PSTAT 126 Project Summary

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-06-11

About the data

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains information about patients and risks for heart disease as well as whether or not they eventually end up developing heart disease. The data set originally contains 1025 observations. For the sake of clearness of the scatter plot, we randomly choose 500 from the original dataset.

Data Description

The `age` variable is a numeric variable that lists the age of the patient in years. The `sex` variable is a binary variable that states the sex of the patient, with 0 representing female and 1 representing male. The `cp` variable is a categorical variable that classifies the type of chest pain the patient is experiencing; 0 represents typical angina, 1 represents atypical angina, 2 represents non-anginal pain, and 3 represents asymptomatic. The `trestbps` variable is a numeric variable that states the resting blood pressure in mmHg.

The `chol` variable is a numeric data which give an overview of person's cholesterol levels. The `fbs` variable is a binary data with 1 means the fast blood sugar is greater than 120mg/dl, and 0 means the fast blood sugar is not. The `restecg` variable is a nominal data with scale of 0 (normal), 1 (abnormality in ST-T waves), and 2 (show left ventricular hypertrophy), which record the resting electrocardiogram results. The `thalach` variable is a numeric data which records the maximum heart rate achieved.

The variable `exang` indicates that if the patient have exercise induced angina (chest pain), which is shown with 0 indicating "no" and 1 indicating "yes". The variable `oldpeak` is a float type data records the ST depression, which is a measure of abnormality of an electrocardiogram, and the measurement is in unit depression. The variable `slope` is the slope of the peak exercise ST segment, which is an electrocardiography read out indicating quality of blood flow to the heart, and the data are in nominal type including 0 (upsloping), 1 (flat) and 2 (downsloping). Finally, the variable `target` is a binary data showing whether the patient is having heart disease (1 = having heart disease and 0 = normal).

Summary Statistics and Graphs

Table 1: Data summary

Name	heart_disease_data
Number of rows	500
Number of columns	14

Column type frequency:

factor	7
numeric	7
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
sex	0	1	FALSE	2	1: 355, 0: 145
cp	0	1	FALSE	4	0: 223, 2: 159, 1: 86, 3: 32
fbs	0	1	FALSE	2	0: 418, 1: 82
restecg	0	1	FALSE	3	0: 263, 1: 233, 2: 4
exang	0	1	FALSE	2	0: 330, 1: 170
slope	0	1	FALSE	3	2: 238, 1: 227, 0: 35
target	0	1	FALSE	2	1: 266, 0: 234

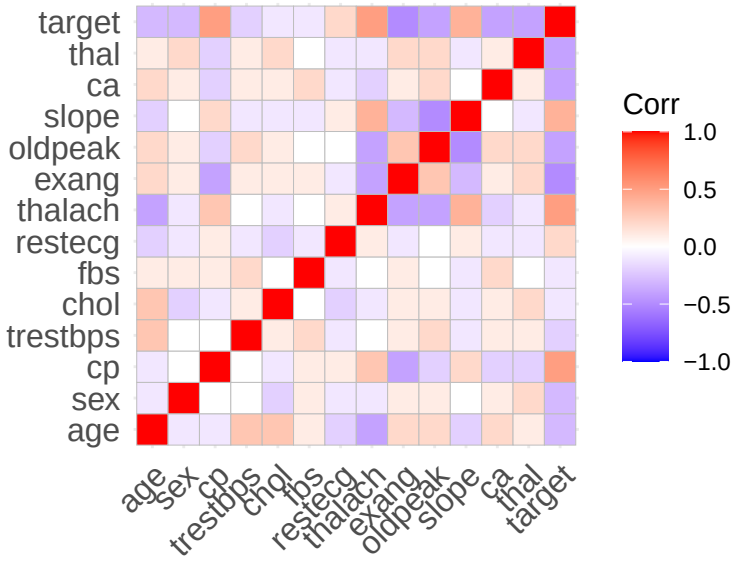
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age	0	1	54.02	9.22	29	47	55.0	60.25	77.0	
trestbps	0	1	132.46	18.06	94	120	130.0	140.00	200.0	
chol	0	1	248.13	55.44	126	211	243.0	282.00	564.0	
thalach	0	1	150.19	23.31	71	137	152.0	168.00	202.0	
oldpeak	0	1	1.03	1.17	0	0	0.8	1.60	6.2	
ca	0	1	0.76	1.07	0	0	0.0	1.00	4.0	
thal	0	1	2.29	0.63	0	2	2.0	3.00	3.0	

Correlation

The following correlation matrix represent the relationship between different variables and the interaction between each other.

Each cell represents a correlation coefficient, the red color, which has a coefficient of 1, represents the strong positive correlation between two variables; whereas, the white color with coefficient of 0 indicates little (no) correlation between two variables, and the purple color with coefficient of -1 indicates a strong negative correlation.



The correlation matrix of numeric data shows that there isn't a strong positive correlation between any two variables, and between maximum heart rate achieved and target; a relatively strong negative correlation between ST slope and oldpeak; and little correlation between sex and chest pain type, cholesterol and fasting blood sugar.

Statistical Approach for Selecting OLS Model

We used a forward step-wise model, meaning we assumed no relationship between the variables, and then added predictors one by one until no other significant predictors were left.

Our criteria for choosing the variables is based on the AIC, meaning that starting with an intercept-only model, we add the variable with the lowest AIC value at a time, then re-examine the new model to find the next variable until no variables significant.

By going through forward step-wise selection, we chose 7 predictors for the model, which are `slope`, `age`, `cp`, `exang`, `target`, `trestbps` and `restecg`.

```
##
## Call:
## lm(formula = thalach ~ slope + age + cp + exang + target + trestbps +
##     restecg, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.480  -9.282   1.679  12.213  49.131
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 169.35644    8.77777  19.294  < 2e-16 ***
```

```
## slope1       -1.16422    3.57973  -0.325 0.745187
## slope2       11.27038    3.70469   3.042 0.002508 **
## age          -0.78128    0.10136  -7.708 1.08e-13 ***
## cp1          10.55720    2.84049   3.717 0.000232 ***
## cp2           5.17212    2.45548   2.106 0.035815 *
## cp3          15.10773    3.79034   3.986 8.03e-05 ***
## exang1       -8.35457    2.22378  -3.757 0.000198 ***
## target1       6.43474    2.32472   2.768 0.005911 **
## trestbps      0.11511    0.05034   2.286 0.022769 *
## restecg1     -3.17946    1.79352  -1.773 0.077055 .
## restecg2    -12.76272   10.06948  -1.267 0.205749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.15 on 388 degrees of freedom
## Multiple R-squared:  0.4909, Adjusted R-squared:  0.4765
## F-statistic: 34.02 on 11 and 388 DF,  p-value: < 2.2e-16
```

Finding R^2 of Predicted Values

About 49% of the variance in maximum heart rate in the test data is accounted for by the slope of the peak exercise ST segment, age, chest pain type, exercise-induced angina, heart disease, resting blood pressure, and resting ecg.

```
## [1] "R-square value on test data: 0.493739305422597"
```

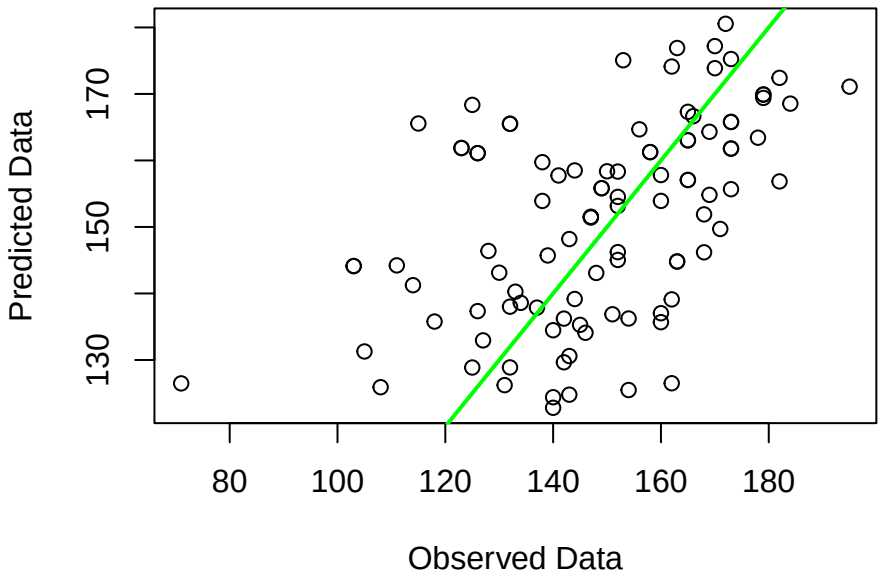


Figure 1: Plotting Predictions for OLS Model

Checking Influential Points of OLS Model

To check outliers, leverage points, and influence points, we created graphs of the residuals, the diagonal values of the hat matrix, and Cook's distance. In each graph, we selected the largest value in red to visualize it on each plot.

The residuals appear randomly scattered with equal variance and no obvious pattern, and based on the Q-Q plot, are approximately normal, indicating that the model is a good fit.

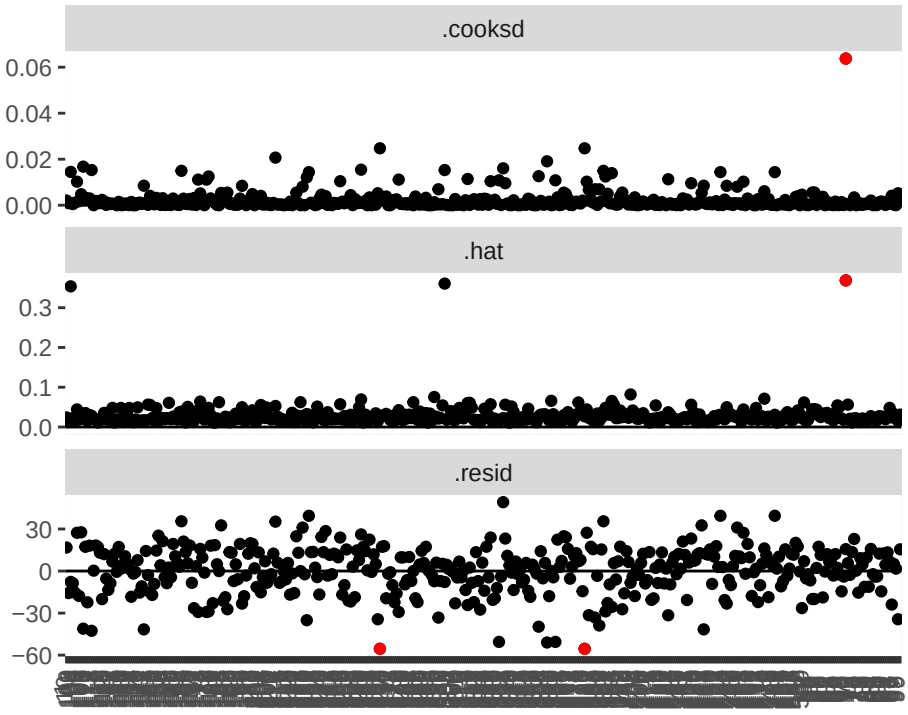


Figure 2: Influential Points Analysis

We then refit the model without the influential point:

```
##
## Call:
## lm(formula = thalach ~ target + age + slope + exang + cp + trestbps +
##     restecg, data = train_data[-unusual_idx, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.044  -9.493   1.371  11.650  47.310
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 170.20376    8.54789  19.912  < 2e-16 ***
## target1       5.82986    2.26682   2.572 0.010490 *
## age          -0.73603    0.09913  -7.425 7.28e-13 ***
## slope1       -0.80053    3.48606  -0.230 0.818496
## slope2       11.07447    3.60713   3.070 0.002291 **
```

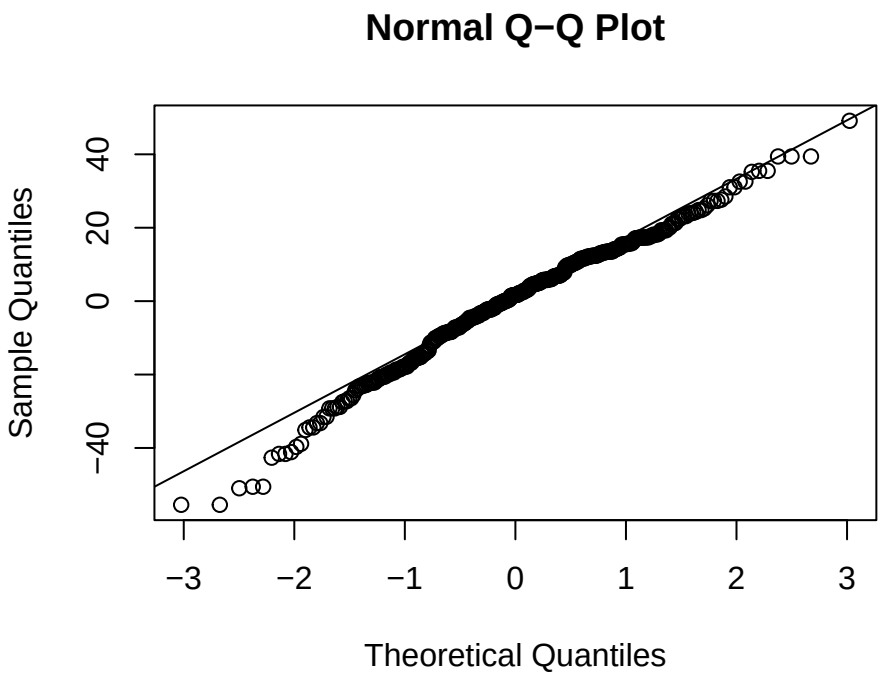


Figure 3: Q-Q Plot of Residuals

```
## exang1       -9.70972    2.18320  -4.447 1.14e-05 ***
## cp1           9.76435    2.77038   3.525 0.000475 ***
## cp2           4.29509    2.39756   1.791 0.074005 .
## cp3          14.44117    3.69288   3.911 0.000109 ***
## trestbps      0.09797    0.04914   1.993 0.046916 *
## restecg1     -2.21404    1.75759  -1.260 0.208537
## restecg2    -13.27747    9.80428  -1.354 0.176448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.7 on 386 degrees of freedom
## Multiple R-squared:  0.4911, Adjusted R-squared:  0.4766
## F-statistic: 33.86 on 11 and 386 DF,  p-value: < 2.2e-16
```

From the plot and the best fit lines (Fig. 4), the lines only change a little for the model with or without the influential points, so we shouldn't be concerned about those point affecting the results.

Innovation

Logistic regression is used to predict the outcome of a binary variable, for example 0 or 1. Suppose the probability of 1 occurs is p and the probability of 1 does not occur is 1-p, then the ratio of 1 occurring vs. 1 not-occurring is simply $\frac{1}{1-p}$.

Since we then need a linear regression to estimate the coefficient, we then apply log map the probability to

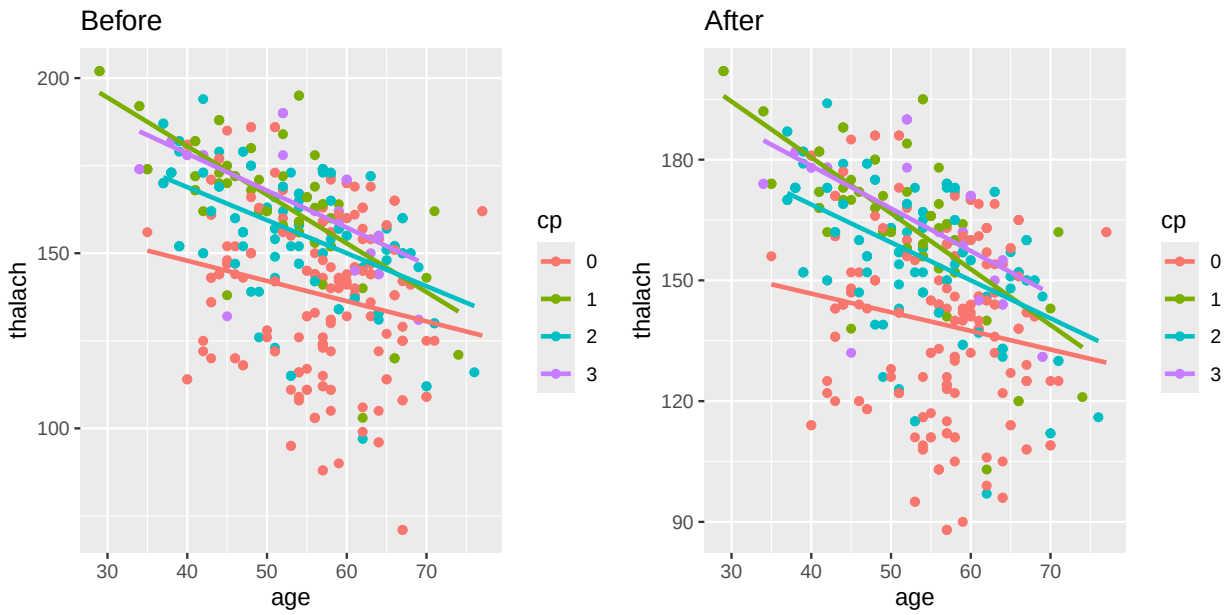


Figure 4: With/Without Influential Points

real number for linear regression, that is

$$Logit(p) = log(\frac{p}{1-p})$$

Apply $logit(p)$ as the new response and fit a linear model with the predictor variables.

$$Logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Then we inverse the logit to get p, namely,

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Technical Conditions

1. The response variable is binary. This fits our model that the response is the target variable, which indicates whether the patient have heart disease.
2. The model should have minimal multicollinearity. The LASSO regression is removing predictors that are having high correlations.
3. Sufficient sample size and independent observation. Our dataset is large and the patients conditions are independent.
4. There shouldn't be prefect separation. We don't have some predictors that can precisely predict the outcome.
5. There shouldn't be highly influential outliners. For this one we didn't particularly remove the outliners.

Method Justification

To explore further models, we decided to use logistic regression since we were interested in seeing how the variables can be used to predict heart disease, or the `target` variable. Since heart disease is a binary variable, OLS, ridge, and LASSO cannot be used since the response variable in these cases must be continuous.

Fitting Logistic Model

First, we used LASSO to determine which variables to include in the model. Based on the results from the LASSO regression, we fit a logistic model using the non-zero coefficients.

```
##
## Call:
## glm(formula = target ~ age + sex + cp + trestbps + fbs + restecg +
##     thalach + exang + slope + ca + thal, family = binomial, data = train_data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.215911   2.691304   1.195  0.23212
## age         -0.013983   0.023098  -0.605  0.54493
## sex1        -2.011873   0.445663  -4.514 6.35e-06 ***
## cp1          0.838787   0.482932   1.737  0.08241 .
## cp2          2.817552   0.506933   5.558 2.73e-08 ***
## cp3          3.350009   0.717658   4.668 3.04e-06 ***
## trestbps    -0.017784   0.009636  -1.846  0.06496 .
## fbs1        -0.539926   0.491566  -1.098  0.27204
## restecg1     0.540507   0.351903   1.536  0.12455
## restecg2    -0.380370   2.182778  -0.174  0.86166
## thalach      0.019662   0.011372   1.729  0.08381 .
## exang1      -0.853594   0.419316  -2.036  0.04178 *
## slope1       0.173169   0.746474   0.232  0.81655
## slope2       2.248040   0.765454   2.937  0.00332 **
## ca          -1.116049   0.193442  -5.769 7.95e-09 ***
## thal        -1.307612   0.284819  -4.591 4.41e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 552.83  on 399  degrees of freedom
## Residual deviance: 235.42  on 384  degrees of freedom
## AIC: 267.42
##
## Number of Fisher Scoring iterations: 6
```

ROC Curve

A tool to visualize logistic regression is a ROC (Receiver Operating Characteristic) curve, in which the true positive rate (sensitivity) is plotted against the false positive rate (1 - specificity) at different threshold values. The closer the ROC curve is to the top-left corner, the better the model is at distinguishing between positive and negative classes. A model with no discriminative power will have an ROC curve along the diagonal line, essentially equivalent to random guessing. As can be seen in Fig. 6, the ROC curve is relatively close to the top-left corner, indicating that the model has high sensitivity and high specificity.

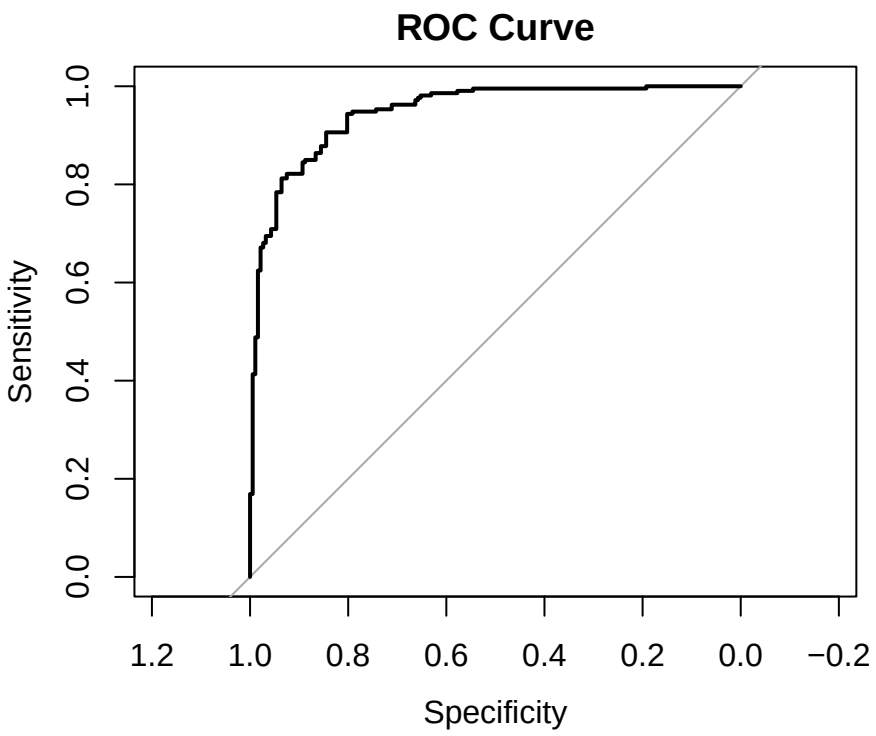


Figure 5: ROC Curve for logisitic model

AUC

The AUC, or area under the curve, is a measure of how well the model is able to discriminate between positive and negative cases. The AUC ranges from 0 to 1, with the following interpretations:

- < 0.5: No discriminative power (equivalent to random guessing).
- 0.5 - 0.7: Poor discrimination.
- 0.7 - 0.8: Acceptable discrimination.
- 0.8 - 0.9: Excellent discrimination.
- > 0.9: Outstanding discrimination.

```
## [1] "AUC: 0.947176822073259"
```

Since the AUC is above 0.9, we can conclude that our model is highly accurate at distinguishing between cases.

Optimal Threshold Value

Based on the ROC curve, we can determine the optimal threshold value, which is the value that maximizes the distance from the diagonal line (representing random chance) or maximizes the area under the ROC curve (AUC). If the predicted probability is less than the threshold, then the response can be classified as 0, or no heart disease. If the predicted probability is greater than threshold, then the response can be classified as 1, or having heart disease.

```
## [1] "Optimal threshold value: 0.524617014252765"
```

Conclusion

We first used forward step-wise model selection to determine our final model, which is

with a $R^2 = 0.4937$, which means that our model captures nearly 50% of variation in the response. We then conducted residual and influential point analysis.