Final Model(?)

Edward J. Lee

2025-04-05

Usual Data Cleaning

```
library(NHANES)  # NHANES dataset
library(dplyr)   # Data wrangling
library(ggplot2) # Visualization
library(car)     # Multicollinearity check (VIF)
library(ggResidpanel) # Advanced diagnostic plots
library(knitr) #for kable
library(gridExtra) #for scatterplot matrix

# if you don't have it installed, do install_packages("NHANES")
data("NHANES")
nrow(NHANES) #10,000 observations
```

```
## [1] 10000
```

```
# remove babies (ages 0-3)
nhanes_filtered <- NHANES %>% filter(Age > 20)
nrow(nhanes_filtered) #7094 observations
```

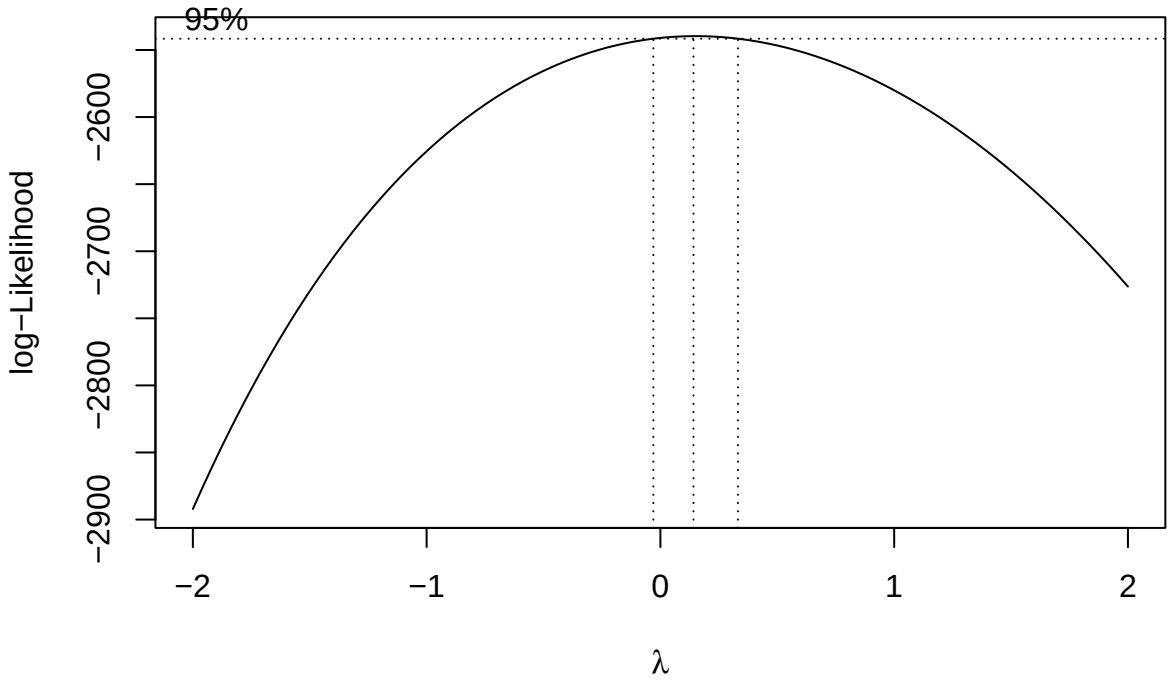
```
## [1] 7094
```

```
# remove NA entries and only select columns of interest
nhanes_data <- nhanes_filtered %>%
  dplyr::select(Height, Age, Weight, BPSysAve, BPDiaAve, BPDia1, BPDia2, BPDia3, BPSys1, BPSys2, BPSys3
                TotChol, SmokeNow, PhysActiveDays) %>%
  na.omit() %>%
  dplyr::filter(
    Height > 0, Weight > 0, BPDia1 > 10, BPDia2 > 10, BPDia3 > 10, BPSys1 > 10, BPSys2 > 10, BPSys3 > 1
                TotChol, SmokeNow, PhysActiveDays)

# categorical predictors
nhanes_data$SmokeNow <- as.factor(nhanes_data$SmokeNow)
nhanes_data <- data.frame(nhanes_data)

# fit the model
model <- lm(TotChol ~ Age + Weight + Height + BPSysAve + BPDiaAve + SmokeNow +
              PhysActiveDays,
```

```
            data = nhanes_data)

n <- nrow(nhanes_data)
```

Box-Cox Transformation and Polynomial Term

```
#POLYNOMIAL "AGE" TERM
pb_data <- nhanes_data %>%
  dplyr::select(Height, Age, Weight, BPSysAve, BPDiaAve,
                TotChol, SmokeNow, PhysActiveDays) %>%
  mutate(pb.Age2 = Age^2)

pb_model <- lm(TotChol~Age+pb.Age2+Height+Weight+BPSysAve+BPDiaAve+
                 SmokeNow+PhysActiveDays, data=pb_data)

#BOX COX TRANSFORMATION
library(MASS) #For BOXCOX

pb.b <- boxcox(pb_model)
```



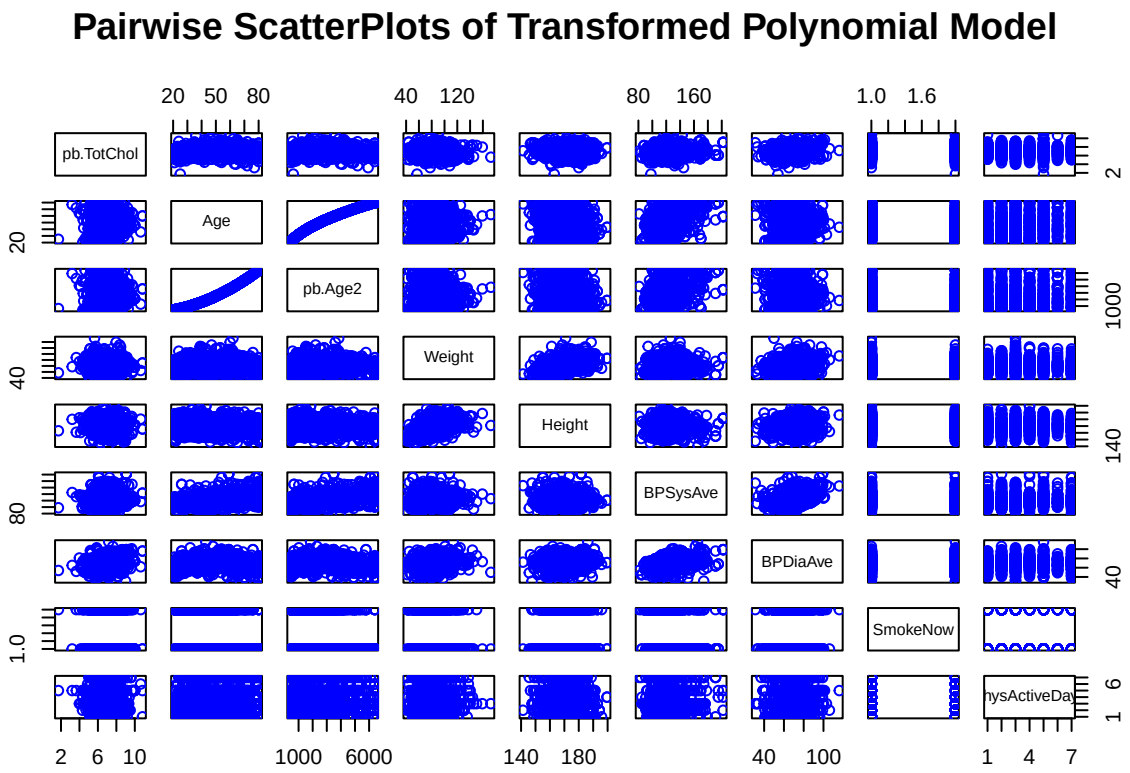

```
pb.lambda <- pb.b$x[which.max(pb.b$y)]

pb.log_product <- sum(log(pb_data$TotChol))
```

```r
pb.geo_mean <- exp(pb.log_product/n)

pb.TotChol <- pb.geo_mean^(1-pb.lambda)*(pb_data$TotChol^pb.lambda - 1)/pb.lambda


p.BXCX.frame <- pb_data %>%
  dplyr::select(-TotChol) %>%
  mutate(pb.TotChol = pb.TotChol)

p.BXCX.model <- lm(pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve +
                         BPDiaAve + SmokeNow + PhysActiveDays,
                       data = p.BXCX.frame)

summary(p.BXCX.model)
```

```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve +
##     BPDiaAve + SmokeNow + PhysActiveDays, data = p.BXCX.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5764 -0.6158 -0.0084  0.6574  3.8416
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.6682540  0.6104391   7.647 4.01e-14 ***
## Age             0.0993535  0.0112143   8.860  < 2e-16 ***
## pb.Age2        -0.0009453  0.0001135  -8.331  < 2e-16 ***
## Weight         -0.0006614  0.0016858  -0.392  0.69487
## Height         -0.0087700  0.0033509  -2.617  0.00897 **
## BPSysAve        0.0057045  0.0019803   2.881  0.00404 **
## BPDiaAve        0.0128515  0.0028416   4.523 6.67e-06 ***
## SmokeNowYes     0.0127777  0.0596913   0.214  0.83053
## PhysActiveDays -0.0128377  0.0154387  -0.832  0.40583
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9849 on 1280 degrees of freedom
## Multiple R-squared:  0.1264, Adjusted R-squared:  0.121
## F-statistic: 23.15 on 8 and 1280 DF,  p-value: < 2.2e-16
```

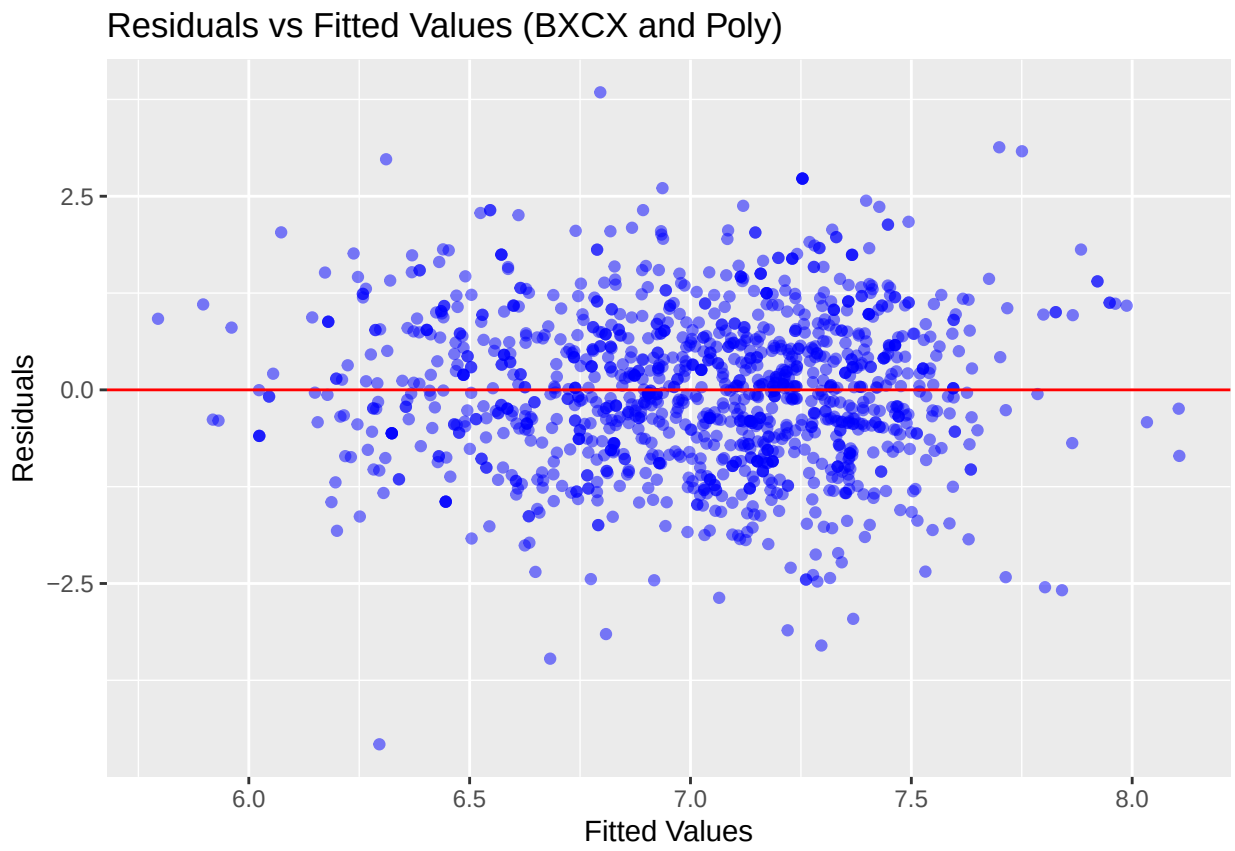
```r
#FITTED AND RESIDUAL VALUES FROM TRANSFORMED
pb.fitted <- fitted(p.BXCX.model)
pb.residuals <- resid(p.BXCX.model)

#DATA FRAME FOR PLOTTING
pb.plot_data <- data.frame(pb.fitted = pb.fitted, pb.residuals = pb.residuals)

#PAIRWISE PLOTS OF ORIGINAL MODEL
pairs(~pb.TotChol+Age+pb.Age2+Weight+Height+
        BPSysAve+BPDiaAve+SmokeNow+PhysActiveDays,
      data = p.BXCX.frame,
```

```
      main = "Pairwise ScatterPlots of Transformed Polynomial Model",
      col = "blue")
```

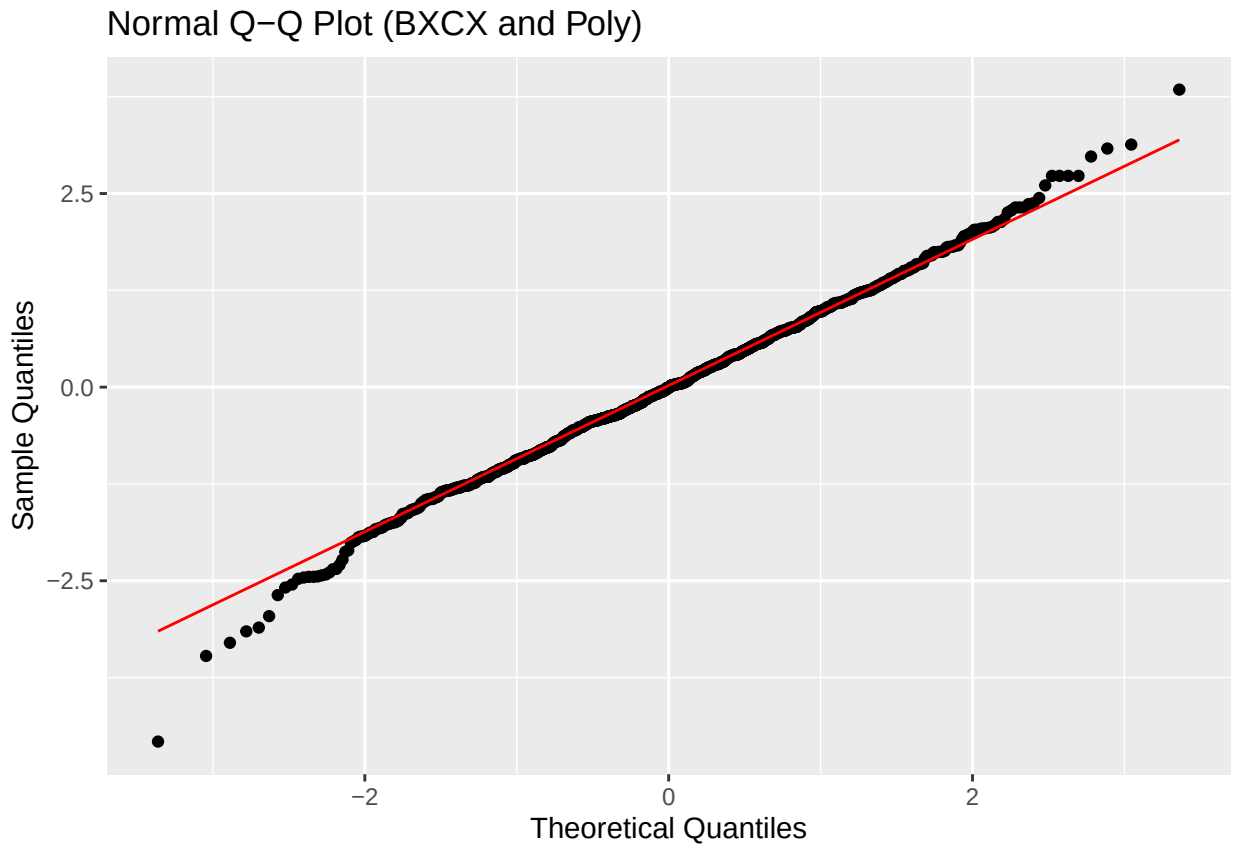


Residual Plots

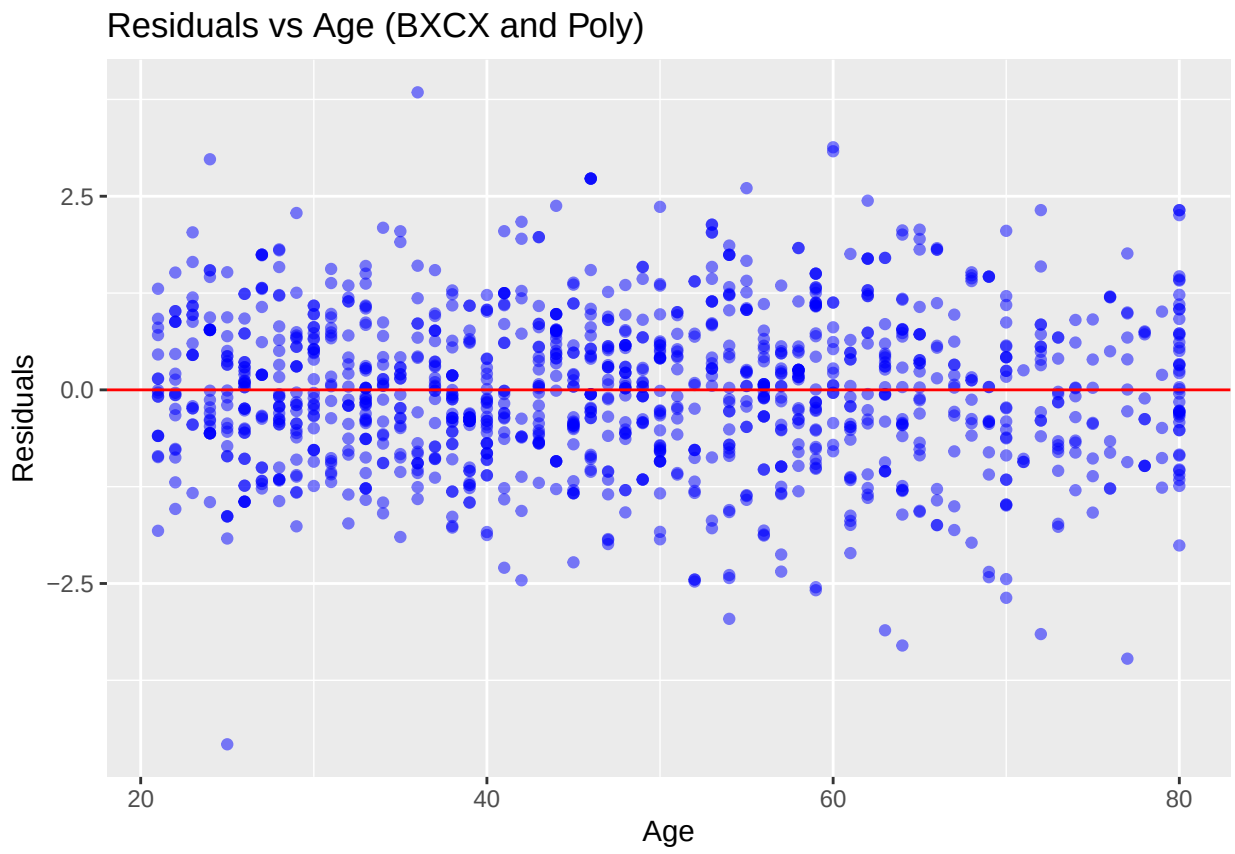
```
#RESIDUALS VS FITTED
res_fitted_plot <- ggplot(data = pb.plot_data,
                          aes(x = pb.fitted, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Fitted Values (BXCX and Poly)",
       x = "Fitted Values", y = "Residuals")

print(res_fitted_plot)
```


Residuals vs Fitted Values (BXCX and Poly)
Residuals
2.5
0.0
−2.5
6.0
6.5
7.0
7.5
8.0
Fitted Values

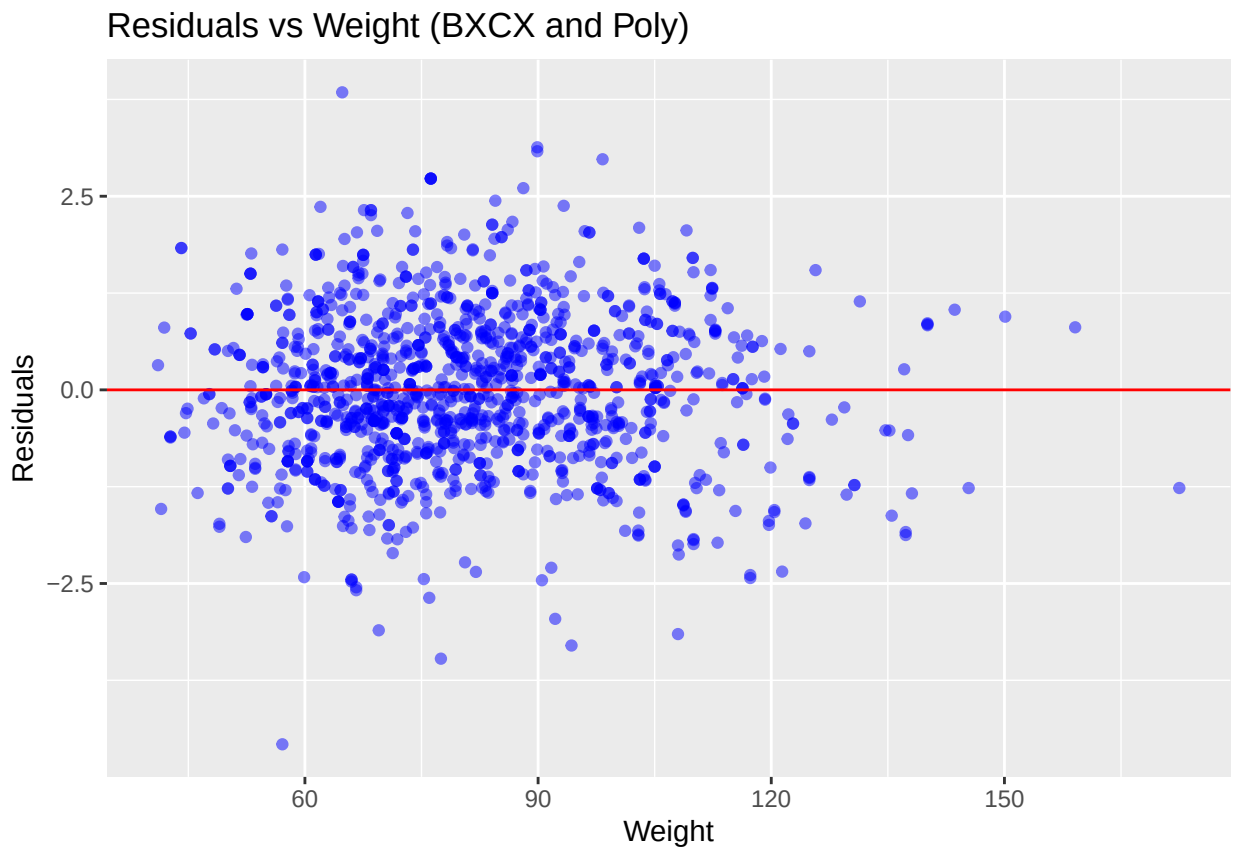
```r
#NORMAL QQ PLOT
qq_plot <- ggplot(data = data.frame(pb.residuals = pb.residuals),
                  aes(sample = pb.residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot (BXCX and Poly)",
       x = "Theoretical Quantiles", y = "Sample Quantiles")

print(qq_plot)
```



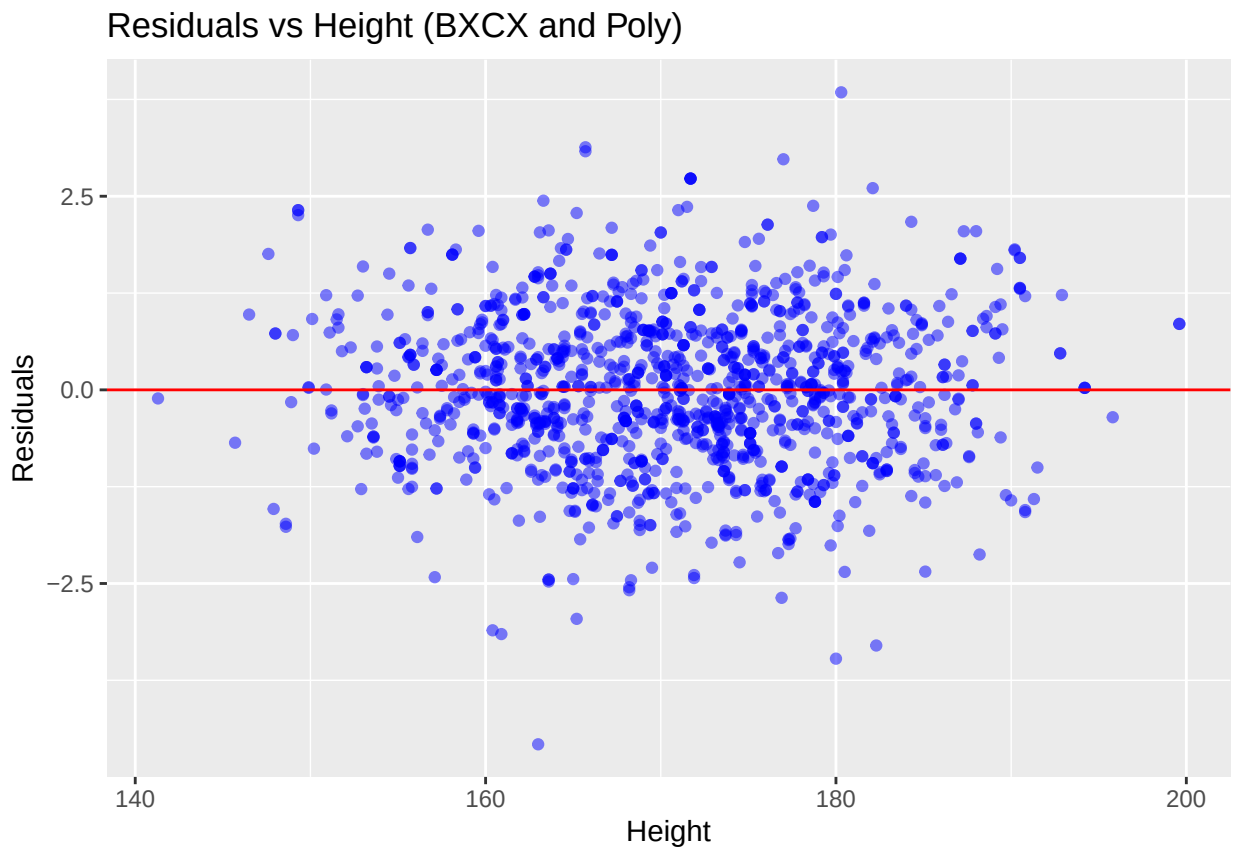

```r
#RESIDUALS VS AGE
res_age_plot <- ggplot(p.BXCX.frame,
                       aes(x = Age, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Age (BXCX and Poly)",
       x = "Age", y = "Residuals")

print(res_age_plot)
```



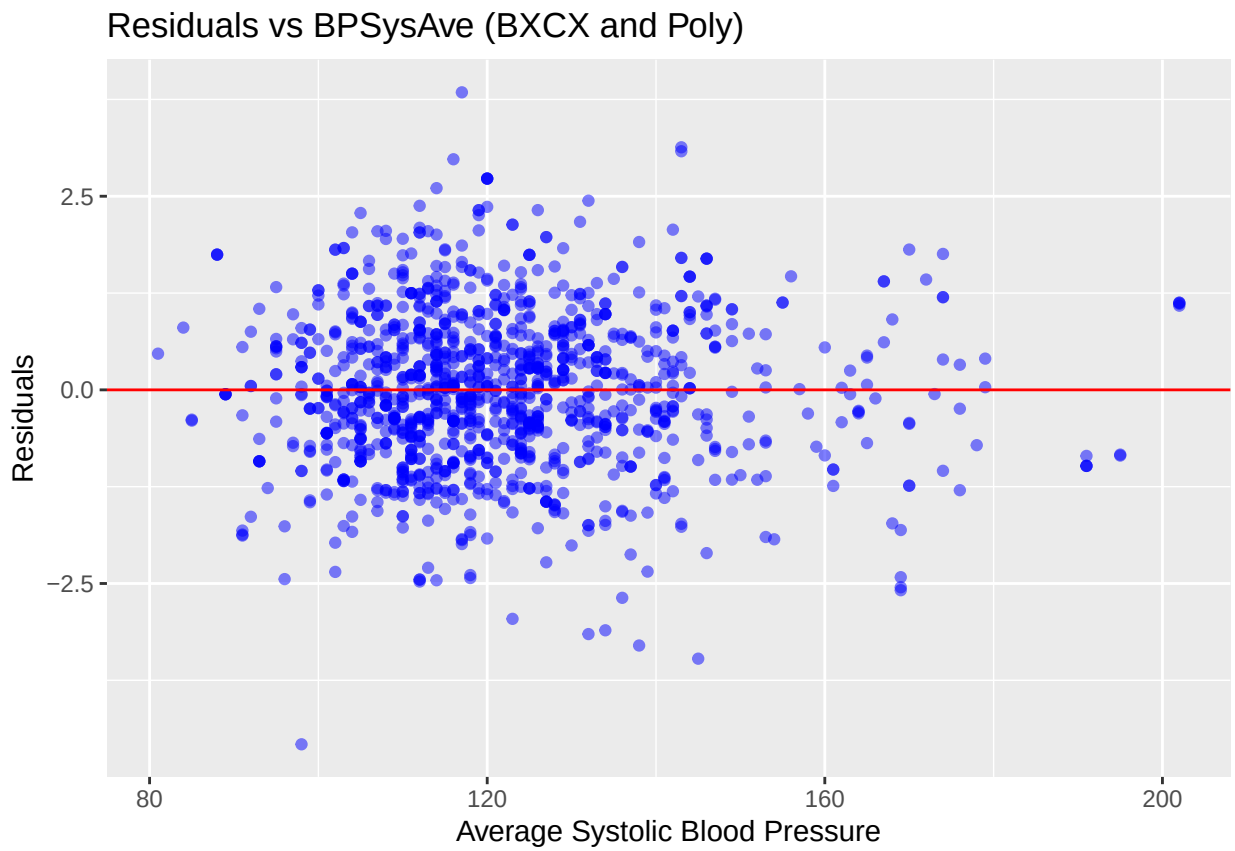

```
#RESIDUALS VS WEIGHT
res_weight_plot <- ggplot(p.BXCX.frame,
                          aes(x = Weight, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Weight (BXCX and Poly)",
       x = "Weight", y = "Residuals")

print(res_weight_plot)
```



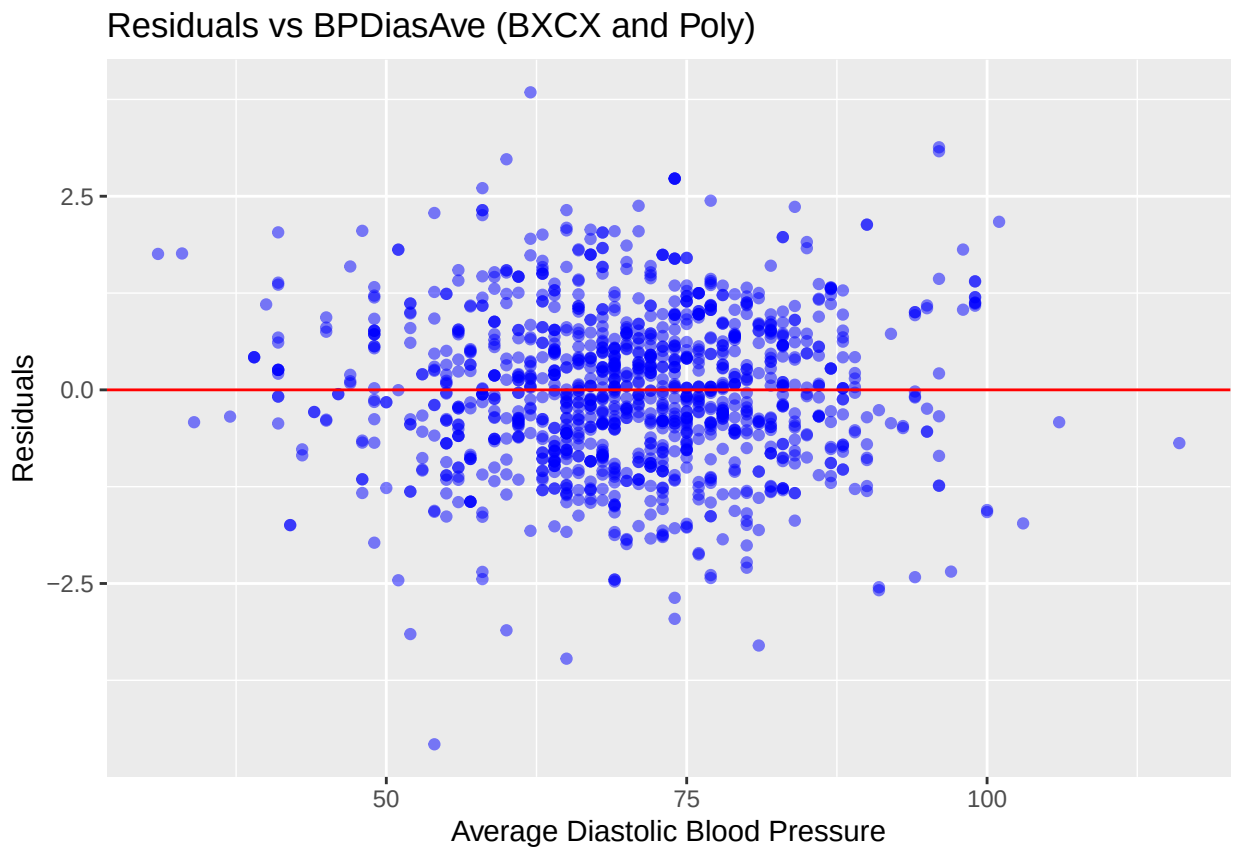

```
#RESIDUALS VS HEIGHT
res_height_plot <- ggplot(p.BXCX.frame,
                          aes(x = Height, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Height (BXCX and Poly)",
       x = "Height", y = "Residuals")

print(res_height_plot)
```



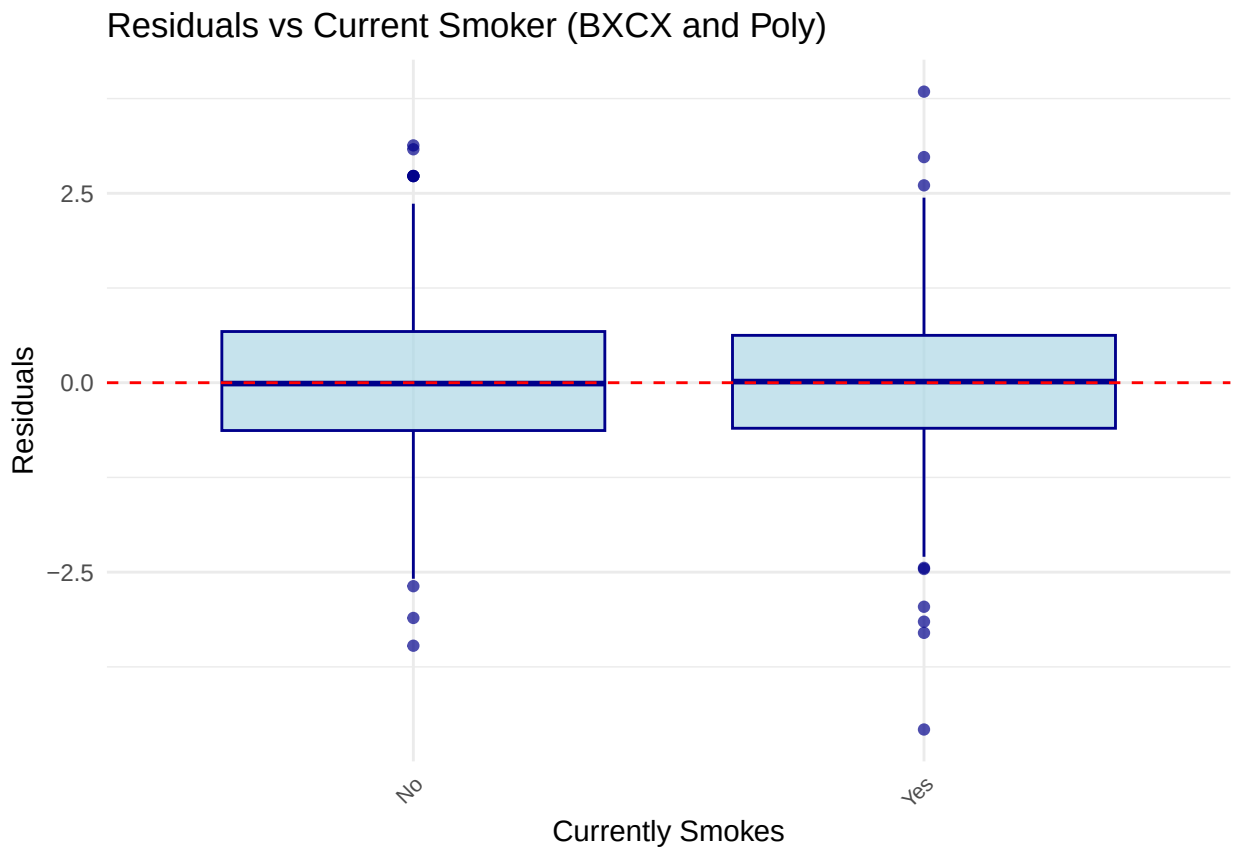

```r
#RESIDUALS VS BPSysAve
res_BPSysAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPSysAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPSysAve (BXCX and Poly)",
       x = "Average Systolic Blood Pressure", y = "Residuals")

print(res_BPSysAve_plot)
```



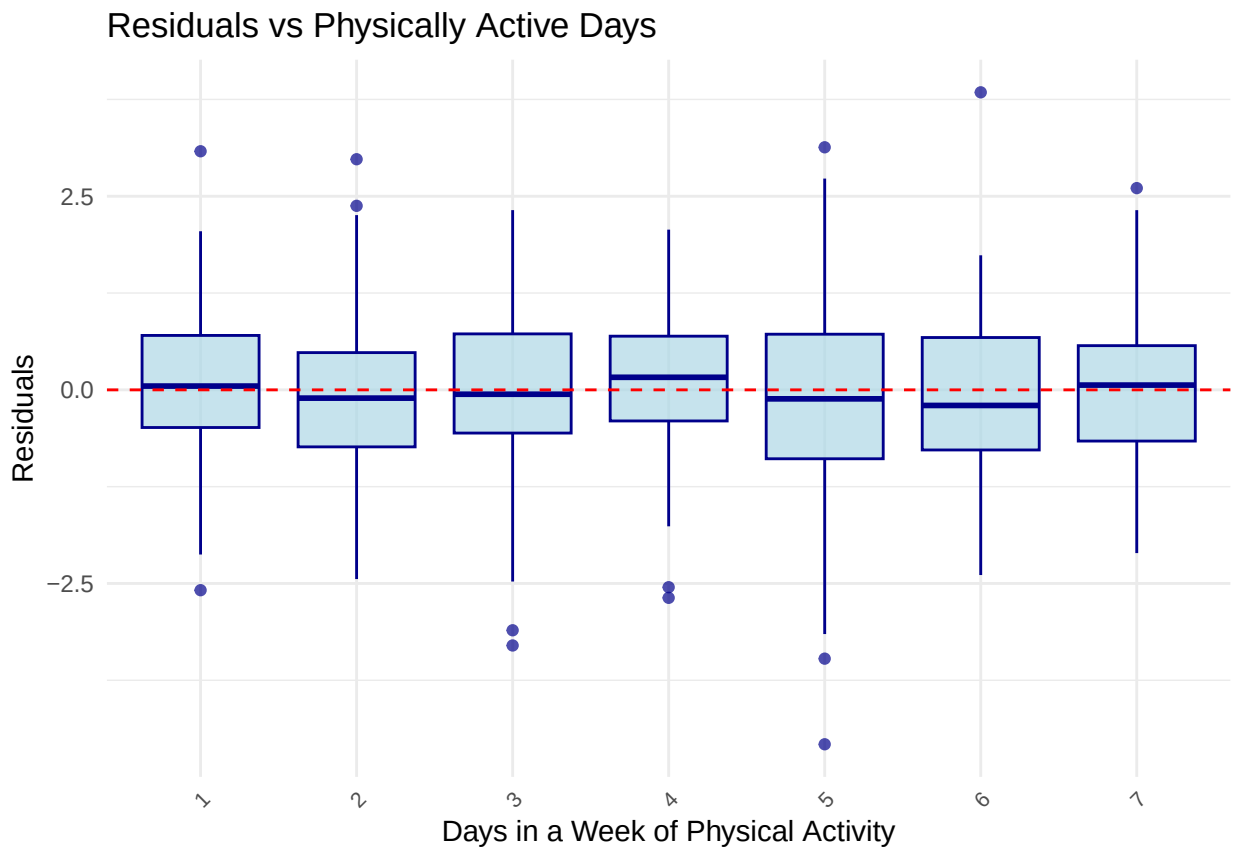

```r
#RESIDUALS VS BPDiaAve
res_BPDiaAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPDiaAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPDiasAve (BXCX and Poly)",
       x = "Average Diastolic Blood Pressure", y = "Residuals")

print(res_BPDiaAve_plot)
```



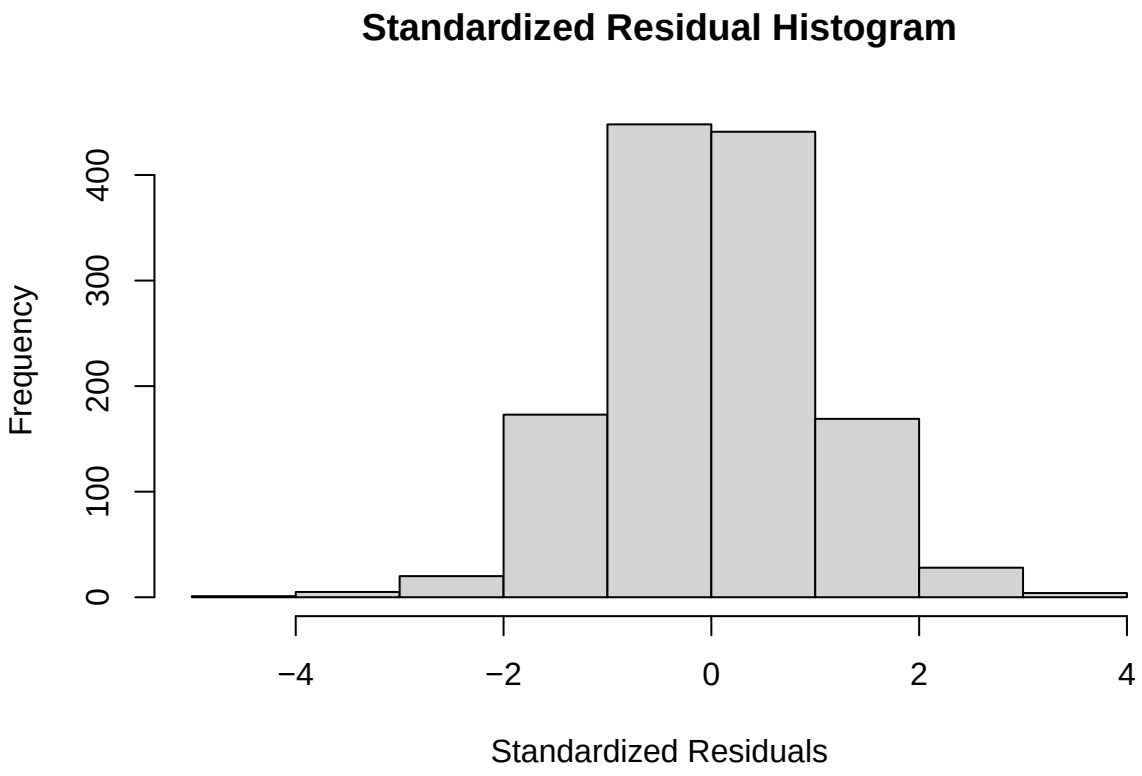

```
#RESIDUALS VS SmokeNow (BOXPLOT)
res_smoke_plot <- ggplot(
  p.BXCX.frame, aes(x = as.factor(SmokeNow), y = pb.residuals)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  ggtitle("Residuals vs Current Smoker (BXCX and Poly)") +
  xlab("Currently Smokes") +
  ylab("Residuals") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

print(res_smoke_plot)
```




```r
#RESIDUALS VS PhysActiveDays (BOXPLOT)
res_active_plot <- ggplot(
  p.BXCX.frame,
  aes(x = as.factor(PhysActiveDays), y = pb.residuals)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  ggtitle("Residuals vs Physically Active Days") +
  xlab("Days in a Week of Physical Activity") +
  ylab("Residuals") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8))

print(res_active_plot)
```




```
tr_stres_values <- rstandard(p.BXCX.model)

tr_stres_plot <- hist(tr_stres_values,
                          xlab = "Standardized Residuals",
                          main = "Standardized Residual Histogram")
```

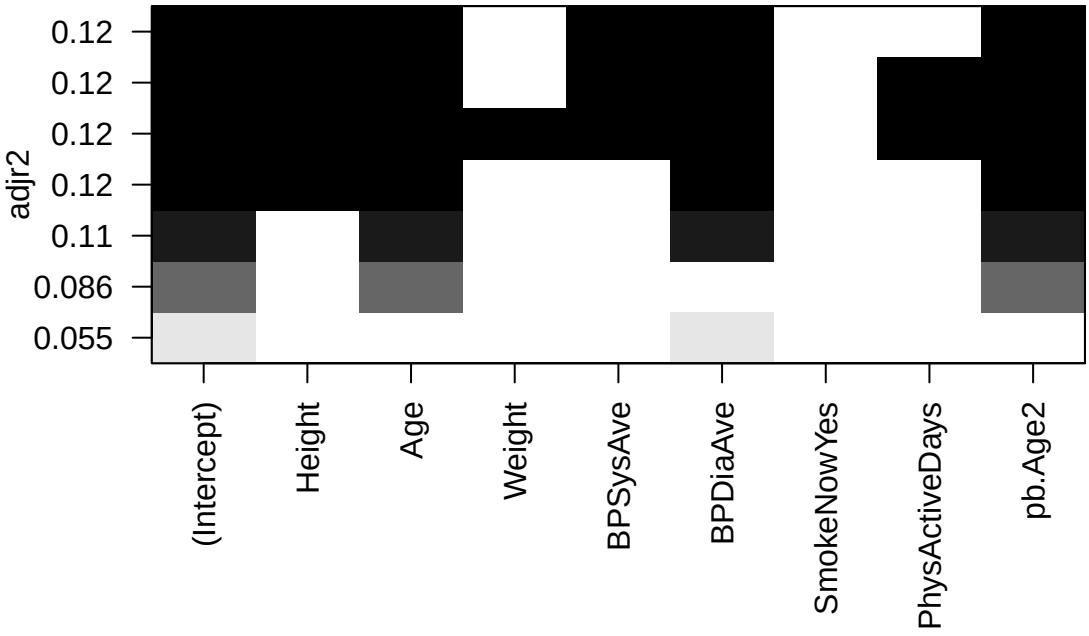



```
library(leaps)

best_subset_p.BXCX <- regsubsets(pb.TotChol~., data=p.BXCX.frame,nvmax=7,                nbest=1,reall

summary(best_subset_p.BXCX)
```

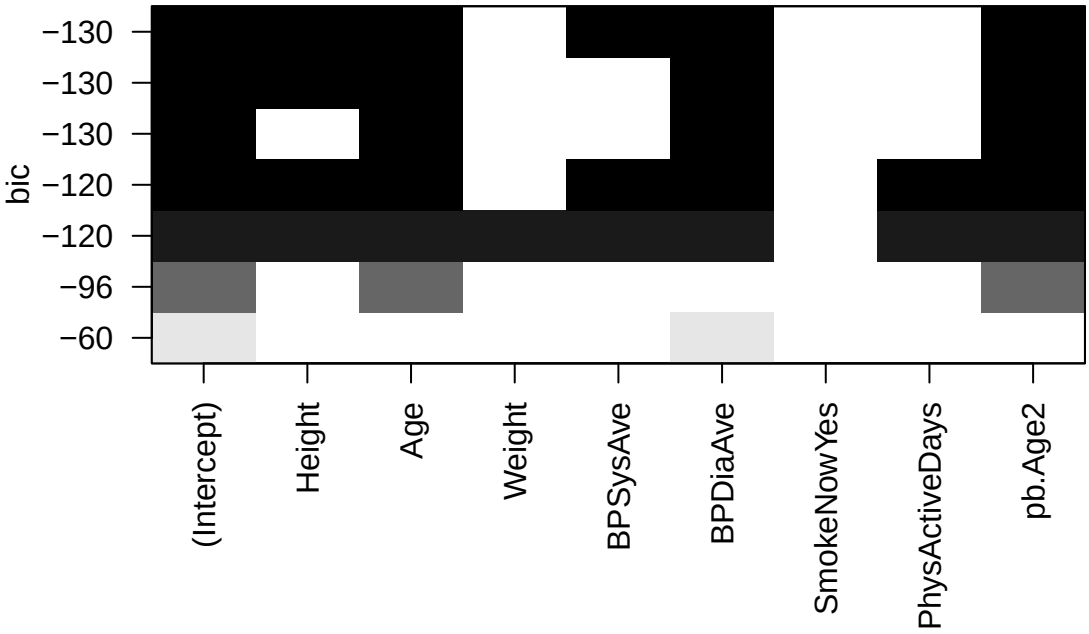
```
## Subset selection object
## Call: regsubsets.formula(pb.TotChol ~ ., data = p.BXCX.frame, nvmax = 7,
##     nbest = 1, really.big = TRUE, method = "exhaustive")
## 8 Variables  (and intercept)
##                Forced in Forced out
## Height             FALSE      FALSE
## Age                FALSE      FALSE
## Weight             FALSE      FALSE
## BPSysAve           FALSE      FALSE
## BPDiaAve           FALSE      FALSE
## SmokeNowYes        FALSE      FALSE
## PhysActiveDays     FALSE      FALSE
## pb.Age2            FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          Height Age Weight BPSysAve BPDiaAve SmokeNowYes PhysActiveDays pb.Age2
## 1  ( 1 ) " "    " " " "    " "      "*"      " "          " "             " "
## 2  ( 1 ) " "    "*" " "    " "      " "      " "          " "             "*"
## 3  ( 1 ) " "    "*" " "    " "      "*"      " "          " "             "*"
```

```
## 4  ( 1 ) "*"    "*" " "    " "      "*"      " "          " "          "*"
## 5  ( 1 ) "*"    "*" " "    "*"      "*"      " "          " "          "*"
## 6  ( 1 ) "*"    "*" " "    "*"      "*"      " "          "*"          "*"
## 7  ( 1 ) "*"    "*" "*"    "*"      "*"      " "          "*"          "*"
```

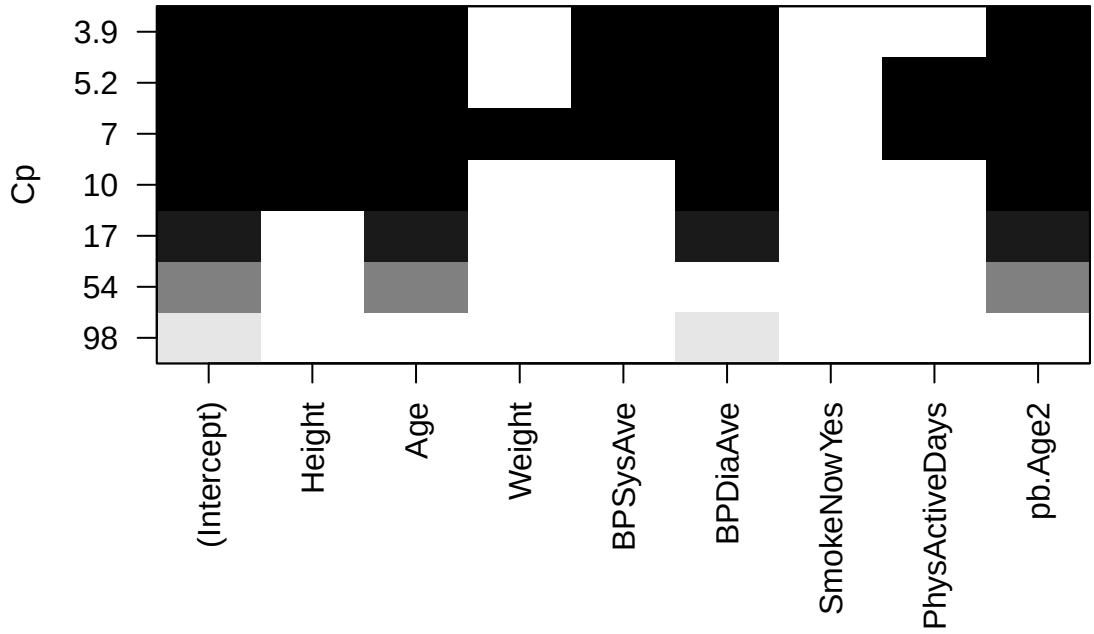
```
plot(best_subset_p.BXCX,scale='adjr2')
```




```
plot(best_subset_p.BXCX,scale='bic');
```




```
plot(best_subset_p.BXCX,scale='Cp')
```




```
AIC_p.BXCX <- step(p.BXCX.model, direction="both")
```

```
## Start:  AIC=-30.3
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     SmokeNow + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - SmokeNow        1     0.044 1241.6 -32.257
## - Weight          1     0.149 1241.7 -32.148
## - PhysActiveDays  1     0.671 1242.3 -31.607
## <none>                        1241.6 -30.303
## - Height          1     6.644 1248.2 -25.423
## - BPSysAve        1     8.049 1249.6 -23.974
## - BPDiaAve        1    19.841 1261.4 -11.868
## - pb.Age2         1    67.328 1308.9  35.767
## - Age             1    76.136 1317.7  44.411
##
## Step:  AIC=-32.26
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - Weight          1     0.168 1241.8 -34.083
## - PhysActiveDays  1     0.671 1242.3 -33.560
## <none>                        1241.6 -32.257
```

```
## + SmokeNow        1     0.044 1241.6 -30.303
## - Height          1     6.630 1248.3 -27.392
## - BPSysAve        1     8.176 1249.8 -25.796
## - BPDiaAve        1    19.809 1261.4 -13.855
## - pb.Age2         1    67.493 1309.1  33.973
## - Age             1    76.096 1317.7  42.416
##
## Step:  AIC=-34.08
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - PhysActiveDays  1     0.647 1242.5 -35.411
## <none>                        1241.8 -34.083
## + Weight          1     0.168 1241.6 -32.257
## + SmokeNow        1     0.063 1241.7 -32.148
## - BPSysAve        1     8.039 1249.8 -27.765
## - Height          1     9.542 1251.3 -26.216
## - BPDiaAve        1    19.650 1261.5 -15.846
## - pb.Age2         1    67.426 1309.2  32.072
## - Age             1    76.087 1317.9  40.571
##
## Step:  AIC=-35.41
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        1242.5 -35.411
## + PhysActiveDays  1     0.647 1241.8 -34.083
## + Weight          1     0.143 1242.3 -33.560
## + SmokeNow        1     0.062 1242.4 -33.475
## - BPSysAve        1     7.955 1250.4 -29.184
## - Height          1     9.401 1251.8 -27.695
## - BPDiaAve        1    19.757 1262.2 -17.075
## - pb.Age2         1    66.884 1309.3  30.176
## - Age             1    75.470 1317.9  38.601
```

```
summary(AIC_p.BXCX)
```

```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve +
##     BPDiaAve, data = p.BXCX.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5880 -0.6276 -0.0100  0.6471  3.8305
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7145172  0.5792640   8.139 9.36e-16 ***
## Age          0.0981881  0.0111223   8.828  < 2e-16 ***
## pb.Age2     -0.0009365  0.0001127  -8.311 2.40e-16 ***
## Height      -0.0093124  0.0029889  -3.116  0.00188 **
## BPSysAve     0.0056363  0.0019665   2.866  0.00422 **
## BPDiaAve     0.0127568  0.0028243   4.517 6.85e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9841 on 1283 degrees of freedom
## Multiple R-squared:  0.1258, Adjusted R-squared:  0.1224
## F-statistic: 36.93 on 5 and 1283 DF,  p-value: < 2.2e-16
```

```
reduced.model <- lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve,
                        data = p.BXCX.frame)
```

```
leverage <- hatvalues(reduced.model)
#leverage

## Threshold
p <- 5
high_lev <- 2*(p+1)/n
#high_lev
```

Find the leverage points

leverage_points_index <- which(leverage > high_lev) leverage_points_index rownames(p.BXCX.frame)[leverage_points_in

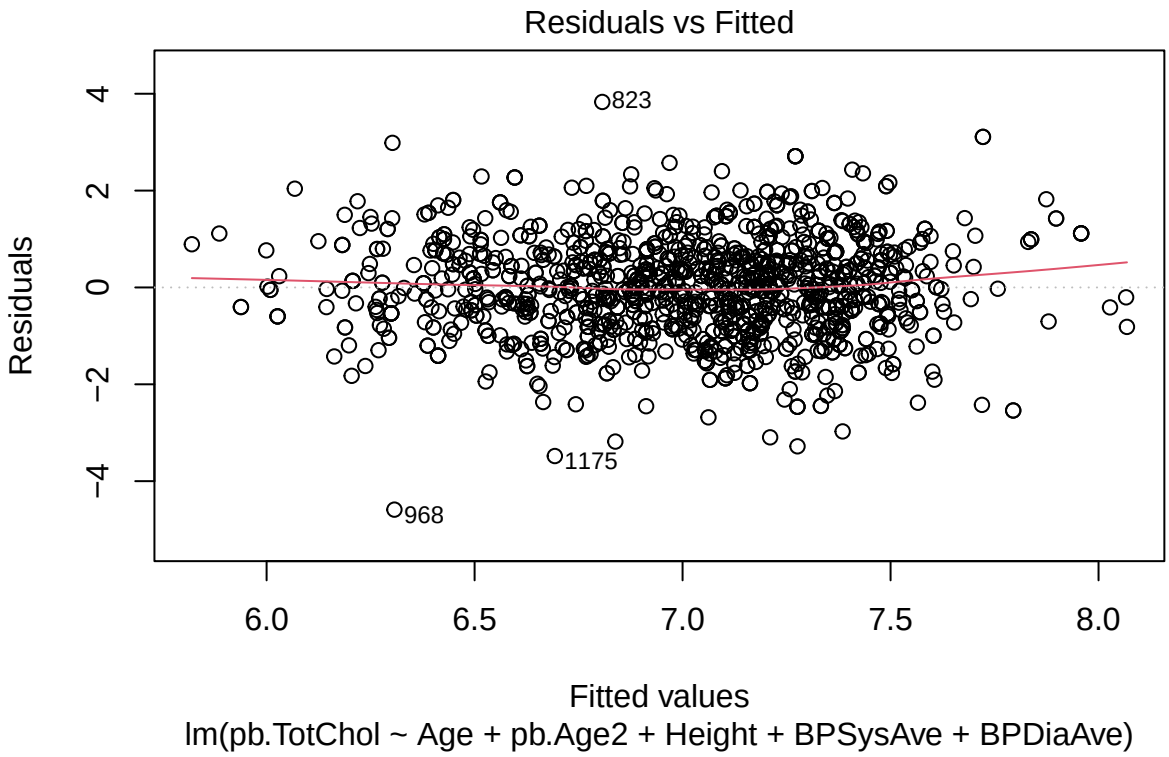
Check if the absolute values of the standardized residuals are greater than 3

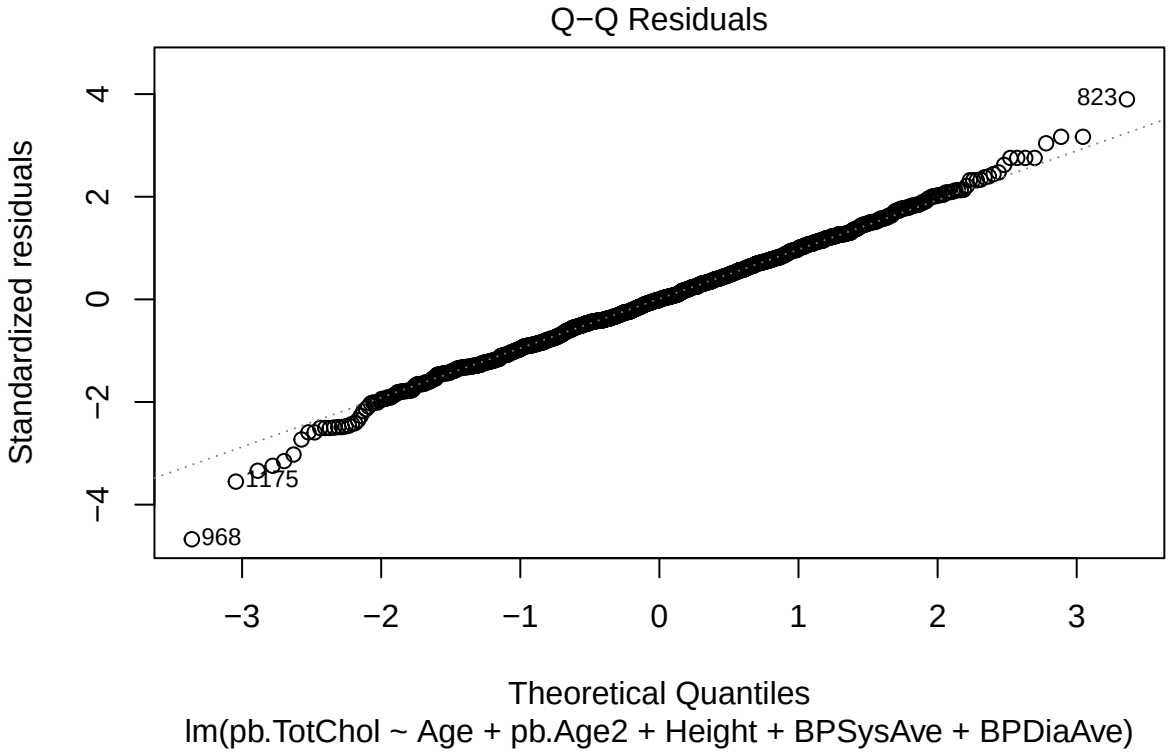
st.residuals <- rstandard(reduced.model) ## standardized residuals st.residuals

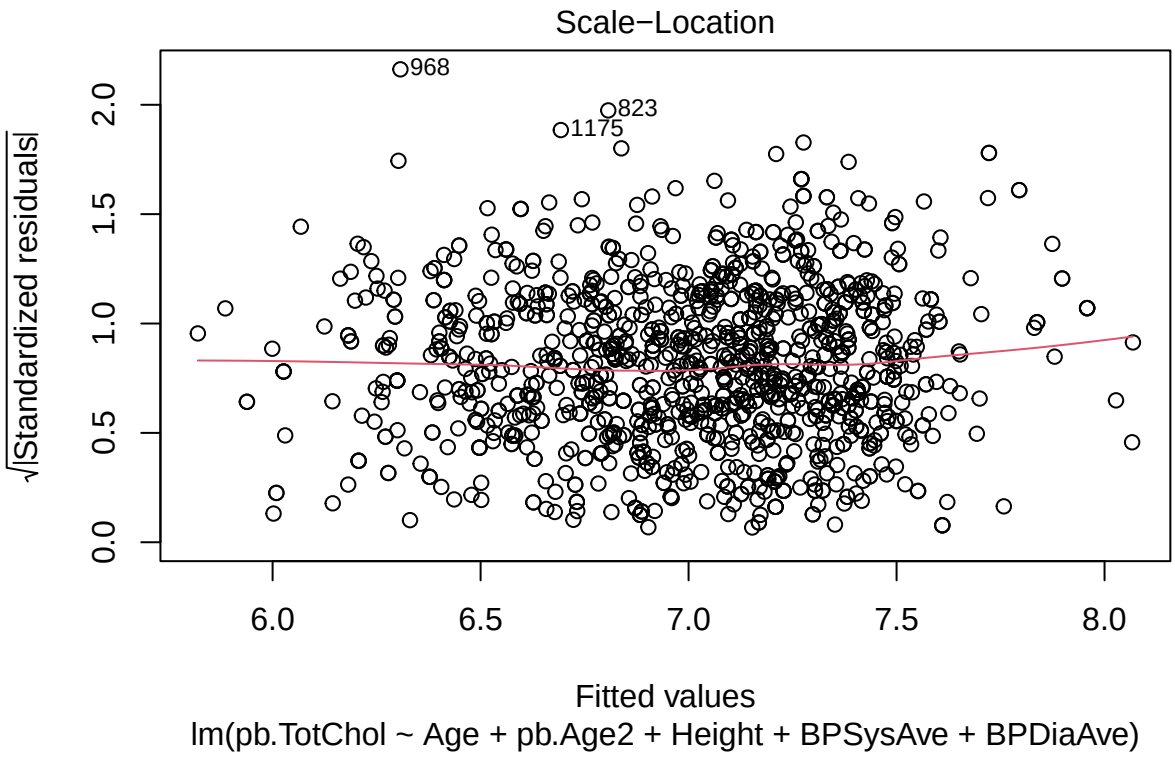
Outliers

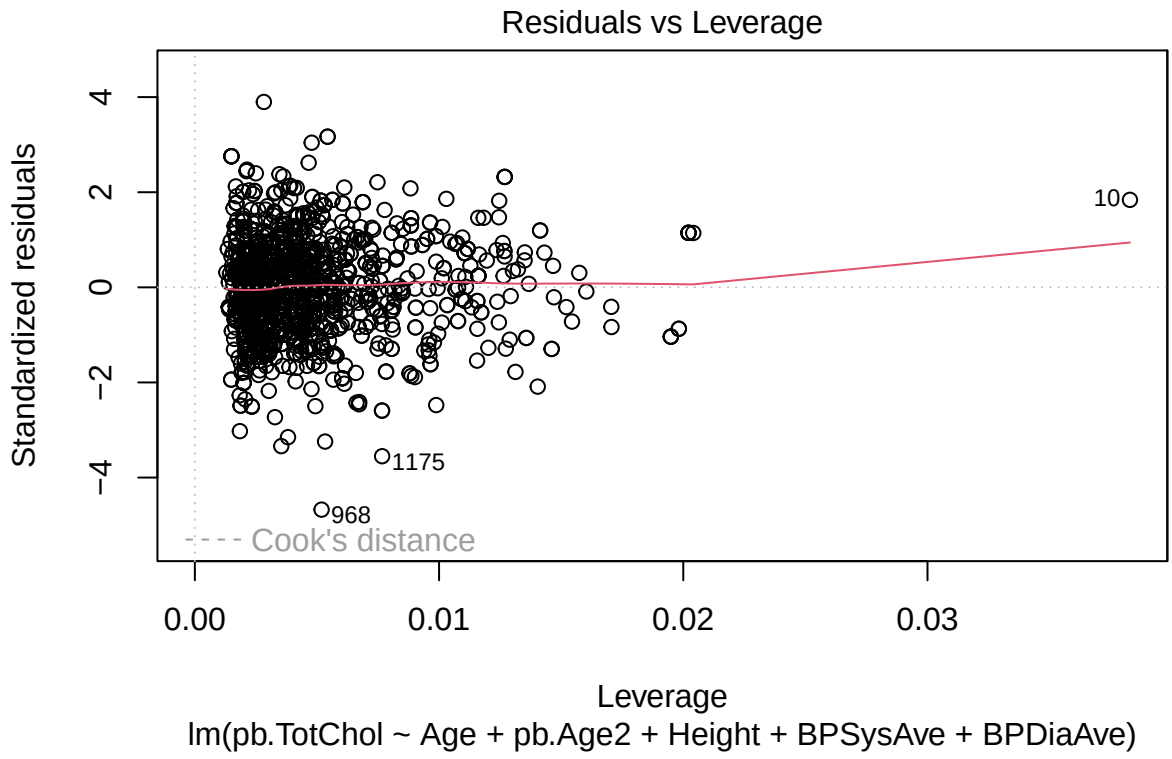
outliers_index <- which(abs(st.residuals)>3) outliers_index

```
#FINDING INFLUENTIAL POINTS USING RESIDUALS VS LEVERAGE
plot(reduced.model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)
823
1175
968
4
2
0
−2
−4
6.0
6.5
7.0
7.5
8.0


Q−Q Residuals
Standardized residuals
4
2
0
−2
−4
823
1175
968
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)


Scale−Location
968
823
1175
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
6.0
6.5
7.0
7.5
8.0
Fitted values
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)




```
#TABLE OF INFLUENTIAL OBSERVATIONS
influential_points <- c(823, 968, 1175)
p.BXCX.frame[influential_points, ]
```

```
##      Height Age Weight BPSysAve BPDiaAve SmokeNow PhysActiveDays pb.Age2
## 823   180.3  36   64.8      117       62      Yes              6    1296
## 968   163.0  25   57.1       98       54      Yes              5     625
## 1175  180.0  77   77.5      145       65       No              5    5929
##      pb.TotChol
## 823   10.637426
## 968    1.719246
## 1175   3.211495
```

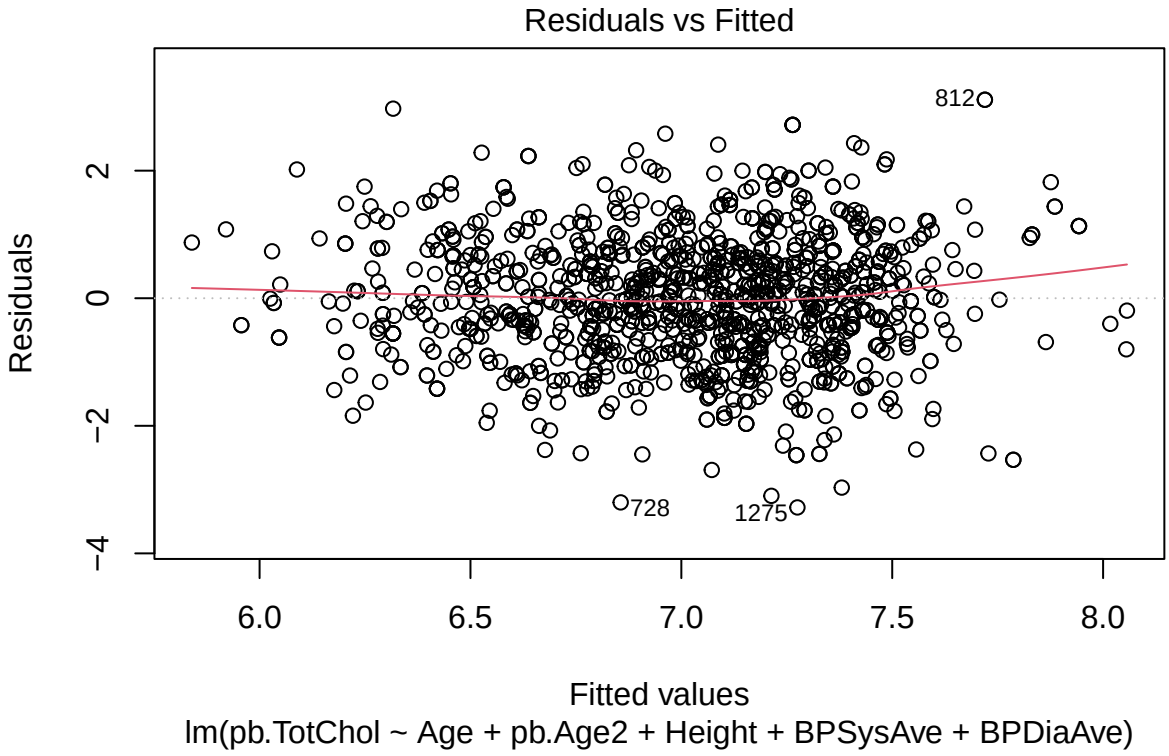
```
reduced.frame <- p.BXCX.frame %>%
  dplyr::filter(!row_number() %in% influential_points)

final.model <- lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve,
                  data = reduced.frame)
```

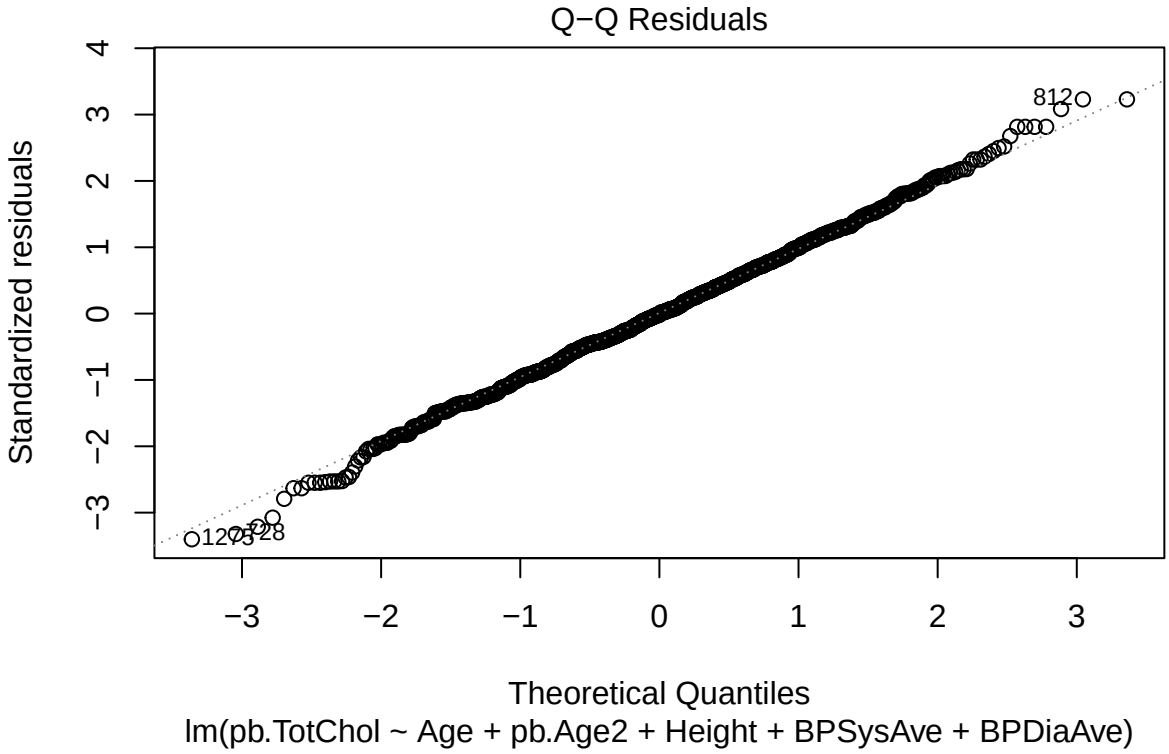
```
summary(final.model)
```

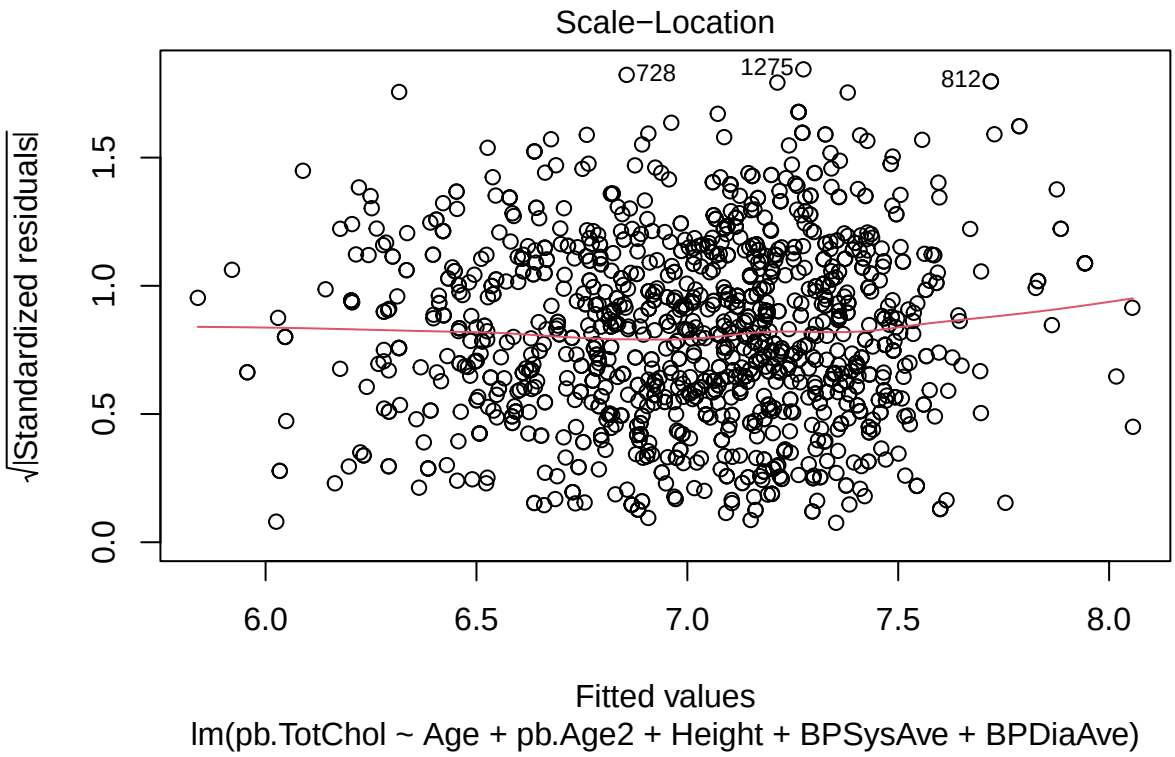
```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve +
```

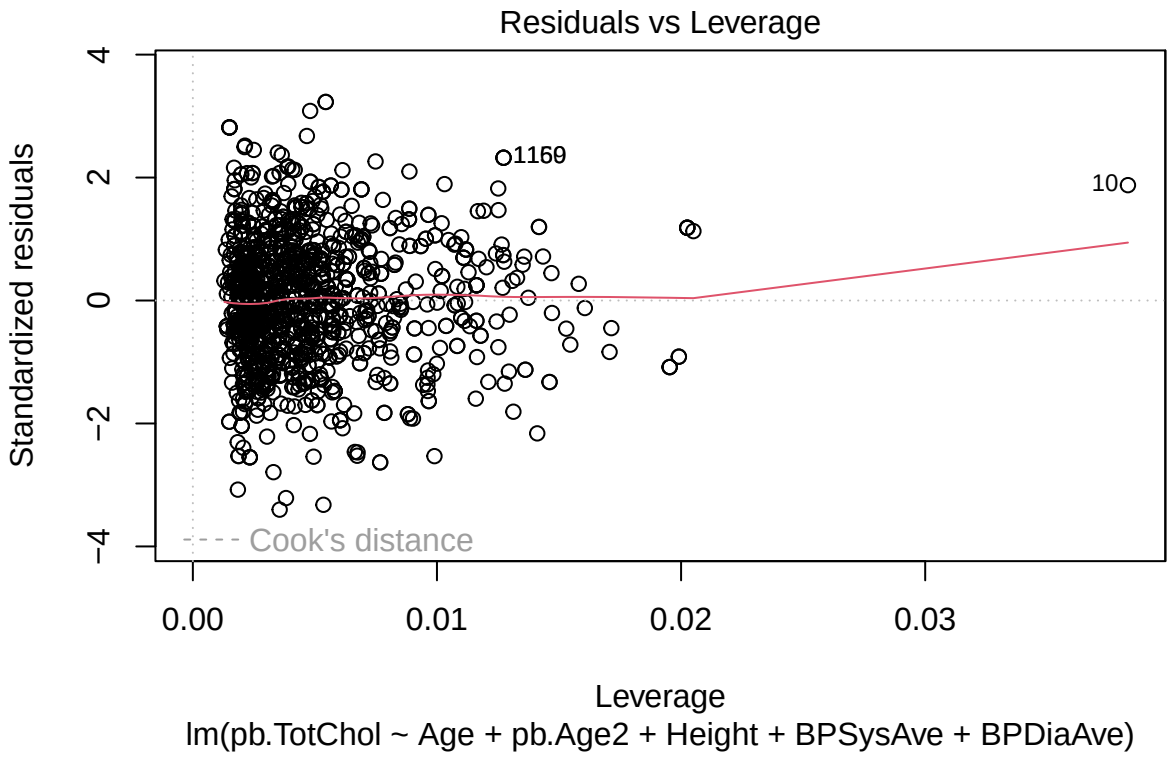
```
##     BPDiaAve, data = reduced.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2796 -0.6170 -0.0132  0.6385  3.1114
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.8544469  0.5695989   8.523  < 2e-16 ***
## Age          0.0943367  0.0109334   8.628  < 2e-16 ***
## pb.Age2     -0.0008959  0.0001108  -8.086 1.42e-15 ***
## Height      -0.0095452  0.0029380  -3.249  0.00119 **
## BPSysAve     0.0054950  0.0019309   2.846  0.00450 **
## BPDiaAve     0.0127714  0.0027741   4.604 4.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.966 on 1280 degrees of freedom
## Multiple R-squared:  0.1253, Adjusted R-squared:  0.1219
## F-statistic: 36.68 on 5 and 1280 DF,  p-value: < 2.2e-16
```

```
plot(final.model)
```

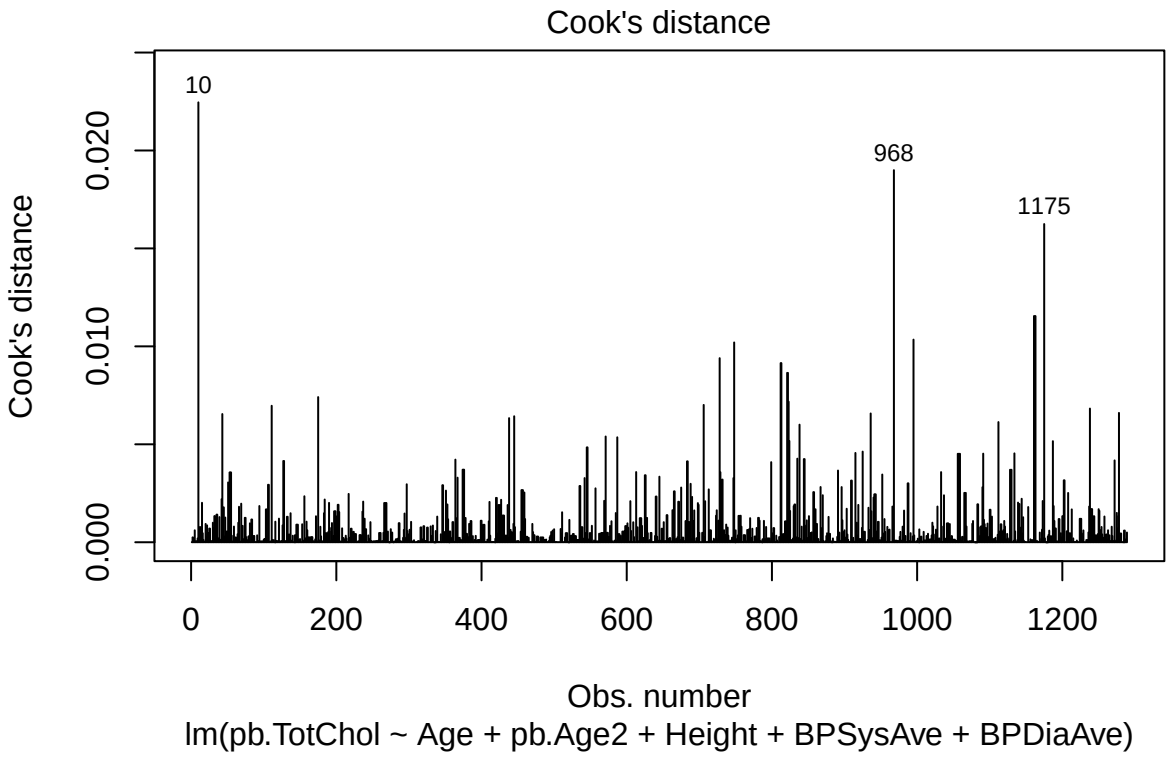



Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)
812
1273
−3
−2
−1
0
1
2
3
4

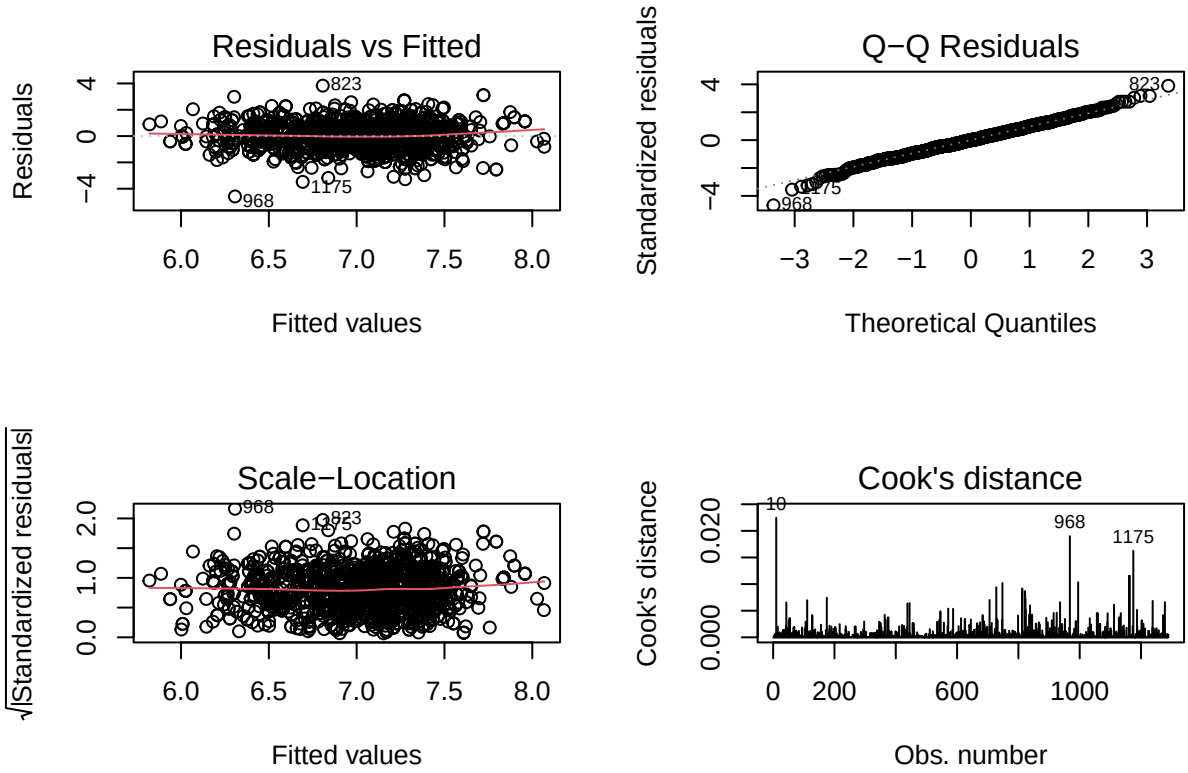

Scale−Location
728
1275
812
√|Standardized residuals|
0.0
0.5
1.0
1.5
6.0
6.5
7.0
7.5
8.0
Fitted values
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)



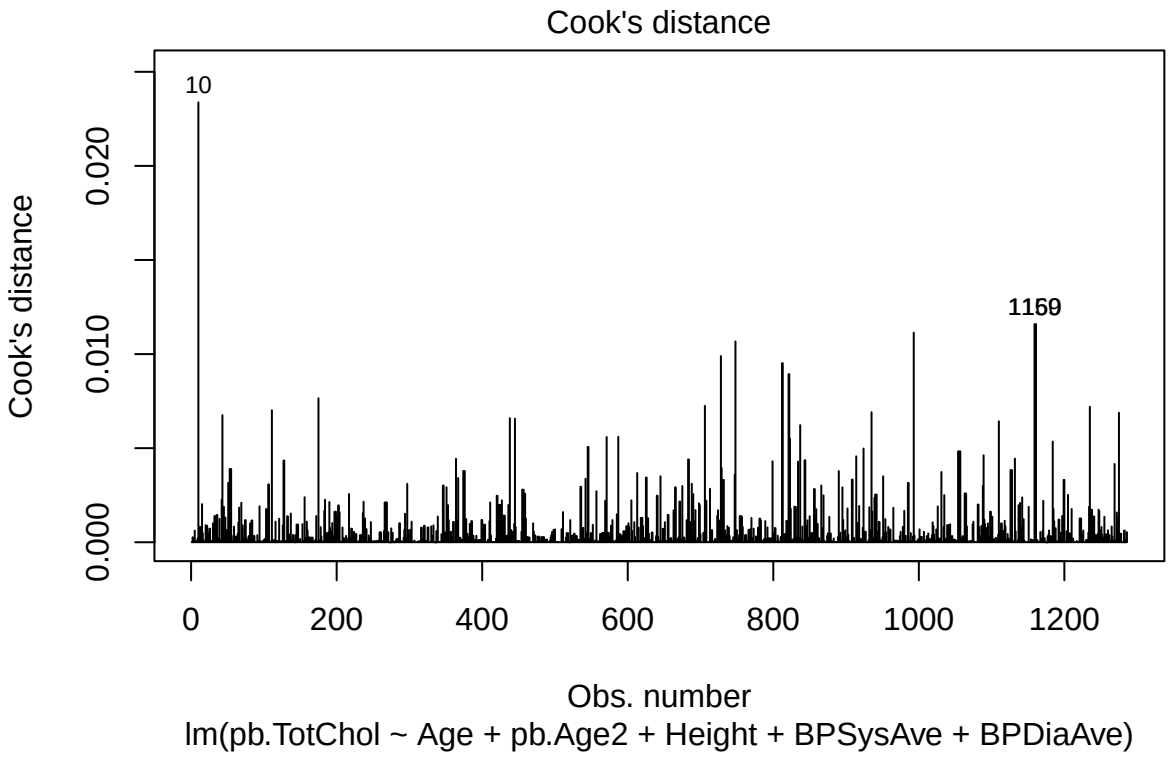

```
cooks <- cooks.distance(reduced.model)
plot(reduced.model, which = 4)
```



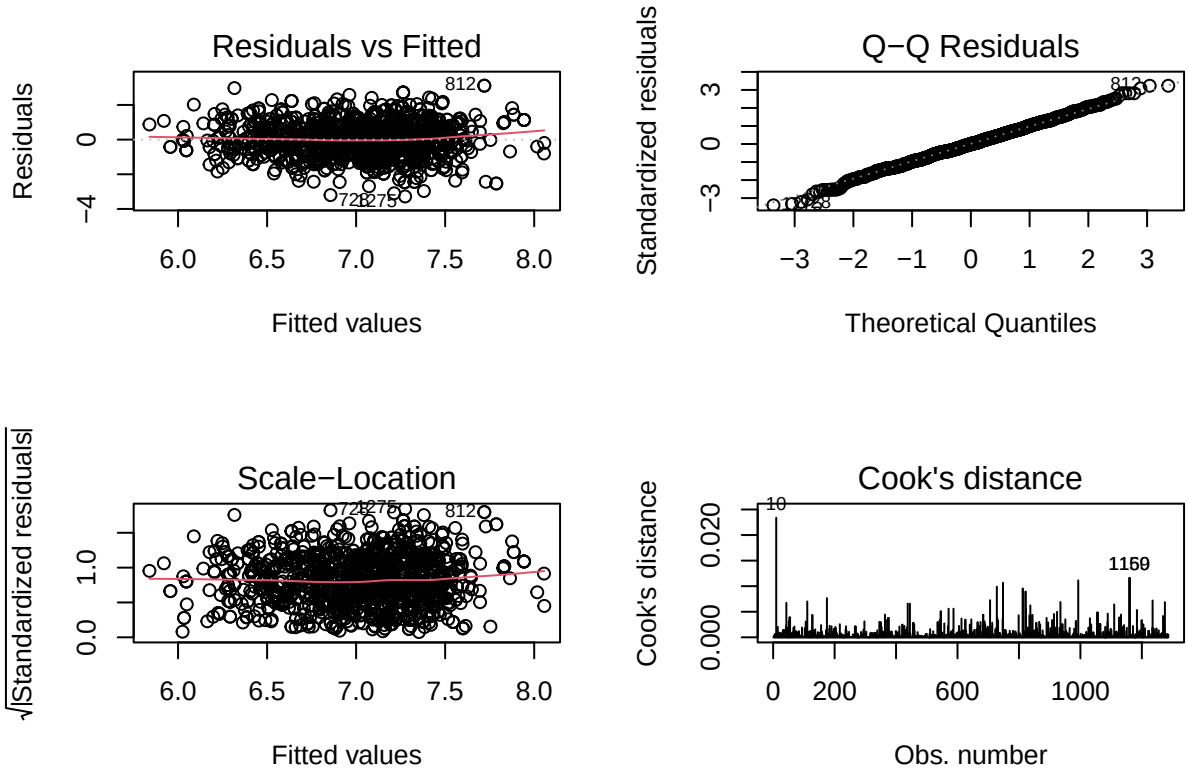

```
# All diagonistic plots
par(mfrow = c(2, 2))
plot(reduced.model, which = c(1, 2, 3, 4))
```




```
cooks <- cooks.distance(final.model)
plot(final.model, which = 4)
```




```
# All diagonistic plots
par(mfrow = c(2, 2))
plot(final.model, which = c(1, 2, 3, 4))
```




```
#Prediction accuracy
set.seed(123)
train_index <- sample(1:nrow(reduced.frame), 0.7 * nrow(reduced.frame))
train_data <- reduced.frame[train_index, ]
test_data <- reduced.frame[-train_index, ]

validation.final.model <- lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve,
                        data = train_data)
predictions <- predict(validation.final.model, newdata = test_data)

# Compare predictions to actual
mean((predictions - test_data$pb.TotChol)^2)  # MSE
```

```
## [1] 0.9767717
```

```
sqrt(mean((predictions - test_data$pb.TotChol)^2))  # RMSE
```

```
## [1] 0.9883176
```

```
library(caret)
#K-Fold (10-Fold)
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(
  pb.TotChol ~ ., data = reduced.frame,
```

```
  method = "lm",
  trControl = train_control
)

print(cv_model)
```

```
## Linear Regression
##
## 1286 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1157, 1158, 1157, 1157, 1158, 1158, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9685399  0.1178192  0.7685106
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```