Project Step 3

Minu Pabbathi, Ziqian Zhao, Paul Zhang

2024-06-02

Introduction

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains information about patients and risk factors for heart disease as well as whether or not they eventually end up developing heart disease. For the sake of convenience, we have randomly chosen 500 observations from the original data set. Here, we are going to further investigate the relationship between the response variable "maximum heart rate" and other variables, identifying the variables that might be good predictors of the response variables, and using different techniques to select the optimal model in a statistical sense.

Model Selection from the Correlation Matrix

From the correlation plot in step 1, we see that there is a relatively strong correlation between `max heart rate (thalach)`, `age`, and `chest pain type (cp)`. Our first model thus involves these two variables and examines how much variation these two variables explain in terms of change in `max heart rate` as well as the interaction between these two variables.

There are total 4 levels of `chest pain type` – `typical angina`, `atypical angina`, `non-anginal pain`, and `asymptomatic` labelled from 0 to 4.

Model 1:

$$\textbf{max heart rate}_i = \beta_0 + \beta_1 \textbf{age}_i + \beta_2 \mathbf{1}\{cp_i = 1\} + \beta_3 \mathbf{1}\{cp_i = 2\} + \beta_4 \mathbf{1}\{cp_i = 3\} + \varepsilon_i$$

Here, we are going to examine how different levels of `chest pain type` would make differences in the response – `max heart rate` when holding `age` fixed; and whether there is an interaction between `age` and `chest pain type`. We visualize the relationship with graph (Fig.1).

From the graph (Fig.1), it can be seen that the best-fit line is shifting a lot, so it is highly likely for `chest pain type` to have significant association with `maximum heart rate` after taking care of `age`.

```
##
## Call:
## lm(formula = thalach ~ age + as.factor(cp), data = heart_disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.344 -11.324   2.852  13.328  45.000
##
## Coefficients:
```

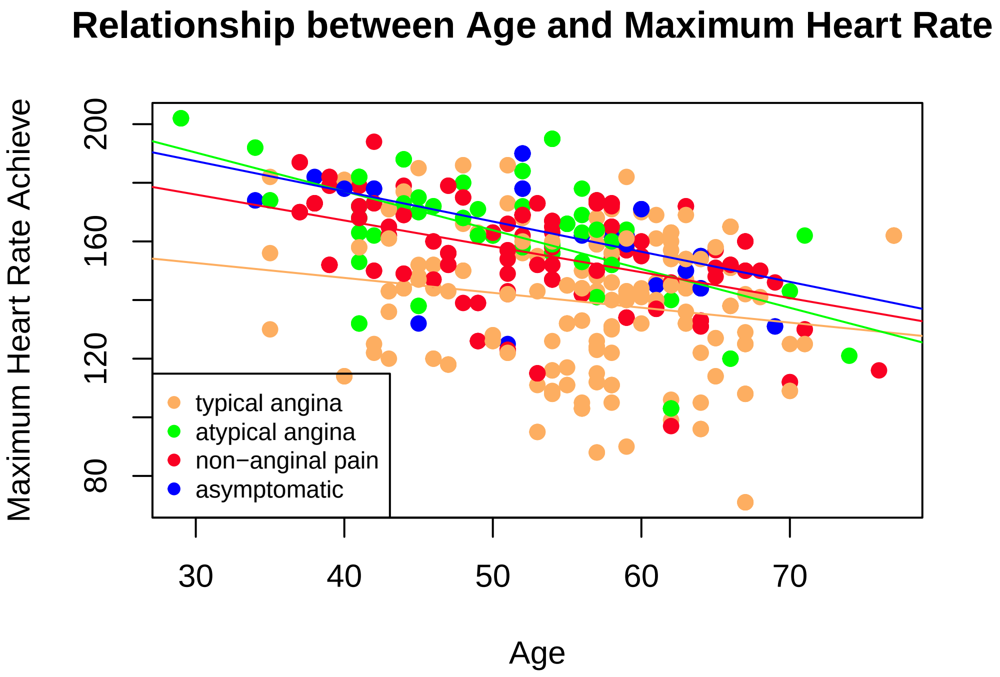

Figure 1: Age vs. Max Heart Rate color coded by chest pain type

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    186.08824    5.70444  32.622  < 2e-16 ***
## age             -0.83200    0.09871  -8.429 3.84e-16 ***
## as.factor(cp)1  19.31619    2.57223   7.510 2.79e-13 ***
## as.factor(cp)2  13.68309    2.06844   6.615 9.65e-11 ***
## as.factor(cp)3  21.42278    3.72605   5.749 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.69 on 495 degrees of freedom
## Multiple R-squared:  0.2922, Adjusted R-squared:  0.2864
## F-statistic: 51.08 on 4 and 495 DF,  p-value: < 2.2e-16
```

From the statistics, it can be seen that both predictors have low p-values which indicate the significance of both predictors. To further interpret the model and coefficient, it is estimated that the difference in `maximum heart rate(MHR)` between people of `atypical angina pain` and `typical angina pain` is about 21.5 bps; the difference in `MHR` between people of `non-anginal pain` and `typical angina pain` is about 14.45 bps; and the difference in `MHR` between people of `asymptomatic` and `typical angina` is about 21.54 bps after adjusting for `age`.

Also, from the plot, we can see that there is some slope changes between the best fit lines, so we are going to examine the interactions between variables.

$$\begin{aligned}\text{max heart rate}_i &= \beta_0 + \beta_1 age_i + \beta_2 \mathbf{1}\{\text{cp}_i = 1\} + \beta_3 \mathbf{1}\{\text{cp}_i = 2\} + \beta_4 \mathbf{1}\{\text{cp}_i = 3\} \\ &+ \beta_5 (age_i \cdot \mathbf{1}\{\text{cp}_i = 1\}) + \beta_6 (age_i \cdot \mathbf{1}\{\text{cp}_i = 2\}) + \beta_7 (age_i \cdot \mathbf{1}\{\text{cp}_i = 3\}) + \varepsilon_i\end{aligned}$$

```
##
```

```
## Call:
## lm(formula = thalach ~ age + as.factor(cp) + age * as.factor(cp),
##     data = heart_disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.829 -11.025   2.872  12.442  44.102
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        167.9039     8.9282  18.806  < 2e-16 ***
## age                 -0.5086     0.1571  -3.238  0.00129 **
## as.factor(cp)1      62.1145    14.5607   4.266 2.39e-05 ***
## as.factor(cp)2      34.5817    12.7009   2.723  0.00670 **
## as.factor(cp)3      50.3697    21.1105   2.386  0.01741 *
## age:as.factor(cp)1  -0.8151     0.2751  -2.963  0.00320 **
## age:as.factor(cp)2  -0.3746     0.2300  -1.629  0.10391
## age:as.factor(cp)3  -0.5211     0.3796  -1.373  0.17047
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 19.56 on 492 degrees of freedom
## Multiple R-squared:  0.3054, Adjusted R-squared:  0.2955
## F-statistic: 30.91 on 7 and 492 DF,  p-value: < 2.2e-16
```

From the statistics, we can see the p-values for the interactions with `cp1`, which indicate that those predictors are not significant and thus we conclude that there is not much interactions between `age` and `chest pain type`. Thus, we are not going to further consider interaction variables.

Another model that we are interested in is the relationship between `max heart rate`, `age`, and `sex`. We first visualize the relationship (Fig.2).

From the graph (Fig.2), we can see that the two best fit lines have a distinct trend, which could be a sign that `sex` is significantly related to `max heart rate` after taking care of `age`. In that case, we decided to analyze it as our second model and examine the interaction between variables.

Model 2:

$$\text{max heart rate}_i = \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3(age_i \cdot sex_i) + \varepsilon_i$$

```
##
## Call:
## lm(formula = thalach ~ age + as.factor(sex) + age * as.factor(sex),
##     data = heart_disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.780 -12.660   4.012  15.016  46.514
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         192.7892    10.2893  18.737   <2e-16 ***
## age                  -0.7208     0.1838  -3.922   0.0001 ***
## as.factor(sex)1      20.9383    12.2920   1.703   0.0891 .
## age:as.factor(sex)1  -0.4874     0.2216  -2.199   0.0283 *
## ---
```

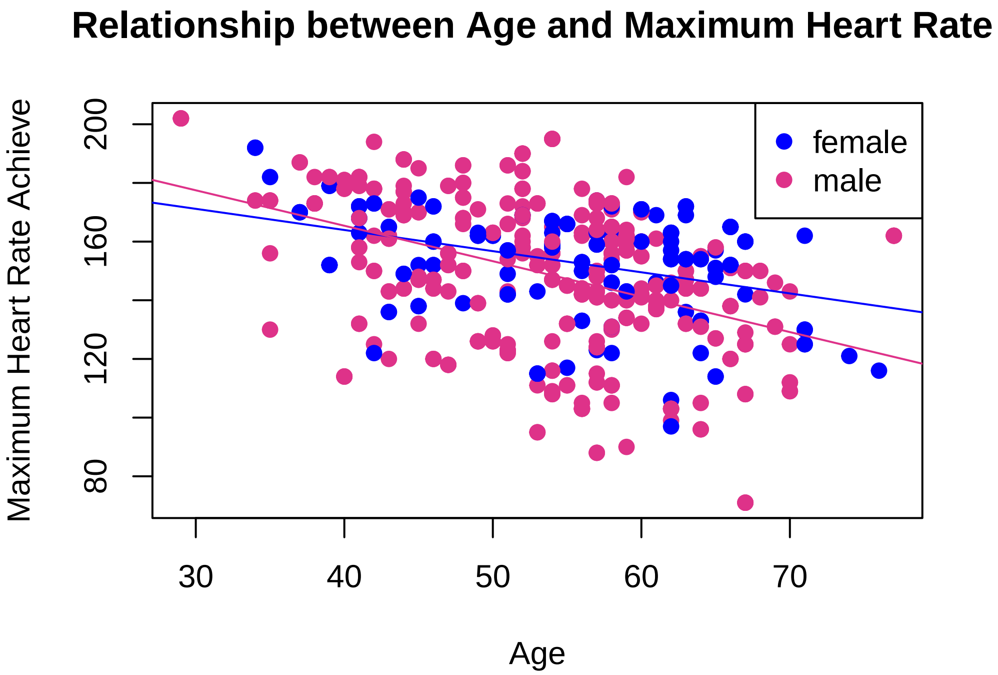

Figure 2: Age vs. Max Heart Rate color coded by gender

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.08 on 496 degrees of freedom
## Multiple R-squared:  0.1872, Adjusted R-squared:  0.1823
## F-statistic: 38.09 on 3 and 496 DF,  p-value: < 2.2e-16
```

From the statistical summary, we can say that the interaction between `age` and `sex` are significant given the p-value is relatively small.

Cross Validation

To compare the models, we conducted 5-fold cross-validation. The data was randomly shuffled and then separated into 5 folds. It was then trained on 4 of the folds and then tested on the 5th fold. This process was repeated so that each fold was used as a testing set once.

```
## [1] "MSE for Model 1: 393.295699277893"
```

```
## [1] "MSE for Model 2: 450.263433113128"
```

The mean MSE for model 1 is 393.2957, and the mean MSE for model 2 is 450.2634. Since the MSE for model 1 is lower, it is more accurate than model 2.

Statistical Approach for Selecting 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.

```
## Start:  AIC=2222.24
## thalach ~ 1
## 
##            Df Sum of Sq    RSS    AIC  F value    Pr(>F)
## + slope     2     48467 150614 2128.6  55.8319 < 2.2e-16 ***
## + target    1     45200 153882 2134.1 102.2180 < 2.2e-16 ***
## + age       1     43130 155951 2138.8  96.2443 < 2.2e-16 ***
## + cp        3     43413 155668 2142.1  32.1645 < 2.2e-16 ***
## + exang     1     32727 166355 2161.4  68.4617 2.790e-15 ***
## + oldpeak   1     26345 172736 2174.6  53.0764 2.165e-12 ***
## + ca        1     12384 186698 2201.8  23.0830 2.311e-06 ***
## + thal      1      3986 195095 2217.2   7.1105  0.008021 **
## + chol      1      2140 196941 2220.4   3.7816  0.052623 .
## + restecg   2      2911 196170 2221.1   2.5746  0.077641 .
## + fbs       1      1490 197591 2221.6   2.6243  0.106143
## <none>                  199081 2222.2
## + trestbps  1       425 198656 2223.5   0.7453  0.388571
## + sex       1       146 198935 2224.0   0.2555  0.613540
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Step:  AIC=2128.59
## thalach ~ slope
## 
##            Df Sum of Sq    RSS    AIC F value    Pr(>F)
## + age       1   23292.6 127322 2071.8 63.2983 2.559e-14 ***
## + cp        3   24468.9 126145 2072.5 22.2423 3.465e-13 ***
## + target    1   16843.1 133771 2089.1 43.5649 1.543e-10 ***
## + exang     1   13570.4 137044 2097.5 34.2617 1.119e-08 ***
## + ca        1    6754.3 143860 2114.5 16.2449 6.849e-05 ***
## + oldpeak   1    5471.4 145143 2117.6 13.0431 0.0003494 ***
## + thal      1    1148.5 149466 2127.9  2.6586 0.1039024
## <none>                  150614 2128.6
## + chol      1     790.6 149824 2128.8  1.8257 0.1775170
## + fbs       1     408.9 150205 2129.6  0.9420 0.3324432
## + sex       1      29.1 150585 2130.5  0.0669 0.7960989
## + trestbps  1       0.3 150614 2130.6  0.0008 0.9781217
## + restecg   2     844.8 149769 2130.6  0.9730 0.3789786
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Step:  AIC=2071.79
## thalach ~ slope + age
## 
##            Df Sum of Sq    RSS    AIC F value    Pr(>F)
## + cp        3   18206.4 109115 2023.8 19.0771 1.829e-11 ***
## + exang     1   11869.7 115452 2039.5 35.4696 6.384e-09 ***
## + target    1   10757.5 116564 2042.9 31.8395 3.495e-08 ***
## + oldpeak   1    2962.2 124359 2065.6  8.2178  0.004403 **
## + ca        1    2744.9 124577 2066.2  7.6016  0.006142 **
## + trestbps  1    1267.2 126054 2070.3  3.4681  0.063413 .
## + sex       1    1104.2 126217 2070.7  3.0183  0.083222 .
## <none>                  127322 2071.8
## + thal      1     591.7 126730 2072.2  1.6107  0.205252
## + restecg   2    1140.3 126181 2072.6  1.5543  0.212818
## + chol      1     189.6 127132 2073.3  0.5145  0.473667
## + fbs       1       0.1 127321 2073.8  0.0001  0.990279
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=2023.78
## thalach ~ slope + age + cp
##
##            Df Sum of Sq    RSS    AIC F value    Pr(>F)
## + exang     1    3469.0 105646 2014.5 11.2300 0.0008948 ***
## + target    1    2528.5 106587 2017.6  8.1130 0.0046607 **
## + oldpeak   1     998.8 108116 2022.6  3.1593 0.0763838 .
## + trestbps  1     968.5 108147 2022.7  3.0628 0.0810015 .
## + sex       1     918.1 108197 2022.8  2.9021 0.0893714 .
## + chol      1     651.0 108464 2023.7  2.0527 0.1528541
## <none>                  109115 2023.8
## + restecg   2    1177.3 107938 2024.0  1.8597 0.1572953
## + ca        1     405.8 108709 2024.5  1.2767 0.2593020
## + thal      1       7.5 109108 2025.8  0.0234 0.8786195
## + fbs       1       0.1 109115 2025.8  0.0003 0.9861814
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=2014.47
## thalach ~ slope + age + cp + exang
##
##            Df Sum of Sq    RSS    AIC F value  Pr(>F)
## + target    1   1579.53 104067 2011.2  5.1757 0.02353 *
## + trestbps  1   1334.28 104312 2012.0  4.3618 0.03749 *
## + oldpeak   1    759.58 104886 2014.0  2.4695 0.11700
## + chol      1    746.53 104900 2014.0  2.4268 0.12021
## + sex       1    676.01 104970 2014.2  2.1961 0.13929
## + restecg   2   1213.33 104433 2014.4  1.9751 0.14033
## <none>                  105646 2014.5
## + ca        1    393.59 105252 2015.2  1.2752 0.25959
## + thal      1     52.22 105594 2016.3  0.1686 0.68159
## + fbs       1     27.61 105618 2016.4  0.0891 0.76547
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=2011.2
## thalach ~ slope + age + cp + exang + target
##
##            Df Sum of Sq    RSS    AIC F value  Pr(>F)
## + trestbps  1   1828.18 102238 2007.0  6.0797 0.01417 *
## + restecg   2   1423.98 102643 2010.4  2.3515 0.09678 .
## + chol      1    598.18 103468 2011.2  1.9656 0.16182
## <none>                  104067 2011.2
## + thal      1    388.89 103678 2011.9  1.2753 0.25957
## + oldpeak   1    346.77 103720 2012.0  1.1367 0.28710
## + sex       1    199.97 103867 2012.5  0.6546 0.41905
## + fbs       1     56.27 104010 2013.0  0.1840 0.66827
## + ca        1     42.96 104024 2013.0  0.1404 0.70811
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=2007
## thalach ~ slope + age + cp + exang + target + trestbps
##
##           Df Sum of Sq    RSS    AIC F value  Pr(>F)
## + restecg  2   1406.61 100832 2006.2  2.3576 0.09621 .
## <none>                 102238 2007.0
## + chol     1    538.75 101700 2007.2  1.7958 0.18112
## + oldpeak  1    385.44 101853 2007.7  1.2829 0.25816
## + thal     1    236.24 102002 2008.2  0.7851 0.37620
## + sex      1    108.96 102129 2008.6  0.3617 0.54797
## + ca       1     92.67 102146 2008.7  0.3076 0.57955
## + fbs      1      0.00 102238 2009.0  0.0000 0.99982
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Step:  AIC=2006.15
## thalach ~ slope + age + cp + exang + target + trestbps + restecg
##
##           Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                 100832 2006.2
## + chol     1    368.31 100463 2006.9  1.2355 0.2671
## + oldpeak  1    218.46 100613 2007.4  0.7317 0.3929
## + thal     1    208.17 100624 2007.4  0.6972 0.4043
## + sex      1    174.81 100657 2007.5  0.5853 0.4448
## + ca       1    119.93 100712 2007.7  0.4013 0.5268
## + fbs      1     14.07 100818 2008.1  0.0470 0.8284
```

```
##
## Call:
## lm(formula = thalach ~ slope + age + cp + exang + target + trestbps +
##     restecg, data = train_data)
##
## Coefficients:
## (Intercept)       slope1       slope2          age          cp1          cp2
##    168.6064      -0.4769      12.5000      -0.8162      10.3982       4.6499
##         cp3       exang1      target1     trestbps     restecg1     restecg2
##     14.3554      -7.2996       6.8524       0.1312      -3.5296     -12.9090
```

By going through forward step-wise selection, we can see that the AIC value for the model is decreasing, which means that model is improved. We finally choose 7 predictors for the model, which are `slope`, `age`, `cp`, `exang`, `target`, `trestbps` and `restecg`.

Fitting Model

```
## 
## Call:
## lm(formula = thalach ~ slope + age + cp + exang + target + trestbps + 
##     restecg, data = train_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -54.652  -9.614   2.136  12.483  48.956 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 168.60639    9.31671  18.097  < 2e-16 ***
## slope1       -0.47691    3.64388  -0.131 0.895948    
## slope2       12.50003    3.78332   3.304 0.001055 ** 
## age          -0.81623    0.10757  -7.588 3.19e-13 ***
## cp1          10.39823    3.07148   3.385 0.000794 ***
## cp2           4.64986    2.65997   1.748 0.081357 .  
## cp3          14.35544    3.95727   3.628 0.000330 ***
## exang1       -7.29962    2.39856  -3.043 0.002523 ** 
## target1       6.85236    2.46722   2.777 0.005786 ** 
## trestbps      0.13117    0.05324   2.464 0.014247 *  
## restecg1     -3.52958    1.92071  -1.838 0.066993 .  
## restecg2    -12.90903   10.14802  -1.272 0.204222    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 17.27 on 338 degrees of freedom
## Multiple R-squared:  0.4935, Adjusted R-squared:  0.477 
## F-statistic: 29.94 on 11 and 338 DF,  p-value: < 2.2e-16
```

Based on the significance tests with 0.1 α-level, the coefficients were mostly all significant. The coefficients that were not found to be significant were still included since the other level of the predictor was significant.

Based on the calculated R^2, about 48% of the variance in maximum heart rate 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.

β **interpretation:**

For predictor `target`, we can say that when the patient has heart disease (`target=1`), their mean maximum heart rate is estimated to be 8.42 bpm higher than those who dose not have heart disease, after accounting for the other predictors in the model.

For predictor `exang`, we can say that when the patient has chest pain that is induced by exercise (`exang=1`), their mean maximum heart rate is estimated to be 8.12 bpm less than those who dose not have exercise induced chest pain, after accounting for the other predictors in the model.

Finding R^2 of Predicted Values

About 52% 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: 0.517275519273223"
```

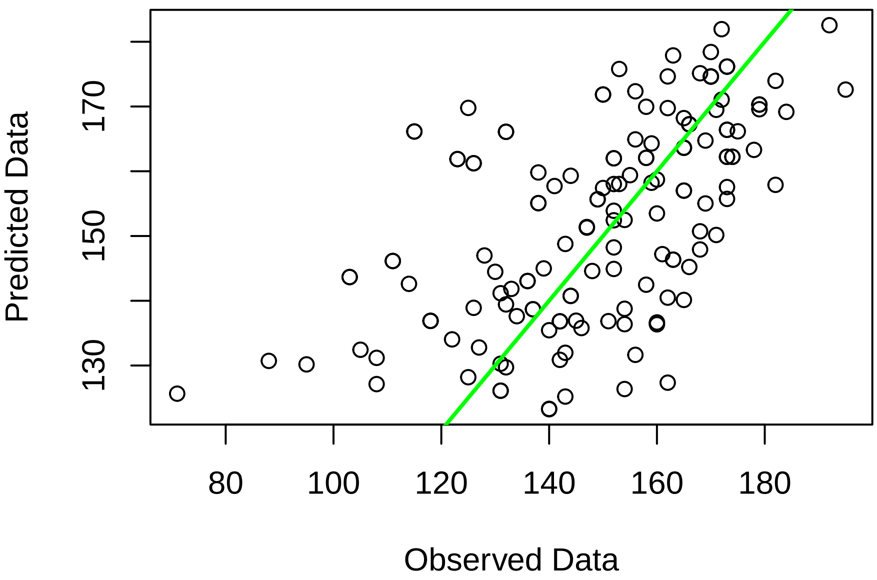

Figure 3: Plotting Prediciton

Checking Influential Points

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.

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
## -52.669  -8.909   1.809  11.272  47.039
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 169.53755    9.04785  18.738  < 2e-16 ***
```

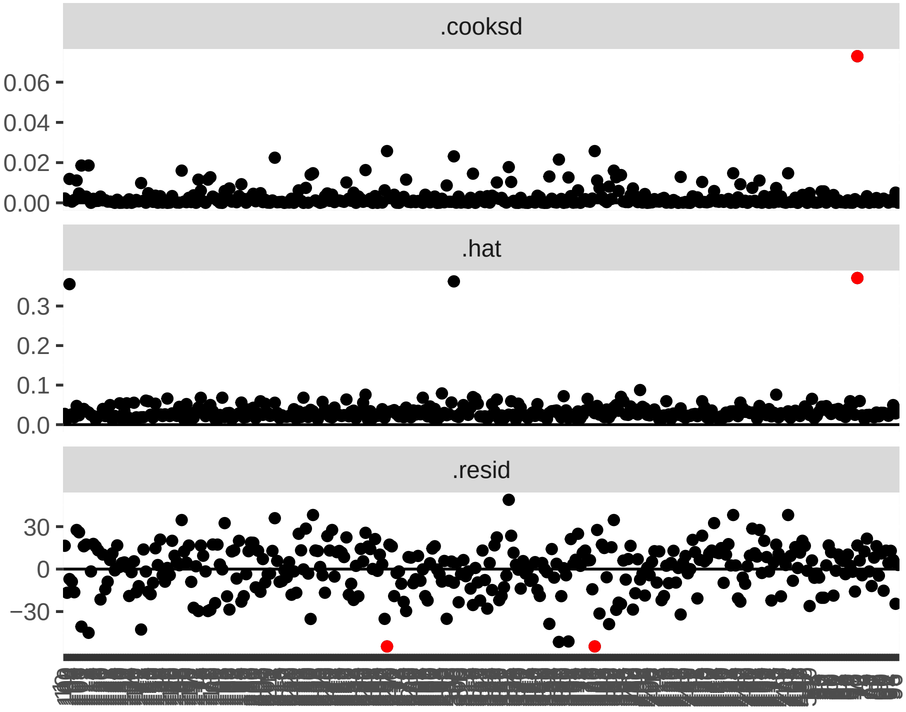

Figure 4: Influential Points Analysis

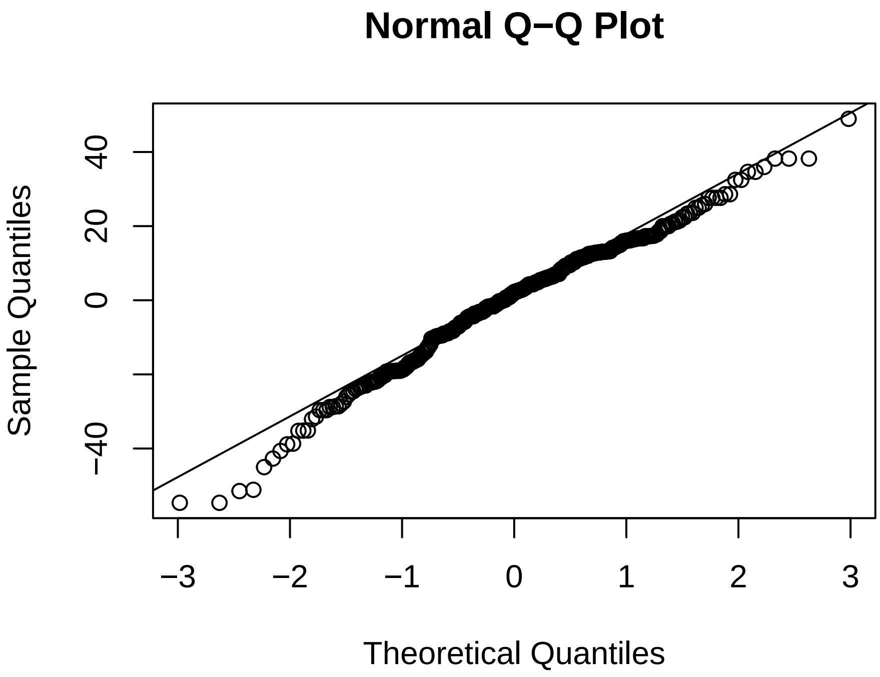

Figure 5: Q-Q Plot of Residuals

```
## target1       6.19683    2.39943   2.583 0.010228 *
## age          -0.76956    0.10490  -7.336 1.66e-12 ***
## slope1       -0.09566    3.53881  -0.027 0.978450
## slope2       12.29236    3.67354   3.346 0.000912 ***
## exang1       -8.78208    2.34962  -3.738 0.000218 ***
## cp1           9.52304    2.98783   3.187 0.001571 **
## cp2           3.64244    2.59130   1.406 0.160754
## cp3          13.63603    3.84516   3.546 0.000446 ***
## trestbps      0.11361    0.05182   2.192 0.029050 *
## restecg1     -2.48593    1.87775  -1.324 0.186441
## restecg2    -13.44567    9.85350  -1.365 0.173305
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 16.77 on 336 degrees of freedom
## Multiple R-squared:  0.4929, Adjusted R-squared:  0.4763
## F-statistic: 29.69 on 11 and 336 DF,  p-value: < 2.2e-16
```

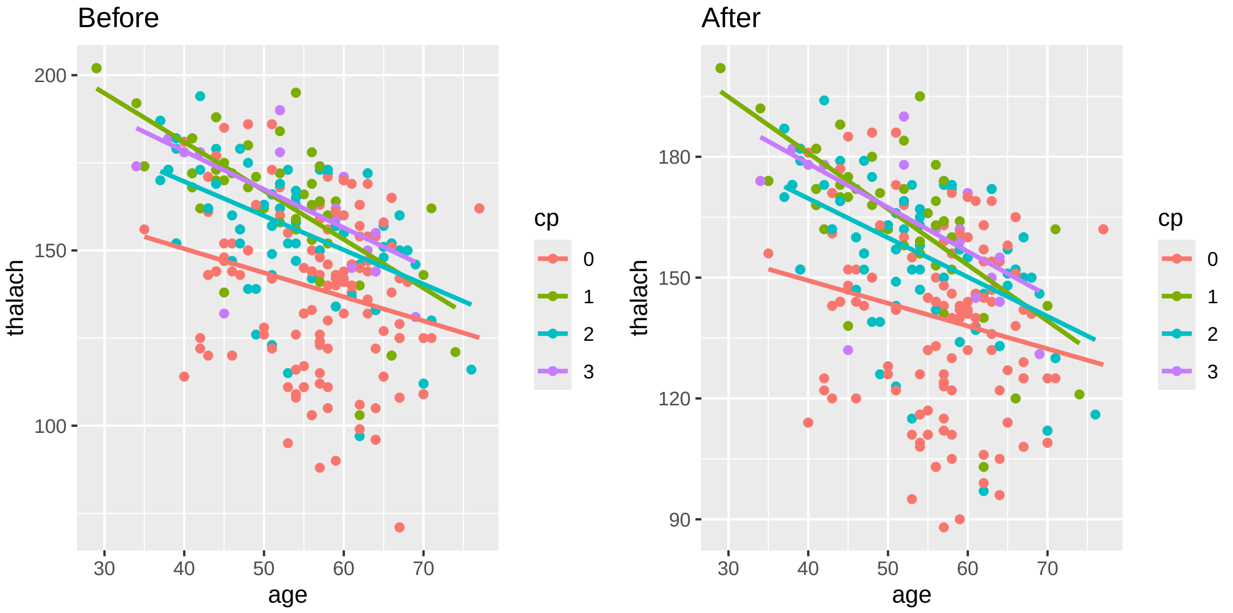

Figure 6: With/Without Influential Points

From the plot and the best fit lines (Fig.6), 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.

Model Interpretation

The final model we selected is

$$\begin{aligned}\textbf{max heart rate}_i = 169.5 &- 0.096 \cdot \mathbf{1}\{slope_i = 1\} + 12.3 \cdot \mathbf{1}\{slope_i = 2\} - 0.77 \cdot \textbf{age}_i \\ &+ 9.52 \cdot \mathbf{1}\{cp_i = 1\} + 3.64 \cdot \mathbf{1}\{cp_i = 2\} + 13.6 \cdot \mathbf{1}\{cp_i = 3\} \\ &- 8.78 \cdot \mathbf{1}\{exang_i = 1\} + 6.2 \cdot \mathbf{1}\{target_i = 1\} + 0.11 \cdot \textbf{trestbps}_i \\ &- 2.5 \cdot \mathbf{1}\{restecg_i = 1\} - 13.4 \cdot \mathbf{1}\{restecg_i = 2\}\end{aligned}$$

From the statistical test, we determine that variables slope, age, chest pain (cp), exercise induced angina (exang), target, resting blood pressure (trestbps), and rest electrocardiographic results (restecg) best explain the variation of maximum heart rate. Here, we are going to explain the coefficients of significance levels.

The difference in maximum heart rate between people at **upsloping** (slope = 2) and **downsloping** (slope = 0) of the peak exercise ST segment is estimated about 12.3 bpm after adjusting for all other variables.

For each unit increase in **age**, a decrease of 0.77 bpm in maximum heart rate is estimated while holding all other variable constant.

There should be expected an difference of 9.52 bpm in maximum heart rate between people with chest pain type of **atypical angina** and **typical angina** after holding all other variables constant; and there should also be an expected difference of 13.6 bpm in maximum heart rate between people without chest pain and **typical angina** after holding all other variable constant.

Comparing to people without **exercise induced angina** (exang = 0), it is expected an decrease of 8.78 bpm in maximum heart rate people with **exercise induced angina** after holding all other variables.

People with **heart disease** (target = 1) are expected to have an approximately 6.20 higher bpm comparing to people who do not have **heart disease** (target = 0) after holding all other variables constant.

For each unit (1 mmHg) increase in **resting blood pressure** , we expect a 0.11 increase in the maximum heart rate, holding all other variables constant.

Confidence Intervals and Prediction Intervals

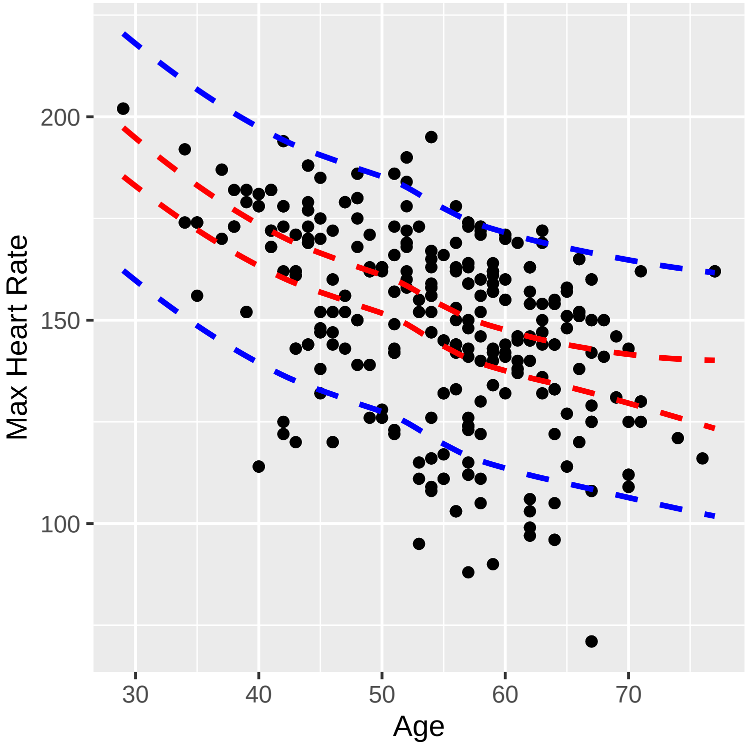

Figure 7: Confidence Intervals and Prediction Intervals

The region between red lines is the confidence interval at 90% significance level and the region between blue lines is the prediction interval at 90% significance level.

Confidence interval at 90% significance level:

```
##       ci.lwr   ci.upr
## 298 141.7386 150.5426
## 467 175.0778 183.9116
## 415 149.6876 163.5641
## 476 158.3448 166.1133
## 103 160.3107 169.5234
## 194 147.2740 160.7344
```

Prediction interval at 90% significance level:

```
##       ci.lwr   ci.upr
## 298 117.3147 174.9665
## 467 150.6665 208.3229
## 415 127.3053 185.9464
## 476 133.4777 190.9805
## 103 136.0592 193.7749
## 194 124.7321 183.2762
```

From the graph (as well as the table), we can see a larger interval for the prediction interval because it also needs to quantifies the uncertainty for both estimates and variation in the response whereas confidence interval quantifies uncertainty due only to the model parameter estimation.

Conclusion

We first examined two models with cross validation to determine the better one. Then we used forward step-wise model selection to determine our final model, which is

$$\begin{aligned}\textbf{max heart rate}_i = 169.5 &- 0.096 \cdot \mathbf{1}\{slope_i = 1\} + 12.3 \cdot \mathbf{1}\{slope_i = 2\} - 0.77 \cdot \textbf{age}_i \\ &+ 9.52 \cdot \mathbf{1}\{cp_i = 1\} + 3.64 \cdot \mathbf{1}\{cp_i = 2\} + 13.6 \cdot \mathbf{1}\{cp_i = 3\} \\ &- 8.78 \cdot \mathbf{1}\{exang_i = 1\} + 6.2 \cdot \mathbf{1}\{target_i = 1\} + 0.11 \cdot \textbf{trestbps}_i \\ &- 2.5 \cdot \mathbf{1}\{restecg_i = 1\} - 13.4 \cdot \mathbf{1}\{restecg_i = 2\}\end{aligned}$$

with a $R^2 = 0.4935$, which means that our model captures nearly 50% of variation in the response. We then tested our model on test data and found a $R^2 = 0.5173$, which indicates that the model captured 51.73% of variation in the response. We then conducted residual and influential point analysis, and compared the fit with and without influential points. The fit overall did not change with or without the influential point. Finally we calculated the confidence and prediction interval for our model, which can be seen above.