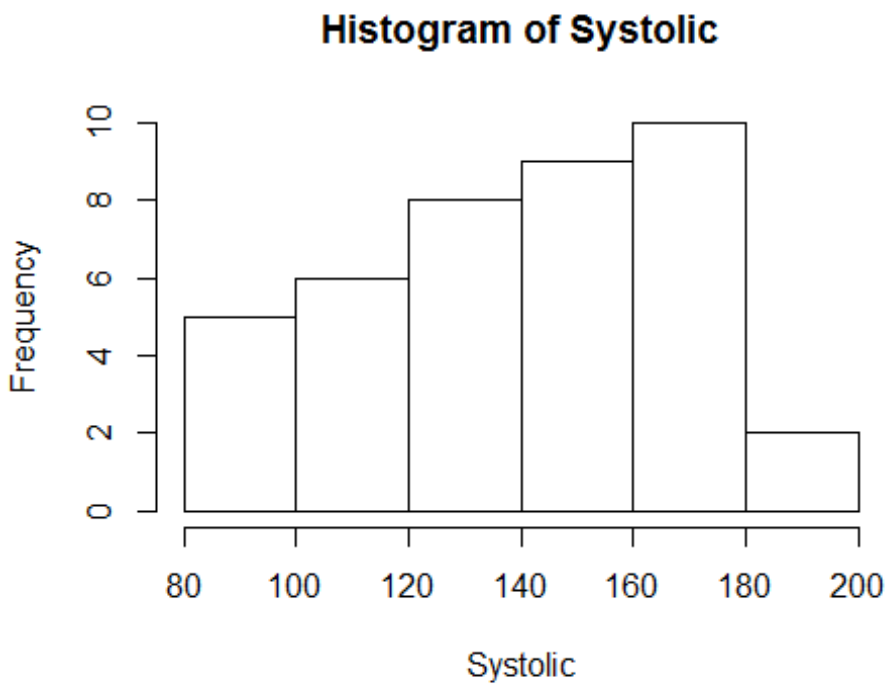
Regression Example

W. Grossmann

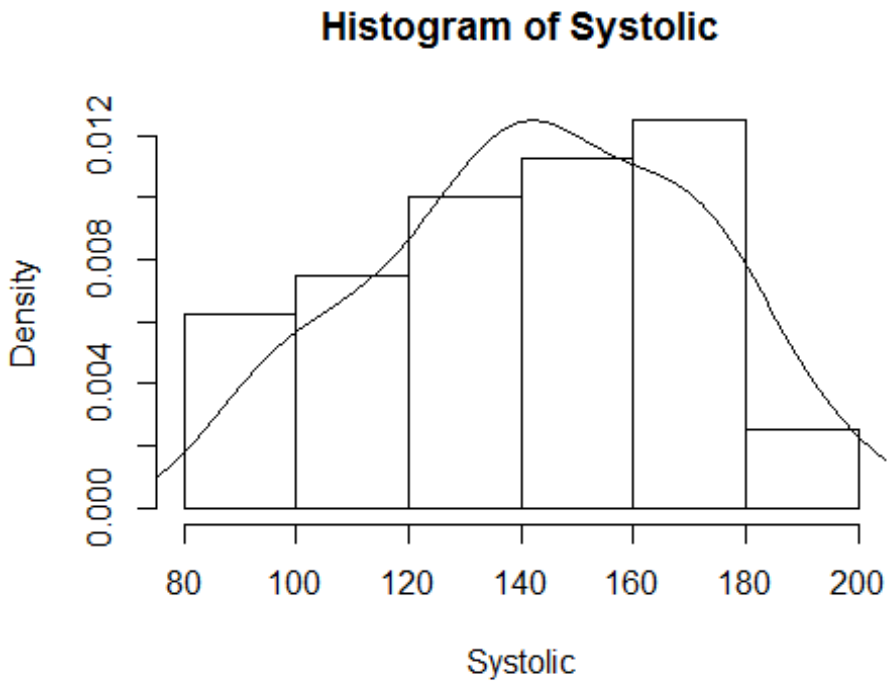
```
#=======================================
# Regression Demo-example
#=======================================

# Step 1: Import the data.
# The easiest way is using the "Import Dataset"-Tab
# in RStudio.
# Otherwise you have to define first the directory
# where you have stored the data and read the data
# using the following command for ","as separator and "." for decimals:
bloodpressure <- read.csv("bloodpressure.csv")
attach(bloodpressure)
summary(bloodpressure)
```

```
##     Systolic       Diastolic        Gender        Age
##  Min.   : 90.0   Min.   : 67.0   female:20   Min.   :18.00
##  1st Qu.:120.0   1st Qu.: 78.0   male  :20   1st Qu.:35.00
##  Median :143.5   Median : 90.0               Median :55.00
##  Mean   :143.6   Mean   : 88.6               Mean   :52.24
##  3rd Qu.:168.0   3rd Qu.: 98.0               3rd Qu.:68.00
##  Max.   :200.0   Max.   :110.0               Max.   :88.00
##                                              NA's   :3
##    WeightInKg       HeightInCm    EyeColour
##  Min.   : 54.00   Min.   :140.0   blue :17
##  1st Qu.: 60.00   1st Qu.:155.0   brown:18
##  Median : 72.00   Median :162.5   green: 1
##  Mean   : 73.08   Mean   :163.6   grey : 4
##  3rd Qu.: 84.50   3rd Qu.:170.0
##  Max.   :100.00   Max.   :193.0
##
```

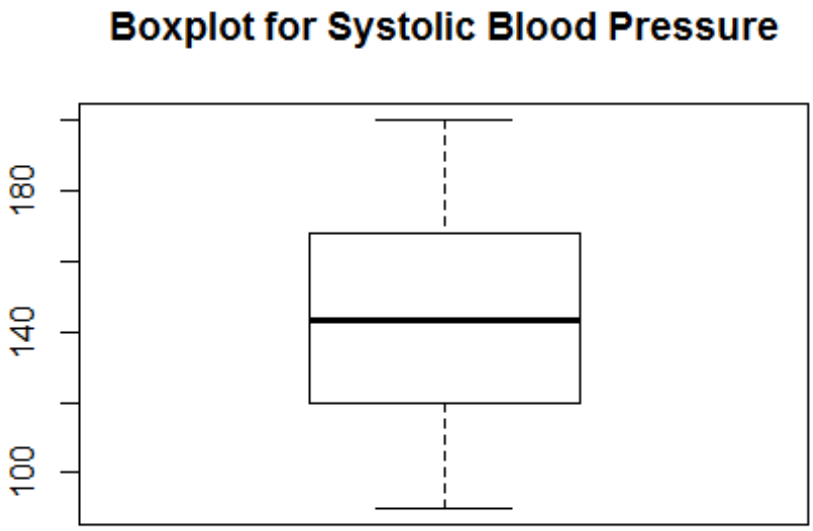
```
#================================================
# Step 1: Data Understanding
# ----------------------------------------------
#
#----------------------------------------------
# Display of Distributions
#----------------------------------------------
#Histogram
hist(Systolic)
```



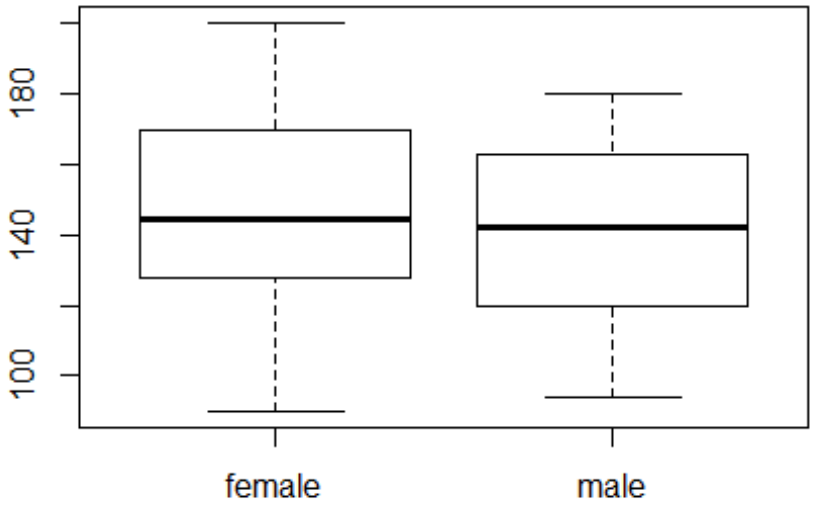

```
#Histogram with smoothed density
hist(Systolic, freq=FALSE)
lines(density(Systolic))
```



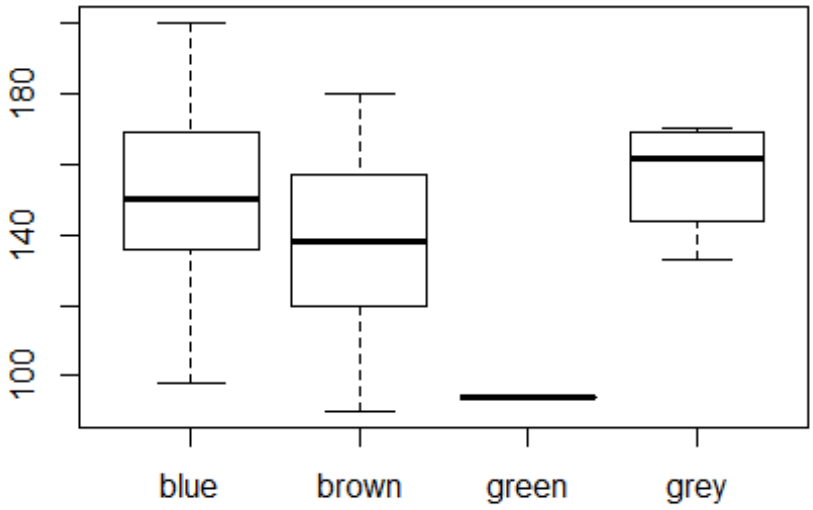
```r
#---------------------------------------
#Boxplot
#---------------------------------------
boxplot(Systolic, main = "Boxplot for Systolic Blood Pressure")
```




```r
#---------------------------------------
#Grouped Boxplots
#---------------------------------------
boxplot(Systolic~Gender)
```

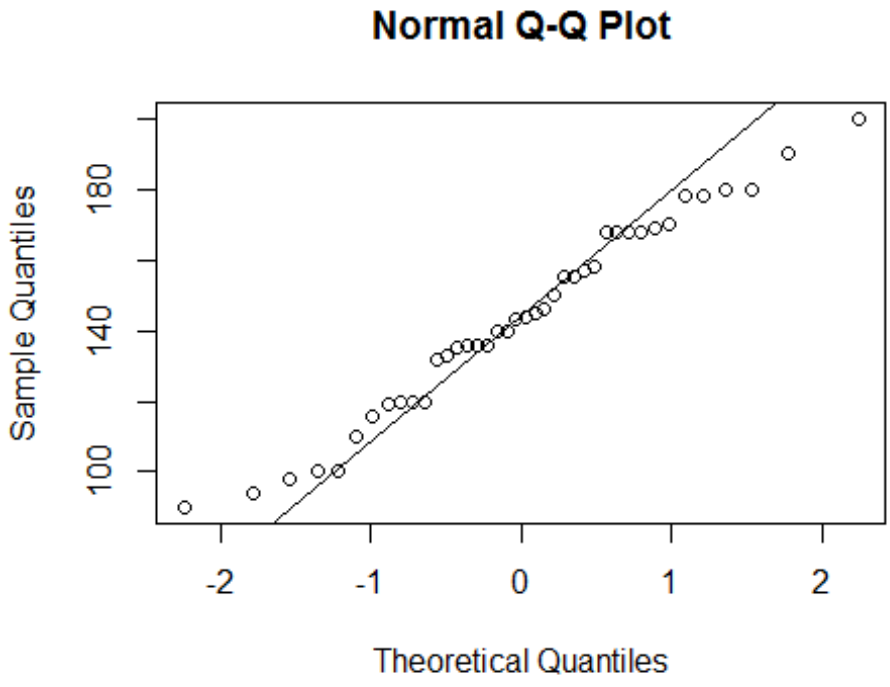



```
boxplot(Systolic~ EyeColour)
```




```
#-----------------------------------------
#Tables for qualitative variables
```

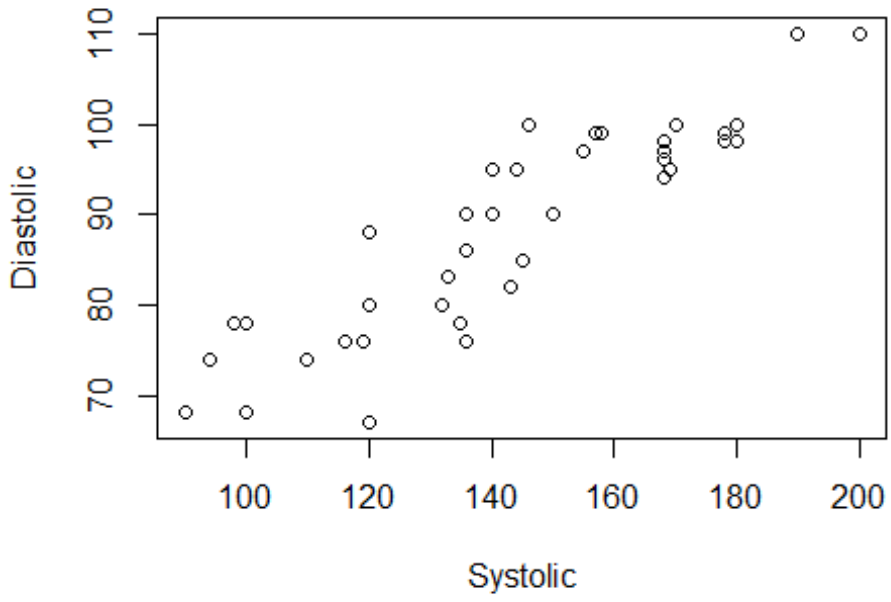
```r
#-----------------------------------------

table(EyeColour)

## EyeColour
##  blue brown green  grey
##    17    18     1     4

table(EyeColour,Gender)

##           Gender
## EyeColour female male
##     blue      10    7
##     brown      9    9
##     green      0    1
##     grey       1    3

#---------------------------------------
#Assessing the normal distribution
#---------------------------------------
qqnorm(Systolic)
qqline(Systolic)
```




```r
#---------------------------------------
#Comparison of Gender Groups
#---------------------------------------
t.test(Systolic~Gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  Systolic by Gender
## t = 0.49693, df = 36.126, p-value = 0.6223
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.70949  22.60949
## sample estimates:
## mean in group female   mean in group male
##               145.85                141.40
```

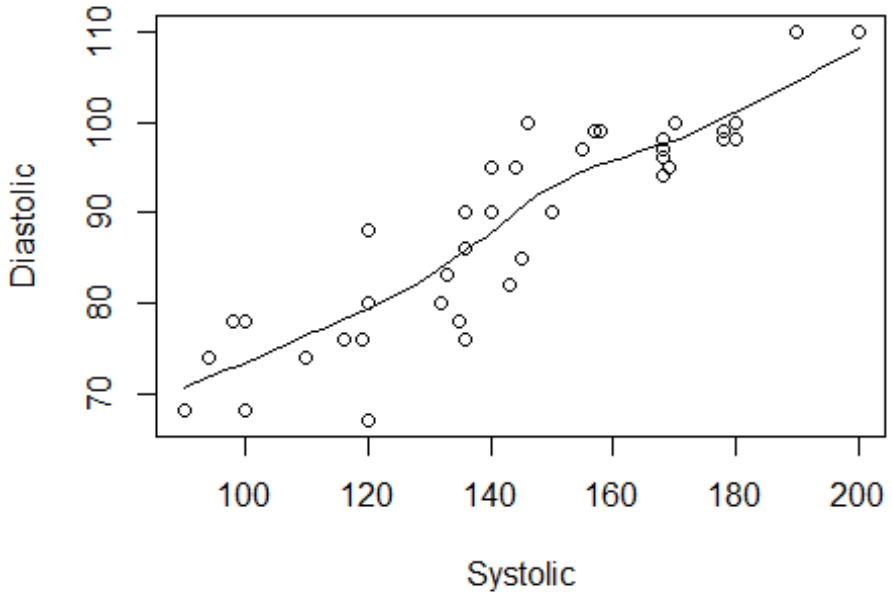
```
#----------------------------------------
# Correlation and Scatterplots
#----------------------------------------
plot(Systolic, Diastolic)
```




```
#----------------------------------------
#Correlation
#----------------------------------------
cor(Systolic, Diastolic)
```

```
## [1] 0.8972328
```

```
#----------------------------------------
# More advance Scatterplot with smoothing curve
#----------------------------------------
scatter.smooth(Systolic, Diastolic)
```




```
#----------------------------------------
#Scatterplot with a regression line
#----------------------------------------
plot(Systolic~Diastolic)
abline(lm(Systolic~Diastolic))



#----------------------------------------
#Correlation Matrix
#----------------------------------------
# Define Data matrix for all quantitative variables:
var_quant<-cbind(Systolic, Diastolic,Age, WeightInKg,HeightInCm)

cor(var_quant) # if data are missing no computation
```

```
##              Systolic  Diastolic Age WeightInKg  HeightInCm
## Systolic    1.0000000  0.8972328  NA 0.35402489 -0.39429205
## Diastolic   0.8972328  1.0000000  NA 0.37309486 -0.49002570
## Age                NA         NA   1         NA          NA
## WeightInKg  0.3540249  0.3730949  NA 1.00000000  0.02126117
## HeightInCm -0.3942921 -0.4900257  NA 0.02126117  1.00000000
```

```
# If you have missing values use the specification:
cor(var_quant, use = "pairwise.complete.obs")
```

```
##              Systolic  Diastolic        Age WeightInKg  HeightInCm
## Systolic    1.0000000  0.8972328  0.8625068 0.35402489 -0.39429205
```

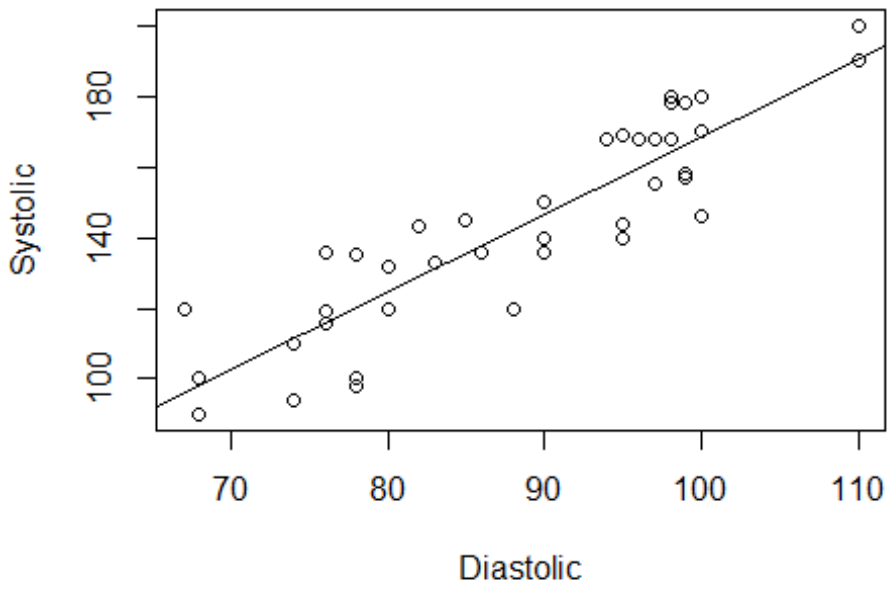
```
## Diastolic   0.8972328  1.0000000  0.7719029 0.37309486 -0.49002570
## Age         0.8625068  0.7719029  1.0000000 0.11698853 -0.29091822
## WeightInKg  0.3540249  0.3730949  0.1169885 1.00000000  0.02126117
## HeightInCm -0.3942921 -0.4900257 -0.2909182 0.02126117  1.00000000
```

```
data1<-na.omit(bloodpressure)
attach(data1)
```

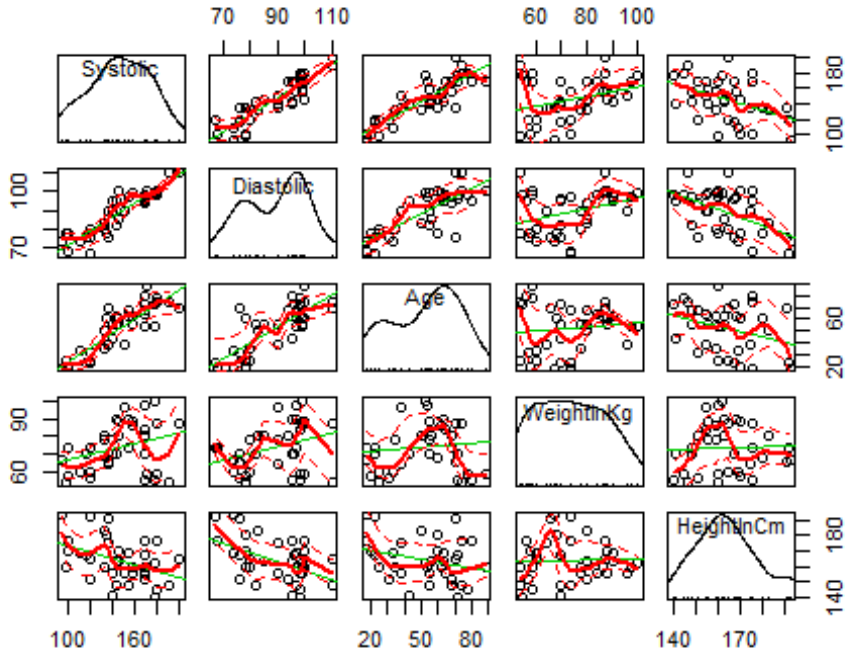
```
## Die folgenden Objekte sind maskiert von bloodpressure:
##
##     Age, Diastolic, EyeColour, Gender, HeightInCm, Systolic,
##     WeightInKg
```

```
#-------------------------------------------
# Scatterplotmatrix
#-------------------------------------------
# For Scatterplotmatrices you have first of all to
# install the library car
# Select in the bottom right pane "Packages" and afterwards
# "Install Packages". Select package "car"

library(car)
```

```
## Warning: package 'car' was built under R version 3.2.5
```




```
scatterplotMatrix(var_quant)
```




```
#=================================================
# Step 2: Model formulation and parameter estimation
#=================================================
# Model with all variables:
mod1<-lm(Systolic~.,data = bloodpressure)
mod1

##
## Call:
## lm(formula = Systolic ~ ., data = bloodpressure)
##
## Coefficients:
##    (Intercept)       Diastolic      Gendermale             Age
##        1.01747         1.20560         1.88302         0.64040
##     WeightInKg      HeightInCm  EyeColourbrown  EyeColourgreen
##        0.13868         -0.03425        -1.34265       -12.32320
##  EyeColourgrey
##       -7.52442

# Model with using only quantitative variables:
mod2<-lm(Systolic~Diastolic+Age+WeightInKg+HeightInCm,data = bloodpressure)


# Model with using quantitative variables and Gender:
mod3<-lm(Systolic~Diastolic+Age+WeightInKg+HeightInCm + Gender,data = bloodpr
essure)
```

```r
#================================================
# Step 3: Modell Assessment
#================================================
#Overall Fit and Test for Parameters

summary(mod1)
```

```
##
## Call:
## lm(formula = Systolic ~ ., data = bloodpressure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6762  -4.3917   0.6043   5.6027  17.9090
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.01747   43.68317   0.023 0.981583
## Diastolic        1.20560    0.31895   3.780 0.000756 ***
## Gendermale       1.88302    4.29790   0.438 0.664657
## Age              0.64040    0.15175   4.220 0.000232 ***
## WeightInKg       0.13868    0.17333   0.800 0.430405
## HeightInCm      -0.03425    0.17915  -0.191 0.849752
## EyeColourbrown  -1.34265    3.92915  -0.342 0.735116
## EyeColourgreen -12.32320   11.92490  -1.033 0.310260
## EyeColourgrey   -7.52442    6.22543  -1.209 0.236902
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.23 on 28 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.8913, Adjusted R-squared:  0.8603
## F-statistic: 28.71 on 8 and 28 DF,  p-value: 1.58e-11
```

```r
summary(mod2)
```

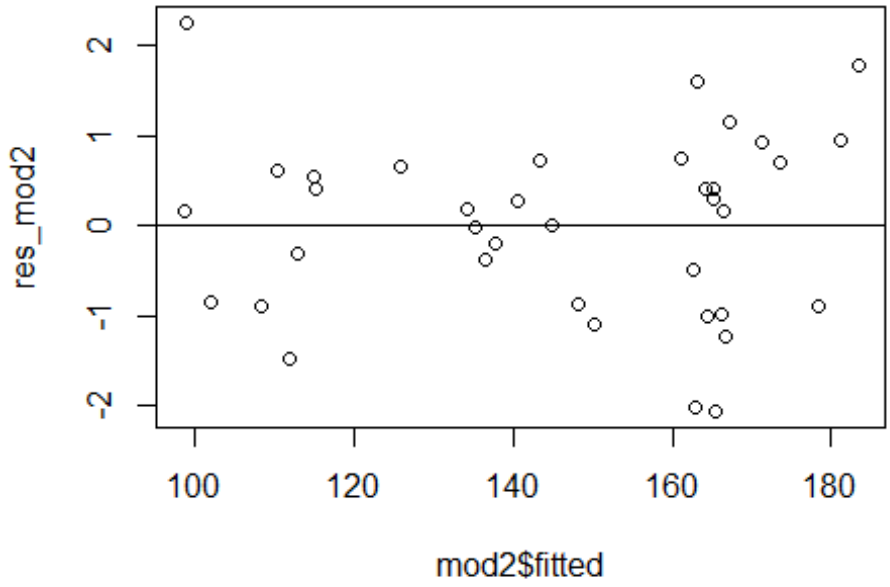
```
##
## Call:
## lm(formula = Systolic ~ Diastolic + Age + WeightInKg + HeightInCm,
##     data = bloodpressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.362  -8.193   1.479   6.194  21.105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.3758    37.1020   0.387 0.700975
## Diastolic     1.1265     0.2902   3.882 0.000487 ***
## Age           0.6459     0.1379   4.683 4.98e-05 ***
```

```
## WeightInKg    0.2103     0.1403   1.499 0.143717
## HeightInCm   -0.1113     0.1502  -0.741 0.464015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.954 on 32 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.8825, Adjusted R-squared:  0.8678
## F-statistic: 60.09 on 4 and 32 DF,  p-value: 1.993e-14
```

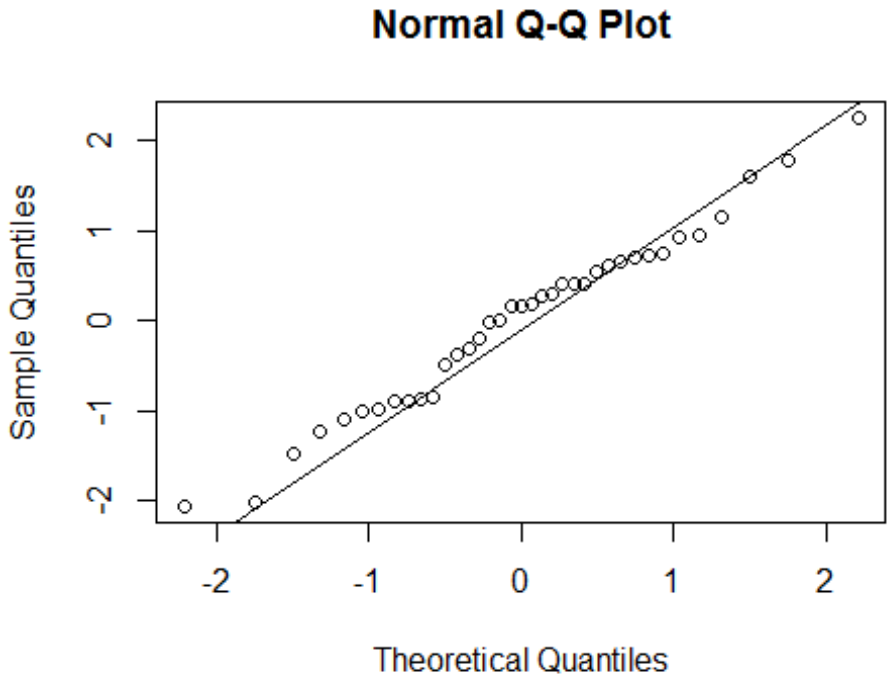
```
summary(mod3)
```

```
##
## Call:
## lm(formula = Systolic ~ Diastolic + Age + WeightInKg + HeightInCm +
##     Gender, data = bloodpressure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.307  -8.177   1.444   6.168  21.101
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.43077   37.79996   0.382 0.705239
## Diastolic    1.12772    0.30090   3.748 0.000732 ***
## Age          0.64569    0.14039   4.599 6.75e-05 ***
## WeightInKg   0.20874    0.16366   1.275 0.211635
## HeightInCm  -0.11180    0.15445  -0.724 0.474582
## Gendermale   0.07913    4.04518   0.020 0.984518
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 31 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.8825, Adjusted R-squared:  0.8636
## F-statistic: 46.57 on 5 and 31 DF,  p-value: 1.664e-13
```

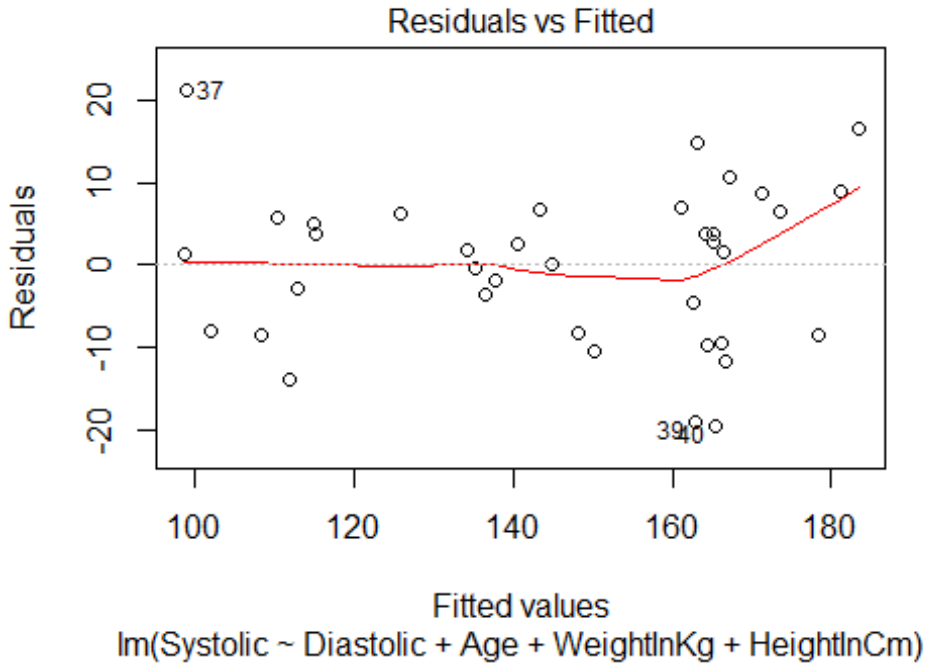
```
#-------------------------------------------------
# Graphical Model Assessment for mod2
#-------------------------------------------------
#Plot for residuals
res_mod2<-(mod2$residuals - mean(mod2$residuals))/sd(mod2$residuals)
plot(res_mod2~mod2$fitted)
abline(h=0)
```

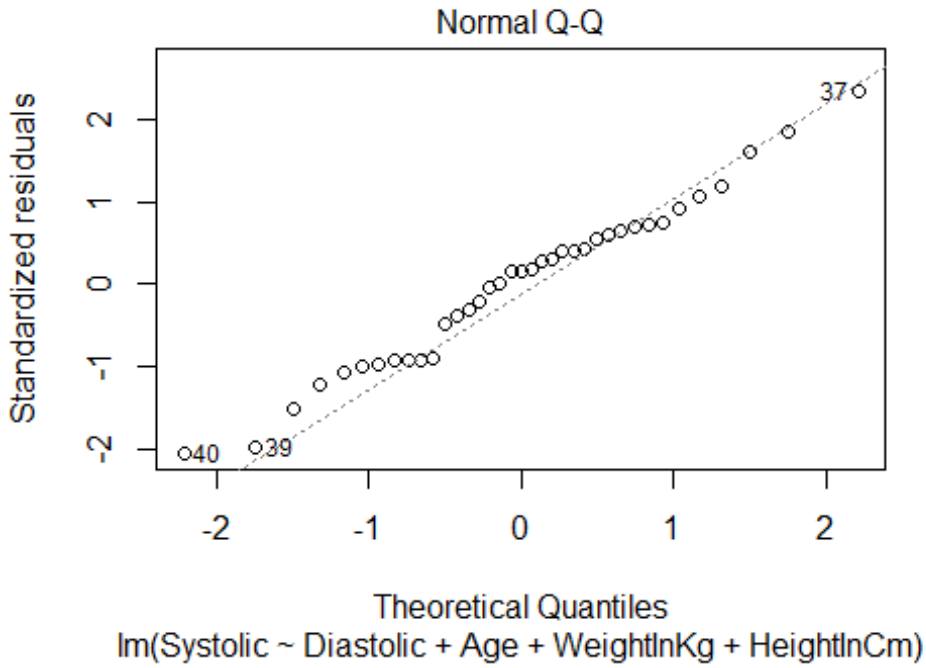


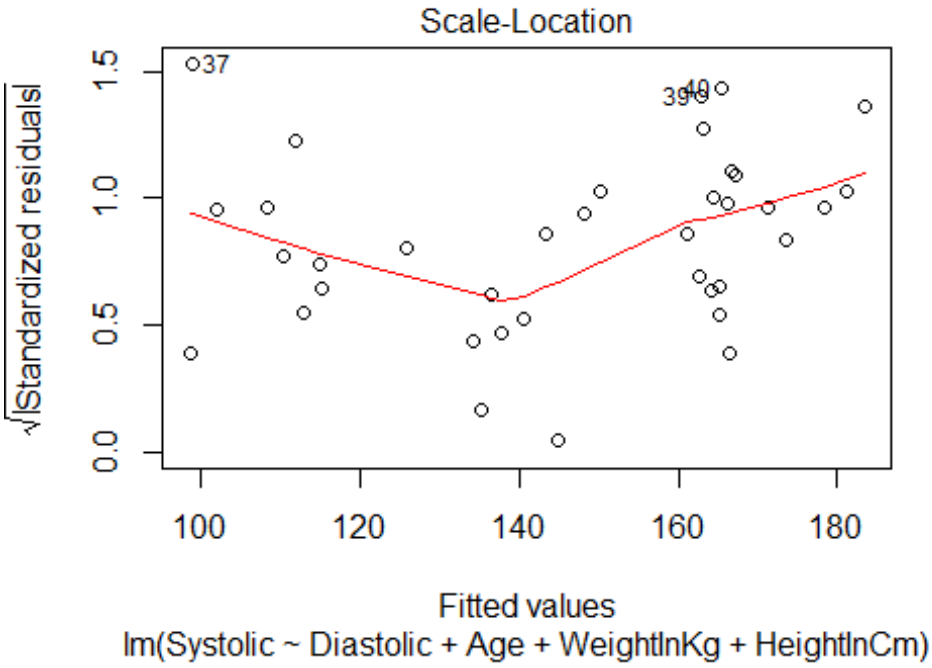

```r
#Checking normal distribution for residuals
qqnorm(res_mod2)
qqline(res_mod2)
```

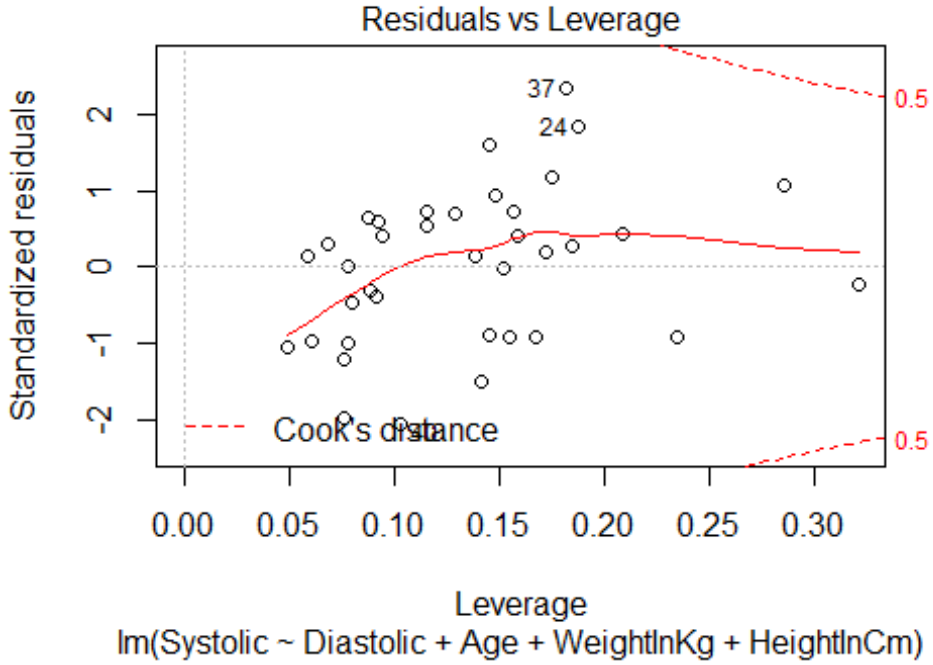


```r
#More diagnostics
plot(mod2)
```










```
#=====================================================
# Prediction for  new values of explanatory variables
#=====================================================
# If you want to predict for many cases the best thing is
# to create a new data set with the explanatoiry variables
```

```r
#for which you want to make prediction
# We use here the data newdata.csv.

newdata <- read.csv("newdata.csv", comment.char="#")
newdata
```

```
##   Diastolic Gender Age WeightInKg HeightInCm EyeColour
## 1        80   male  25         60        165     brown
## 2        76 female  31         57        151      blue
```

```r
pred1<-predict(mod1,newdata)
pred1
```

```
##        1        2
## 116.6853 115.2284
```

```r
pred2<-predict(mod2,newdat=newdata)
pred2
```

```
##        1        2
## 114.8938 115.1906
```

```r
#================================================
# Model Selection
#================================================
drop1(mod1)
```

```
## Single term deletions
##
## Model:
## Systolic ~ Diastolic + Gender + Age + WeightInKg + HeightInCm +
##     EyeColour
##            Df Sum of Sq    RSS    AIC
## <none>                  2932.4 179.79
## Diastolic   1   1496.32 4428.7 193.04
## Gender      1     20.10 2952.5 178.04
## Age         1   1865.26 4797.7 196.00
## WeightInKg  1     67.04 2999.4 178.62
## HeightInCm  1      3.83 2936.2 177.84
## EyeColour   3    238.00 3170.4 176.68
```

```r
mod4<-lm(Systolic~Diastolic+Age)
summary(mod4)
```

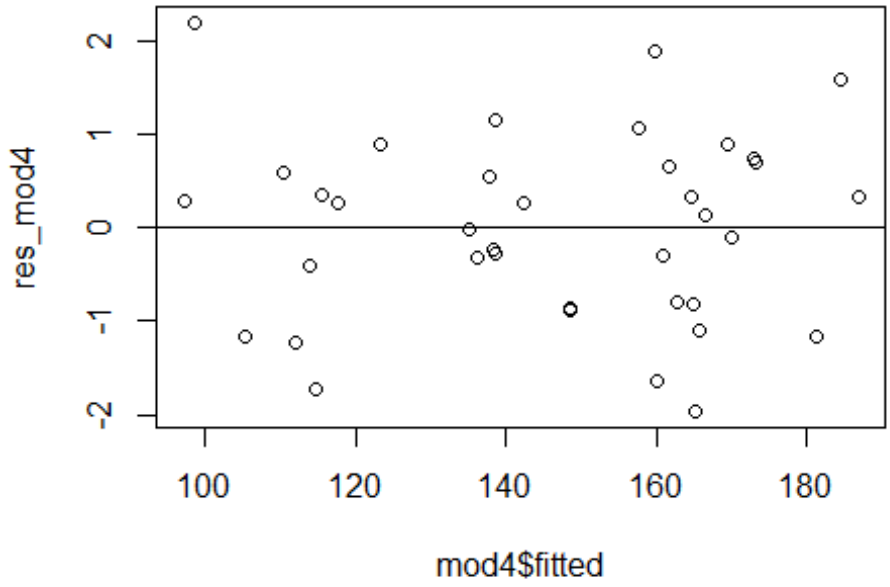
```
##
## Call:
## lm(formula = Systolic ~ Diastolic + Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.084  -7.882   1.402   6.450  21.377
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.0572    15.9605  -0.380    0.707
## Diastolic     1.3631     0.2302   5.920 1.09e-06 ***
## Age           0.5805     0.1309   4.434 9.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10 on 34 degrees of freedom
## Multiple R-squared:  0.8739, Adjusted R-squared:  0.8665
## F-statistic: 117.8 on 2 and 34 DF,  p-value: 5.152e-16
```

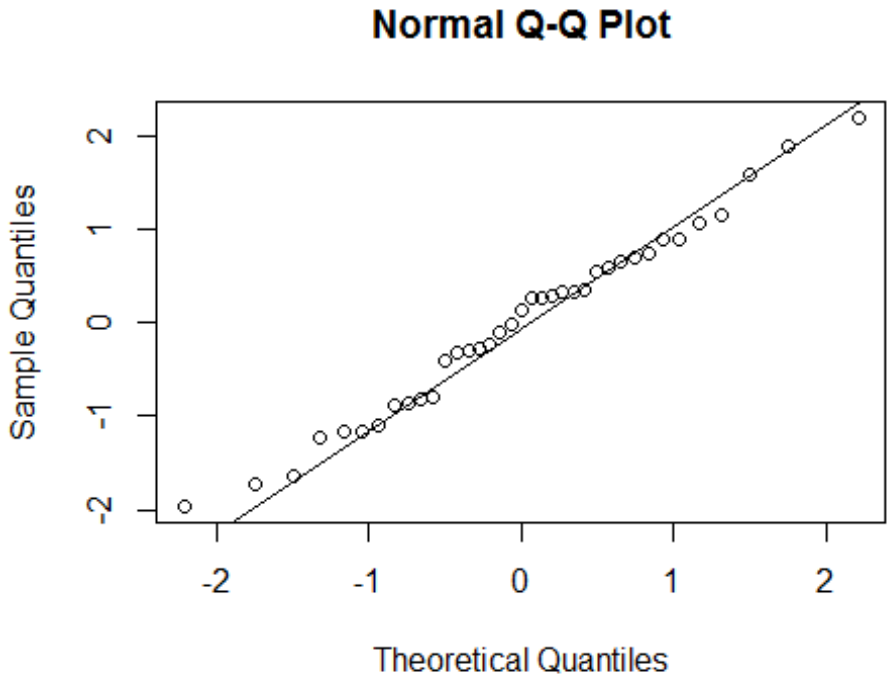
```
drop1(mod4)
```

```
## Single term deletions
##
## Model:
## Systolic ~ Diastolic + Age
##           Df Sum of Sq    RSS    AIC
## <none>                 3402.7 173.29
## Diastolic  1    3507.6 6910.3 197.50
## Age        1    1967.7 5370.3 188.18
```

```
#-----------------------------------------------
# Checking Residuals for mod4
#-----------------------------------------------
#Plot for residuals:
res_mod4<-(mod4$residuals - mean(mod4$residuals))/sd(mod4$residuals)
plot(res_mod4~mod4$fitted)
abline(h=0)
```




```r
#Checking normal distribution for residuals:
qqnorm(res_mod4)
qqline(res_mod4)
```



```r
#=============================================
# Automatic model selection
#=============================================
#Backward selection
reduce.mod<-step(mod1, direction = "backward")
```

```
## Start:  AIC=179.79
## Systolic ~ Diastolic + Gender + Age + WeightInKg + HeightInCm +
##     EyeColour
##
##              Df Sum of Sq    RSS    AIC
## - EyeColour   3    238.00 3170.4 176.68
## - HeightInCm  1      3.83 2936.2 177.84
## - Gender      1     20.10 2952.5 178.04
## - WeightInKg  1     67.04 2999.4 178.62
## <none>                    2932.4 179.79
## - Diastolic   1   1496.32 4428.7 193.04
## - Age         1   1865.26 4797.7 196.00
##
## Step:  AIC=176.68
## Systolic ~ Diastolic + Gender + Age + WeightInKg + HeightInCm
##
##              Df Sum of Sq    RSS    AIC
## - Gender      1      0.04 3170.4 174.68
## - HeightInCm  1     53.59 3224.0 175.30
## - WeightInKg  1    166.36 3336.8 176.57
## <none>                    3170.4 176.68
## - Diastolic   1   1436.50 4606.9 188.50
## - Age         1   2163.46 5333.9 193.92
##
## Step:  AIC=174.68
## Systolic ~ Diastolic + Age + WeightInKg + HeightInCm
##
##              Df Sum of Sq    RSS    AIC
## - HeightInCm  1     54.42 3224.9 173.31
## <none>                    3170.4 174.68
## - WeightInKg  1    222.58 3393.0 175.19
## - Diastolic   1   1493.15 4663.6 186.96
## - Age         1   2172.80 5343.2 191.99
##
## Step:  AIC=173.31
## Systolic ~ Diastolic + Age + WeightInKg
##
##              Df Sum of Sq    RSS    AIC
## - WeightInKg  1    177.82 3402.7 173.29
## <none>                    3224.9 173.31
## - Age         1   2136.24 5361.1 190.11
## - Diastolic   1   2494.94 5719.8 192.51
##
## Step:  AIC=173.29
## Systolic ~ Diastolic + Age
```

```
##
##             Df Sum of Sq    RSS    AIC
## <none>                    3402.7 173.29
## - Age        1    1967.7 5370.3 188.18
## - Diastolic  1    3507.6 6910.3 197.50

#-----------------------------------------------
# Forward Selection
#-----------------------------------------------
mod.null<- lm(Systolic~1)
mod.forw<-step(mod.null, direction = "forward",
               scope = ~Diastolic + Age+WeightInKg+Gender, trace= 0 )
summary(mod.forw)

##
## Call:
## lm(formula = Systolic ~ Diastolic + Age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.084  -7.882   1.402   6.450  21.377
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.0572    15.9605  -0.380    0.707
## Diastolic     1.3631     0.2302   5.920 1.09e-06 ***
## Age           0.5805     0.1309   4.434 9.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10 on 34 degrees of freedom
## Multiple R-squared:  0.8739, Adjusted R-squared:  0.8665
## F-statistic: 117.8 on 2 and 34 DF,  p-value: 5.152e-16

#================================================
#Testing the model with independent data
#================================================

# Select randomly 70% of the data (in our case approximately 30)

set.seed(1)
## fixing the seed value for the random selection guarantees the
## same results in repeated runs
N=length(Systolic)
n1<-30 # Size of training data
n2<-N-n1 # Size of test data
N

## [1] 37

train<-sample(1:N,n1) #Selection of training data
train
```

```
##  [1] 10 14 21 31  7 29 30 20 19  2  6  5 18 37 25 11 16 34  8 36 35  4 24
## [24] 28 17 26  1 13 32  3
```

```
bloodpressure[train,]
```

```
##    Systolic Diastolic Gender Age WeightInKg HeightInCm EyeColour
## 10      143        82 female  55         70        148     brown
## 14      140        90   male  55         78        156     brown
## 21      168        96   male  72         70        167      grey
## 31      140        95 female  43         70        141      blue
## 7        90        68 female  NA         54        169     brown
## 29      178        99   male  53        100        162      blue
## 30       98        78 female  25         54        160      blue
## 20      155        97   male  63         90        168      grey
## 19      133        83   male  50         80        183      grey
## 2       119        76 female  31         57        151     brown
## 6       100        78 female  20         58        170     brown
## 5       132        80   male  35         78        159     brown
## 18       94        74   male  18         67        193     green
## 37      120        67   male  23         74        192      blue
## 25      100        68   male  18         74        176      blue
## 11      116        76   male  22         70        167     brown
## 16      120        88   male  NA         76        154     brown
## 34      180       100   male  74         87        176      blue
## 8       168        98   male  64         84        155     brown
## 36      168        97   male  54         97        160      blue
## 35      169        95 female  80         56        176      blue
## 4       178        98 female  72         55        140     brown
## 24      200       110 female  70         87        165      blue
## 28      150        90 female  38         96        155      blue
## 17      136        90 female  38         64        175     brown
## 26      110        74 female  33         60        169      blue
## 1       120        80   male  25         60        165     brown
## 13      168        94 female  68         80        151     brown
## 32      136        76 female  70         66        191      blue
## 3       145        85   male  56         80        163     brown
```

```
#-------------------------------------
# Regression with training data
#-------------------------------------
m_train<-lm(Systolic~Diastolic+Age,data=bloodpressure[train,])
summary(m_train)
```

```
##
## Call:
## lm(formula = Systolic ~ Diastolic + Age, data = bloodpressure[train,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.877  -4.985   1.275   4.852  22.922
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.9562    17.4245  -1.031 0.312635
## Diastolic     1.5067     0.2533   5.947 3.29e-06 ***
## Age           0.6126     0.1409   4.346 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.313 on 25 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.9028, Adjusted R-squared:  0.895
## F-statistic:   116 on 2 and 25 DF,  p-value: 2.23e-13
```

```
#----------------------------------------
#Prediction for the test cases and comparison
#----------------------------------------
pred_test<-predict(m_train,bloodpressure[-train,])
pred_test
```

```
##        9       12       15       22       23       27       33       38
## 177.4750 136.3159 169.1808 186.6138 193.1046       NA 169.8428 164.8929
##       39       40
## 163.7667 169.4623
```

```
Syst<-Systolic[-train]

diff<-Syst-pred_test
```

```
## Warning in Syst - pred_test: Länge des längeren Objektes
##       ist kein Vielfaches der Länge des kürzeren Objektes
```

```
#-------------------------------------------
# Basic Statistics for evaluation
#-------------------------------------------
Mu<-mean(diff)
MSE<-(sum(diff**2)/n2)
RMSE<-sqrt(MSE)
cbind(Mu,MSE,RMSE)
```

```
##      Mu MSE RMSE
## [1,] NA  NA   NA
```

```
SD_Train<-sd(m_train$residuals)
SD_all<-sd(mod4$residuals)
cbind(RMSE,SD_Train,SD_all)
```

```
##      RMSE SD_Train  SD_all
## [1,]   NA 8.961272 9.72208
```

```
#----------------------------------------
#Plotting Differences
#----------------------------------------
names<-labels(diff)
```

```
plot(names, diff, type = "n")
text(names, diff, labels = names)
abline(h=0)
```