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,
                                     Height > 0,
                                     Weight > 0,
                                     BPDia1 > 10,
                                     BPDia2 > 10,
                                     BPDia3 > 10,
                                     BPDiaAve > 10,
                                     BPSys1 > 10,
                                     BPSys2 > 10,
                                     BPSys3 > 10,
                                     BPSysAve > 10,
                                     TotChol > 0)

nrow(nhanes_filtered) #7094 observations
```

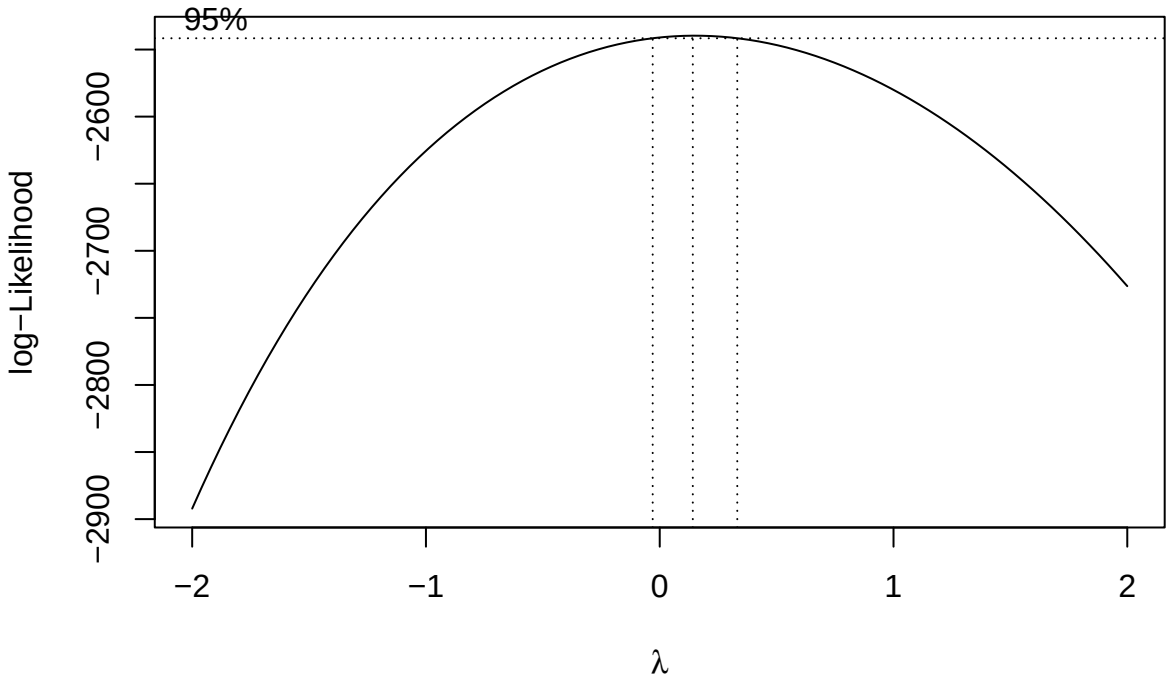
```
## [1] 5989
```

```
# 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)
```

```r
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

```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)

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

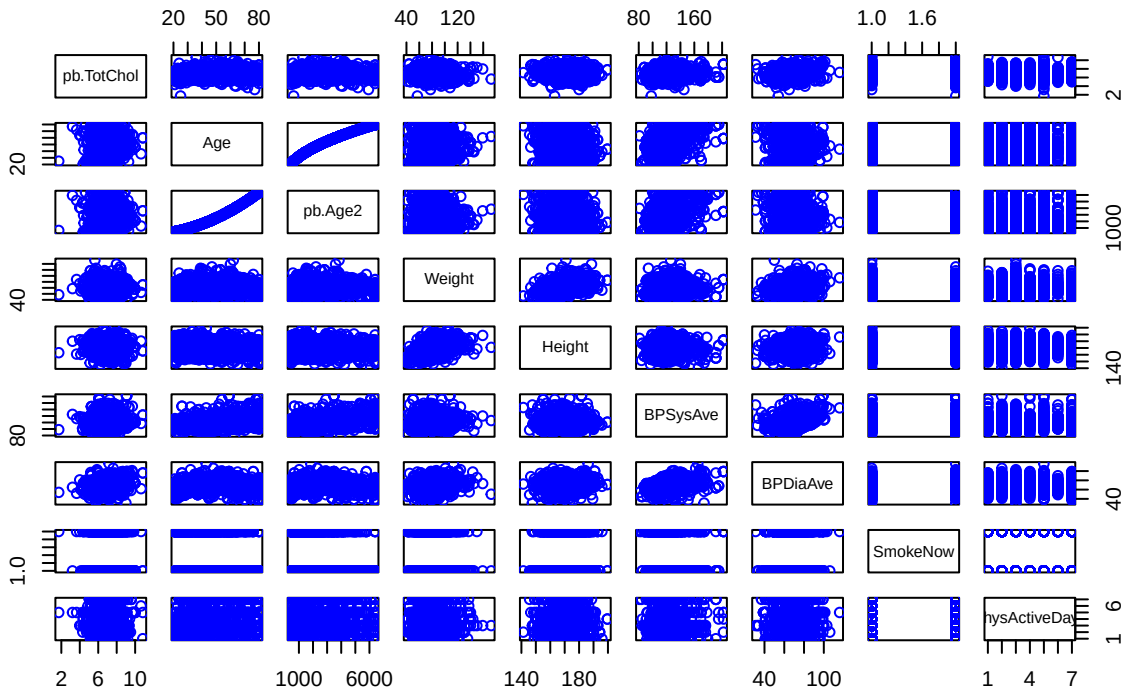



```r
pb.lambda <- pb.b$x[which.max(pb.b$y)]

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
```

```r
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)
```

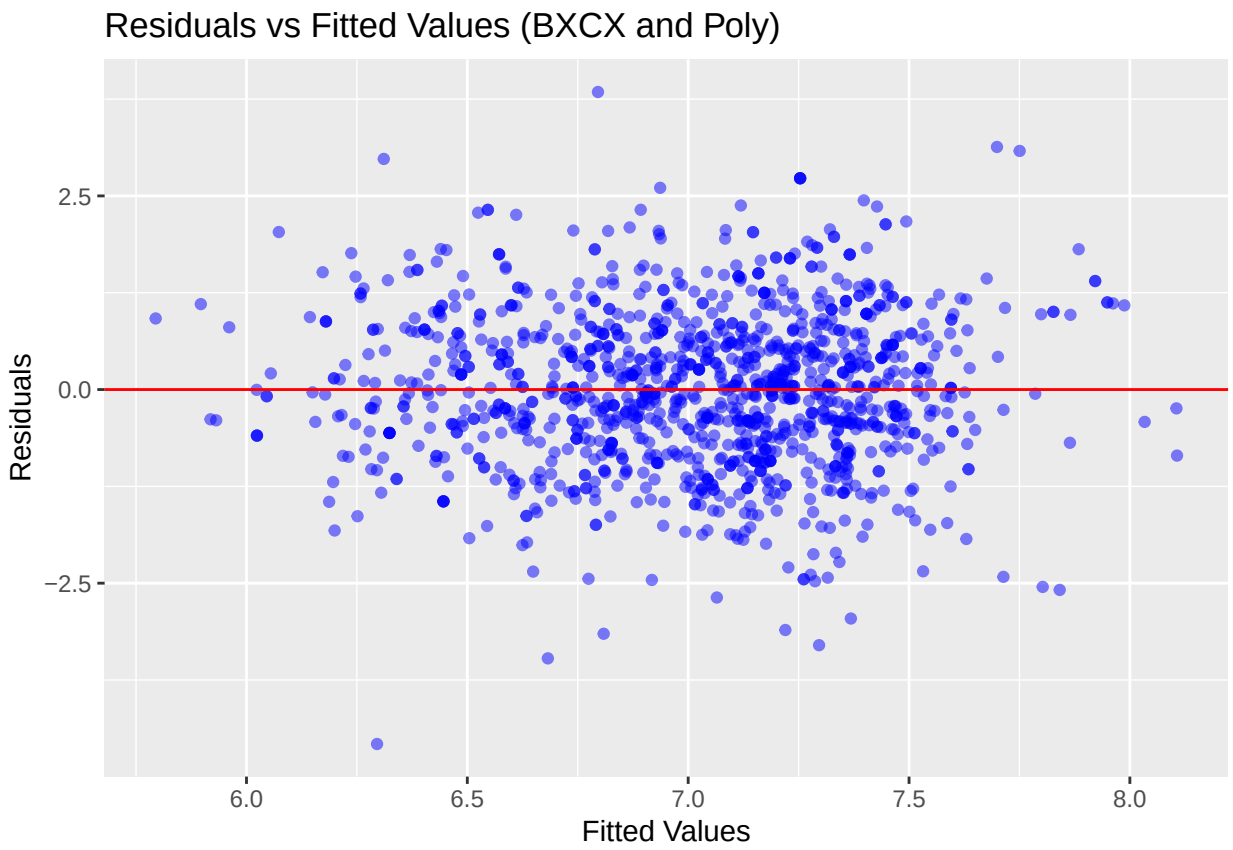
```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve +
##     BPDiaAve + SmokeNow + PhysActiveDays, data = p.BXCX.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5764 -0.6158 -0.0084  0.6574  3.8416
##
## Coefficients:
##                  Estimate 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")
```

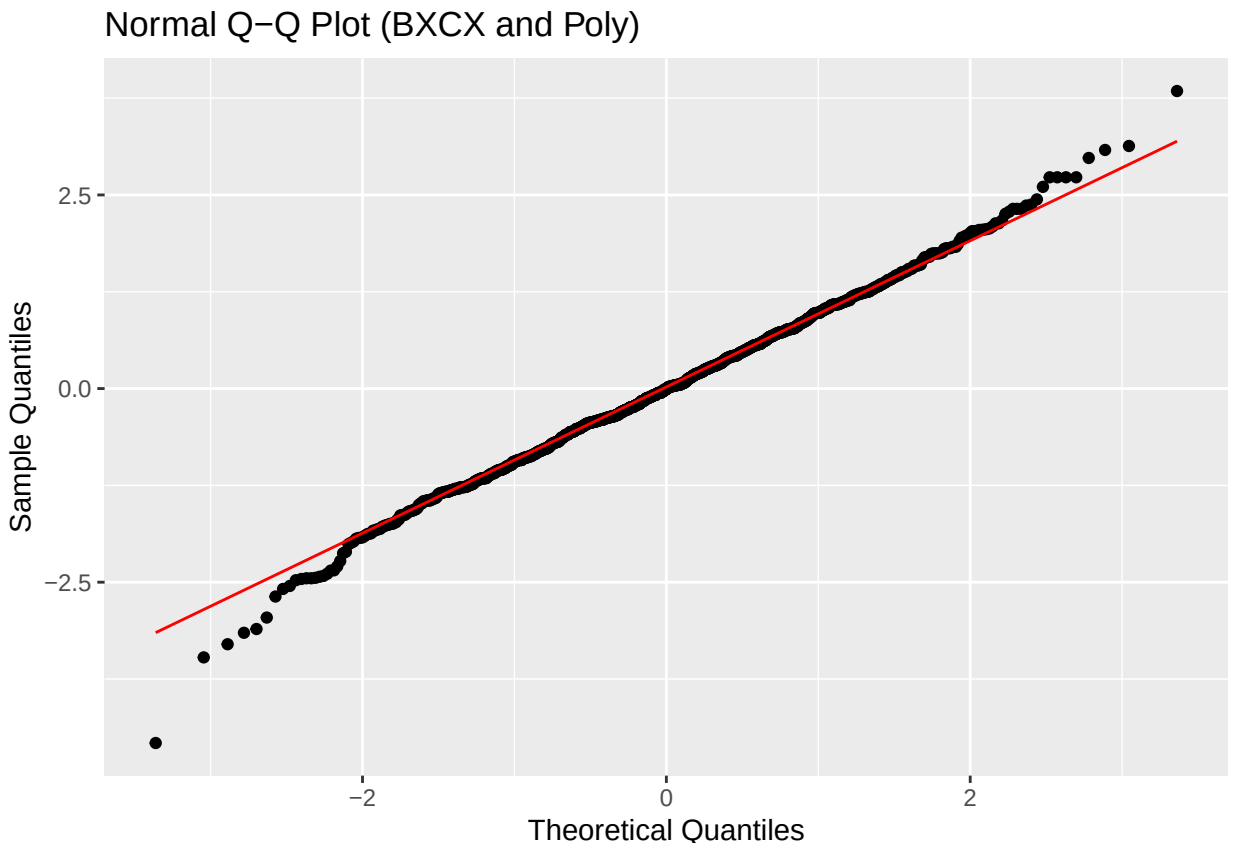


Residual Plots

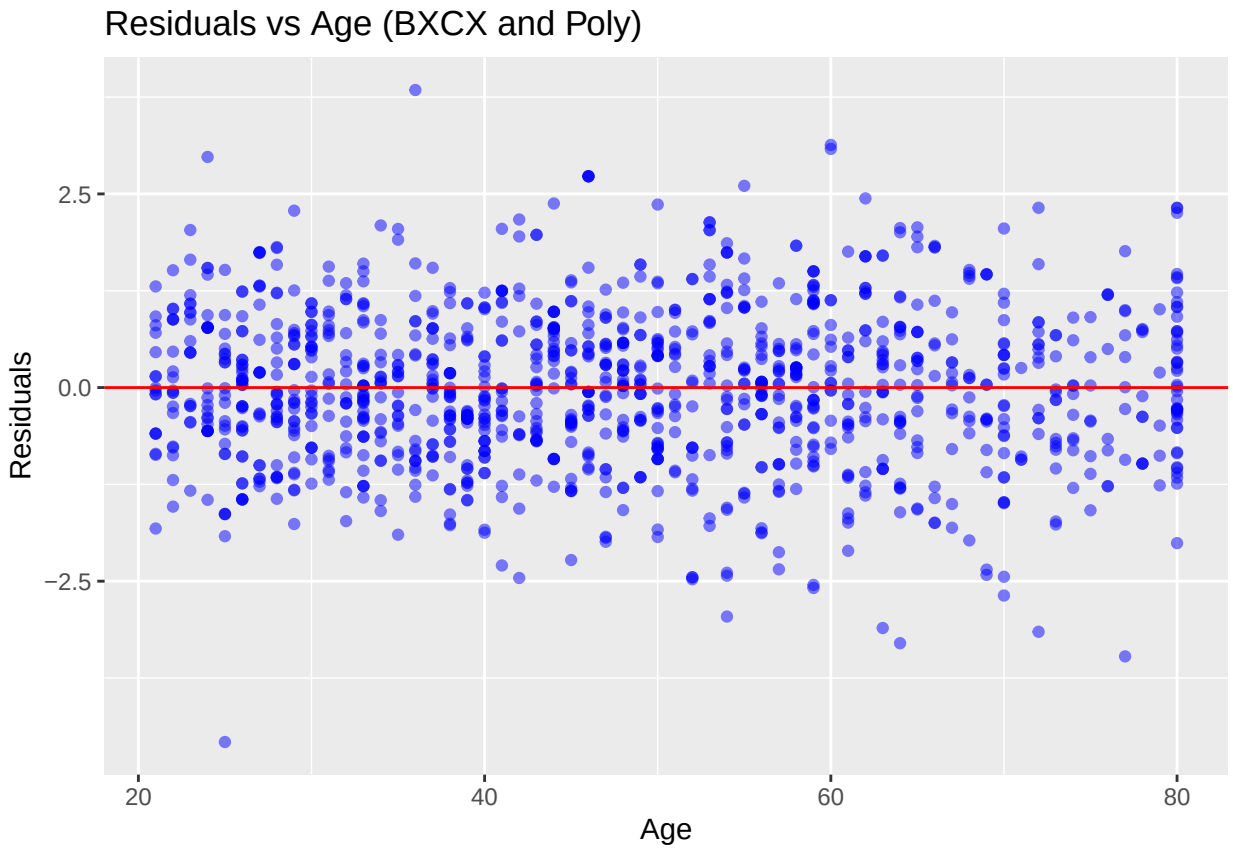
```
#RESIDUALS VS FITTED
res_fitted_plot <- ggplot(data = pb.plot_data,
                          aes(x = pb.fitted, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Fitted Values (BXCX and Poly)",
       x = "Fitted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

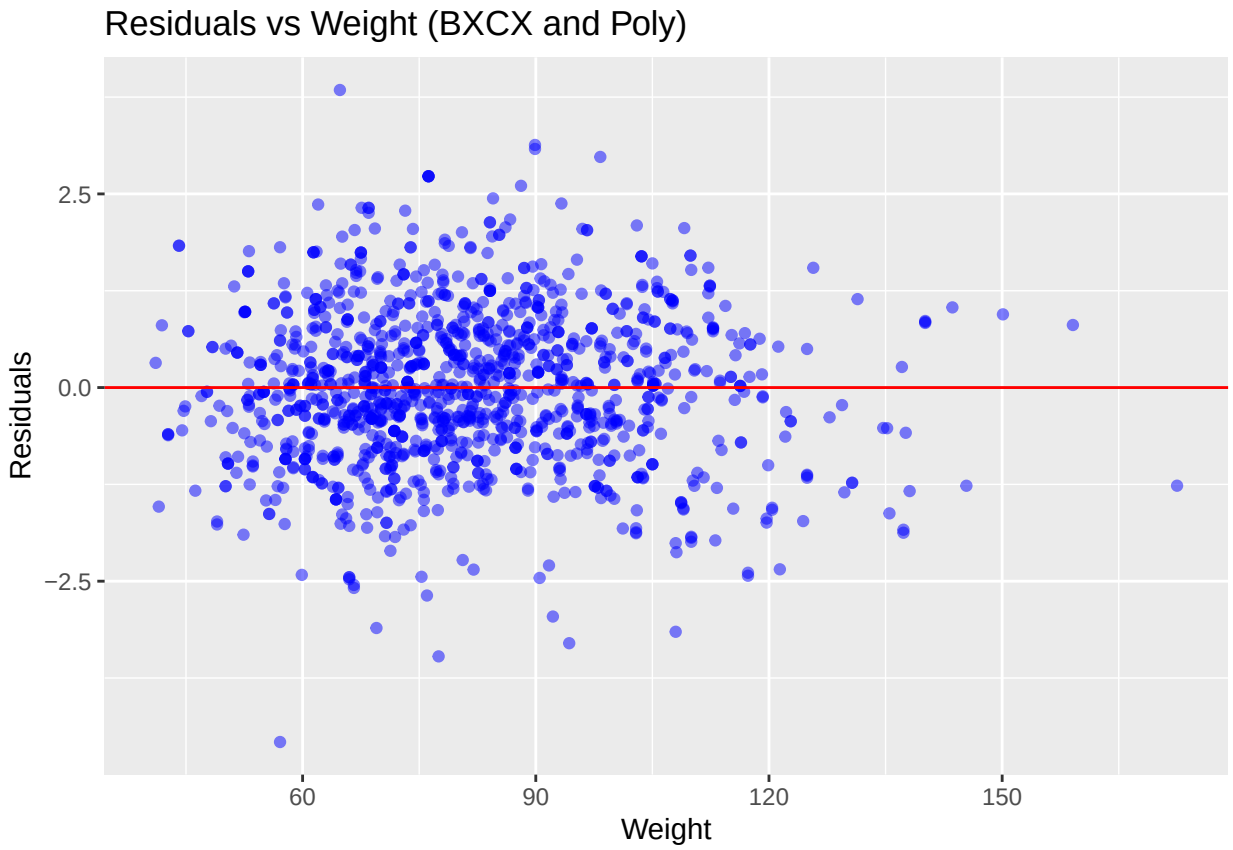
```r
#NORMAL QQ PLOT
qq_plot <- ggplot(data = data.frame(pb.residuals = pb.residuals),
                  aes(sample = pb.residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot (BXCX and Poly)",
       x = "Theoretical Quantiles", y = "Sample Quantiles")

print(qq_plot)
```



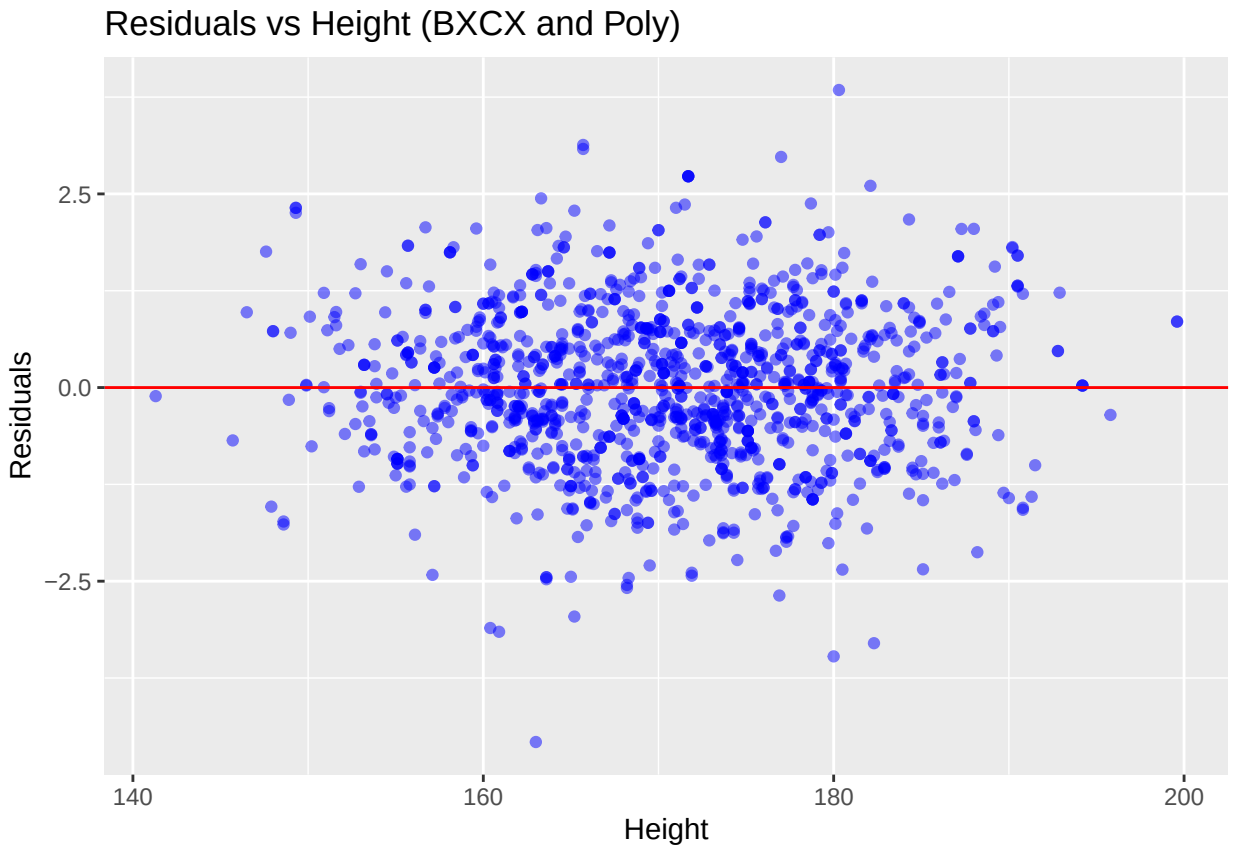

```r
#RESIDUALS VS AGE
res_age_plot <- ggplot(p.BXCX.frame,
                       aes(x = Age, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Age (BXCX and Poly)",
       x = "Age", y = "Residuals")

print(res_age_plot)
```



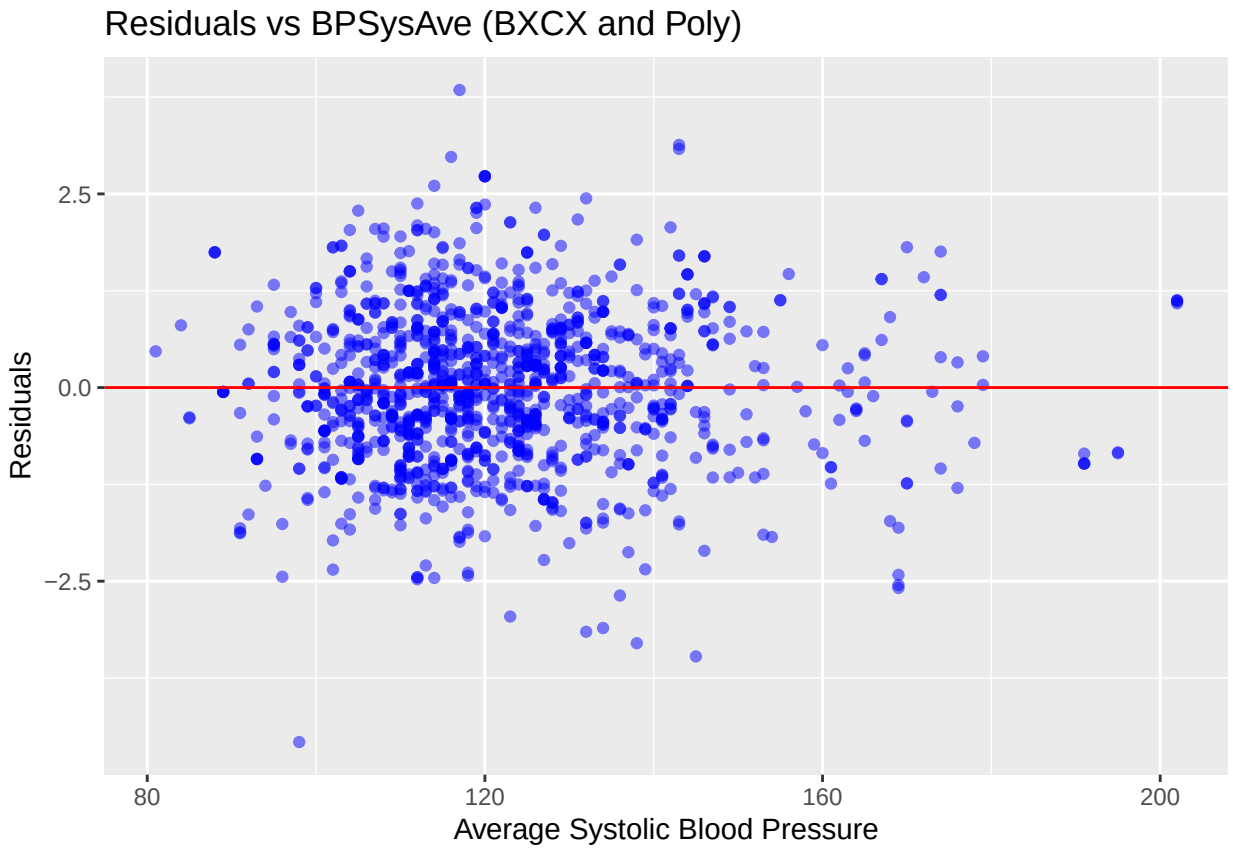

```
#RESIDUALS VS WEIGHT
res_weight_plot <- ggplot(p.BXCX.frame,
                          aes(x = Weight, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Weight (BXCX and Poly)",
       x = "Weight", y = "Residuals")

print(res_weight_plot)
```



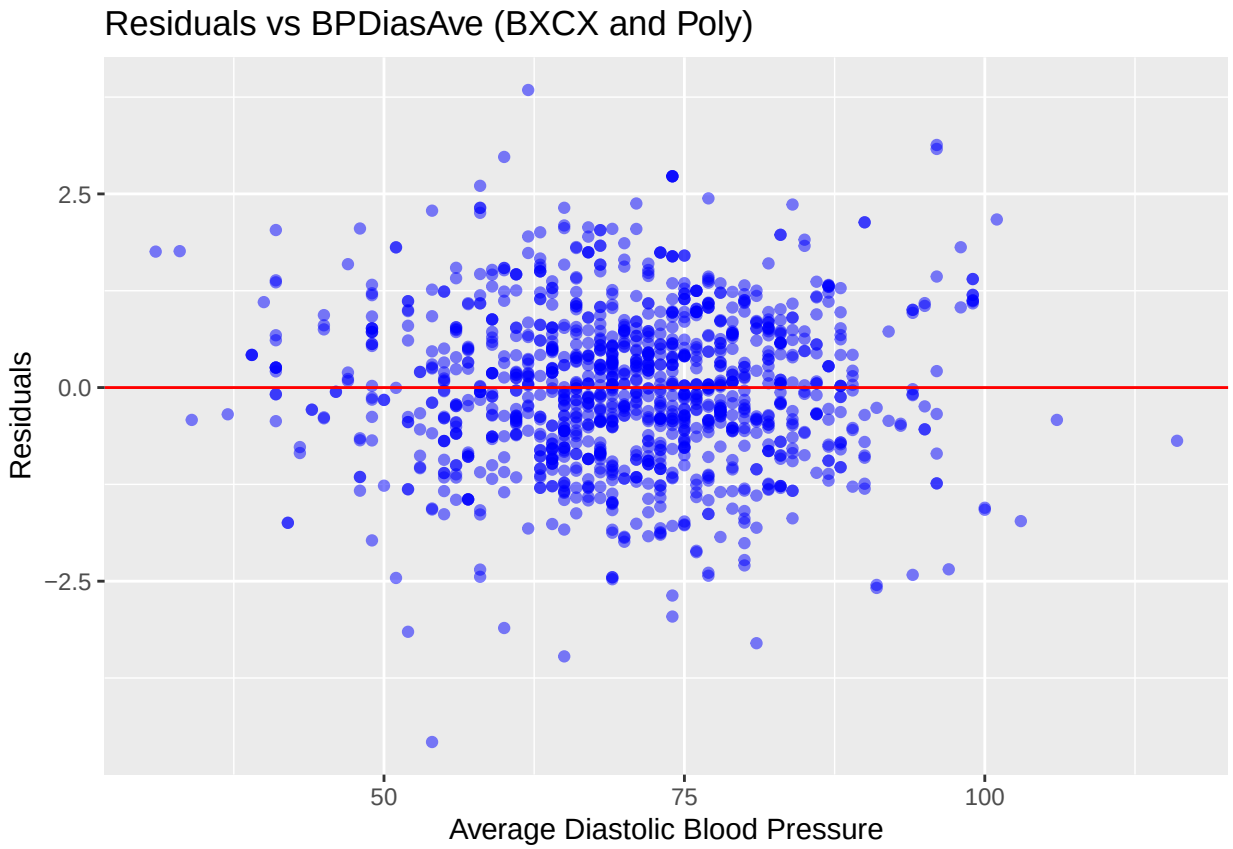

```
#RESIDUALS VS HEIGHT
res_height_plot <- ggplot(p.BXCX.frame,
                          aes(x = Height, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs Height (BXCX and Poly)",
       x = "Height", y = "Residuals")

print(res_height_plot)
```



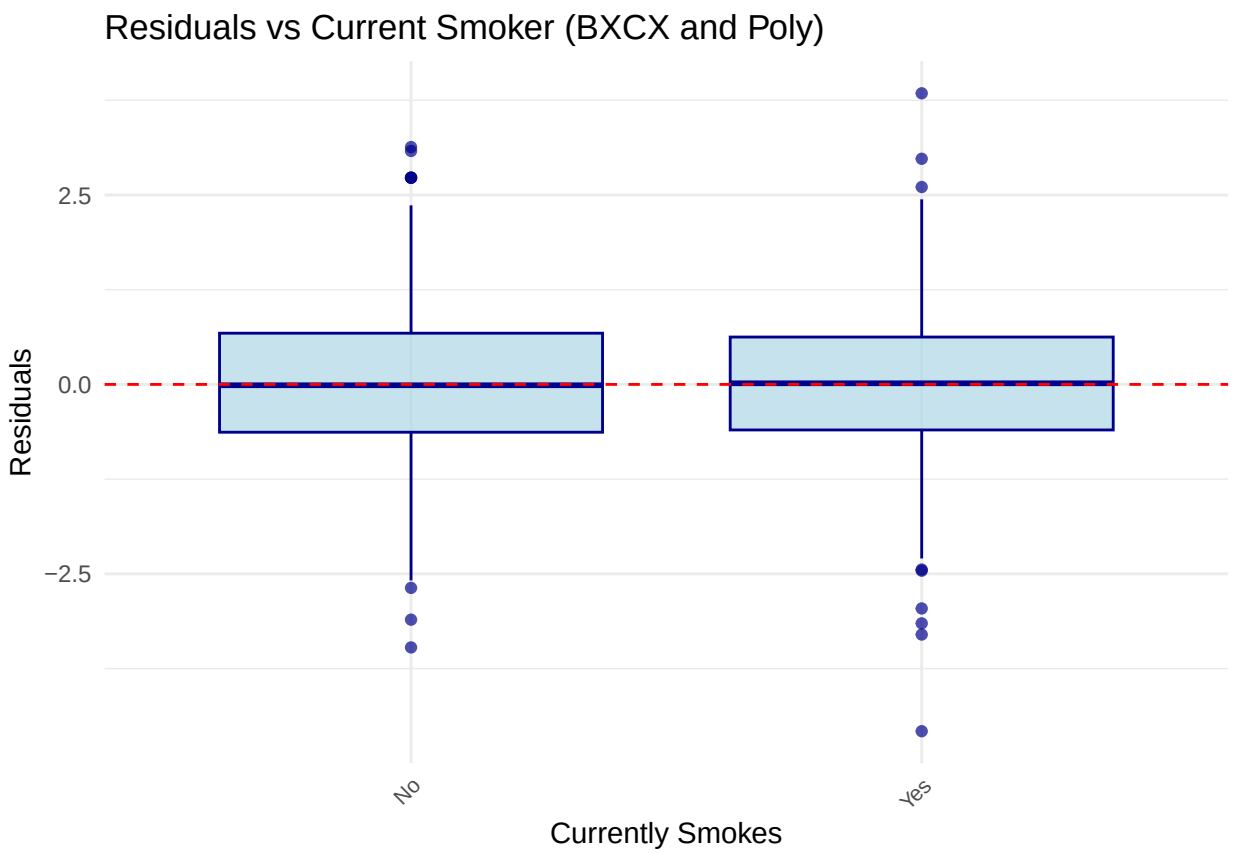

```r
#RESIDUALS VS BPSysAve
res_BPSysAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPSysAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPSysAve (BXCX and Poly)",
       x = "Average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

print(res_BPSysAve_plot)
```



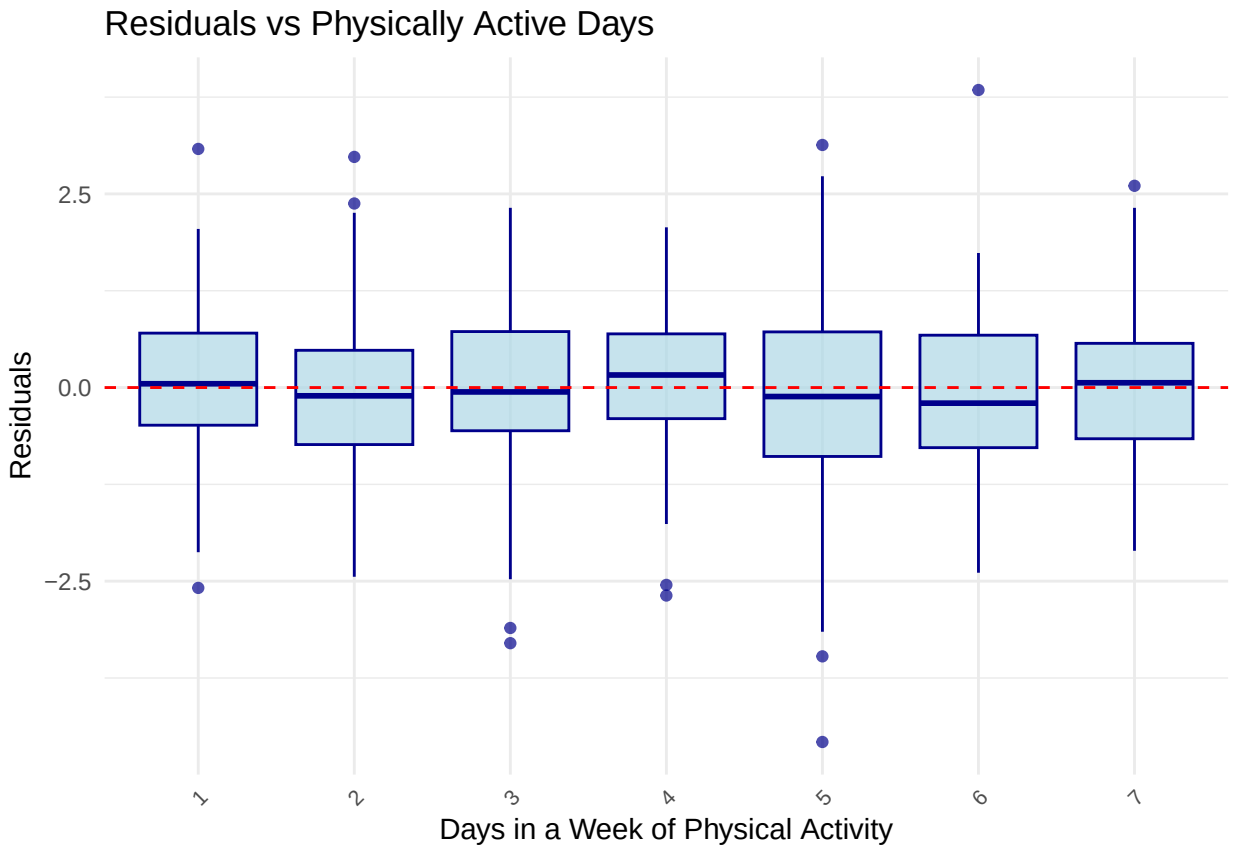

```
#RESIDUALS VS BPDiaAve
res_BPDiaAve_plot <- ggplot(p.BXCX.frame,
                            aes(x = BPDiaAve, y = pb.residuals)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residuals vs BPDiasAve (BXCX and Poly)",
       x = "Average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

print(res_BPDiaAve_plot)
```



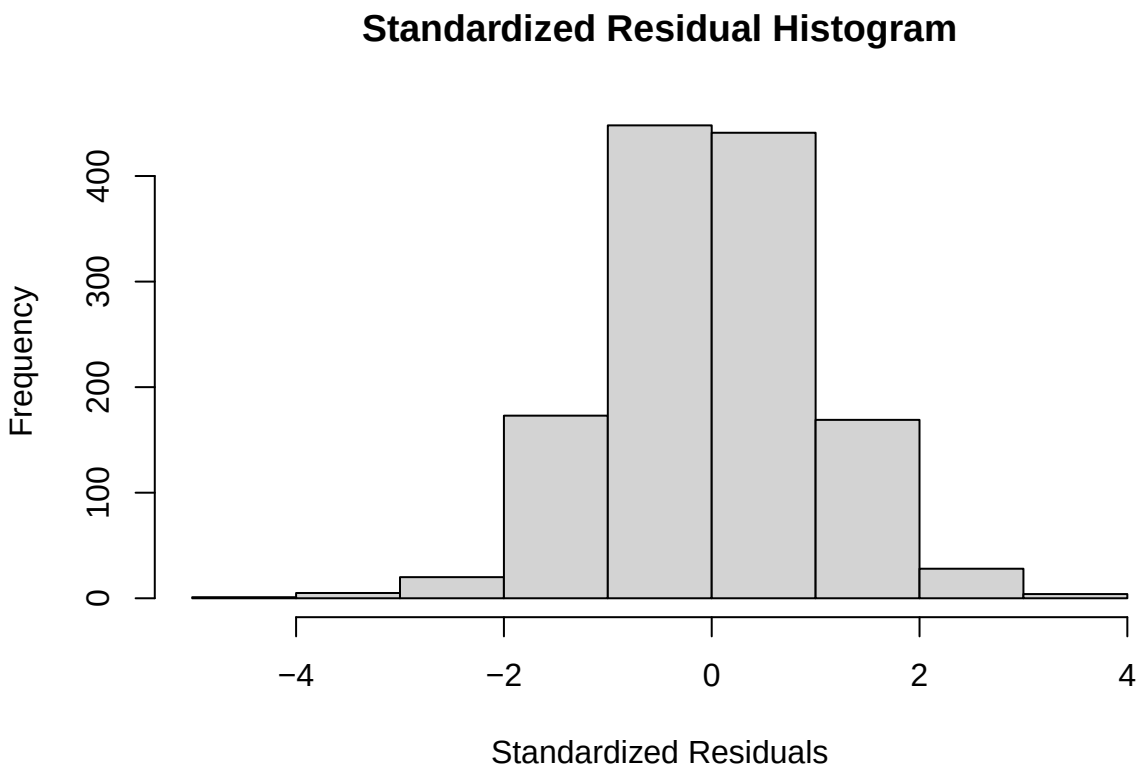

```r
#RESIDUALS VS SmokeNow (BOXPLOT)
res_smoke_plot <- ggplot(
  p.BXCX.frame, aes(x = as.factor(SmokeNow), y = pb.residuals)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  theme_minimal() +
  ggtitle("Residuals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




```
leverage <- hatvalues(p.BXCX.model)


##LEVERAGE POINTS
p <- 8
high_lev <- 2*(p+1)/n

leverage_points <- p.BXCX.frame[leverage > high_lev,]
leverage_points <- leverage_points %>%
  mutate(row = row.names(leverage_points))


#FINDING OUTLIERS
st.residuals <- rstandard(p.BXCX.model)

outlier_points <- p.BXCX.frame[abs(st.residuals) > 4,]

#COOKS DISTANCE
cooks_value <- cooks.distance(p.BXCX.model)

f_value <- qf(0.50, 8, 1280)

cooks_points <- p.BXCX.frame[cooks_value > f_value,]

#DFFITS
dffits_cutoff <- 2*(sqrt((p+1)/n))

dffits_value = dffits(p.BXCX.model)

dffits_points <- p.BXCX.frame[(abs(dffits_value) > dffits_cutoff),]
```

```r
dffits_points <- dffits_points %>%
  mutate(row = row.names(dffits_points))

#DFBETAS
dfbetas_cutoff <- 2/sqrt(n)

dfbeta_frame <- as.data.frame(dfbetas(p.BXCX.model))

dfbeta_points <- round(dfbeta_frame[apply(
  abs(dfbeta_frame)>dfbetas_cutoff,1,any),],4)
dfbeta_points <- dfbeta_points %>%
  mutate(row = row.names(dfbeta_points))

#Problematic observations
influential_points <- c(728,823)
p.BXCX.frame[influential_points, ]
```

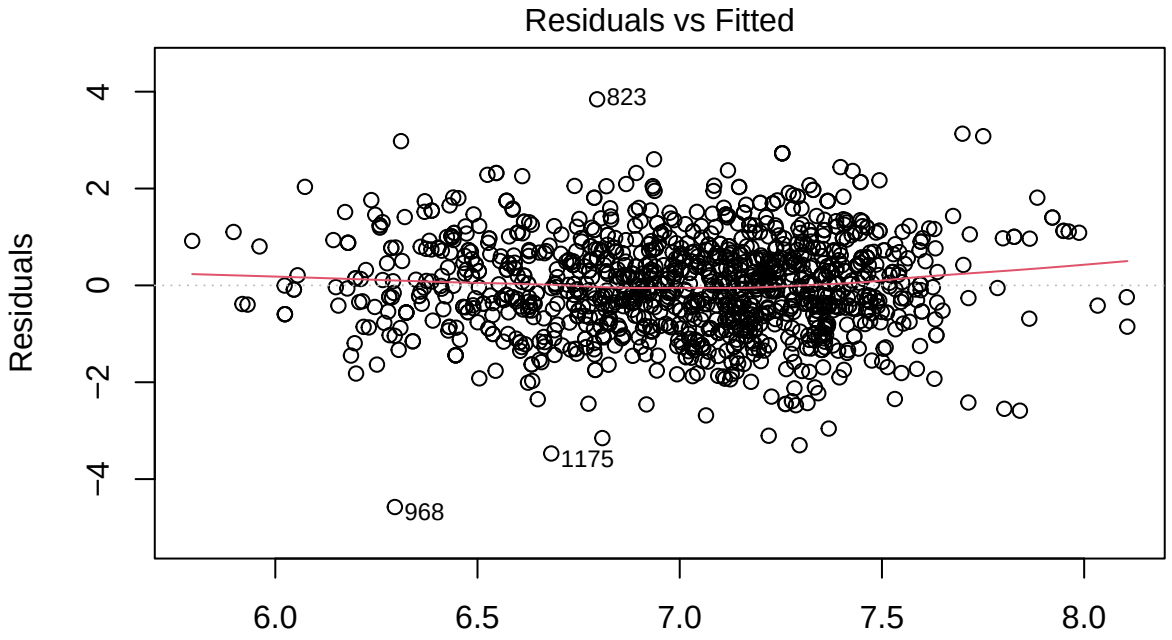
```
##     Height Age Weight BPSysAve BPDiaAve SmokeNow PhysActiveDays pb.Age2
## 728  160.9  72  108.0      132       52      Yes              5    5184
## 823  180.3  36   64.8      117       62      Yes              6    1296
##     pb.TotChol
## 728    3.65555
## 823   10.63743
```

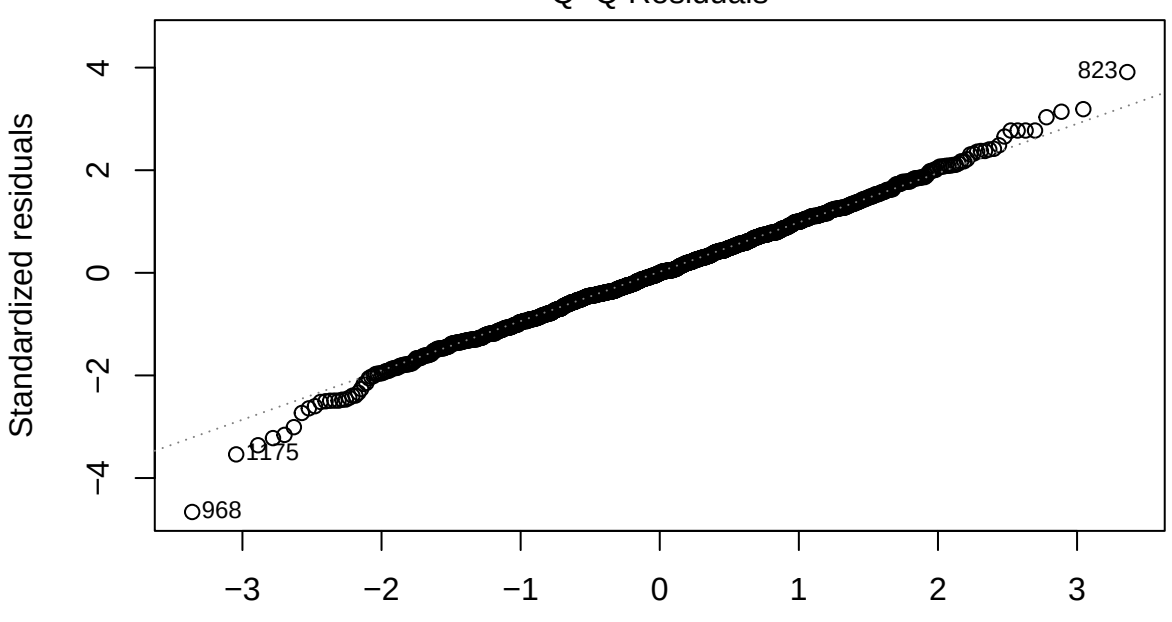
```r
clean.frame <- p.BXCX.frame %>%
dplyr::filter(!row_number() %in% influential_points)

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

summary(clean_model)
```

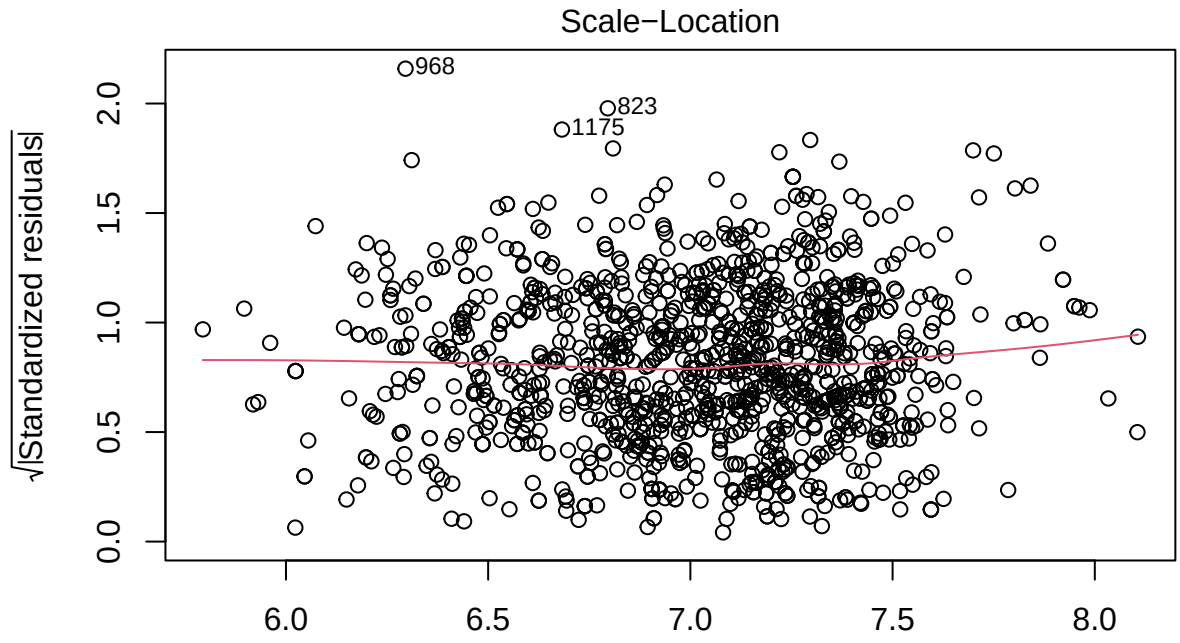
```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve +
##     BPDiaAve + SmokeNow + PhysActiveDays, data = clean.frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5691 -0.6185  0.0030  0.6555  3.1272
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.828e+00  6.056e-01   7.973 3.40e-15 ***
## Age             9.839e-02  1.111e-02   8.855  < 2e-16 ***
## pb.Age2        -9.317e-04  1.124e-04  -8.287 2.91e-16 ***
## Weight         -3.742e-05  1.675e-03  -0.022  0.98218
## Height         -9.841e-03  3.326e-03  -2.959  0.00315 **
## BPSysAve        5.646e-03  1.962e-03   2.878  0.00407 **
## BPDiaAve        1.275e-02  2.818e-03   4.523 6.67e-06 ***
## SmokeNowYes     1.781e-02  5.923e-02   0.301  0.76375
## PhysActiveDays -1.413e-02  1.531e-02  -0.923  0.35604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

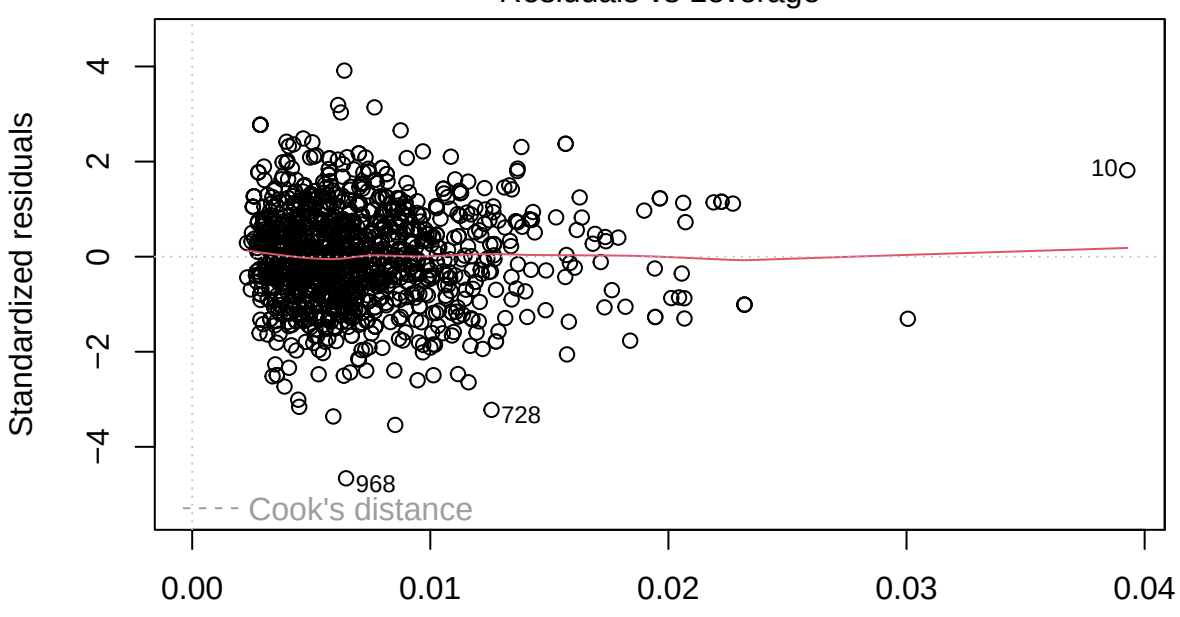
```
##
## Residual standard error: 0.9757 on 1278 degrees of freedom
## Multiple R-squared:  0.129,  Adjusted R-squared:  0.1236
## F-statistic: 23.66 on 8 and 1278 DF,  p-value: < 2.2e-16
```

```
plots <- plot(p.BXCX.model)
```





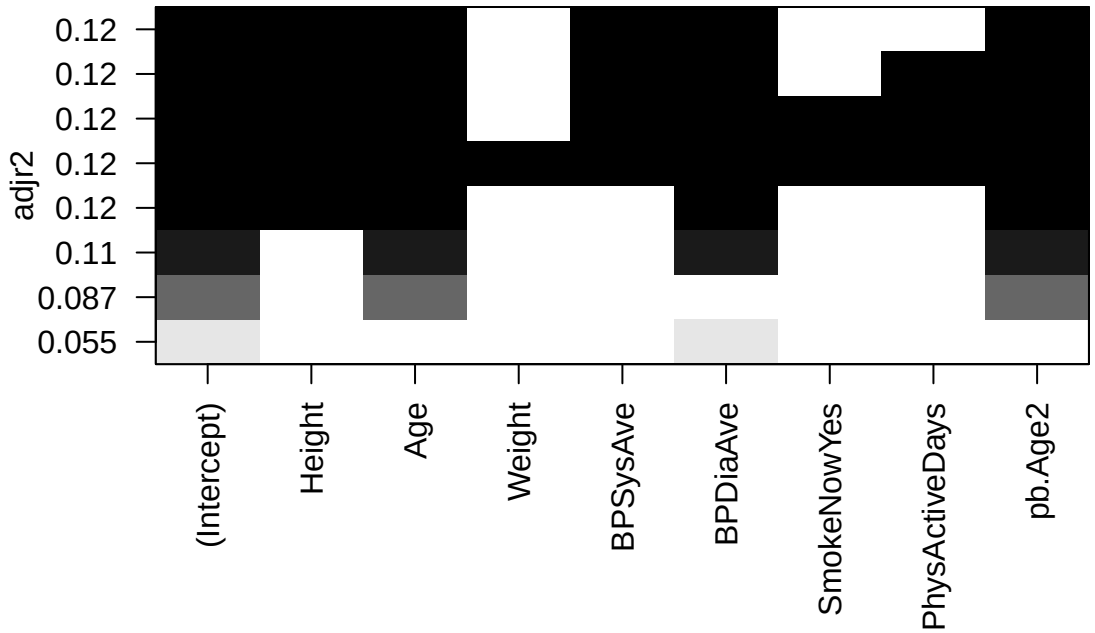




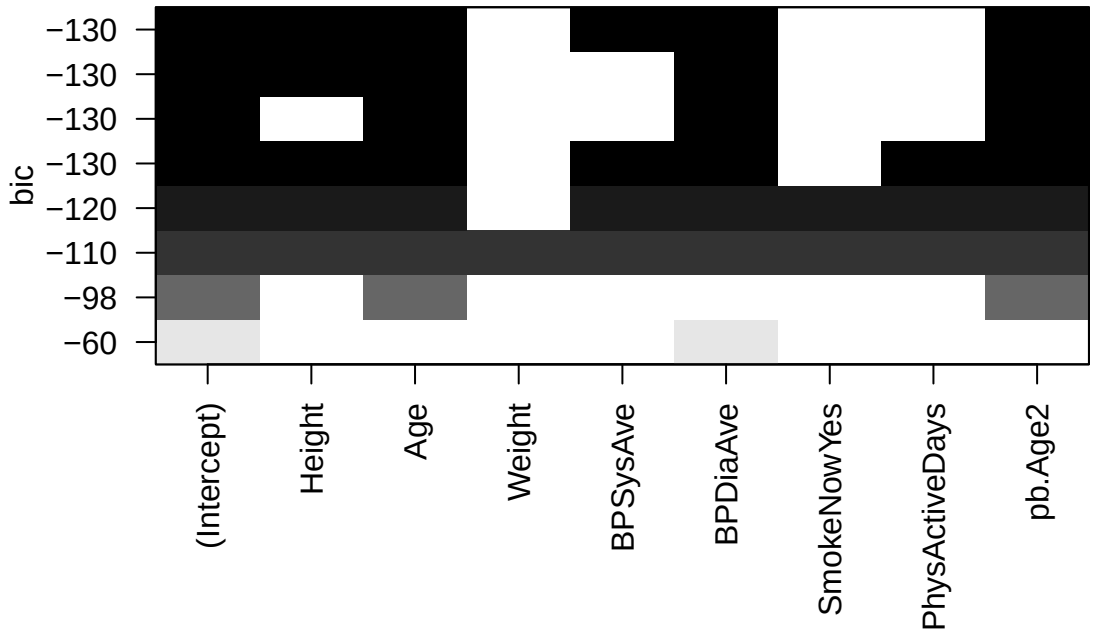

```
library(leaps)

best_subset <- regsubsets(pb.TotChol~., data=clean.frame,nvmax=8,                    nbest=1,really.big=TR

summary(best_subset)

## Subset selection object
```

```
## Call: regsubsets.formula(pb.TotChol ~ ., data = clean.frame, nvmax = 8,
##     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 8
## 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 ) "*"    "*" " "    "*"      "*"      "*"         "*"            "*"
## 8  ( 1 ) "*"    "*" "*"    "*"      "*"      "*"         "*"            "*"
```

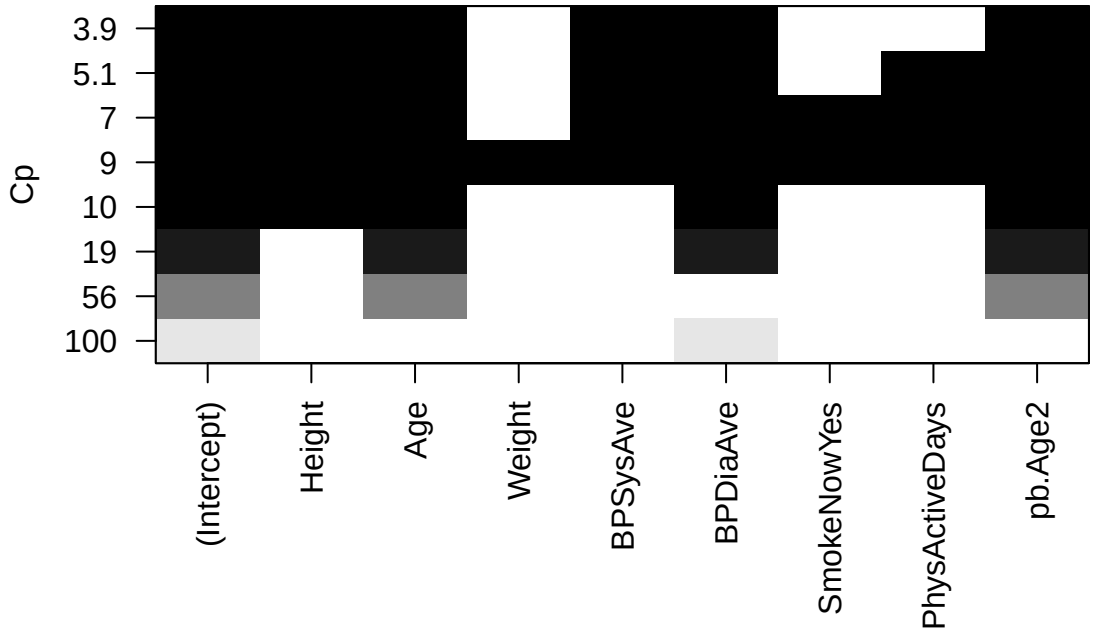
```
plot(best_subset,scale='adjr2')
```




```
plot(best_subset,scale='bic');
```




```
plot(best_subset,scale='Cp')
```




```
AIC <- step(clean_model, direction="both")
```

```
## Start:  AIC=-54.33
## pb.TotChol ~ Age + pb.Age2 + Weight + Height + BPSysAve + BPDiaAve +
##     SmokeNow + PhysActiveDays
##
##                   Df Sum of Sq    RSS     AIC
## - Weight           1     0.000 1216.7 -56.334
## - SmokeNow         1     0.086 1216.8 -56.244
## - PhysActiveDays   1     0.811 1217.5 -55.476
## <none>                         1216.7 -54.335
## - BPSysAve         1     7.884 1224.5 -48.022
```

```
## - Height          1     8.333 1225.0 -47.550
## - BPDiaAve        1    19.475 1236.1 -35.897
## - pb.Age2         1    65.377 1282.0  11.028
## - Age             1    74.647 1291.3  20.300
##
## Step:  AIC=-56.33
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + SmokeNow +
##     PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - SmokeNow        1     0.088 1216.8 -58.241
## - PhysActiveDays  1     0.811 1217.5 -57.476
## <none>                        1216.7 -56.334
## + Weight          1     0.000 1216.7 -54.335
## - BPSysAve        1     7.936 1224.6 -49.967
## - Height          1    10.536 1227.2 -47.237
## - BPDiaAve        1    19.546 1236.2 -37.823
## - pb.Age2         1    65.904 1282.6   9.557
## - Age             1    75.216 1291.9  18.868
##
## Step:  AIC=-58.24
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve + PhysActiveDays
##
##                  Df Sum of Sq    RSS     AIC
## - PhysActiveDays  1     0.811 1217.6 -59.384
## <none>                        1216.8 -58.241
## + SmokeNow        1     0.088 1216.7 -56.334
## + Weight          1     0.003 1216.8 -56.244
## - BPSysAve        1     8.071 1224.8 -51.731
## - Height          1    10.615 1227.4 -49.062
## - BPDiaAve        1    19.459 1236.2 -39.821
## - pb.Age2         1    66.037 1282.8   7.779
## - Age             1    75.131 1291.9  16.872
##
## Step:  AIC=-59.38
## pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve + BPDiaAve
##
##                  Df Sum of Sq    RSS     AIC
## <none>                        1217.6 -59.384
## + PhysActiveDays  1     0.811 1216.8 -58.241
## + SmokeNow        1     0.088 1217.5 -57.476
## + Weight          1     0.000 1217.6 -57.384
## - BPSysAve        1     7.982 1225.5 -52.974
## - Height          1    10.444 1228.0 -50.391
## - BPDiaAve        1    19.562 1237.1 -40.870
## - pb.Age2         1    65.411 1283.0   5.965
## - Age             1    74.398 1292.0  14.949
```

```
summary(AIC)
```

```
##
## Call:
## lm(formula = pb.TotChol ~ Age + pb.Age2 + Height + BPSysAve +
##     BPDiaAve, data = clean.frame)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5880 -0.6170 -0.0140  0.6438  3.1057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.8098341  0.5741846   8.377  < 2e-16 ***
## Age          0.0974977  0.0110200   8.847  < 2e-16 ***
## pb.Age2     -0.0009263  0.0001117  -8.296 2.70e-16 ***
## Height      -0.0098211  0.0029628  -3.315 0.000943 ***
## BPSysAve     0.0056469  0.0019487   2.898 0.003821 **
## BPDiaAve     0.0127101  0.0028016   4.537 6.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9749 on 1281 degrees of freedom
## Multiple R-squared:  0.1284, Adjusted R-squared:  0.125
## F-statistic: 37.73 on 5 and 1281 DF,  p-value: < 2.2e-16
```

```
#PREDICTION ACCURACY
set.seed(123)
train_index <- sample(1:nrow(clean.frame), 0.7 * nrow(clean.frame))
train_data <- clean.frame[train_index, ]
test_data <- clean.frame[-train_index, ]

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

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

```
## [1] 0.9542581
```

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

```
## [1] 0.9768613
```

```
library(caret)

#K-Fold (10-Fold)
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(
  pb.TotChol ~ ., data = clean.frame,
  method = "lm",
  trControl = train_control
)

print(cv_model)
```

```
## Linear Regression
##
## 1287 samples
##    8 predictor
##
## No pre-processing
```

```
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1158, 1159, 1158, 1158, 1158, 1159, ...
## Resampling results:
##
##   RMSE       Rsquared   MAE
##   0.9770068  0.1348559  0.771403
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```