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
```

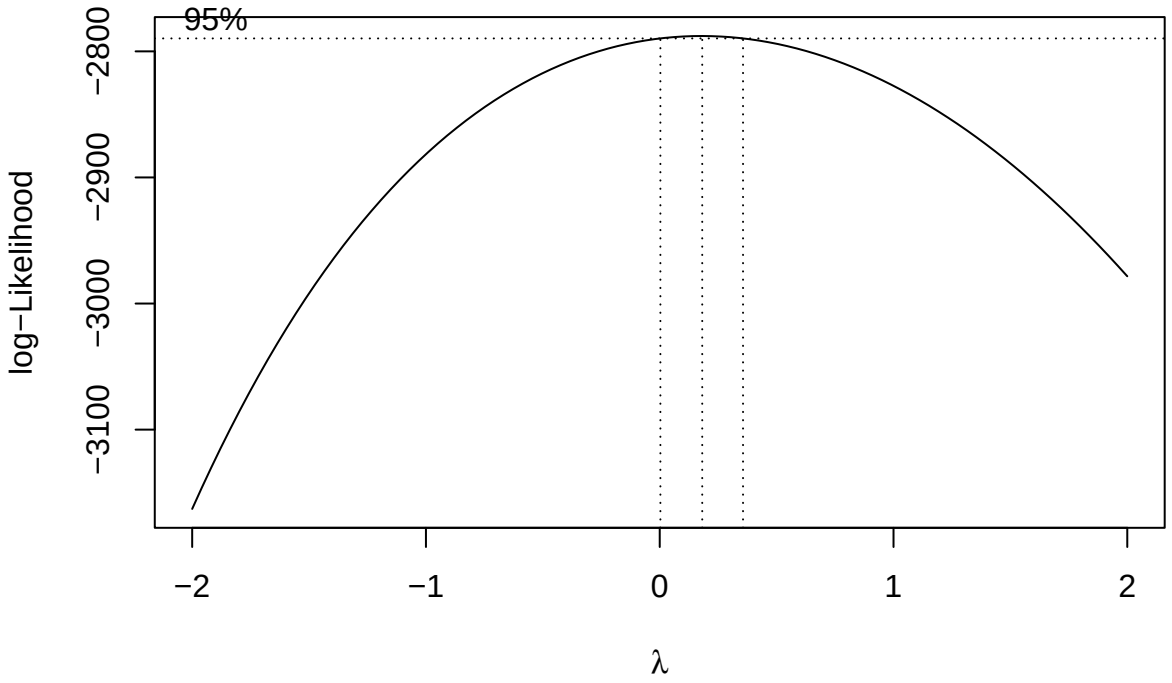
```
## [1] 7094
```

```
# 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()

# 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)
```

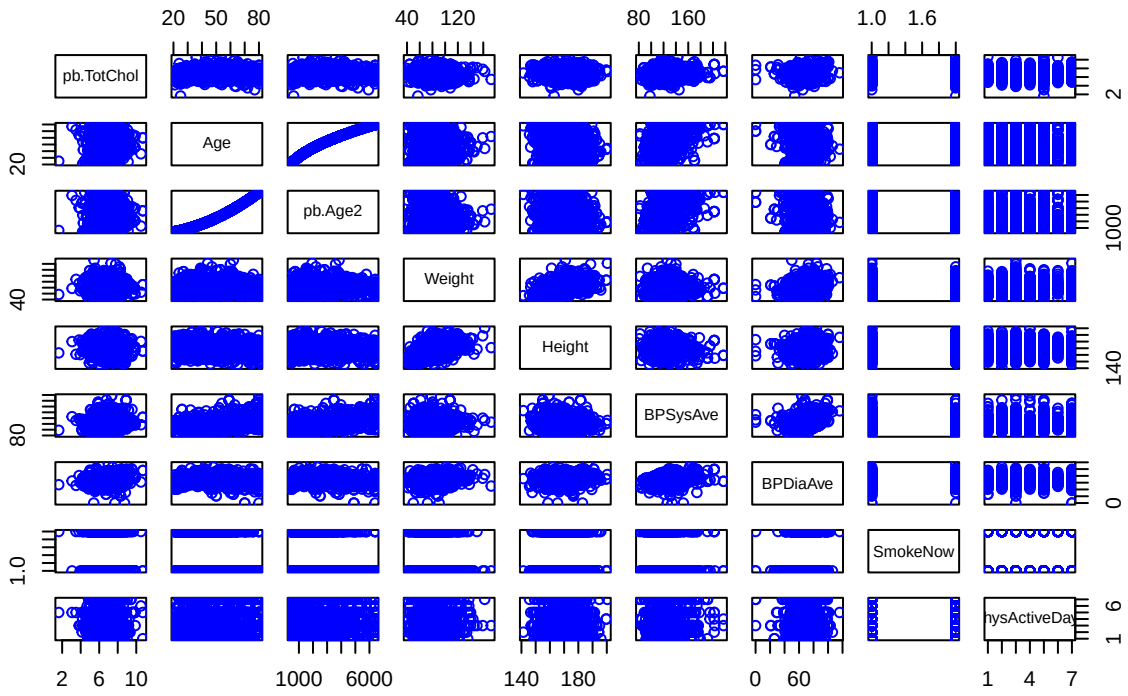



```
pb.lambda <- pb.b$x[which.max(pb.b$y)]

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

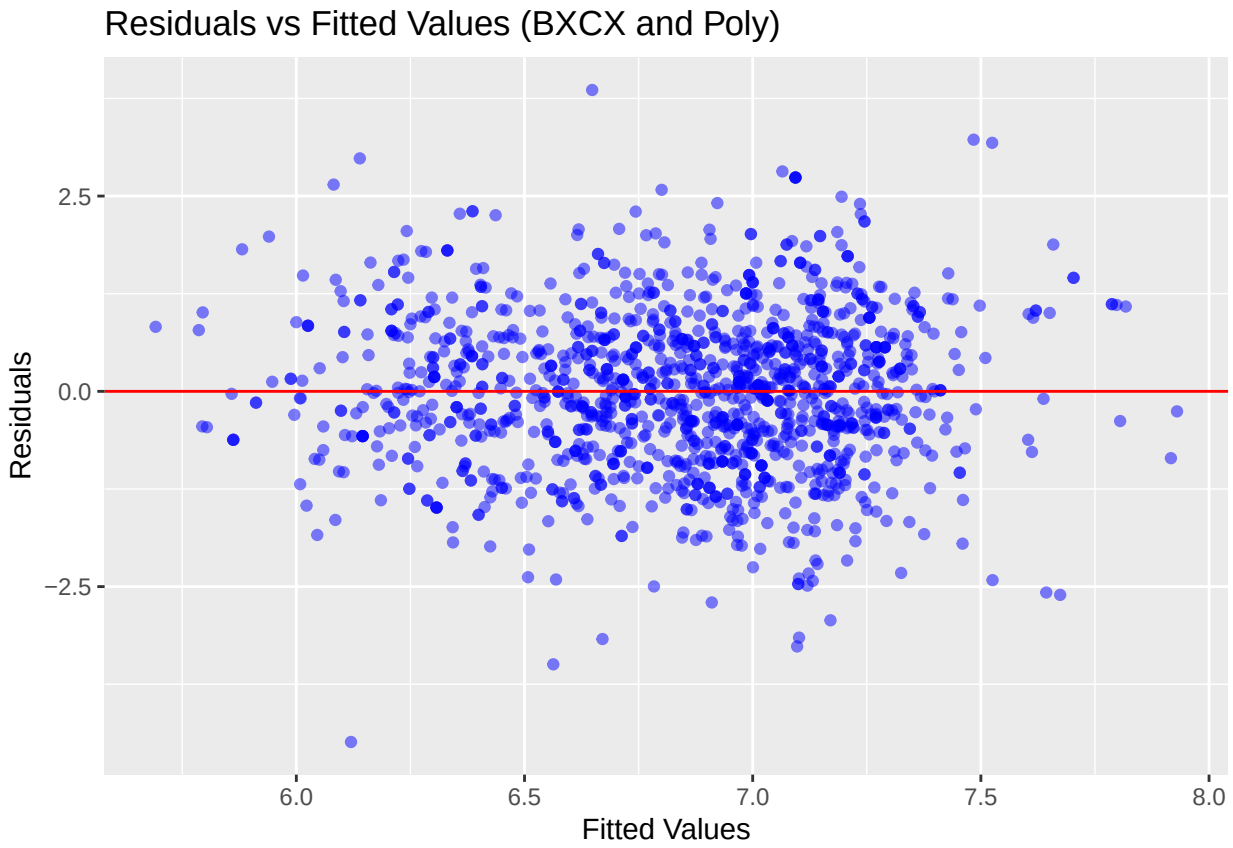
```
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve +
##     BPDiaAve + SmokeNow + PhysActiveDays, data = p.BXCX.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4892 -0.6298  0.0122  0.6579  3.8580
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.3570886  0.5883660   7.405 2.27e-13 ***
## Age             0.1023201  0.0106377   9.619  < 2e-16 ***
## pb.Age2        -0.0009777  0.0001074  -9.107  < 2e-16 ***
## Weight         -0.0007650  0.0016022  -0.477 0.633088
## Height         -0.0078447  0.0032064  -2.447 0.014545 *
## BPSysAve        0.0070900  0.0018160   3.904 9.91e-05 ***
## BPDiaAve        0.0095125  0.0024571   3.871 0.000113 ***
## SmokeNowYes    -0.0196813  0.0575264  -0.342 0.732308
## PhysActiveDays -0.0101337  0.0148490  -0.682 0.495070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9873 on 1380 degrees of freedom
## Multiple R-squared:  0.1215, Adjusted R-squared:  0.1164
## F-statistic: 23.85 on 8 and 1380 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")
```

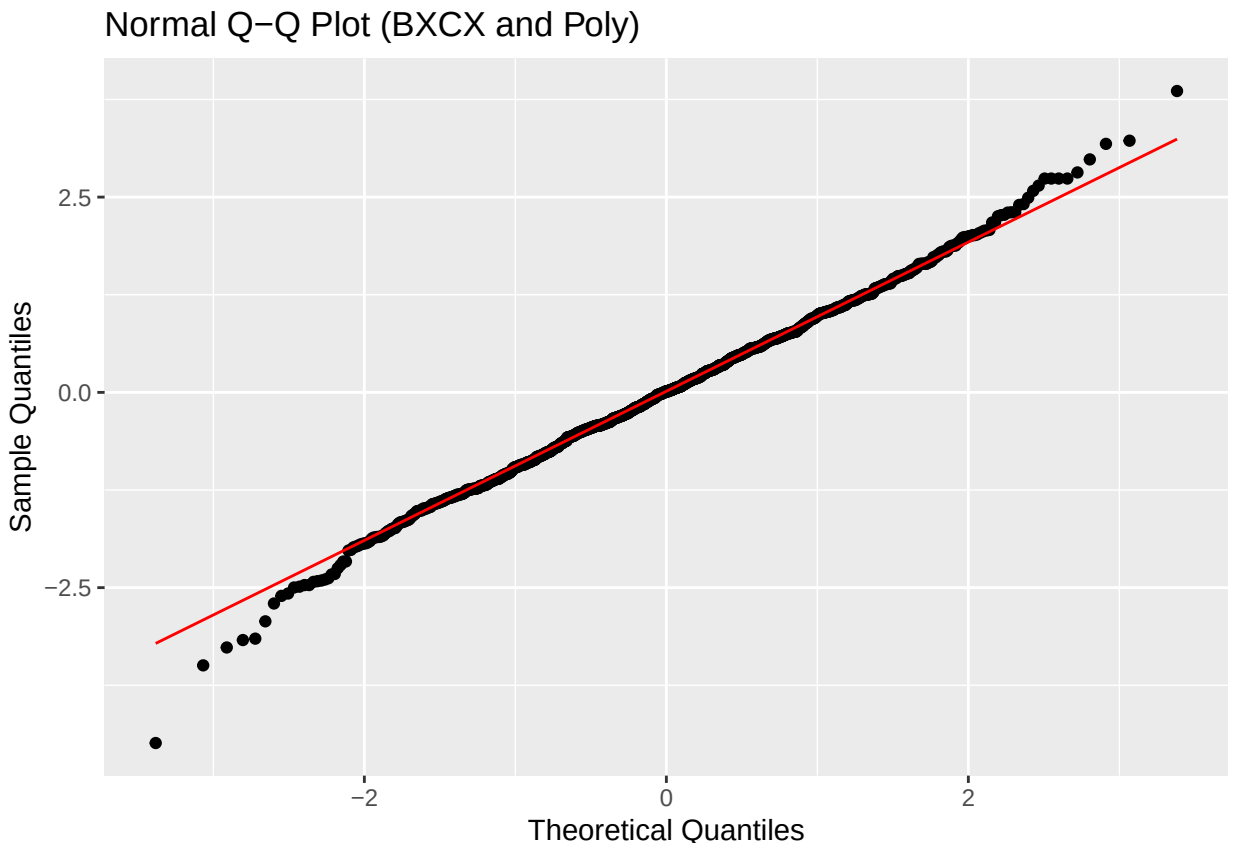


Residual Plots

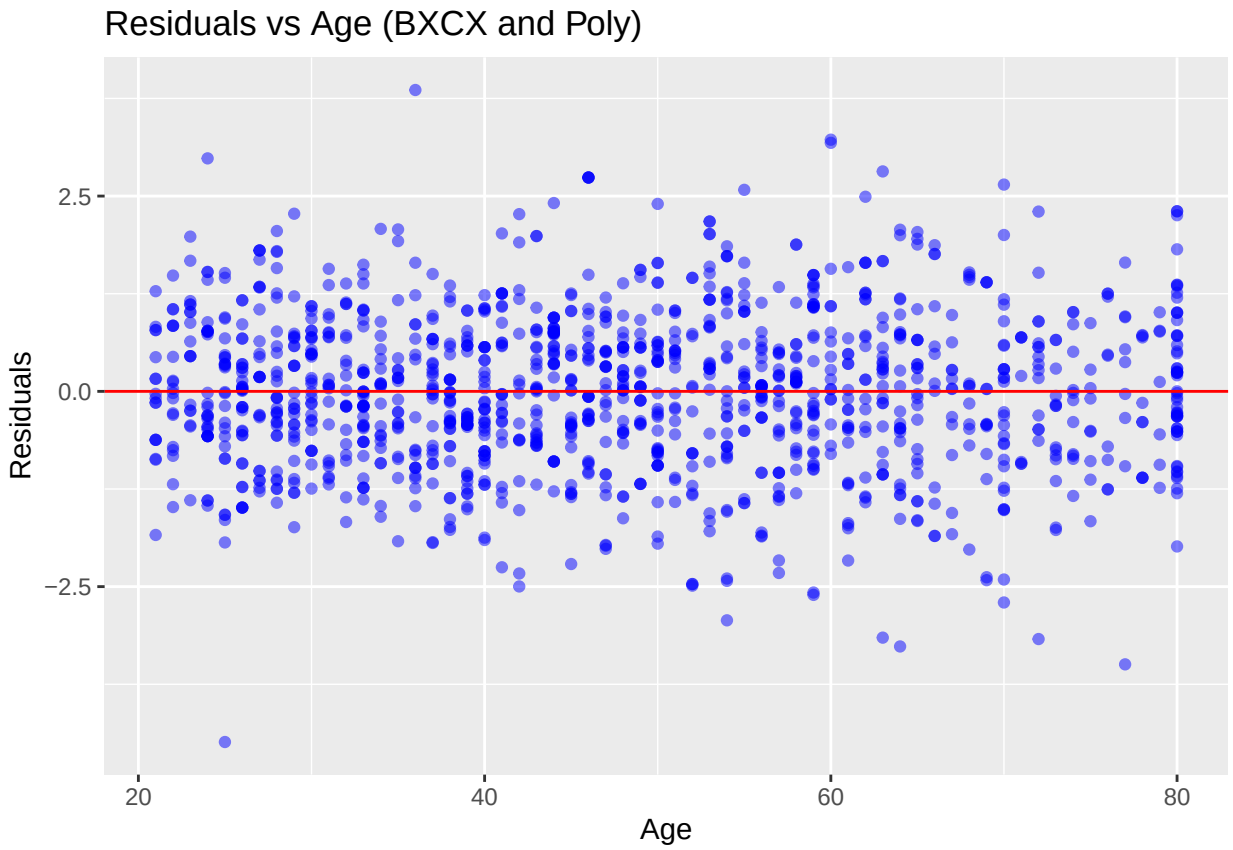
```
#RESIDUALS VS FITTED
res_fitted_plot <- ggplot(data = pb.plot_data,
                          aes(x = pb.fitted, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Fitted Values (BXCX and Poly)",
       x = "Fitted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

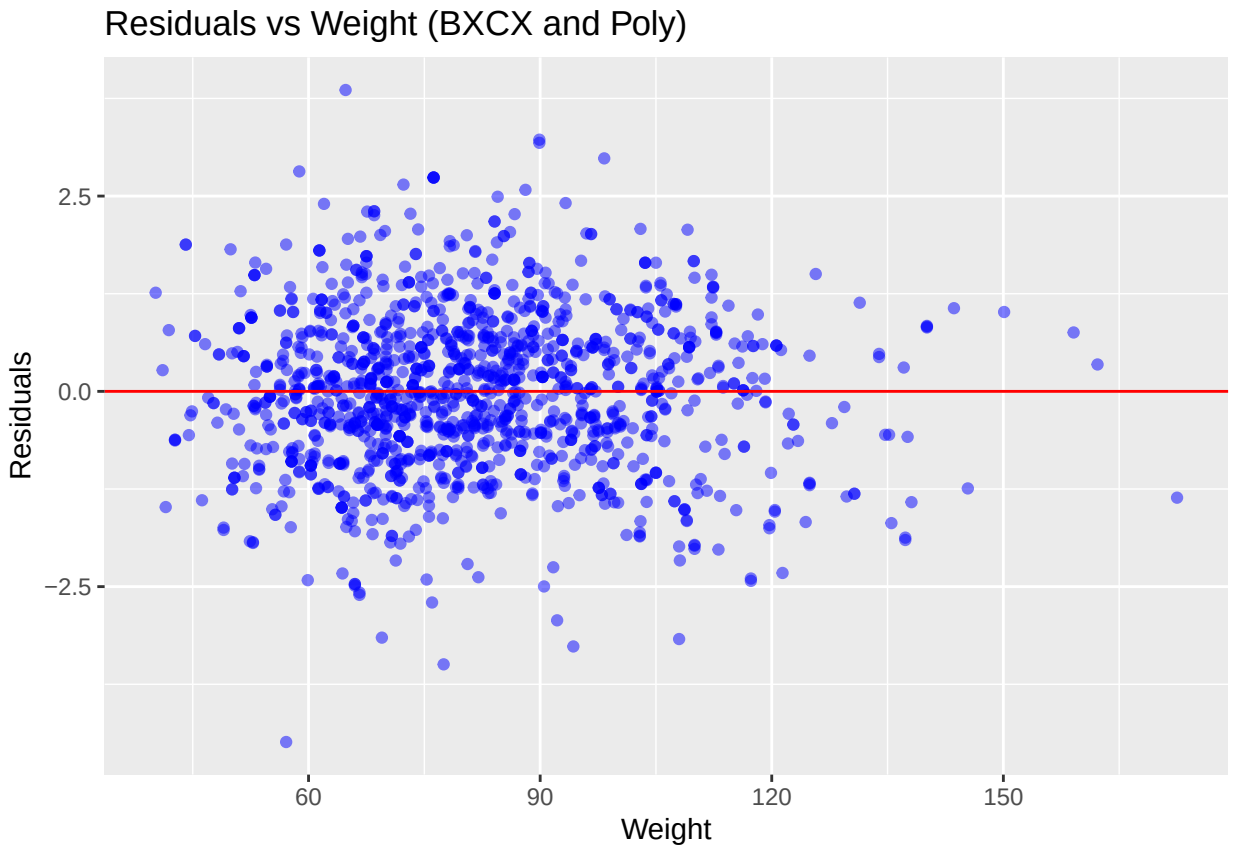
```r
#NORMAL QQ PLOT
qq_plot <- ggplot(data = data.frame(pb.residuals = pb.residuals),
                  aes(sample = pb.residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot (BXCX and Poly)",
       x = "Theoretical Quantiles", y = "Sample Quantiles")

print(qq_plot)
```



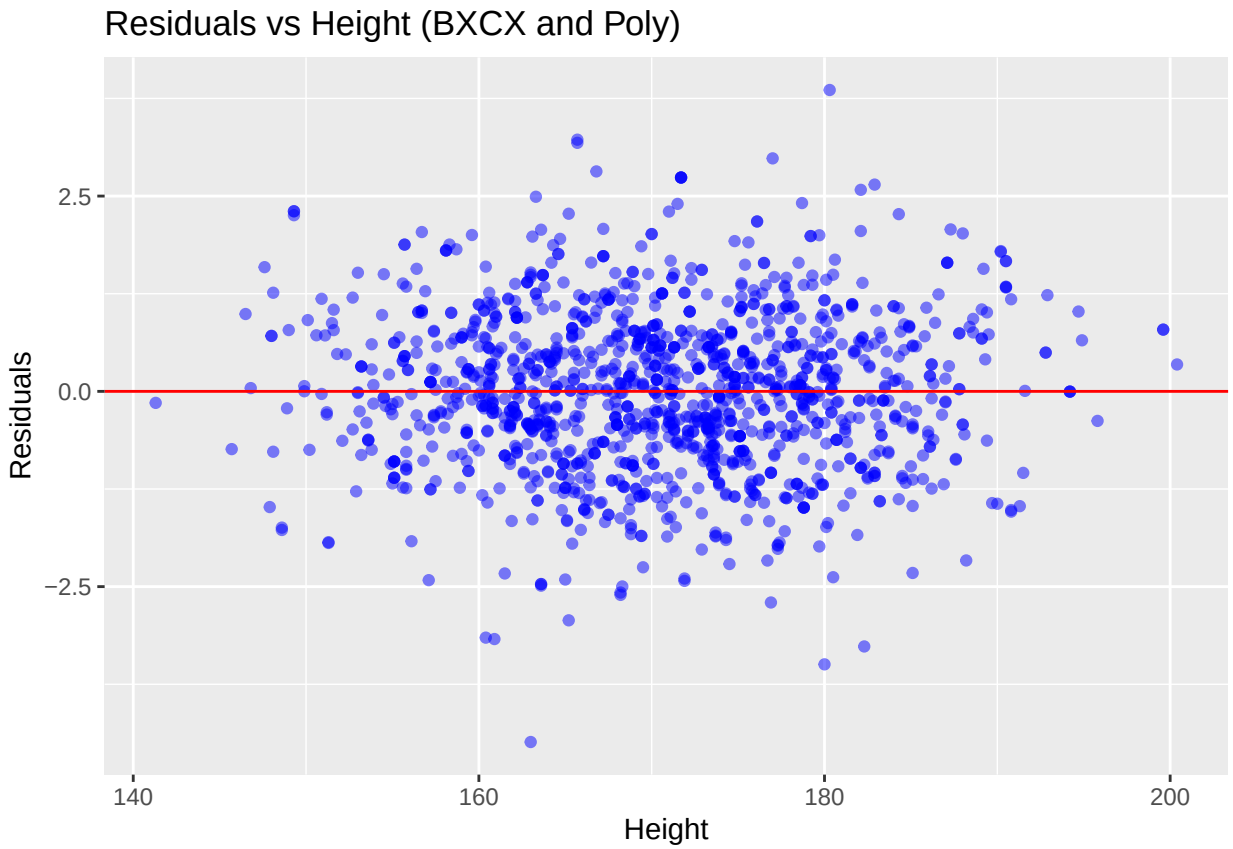

```r
#RESIDUALS VS AGE
res_age_plot <- ggplot(p.BXCX.frame,
                       aes(x = Age, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Age (BXCX and Poly)",
       x = "Age", y = "Residuals")

print(res_age_plot)
```



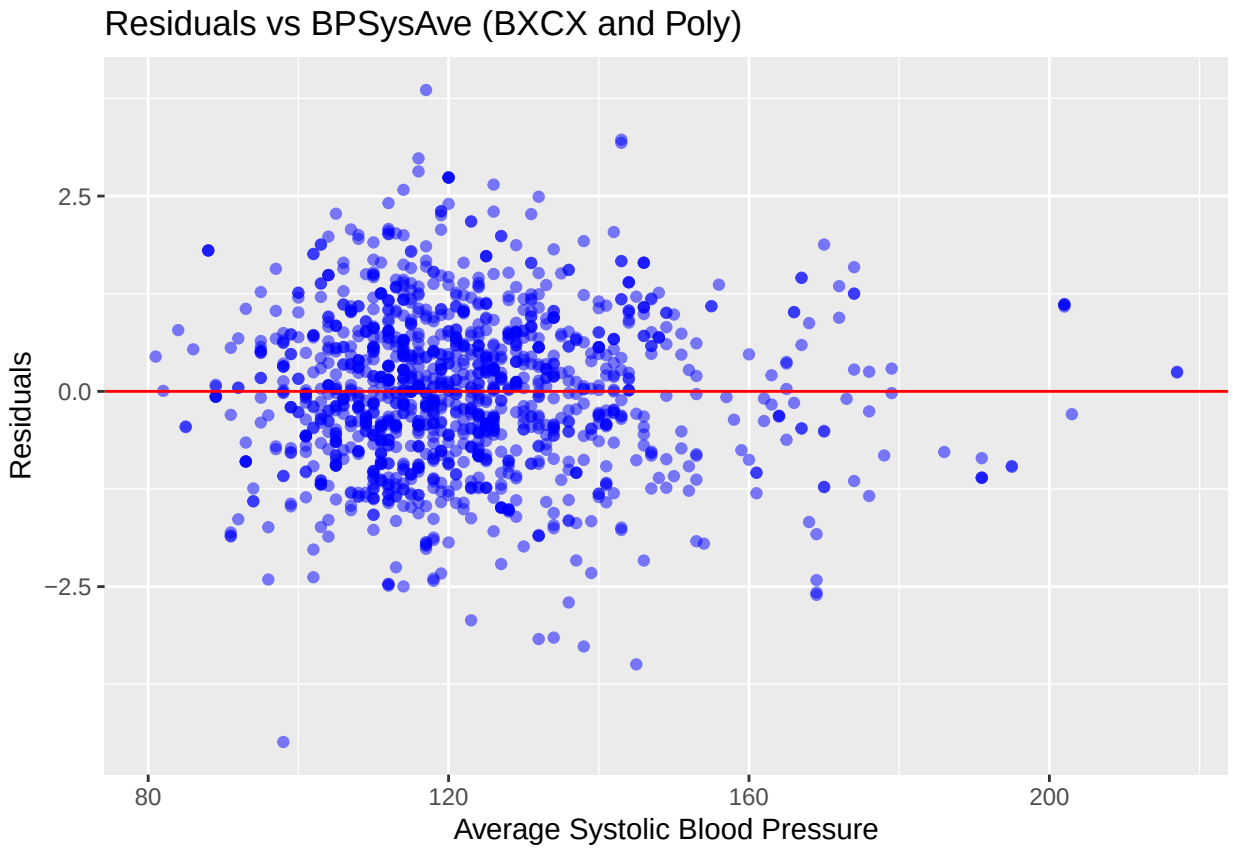

```
#RESIDUALS VS WEIGHT
res_weight_plot <- ggplot(p.BXCX.frame,
                          aes(x = Weight, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Weight (BXCX and Poly)",
       x = "Weight", y = "Residuals")

print(res_weight_plot)
```



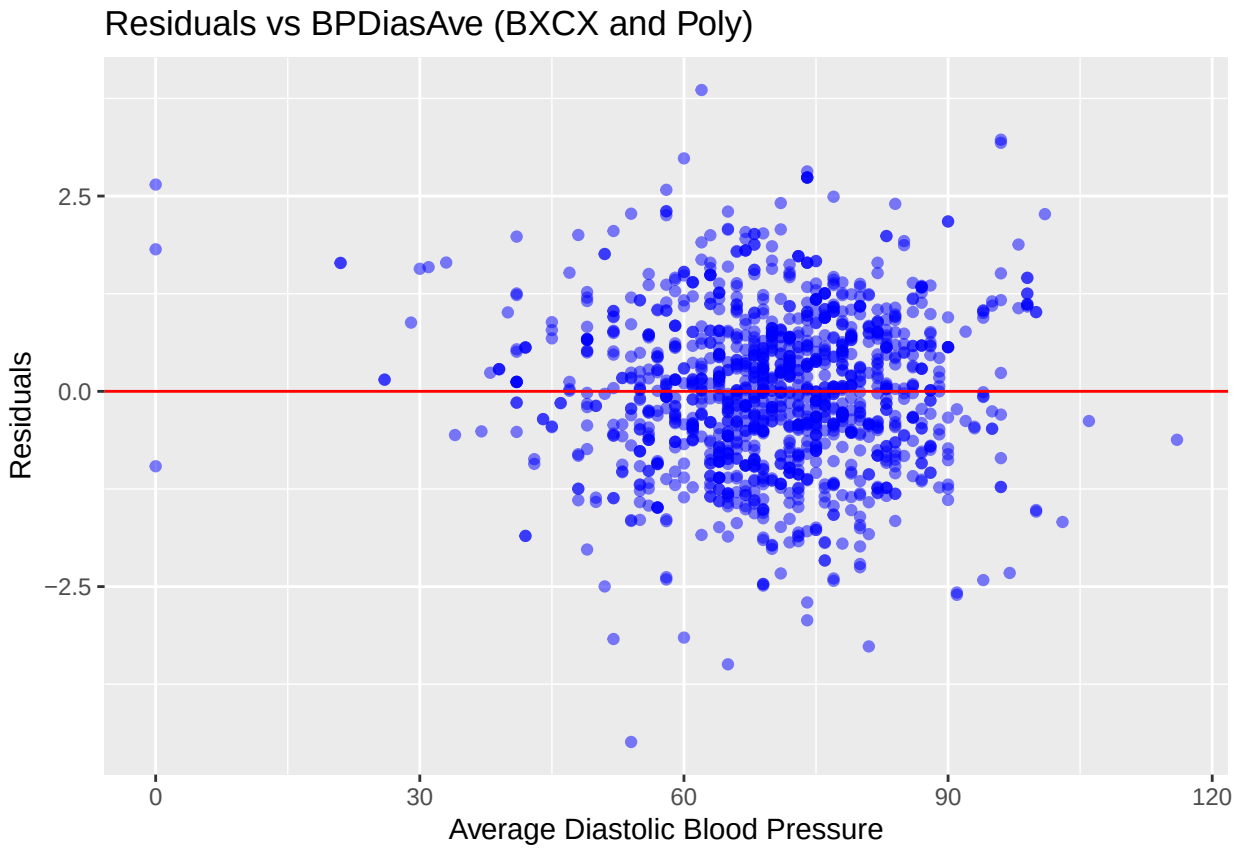

```
#RESIDUALS VS HEIGHT
res_height_plot <- ggplot(p.BXCX.frame,
                          aes(x = Height, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Height (BXCX and Poly)",
       x = "Height", y = "Residuals")

print(res_height_plot)
```



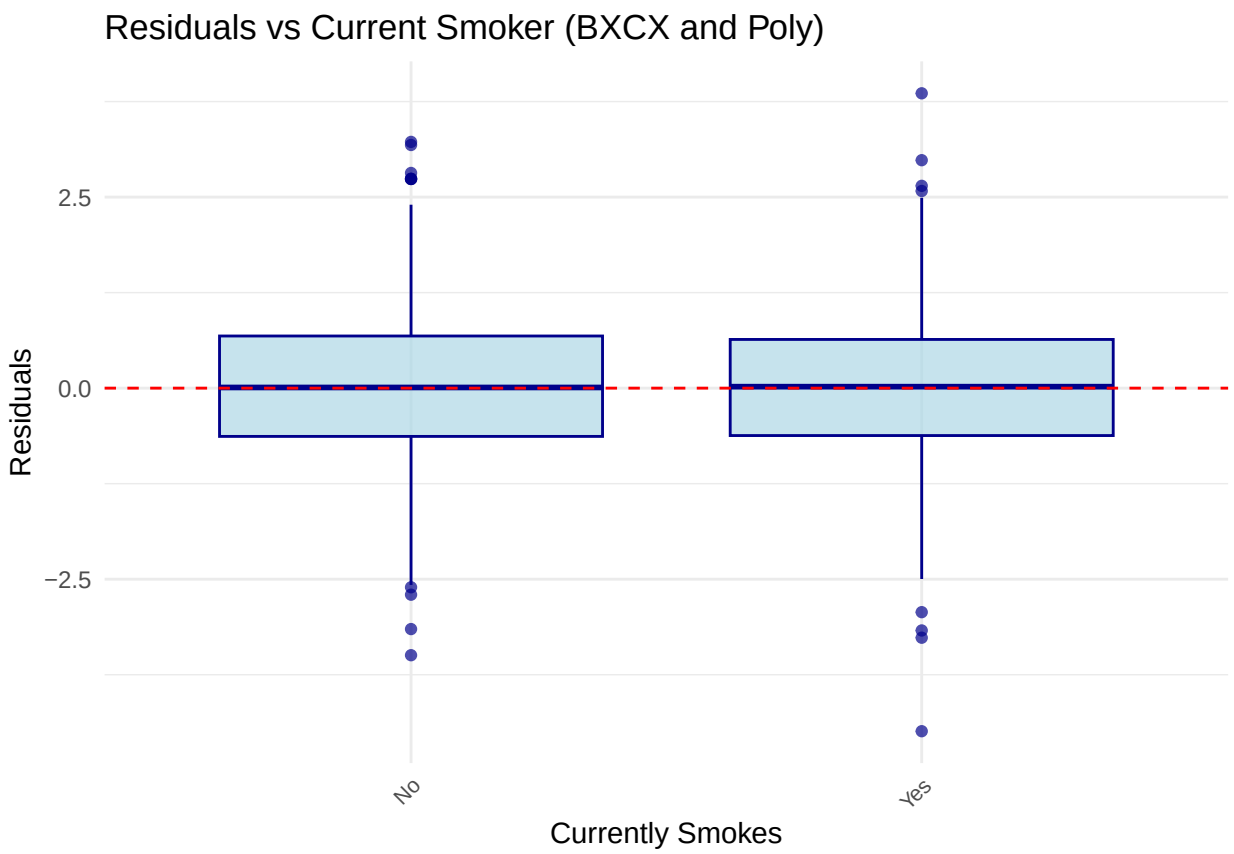

```
#RESIDUALS VS BPSysAve
res_BPSysAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPSysAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPSysAve (BXCX and Poly)",
       x = "Average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

print(res_BPSysAve_plot)
```



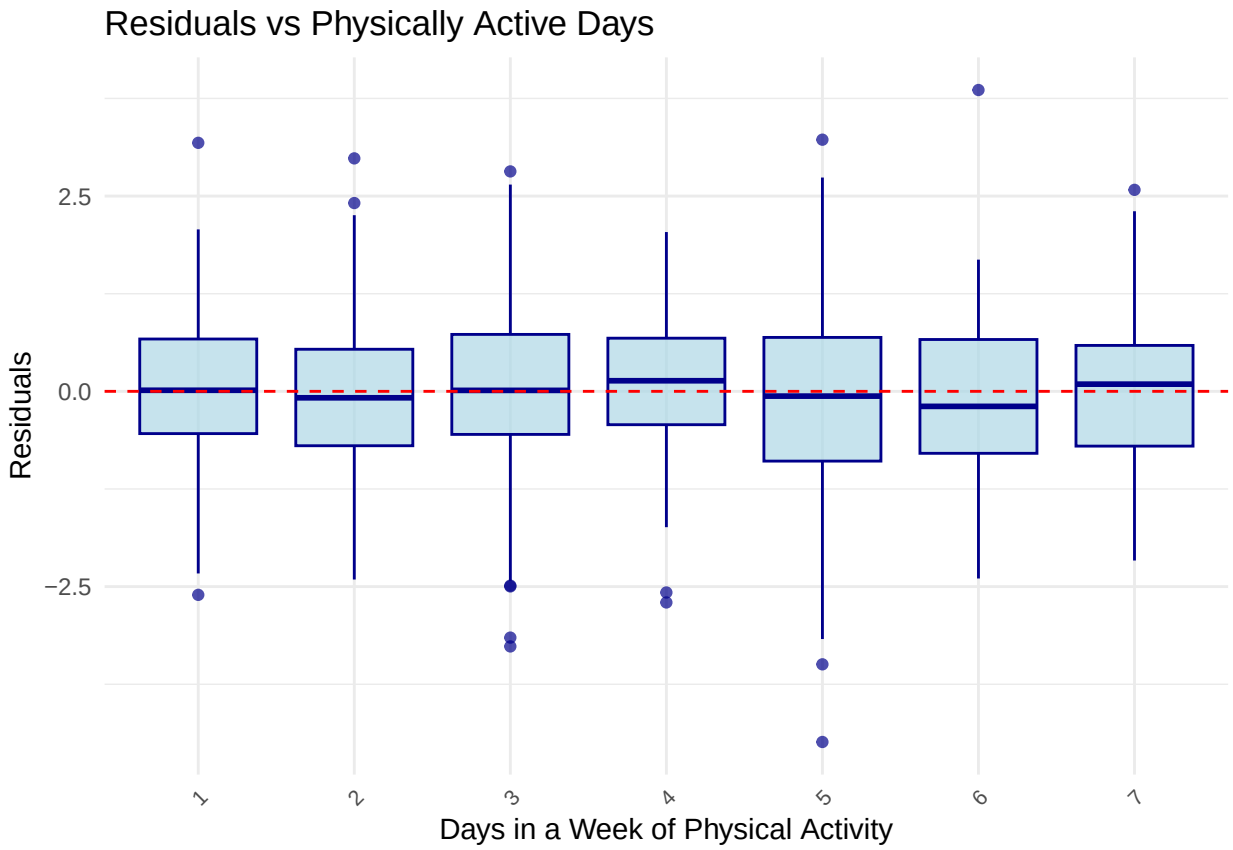

```
#RESIDUALS VS BPDiaAve
res_BPDiaAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPDiaAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPDiasAve (BXCX and Poly)",
       x = "Average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

print(res_BPDiaAve_plot)
```



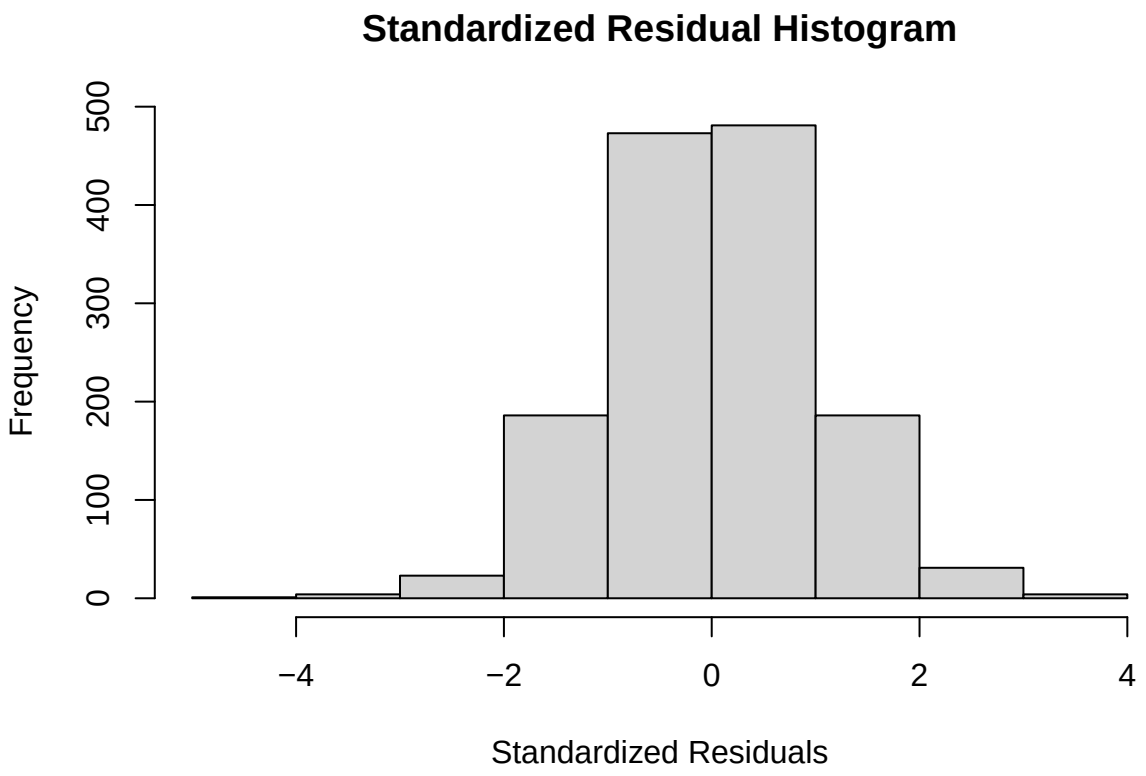

```r
#RESIDUALS VS SmokeNow (BOXPLOT)
res_smoke_plot <- ggplot(
  p.BXCX.frame, aes(x = as.factor(SmokeNow), y = pb.residuals)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  ggtitle("Residuals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



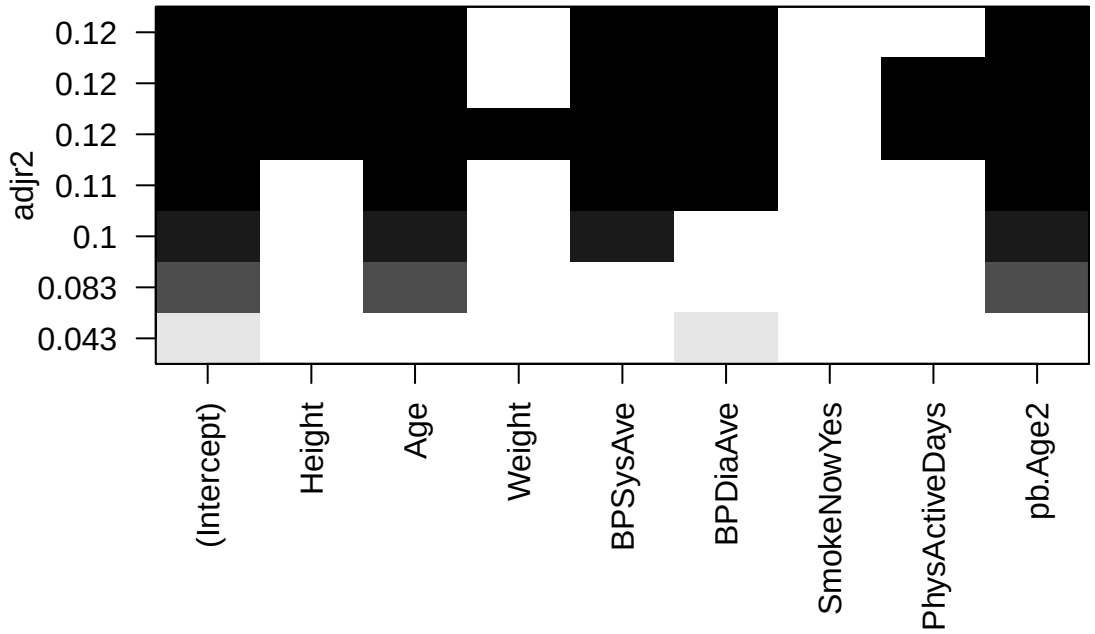

```
tr_stres_values <- rstandard(p.BXCX.model)

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



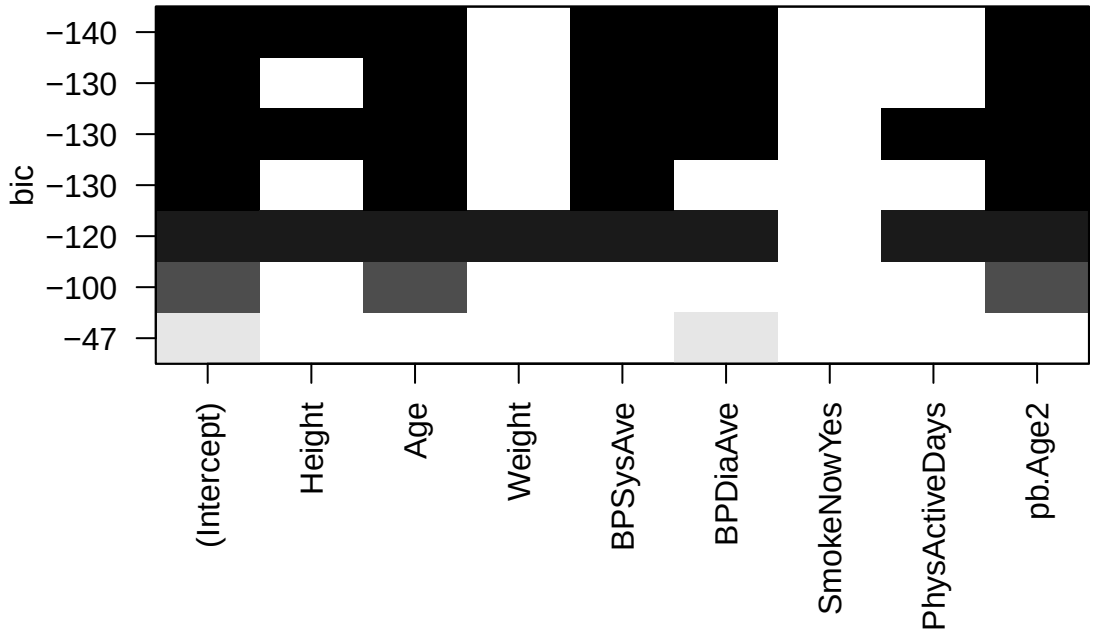

```
library(leaps)

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 ) "*"    "*" "*"    "*"      "*"      " "         "*"            "*"
```

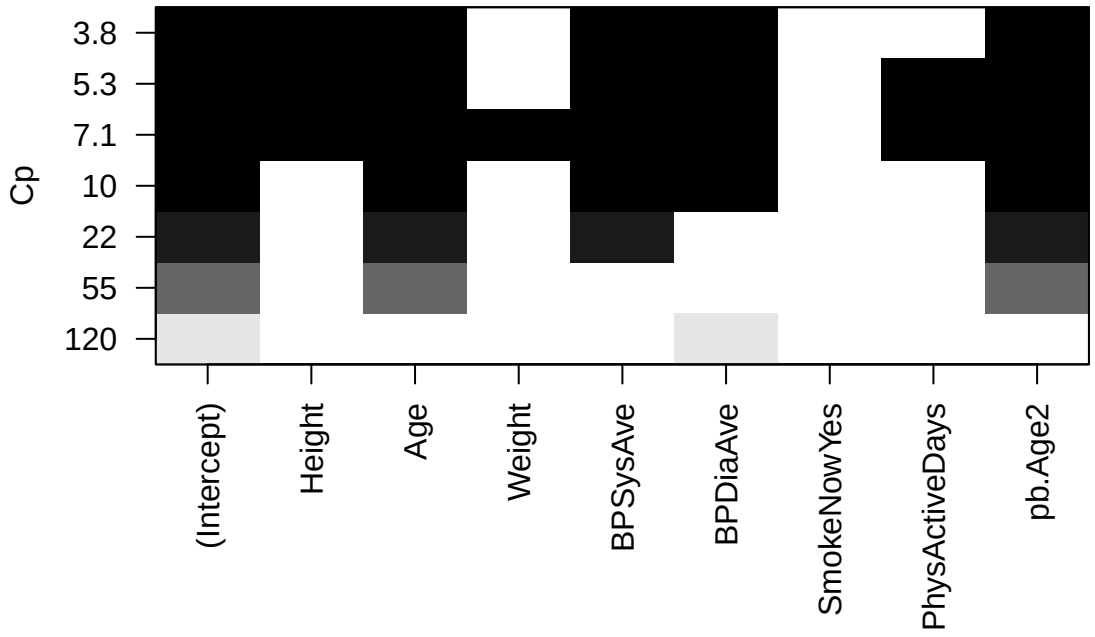
```
plot(best_subset_p.BXCX,scale='adjr2')
```




```
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=-26.48
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     SmokeNow + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - SmokeNow        1     0.114 1345.3 -28.365
## - Weight          1     0.222 1345.4 -28.253
## - PhysActiveDays  1     0.454 1345.7 -28.014
## <none>                        1345.2 -26.482
## - Height          1     5.835 1351.1 -22.471
## - BPDiaAve        1    14.610 1359.8 -13.478
## - BPSysAve        1    14.858 1360.1 -13.225
## - pb.Age2         1    80.843 1426.1  52.579
## - Age             1    90.187 1435.4  61.651
##
## Step:  AIC=-28.36
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - Weight          1     0.199 1345.5 -30.159
## - PhysActiveDays  1     0.452 1345.8 -29.898
## <none>                        1345.3 -28.365
## + SmokeNow        1     0.114 1345.2 -26.482
## - Height          1     5.833 1351.2 -24.355
## - BPDiaAve        1    14.747 1360.1 -15.222
## - BPSysAve        1    14.765 1360.1 -15.203
## - pb.Age2         1    80.740 1426.1  50.590
## - Age             1    90.533 1435.9  60.096
##
## Step:  AIC=-30.16
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + PhysActiveDays
```

```
##
##                   Df Sum of Sq    RSS     AIC
## - PhysActiveDays  1     0.432 1346.0 -31.713
## <none>                        1345.5 -30.159
## + Weight          1     0.199 1345.3 -28.365
## + SmokeNow        1     0.091 1345.4 -28.253
## - Height          1     8.591 1354.1 -23.319
## - BPDiaAve        1    14.551 1360.1 -17.219
## - BPSysAve        1    14.579 1360.1 -17.191
## - pb.Age2         1    80.612 1426.2  48.659
## - Age             1    90.449 1436.0  58.207
##
## Step:  AIC=-31.71
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve
##
##                   Df Sum of Sq    RSS     AIC
## <none>                        1346.0 -31.713
## + PhysActiveDays  1     0.432 1345.5 -30.159
## + Weight          1     0.179 1345.8 -29.898
## + SmokeNow        1     0.090 1345.9 -29.806
## - Height          1     8.514 1354.5 -24.954
## - BPSysAve        1    14.473 1360.4 -18.858
## - BPDiaAve        1    14.691 1360.7 -18.635
## - pb.Age2         1    80.243 1426.2  46.720
## - Age             1    90.027 1436.0  56.217
```

```
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
## -4.5043 -0.6322  0.0074  0.6457  3.8423
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.3844636  0.5574336   7.865 7.36e-15 ***
## Age          0.1016419  0.0105680   9.618  < 2e-16 ***
## pb.Age2     -0.0009697  0.0001068  -9.080  < 2e-16 ***
## Height      -0.0084564  0.0028590  -2.958 0.003151 **
## BPSysAve     0.0069601  0.0018049   3.856 0.000120 ***
## BPDiaAve     0.0094766  0.0024392   3.885 0.000107 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9865 on 1383 degrees of freedom
## Multiple R-squared:  0.121,  Adjusted R-squared:  0.1178
## F-statistic: 38.07 on 5 and 1383 DF,  p-value: < 2.2e-16
```

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

```r
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_inde

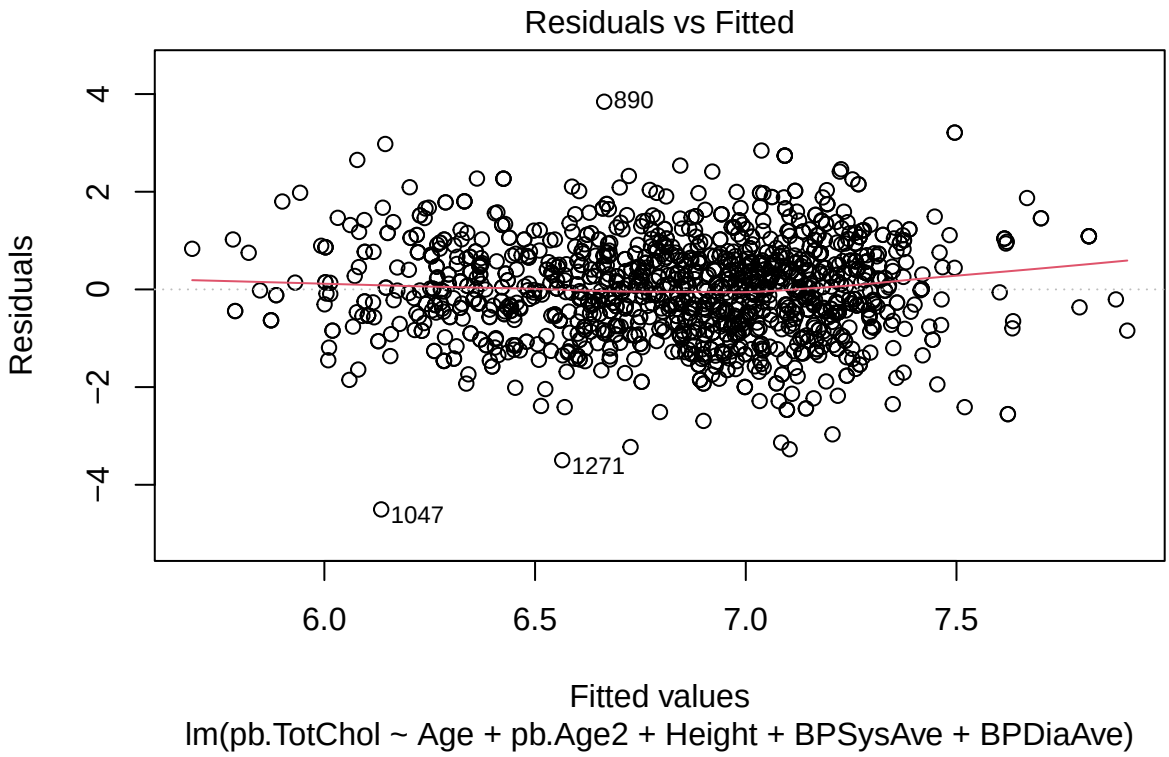
Check if the absolute values of the standardized residuals are greater than 3

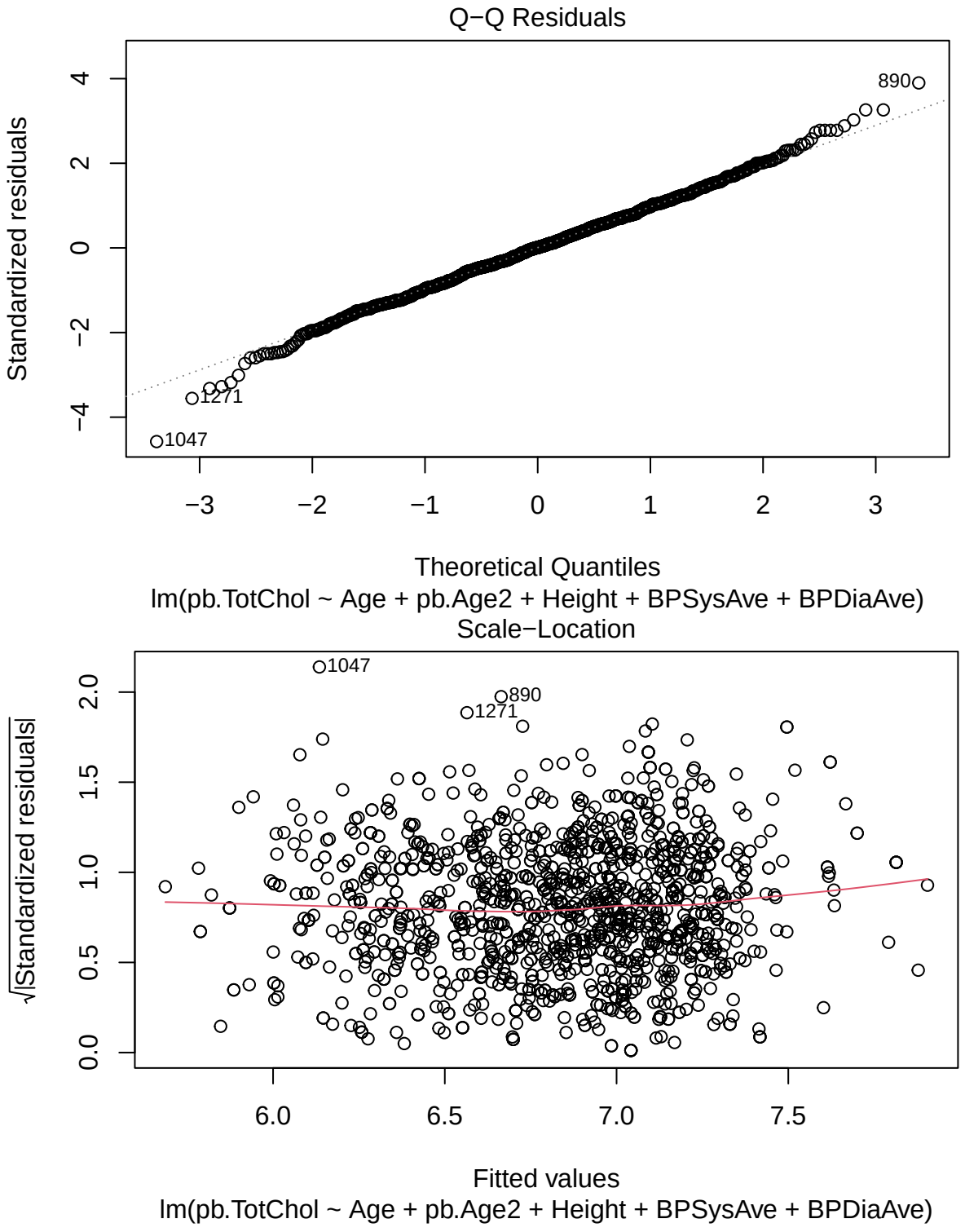
st.residuals <- rstandard(reduced.model) ## standardized residuals st.residuals

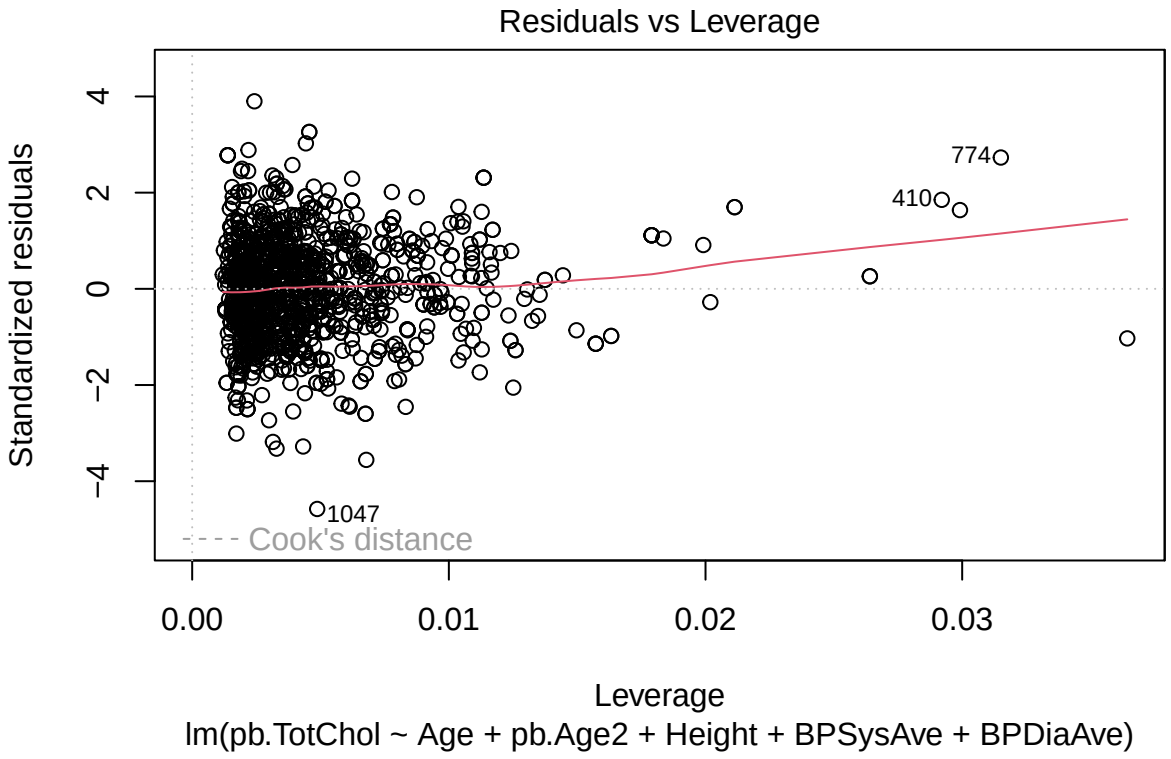
Outliers

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

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




Q−Q Residuals
Standardized residuals
890
1271
1047
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)
Scale−Location
√|Standardized residuals|
1047
890
1271
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)




```
#TABLE OF INFLUENTIAL OBSERVATIONS
influential_points <- c(410, 774, 1047)
p.BXCX.frame[influential_points, ]
```

```
##      Height Age Weight BPSysAve BPDiaAve SmokeNow PhysActiveDays pb.Age2
## 410   158.7  80   49.9      134        0       No              7    6400
## 774   182.9  70   72.3      126        0      Yes              3    4900
## 1047  163.0  25   57.1       98       54      Yes              5     625
##      pb.TotChol
## 410    7.699608
## 774    8.728813
## 1047   1.630585
```

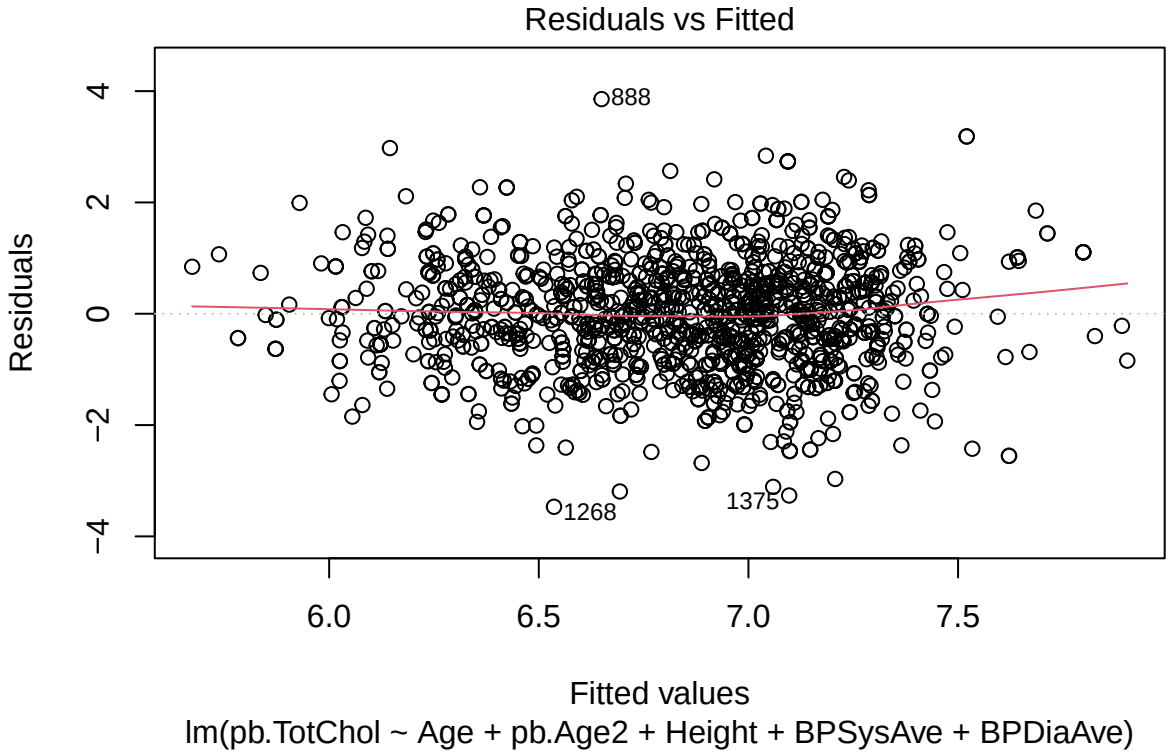
```
reduced.frame <- p.BXCX.frame %>%
  dplyr::filter(!row_number() %in% c(410, 774, 1047))

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

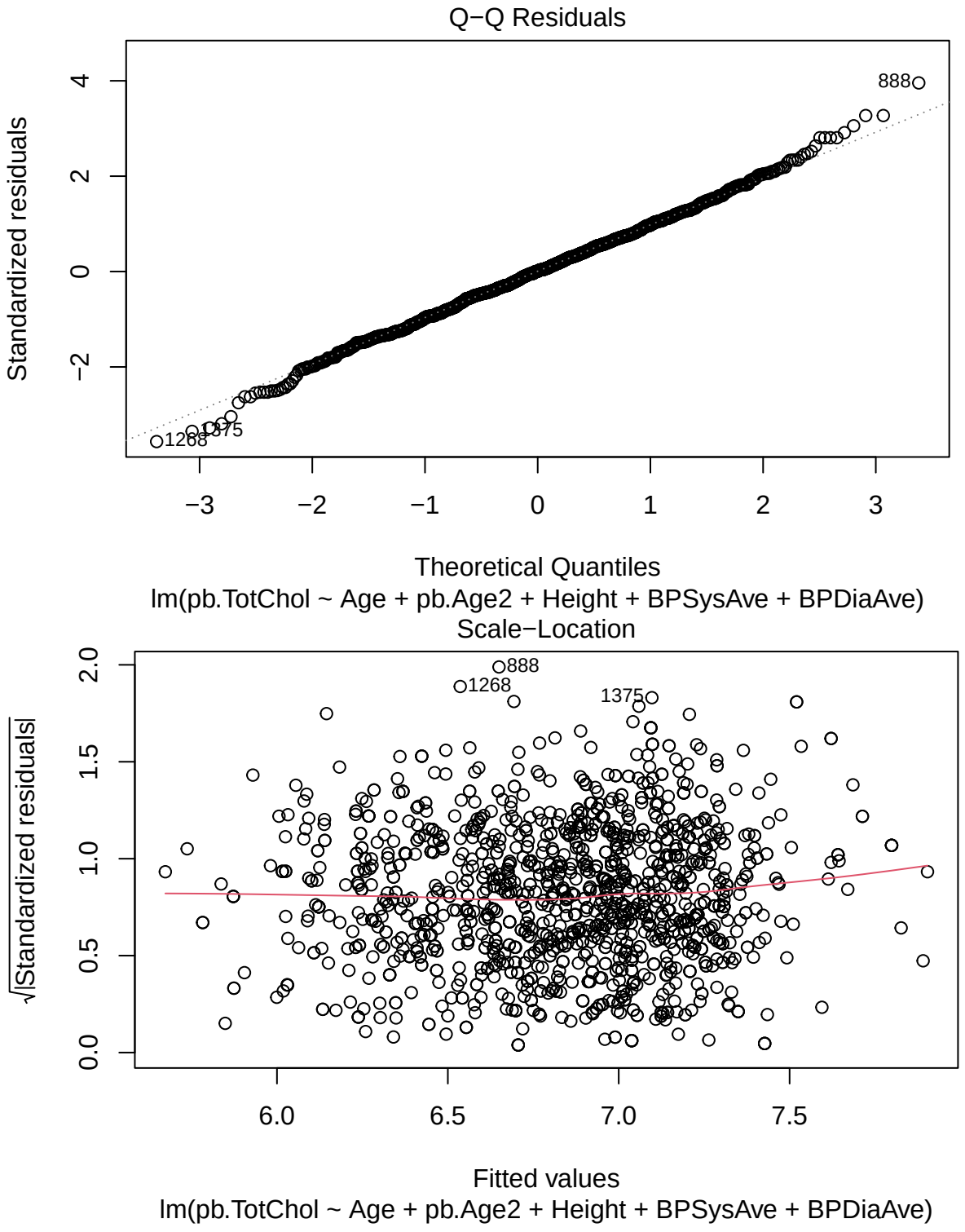
```
summary(final.model)
```

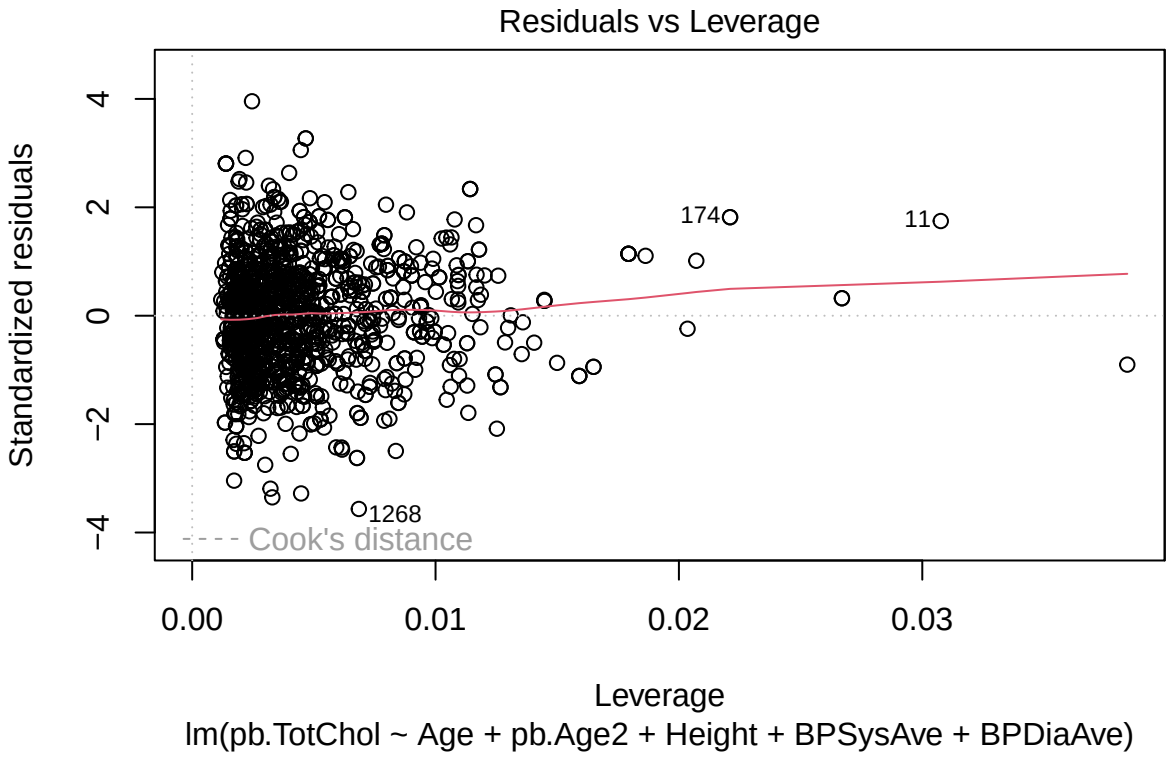
```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve +
##     BPDiaAve, data = reduced.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4681 -0.6309  0.0054  0.6454  3.8566
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.5210109  0.5529059   8.177 6.54e-16 ***
## Age          0.0989894  0.0104711   9.454  < 2e-16 ***
## pb.Age2     -0.0009468  0.0001058  -8.949  < 2e-16 ***
## Height      -0.0091079  0.0028338  -3.214  0.00134 **
## BPSysAve     0.0063692  0.0017914   3.555  0.00039 ***
## BPDiaAve     0.0111139  0.0024746   4.491 7.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9761 on 1380 degrees of freedom
## Multiple R-squared:  0.1233, Adjusted R-squared:  0.1201
## F-statistic: 38.82 on 5 and 1380 DF,  p-value: < 2.2e-16
```

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




Q−Q Residuals
Standardized residuals
888
1268
1375
−3
−2
−1
0
1
2
3
−2
0
2
4
Theoretical Quantiles
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)
Scale−Location
√|Standardized residuals|
888
1268
1375
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)


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
174
11
1268
Cook's distance
0.00
0.01
0.02
0.03
Leverage
lm(pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve)