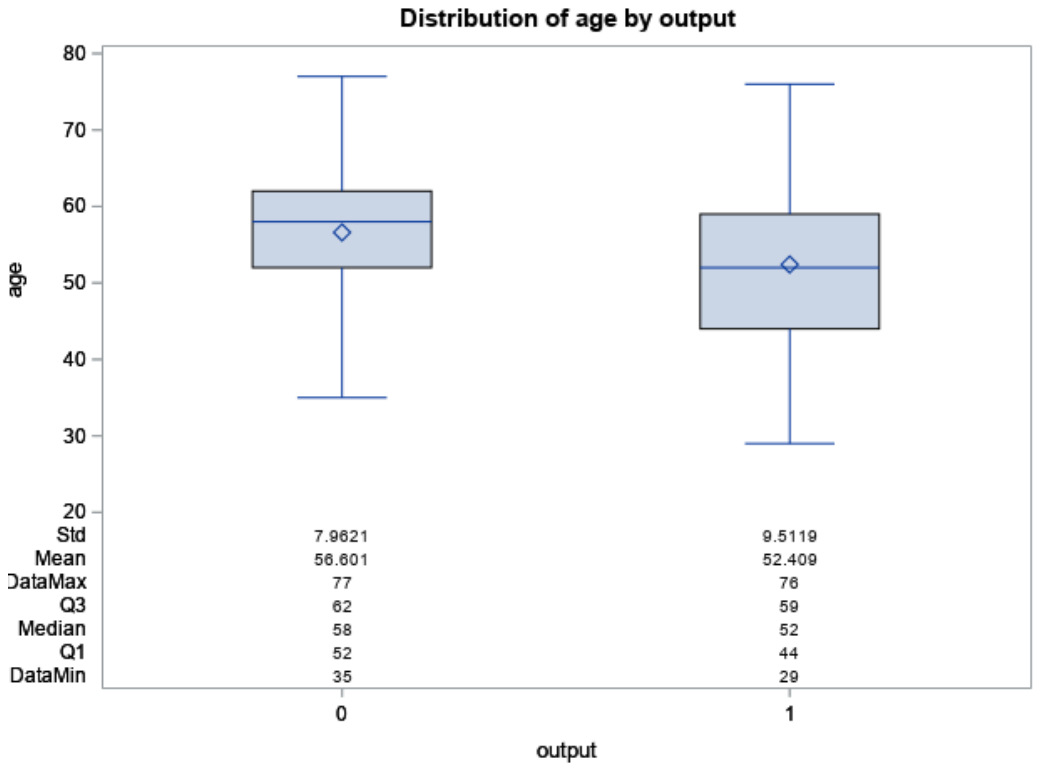
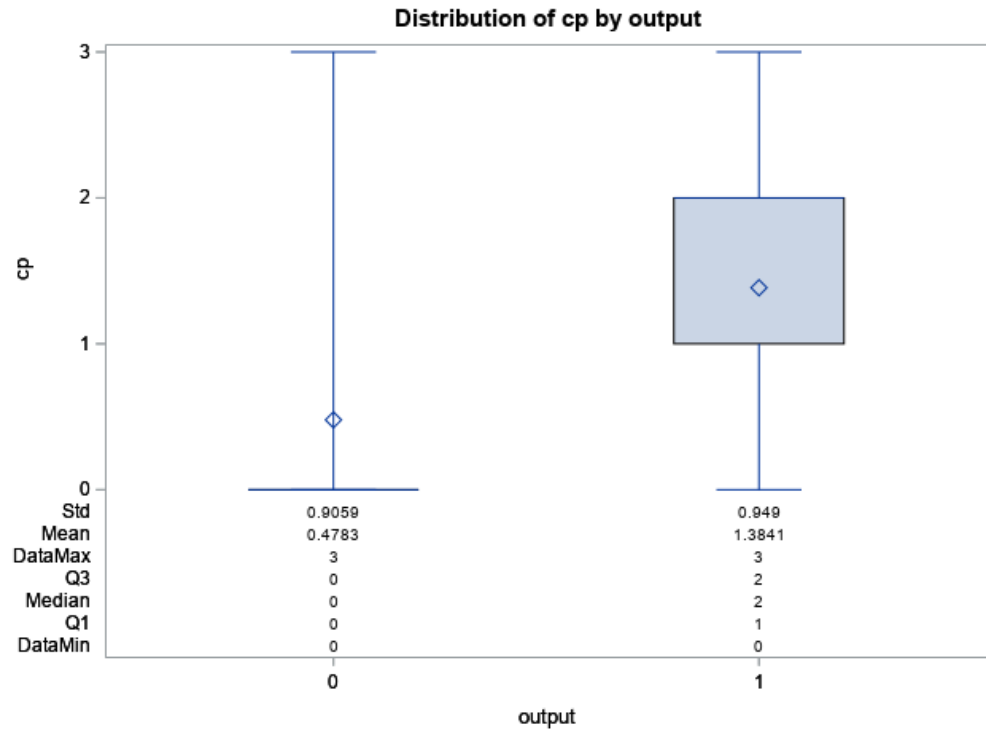


Figure 34



Code:

```
proc import datafile = "heart.csv" out = heart replace;
delimiter = ",";
getnames = yes;
run;
proc print;
run;

title "Descriptives";
proc means min p25 p50 p75 max std;
var age trtbps chol thalachh oldpeak;
run;

title "Histogram W/ normal curve";
proc univariate normal;
var age trtbps chol thalachh oldpeak;
histogram / normal (mu = est sigma = est);
run;

title "Boxplots vs IDVs";
proc sort;
by output;
run;
proc boxplot;
plot trtbps*output;
insetgroup min q1 q2 q3 max mean stddev;
plot chol*output;
insetgroup min q1 q2 q3 max mean stddev;
plot thalachh*output;
insetgroup min q1 q2 q3 max mean stddev;
plot oldpeak*output;
insetgroup min q1 q2 q3 max mean stddev;
plot age*output;
insetgroup min q1 q2 q3 max mean stddev;
plot sex*output;
insetgroup min q1 q2 q3 max mean stddev;
plot cp*output;
insetgroup min q1 q2 q3 max mean stddev;
plot fbs*output;
insetgroup min q1 q2 q3 max mean stddev;
plot restecg*output;
insetgroup min q1 q2 q3 max mean stddev;
plot thalachh*output;
insetgroup min q1 q2 q3 max mean stddev;
plot exng*output;
insetgroup min q1 q2 q3 max mean stddev;
```

```
plot oldpeak*output;
insetgroup min q1 q2 q3 max mean stddev;
plot caa*output;
insetgroup min q1 q2 q3 max mean stddev;
run;

* scatter plot;
proc sgscatter;
plot (age)*(sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output)/pbspline;
run;

proc gplot;
plot age*(sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall output);
run;

* correlation table;
proc corr data = heart;
var output sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall age;
run;

proc reg data = heart;
model output =  sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall age/vif;
run;

proc reg data = heart;
model output = thalachh cp caa sex slp exng thall trtbps;
run;

proc logistic data = heart;
model output (event = '1') = sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
age/rsquare stb corrb;
run;

proc reg data = heart;
model output = sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall age/influence R
;
run;

data heart;
set heart;
if _n_ = 282 then delete;
run;

* removed high p-values variables;
proc logistic data = heart;
model output (event = '1') = sex cp trtbps thalachh exng oldpeak caa thall/rsquare stb corrb;
```

```
run;

* temporary;
* need to check with predicting variable again (age or output);
* check for insignificant variables;
* age dropped due to high p value;
proc reg data = heart;
model output =  sex cp trtbps thalachh exng oldpeak caa thall/vif;
run;

* new dataset;
data heart_new;

set heart;
drop chol fbs restecg slp age;
run;
proc print data=heart_new;
run;

* correlation table;
proc corr data = heart_new;
var output sex cp trtbps thalachh exng oldpeak caa thall;
run;

* check for outiers/influencial points;
proc reg data = heart_new;
model output = sex cp trtbps thalachh exng oldpeak caa thall/influence R ;
run;

* studentized plots;
title "Studentized plots"
proc reg data = heart_new;
model output =  sex cp trtbps thalachh exng oldpeak caa thall;
plot student.*(sex cp trtbps thalachh exng oldpeak caa thall);
plot student.*predicted.;
plot npp.*student.;
run;


* train and test set;
title"split data to train and test set";
proc surveyselect data = heart out = train seed=47274
samprate = 0.75 outall;
run;

proc freq data = train;
tables selected;
```

```
run;

* new y for train set;
data train;
set train;
if selected then train_y = output;
run;
proc print data = train;
run;

* model selection;
* stepwise selection method;
proc logistic data = train;
model train_y (event = '1') = sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
age/ selection = stepwise rsquare;
run;

* forward selection;
proc logistic data = train;
model train_y (event = '1') = sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
age/ selection = forward rsquare;
run;

* backward selection;
proc logistic data = train;
model train_y (event = '1') = sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
age/ selection = backward rsquare;
run;

* final model;
title "Fianl model";
proc logistic data = train;
model train_y (event = '1') = thalachh cp caa sex slp exng thall trtbps/ stb rsquare iplots;
run;

* classification table with prob 0.1 to 0.8 increase by 0.05;
proc logistic data= train;
model train_y (event = '1') = thalachh cp caa sex slp exng thall trtbps/ctable pprob=(0.1 to 0.8 by
0.05);
output out=pred(where= (train_y= .)) p= phat lower=lcl upper=ucl;
run;

* compute predicted y in testing set for pred_prob > 0.65;
data probs;
set pred;
pred_y = 0;
```

```
threshold = 0.65;
if phat > threshold then pred_y = 1;
run;

title 'Model 1';
Proc reg data=train;
model train_y = thalachh cp caa sex slp exng thall trtbps;
output out=outm1(where=(train_y=.)) p=yhat;
run;

* classification matrix;
proc freq data = probs;
tables output*pred_y / norow nocol noperccent;
run;

proc corr data=outm1;
var output yhat;
run;

proc sort data = train;
by output;
run;
proc boxplot;
plot age*output;
insetgroup min q1 q2 q3 max mean stddev;
plot cp*output;
insetgroup min q1 q2 q3 max mean stddev;
run;
```