DATA 605 Discussion 12

Vishal Arora

11/16/2019

Heart Disease

Data Context : This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Downladed from : https://www.kaggle.com/ronitf/heart-disease-uci

The terms choosen are as follows:

Dichotomous Term is sex (1 = male; 0 = female) Quadratic Term: Age Interaction sex vs. cholesterol (quantitative)

Read data

```
heart <- read.csv("heart.csv",header=TRUE, sep=",")
head(heart, 10)
```

```
##    ï..age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca
## 1      63   1  3      145  233   1       0     150     0     2.3     0  0
## 2      37   1  2      130  250   0       1     187     0     3.5     0  0
## 3      41   0  1      130  204   0       0     172     0     1.4     2  0
## 4      56   1  1      120  236   0       1     178     0     0.8     2  0
## 5      57   0  0      120  354   0       1     163     1     0.6     2  0
## 6      57   1  0      140  192   0       1     148     0     0.4     1  0
## 7      56   0  1      140  294   0       0     153     0     1.3     1  0
## 8      44   1  1      120  263   0       1     173     0     0.0     2  0
## 9      52   1  2      172  199   1       1     162     0     0.5     2  0
## 10     57   1  2      150  168   0       1     174     0     1.6     2  0
##    thal target
## 1     1      1
## 2     2      1
## 3     2      1
## 4     2      1
## 5     2      1
## 6     1      1
## 7     2      1
## 8     3      1
## 9     3      1
## 10    2      1
```

Apply Linear Model

```
# Quadratic Term
age2 <- heart$ï..age^2

# Dichotomous vs. quantative
sx_chl <- heart$sex * heart$chol

# first model
modl <- lm(thalach ~ sex + ï..age
 + age2 + cp + trestbps+fbs+restecg+exang+oldpeak+slope+ca+thal+target+ chol + sx_chl, heart)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ sex + ï..age + age2 + cp + trestbps +
##     fbs + restecg + exang + oldpeak + slope + ca + thal + target +
##     chol + sx_chl, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.300 -10.379   1.885  11.830  48.304
##
## Coefficients:
##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 164.889971  32.913932   5.010 0.000000954 ***
## sex           0.315713  10.883667   0.029    0.976878
## ï..age       -1.319470   1.184114  -1.114    0.266078
## age2          0.004353   0.010980   0.396    0.692068
## cp            2.124421   1.199762   1.771    0.077672 .
## trestbps      0.124942   0.064708   1.931    0.054485 .
## fbs           1.960890   3.104881   0.632    0.528183
## restecg      -1.418114   2.068313  -0.686    0.493495
## exang        -9.491782   2.626219  -3.614    0.000356 ***
## oldpeak      -0.886666   1.202700  -0.737    0.461585
## slope         7.556577   2.181148   3.464    0.000612 ***
## ca           -0.452962   1.172178  -0.386    0.699466
## thal          2.165220   1.870993   1.157    0.248130
## target        7.944773   3.002587   2.646    0.008595 **
## chol          0.036010   0.030141   1.195    0.233190
## sx_chl        0.003584   0.042354   0.085    0.932622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.27 on 287 degrees of freedom
## Multiple R-squared:  0.3956, Adjusted R-squared:  0.364
## F-statistic: 12.52 on 15 and 287 DF,  p-value: < 2.2e-16
```

Using Backwards Elimination

Removing variable with highest p value - one at a time. Starting with sex:

```r
modl <- update(modl, .~. -sex)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + age2 + cp + trestbps + fbs +
##     restecg + exang + oldpeak + slope + ca + thal + target +
##     chol + sx_chl, data = heart)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.28 -10.35   1.90  11.80  48.30
##
## Coefficients:
##               Estimate Std. Error t value   Pr(>|t|)
## (Intercept) 164.993433  32.663303   5.051 0.00000078 ***
## ï..age       -1.317503   1.180117  -1.116   0.265174
## age2          0.004334   0.010942   0.396   0.692310
## cp            2.125509   1.197094   1.776   0.076862 .
## trestbps      0.124899   0.064578   1.934   0.054084 .
## fbs           1.969747   3.084466   0.639   0.523589
## restecg      -1.423273   2.057077  -0.692   0.489563
## exang        -9.491102   2.621555  -3.620   0.000347 ***
## oldpeak      -0.882572   1.192316  -0.740   0.459772
## slope         7.567227   2.146290   3.526   0.000491 ***
## ca           -0.454019   1.169577  -0.388   0.698161
## thal          2.168395   1.864546   1.163   0.245808
## target        7.944107   2.997286   2.650   0.008483 **
## chol          0.035387   0.021119   1.676   0.094909 .
## sx_chl        0.004780   0.009733   0.491   0.623737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.23 on 288 degrees of freedom
## Multiple R-squared:  0.3956, Adjusted R-squared:  0.3662
## F-statistic: 13.46 on 14 and 288 DF,  p-value: < 2.2e-16
```

Removing ca: number of major vessels

```r
modl <- update(modl, .~. -ca)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + age2 + cp + trestbps + fbs +
##     restecg + exang + oldpeak + slope + thal + target + chol +
##     sx_chl, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.693 -10.107   1.754  11.778  47.909
##
```

```
## Coefficients:
##               Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 164.238103  32.557351   5.045 0.000000804 ***
## ï..age       -1.291530   1.176486  -1.098    0.273212
## age2          0.004000   0.010892   0.367    0.713731
## cp            2.154477   1.193009   1.806    0.071971 .
## trestbps      0.125473   0.064466   1.946    0.052583 .
## fbs           1.800014   3.048828   0.590    0.555387
## restecg      -1.420493   2.054040  -0.692    0.489769
## exang        -9.406702   2.608682  -3.606    0.000366 ***
## oldpeak      -0.933972   1.183199  -0.789    0.430548
## slope         7.467407   2.127697   3.510    0.000520 ***
## thal          2.154899   1.861480   1.158    0.247972
## target        8.255172   2.883938   2.862    0.004511 **
## chol          0.035409   0.021088   1.679    0.094213 .
## sx_chl        0.004602   0.009708   0.474    0.635832
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.21 on 289 degrees of freedom
## Multiple R-squared:  0.3953, Adjusted R-squared:  0.3681
## F-statistic: 14.53 on 13 and 289 DF,  p-value: < 2.2e-16
```

Remove age^2

```r
modl <- update(modl, .~. -age2)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + fbs + restecg +
##     exang + oldpeak + slope + thal + target + chol + sx_chl,
##     data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.889 -10.012   1.787  11.560  47.931
##
## Coefficients:
##               Estimate Std. Error t value       Pr(>|t|)
## (Intercept) 153.131264  12.028995  12.730        < 2e-16 ***
## ï..age       -0.862025   0.126517  -6.814 0.000000000055 ***
## cp            2.180698   1.189092   1.834       0.067691 .
## trestbps      0.125405   0.064370   1.948       0.052356 .
## fbs           1.692430   3.030189   0.559       0.576918
## restecg      -1.407098   2.050650  -0.686       0.493153
## exang        -9.404929   2.604783  -3.611       0.000360 ***
## oldpeak      -0.931317   1.181410  -0.788       0.431159
## slope         7.501198   2.122533   3.534       0.000476 ***
## thal          2.135826   1.857978   1.150       0.251279
## target        8.278011   2.878963   2.875       0.004335 **
## chol          0.035008   0.021028   1.665       0.097034 .
## sx_chl        0.004474   0.009687   0.462       0.644509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.18 on 290 degrees of freedom
## Multiple R-squared:  0.395,  Adjusted R-squared:   0.37
## F-statistic: 15.78 on 12 and 290 DF,  p-value: < 2.2e-16
```

Remove: Sex vs Cholest.

```
modl <- update(modl, .~. -sx_chl)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + fbs + restecg +
##     exang + oldpeak + slope + thal + target + chol, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.491 -10.335   1.681  11.419  47.116
##
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)
## (Intercept) 154.31610   11.73640  13.149         < 2e-16 ***
## ï..age       -0.86896    0.12545  -6.927 0.0000000000277 ***
## cp            2.23120    1.18245   1.887        0.060165 .
## trestbps      0.12324    0.06411   1.922        0.055543 .
## fbs           1.75072    3.02346   0.579        0.563007
## restecg      -1.46094    2.04457  -0.715        0.475462
## exang        -9.34516    2.59805  -3.597        0.000378 ***
## oldpeak      -0.93254    1.17981  -0.790        0.429928
## slope         7.54522    2.11752   3.563        0.000428 ***
## thal          2.25434    1.83769   1.227        0.220916
## target        7.93032    2.77505   2.858        0.004575 **
## chol          0.03507    0.02100   1.670        0.096035 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.16 on 291 degrees of freedom
## Multiple R-squared:  0.3946, Adjusted R-squared:  0.3717
## F-statistic: 17.24 on 11 and 291 DF,  p-value: < 2.2e-16
```

Remove fbs:fasting blood sugar

```
modl <- update(modl, .~. -fbs)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + restecg + exang +
##     oldpeak + slope + thal + target + chol, data = heart)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -58.072 -10.559   1.544  11.634  46.962 
## 
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)    
## (Intercept) 153.84255   11.69453  13.155         < 2e-16 ***
## ï..age       -0.86332    0.12493  -6.910 0.0000000000304 ***
## cp            2.30385    1.17444   1.962        0.050753 .  
## trestbps      0.12846    0.06340   2.026        0.043667 *  
## restecg      -1.52824    2.03893  -0.750        0.454140    
## exang        -9.27717    2.59244  -3.579        0.000404 ***
## oldpeak      -0.97787    1.17587  -0.832        0.406308    
## slope         7.48124    2.11223   3.542        0.000462 ***
## thal          2.21444    1.83430   1.207        0.228316    
## target        7.88037    2.77055   2.844        0.004765 ** 
## chol          0.03482    0.02097   1.660        0.097911 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.14 on 292 degrees of freedom
## Multiple R-squared:  0.3939, Adjusted R-squared:  0.3731 
## F-statistic: 18.97 on 10 and 292 DF,  p-value: < 2.2e-16
```

Remove restecg: resting electrocardiographic results

```r
modl <- update(modl, .~. -restecg)
summary(modl)
```

```
## 
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + exang + oldpeak + 
##     slope + thal + target + chol, data = heart)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -58.930 -10.362   1.935  11.878  46.102 
## 
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)    
## (Intercept) 152.29332   11.50182  13.241         < 2e-16 ***
## ï..age       -0.85946    0.12473  -6.890 0.0000000000341 ***
## cp            2.31659    1.17344   1.974        0.049299 *  
## trestbps      0.13174    0.06321   2.084        0.037993 *  
## exang        -9.25423    2.59032  -3.573        0.000413 ***
## oldpeak      -1.01680    1.17384  -0.866        0.387080    
## slope         7.39037    2.10717   3.507        0.000524 ***
## thal          2.15370    1.83114   1.176        0.240489    
## target        7.68187    2.75580   2.788        0.005658 ** 
## chol          0.03684    0.02078   1.773        0.077348 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.12 on 293 degrees of freedom
## Multiple R-squared:  0.3927, Adjusted R-squared:  0.374
## F-statistic: 21.05 on 9 and 293 DF,  p-value: < 2.2e-16
```

Remove oldpeak: ST depression induced by exercise relative to rest

```
modl <- update(modl, .~. -oldpeak)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + exang + slope +
##     thal + target + chol, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.380 -10.718   1.829  11.223  45.912
##
## Coefficients:
##              Estimate Std. Error t value       Pr(>|t|)
## (Intercept) 151.24588   11.43322  13.229        < 2e-16 ***
## ï..age       -0.86624    0.12443  -6.961 0.000000000022 ***
## cp            2.27798    1.17209   1.944       0.052908 .
## trestbps      0.12630    0.06287   2.009       0.045450 *
## exang        -9.41774    2.58233  -3.647       0.000314 ***
## slope         8.28532    1.83572   4.513 0.000009239010 ***
## thal          2.01228    1.82307   1.104       0.270588
## target        8.17002    2.69642   3.030       0.002663 **
## chol          0.03674    0.02077   1.769       0.078012 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.11 on 294 degrees of freedom
## Multiple R-squared:  0.3911, Adjusted R-squared:  0.3746
## F-statistic: 23.61 on 8 and 294 DF,  p-value: < 2.2e-16
```

Remove thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

```
modl <- update(modl, .~. -thal)
summary(modl)
```

```
##
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + exang + slope +
##     target + chol, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.288 -10.140   1.905  11.671  44.759
##
```

```
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)    
## (Intercept) 155.87051   10.64180  14.647         < 2e-16 ***
## ï..age       -0.86973    0.12444  -6.989 0.0000000000185 ***
## cp            2.28549    1.17251   1.949         0.05221 .  
## trestbps      0.12711    0.06288   2.021         0.04415 *  
## exang        -9.23031    2.57770  -3.581         0.00040 ***
## slope         8.32783    1.83600   4.536 0.0000083551781 ***
## target        7.37733    2.59998   2.837         0.00486 ** 
## chol          0.03843    0.02072   1.854         0.06467 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.12 on 295 degrees of freedom
## Multiple R-squared:  0.3886, Adjusted R-squared:  0.3741 
## F-statistic: 26.79 on 7 and 295 DF,  p-value: < 2.2e-16
```

Remove chol: cholesterol

```
modl <- update(modl, .~. -chol)
summary(modl)
```

```
## 
## Call:
## lm(formula = thalach ~ ï..age + cp + trestbps + exang + slope + 
##     target, data = heart)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -61.079  -9.939   2.033  11.434  45.323 
## 
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)    
## (Intercept) 161.75331   10.19977  15.859         < 2e-16 ***
## ï..age       -0.82749    0.12284  -6.736 0.0000000000846 ***
## cp            2.17615    1.17584   1.851        0.065205 .  
## trestbps      0.13563    0.06297   2.154        0.032063 *  
## exang        -9.10341    2.58738  -3.518        0.000502 ***
## slope         8.51057    1.84089   4.623 0.0000056542036 ***
## target        7.32599    2.61053   2.806        0.005343 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 18.2 on 296 degrees of freedom
## Multiple R-squared:  0.3815, Adjusted R-squared:  0.3689 
## F-statistic: 30.43 on 6 and 296 DF,  p-value: < 2.2e-16
```

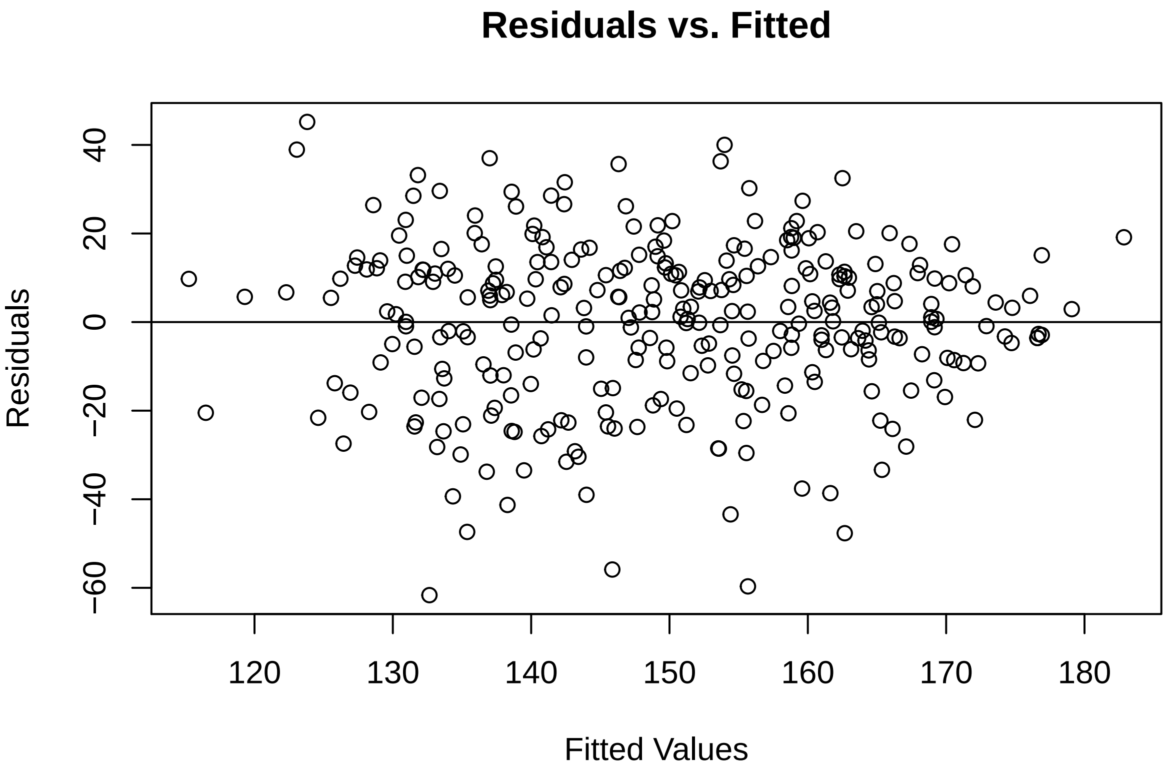
Remove cp: chest pain type

```
modl <- update(modl, .~. -cp)
summary(modl)
```

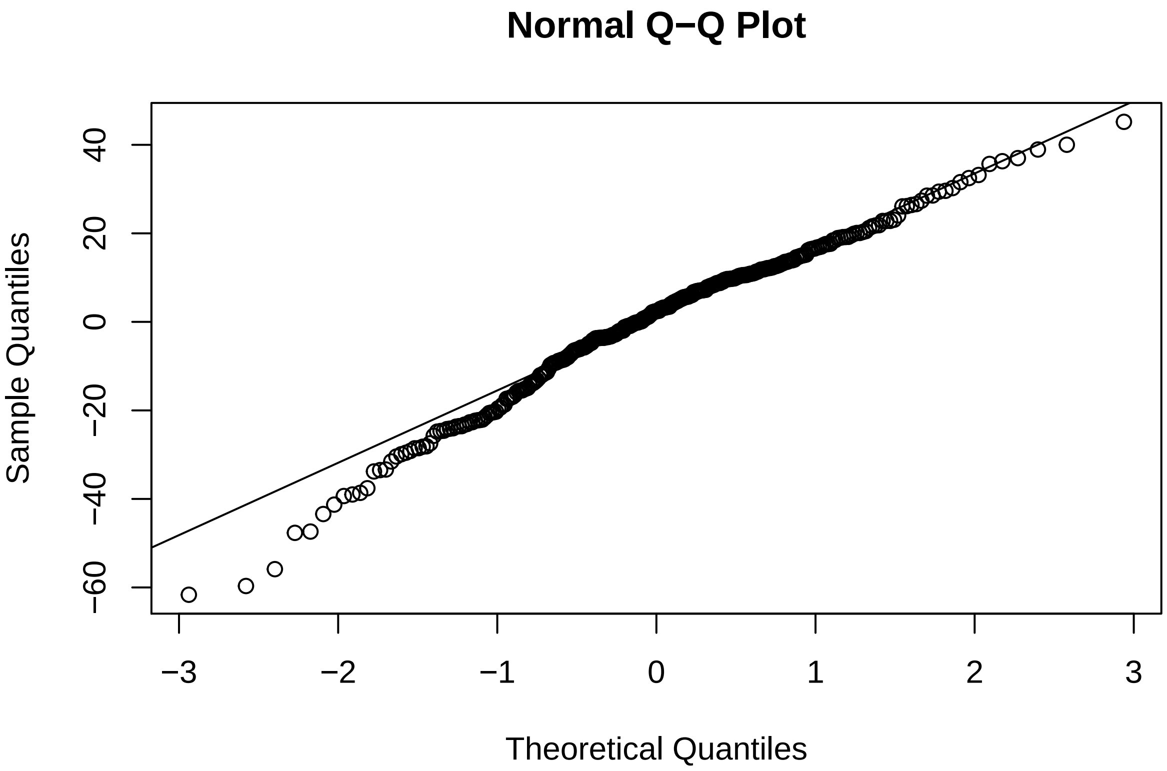
```
##
## Call:
## lm(formula = thalach ~ ï..age + trestbps + exang + slope + target,
##     data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -61.647 -10.178   2.408  11.865  45.190
##
## Coefficients:
##              Estimate Std. Error t value        Pr(>|t|)
## (Intercept) 161.87607   10.24111  15.806         < 2e-16 ***
## ï..age       -0.82850    0.12334  -6.717 0.0000000000944 ***
## trestbps      0.14983    0.06276   2.387        0.017598 *
## exang       -10.35249    2.50799  -4.128 0.0000476278629 ***
## slope         8.30084    1.84488   4.499 0.0000097920741 ***
## target        8.92310    2.47380   3.607        0.000363 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.27 on 297 degrees of freedom
## Multiple R-squared:  0.3743, Adjusted R-squared:  0.3638
## F-statistic: 35.54 on 5 and 297 DF,  p-value: < 2.2e-16
```

Plot

```r
plot(modl$fitted.values, modl$residuals, xlab="Fitted Values", ylab="Residuals", main="Residuals vs. Fi
abline(h=0)
```



```r
qqnorm(modl$residuals)
qqline(modl$residuals)
```



Conclusion

We included 5 variables in the final model that are used to predict the target variable, thalach: max heart rate achieved.

The results show: - The residuals median is 2.408 which is not very close to zero, indicating poor fit. - As min and max are not similar values, this indicates variability is not consistent. - The `R^2` values explains only 37.43% of variability in the data - which again indicate poor fit. - The normal q-q plot shows us that the residuals do not follow a normal distribution.

So we may conclude this is as poorly fitted model.