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
```

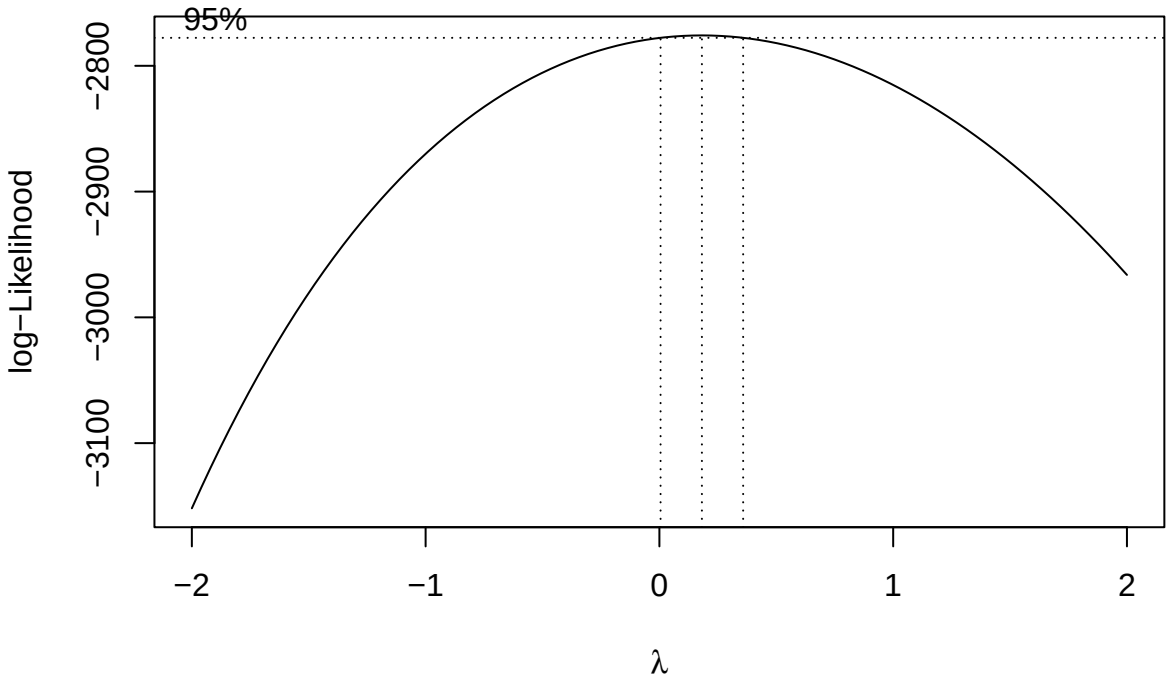
```
## [1] 7094
```

```
# remove NA entries and only select columns of interest
nhanes_data <- nhanes_filtered %>%
  dplyr::select(Height, Age, Weight, BPSysAve, BPDiaAve,
                TotChol, SmokeNow, PhysActiveDays) %>%
  na.omit() %>%
  dplyr::filter(
    Height > 0,
    Weight > 0,
    BPSysAve > 1,    # Realistic minimum for systolic BP
    BPDiaAve > 1,    # Realistic minimum for diastolic BP
    TotChol > 0
  )

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

```r
# 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

```r
#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)
```

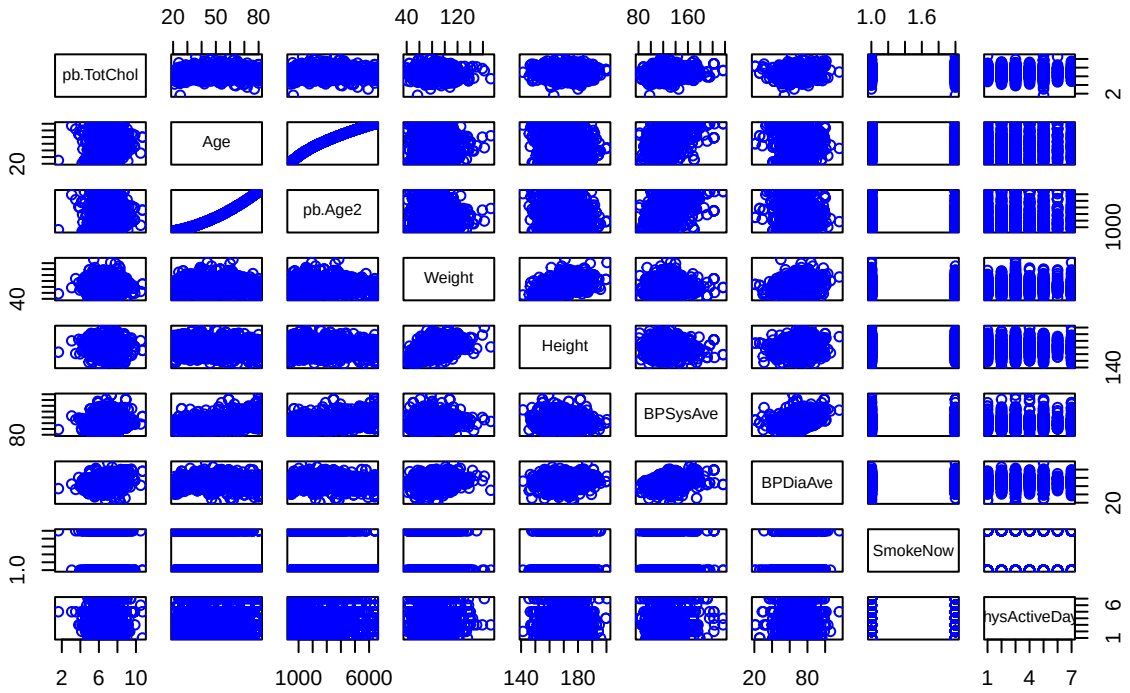


```r
pb.lambda <- pb.b$x[which.max(pb.b$y)]

pb.log_product <- sum(log(pb_data$TotChol))
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
## -4.4740 -0.6222  0.0117  0.6571  3.8785
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.3898163  0.5871066   7.477 1.35e-13 ***
## Age             0.1016266  0.0106229   9.567  < 2e-16 ***
## pb.Age2        -0.0009730  0.0001072  -9.074  < 2e-16 ***
## Weight         -0.0006222  0.0015980  -0.389  0.69708
## Height         -0.0083458  0.0032007  -2.608  0.00922 **
## BPSysAve        0.0067861  0.0018256   3.717  0.00021 ***
## BPDiaAve        0.0109084  0.0025547   4.270 2.09e-05 ***
## SmokeNowYes    -0.0254010  0.0573919  -0.443  0.65813
## PhysActiveDays -0.0101659  0.0148299  -0.685  0.49314
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.984 on 1377 degrees of freedom
## Multiple R-squared:  0.1255, Adjusted R-squared:  0.1204
## F-statistic: 24.69 on 8 and 1377 DF,  p-value: < 2.2e-16
```

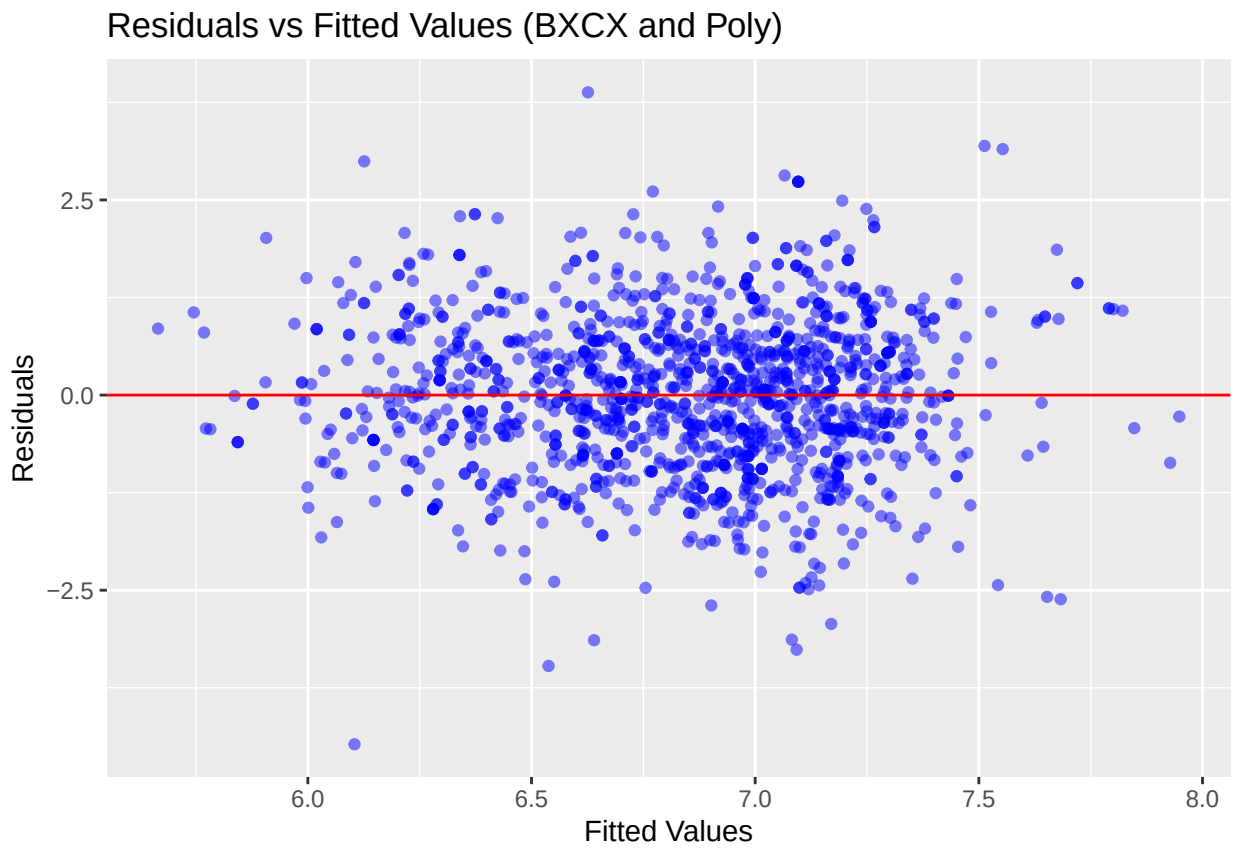
```r
#FITTED AND RESIDUAL VALUES FROM TRANSFORMED
pb.fitted <- fitted(p.BXCX.model)
pb.residuals <- resid(p.BXCX.model)

#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")
```

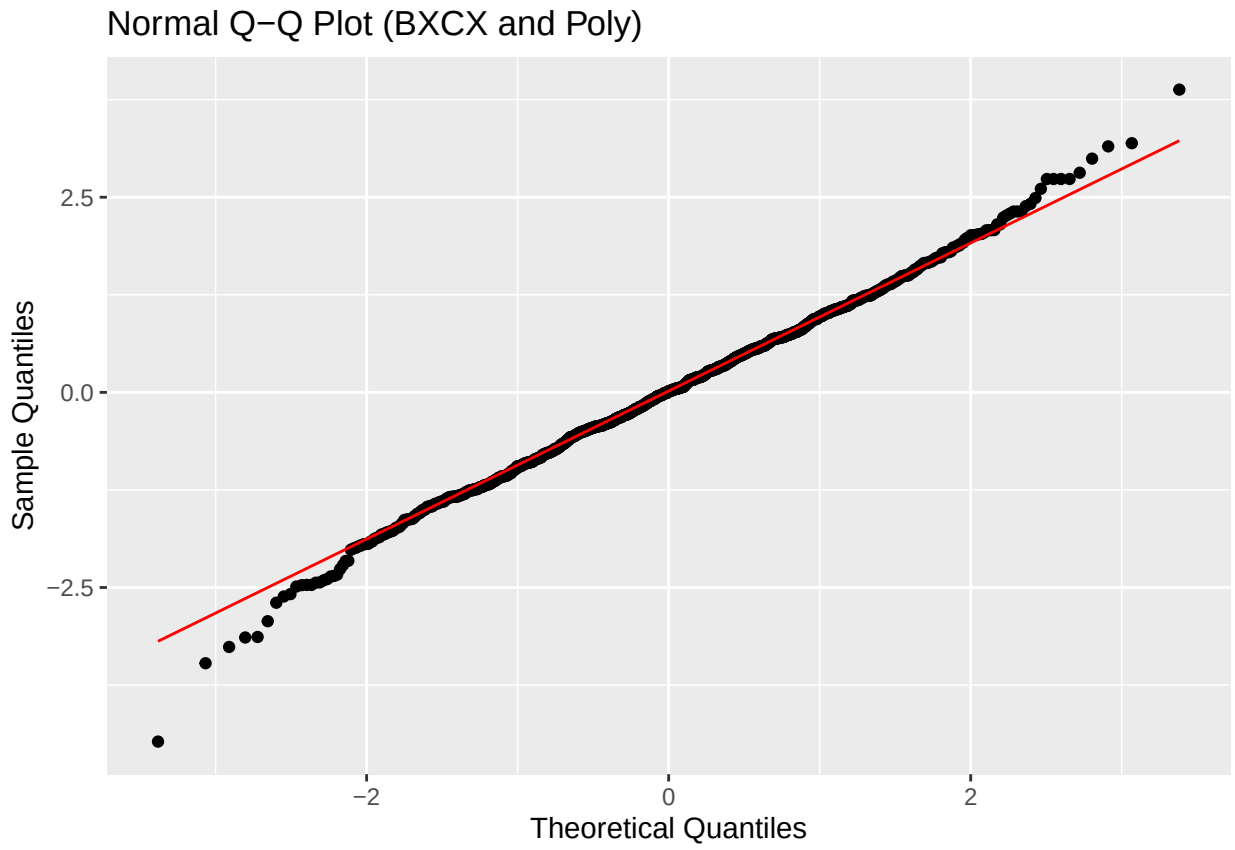


Residual Plots

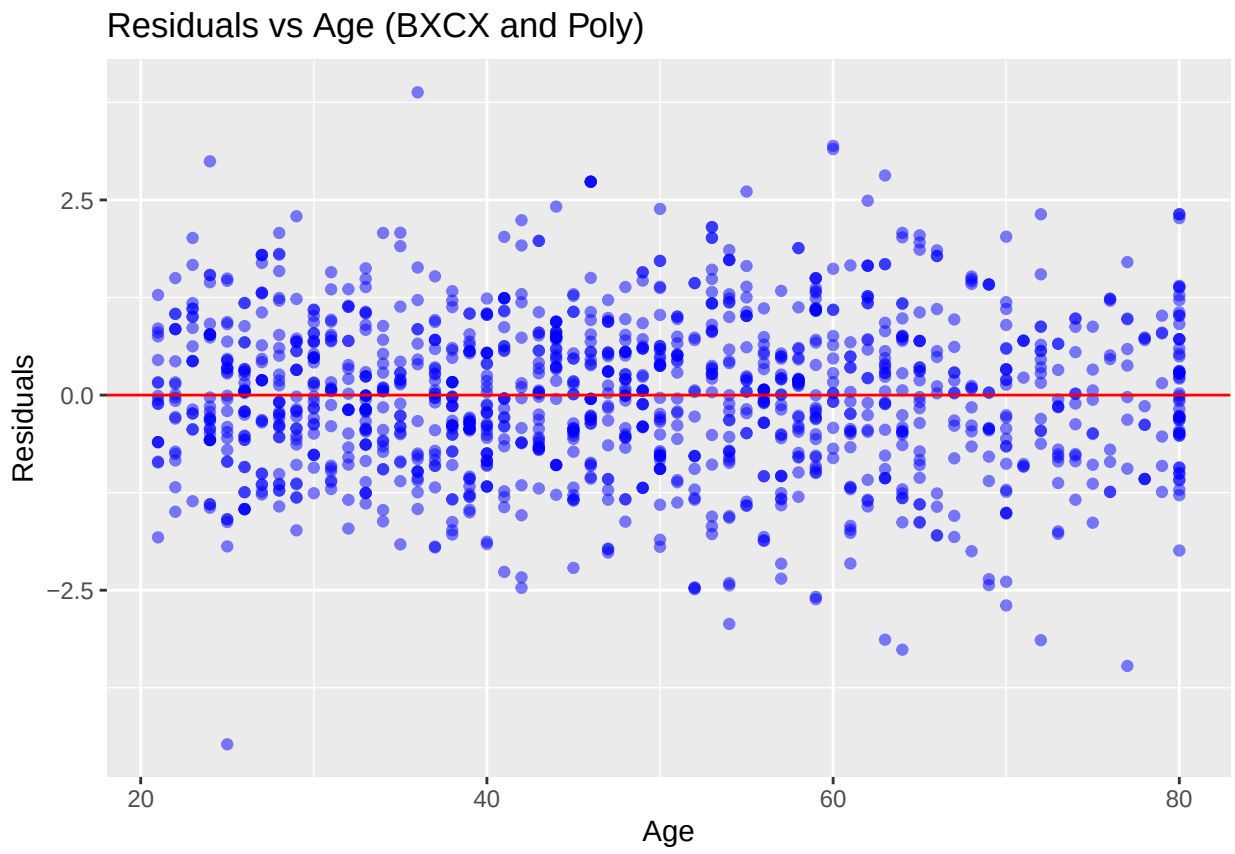
```
#RESIDUALS VS FITTED
res_fitted_plot <- ggplot(data = pb.plot_data,
                          aes(x = pb.fitted, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Fitted Values (BXCX and Poly)",
       x = "Fitted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

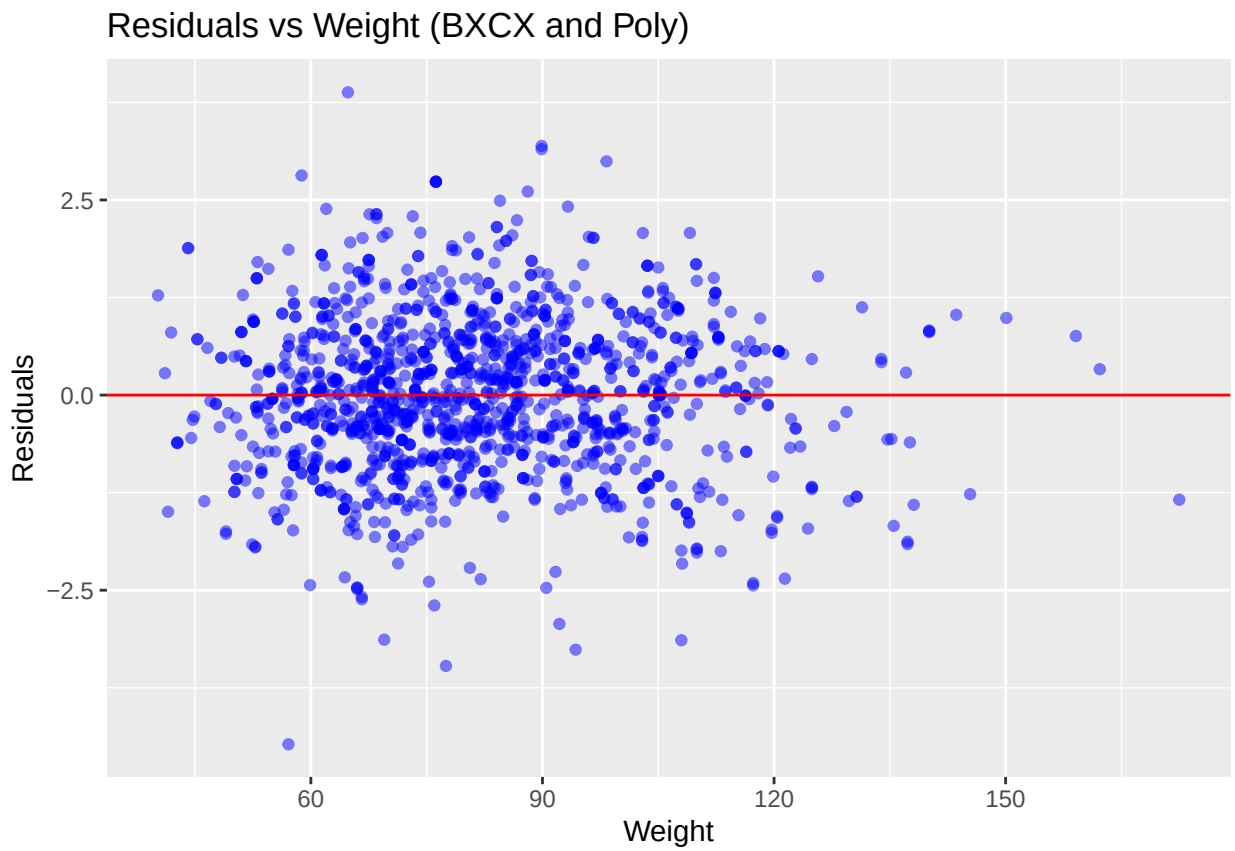
```r
#NORMAL QQ PLOT
qq_plot <- ggplot(data = data.frame(pb.residuals = pb.residuals),
                  aes(sample = pb.residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot (BXCX and Poly)",
       x = "Theoretical Quantiles", y = "Sample Quantiles")

print(qq_plot)
```



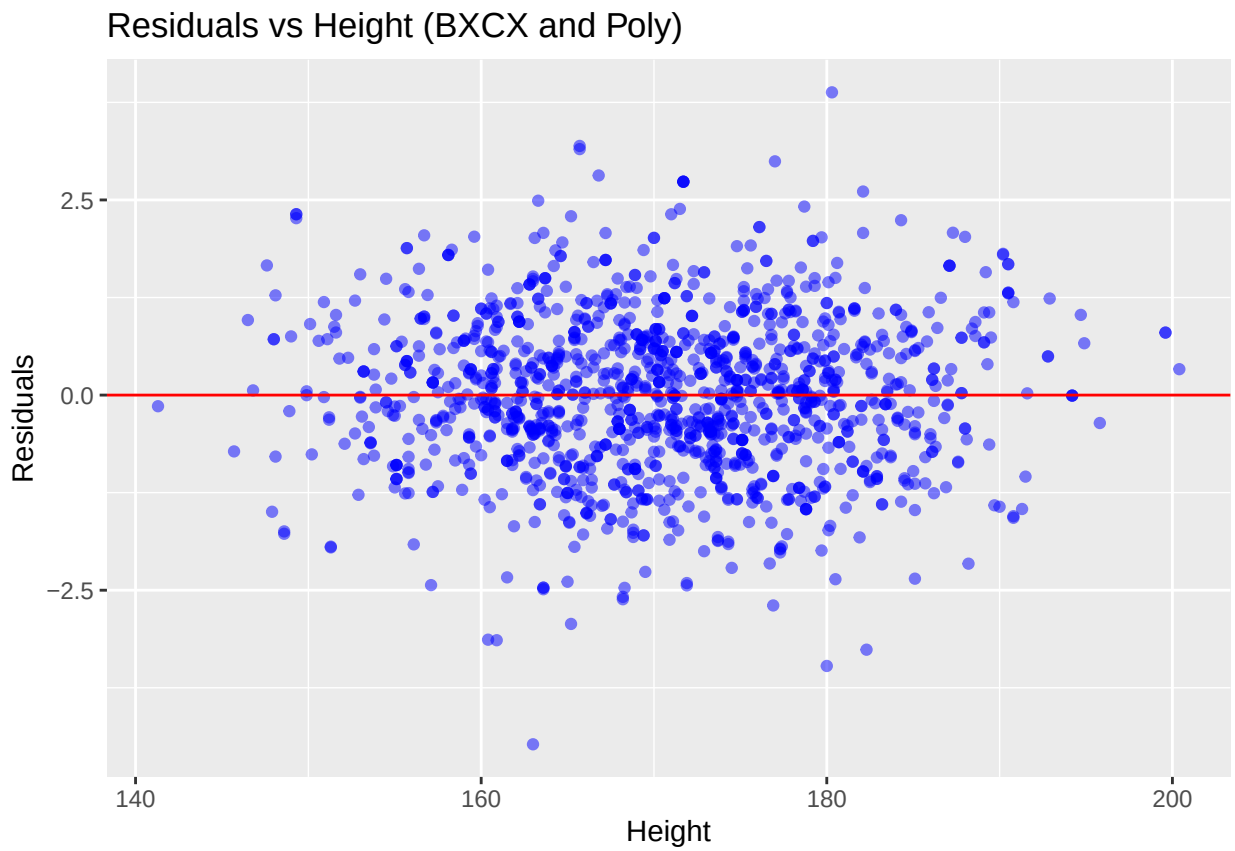

```r
#RESIDUALS VS AGE
res_age_plot <- ggplot(p.BXCX.frame,
                       aes(x = Age, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Age (BXCX and Poly)",
       x = "Age", y = "Residuals")

print(res_age_plot)
```



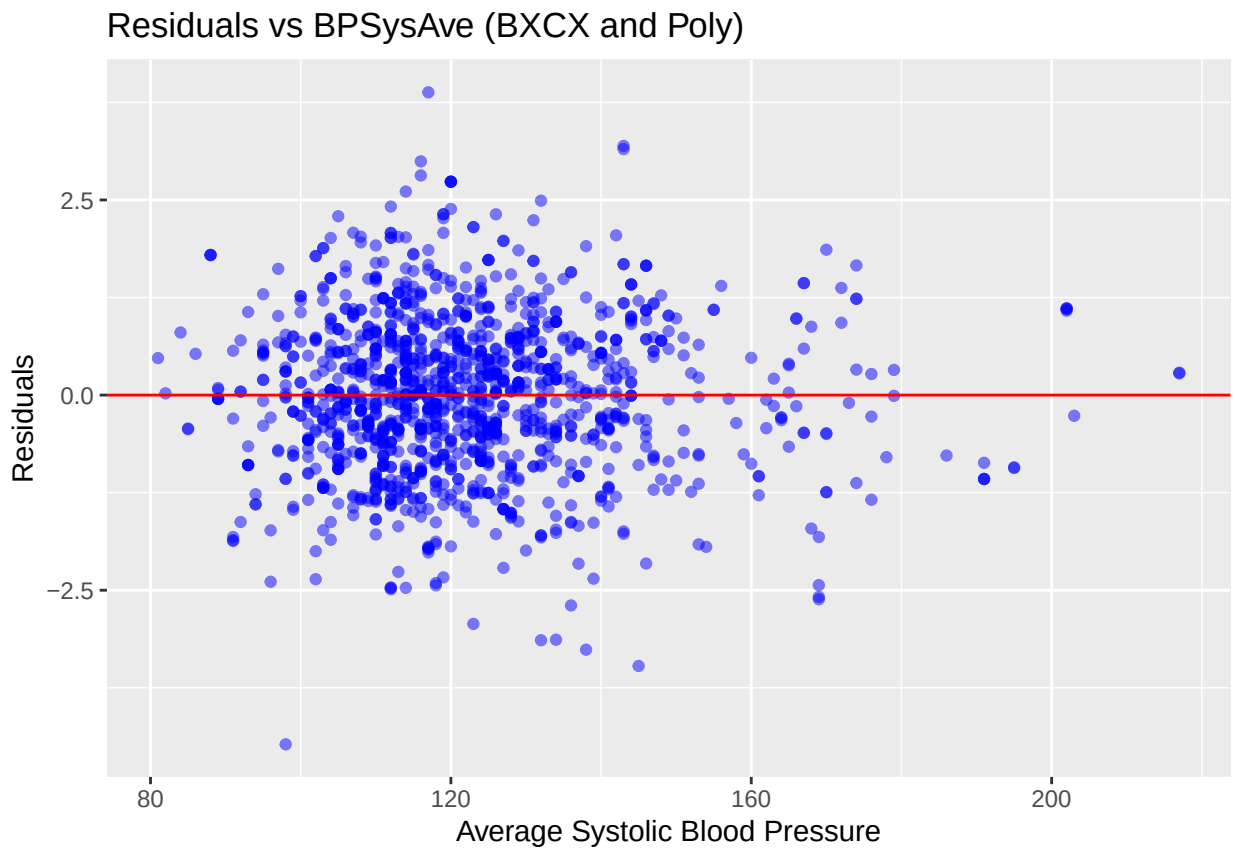

```
#RESIDUALS VS WEIGHT
res_weight_plot <- ggplot(p.BXCX.frame,
                          aes(x = Weight, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Weight (BXCX and Poly)",
       x = "Weight", y = "Residuals")

print(res_weight_plot)
```



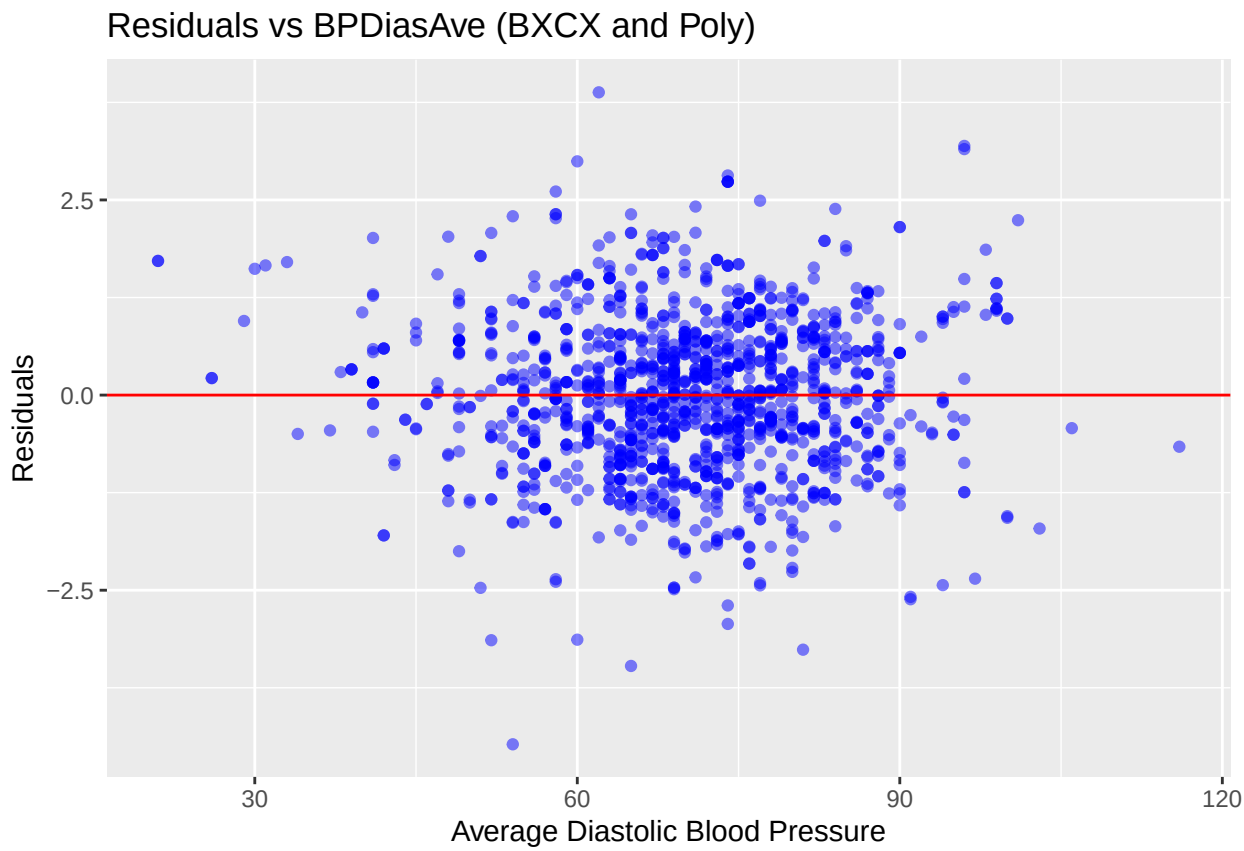

```
#RESIDUALS VS HEIGHT
res_height_plot <- ggplot(p.BXCX.frame,
                          aes(x = Height, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Height (BXCX and Poly)",
       x = "Height", y = "Residuals")

print(res_height_plot)
```



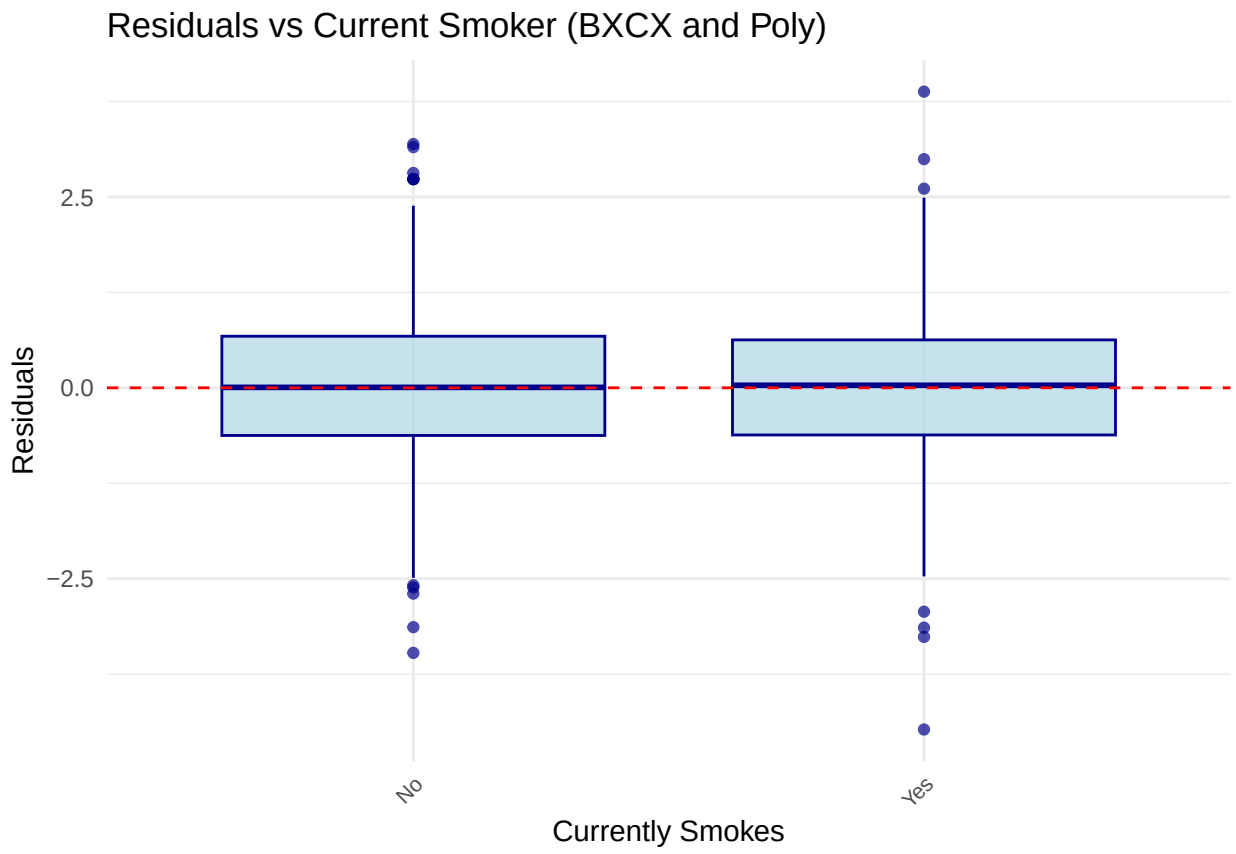

```
#RESIDUALS VS BPSysAve
res_BPSysAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPSysAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPSysAve (BXCX and Poly)",
       x = "Average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

print(res_BPSysAve_plot)
```



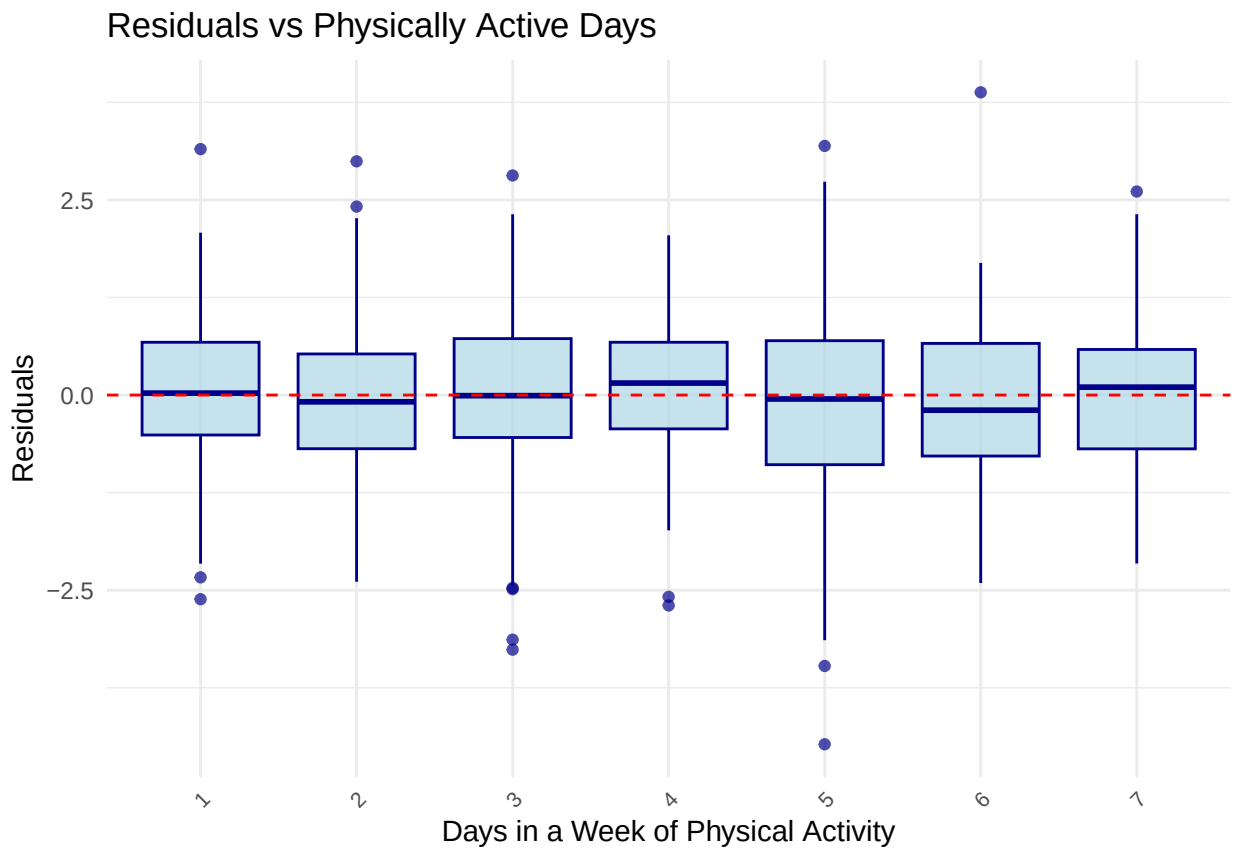

```
#RESIDUALS VS BPDiaAve
res_BPDiaAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPDiaAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPDiasAve (BXCX and Poly)",
       x = "Average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

print(res_BPDiaAve_plot)
```



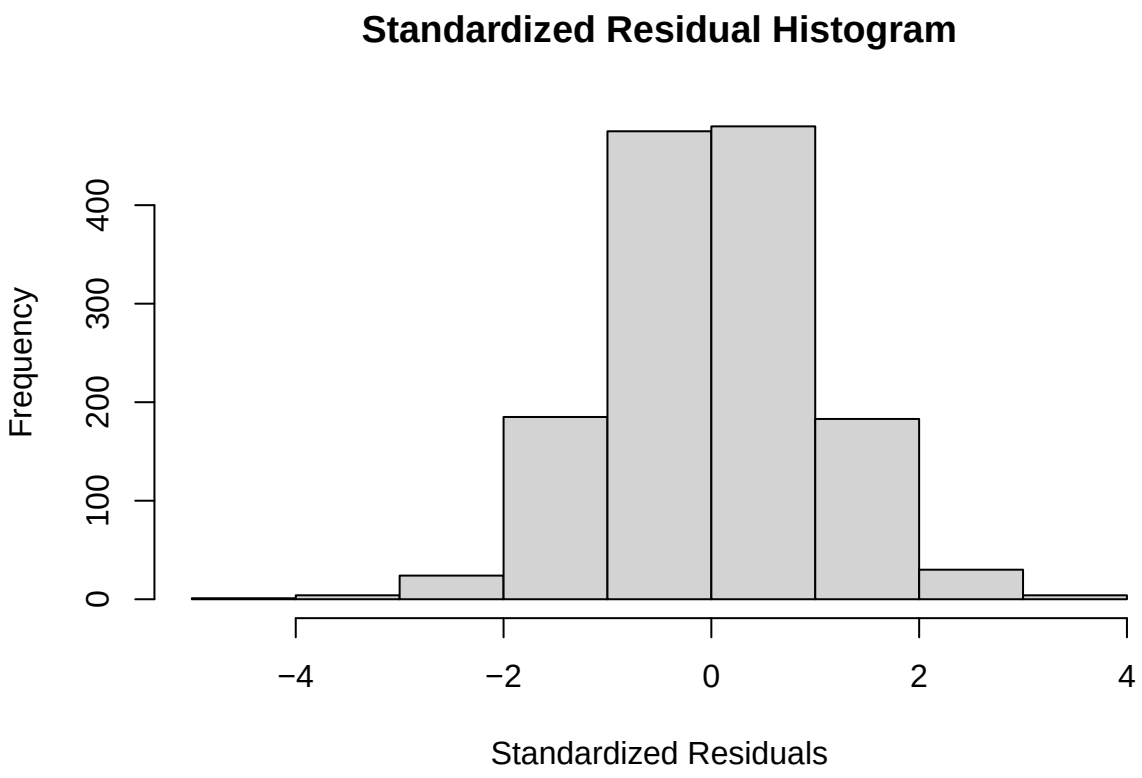

```
#RESIDUALS VS SmokeNow (BOXPLOT)
res_smoke_plot <- ggplot(
  p.BXCX.frame, aes(x = as.factor(SmokeNow), y = pb.residuals)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  ggtitle("Residuals 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)
```




```
#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")
```

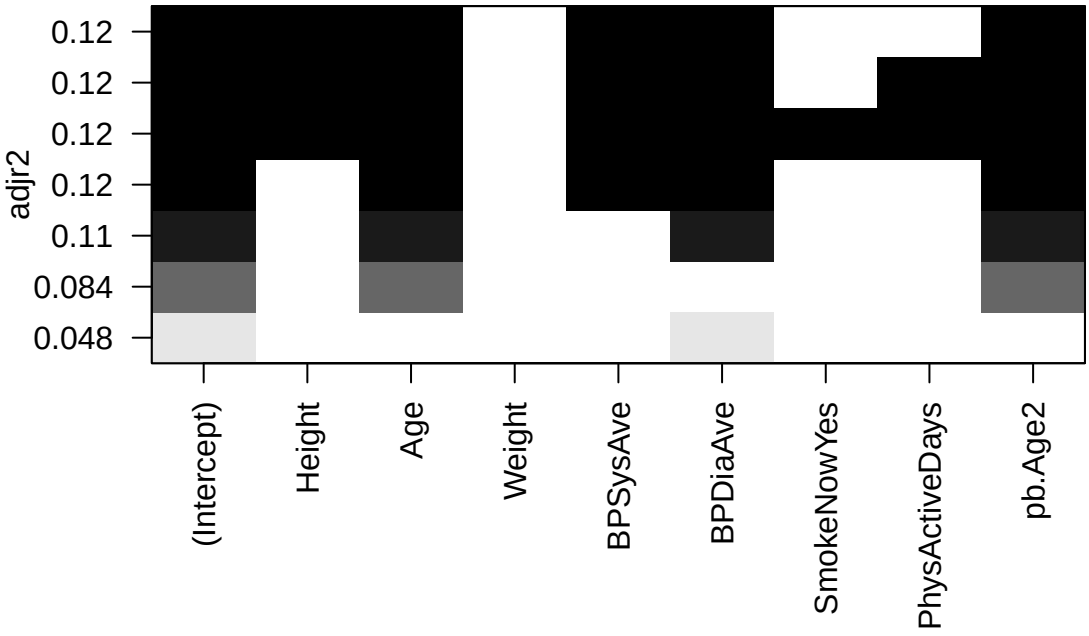



```
library(leaps)

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

summary(best_subset_p.BXCX)
```

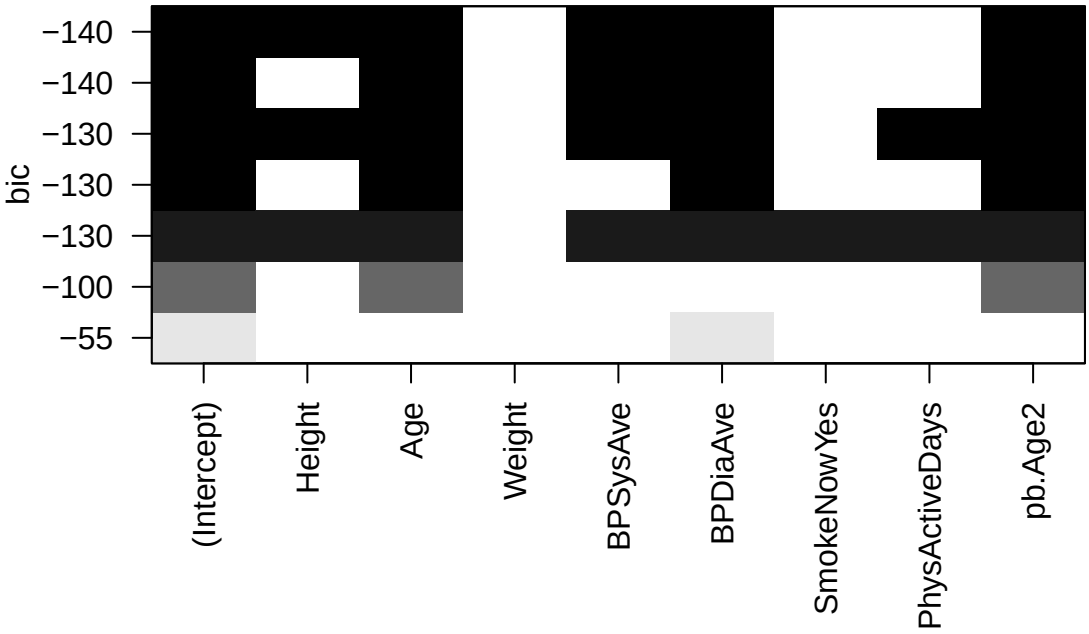
```
## Subset selection object
## Call: regsubsets.formula(pb.TotChol ~ ., data = p.BXCX.frame, nvmax = 7,
##     nbest = 1, really.big = TRUE, method = "exhaustive")
## 8 Variables  (and intercept)
##                Forced in Forced out
## Height             FALSE      FALSE
## Age                FALSE      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 ) "*"    "*" " "    "*"      "*"      "*"         "*"           "*"
```

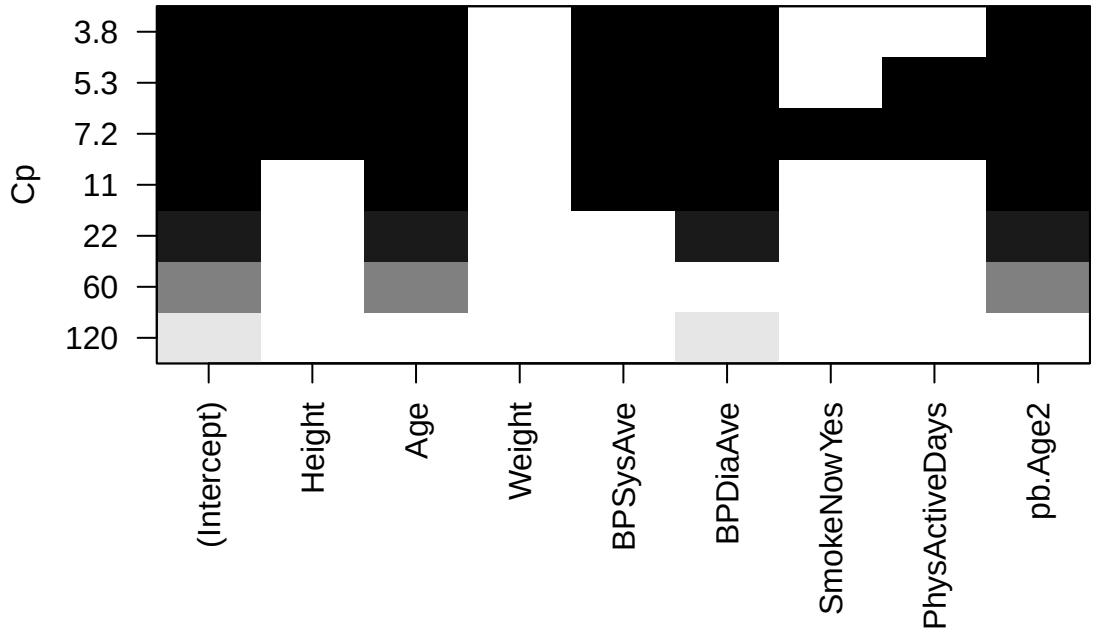
```
plot(best_subset_p.BXCX,scale='adjr2')
```




```
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=-35.84
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     SmokeNow + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - Weight          1     0.147 1333.3 -37.687
## - SmokeNow        1     0.190 1333.4 -37.642
## - PhysActiveDays  1     0.455 1333.7 -37.366
## <none>                        1333.2 -35.839
## - Height          1     6.583 1339.8 -31.012
## - BPSysAve        1    13.378 1346.6 -24.001
## - BPDiaAve        1    17.652 1350.8 -19.608
## - pb.Age2         1    79.723 1412.9  42.658
## - Age             1    88.611 1421.8  51.349
##
## Step:  AIC=-37.69
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + SmokeNow +
##     PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - SmokeNow        1     0.165 1333.5 -39.515
## - PhysActiveDays  1     0.437 1333.8 -39.232
## <none>                        1333.3 -37.687
```

```
## + Weight          1     0.147 1333.2 -35.839
## - Height          1     9.393 1342.7 -29.957
## - BPSysAve        1    13.235 1346.6 -25.996
## - BPDiaAve        1    17.506 1350.8 -21.607
## - pb.Age2         1    79.682 1413.0  40.762
## - Age             1    88.586 1421.9  49.468
##
## Step:  AIC=-39.51
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - PhysActiveDays  1     0.438 1333.9 -41.060
## <none>                        1333.5 -39.515
## + SmokeNow        1     0.165 1333.3 -37.687
## + Weight          1     0.122 1333.4 -37.642
## - Height          1     9.303 1342.8 -31.880
## - BPSysAve        1    13.132 1346.6 -27.932
## - BPDiaAve        1    17.710 1351.2 -23.229
## - pb.Age2         1    79.574 1413.1  38.818
## - Age             1    89.082 1422.6  48.113
##
## Step:  AIC=-41.06
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        1333.9 -41.060
## + PhysActiveDays  1     0.438 1333.5 -39.515
## + SmokeNow        1     0.166 1333.8 -39.232
## + Weight          1     0.106 1333.8 -39.170
## - Height          1     9.226 1343.2 -33.507
## - BPSysAve        1    13.058 1347.0 -29.558
## - BPDiaAve        1    17.783 1351.7 -24.705
## - pb.Age2         1    79.209 1413.2  36.889
## - Age             1    88.659 1422.6  46.127
```

```
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
## -4.4926 -0.6148  0.0095  0.6439  3.8566
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.3948047  0.5562642   7.901 5.63e-15 ***
## Age          0.1010867  0.0105551   9.577  < 2e-16 ***
## pb.Age2     -0.0009657  0.0001067  -9.052  < 2e-16 ***
## Height      -0.0088162  0.0028536  -3.089 0.002045 **
## BPSysAve     0.0066715  0.0018151   3.675 0.000247 ***
## BPDiaAve     0.0108831  0.0025373   4.289 1.92e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9832 on 1380 degrees of freedom
## Multiple R-squared:  0.125,  Adjusted R-squared:  0.1218
## F-statistic: 39.41 on 5 and 1380 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

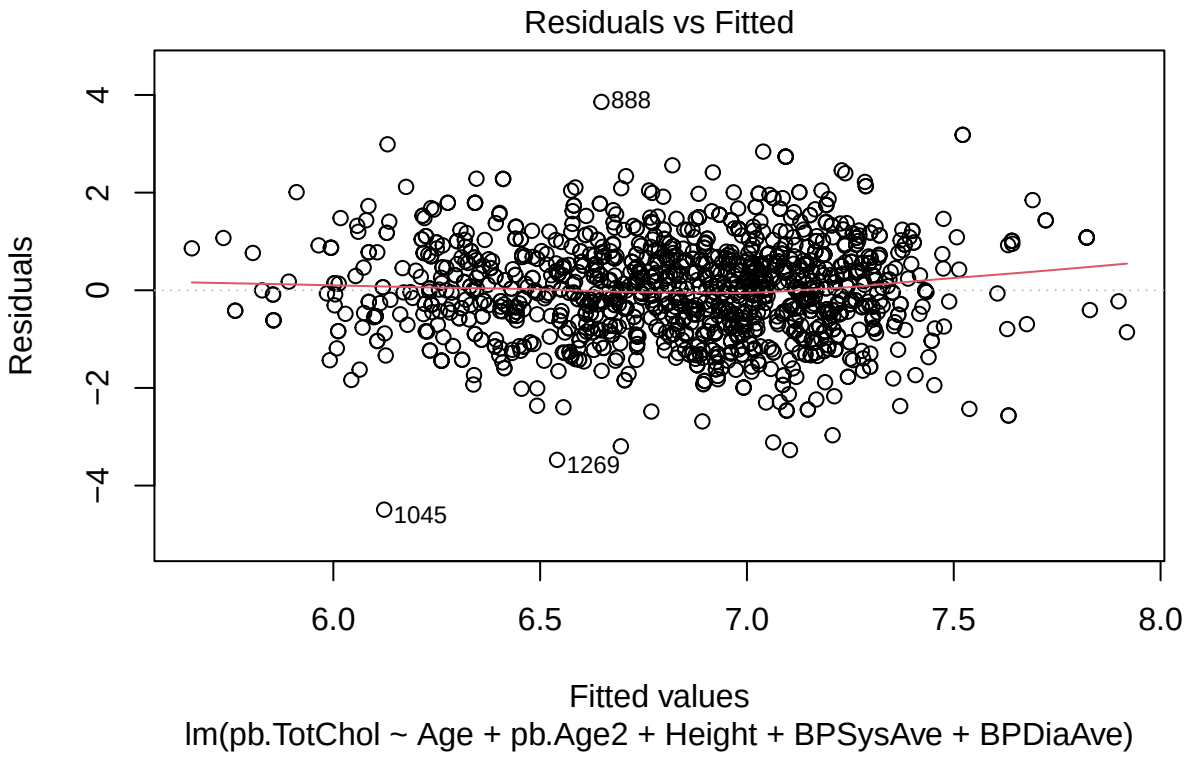
Check if the absolute values of the standardized residuals are greater than 3

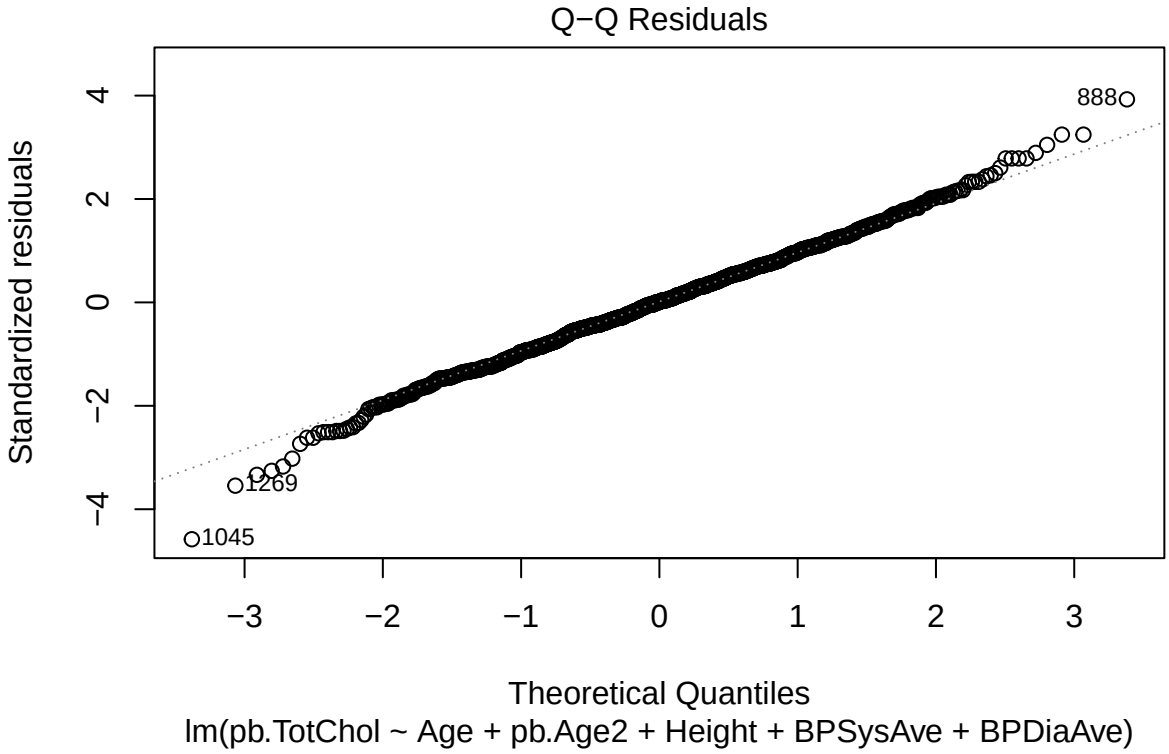
st.residuals <- rstandard(reduced.model) ## standardized residuals st.residuals

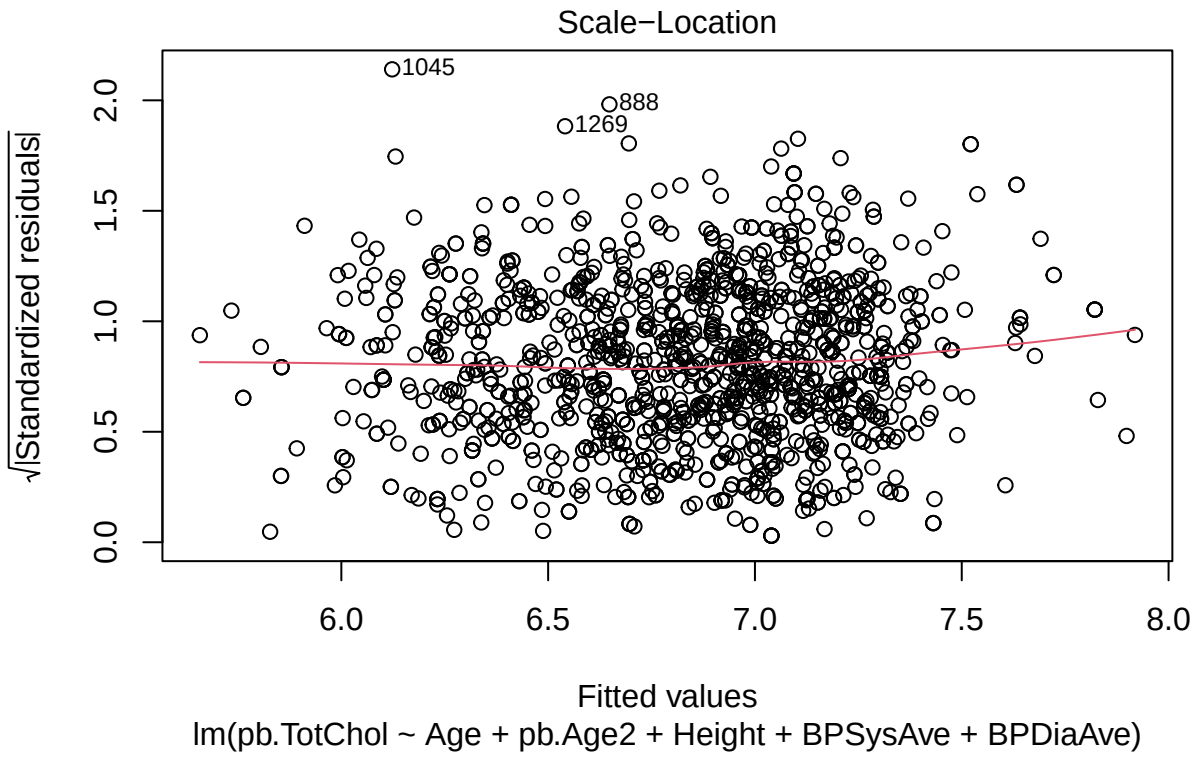
Outliers

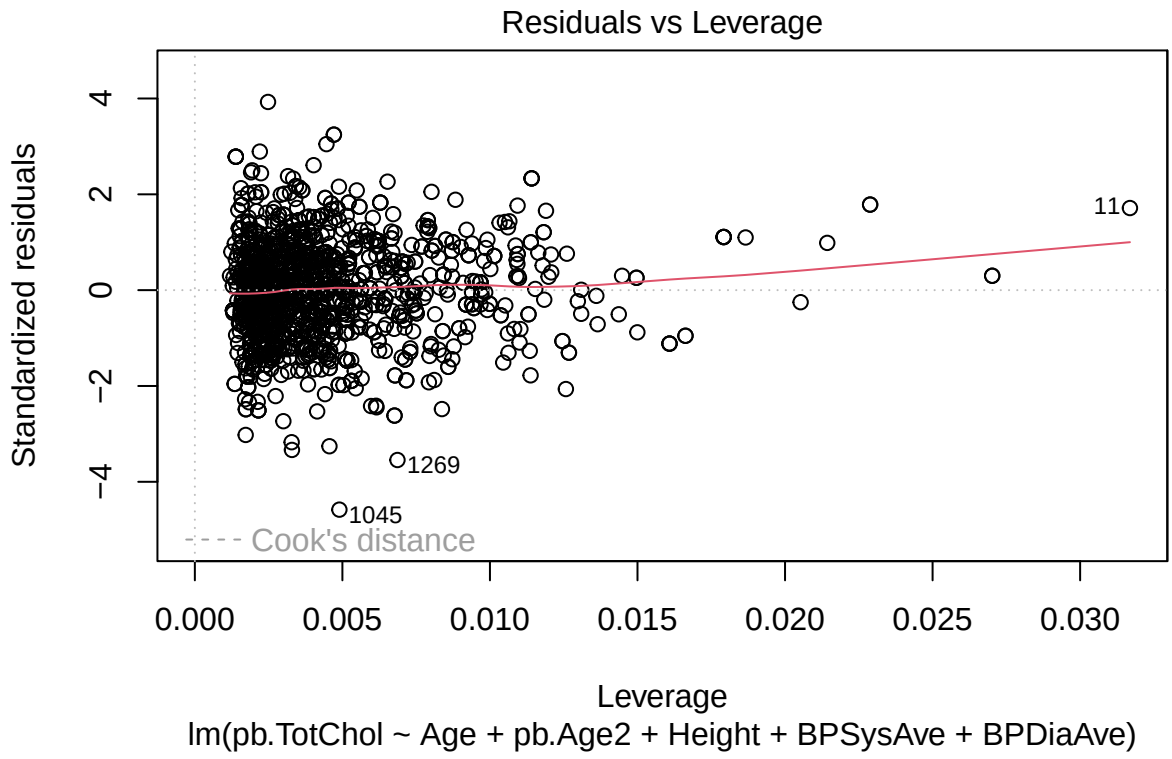
outliers_index <- which(abs(st.residuals)>3) outliers_index

```
#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)
888
1269
1045
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
888
1269
1045
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1045
888
1269
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(11, 1045, 1269)
p.BXCX.frame[influential_points, ]
```

```
##      Height Age Weight BPSysAve BPDiaAve SmokeNow PhysActiveDays pb.Age2
## 11    147.6  61   61.8      174       31      Yes              3    3721
## 1045  163.0  25   57.1       98       54      Yes              5     625
## 1269  180.0  77   77.5      145       65       No              5    5929
##      pb.TotChol
## 11     8.822565
## 1045   1.630341
## 1269   3.067699
```

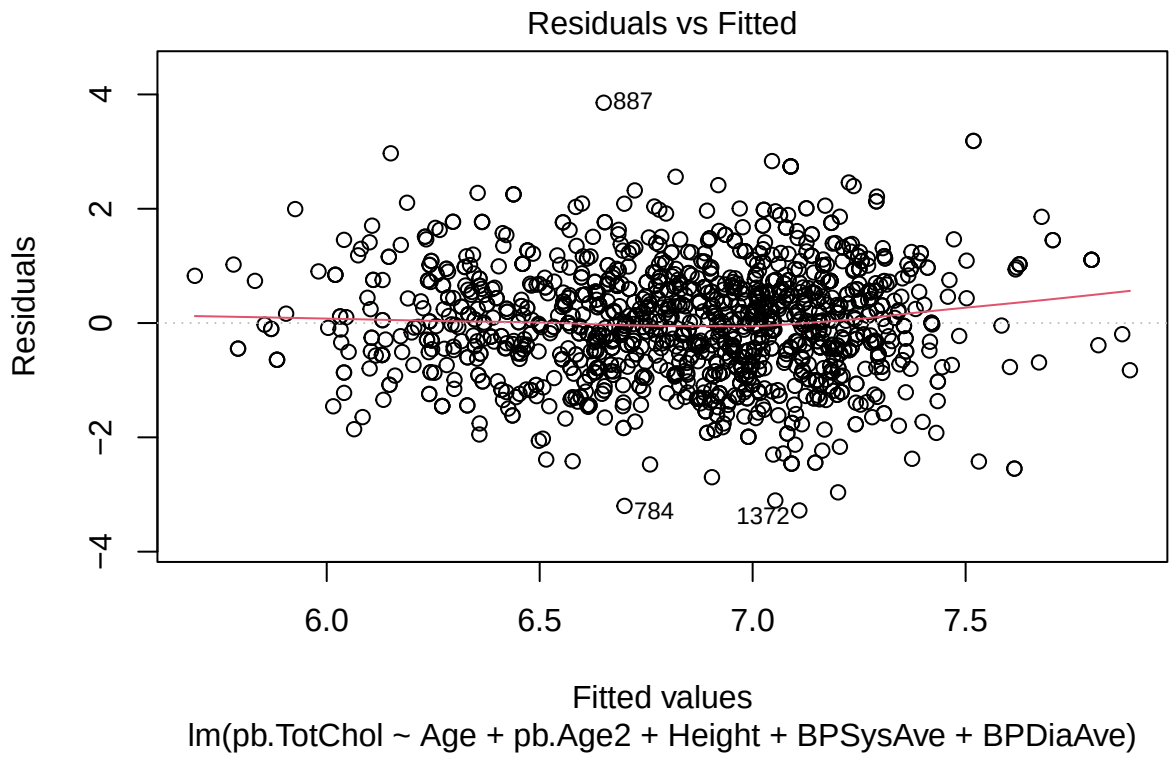
```
reduced.frame <- p.BXCX.frame %>%
  dplyr::filter(!row_number() %in% influential_points)

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

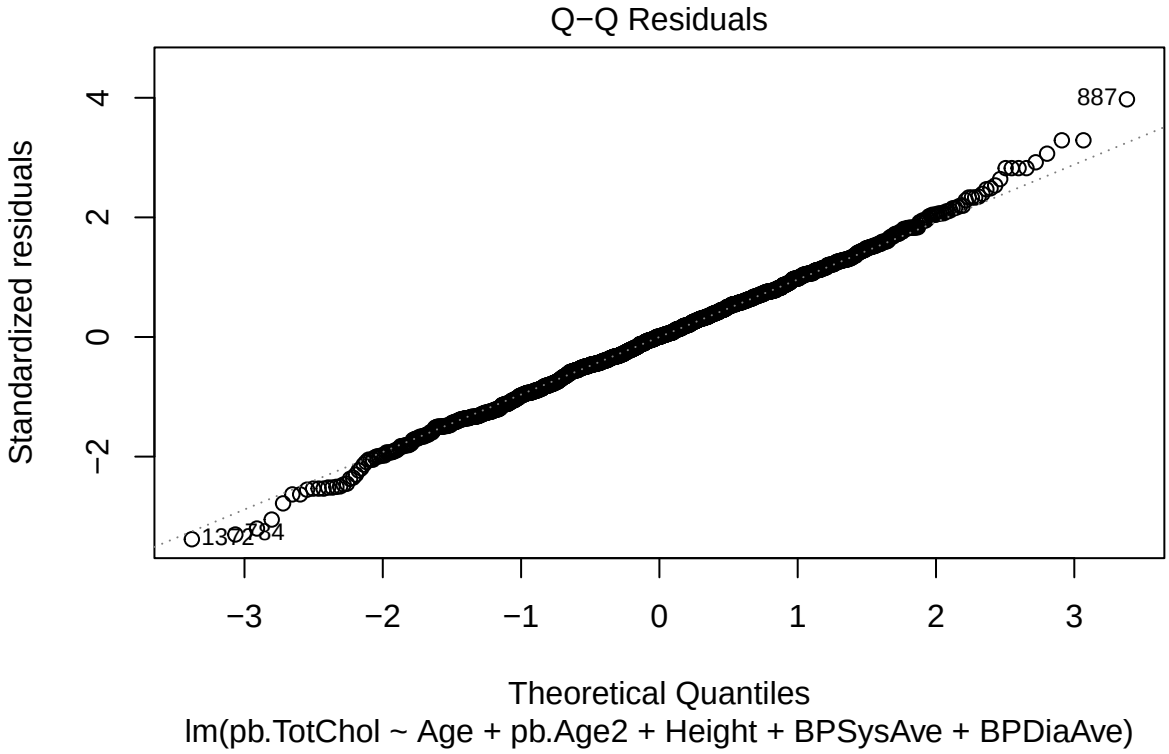
```
summary(final.model)
```

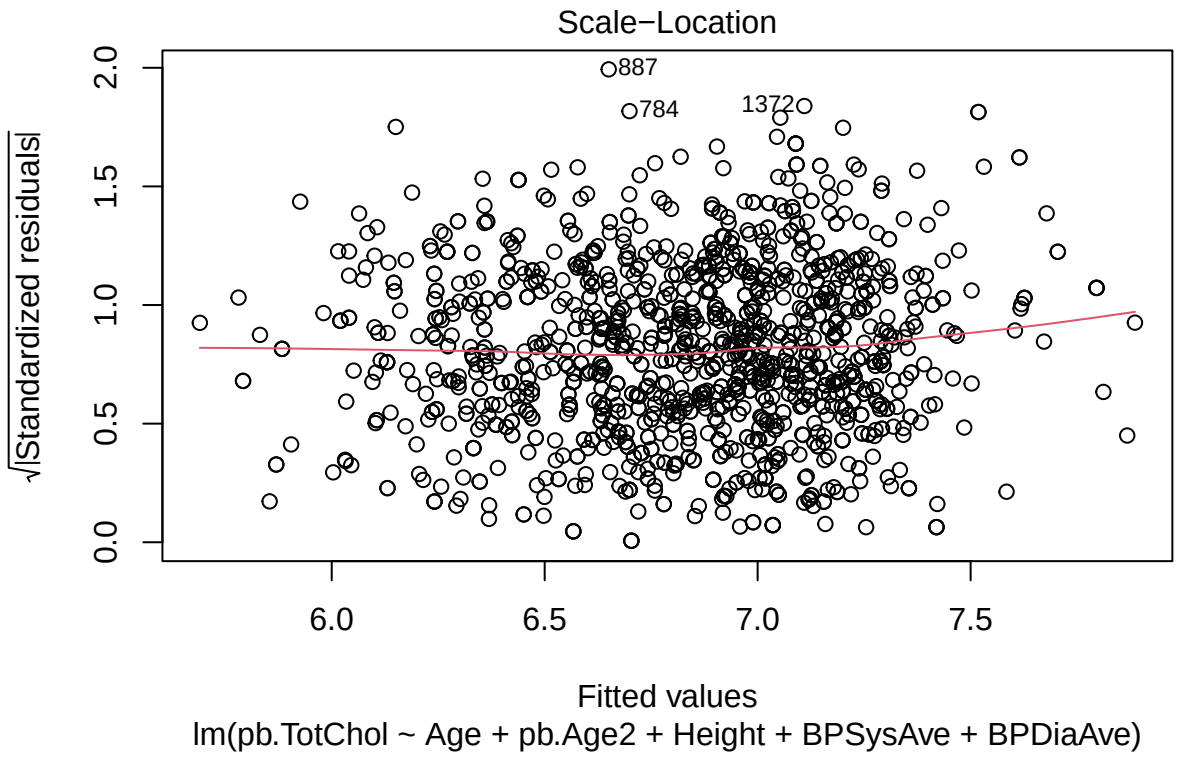
```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve +
```

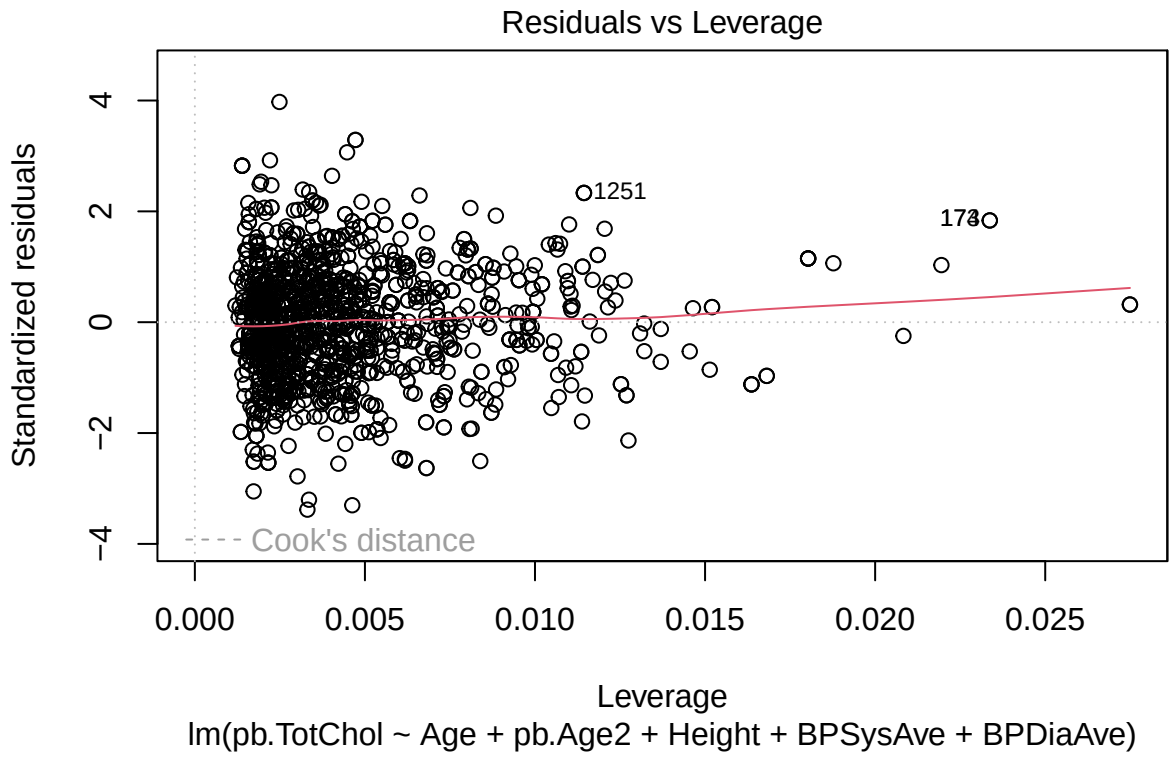
```
##     BPDiaAve, data = reduced.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2786 -0.6295  0.0051  0.6279  3.8544
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.4544920  0.5503688   8.094 1.26e-15 ***
## Age          0.0968570  0.0104629   9.257  < 2e-16 ***
## pb.Age2     -0.0009212  0.0001058  -8.705  < 2e-16 ***
## Height      -0.0084156  0.0028267  -2.977 0.002960 **
## BPSysAve     0.0061796  0.0018100   3.414 0.000658 ***
## BPDiaAve     0.0112450  0.0025318   4.441 9.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9712 on 1377 degrees of freedom
## Multiple R-squared:  0.1219, Adjusted R-squared:  0.1187
## F-statistic: 38.25 on 5 and 1377 DF,  p-value: < 2.2e-16
```

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

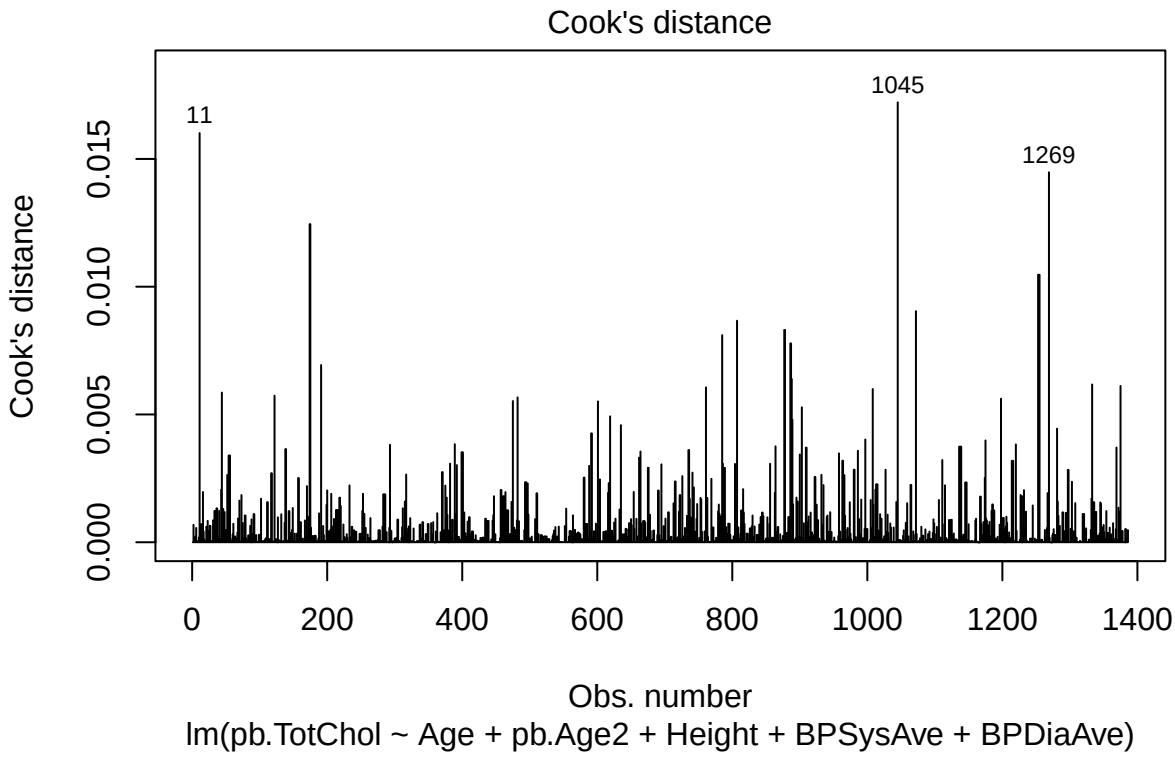



Q−Q Residuals
Standardized residuals
4
2
0
−2
887
1372
784
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)

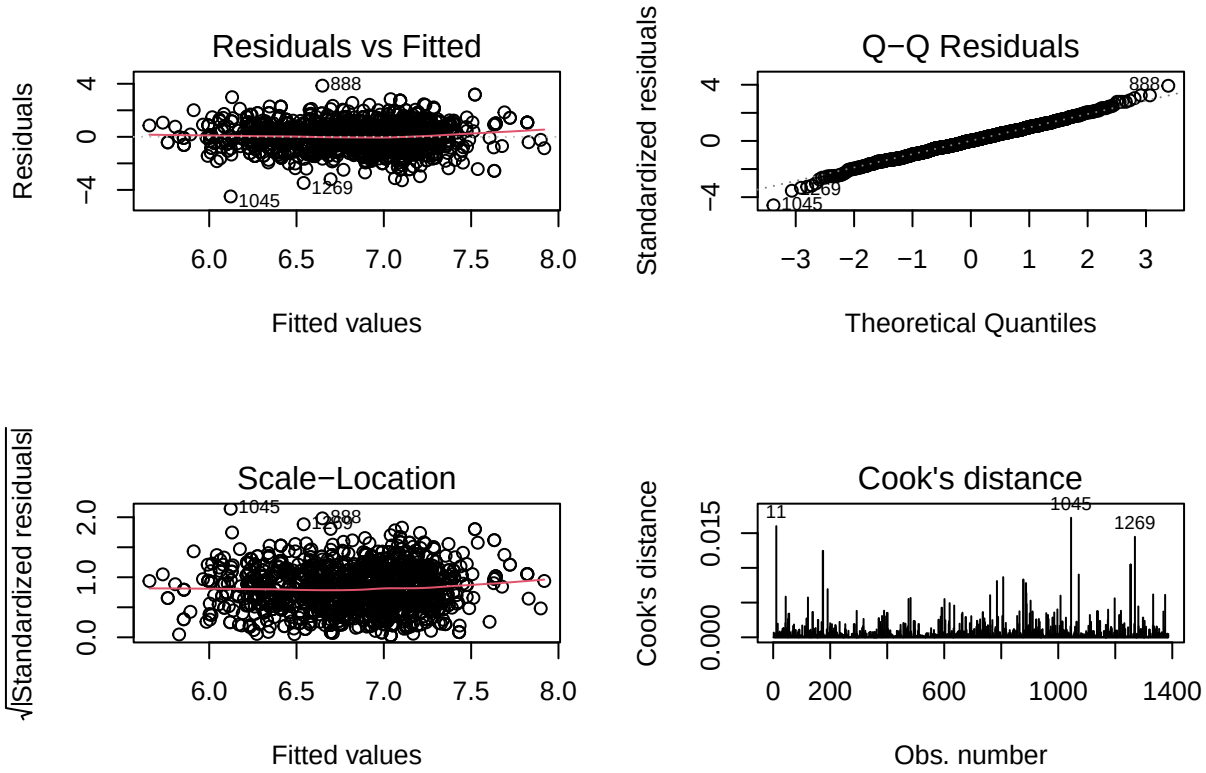

Scale−Location
887
784
1372
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
6.0
6.5
7.0
7.5
Fitted values
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)



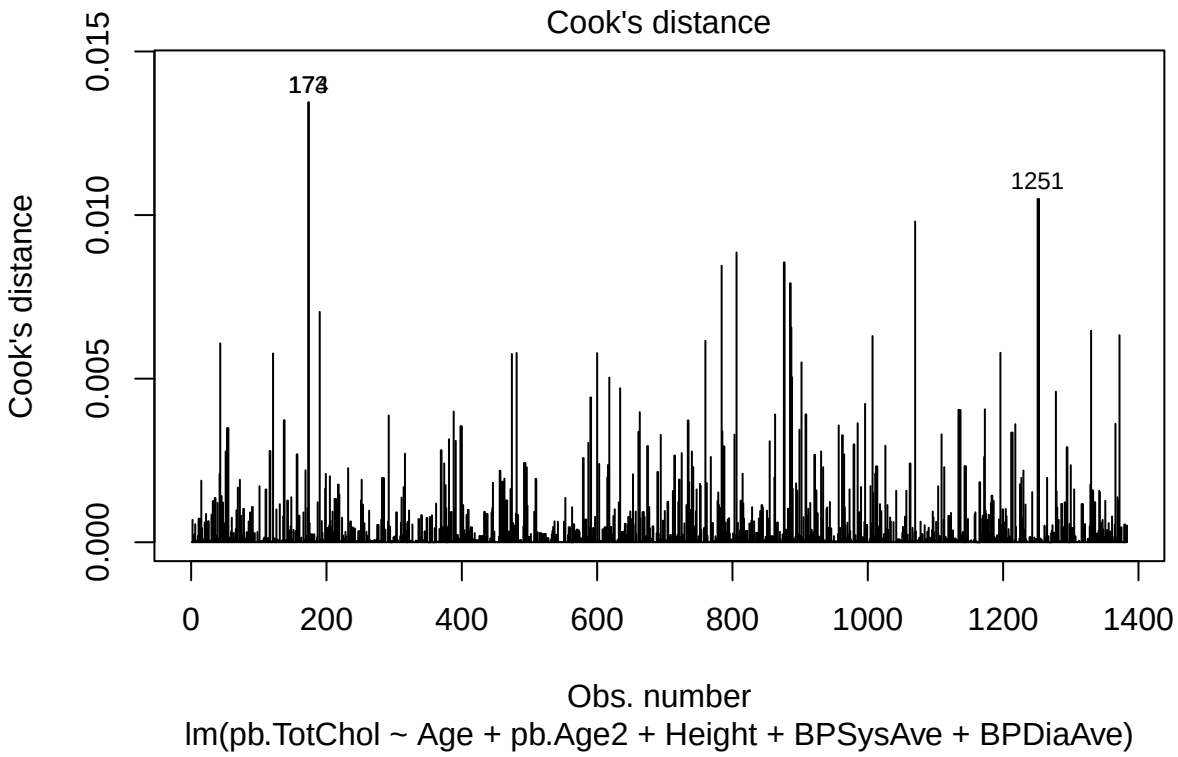

```
cooks <- cooks.distance(reduced.model)
plot(reduced.model, which = 4)
```



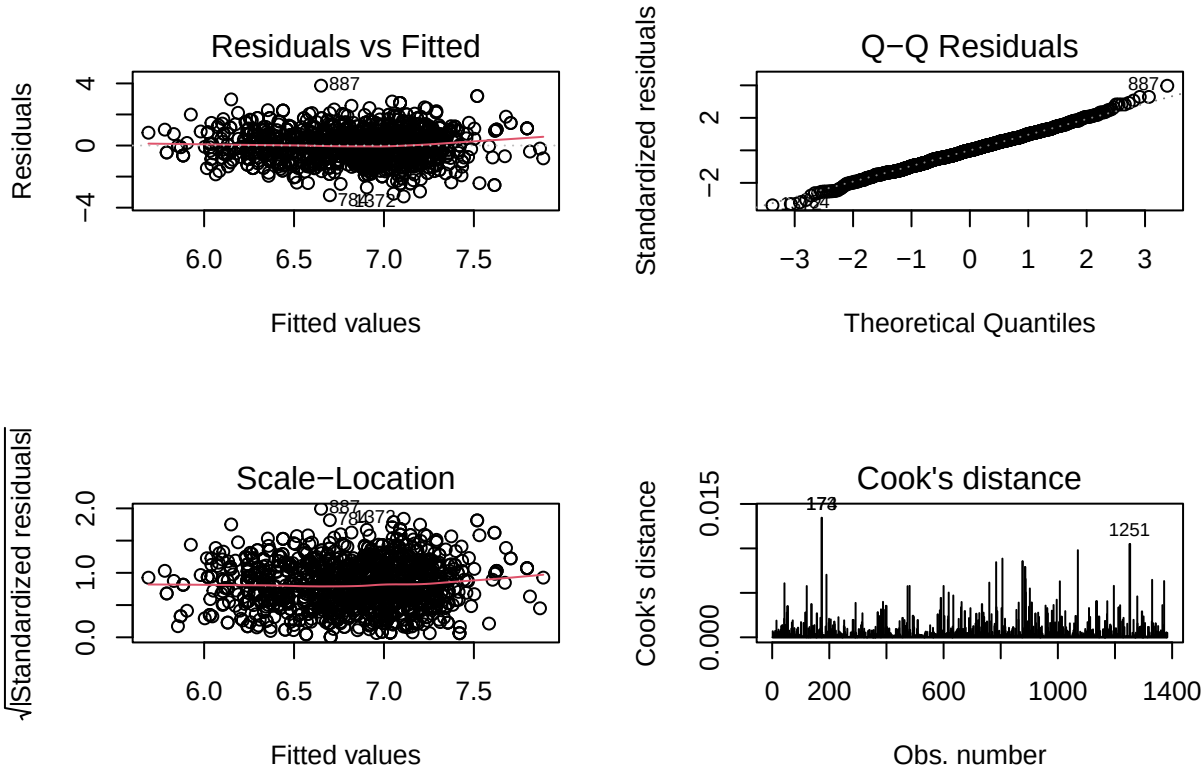

```
# All diagonistic plots
par(mfrow = c(2, 2))
plot(reduced.model, which = c(1, 2, 3, 4))
```




```
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))
```


Residuals vs Fitted
Residuals
Fitted values
887
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Cook's distance
1251
Obs. number