# Final Data Analysis

*Sarah Adilijiang*

**Part 1**

Check the current data types and then change the data types of catogorical variables “sex”, “chestpain”, “fbs”, “resteccg”, “exang”, “slope”, “extest” into factors.

```
# check the current data types of each variables
data_ori = read.table('heart.txt', header = TRUE)
str(data_ori)
```

```
## 'data.frame':    303 obs. of  13 variables:
##  $ age      : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex      : int  1 1 1 1 0 1 0 0 1 1 ...
##  $ chestpain: int  1 4 4 3 2 2 4 4 4 4 ...
##  $ restbp   : int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol     : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs      : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg  : int  2 2 2 0 2 0 2 0 2 2 ...
##  $ exang    : int  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope    : int  3 2 2 3 1 1 3 1 2 3 ...
##  $ fluoro   : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ extest   : int  1 0 2 0 0 0 0 0 2 2 ...
##  $ maxhr    : int  150 108 129 187 172 178 160 163 147 155 ...
```

```
# change the data type of categorical variables into factors
data = data_ori
data$sex = as.factor(data$sex)
data$chestpain = as.factor(data$chestpain)
data$fbs = as.factor(data$fbs)
data$restecg = as.factor(data$restecg)
data$exang = as.factor(data$exang)
data$slope = as.factor(data$slope)
data$extest = as.factor(data$extest)

# check the data types after the changing
str(data)
```

```
## 'data.frame':    303 obs. of  13 variables:
##  $ age      : int  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex      : Factor w/ 2 levels "0","1": 2 2 2 2 1 2 1 1 2 2 ...
##  $ chestpain: Factor w/ 4 levels "1","2","3","4": 1 4 4 3 2 2 4 4 4 4 ...
##  $ restbp   : int  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol     : int  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg  : Factor w/ 3 levels "0","1","2": 3 3 3 1 3 1 3 1 3 3 ...
##  $ exang    : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak  : num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope    : Factor w/ 3 levels "1","2","3": 3 2 2 3 1 1 3 1 2 3 ...
##  $ fluoro   : int  0 3 2 0 0 0 2 0 1 0 ...
##  $ extest   : Factor w/ 3 levels "0","1","2": 2 1 3 1 1 1 1 1 3 3 ...
##  $ maxhr    : int  150 108 129 187 172 178 160 163 147 155 ...
```

**Part 2**

Address with the missing values.

```
# check the number of missing values
data[is.na(data)]
```

```
## [1] NA NA NA NA NA NA
```

```
sum(is.na(data))
```

```
## [1] 6
```

```
nrow(data)
```

```
## [1] 303
```

```
sum(is.na(data))/nrow(data)
```

```
## [1] 0.01980198
```

```
# look at summary of each covariates and find the missing values
summary(data)
```

```
##       age        sex     chestpain     restbp           chol       fbs
##  Min.   :29.00   0: 97   1: 23     Min.   : 94.0   Min.   :126.0   0:258
##  1st Qu.:48.00   1:206   2: 50     1st Qu.:120.0   1st Qu.:211.0   1: 45
##  Median :56.00           3: 86     Median :130.0   Median :241.0
##  Mean   :54.44           4:144     Mean   :131.7   Mean   :246.7
##  3rd Qu.:61.00                     3rd Qu.:140.0   3rd Qu.:275.0
##  Max.   :77.00                     Max.   :200.0   Max.   :564.0
##
##  restecg exang      oldpeak     slope       fluoro        extest
##  0:151   0:204   Min.   :0.00   1:142   Min.   :0.0000   0   :166
##  1:  4   1: 99   1st Qu.:0.00   2:140   1st Qu.:0.0000   1   : 18
##  2:148           Median :0.80   3: 21   Median :0.0000   2   :117
##                  Mean   :1.04           Mean   :0.6722   NA's:  2
##                  3rd Qu.:1.60           3rd Qu.:1.0000
##                  Max.   :6.20           Max.   :3.0000
##                                         NA's   :4
##      maxhr
##  Min.   : 71.0
##  1st Qu.:133.5
##  Median :153.0
##  Mean   :149.6
##  3rd Qu.:166.0
##  Max.   :202.0
##
```

```
# locate the missing values
which(is.na(data$fluoro))
```

```
## [1] 167 193 288 303
```

```
which(is.na(data$extest))
```

```
## [1]  88 267
```

There're six missing values in this data set, which are located at different data points in the variabes "fluoro" and "extest". Variable "extest" is a factor variable, so it's not appropriate to use mean or regression to impute the missing values for it. Though variable "fluoro" is a quantitative variable, it only has four possible integer

values: 0,1,2,3, and more than half of the values are 0 (median of "fluoro" is 0.0000). No matter using mean (0.6722) or regression, it will not properly give a precise integer value.

Most importantly, these six missing values only occupy 2% of the total number of data points (303). Therefore, the best way to handle these small number of missing values in the data set is to simply delete the data points that contain these missing values.

```
# delete data points with missing values
data2 = data[-c(167,193,288,303,88,267), ]
rownames(data2) = as.character(seq(1,297))  # indices are reordered

# check if there is any missing values now
sum(is.na(data2))
```

```
## [1] 0
```

### Part 3

Check constant variance and normality of errors.

For now, we fit a linear model with no interaction terms and include all the variables to see whether a linear model is appropriate for this data set or not. First, we check if the errors have constant variance, and then we check if the errors basically follow a normal distribution.

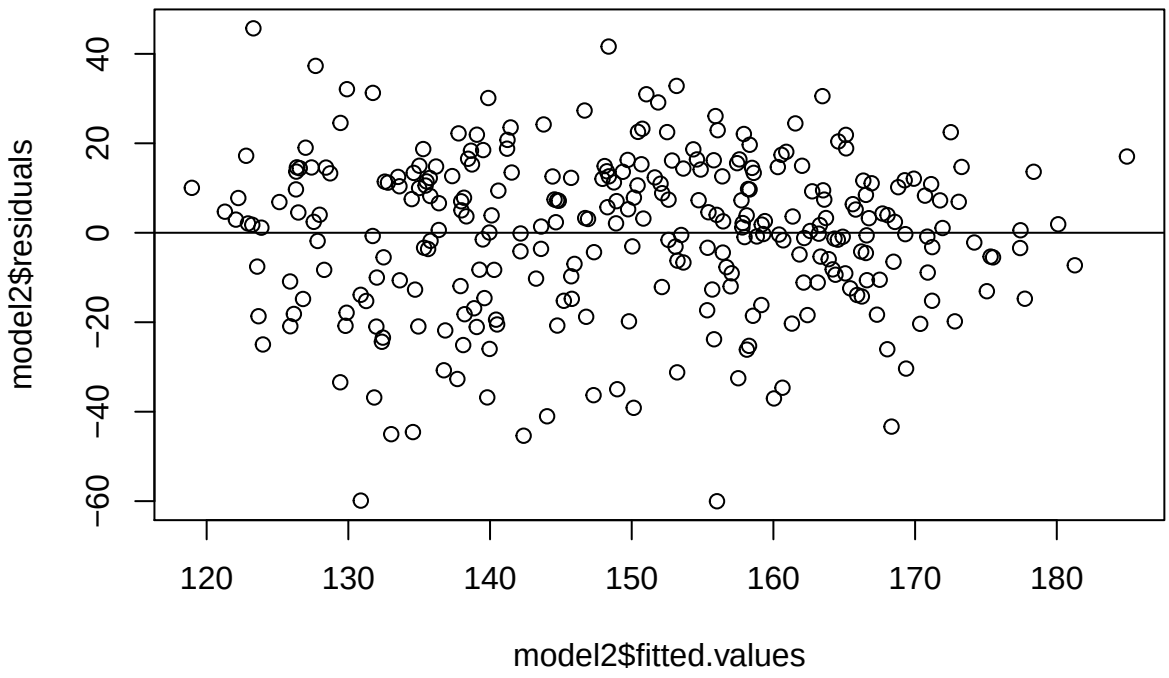
(1) Constant variance:

```
# plot residuals against fitted y
model2 = lm(maxhr~., data2)
plot(model2$fitted.values, model2$residuals)
abline(h=0)

# use a formal test: Breusch-Pagan test to check the heteroscedasticity
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```
bptest(model2)
```

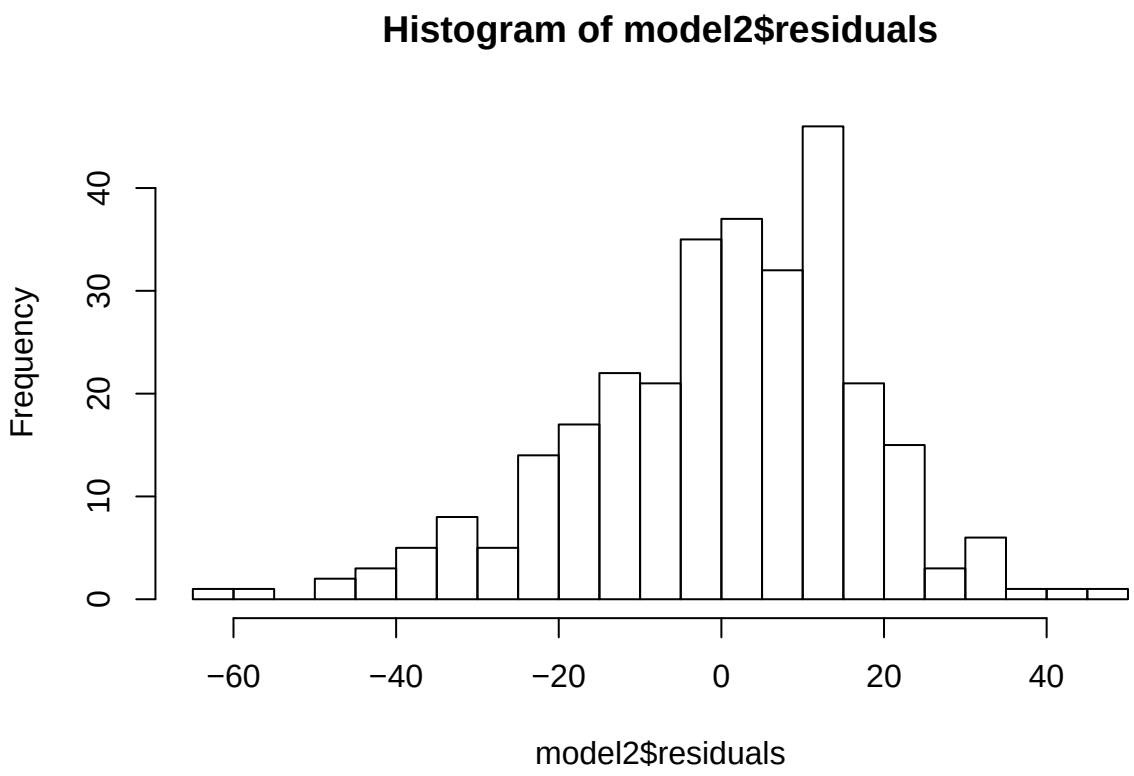
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 18.503, df = 17, p-value = 0.3578
```

There is no significant pattern of heteroscedasticity or nonlinearity in the residuals vs fitted values plot.

And the Breusch-Pagan test also indicates that there is no significant evidence for heteroscedasticity in this model. Breusch-Pagan test's Null Hypothesis is homoscedasticity of the regression model, the Alternative being a heteroscedastic model. Here the p-value $= 0.3578 > 0.1$, so we Do Not Reject Null Hypothesis (homoscedasticity) at $\alpha = 10\%$ significance level or smaller. Therefore, there is no significant evidence for heteroscedasticity.

(2) Normality of errors:

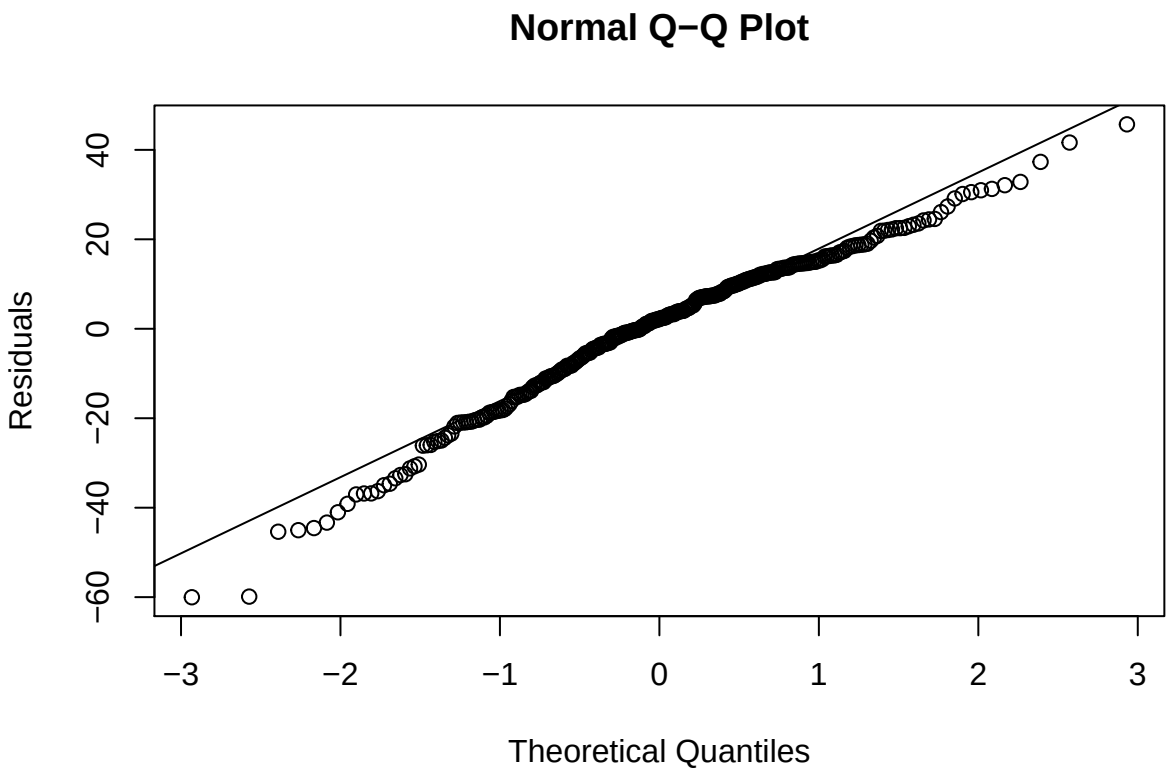
```
# look at the histogram of the residuals
hist(model2$residuals, breaks = 20)
```




```
# Shapiro-Wilk test
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.97721, p-value = 0.000114
```

```
# Q-Q plot
qqnorm(model2$residuals, ylab = "Residuals")
qqline(model2$residuals)
```




The main part of the histogram of residuals seems to follow a symmetric, bell-shape. But it is a little right skewed and the left tail is a little longer than the right tail. So though the main portion of residuals look normal, the whole distribution seems not to be quite normal.

The Shapiro-Wilk test's Null Hypothesis is that data follows a normal distribution. Here the Shapiro-Wilk test has a p-value $= 0.000114 < 0.001$, so we Reject Null Hypothesis at $\alpha = 0.1\%$ significance level. Therefore, the Shapiro-Wilk test also indicates that the residuals do not follow a normal distribution.

However, the most part of Q-Q plot approximately follows the line. Though the left tail and right tail do not follow the line, it looks like a short-tailed nonnormality error problem. When nonnormality is found, the resolution depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.

Conclusion:

According to the two examinations above, we can say that a linear model is basically appropriate for this data set. The linear model does not have significant heteroscedasticity or nonlinearity problem, and the short-tailed nonnormality problem is not serious and can reasonably be ignored.

**Part 4**

Check individual points.

(1) Large leverage points.

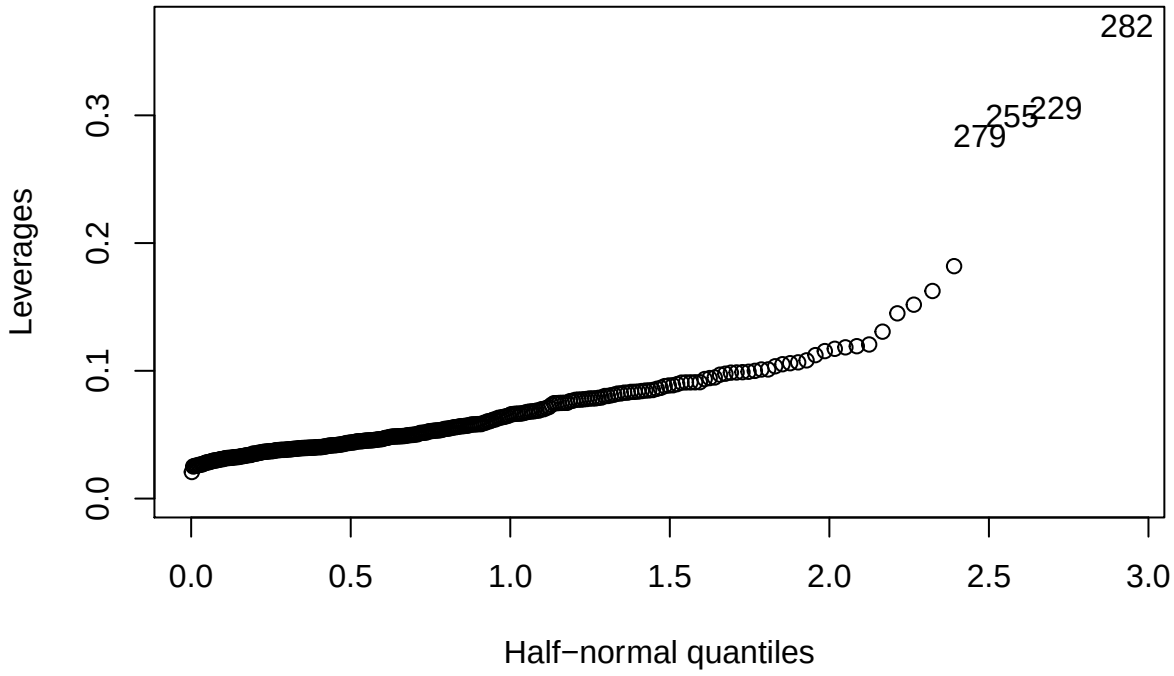
```
# find large leverage points
diag_H = hatvalues(model2)    # i.e. leverages
sum(diag_H > 2 * mean(diag_H))
```

```
## [1] 9
```

```
diag_H[diag_H > 2 * mean(diag_H)]
```

```
##         1        91       126       152       186       229       255
## 0.1449929 0.1624795 0.1306628 0.1819268 0.1517779 0.3057359 0.2994398
##       279       282
## 0.2834849 0.3701956
```

```
# find large leverage points via half-normal plot
library(faraway)
leverages = influence(model2)$hat
halfnorm(leverages, nlab = 4, ylab = "Leverages")
```




There are nie observations that have hat values which are more than twice the mean of leverage values. From the half-norm plot, we can see that among the nine large leverage points there are four ones that have hat values much higher than the others, they are the 279th, 255th, 229th and 282th observations. (Note that here the indices of observations are reordered from 1 to 297 after removing the missing values.)

(2) Outliers.

```
# find potential outliers
jack <- rstudent(model2)
jack[which.max(abs(jack))]
```

```
##       242
## -3.517403
```

```
# Here we use 5% significance level to perform the t-test
alpha = 0.05
```

```r
n = nrow(data2)
p = length(model2$coefficients)

# t-test without Bonferroni correction
t = qt(1-alpha/2, df = n-p-1)
jack[abs(jack) > t]
```

```
##        28        47        80       112       114       137       170
## -2.082132 -2.109487 -2.075067  2.498231 -2.613802 -2.369866 -2.241598
##       174       197       211       221       222       230       234
## -2.574674 -2.525711  2.184904 -2.096333  2.640857 -2.022217 -2.166588
##       242       243       292
## -3.517403 -3.499560 -2.648321
```

```r
sum(abs(jack) > t)
```

```
## [1] 17
```

```r
# t-test with Bonferroni correction
t = qt(1-(alpha/2)/n, df = n-p-1)
jack[abs(jack) > t]
```

```
## named numeric(0)
```

```r
sum(abs(jack) > t)
```

```
## [1] 0
```

```r
# outlier test
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

```r
outlierTest(model2)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 242 -3.517403         0.00050882      0.15112
```

Seventeen observations seem to be outliers to the regression model under the looser measurement without Bonferroni correction.

However, when using the Bonferroni correction, there are no outliers any more.

(3) Influential points.

```r
# find influential points with large Cook's Distance
cook = cooks.distance(model2)
n = nrow(data2)
cook[cook > 4/n]
```

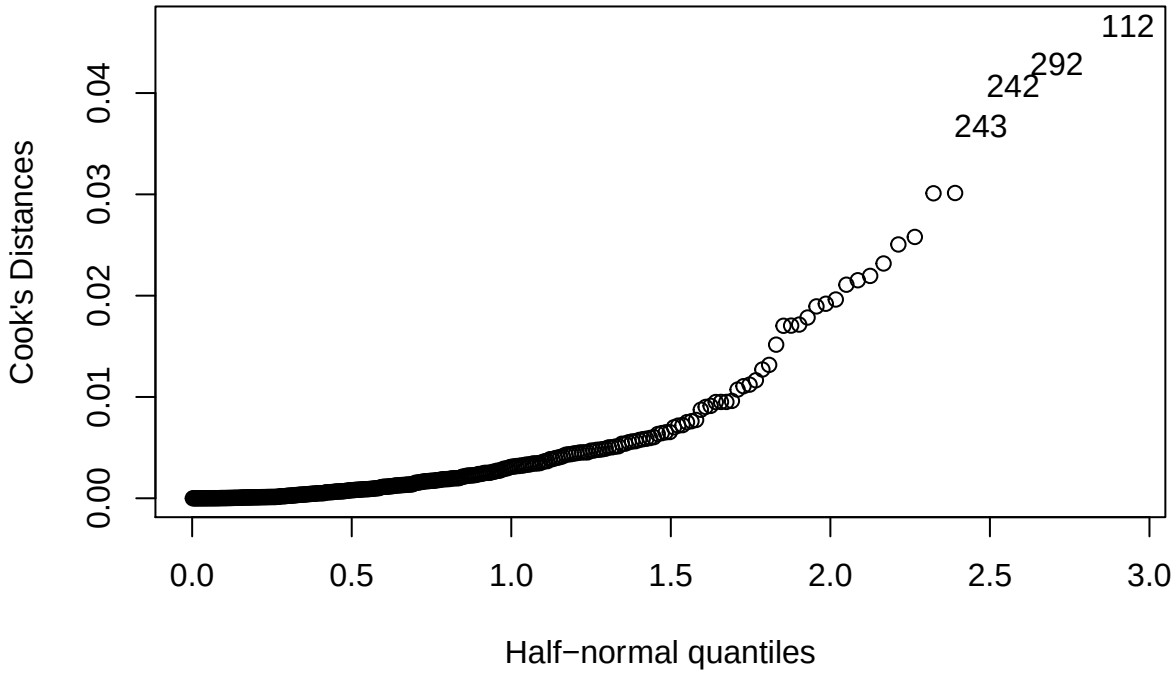
```
##         28         60        112        114        137        154
## 0.03010455 0.01784120 0.04667887 0.01704414 0.01714450 0.01920179
```

```
##         161         185         191         197         209         211
## 0.01702737 0.02151969 0.01516237 0.02505418 0.02195008 0.02317553
##         222         229         230         234         242         243
## 0.01962043 0.02107968 0.01893767 0.02579506 0.04069905 0.03678542
##         282         292
## 0.03013332 0.04288645
```

```
sum(cook > 4/n)
```

```
## [1] 20
```

```
# find influential points with large Cook's Distance via half-normal plot
halfnorm(cook, nlab = 4, ylab = "Cook's Distances")
```




Generally, a Cook's Distance $D_i$ is considered large if $D_i > 4/n$. Here there are twenty observations that have large Cook's Distances thus have large influence on the fitted model. The half-norm plot shows that the highest influential points are the 243th, 242th, 292th, and 112th observations.

Conclusion:

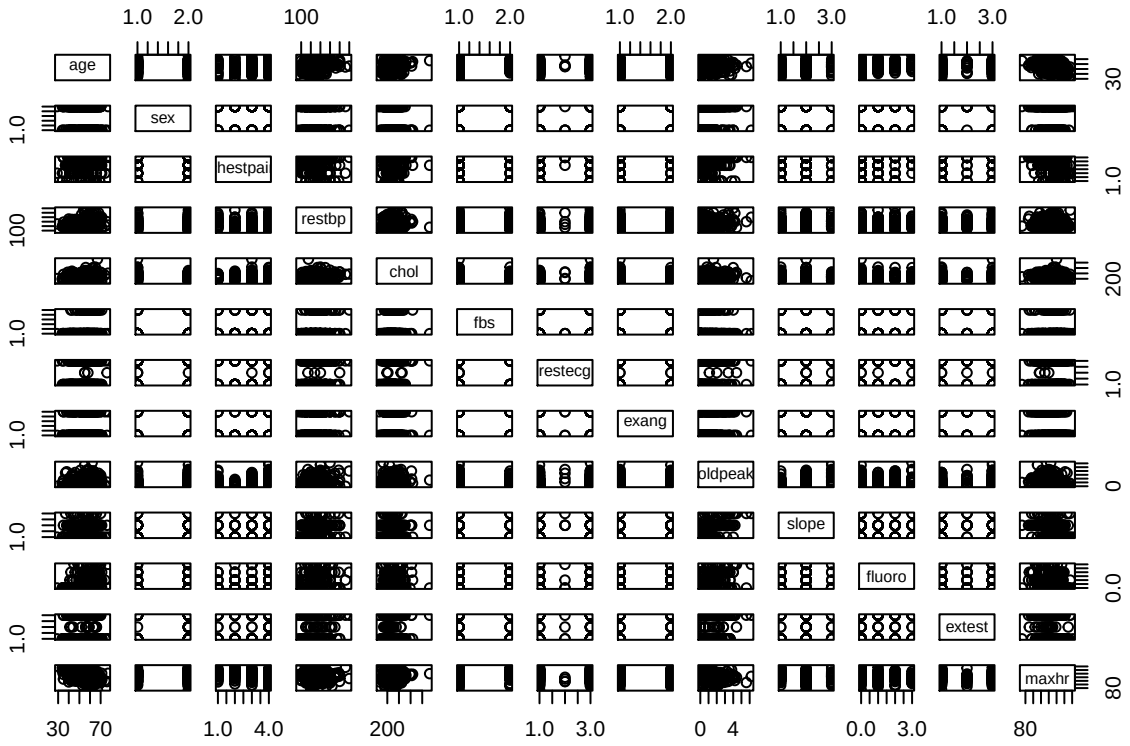
None of the detected large leverage points are outliers even under the looser measurement without Bonferroni correction. So there's no need to remove these points. And the largest influential points (243th, 242th, 292th, and 112th observations) are all just detected outliers without Bonferroni correction, thus there's no significant necessity to remove these points either.

As a result, there's no significant reason to remove any individual points in this linear model.

**Part 5**

Check the correlations between covariates.

```
# have a look at pairwise scatterplots
pairs(data2)
```




```
# check the gvif values of each variables
library(car)
car::vif(model2)
```

```
##                GVIF Df GVIF^(1/(2*Df))
## age        1.324564  1        1.150897
## sex        1.341676  1        1.158307
## chestpain 1.638288  3        1.085755
## restbp     1.216026  1        1.102736
## chol       1.140080  1        1.067745
## fbs        1.126959  1        1.061583
## restecg    1.175698  2        1.041295
## exang      1.374633  1        1.172447
## oldpeak    1.858347  1        1.363212
## slope      1.792841  2        1.157139
## fluoro     1.376063  1        1.173057
## extest     1.666921  2        1.136263
```

Since the correlation matrix cannot be used for categorical variables, here I used pairwise scatterplots to have a look at potential collinearity problems. Except that some variablies are skewed to one side, there seems not be significant high correlations between covariates.

To formally check if there is any collinearity problem, I checked the variance inflation factors (VIF). It is suggested that the straightforward VIF can't be used if there are variables with more than one degree of

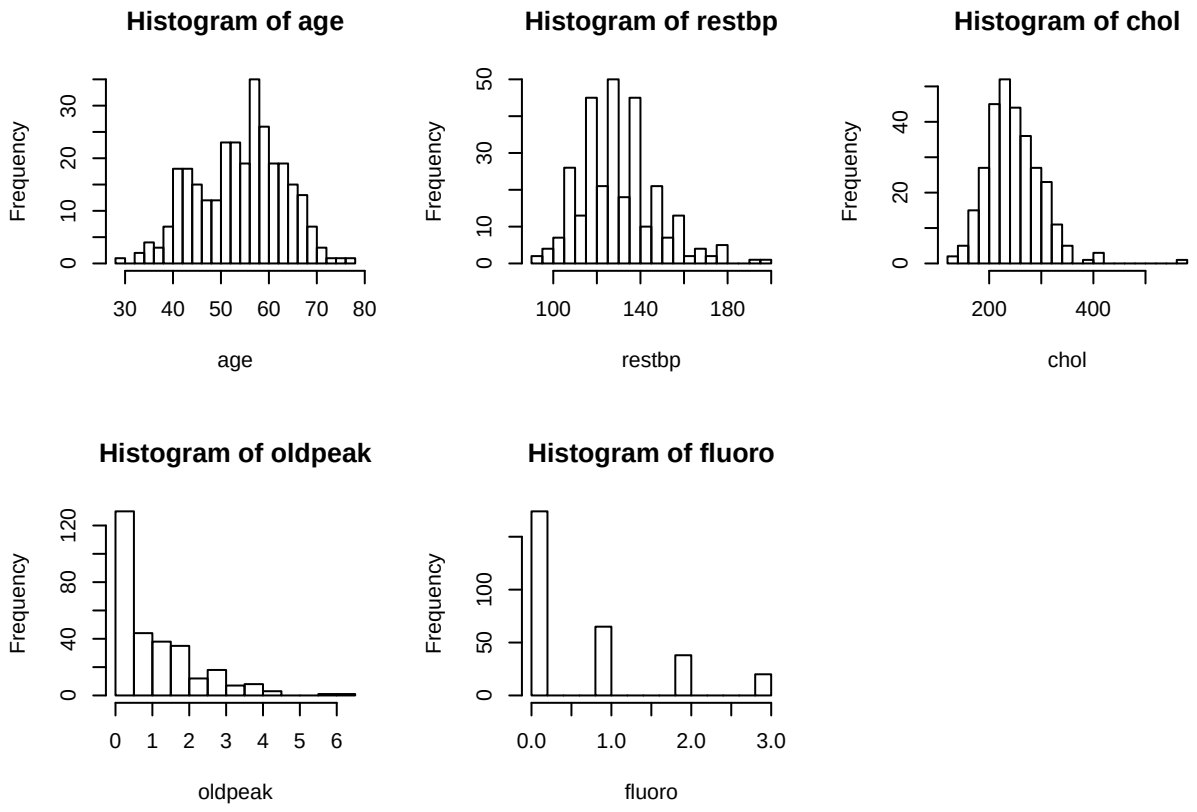
freedom (e.g. categorical variables with more than two levels) and instead we should use the GVIF (generalized variance inflation factor) function in the car package. For continous variables, the GVIF values are the same as VIF values, however, for categorical variables, GVIF values are the VIFs corrected by the number of degrees of freedom (df) of the categorical variables.

Here we see that none of GVIF values is greater than 5, thus there is no significant variance inflation prolems as well as collinearity problems in this data set.

### Part 6

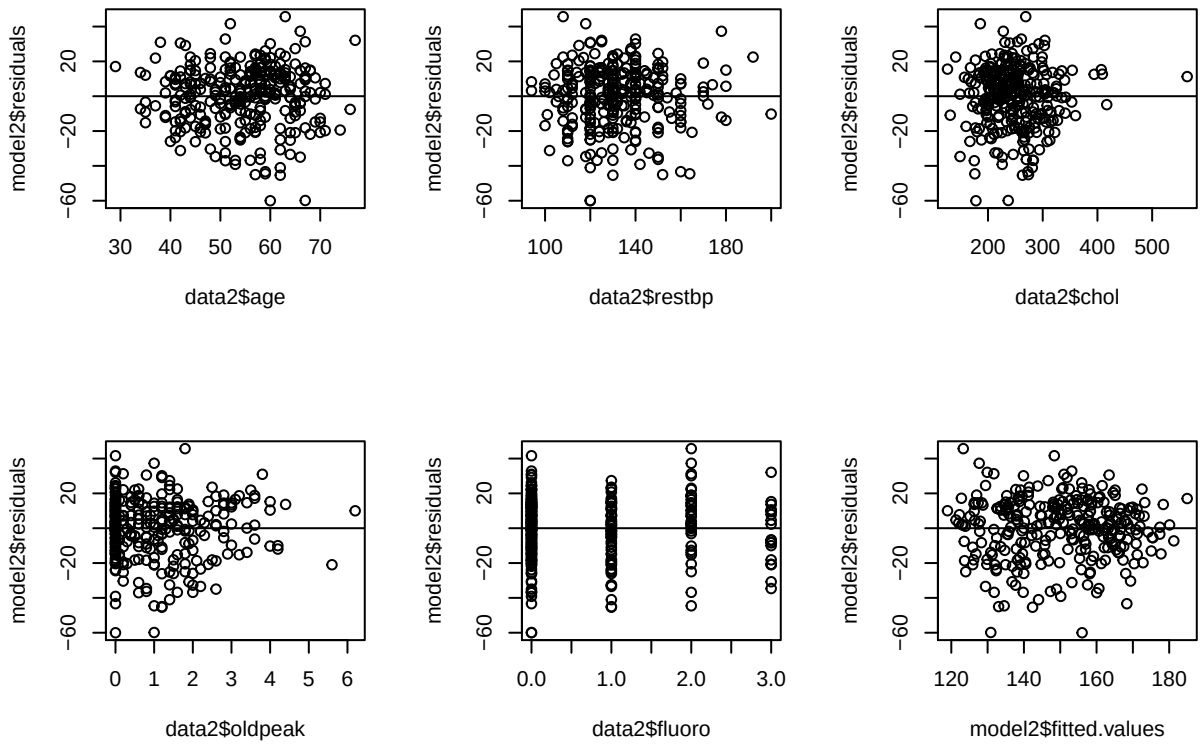
Check if there is any need of transformations.

```
# plot the histograms of quantitative variables
par(mfrow = c(2,3))
hist(data2$age, main="Histogram of age", xlab="age", breaks=20)
hist(data2$restbp, main="Histogram of restbp", xlab="restbp", breaks=20)
hist(data2$chol, main="Histogram of chol", xlab="chol", breaks=20)
hist(data2$oldpeak, main="Histogram of oldpeak", xlab="oldpeak", breaks=20)
hist(data2$fluoro, main="Histogram of fluoro", xlab="fluoro", breaks=20)

# plot residuals against quantitative variables
par(mfrow = c(2,3))
```




```
plot(data2$age, model2$residuals, abline(h=0))
plot(data2$restbp, model2$residuals, abline(h=0))
plot(data2$chol, model2$residuals, abline(h=0))
plot(data2$oldpeak, model2$residuals, abline(h=0))
plot(data2$fluoro, model2$residuals, abline(h=0))
```

```r
plot(model2$fitted.values, model2$residuals, abline(h=0))
```

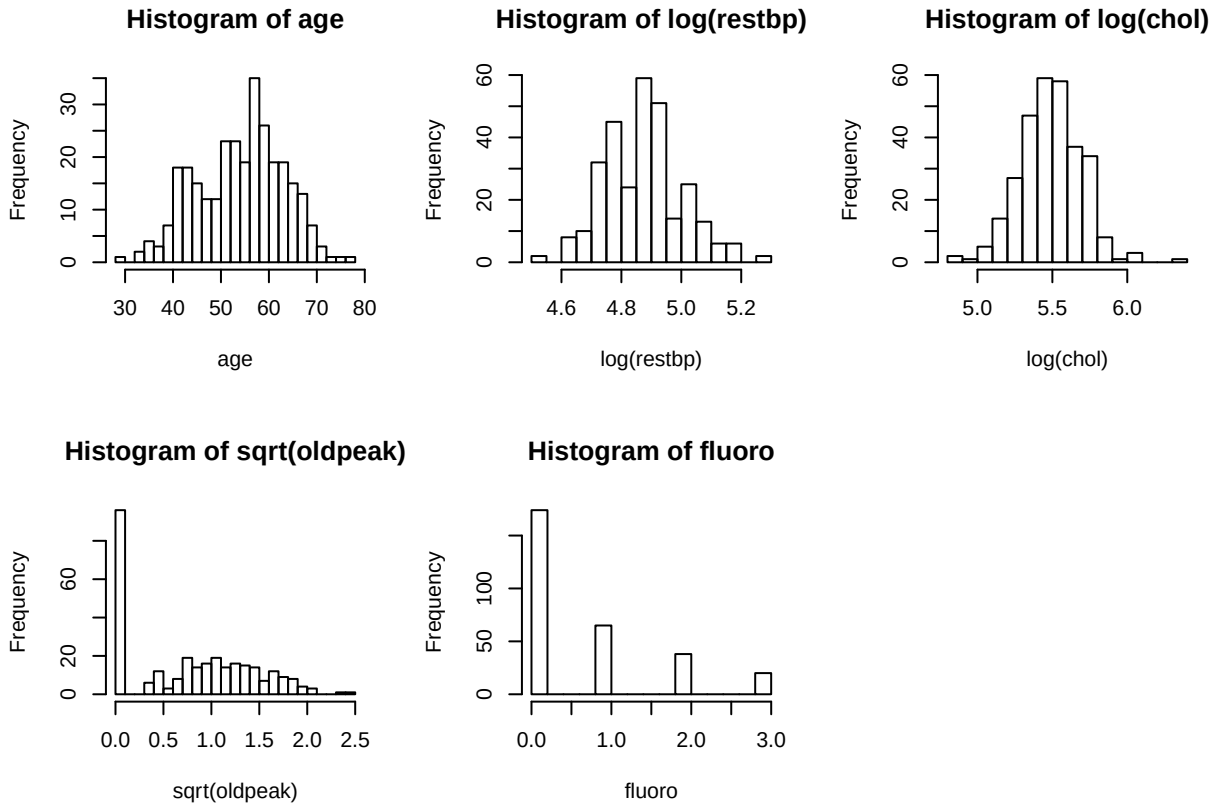



In the histograms of quantitative variables, we can see that the distribution of "restbp", "chol", "oldpeak" and "fluoro" is skewed to the left. And in the residuals vs quantitative variables plot, the distribution of residuals are also skewed to the left for variables "restbp", "chol" and "oldpeak", especially for "chol". For variable "fluro", though most of the values are at located at zero, the corresponding residuals at each value seems to be normally distributed with similar variance.

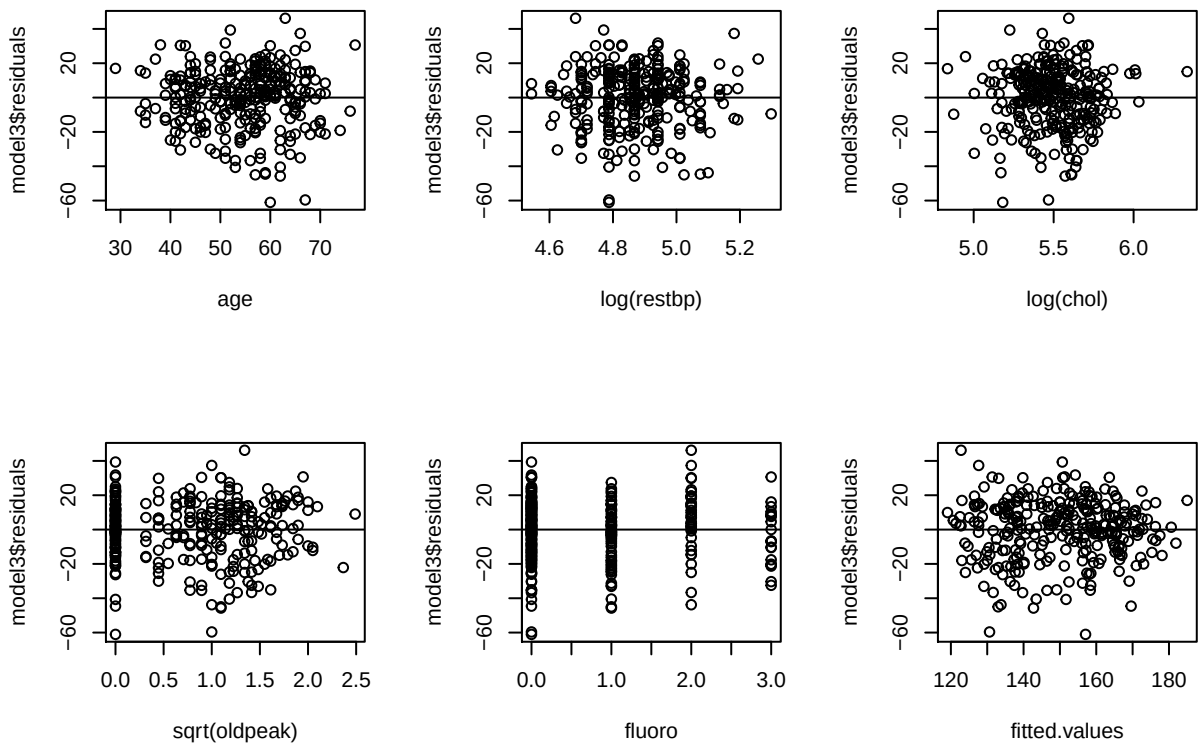
Therefore, here I considered some transformations of quantitative variables "restbp", "chol" and "oldpeak". Since "restbp" and "chol" are positive values, I will perform log transformations of these two variables. Since "oldpeak" is nonnegative and includes quite a lot of zero values, I will perform a square root transformation of this variable.

```r
# variable transformations
data3 = data2
data3$restbp = log(data3$restbp)
data3$chol = log(data3$chol)
data3$oldpeak = sqrt(data3$oldpeak)

# fit the new model with transformed variables
model3 = lm(maxhr~., data3)

# plot the histograms of quantitative variables
par(mfrow = c(2,3))
hist(data3$age, main="Histogram of age", xlab="age", breaks=20)
hist(data3$restbp, main="Histogram of log(restbp)", xlab="log(restbp)", breaks=20)
hist(data3$chol, main="Histogram of log(chol)", xlab="log(chol)", breaks=20)
```

```r
hist(data3$oldpeak, main="Histogram of sqrt(oldpeak)", xlab="sqrt(oldpeak)", breaks=20)
hist(data3$fluoro, main="Histogram of fluoro", xlab="fluoro", breaks=20)

# plot residuals against quantitative variables
par(mfrow = c(2,3))
```




```r
plot(data3$age, model3$residuals, abline(h=0), xlab="age")
plot(data3$restbp, model3$residuals, abline(h=0), xlab="log(restbp)")
plot(data3$chol, model3$residuals, abline(h=0), xlab="log(chol)")
plot(data3$oldpeak, model3$residuals, abline(h=0), xlab="sqrt(oldpeak)")
plot(data3$fluoro, model3$residuals, abline(h=0), xlab="fluoro")
plot(model3$fitted.values, model3$residuals, abline(h=0), xlab="fitted.values")
```




After the transformations, the distribution of "log(restbp)" and "log(chol)" are centered to the middle. Though most of "sqrt(oldpeak)" values are still located at zero, the positive values are centered to the middle as well. And in the residuals vs quantitative variables plot, the distribution of residuals of "log(restbp)", "log(chol)", and positive values of "sqrt(oldpeak)" are more spreaded and looks normally distributed. On the other hand, the distribution of residuals against the fitted values have no significant changes.

From now on, we will use the transformed data set. And when using the terms "restbp", "chol" and "oldpeak", we are talking about their transformed variables.

### Part 7

Model selections using AIC & BIC.

(1) Backward elimination using BIC.

```
# backward elimination using BIC
n = nrow(data3)
fit_backward_bic = step(model3, direction="backward", k=log(n))
```

```
## Start:  AIC=1796.75
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + restecg +
##     exang + oldpeak + slope + fluoro + extest
##
##             Df Sum of Sq    RSS    AIC
## - restecg    2     361.1  89533 1786.6
## - extest     2     465.1  89637 1786.9
## - chestpain  3    3025.8  92198 1789.6
## - fbs        1       7.5  89179 1791.1
```

```
## - sex        1      50.4  89222 1791.2
## - fluoro     1      84.5  89256 1791.3
## - chol       1     766.2  89938 1793.6
## - restbp     1     779.1  89951 1793.6
## - oldpeak    1    1050.8  90223 1794.5
## <none>                    89172 1796.8
## - exang      1    3859.6  93031 1803.6
## - slope      2    6435.3  95607 1806.1
## - age        1   11499.6 100671 1827.1
##
## Step:  AIC=1786.56
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + exang +
##     oldpeak + slope + fluoro + extest
##
##             Df Sum of Sq    RSS    AIC
## - extest     2     530.1  90063 1776.9
## - chestpain  3    3027.0  92560 1779.4
## - fbs        1      17.2  89550 1780.9
## - sex        1      21.6  89555 1780.9
## - fluoro     1      74.3  89607 1781.1
## - restbp     1     833.0  90366 1783.6
## - chol       1     877.8  90411 1783.8
## - oldpeak    1    1174.0  90707 1784.7
## <none>                    89533 1786.6
## - exang      1    3866.1  93399 1793.4
## - slope      2    6315.0  95848 1795.4
## - age        1   11535.5 101068 1816.9
##
## Step:  AIC=1776.93
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + exang +
##     oldpeak + slope + fluoro
##
##             Df Sum of Sq    RSS    AIC
## - chestpain  3    3252.9  93316 1770.4
## - fbs        1       5.4  90068 1771.2
## - sex        1      63.3  90126 1771.4
## - fluoro     1      88.9  90152 1771.5
## - restbp     1     768.9  90832 1773.8
## - chol       1    1027.7  91091 1774.6
## - oldpeak    1    1150.3  91213 1775.0
## <none>                    90063 1776.9
## - exang      1    3855.1  93918 1783.7
## - slope      2    7048.0  97111 1787.9
## - age        1   11649.4 101712 1807.4
##
## Step:  AIC=1770.39
## maxhr ~ age + sex + restbp + chol + fbs + exang + oldpeak + slope +
##     fluoro
##
##           Df Sum of Sq    RSS    AIC
## - sex      1      40.6  93356 1764.8
## - fbs      1      82.2  93398 1765.0
## - fluoro   1     462.9  93779 1766.2
## - chol     1     952.9  94269 1767.7
```

```
## - restbp   1     989.1  94305 1767.8
## - oldpeak  1    1307.1  94623 1768.8
## <none>                  93316 1770.4
## - slope    2    7614.1 100930 1782.3
## - exang    1    8218.9 101535 1789.8
## - age      1   11863.1 105179 1800.2
##
## Step:  AIC=1764.82
## maxhr ~ age + restbp + chol + fbs + exang + oldpeak + slope +
##     fluoro
##
##           Df Sum of Sq    RSS    AIC
## - fbs      1      77.6  93434 1759.4
## - fluoro   1     498.3  93855 1760.7
## - restbp   1    1008.2  94365 1762.3
## - chol     1    1051.3  94408 1762.5
## - oldpeak  1    1360.2  94717 1763.4
## <none>                  93356 1764.8
## - slope    2    7573.6 100930 1776.6
## - exang    1    8542.9 101899 1785.1
## - age      1   11853.3 105210 1794.6
##
## Step:  AIC=1759.37
## maxhr ~ age + restbp + chol + exang + oldpeak + slope + fluoro
##
##           Df Sum of Sq    RSS    AIC
## - fluoro   1     454.0  93888 1755.1
## - chol     1    1036.2  94470 1757.0
## - restbp   1    1113.4  94547 1757.2
## - oldpeak  1    1405.4  94839 1758.1
## <none>                  93434 1759.4
## - slope    2    7658.6 101093 1771.4
## - exang    1    8565.6 102000 1779.7
## - age      1   11782.6 105217 1789.0
##
## Step:  AIC=1755.12
## maxhr ~ age + restbp + chol + exang + oldpeak + slope
##
##           Df Sum of Sq    RSS    AIC
## - chol     1     974.5  94863 1752.5
## - restbp   1    1152.2  95040 1753.0
## - oldpeak  1    1805.1  95693 1755.1
## <none>                  93888 1755.1
## - slope    2    7651.4 101540 1767.0
## - exang    1    8873.9 102762 1776.2
## - age      1   14494.4 108382 1792.1
##
## Step:  AIC=1752.49
## maxhr ~ age + restbp + exang + oldpeak + slope
##
##           Df Sum of Sq    RSS    AIC
## - restbp   1    1372.4  96235 1751.1
## - oldpeak  1    1810.3  96673 1752.4
## <none>                  94863 1752.5
```

```
## - slope    2    7710.1 102573 1764.3
## - exang    1    8588.2 103451 1772.5
## - age      1   13663.5 108526 1786.8
##
## Step:  AIC=1751.06
## maxhr ~ age + exang + oldpeak + slope
##
##           Df Sum of Sq    RSS    AIC
## - oldpeak  1    1564.6  97800 1750.2
## <none>                  96235 1751.1
## - slope    2    8106.3 104341 1763.7
## - exang    1    8506.8 104742 1770.5
## - age      1   12332.1 108567 1781.2
##
## Step:  AIC=1750.16
## maxhr ~ age + exang + slope
##
##         Df Sum of Sq    RSS    AIC
## <none>                97800 1750.2
## - exang  1    9992.5 107792 1773.4
## - slope  2   14878.4 112678 1780.8
## - age    1   14298.9 112099 1785.0
```

```
fit_backward_bic
```

```
##
## Call:
## lm(formula = maxhr ~ age + exang + slope, data = data3)
##
## Coefficients:
## (Intercept)          age       exang1       slope2       slope3
##    204.3494      -0.7824     -12.8960     -15.4209     -10.5850
```

(2) Forward selection using BIC.

```
# forward selection using BIC
n = nrow(data3)
fit_start = lm(maxhr~1, data3)
fit_forward_bic = step(fit_start,
                       maxhr~age+sex+chestpain+restbp+chol+fbs+restecg+exang+
                           oldpeak+slope+fluoro+extest,
                       direction="forward", k=log(n))
```

```
## Start:  AIC=1865.66
## maxhr ~ 1
##
##             Df Sum of Sq    RSS    AIC
## + slope      2     32504 123285 1807.5
## + oldpeak    1     25287 130503 1818.8
## + age        1     24253 131536 1821.1
## + exang      1     23016 132773 1823.9
## + chestpain  3     23983 131807 1833.1
## + fluoro     1     11250 144539 1849.1
## + extest     2     13882 141907 1849.3
## <none>                   155789 1865.7
## + sex        1       570 155219 1870.3
```

```
## + restbp     1       335 155454 1870.7
## + restecg    2      3119 152670 1871.0
## + fbs        1        10 155780 1871.3
## + chol       1         0 155789 1871.4
##
## Step:  AIC=1807.55
## maxhr ~ slope
##
##             Df Sum of Sq    RSS    AIC
## + age        1   15492.8 107792 1773.4
## + exang      1   11186.4 112099 1785.0
## + chestpain  3   12445.0 110840 1793.0
## + fluoro     1    6130.4 117155 1798.1
## + oldpeak    1    5825.0 117460 1798.9
## <none>                   123285 1807.5
## + extest     2    3985.3 119300 1809.2
## + sex        1     498.7 122786 1812.0
## + fbs        1      30.4 123254 1813.2
## + restbp     1      28.1 123257 1813.2
## + chol       1      13.2 123272 1813.2
## + restecg    2     923.9 122361 1816.7
##
## Step:  AIC=1773.36
## maxhr ~ slope + age
##
##             Df Sum of Sq    RSS    AIC
## + exang      1    9992.5  97800 1750.2
## + chestpain  3    9885.7  97906 1761.9
## + oldpeak    1    3050.4 104742 1770.5
## <none>                   107792 1773.4
## + fluoro     1    1433.0 106359 1775.1
## + sex        1    1185.5 106607 1775.8
## + restbp     1     957.9 106834 1776.4
## + extest     2    2883.0 104909 1776.7
## + chol       1     823.4 106969 1776.8
## + fbs        1     137.0 107655 1778.7
## + restecg    2     607.0 107185 1783.1
##
## Step:  AIC=1750.16
## maxhr ~ slope + age + exang
##
##             Df Sum of Sq   RSS    AIC
## <none>                   97800 1750.2
## + oldpeak    1    1564.6 96235 1751.1
## + chol       1    1178.2 96621 1752.2
## + restbp     1    1126.7 96673 1752.4
## + fluoro     1     779.4 97020 1753.5
## + chestpain  3    4175.1 93624 1754.3
## + sex        1     381.5 97418 1754.7
## + fbs        1     121.7 97678 1755.5
## + extest     2    1161.7 96638 1758.0
## + restecg    2     667.9 97132 1759.5
```

```
fit_forward_bic
```

```
##
## Call:
## lm(formula = maxhr ~ slope + age + exang, data = data3)
##
## Coefficients:
## (Intercept)       slope2       slope3          age       exang1
##    204.3494     -15.4209     -10.5850      -0.7824     -12.8960
```

(3) Backward elimination using AIC.

```
# backward elimination using AIC
fit_backward_aic = step(model3, direction="backward")
```

```
## Start:  AIC=1730.26
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + restecg +
##     exang + oldpeak + slope + fluoro + extest
##
##             Df Sum of Sq    RSS    AIC
## - restecg    2     361.1  89533 1727.5
## - extest     2     465.1  89637 1727.8
## - fbs        1       7.5  89179 1728.3
## - sex        1      50.4  89222 1728.4
## - fluoro     1      84.5  89256 1728.5
## <none>                    89172 1730.3
## - chol       1     766.2  89938 1730.8
## - restbp     1     779.1  89951 1730.8
## - oldpeak    1    1050.8  90223 1731.7
## - chestpain  3    3025.8  92198 1734.2
## - exang      1    3859.6  93031 1740.8
## - slope      2    6435.3  95607 1747.0
## - age        1   11499.6 100671 1764.3
##
## Step:  AIC=1727.46
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + exang +
##     oldpeak + slope + fluoro + extest
##
##             Df Sum of Sq    RSS    AIC
## - extest     2     530.1  90063 1725.2
## - fbs        1      17.2  89550 1725.5
## - sex        1      21.6  89555 1725.5
## - fluoro     1      74.3  89607 1725.7
## <none>                    89533 1727.5
## - restbp     1     833.0  90366 1728.2
## - chol       1     877.8  90411 1728.4
## - oldpeak    1    1174.0  90707 1729.3
## - chestpain  3    3027.0  92560 1731.3
## - exang      1    3866.1  93399 1738.0
## - slope      2    6315.0  95848 1743.7
## - age        1   11535.5 101068 1761.5
##
## Step:  AIC=1725.22
## maxhr ~ age + sex + chestpain + restbp + chol + fbs + exang +
##     oldpeak + slope + fluoro
```

```
##
##             Df Sum of Sq    RSS    AIC
## - fbs        1       5.4  90068 1723.2
## - sex        1      63.3  90126 1723.4
## - fluoro     1      88.9  90152 1723.5
## <none>                    90063 1725.2
## - restbp     1     768.9  90832 1725.7
## - chol       1    1027.7  91091 1726.6
## - oldpeak    1    1150.3  91213 1727.0
## - chestpain  3    3252.9  93316 1729.8
## - exang      1    3855.1  93918 1735.7
## - slope      2    7048.0  97111 1743.6
## - age        1   11649.4 101712 1759.3
##
## Step:  AIC=1723.23
## maxhr ~ age + sex + chestpain + restbp + chol + exang + oldpeak +
##     slope + fluoro
##
##             Df Sum of Sq    RSS    AIC
## - sex        1      61.8  90130 1721.4
## - fluoro     1      84.2  90153 1721.5
## <none>                    90068 1723.2
## - restbp     1     803.3  90872 1723.9
## - chol       1    1025.8  91094 1724.6
## - oldpeak    1    1167.1  91235 1725.1
## - chestpain  3    3329.7  93398 1728.0
## - exang      1    3849.8  93918 1733.7
## - slope      2    7071.2  97140 1741.7
## - age        1   11659.2 101728 1757.4
##
## Step:  AIC=1721.44
## maxhr ~ age + chestpain + restbp + chol + exang + oldpeak + slope +
##     fluoro
##
##             Df Sum of Sq    RSS    AIC
## - fluoro     1     102.9  90233 1719.8
## <none>                    90130 1721.4
## - restbp     1     829.3  90959 1722.2
## - chol       1    1143.9  91274 1723.2
## - oldpeak    1    1214.8  91345 1723.4
## - chestpain  3    3303.9  93434 1726.1
## - exang      1    4035.4  94166 1732.5
## - slope      2    7011.3  97142 1739.7
## - age        1   11612.6 101743 1755.4
##
## Step:  AIC=1719.78
## maxhr ~ age + chestpain + restbp + chol + exang + oldpeak + slope
##
##             Df Sum of Sq    RSS    AIC
## <none>                    90233 1719.8
## - restbp     1     837.7  91071 1720.5
## - chol       1    1118.2  91351 1721.4
## - oldpeak    1    1396.1  91629 1722.3
## - chestpain  3    3655.0  93888 1725.6
```

```
## - exang      1    4019.5  94253 1730.7
## - slope      2    6987.3  97220 1737.9
## - age        1   13405.0 103638 1758.9
```

```
fit_backward_aic
```

```
##
## Call:
## lm(formula = maxhr ~ age + chestpain + restbp + chol + exang +
##     oldpeak + slope, data = data3)
##
## Coefficients:
## (Intercept)          age   chestpain2   chestpain3   chestpain4
##     92.9148      -0.8116      -3.4779      -4.3887     -11.0394
##      restbp         chol       exang1      oldpeak       slope2
##     13.6494       9.8088      -9.0496      -4.3561     -11.9140
##      slope3
##     -6.0219
```

(4) Forward selection using AIC.

```
# forward selection using AIC
fit_start = lm(maxhr~1, data3)
fit_forward_aic = step(fit_start,
                       maxhr~age+sex+chestpain+restbp+chol+fbs+restecg+exang+
                           oldpeak+slope+fluoro+extest,
                       direction="forward")
```

```
## Start:  AIC=1861.97
## maxhr ~ 1
##
##             Df Sum of Sq    RSS    AIC
## + slope      2     32504 123285 1796.5
## + oldpeak    1     25287 130503 1811.4
## + age        1     24253 131536 1813.7
## + exang      1     23016 132773 1816.5
## + chestpain  3     23983 131807 1818.3
## + extest     2     13882 141907 1838.2
## + fluoro     1     11250 144539 1841.7
## + restecg    2      3119 152670 1860.0
## <none>                   155789 1862.0
## + sex        1       570 155219 1862.9
## + restbp     1       335 155454 1863.3
## + fbs        1        10 155780 1864.0
## + chol       1         0 155789 1864.0
##
## Step:  AIC=1796.47
## maxhr ~ slope
##
##             Df Sum of Sq    RSS    AIC
## + age        1   15492.8 107792 1758.6
## + exang      1   11186.4 112099 1770.2
## + chestpain  3   12445.0 110840 1770.9
## + fluoro     1    6130.4 117155 1783.3
## + oldpeak    1    5825.0 117460 1784.1
## + extest     2    3985.3 119300 1790.7
```

```
## <none>                   123285 1796.5
## + sex        1     498.7 122786 1797.3
## + restecg    2     923.9 122361 1798.2
## + fbs        1      30.4 123254 1798.4
## + restbp     1      28.1 123257 1798.4
## + chol       1      13.2 123272 1798.4
##
## Step:  AIC=1758.59
## maxhr ~ slope + age
##
##             Df Sum of Sq    RSS    AIC
## + exang      1    9992.5  97800 1731.7
## + chestpain  3    9885.7  97906 1736.0
## + oldpeak    1    3050.4 104742 1752.1
## + extest     2    2883.0 104909 1754.5
## + fluoro     1    1433.0 106359 1756.6
## + sex        1    1185.5 106607 1757.3
## + restbp     1     957.9 106834 1757.9
## + chol       1     823.4 106969 1758.3
## <none>                   107792 1758.6
## + fbs        1     137.0 107655 1760.2
## + restecg    2     607.0 107185 1760.9
##
## Step:  AIC=1731.69
## maxhr ~ slope + age + exang
##
##             Df Sum of Sq   RSS    AIC
## + chestpain  3    4175.1 93624 1724.7
## + oldpeak    1    1564.6 96235 1728.9
## + chol       1    1178.2 96621 1730.1
## + restbp     1    1126.7 96673 1730.2
## + fluoro     1     779.4 97020 1731.3
## <none>                   97800 1731.7
## + extest     2    1161.7 96638 1732.1
## + sex        1     381.5 97418 1732.5
## + fbs        1     121.7 97678 1733.3
## + restecg    2     667.9 97132 1733.7
##
## Step:  AIC=1724.73
## maxhr ~ slope + age + exang + chestpain
##
##           Df Sum of Sq   RSS    AIC
## + chol     1   1299.15 92325 1722.6
## + oldpeak  1   1218.70 92406 1722.8
## + restbp   1    894.11 92730 1723.9
## <none>                 93624 1724.7
## + sex      1    383.63 93241 1725.5
## + fluoro   1    236.77 93388 1726.0
## + extest   2    730.60 92894 1726.4
## + restecg  2    669.07 92955 1726.6
## + fbs      1     25.35 93599 1726.7
##
## Step:  AIC=1722.58
## maxhr ~ slope + age + exang + chestpain + chol
```

```
##
##           Df Sum of Sq   RSS    AIC
## + oldpeak  1   1254.53 91071 1720.5
## + restbp   1    696.12 91629 1722.3
## <none>                 92325 1722.6
## + fluoro   1    285.03 92040 1723.7
## + sex      1    190.31 92135 1724.0
## + fbs      1     26.09 92299 1724.5
## + extest   2    531.32 91794 1724.9
## + restecg  2    521.60 91804 1724.9
##
## Step:  AIC=1720.52
## maxhr ~ slope + age + exang + chestpain + chol + oldpeak
##
##           Df Sum of Sq   RSS    AIC
## + restbp   1    837.68 90233 1719.8
## <none>                 91071 1720.5
## + fluoro   1    111.33 90959 1722.2
## + sex      1    110.71 90960 1722.2
## + fbs      1     17.40 91053 1722.5
## + extest   2    499.44 90571 1722.9
## + restecg  2    406.32 90664 1723.2
##
## Step:  AIC=1719.78
## maxhr ~ slope + age + exang + chestpain + chol + oldpeak + restbp
##
##           Df Sum of Sq   RSS    AIC
## <none>                 90233 1719.8
## + fluoro   1    102.94 90130 1721.4
## + sex      1     80.49 90153 1721.5
## + fbs      1      0.07 90233 1721.8
## + extest   2    591.63 89641 1721.8
## + restecg  2    366.63 89866 1722.6
```

```
fit_forward_aic
```

```
##
## Call:
## lm(formula = maxhr ~ slope + age + exang + chestpain + chol +
##     oldpeak + restbp, data = data3)
##
## Coefficients:
## (Intercept)       slope2       slope3          age       exang1
##     92.9148     -11.9140      -6.0219      -0.8116      -9.0496
##  chestpain2   chestpain3   chestpain4         chol      oldpeak
##     -3.4779      -4.3887     -11.0394       9.8088      -4.3561
##      restbp
##     13.6494
```

Discussions:

AIC and BIC are both penalized-likelihood criteria. The AIC or BIC for a model is usually written in the form $-2log(L) + k \times p$, where L is the likelihood function, p is the number of parameters in the model, and k is 2 for AIC and log(n) for BIC. So the step() function in R use k=2 as default to compute AIC values, and it computes BIC values when we set k = log(n), where n is the number of data points.

Thus, BIC penalizes model complexity more heavily, so BIC has a larger chance than AIC, for any given n, of choosing too small a model. On the other hand, AIC always has a chance of choosing too big a model, regardless of n.

Though these methods may have multiple testing issues for large number of covariates, here the number of covariates is 12, which is much less than the number of data points 297. So we are not concerning about multiple tesing issues here.

Here, both the BIC forward selection and backward elimination methods get the same final model. In the backward elimination, it removes variables in the order of "restecg" → "extest" → "chestpain" → "sex" → "fbs" → "fluoro" → "chol" → "restbp" → "oldpeak". In the forward selection, it adds variables in the order of "slope" → "age" → "exang". Thus, we can write the final model selected by BIC in the following order: $maxhr \sim slope + age + exang$

Similarly, both the AIC forward selection and backward elimination methods get the same final model. In the backward elimination, it removes variables in the order of "restecg" → "extest" → "fbs" → "sex" → "fluoro". In the forward selection, it adds variables in the order of "slope" → "age" → "exang" → "chestpain" → "chol" → "oldpeak" → "restbp". Thus, we can write the final model selected by AIC in the following order: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + restbp$

Now we compare the two models:

```
reduced = lm(maxhr ~ slope+age+exang, data3)
larger = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak+restbp, data3)
anova(reduced, larger)
```

```
## Analysis of Variance Table
##
## Model 1: maxhr ~ slope + age + exang
## Model 2: maxhr ~ slope + age + exang + chestpain + chol + oldpeak + restbp
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    292 97800
## 2    286 90233  6    7566.5 3.9971 0.0007355 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of F test is $0.0007355 < 0.001$, so we Reject Null Hypothesis at $\alpha = 0.1\%$ significance level. Thus, so far, we prefer the larger model selected by AIC here:

$$maxhr \sim slope + age + exang + chestpain + chol + oldpeak + restbp$$

### Part 8

Variable selections using Lasso regularization.

```
# standardization (z-score normalization) of quantitative covariates
n = nrow(data3)
data3_quant = data3[c(1,4,5,9,11)]
data3_std = scale(data3_quant, center = colMeans(data3_quant), scale=FALSE)  # or center=TRUE
data3_std = scale(data3_std, center=FALSE,
                  scale = sqrt(colSums(data3_std^2)/n) ) # if use scale=TRUE, it's dividing by (n-1)

data3_categ = data3[c(2,3,6,7,8,10,12)]
data3_std = cbind(data.frame(data3_std), data3_categ, data3$maxhr)
colnames(data3_std)[13] = "maxhr"

# the new colnames are in the following sequences
colnames(data3_std)
```

```
##  [1] "age"       "restbp"    "chol"      "oldpeak"   "fluoro"
##  [6] "sex"       "chestpain" "fbs"       "restecg"   "exang"
## [11] "slope"     "extest"    "maxhr"
```

```
# Check the standardization results
colMeans(data3_std[c(1:5)])
```

```
##           age        restbp          chol       oldpeak        fluoro
## -1.240123e-16 -2.950689e-15 -1.158036e-15 -9.546236e-17 -8.859355e-17
```

```
colSums(data3_std[c(1:5)]^2)
```

```
##     age  restbp    chol oldpeak  fluoro
##     297     297     297     297     297
```

Here we first standardized the quantitative covariates, so that now each quantitative covariate have zero mean and $\sum_i X_{ij}^2 = n$.

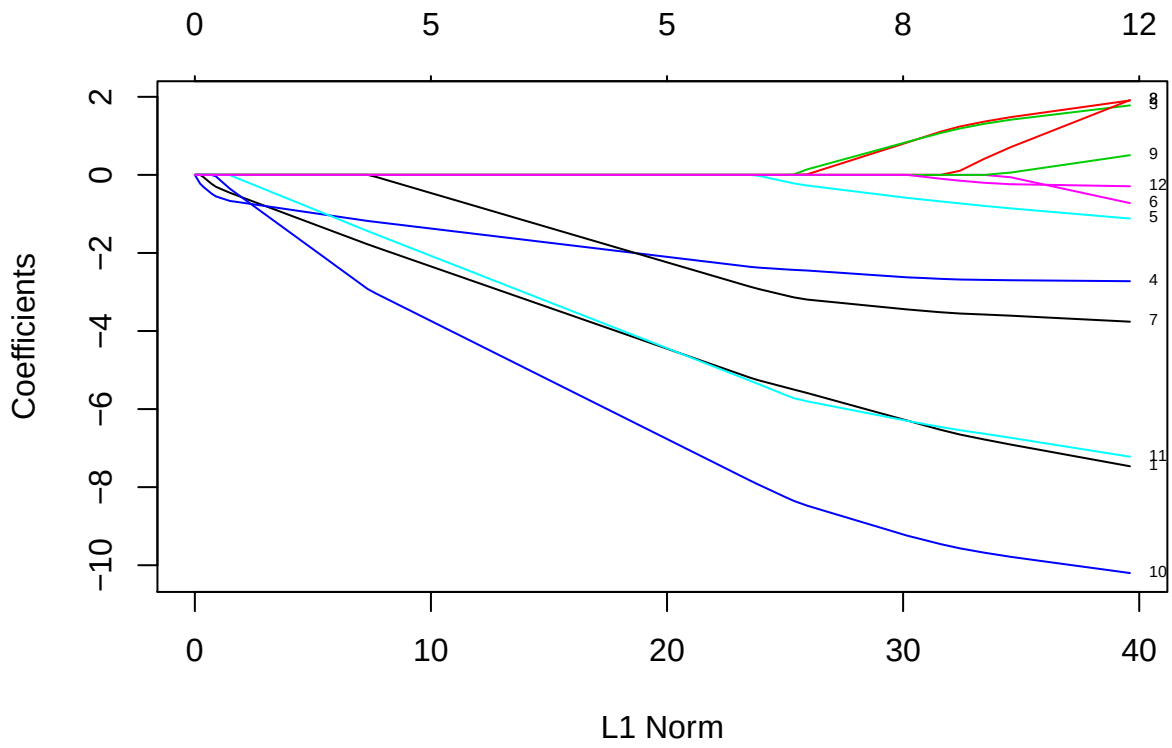
```
# lasso regression
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
model_lasso = glmnet(x = as.matrix(data3_std[-13]), y = as.matrix(data3_std[13]),
                     lambda = seq(0,10,by=0.1))
plot.glmnet(model_lasso, xvar="norm", label=TRUE)
```

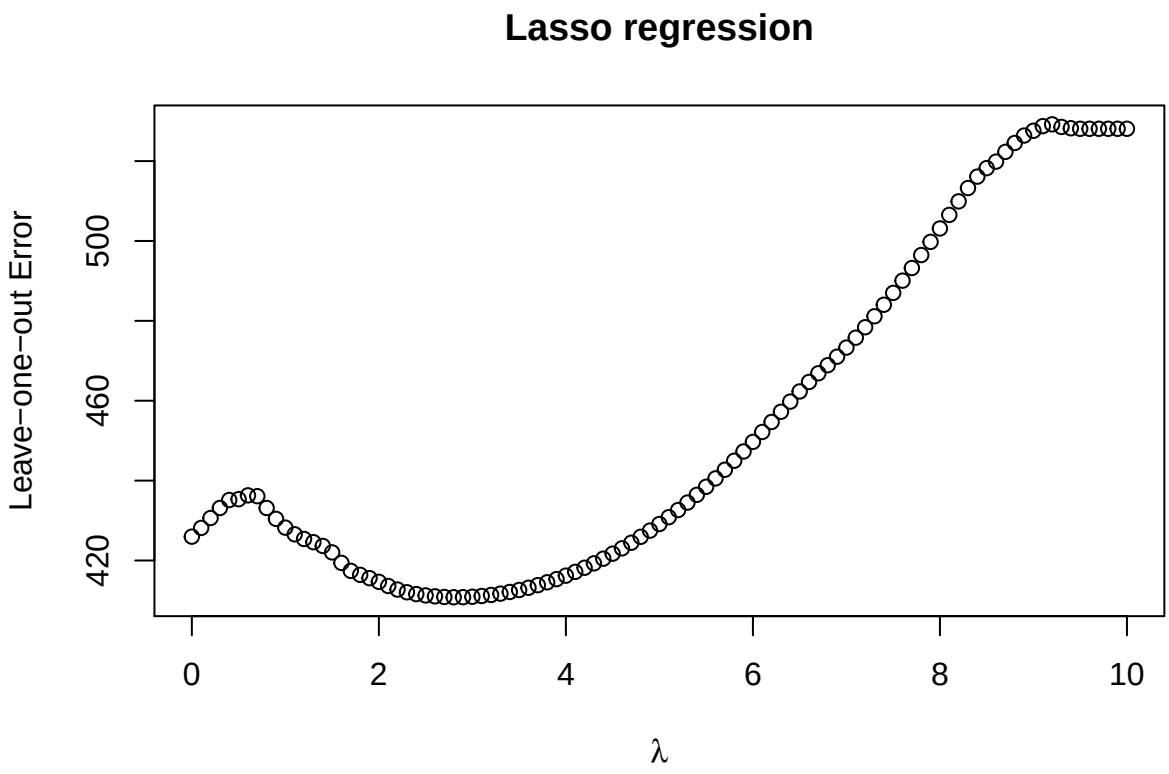


```
# plot.glmnet(model_lasso, xvar="lambda", label=TRUE)
```

The lasso regularization can reach sparsity, thus it can force the coefficients of covariates to zero values by increasing the $\lambda$ value. Though it has shrinkage bias of coefficients, we can still use it as an effective tool to select covariates. Here the lasso regularization has already removed all the covariates before $\lambda = 10$.

Next, we need to choose the best $\lambda$ value by leave-one-out cross validations.

```
# leave-one-out cross-validation of lasso regression
n = nrow(data3_std)
lambdas = seq(0,10,by=0.1)
ave_square_error = rep(0,length(lambdas))

for (k in 1:length(lambdas)){
    store_pred_error = rep(0,n)
    for (i in 1:n) {
        model_lasso = glmnet(x = as.matrix(data3_std[-i,-13]), y = as.matrix(data3_std[-i,13]),
                             lambda=lambdas[k])
        xi = data3_std[i,-13]
        betahat = rbind(model_lasso$a0, as.matrix(model_lasso$beta))
        fitted_yi = unlist(c(1,xi)) %*% as.vector(betahat)
        store_pred_error[i] = data3_std[i,13] - fitted_yi
    }
    ave_square_error[k] = sum(store_pred_error^2)/n
}

# plot leave-one-out error against lambda
plot(lambdas, ave_square_error, main="Lasso regression",
     xlab=expression(lambda), ylab="Leave-one-out Error")
```




```
# find the best lambda value
best_lambda_lasso = lambdas[which.min(ave_square_error)]
print(paste0("best lambda = ",best_lambda_lasso))
```

```
## [1] "best lambda = 2.8"
```

```
print(paste0("minimum average of prediction error = ",min(ave_square_error)))
```

```
## [1] "minimum average of prediction error = 410.799112041463"
```

According to the plots, as $\lambda$ increases from 0 to 2.8, the average of squared leave-one-out prediction error first increases a little and then decreases to a minimal value. Then, as $\lambda$ increases from 2.8 to 10, the average of squared leave-one-out prediction error substantially increases again and reaches a plateau near 10. Therefore, lasso regularization dose offer substantial improvement of the prediction error by removing covariates. And here it reduces the prediction error to the most extent at $\lambda = 2.8$.

Next, we use this best $\lambda$ value (2.8) to fit the lasso regression model and get the coefficients of each covariates.

```
# fit the lasso regression model with the best lambda value and get the coefficients
model_lasso = glmnet(x = as.matrix(data3_std[-13]), y = as.matrix(data3_std[13]),
                     lambda = best_lambda_lasso)
betahat = rbind(model_lasso$a0, as.matrix(model_lasso$beta))
colnames(betahat) = model_lasso$lambda
rownames(betahat)[1] = "(Intercept)"
betahat
```

```
##                    2.8
## (Intercept) 167.088374
## age          -4.673941
```

```
## restbp        0.000000
## chol          0.000000
## oldpeak      -2.175929
## fluoro        0.000000
## sex           0.000000
## chestpain    -2.426029
## fbs           0.000000
## restecg       0.000000
## exang        -7.082897
## slope        -4.688217
## extest        0.000000
```

So the variables selected by lasso regularization are "age", "oldpeak", "chestpain", "exang", and "slope". Thus, we can write the selected model as:

$$maxhr \sim slope + age + exang + chestpain + oldpeak$$

**Part 9**

Compare the two selected models from AIC and Lasso regularization:

AIC: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + restbp$

Lasso: $maxhr \sim slope + age + exang + chestpain + oldpeak$

```
reduced = lm(maxhr ~ slope+age+exang+chestpain+oldpeak, data3)
larger  = lm(maxhr ~ slope+age+exang+chestpain+oldpeak +chol+restbp, data3)
anova(reduced, larger)
```

```
## Analysis of Variance Table
##
## Model 1: maxhr ~ slope + age + exang + chestpain + oldpeak
## Model 2: maxhr ~ slope + age + exang + chestpain + oldpeak + chol + restbp
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    288 92406
## 2    286 90233  2    2172.7 3.4432 0.03329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(maxhr ~ slope+age+exang+chestpain+oldpeak +chol+restbp, data3))
```

```
## Analysis of Variance Table
##
## Response: maxhr
##            Df Sum Sq Mean Sq F value    Pr(>F)
## slope       2  32504 16252.2 51.5125 < 2.2e-16 ***
## age         1  15493 15492.8 49.1055 1.743e-11 ***
## exang       1   9993  9992.5 31.6719 4.352e-08 ***
## chestpain   3   4175  1391.7  4.4111  0.004727 **
## oldpeak     1   1219  1218.7  3.8627  0.050337 .
## chol        1   1335  1335.0  4.2313  0.040591 *
## restbp      1    838   837.7  2.6551  0.104321
## Residuals 286  90233   315.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(maxhr ~ slope+age+exang+chestpain+oldpeak +restbp+chol, data3))
```

```
## Analysis of Variance Table
##
## Response: maxhr
##            Df Sum Sq Mean Sq F value    Pr(>F)
## slope       2  32504 16252.2 51.5125 < 2.2e-16 ***
## age         1  15493 15492.8 49.1055 1.743e-11 ***
## exang       1   9993  9992.5 31.6719 4.352e-08 ***
## chestpain   3   4175  1391.7  4.4111  0.004727 **
## oldpeak     1   1219  1218.7  3.8627  0.050337 .
## restbp      1   1054  1054.5  3.3422  0.068567 .
## chol        1   1118  1118.2  3.5442  0.060768 .
## Residuals 286  90233   315.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of F test is $0.03329 < 0.05$, so we Reject Null Hypothesis at $\alpha = 5\%$ significance level. Thus, we prefer the larger model here selected by AIC: $maxhr \sim slope+age+exang+chestpain+chol+oldpeak+restbp$

Then when we use anova() function to see if we can remove any one of "chol" and "restbp" with different orders, we find that we can remove "restbp" while keeping "chol" in the model, however, we cannot remove "chol" while keeping "restbp" in the model.

Therefore, so far, we will prefer the model between the sizes of two selected models:

$$maxhr \sim slope + age + exang + chestpain + chol + oldpeak$$

Then next, we can consider the interaction terms and see if we can still remove the following covariates from the model in the order of:

$$restecg \rightarrow extest \rightarrow fbs \rightarrow sex \rightarrow fluoro \rightarrow restbp$$

Note that since lasso may have shrinkage bias of coefficients, the above order is the one indicated by AIC, not by lasso.

Let's have a look at the summary of the current smaller selected model.

```
# reoder the sequence and have a look at the current smaller selected model
model3 = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak, data3)
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##            Df Sum Sq Mean Sq F value    Pr(>F)
## slope       2  32504 16252.2 51.2171 < 2.2e-16 ***
## age         1  15493 15492.8 48.8239 1.959e-11 ***
## exang       1   9993  9992.5 31.4903 4.721e-08 ***
## chestpain   3   4175  1391.7  4.3858  0.004887 **
## chol        1   1299  1299.2  4.0941  0.043959 *
## oldpeak     1   1255  1254.5  3.9535  0.047723 *
## Residuals 287  91071   317.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the sequence of variables and have a look at the full model without interaction terms
model3 = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg, data3)
anova(model3)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##            Df Sum Sq Mean Sq F value    Pr(>F)
## slope       2  32504 16252.2 50.8498 < 2.2e-16 ***
## age         1  15493 15492.8 48.4738 2.398e-11 ***
## exang       1   9993  9992.5 31.2645 5.363e-08 ***
## chestpain   3   4175  1391.7  4.3544  0.005115 **
## chol        1   1299  1299.2  4.0648  0.044745 *
## oldpeak     1   1255  1254.5  3.9252  0.048550 *
## restbp      1    838   837.7  2.6209  0.106593
## fluoro      1    103   102.9  0.3221  0.570809
## sex         1     62    61.8  0.1933  0.660556
## fbs         1      5     5.4  0.0169  0.896714
## extest      2    530   265.1  0.8293  0.437413
## restecg     2    361   180.6  0.5650  0.569029
## Residuals 279  89172   319.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

The summary shows that in the smaller selected model, all the selected variabbles are significant predictors. And in the full model without considering interaction terms, it's also true that only the covariates from the selected smaller model are significant predictors.

### Part 10

Consider interaction terms.

Since there are many covariates, and this is not like a chemical reaction problem, thus here we can only consider the two-way interactions.

Also, since there are many covariates including categorical covariates, and many categorical covariates have several levels, there will be multiple testing issues when we are considering so many interaction terms. Some interaction terms will become significant while they are actually false positive. Therefore, first, we will try to break the problem into different steps. Second, we will try to reduce the size of the models step by step.

(1) The smaller seleceted model with all two-way interaction terms.

```
# include all two-way interactions within the smaller model
model_small_inter = lm(maxhr ~ (slope+age+exang+chestpain+chol+oldpeak)**2, data3)
anova(model_small_inter)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## slope               2  32504 16252.2 52.3537 < 2.2e-16 ***
## age                 1  15493 15492.8 49.9074 1.533e-11 ***
## exang               1   9993  9992.5 32.1891 3.788e-08 ***
## chestpain           3   4175  1391.7  4.4831  0.004359 **
## chol                1   1299  1299.2  4.1850  0.041809 *
## oldpeak             1   1255  1254.5  4.0413  0.045454 *
## slope:age           2   1781   890.7  2.8692  0.058579 .
## slope:exang         2    892   445.9  1.4364  0.239689
## slope:chestpain     6   4545   757.5  2.4402  0.025991 *
## slope:chol          2    575   287.3  0.9256  0.397607
```

```
## slope:oldpeak       2     50    25.2  0.0813  0.921918
## age:exang           1    328   327.7  1.0556  0.305186
## age:chestpain       3   1443   481.0  1.5495  0.202213
## age:chol            1     72    71.7  0.2308  0.631322
## age:oldpeak         1     14    14.4  0.0464  0.829666
## exang:chestpain     3    198    65.9  0.2124  0.887745
## exang:chol          1    184   184.1  0.5929  0.442009
## exang:oldpeak       1      4     4.4  0.0141  0.905731
## chestpain:chol      3    754   251.4  0.8099  0.489405
## chestpain:oldpeak   3    798   266.0  0.8568  0.464093
## chol:oldpeak        1    272   272.3  0.8771  0.349895
## Residuals         255  79160   310.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# only add slope:chestpain, slope:age
model_small_inter = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
                       +slope:chestpain +slope:age, data3)
anova(model_small_inter)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 53.6192 < 2.2e-16 ***
## age                1  15493 15492.8 51.1138 7.655e-12 ***
## exang              1   9993  9992.5 32.9672 2.444e-08 ***
## chestpain          3   4175  1391.7  4.5915  0.003724 **
## chol               1   1299  1299.2  4.2862  0.039344 *
## oldpeak            1   1255  1254.5  4.1389  0.042853 *
## slope:chestpain    6   6058  1009.7  3.3314  0.003479 **
## slope:age          2    446   223.1  0.7361  0.479891
## Residuals        279  84566   303.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# only add
model_small_inter = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
                       +slope:chestpain, data3)
anova(model_small_inter)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 53.7201 < 2.2e-16 ***
## age                1  15493 15492.8 51.2099 7.243e-12 ***
## exang              1   9993  9992.5 33.0293 2.360e-08 ***
## chestpain          3   4175  1391.7  4.6002  0.003678 **
## chol               1   1299  1299.2  4.2942  0.039154 *
## oldpeak            1   1255  1254.5  4.1467  0.042653 *
## slope:chestpain    6   6058  1009.7  3.3376  0.003424 **
## Residuals        281  85012   302.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The smaller model with all the two-way interaction terms have only two potential significant interaction terms: "slope:age" and "slope:chestpain". When only adding these two interactions in the decreased order of their significance level (adding the more significant interaction first), the anova test shows that "slope:age" can be removed from the model. Therefore, for the selected smaller model, we can add one interaction term within themselves so far:

$$maxhr \sim slope + age + exang + chestpain + chol + oldpeak + slope : chestpain$$

Next, we will have a look at two-way interaction terms of covariates in the order of:

$$restecg \rightarrow extest \rightarrow fbs \rightarrow sex \rightarrow fluoro \rightarrow restbp$$

(2) Interaction terms of "restecg".

```
# interaction terms of "restecg"
model3_restecg = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:age+restecg:exang+restecg:chestpain
            +restecg:chol+restecg:oldpeak+restecg:restbp+restecg:fluoro
            +restecg:sex+restecg:fbs+restecg:extest, data3)
anova(model3_restecg)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## slope                2  32504 16252.2 54.6148 < 2.2e-16 ***
## age                  1  15493 15492.8 52.0628 6.139e-12 ***
## exang                1   9993  9992.5 33.5793 2.011e-08 ***
## chestpain            3   4175  1391.7  4.6768  0.003368 **
## chol                 1   1299  1299.2  4.3657  0.037660 *
## oldpeak              1   1255  1254.5  4.2158  0.041070 *
## restbp               1    838   837.7  2.8150  0.094614 .
## fluoro               1    103   102.9  0.3459  0.556941
## sex                  1     62    61.8  0.2076  0.649066
## fbs                  1      5     5.4  0.0181  0.892990
## extest               2    530   265.1  0.8907  0.411626
## restecg              2    361   180.6  0.6068  0.545880
## slope:chestpain      6   6689  1114.9  3.7465  0.001366 **
## slope:restecg        3   2073   691.0  2.3220  0.075627 .
## age:restecg          2     94    46.8  0.1573  0.854571
## exang:restecg        1   1824  1824.3  6.1304  0.013938 *
## chestpain:restecg    3    231    77.0  0.2589  0.854946
## chol:restecg         2    945   472.3  1.5873  0.206496
## oldpeak:restecg      1     24    24.4  0.0818  0.775058
## restbp:restecg       1     11    10.9  0.0368  0.848040
## fluoro:restecg       1      3     3.3  0.0110  0.916615
## sex:restecg          1    103   102.8  0.3454  0.557266
## fbs:restecg          1   1015  1015.1  3.4112  0.065913 .
## extest:restecg       2    277   138.4  0.4649  0.628716
## Residuals          255  75883   297.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# reoder the interaction terms according to their significance level
model3_restecg = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:exang+restecg:slope+restecg:fbs
            +restecg:age+restecg:chestpain+restecg:chol
            +restecg:oldpeak+restecg:restbp+restecg:fluoro
            +restecg:sex+restecg:extest, data3)
anova(model3_restecg)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## slope                2  32504 16252.2 54.6148 < 2.2e-16 ***
## age                  1  15493 15492.8 52.0628 6.139e-12 ***
## exang                1   9993  9992.5 33.5793 2.011e-08 ***
## chestpain            3   4175  1391.7  4.6768  0.003368 **
## chol                 1   1299  1299.2  4.3657  0.037660 *
## oldpeak              1   1255  1254.5  4.2158  0.041070 *
## restbp               1    838   837.7  2.8150  0.094614 .
## fluoro               1    103   102.9  0.3459  0.556941
## sex                  1     62    61.8  0.2076  0.649066
## fbs                  1      5     5.4  0.0181  0.892990
## extest               2    530   265.1  0.8907  0.411626
## restecg              2    361   180.6  0.6068  0.545880
## slope:chestpain      6   6689  1114.9  3.7465  0.001366 **
## exang:restecg        2    970   485.2  1.6306  0.197853
## slope:restecg        3   3013  1004.3  3.3750  0.019008 *
## fbs:restecg          1    907   907.1  3.0482  0.082030 .
## age:restecg          1      1     1.4  0.0047  0.945118
## chestpain:restecg    3    276    91.8  0.3086  0.819140
## chol:restecg         2    948   473.9  1.5926  0.205408
## oldpeak:restecg      1     22    22.4  0.0752  0.784190
## restbp:restecg       1      1     0.7  0.0023  0.962157
## fluoro:restecg       1     47    46.5  0.1563  0.692913
## sex:restecg          1    138   138.1  0.4641  0.496311
## extest:restecg       2    277   138.4  0.4649  0.628716
## Residuals          255  75883   297.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# only add "restecg:slope", "restecg:fbs", "restecg:exang"
model3_restecg = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang, data3)
anova(model3_restecg)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 55.9252 < 2.2e-16 ***
## age                1  15493 15492.8 53.3120 3.274e-12 ***
## exang              1   9993  9992.5 34.3850 1.332e-08 ***
## chestpain          3   4175  1391.7  4.7890  0.002878 **
```

```
## chol              1   1299  1299.2  4.4705  0.035412 *
## oldpeak           1   1255  1254.5  4.3169  0.038690 *
## restbp            1    838   837.7  2.8825  0.090711 .
## fluoro            1    103   102.9  0.3542  0.552227
## sex               1     62    61.8  0.2126  0.645150
## fbs               1      5     5.4  0.0186  0.891717
## extest            2    530   265.1  0.9121  0.402924
## restecg           2    361   180.6  0.6214  0.537991
## slope:chestpain   6   6689  1114.9  3.8364  0.001094 **
## slope:restecg     3   2073   691.0  2.3777  0.070243 .
## fbs:restecg       1    831   831.3  2.8607  0.091933 .
## exang:restecg     2   1986   993.2  3.4175  0.034235 *
## Residuals       267  77592   290.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, so far, we consider not removing the variable "restecg" from the model, and add three of its interaction terms to the model: "restecg:slope", "restecg:fbs", "restecg:exang".

(4) Interaction terms of "extest".

```
# interaction terms of "extest"
model3_extest = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang
            +extest:slope+extest:age+extest:exang+extest:chestpain
            +extest:chol+extest:oldpeak+extest:restbp+extest:fluoro
            +extest:sex+extest:fbs+extest:restecg, data3)
anova(model3_extest)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 56.7784 < 2.2e-16 ***
## age                1  15493 15492.8 54.1254 3.031e-12 ***
## exang              1   9993  9992.5 34.9096 1.187e-08 ***
## chestpain          3   4175  1391.7  4.8620 0.0026624 **
## chol               1   1299  1299.2  4.5387 0.0341627 *
## oldpeak            1   1255  1254.5  4.3828 0.0373619 *
## restbp             1    838   837.7  2.9265 0.0884397 .
## fluoro             1    103   102.9  0.3596 0.5492740
## sex                1     62    61.8  0.2158 0.6426882
## fbs                1      5     5.4  0.0189 0.8909119
## extest             2    530   265.1  0.9260 0.3975478
## restecg            2    361   180.6  0.6308 0.5330351
## slope:chestpain    6   6689  1114.9  3.8950 0.0009917 ***
## slope:restecg      3   2073   691.0  2.4139 0.0673047 .
## fbs:restecg        1    831   831.3  2.9044 0.0896446 .
## exang:restecg      2   1986   993.2  3.4697 0.0327108 *
## slope:extest       4   1077   269.2  0.9404 0.4411956
## age:extest         2   1145   572.5  2.0000 0.1376021
## exang:extest       2    259   129.3  0.4516 0.6371838
## chestpain:extest   6   1884   313.9  1.0968 0.3649104
## chol:extest        2     26    13.0  0.0454 0.9556026
## oldpeak:extest     2    124    62.2  0.2174 0.8047592
```

```
## restbp:extest      2    463   231.6  0.8092 0.4464486
## fluoro:extest      2    238   119.0  0.4156 0.6604463
## sex:extest         2   2262  1131.2  3.9519 0.0204919 *
## fbs:extest         2   1106   552.9  1.9316 0.1471844
## extest:restecg     3    883   294.4  1.0286 0.3805940
## Residuals        238  68125   286.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms according to their significance level
model3_extest = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang
            +extest:sex+extest:slope+extest:age+extest:exang
            +extest:chestpain+extest:chol+extest:oldpeak+extest:restbp
            +extest:fluoro+extest:fbs+extest:restecg, data3)
anova(model3_extest)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## slope               2  32504 16252.2 56.7784 < 2.2e-16 ***
## age                 1  15493 15492.8 54.1254 3.031e-12 ***
## exang               1   9993  9992.5 34.9096 1.187e-08 ***
## chestpain           3   4175  1391.7  4.8620 0.0026624 **
## chol                1   1299  1299.2  4.5387 0.0341627 *
## oldpeak             1   1255  1254.5  4.3828 0.0373619 *
## restbp              1    838   837.7  2.9265 0.0884397 .
## fluoro              1    103   102.9  0.3596 0.5492740
## sex                 1     62    61.8  0.2158 0.6426882
## fbs                 1      5     5.4  0.0189 0.8909119
## extest              2    530   265.1  0.9260 0.3975478
## restecg             2    361   180.6  0.6308 0.5330351
## slope:chestpain     6   6689  1114.9  3.8950 0.0009917 ***
## slope:restecg       3   2073   691.0  2.4139 0.0673047 .
## fbs:restecg         1    831   831.3  2.9044 0.0896446 .
## exang:restecg       2   1986   993.2  3.4697 0.0327108 *
## sex:extest          2    504   252.1  0.8806 0.4158944
## slope:extest        4   1213   303.1  1.0590 0.3775240
## age:extest          2   1062   531.0  1.8553 0.1586679
## exang:extest        2    347   173.5  0.6061 0.5463291
## chestpain:extest    6   1837   306.1  1.0694 0.3815491
## chol:extest         2     24    11.9  0.0415 0.9593396
## oldpeak:extest      2    502   251.1  0.8773 0.4172267
## restbp:extest       2   1179   589.4  2.0592 0.1298182
## fluoro:extest       2    811   405.4  1.4162 0.2446812
## fbs:extest          2   1106   552.9  1.9316 0.1471844
## extest:restecg      3    883   294.4  1.0286 0.3805940
## Residuals         238  68125   286.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now none of interaction terms of "extest" is significant any more. Thus, so far, we can consider removing the variable "extest" from the model in the end. For now we keep it in the model to see if there are other ways of

interactions with it.

(5) Interaction terms of "fbs".

```
# interaction terms of "fbs"
model3_fbs = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang
            +fbs:slope+fbs:age+fbs:exang+fbs:chestpain
            +fbs:chol+fbs:oldpeak+fbs:restbp+fbs:fluoro
            +fbs:sex+fbs:extest, data3)
anova(model3_fbs)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 55.2772 < 2.2e-16 ***
## age                1  15493 15492.8 52.6943 4.784e-12 ***
## exang              1   9993  9992.5 33.9866 1.685e-08 ***
## chestpain          3   4175  1391.7  4.7335  0.003126 **
## chol               1   1299  1299.2  4.4187  0.036536 *
## oldpeak            1   1255  1254.5  4.2669  0.039880 *
## restbp             1    838   837.7  2.8491  0.092657 .
## fluoro             1    103   102.9  0.3501  0.554565
## sex                1     62    61.8  0.2101  0.647092
## fbs                1      5     5.4  0.0184  0.892348
## extest             2    530   265.1  0.9015  0.407244
## restecg            2    361   180.6  0.6142  0.541903
## slope:chestpain    6   6689  1114.9  3.7920  0.001233 **
## slope:restecg      3   2073   691.0  2.3501  0.072938 .
## fbs:restecg        1    831   831.3  2.8276  0.093892 .
## exang:restecg      2   1986   993.2  3.3779  0.035664 *
## slope:fbs          2    599   299.6  1.0189  0.362483
## age:fbs            1      5     4.5  0.0155  0.901168
## exang:fbs          1    570   569.7  1.9378  0.165127
## chestpain:fbs      3    569   189.7  0.6451  0.586688
## chol:fbs           1      9     8.5  0.0291  0.864789
## oldpeak:fbs        1    283   283.2  0.9633  0.327300
## restbp:fbs         1     16    16.4  0.0557  0.813678
## fluoro:fbs         1     79    79.4  0.2701  0.603688
## sex:fbs            1    727   726.9  2.4722  0.117122
## fbs:extest         2    350   174.9  0.5948  0.552448
## Residuals        253  74385   294.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that only the previously added interaction term "restecg:fbs" is significant for variable "fbs". Thus, so far, we consider not removing the variable "fbs" from the model, and keeping its previously added interaction term to the model: "restecg:fbs".

(6) Interaction terms of "sex".

```
# interaction terms of "sex"
model3_sex = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang
```

```
            +sex:slope+sex:age+sex:exang+sex:chestpain
            +sex:chol+sex:oldpeak+sex:restbp+sex:fluoro
            +sex:fbs+sex:extest+sex:restecg, data3)
anova(model3_sex)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 56.5520 < 2.2e-16 ***
## age                1  15493 15492.8 53.9095 2.897e-12 ***
## exang              1   9993  9992.5 34.7704 1.187e-08 ***
## chestpain          3   4175  1391.7  4.8426  0.002705 **
## chol               1   1299  1299.2  4.5206  0.034462 *
## oldpeak            1   1255  1254.5  4.3653  0.037681 *
## restbp             1    838   837.7  2.9148  0.089001 .
## fluoro             1    103   102.9  0.3582  0.550039
## sex                1     62    61.8  0.2149  0.643328
## fbs                1      5     5.4  0.0188  0.891122
## extest             2    530   265.1  0.9223  0.398929
## restecg            2    361   180.6  0.6283  0.534322
## slope:chestpain    6   6689  1114.9  3.8794  0.001008 **
## slope:restecg      3   2073   691.0  2.4043  0.067998 .
## fbs:restecg        1    831   831.3  2.8928  0.090210 .
## exang:restecg      2   1986   993.2  3.4558  0.033065 *
## slope:sex          2    917   458.4  1.5950  0.204943
## age:sex            1      4     3.9  0.0136  0.907386
## exang:sex          1   1704  1704.5  5.9309  0.015572 *
## chestpain:sex      3    562   187.3  0.6518  0.582480
## chol:sex           1    131   131.0  0.4559  0.500160
## oldpeak:sex        1     23    23.3  0.0810  0.776117
## restbp:sex         1     22    21.8  0.0760  0.782992
## fluoro:sex         1    260   260.2  0.9054  0.342239
## sex:fbs            1   1083  1082.7  3.7676  0.053371 .
## sex:extest         2    401   200.3  0.6971  0.499007
## sex:restecg        1     64    63.9  0.2223  0.637705
## Residuals        252  72421   287.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms according to their significance level
model3_sex = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang
            +sex:exang+sex:fbs+sex:slope+sex:age+sex:chestpain
            +sex:chol+sex:oldpeak+sex:restbp+sex:fluoro
            +sex:extest+sex:restecg, data3)
anova(model3_sex)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 56.5520 < 2.2e-16 ***
## age                1  15493 15492.8 53.9095 2.897e-12 ***
```

```
## exang             1   9993  9992.5 34.7704 1.187e-08 ***
## chestpain         3   4175  1391.7  4.8426  0.002705 **
## chol              1   1299  1299.2  4.5206  0.034462 *
## oldpeak           1   1255  1254.5  4.3653  0.037681 *
## restbp            1    838   837.7  2.9148  0.089001 .
## fluoro            1    103   102.9  0.3582  0.550039
## sex               1     62    61.8  0.2149  0.643328
## fbs               1      5     5.4  0.0188  0.891122
## extest            2    530   265.1  0.9223  0.398929
## restecg           2    361   180.6  0.6283  0.534322
## slope:chestpain   6   6689  1114.9  3.8794  0.001008 **
## slope:restecg     3   2073   691.0  2.4043  0.067998 .
## fbs:restecg       1    831   831.3  2.8928  0.090210 .
## exang:restecg     2   1986   993.2  3.4558  0.033065 *
## exang:sex         1   2261  2261.3  7.8686  0.005422 **
## sex:fbs           1    901   900.6  3.1337  0.077898 .
## slope:sex         2    272   135.8  0.4727  0.623893
## age:sex           1      0     0.1  0.0005  0.981899
## chestpain:sex     3    506   168.7  0.5872  0.623919
## chol:sex          1    118   118.0  0.4105  0.522274
## oldpeak:sex       1     29    29.3  0.1018  0.749962
## restbp:sex        1     73    72.9  0.2535  0.615035
## fluoro:sex        1    546   546.1  1.9004  0.169255
## sex:extest        2    401   200.3  0.6971  0.499007
## sex:restecg       1     64    63.9  0.2223  0.637705
## Residuals       252  72421   287.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, so far, we consider not removing the variable "sex" from the model, and add two of its interaction terms to the model: "sex:exang", "sex:fbs".

(7) Interaction terms of "fluoro".

```
# interaction terms of "fluoro"
model3_fluoro = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang+sex:exang+sex:fbs
            +fluoro:slope+fluoro:age+fluoro:exang+fluoro:chestpain
            +fluoro:chol+fluoro:oldpeak+fluoro:restbp+fluoro:sex
            +fluoro:fbs+fluoro:extest+fluoro:restecg, data3)
anova(model3_fluoro)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 58.7430 < 2.2e-16 ***
## age                1  15493 15492.8 55.9981 1.247e-12 ***
## exang              1   9993  9992.5 36.1175 6.564e-09 ***
## chestpain          3   4175  1391.7  5.0303 0.0021112 **
## chol               1   1299  1299.2  4.6957 0.0311861 *
## oldpeak            1   1255  1254.5  4.5344 0.0342005 *
## restbp             1    838   837.7  3.0277 0.0830871 .
## fluoro             1    103   102.9  0.3721 0.5424243
## sex                1     62    61.8  0.2233 0.6369806
```

```
## fbs                1      5     5.4  0.0195 0.8890476
## extest             2    530   265.1  0.9581 0.3850441
## restecg            2    361   180.6  0.6527 0.5215468
## slope:chestpain    6   6689  1114.9  4.0297 0.0007143 ***
## slope:restecg      3   2073   691.0  2.4975 0.0602747 .
## fbs:restecg        1    831   831.3  3.0049 0.0842528 .
## exang:restecg      2   1986   993.2  3.5897 0.0290440 *
## exang:sex          1   2261  2261.3  8.1735 0.0046108 **
## sex:fbs            1    901   900.6  3.2551 0.0724094 .
## slope:fluoro       2    588   293.8  1.0620 0.3473280
## age:fluoro         1   1529  1529.5  5.5281 0.0194932 *
## exang:fluoro       1    419   419.2  1.5150 0.2195371
## chestpain:fluoro   3     49    16.3  0.0590 0.9811831
## chol:fluoro        1      6     6.2  0.0224 0.8810372
## oldpeak:fluoro     1      1     1.2  0.0043 0.9477293
## restbp:fluoro      1      0     0.4  0.0014 0.9698749
## fluoro:sex         1    632   632.2  2.2850 0.1318995
## fluoro:fbs         1    145   145.0  0.5239 0.4698474
## fluoro:extest      2   1874   937.1  3.3872 0.0353684 *
## fluoro:restecg     2    296   147.8  0.5343 0.5867753
## Residuals        249  68890   276.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms according to their significance level
model3_fluoro = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang+sex:exang+sex:fbs
            +fluoro:age+fluoro:extest+fluoro:slope+fluoro:exang
            +fluoro:chestpain+fluoro:chol+fluoro:oldpeak+fluoro:restbp
            +fluoro:sex+fluoro:fbs+fluoro:restecg, data3)
anova(model3_fluoro)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 58.7430 < 2.2e-16 ***
## age                1  15493 15492.8 55.9981 1.247e-12 ***
## exang              1   9993  9992.5 36.1175 6.564e-09 ***
## chestpain          3   4175  1391.7  5.0303 0.0021112 **
## chol               1   1299  1299.2  4.6957 0.0311861 *
## oldpeak            1   1255  1254.5  4.5344 0.0342005 *
## restbp             1    838   837.7  3.0277 0.0830871 .
## fluoro             1    103   102.9  0.3721 0.5424243
## sex                1     62    61.8  0.2233 0.6369806
## fbs                1      5     5.4  0.0195 0.8890476
## extest             2    530   265.1  0.9581 0.3850441
## restecg            2    361   180.6  0.6527 0.5215468
## slope:chestpain    6   6689  1114.9  4.0297 0.0007143 ***
## slope:restecg      3   2073   691.0  2.4975 0.0602747 .
## fbs:restecg        1    831   831.3  3.0049 0.0842528 .
## exang:restecg      2   1986   993.2  3.5897 0.0290440 *
## exang:sex          1   2261  2261.3  8.1735 0.0046108 **
## sex:fbs            1    901   900.6  3.2551 0.0724094 .
```

```
## age:fluoro         1   1261  1261.0  4.5578 0.0337448 *
## fluoro:extest      2   1244   621.8  2.2475 0.1077974
## slope:fluoro       2   1255   627.4  2.2678 0.1056742
## exang:fluoro       1    247   247.3  0.8939 0.3453378
## chestpain:fluoro   3     37    12.4  0.0447 0.9874091
## chol:fluoro        1      0     0.2  0.0008 0.9770367
## oldpeak:fluoro     1     96    96.2  0.3477 0.5559245
## restbp:fluoro      1      8     8.3  0.0299 0.8628561
## fluoro:sex         1    919   918.7  3.3205 0.0696209 .
## fluoro:fbs         1    177   177.0  0.6399 0.4245031
## fluoro:restecg     2    296   147.8  0.5343 0.5867753
## Residuals        249  68890   276.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# reoder the interaction terms again according to their significance level
model3_fluoro = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang+sex:exang+sex:fbs
            +fluoro:age+fluoro:extest+fluoro:sex+fluoro:slope+fluoro:exang
            +fluoro:chestpain+fluoro:chol+fluoro:oldpeak+fluoro:restbp
            +fluoro:fbs+fluoro:restecg, data3)
anova(model3_fluoro)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 58.7430 < 2.2e-16 ***
## age                1  15493 15492.8 55.9981 1.247e-12 ***
## exang              1   9993  9992.5 36.1175 6.564e-09 ***
## chestpain          3   4175  1391.7  5.0303 0.0021112 **
## chol               1   1299  1299.2  4.6957 0.0311861 *
## oldpeak            1   1255  1254.5  4.5344 0.0342005 *
## restbp             1    838   837.7  3.0277 0.0830871 .
## fluoro             1    103   102.9  0.3721 0.5424243
## sex                1     62    61.8  0.2233 0.6369806
## fbs                1      5     5.4  0.0195 0.8890476
## extest             2    530   265.1  0.9581 0.3850441
## restecg            2    361   180.6  0.6527 0.5215468
## slope:chestpain    6   6689  1114.9  4.0297 0.0007143 ***
## slope:restecg      3   2073   691.0  2.4975 0.0602747 .
## fbs:restecg        1    831   831.3  3.0049 0.0842528 .
## exang:restecg      2   1986   993.2  3.5897 0.0290440 *
## exang:sex          1   2261  2261.3  8.1735 0.0046108 **
## sex:fbs            1    901   900.6  3.2551 0.0724094 .
## age:fluoro         1   1261  1261.0  4.5578 0.0337448 *
## fluoro:extest      2   1244   621.8  2.2475 0.1077974
## fluoro:sex         1    446   445.9  1.6117 0.2054440
## slope:fluoro       2   1393   696.4  2.5172 0.0827401 .
## exang:fluoro       1    386   385.7  1.3940 0.2388644
## chestpain:fluoro   3    124    41.4  0.1497 0.9298372
## chol:fluoro        1     16    15.6  0.0564 0.8124237
## oldpeak:fluoro     1    162   162.2  0.5864 0.4445350
## restbp:fluoro      1     36    36.2  0.1309 0.7178292
```

```
## fluoro:fbs         1    177   177.0  0.6399 0.4245031
## fluoro:restecg     2    296   147.8  0.5343 0.5867753
## Residuals        249  68890   276.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms again according to their significance level
model3_fluoro = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang+sex:exang+sex:fbs
            +fluoro:age+fluoro:extest+fluoro:slope+fluoro:sex+fluoro:exang
            +fluoro:chestpain+fluoro:chol+fluoro:oldpeak+fluoro:restbp
            +fluoro:fbs+fluoro:restecg, data3)
anova(model3_fluoro)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 58.7430 < 2.2e-16 ***
## age                1  15493 15492.8 55.9981 1.247e-12 ***
## exang              1   9993  9992.5 36.1175 6.564e-09 ***
## chestpain          3   4175  1391.7  5.0303 0.0021112 **
## chol               1   1299  1299.2  4.6957 0.0311861 *
## oldpeak            1   1255  1254.5  4.5344 0.0342005 *
## restbp             1    838   837.7  3.0277 0.0830871 .
## fluoro             1    103   102.9  0.3721 0.5424243
## sex                1     62    61.8  0.2233 0.6369806
## fbs                1      5     5.4  0.0195 0.8890476
## extest             2    530   265.1  0.9581 0.3850441
## restecg            2    361   180.6  0.6527 0.5215468
## slope:chestpain    6   6689  1114.9  4.0297 0.0007143 ***
## slope:restecg      3   2073   691.0  2.4975 0.0602747 .
## fbs:restecg        1    831   831.3  3.0049 0.0842528 .
## exang:restecg      2   1986   993.2  3.5897 0.0290440 *
## exang:sex          1   2261  2261.3  8.1735 0.0046108 **
## sex:fbs            1    901   900.6  3.2551 0.0724094 .
## age:fluoro         1   1261  1261.0  4.5578 0.0337448 *
## fluoro:extest      2   1244   621.8  2.2475 0.1077974
## slope:fluoro       2   1255   627.4  2.2678 0.1056742
## fluoro:sex         1    584   583.9  2.1104 0.1475626
## exang:fluoro       1    386   385.7  1.3940 0.2388644
## chestpain:fluoro   3    124    41.4  0.1497 0.9298372
## chol:fluoro        1     16    15.6  0.0564 0.8124237
## oldpeak:fluoro     1    162   162.2  0.5864 0.4445350
## restbp:fluoro      1     36    36.2  0.1309 0.7178292
## fluoro:fbs         1    177   177.0  0.6399 0.4245031
## fluoro:restecg     2    296   147.8  0.5343 0.5867753
## Residuals        249  68890   276.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, so far, we consider not removing the variable "fluoro" from the model, and add one of its interaction terms to the model: "fluoro:age".

(8) Interaction terms of "restbp".

```r
# interaction terms of "restbp"
model3_restbp = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+extest+restecg + slope:chestpain
            +restecg:slope+restecg:fbs+restecg:exang+sex:exang+sex:fbs+fluoro:age
            +restbp:slope+restbp:age+restbp:exang+restbp:chestpain
            +restbp:chol+restbp:oldpeak+restbp:fluoro+restbp:sex
            +restbp:fbs+restbp:extest+restbp:restecg, data3)
anova(model3_restbp)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## slope               2  32504 16252.2 58.6467 < 2.2e-16 ***
## age                 1  15493 15492.8 55.9063 1.309e-12 ***
## exang               1   9993  9992.5 36.0583 6.772e-09 ***
## chestpain           3   4175  1391.7  5.0220 0.0021361 **
## chol                1   1299  1299.2  4.6880 0.0313272 *
## oldpeak             1   1255  1254.5  4.5270 0.0343505 *
## restbp              1    838   837.7  3.0228 0.0833434 .
## fluoro              1    103   102.9  0.3715 0.5427573
## sex                 1     62    61.8  0.2229 0.6372584
## fbs                 1      5     5.4  0.0195 0.8891384
## extest              2    530   265.1  0.9565 0.3856502
## restecg             2    361   180.6  0.6516 0.5221057
## slope:chestpain     6   6689  1114.9  4.0231 0.0007264 ***
## slope:restecg       3   2073   691.0  2.4934 0.0606067 .
## fbs:restecg         1    831   831.3  2.9999 0.0845110 .
## exang:restecg       2   1986   993.2  3.5838 0.0292164 *
## exang:sex           1   2261  2261.3  8.1601 0.0046454 **
## sex:fbs             1    901   900.6  3.2498 0.0726471 .
## age:fluoro          1   1261  1261.0  4.5504 0.0338935 *
## slope:restbp        2   1315   657.7  2.3735 0.0952700 .
## age:restbp          1    204   203.7  0.7352 0.3920462
## exang:restbp        1     14    14.0  0.0503 0.8226565
## chestpain:restbp    3    581   193.6  0.6987 0.5536193
## chol:restbp         1     68    68.0  0.2454 0.6207529
## oldpeak:restbp      1    121   120.9  0.4361 0.5096017
## restbp:fluoro       1     41    41.0  0.1479 0.7008602
## restbp:sex          1     68    68.3  0.2466 0.6199167
## restbp:fbs          1    336   336.3  1.2134 0.2717310
## restbp:extest       2   1044   522.2  1.8845 0.1540770
## restbp:restecg      2    650   325.0  1.1728 0.3111954
## Residuals         248  68726   277.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, so far, we consider not removing the variable "restbp" from the model, and add one of its interaction terms to the model: "restbp:slope".

Conclusion:

So far, after examing the interaction terms of the other covariates (not the ones in the selected smaller model), we have added seven interaction terms into the model: "restecg:slope", "restecg:fbs", "restecg:exang", "sex:exang", "sex:fbs", "fluoro:age", "restbp:slope". There is no interaction terms of variable "extest", so now

we can remove this variable from the model.

Next, we try to further explore the current model with all the eight interaction terms and see if we can further reduce the size of the model.

(9) Explore the current model.

```
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +restecg:slope +restecg:fbs +restecg:exang
           +sex:exang +sex:fbs +fluoro:age +restbp:slope, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 59.4758 < 2.2e-16 ***
## age                1  15493 15492.8 56.6966 8.072e-13 ***
## exang              1   9993  9992.5 36.5681 5.016e-09 ***
## chestpain          3   4175  1391.7  5.0930 0.0019210 **
## chol               1   1299  1299.2  4.7543 0.0301081 *
## oldpeak            1   1255  1254.5  4.5910 0.0330546 *
## restbp             1    838   837.7  3.0655 0.0811312 .
## fluoro             1    103   102.9  0.3767 0.5398869
## sex                1     62    61.8  0.2260 0.6348648
## fbs                1      5     5.4  0.0197 0.8883558
## restecg            2    426   213.1  0.7798 0.4595622
## slope:chestpain    6   6877  1146.1  4.1943 0.0004762 ***
## slope:restecg      3   1979   659.8  2.4145 0.0669861 .
## fbs:restecg        1    884   883.7  3.2341 0.0732635 .
## exang:restecg      2   1855   927.4  3.3938 0.0350512 *
## exang:sex          1   2217  2216.9  8.1129 0.0047411 **
## sex:fbs            1    964   964.4  3.5294 0.0613932 .
## age:fluoro         1   1241  1241.1  4.5419 0.0339987 *
## slope:restbp       2   1480   739.9  2.7076 0.0685478 .
## Residuals        264  72140   273.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms again according to their significance level
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +sex:exang +fluoro:age +restecg:exang
           +sex:fbs +restecg:slope +restbp:slope +restecg:fbs, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 59.4758 < 2.2e-16 ***
## age                1  15493 15492.8 56.6966 8.072e-13 ***
## exang              1   9993  9992.5 36.5681 5.016e-09 ***
## chestpain          3   4175  1391.7  5.0930 0.0019210 **
## chol               1   1299  1299.2  4.7543 0.0301081 *
## oldpeak            1   1255  1254.5  4.5910 0.0330546 *
```

```
## restbp            1    838   837.7  3.0655 0.0811312 .
## fluoro            1    103   102.9  0.3767 0.5398869
## sex               1     62    61.8  0.2260 0.6348648
## fbs               1      5     5.4  0.0197 0.8883558
## restecg           2    426   213.1  0.7798 0.4595622
## slope:chestpain   6   6877  1146.1  4.1943 0.0004762 ***
## exang:sex         1   2177  2176.5  7.9650 0.0051318 **
## age:fluoro        1   1105  1104.6  4.0424 0.0453887 *
## exang:restecg     2   1043   521.6  1.9087 0.1503135
## sex:fbs           1   1490  1489.7  5.4516 0.0203006 *
## slope:restecg     3   2609   869.8  3.1831 0.0244389 *
## slope:restbp      2   1548   773.8  2.8319 0.0606898 .
## fbs:restecg       1    649   649.1  2.3753 0.1244667
## Residuals       264  72140   273.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# remove "restecg:fbs"
# reoder the interaction terms again according to their significance level
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +sex:exang +fluoro:age +sex:fbs
           +restecg:slope +restbp:slope +restecg:exang, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 59.1687 < 2.2e-16 ***
## age                1  15493 15492.8 56.4039 9.042e-13 ***
## exang              1   9993  9992.5 36.3793 5.438e-09 ***
## chestpain          3   4175  1391.7  5.0667 0.0019884 **
## chol               1   1299  1299.2  4.7298 0.0305295 *
## oldpeak            1   1255  1254.5  4.5673 0.0335033 *
## restbp             1    838   837.7  3.0497 0.0819113 .
## fluoro             1    103   102.9  0.3748 0.5409323
## sex                1     62    61.8  0.2249 0.6357382
## fbs                1      5     5.4  0.0196 0.8886421
## restecg            2    426   213.1  0.7757 0.4614013
## slope:chestpain    6   6877  1146.1  4.1726 0.0005003 ***
## exang:sex          1   2177  2176.5  7.9239 0.0052448 **
## age:fluoro         1   1105  1104.6  4.0215 0.0459409 *
## sex:fbs            1   1400  1400.2  5.0975 0.0247744 *
## slope:restecg      3   1946   648.7  2.3617 0.0717294 .
## slope:restbp       2   1295   647.3  2.3565 0.0967311 .
## exang:restecg      2   2049  1024.5  3.7299 0.0252633 *
## Residuals        265  72789   274.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# reoder the interaction terms again according to their significance level
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +sex:exang +sex:fbs +fluoro:age
```

```
            +restecg:exang +restecg:slope +restbp:slope, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 59.1687 < 2.2e-16 ***
## age                1  15493 15492.8 56.4039 9.042e-13 ***
## exang              1   9993  9992.5 36.3793 5.438e-09 ***
## chestpain          3   4175  1391.7  5.0667 0.0019884 **
## chol               1   1299  1299.2  4.7298 0.0305295 *
## oldpeak            1   1255  1254.5  4.5673 0.0335033 *
## restbp             1    838   837.7  3.0497 0.0819113 .
## fluoro             1    103   102.9  0.3748 0.5409323
## sex                1     62    61.8  0.2249 0.6357382
## fbs                1      5     5.4  0.0196 0.8886421
## restecg            2    426   213.1  0.7757 0.4614013
## slope:chestpain    6   6877  1146.1  4.1726 0.0005003 ***
## exang:sex          1   2177  2176.5  7.9239 0.0052448 **
## sex:fbs            1   1286  1285.7  4.6807 0.0313975 *
## age:fluoro         1   1219  1219.1  4.4383 0.0360815 *
## exang:restecg      2   1133   566.3  2.0618 0.1292592
## slope:restecg      3   2609   869.8  3.1666 0.0249691 *
## slope:restbp       2   1548   773.8  2.8173 0.0615580 .
## Residuals        265  72789   274.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# reoder the interaction terms again according to their significance level
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +restbp+fluoro+sex+fbs+restecg + slope:chestpain
            +sex:exang +sex:fbs +fluoro:age
            +restecg:slope +restbp:slope +restecg:exang, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 59.1687 < 2.2e-16 ***
## age                1  15493 15492.8 56.4039 9.042e-13 ***
## exang              1   9993  9992.5 36.3793 5.438e-09 ***
## chestpain          3   4175  1391.7  5.0667 0.0019884 **
## chol               1   1299  1299.2  4.7298 0.0305295 *
## oldpeak            1   1255  1254.5  4.5673 0.0335033 *
## restbp             1    838   837.7  3.0497 0.0819113 .
## fluoro             1    103   102.9  0.3748 0.5409323
## sex                1     62    61.8  0.2249 0.6357382
## fbs                1      5     5.4  0.0196 0.8886421
## restecg            2    426   213.1  0.7757 0.4614013
## slope:chestpain    6   6877  1146.1  4.1726 0.0005003 ***
## exang:sex          1   2177  2176.5  7.9239 0.0052448 **
## sex:fbs            1   1286  1285.7  4.6807 0.0313975 *
## age:fluoro         1   1219  1219.1  4.4383 0.0360815 *
```

```
## slope:restecg     3   1946   648.7  2.3617 0.0717294 .
## slope:restbp      2   1295   647.3  2.3565 0.0967311 .
## exang:restecg     2   2049  1024.5  3.7299 0.0252633 *
## Residuals       265  72789   274.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After several trials, only the interaction term "restecg:fbs" got removed from the model.

Now we get the model with seven interaction terms: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + restbp + fluoro + sex + fbs + restecg + slope : chestpain + exang : sex + sex : fbs + age : fluoro + slope : restecg + slope : restbp + exang : restecg$

Then, let's have a look at the model summary.

```
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +sex:exang +sex:fbs +fluoro:age
           +restecg:slope +restbp:slope +restecg:exang, data3)
summary(model)
```

```
##
## Call:
## lm(formula = maxhr ~ slope + age + exang + chestpain + chol +
##     oldpeak + restbp + fluoro + sex + fbs + restecg + slope:chestpain +
##     sex:exang + sex:fbs + fluoro:age + restecg:slope + restbp:slope +
##     restecg:exang, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.741  -9.329   0.780  11.273  38.475
##
## Coefficients: (1 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         20.8838    61.8690   0.338  0.73597
## slope2             178.5344    82.1194   2.174  0.03058 *
## slope3             -52.5710   186.7328  -0.282  0.77852
## age                 -1.0265     0.1451  -7.072 1.35e-11 ***
## exang1              12.8224     5.2509   2.442  0.01526 *
## chestpain2           3.0097     6.4531   0.466  0.64132
## chestpain3          -0.8935     6.2499  -0.143  0.88642
## chestpain4           1.4782     6.2288   0.237  0.81260
## chol                 8.6326     5.1139   1.688  0.09257 .
## oldpeak             -2.6707     2.0932  -1.276  0.20312
## restbp              30.6115    11.8212   2.590  0.01014 *
## fluoro             -17.7481     7.7603  -2.287  0.02298 *
## sex1                -1.1613     2.7147  -0.428  0.66917
## fbs1                -8.7701     5.4850  -1.599  0.11103
## restecg1             6.1554    20.2140   0.305  0.76098
## restecg2            -0.9883     3.1071  -0.318  0.75067
## slope2:chestpain2 -18.0354     9.7422  -1.851  0.06524 .
## slope3:chestpain2  28.3158    19.8837   1.424  0.15560
## slope2:chestpain3  -7.7950     8.6508  -0.901  0.36837
## slope3:chestpain3  22.9900    15.5425   1.479  0.14028
## slope2:chestpain4 -25.6684     8.4357  -3.043  0.00258 **
## slope3:chestpain4   0.8568    13.2652   0.065  0.94855
```

```
## exang1:sex1       -17.2514     5.2822  -3.266  0.00123 **
## sex1:fbs1          14.1204     6.4538   2.188  0.02955 *
## age:fluoro          0.3199     0.1323   2.419  0.01626 *
## slope2:restecg1   -11.4056    26.8651  -0.425  0.67151
## slope3:restecg1         NA         NA      NA       NA
## slope2:restecg2    14.0276     4.3477   3.226  0.00141 **
## slope3:restecg2    -0.1634     9.4493  -0.017  0.98621
## slope2:restbp     -37.5466    16.6607  -2.254  0.02504 *
## slope3:restbp       6.8312    37.2295   0.183  0.85455
## exang1:restecg1   -11.5501    21.4126  -0.539  0.59006
## exang1:restecg2   -12.2856     4.5117  -2.723  0.00690 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.57 on 265 degrees of freedom
## Multiple R-squared:  0.5328, Adjusted R-squared:  0.4781
## F-statistic: 9.748 on 31 and 265 DF,  p-value: < 2.2e-16
```

Here we see that one of interaction term of "slope:restecg" has NA values for coefficients, which indicates that this term may be linearly related to the other terms. Thus, we should remove the whole interaction term of "slope:restecg" to address with this problem.

Therefore, now we have six interaction terms left: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + restbp + fluoro + sex + fbs + restecg + slope : chestpain + exang : sex + sex : fbs + age : fluoro + slope : restbp + exang : restecg$

Then, we have a look at the anova tests again.

```
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +restbp+fluoro+sex+fbs+restecg + slope:chestpain
           +sex:exang +sex:fbs +fluoro:age
           +restbp:slope +restecg:exang, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 57.3168 < 2.2e-16 ***
## age                1  15493 15492.8 54.6386 1.852e-12 ***
## exang              1   9993  9992.5 35.2406 8.993e-09 ***
## chestpain          3   4175  1391.7  4.9081 0.0024531 **
## chol               1   1299  1299.2  4.5817 0.0332171 *
## oldpeak            1   1255  1254.5  4.4244 0.0363615 *
## restbp             1    838   837.7  2.9542 0.0868073 .
## fluoro             1    103   102.9  0.3631 0.5473238
## sex                1     62    61.8  0.2178 0.6410697
## fbs                1      5     5.4  0.0190 0.8903862
## restecg            2    426   213.1  0.7515 0.4726640
## slope:chestpain    6   6877  1146.1  4.0420 0.0006763 ***
## exang:sex          1   2177  2176.5  7.6759 0.0059879 **
## sex:fbs            1   1286  1285.7  4.5342 0.0341357 *
## age:fluoro         1   1219  1219.1  4.2994 0.0390816 *
## slope:restbp       2    909   454.7  1.6037 0.2030734
## exang:restecg      2   1178   588.9  2.0770 0.1273191
## Residuals        268  75992   283.6
```

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

```
# remove "slope:restbp", "exang:restecg" and variables "restecg", "restbp"
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +fluoro+sex+fbs + slope:chestpain
           +sex:exang +sex:fbs +fluoro:age, data3)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: maxhr
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## slope              2  32504 16252.2 56.0849 < 2.2e-16 ***
## age                1  15493 15492.8 53.4643 2.881e-12 ***
## exang              1   9993  9992.5 34.4832 1.237e-08 ***
## chestpain          3   4175  1391.7  4.8027  0.002813 **
## chol               1   1299  1299.2  4.4833  0.035126 *
## oldpeak            1   1255  1254.5  4.3293  0.038388 *
## fluoro             1    111   111.3  0.3842  0.535888
## sex                1     88    87.7  0.3028  0.582578
## fbs                1     40    39.8  0.1374  0.711166
## slope:chestpain    6   6403  1067.1  3.6826  0.001550 **
## exang:sex          1   2358  2357.7  8.1361  0.004669 **
## sex:fbs            1   1098  1098.3  3.7900  0.052577 .
## age:fluoro         1   1284  1284.0  4.4310  0.036199 *
## Residuals        275  79689   289.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, we can remove the interaction terms "slope:restbp" and "exang:restecg" from the model. So there is no any interaction terms of "restecg" and "restbp" left in the model thus we can remove them as well.

Now we get the model with four interaction terms: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + fluoro + sex + fbs + slope : chestpain + exang : sex + sex : fbs + age : fluoro$

Then, let's have a look at the model summary.

```
model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
           +fluoro+sex+fbs + slope:chestpain
           +sex:exang +sex:fbs +fluoro:age, data3)
summary(model)
```

```
##
## Call:
## lm(formula = maxhr ~ slope + age + exang + chestpain + chol +
##     oldpeak + fluoro + sex + fbs + slope:chestpain + sex:exang +
##     sex:fbs + fluoro:age, data = data3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.545 -10.677   1.226  11.731  38.886
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       149.7613    28.2627   5.299 2.39e-07 ***
## slope2              1.0057     7.8397   0.128  0.89802
```

```
## slope3            -17.4267    11.7668  -1.481  0.13975
## age                -0.8983     0.1408  -6.380 7.48e-10 ***
## exang1              4.4066     4.5229   0.974  0.33078
## chestpain2          0.5794     6.5153   0.089  0.92921
## chestpain3         -2.9517     6.3271  -0.467  0.64122
## chestpain4         -1.2646     6.3066  -0.201  0.84122
## chol               11.2922     5.0954   2.216  0.02750 *
## oldpeak            -1.7160     2.0905  -0.821  0.41243
## fluoro            -16.3719     7.8913  -2.075  0.03895 *
## sex1               -0.2087     2.7485  -0.076  0.93954
## fbs1               -9.0436     5.4980  -1.645  0.10114
## slope2:chestpain2 -13.7316     9.8824  -1.390  0.16580
## slope3:chestpain2  21.7392    17.0034   1.279  0.20215
## slope2:chestpain3  -7.3204     8.7885  -0.833  0.40559
## slope3:chestpain3  19.9654    14.1890   1.407  0.16053
## slope2:chestpain4 -23.5953     8.5740  -2.752  0.00632 **
## slope3:chestpain4   1.1221    13.0059   0.086  0.93131
## exang1:sex1       -15.7323     5.0572  -3.111  0.00206 **
## sex1:fbs1          13.2356     6.4573   2.050  0.04134 *
## age:fluoro          0.2824     0.1342   2.105  0.03620 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.02 on 275 degrees of freedom
## Multiple R-squared:  0.4885, Adjusted R-squared:  0.4494
## F-statistic: 12.51 on 21 and 275 DF,  p-value: < 2.2e-16
```

Conclusion:

Therefore, we now get our final model: $maxhr \sim slope + age + exang + chestpain + chol + oldpeak + fluoro + sex + fbs + slope : chestpain + exang : sex + sex : fbs + age : fluoro$

### Part 11

Prediction error.

Now we use leave-one-out cross validation to calculate the average of prediction errors.

```
# leave-one-out cross-validation to compute the average of prediction error
n = nrow(data3)
store_pred_error = rep(0,n)

for (i in 1:n) {
    model = lm(maxhr ~ slope+age+exang+chestpain+chol+oldpeak
            +fluoro+sex+fbs + slope:chestpain
            +sex:exang +sex:fbs +fluoro:age, data3)
    xi = data.frame(data3[i,-13])
    fitted_yi = predict(model, xi)[1]
    store_pred_error[i] = data3[i,13] - fitted_yi
}
ave_square_error = sum(store_pred_error^2)/n
print(paste0("average of prediction error = ",ave_square_error))
```

```
## [1] "average of prediction error = 268.313476634463"
```

Note that this average of prediction error is much smaller than the previous lasso regression model with best lambda value, so now our model is performing better.