# Zhang_Yue_FinalProject

Yue Zhang

2024-11-22

```
getwd()
```

```
## [1] "/Users/yuezhang/Documents/Biostat/Biostatistics/PHL_1700/Code"
```

```
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggthemes)
library(ggplot2)
library(readxl)
library(lmtest)
library(mfx)
library(pROC)
library(haven)
library(car)
library(PMCMRplus)
library(VGAM)
library(describedata)
library(olsrr)

mytheme = theme_clean(base_size = 12) +
  theme(
    axis.text = element_text(color = "black"),
    legend.position = "right",
    axis.text.x = element_text(
      angle = 45,
      vjust = 0.5,
      hjust = 0.5
    ),
    plot.title = element_text(size = 12)
  )
theme_set(mytheme)
```

#Load Data

```
bwt = read.csv("/Users/yuezhang/Documents/Biostat/Biostatistics/PHL_1700/Data/Raw/Birthweight data Chen
```

#Data Exploration

```
#Check Missing Data
bwt %>% summarize(across(everything(), ~sum(is.na(.))))
```

```
##   id low age lwt smoke ptl ht ui bwt race
## 1  0   0   0   0     0   0  0  0   0    0
```

```r
#Descriptive Statistics for Low Birth Weight
summary_low = bwt %>% filter(low == 1) %>% summarize(
  age_mean = mean(age, na.rm = TRUE),
  age_sd = sd(age, na.rm = TRUE),
  lwt_mean = mean(lwt, na.rm = TRUE),
  lwt_sd = sd(lwt, na.rm = TRUE),
  race_prop = list(prop.table(table(race))),
  smoke_prop = mean(smoke == 1, na.rm = TRUE),
  ptl_prop = mean(ptl == 1, na.rm = TRUE),
  ht_prop = mean(ht == 1, na.rm = TRUE),
  ui_prop = mean(ui == 1, na.rm = TRUE)
)

print(summary_low)
```

```
##   age_mean   age_sd lwt_mean   lwt_sd                        race_prop
## 1 22.36839 5.805274 143.3286 28.67558 0.4944238, 0.1078067, 0.3977695
##   smoke_prop ptl_prop   ht_prop   ui_prop
## 1  0.4832714 0.197026 0.0929368 0.2416357
```

```r
print(summary_low$race_prop)
```

```
## [[1]]
## race
##         1         2         3
## 0.4944238 0.1078067 0.3977695
```

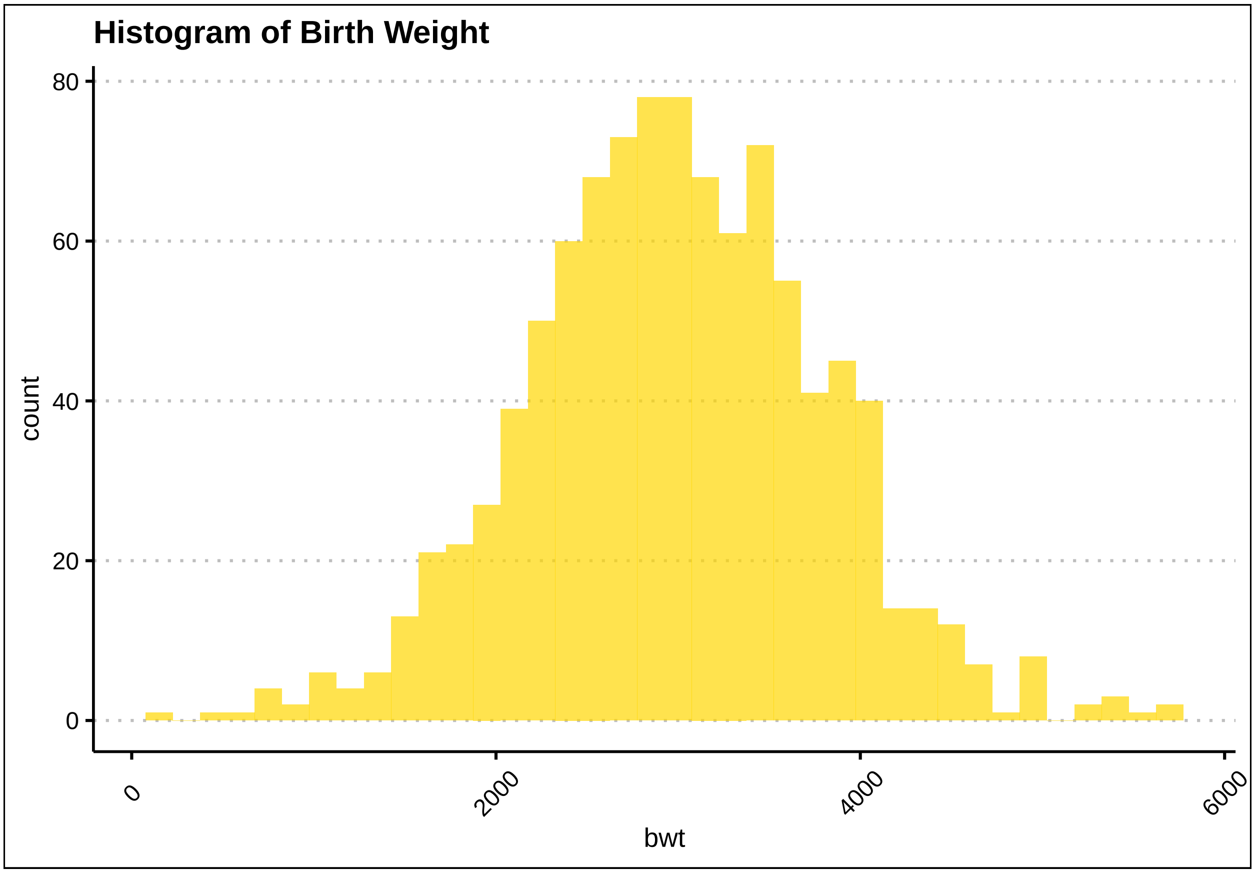
```r
low = bwt %>% filter(low == 1)
cau = low %>% filter(race == 1)
aa = low %>% filter(race == 2)
other = low %>% filter(race ==3)
smoke = low %>% filter(smoke ==1)
pl = low %>% filter(ptl == 1)
hy = low %>% filter(ht == 1)
u = low %>% filter(ui == 1)

high = bwt %>% filter(low == 0)
cau2 = high %>% filter(race == 1)
aa2 = high %>% filter(race == 2)
other2 = high %>% filter(race ==3)
smoke2 = high %>% filter(smoke ==1)
pl2 = high %>% filter(ptl == 1)
hy2 = high %>% filter(ht == 1)
u2 = high %>% filter(ui == 1)

#Descriptive Statistics for High Birth Weight
summary_high = bwt %>% filter(low == 0) %>% summarize(
  age_mean = mean(age, na.rm = TRUE),
  age_sd = sd(age, na.rm = TRUE),
  lwt_mean = mean(lwt, na.rm = TRUE),
  lwt_sd = sd(lwt, na.rm = TRUE),
  race_prop = list(prop.table(table(race))),
  smoke_prop = mean(smoke == 1, na.rm = TRUE),
  ptl_prop = mean(ptl == 1, na.rm = TRUE),
  ht_prop = mean(ht == 1, na.rm = TRUE),
  ui_prop = mean(ui == 1, na.rm = TRUE)
)

print(summary_high)
```

```
##   age_mean   age_sd lwt_mean   lwt_sd                           race_prop
## 1 23.04547 5.497755 153.1519 32.25093 0.5745554, 0.1244870, 0.3009576
##   smoke_prop  ptl_prop    ht_prop   ui_prop
## 1  0.3584131 0.1600547 0.05471956 0.1025992
```

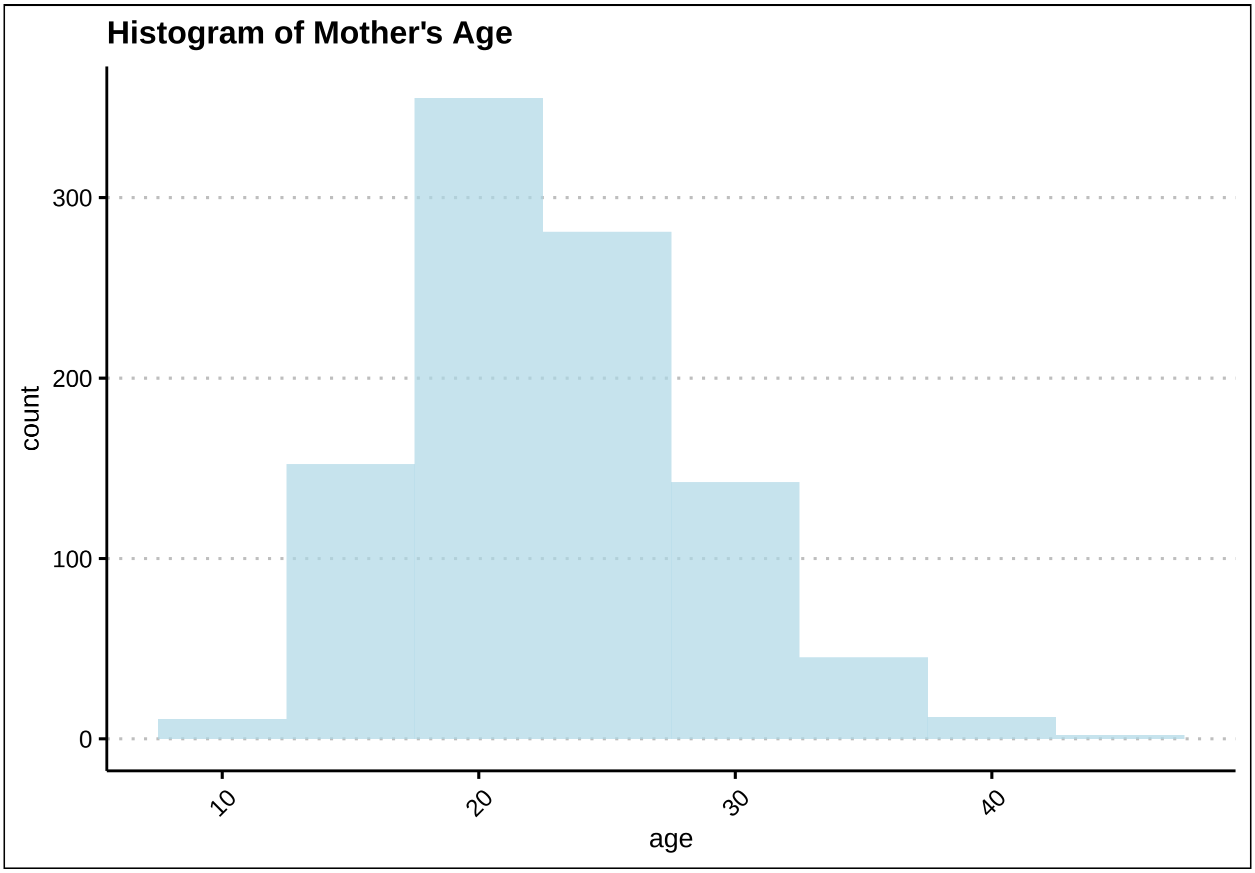
```
print(summary_high$race_prop)
```

```
## [[1]]
## race
##         1         2         3
## 0.5745554 0.1244870 0.3009576
```

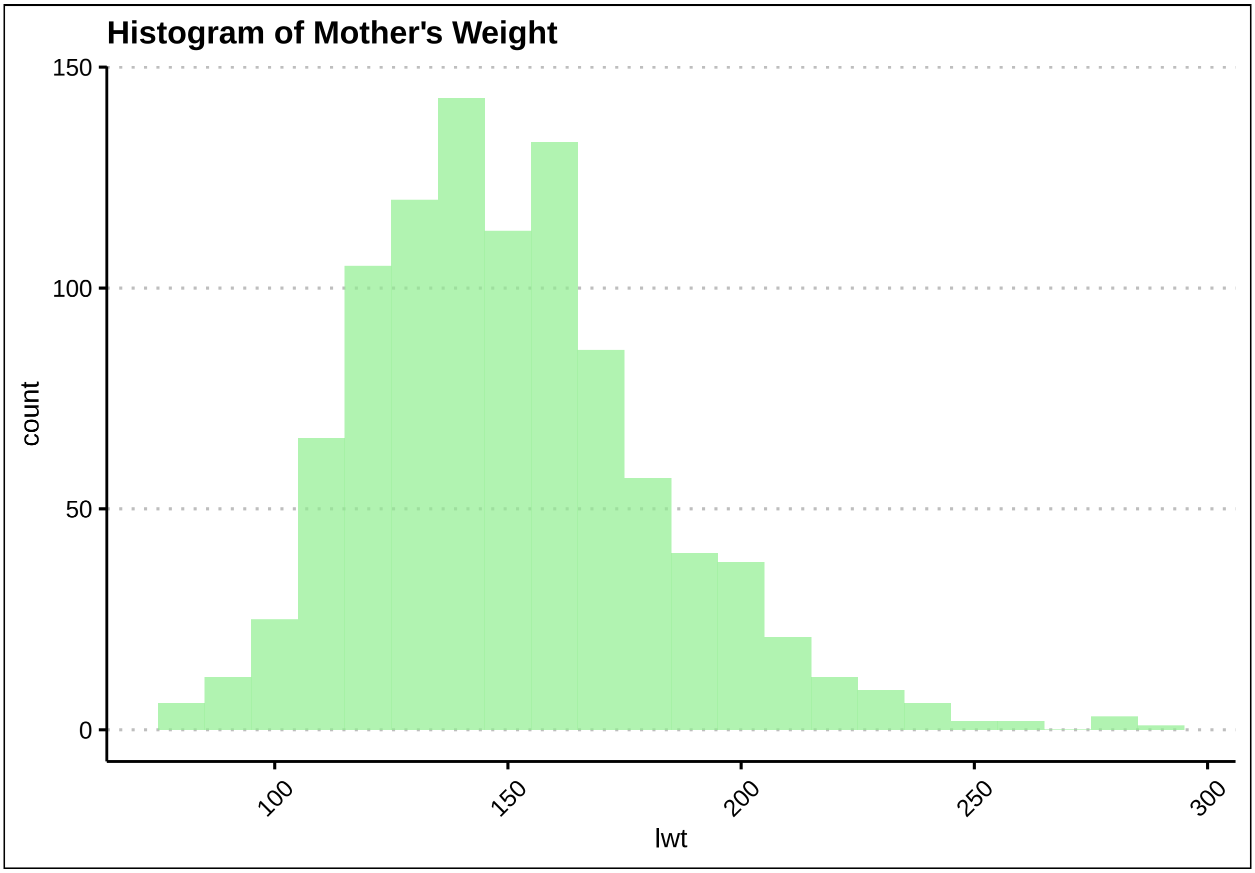
```
#Histogram of Birth Weight
ggplot(bwt, aes(x = bwt)) +
  geom_histogram(binwidth = 150, fill = "gold", alpha = 0.7) +
  ggtitle("Histogram of Birth Weight")
```



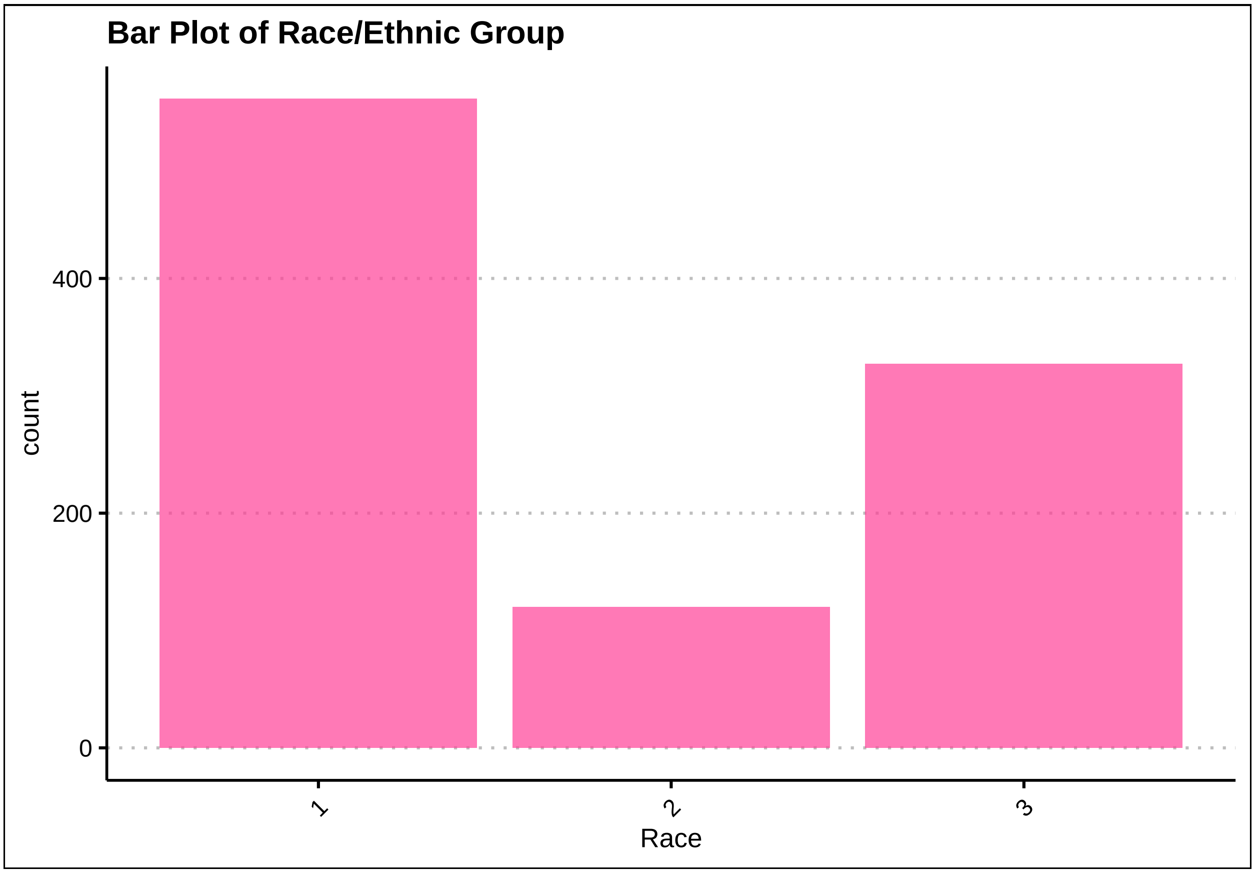
```
#Histogram of Mother's Age
ggplot(bwt, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "lightblue", alpha = 0.7) +
  ggtitle("Histogram of Mother's Age")
```



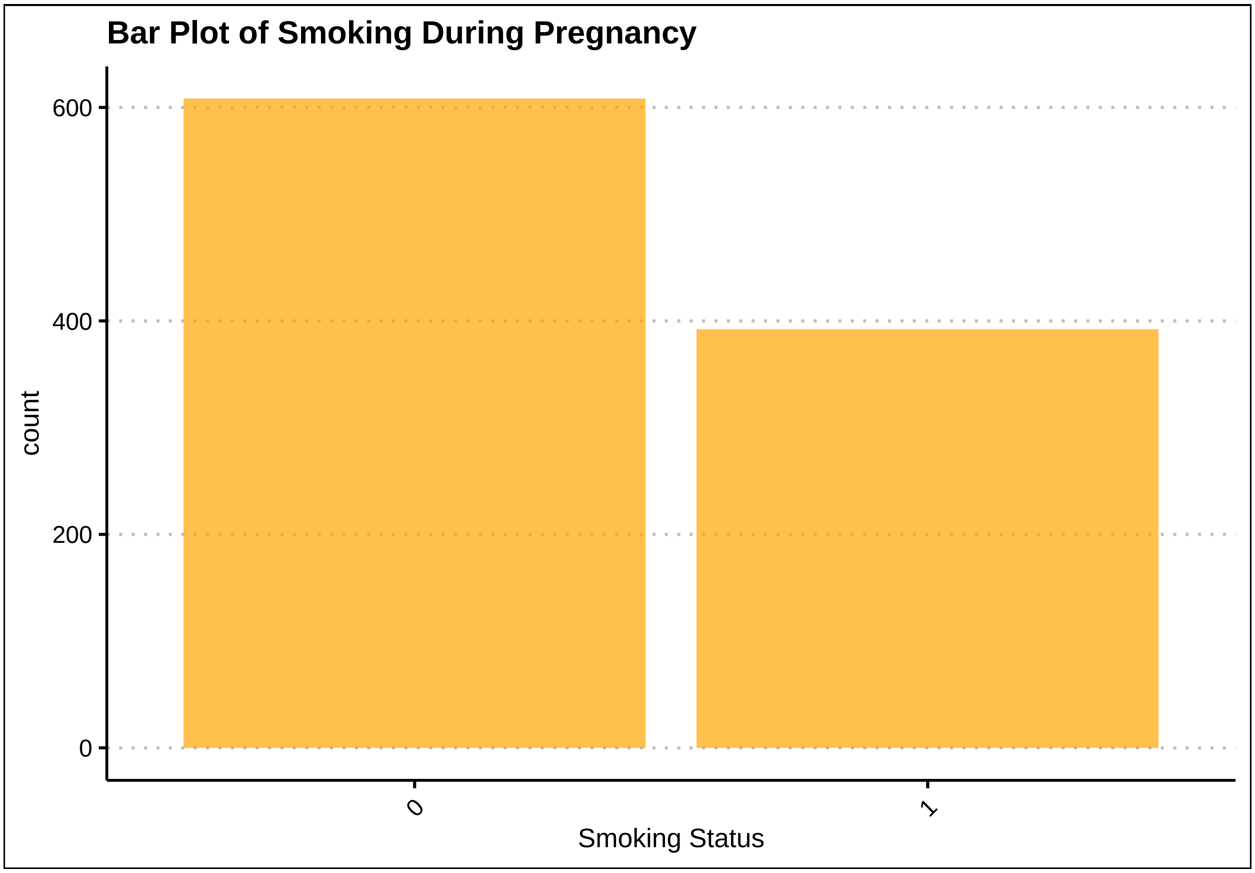
```
#Histogram of Mother's Weight at Last Menstrual Period
ggplot(bwt, aes(x = lwt)) +
  geom_histogram(binwidth = 10, fill = "lightgreen", alpha = 0.7) +
  ggtitle("Histogram of Mother's Weight")
```



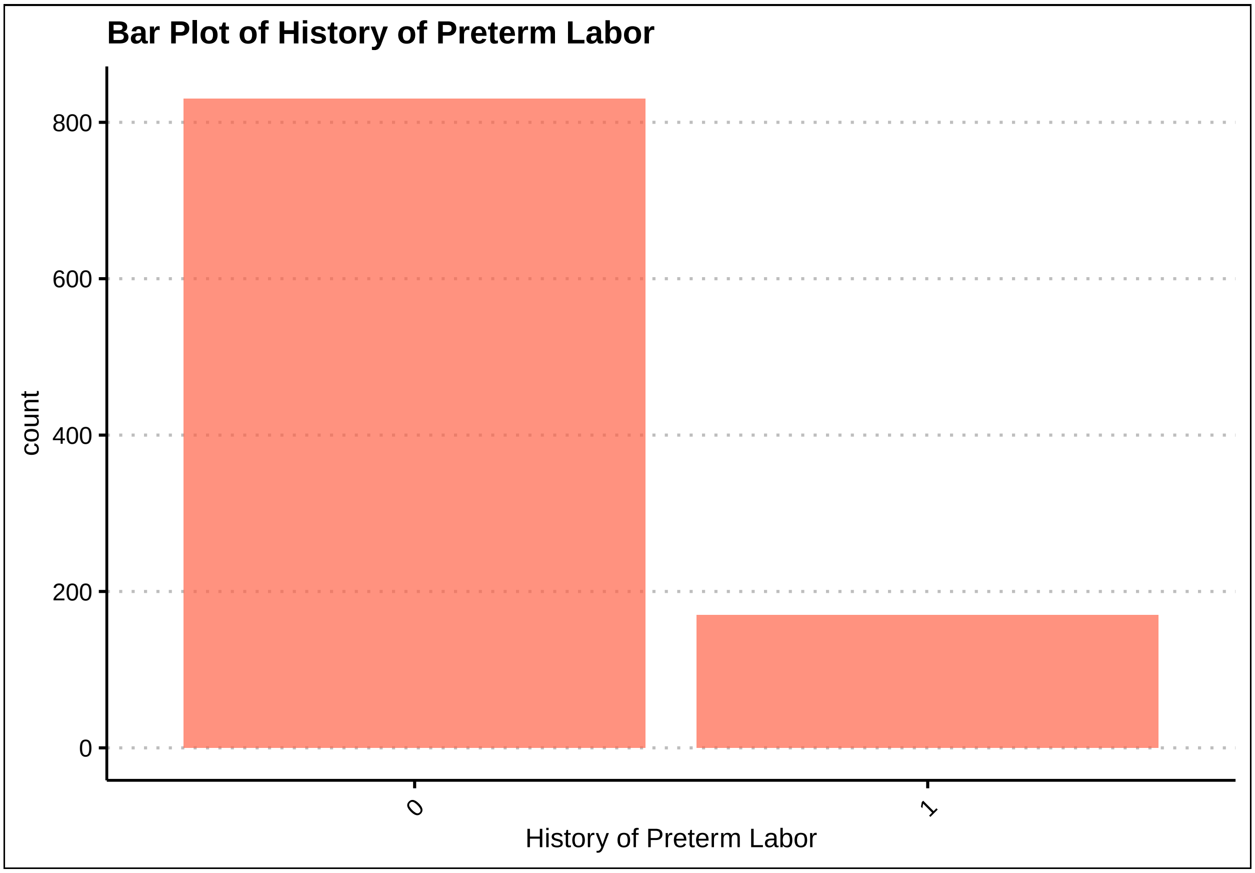
```
#Bar Plot of Race
ggplot(bwt, aes(x = factor(race))) +
  geom_bar(fill = "violetred1", alpha = 0.7) +
  xlab("Race") +
  ggtitle("Bar Plot of Race/Ethnic Group")
```



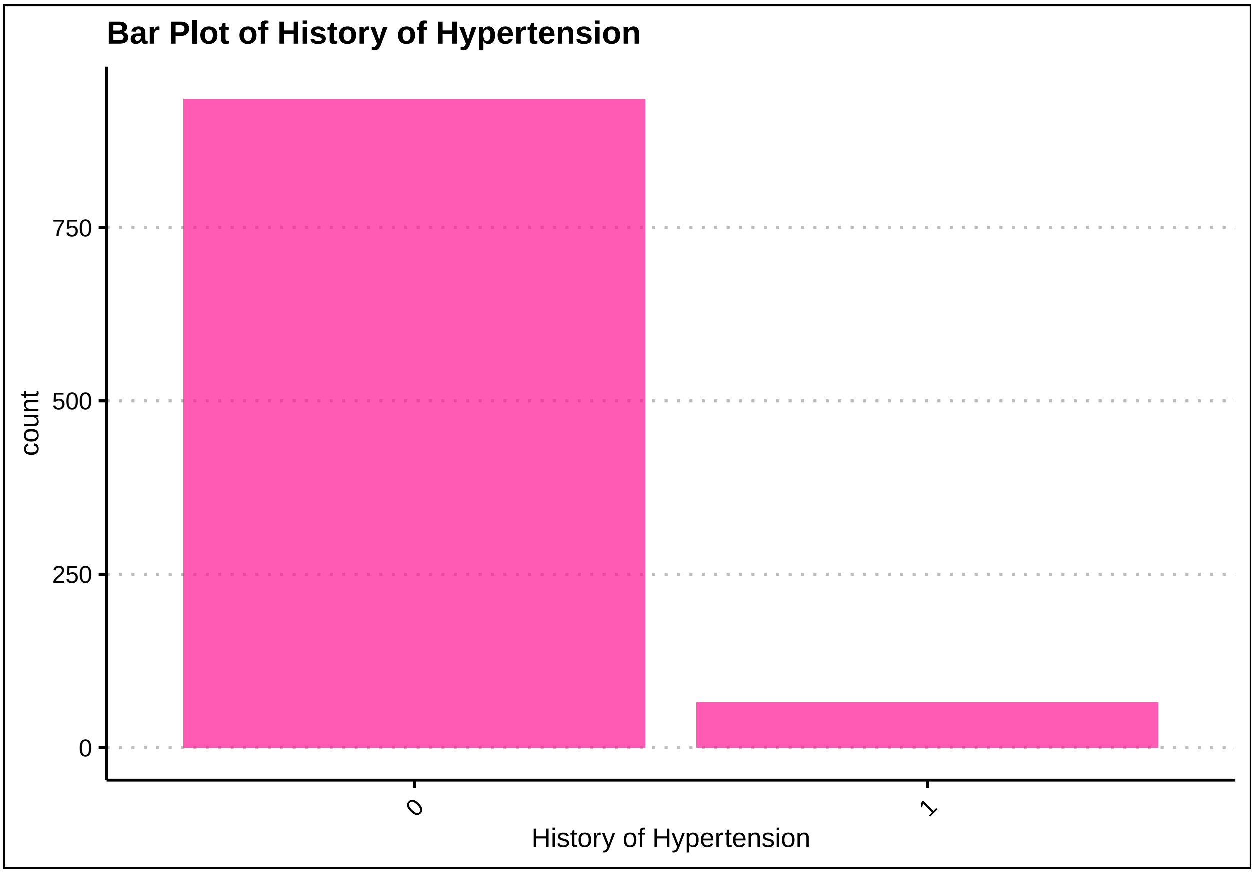
```
#Bar Plot of Smoking Status
ggplot(bwt, aes(x = factor(smoke))) +
  geom_bar(fill = "orange", alpha = 0.7) +
  xlab("Smoking Status") +
  ggtitle("Bar Plot of Smoking During Pregnancy")
```



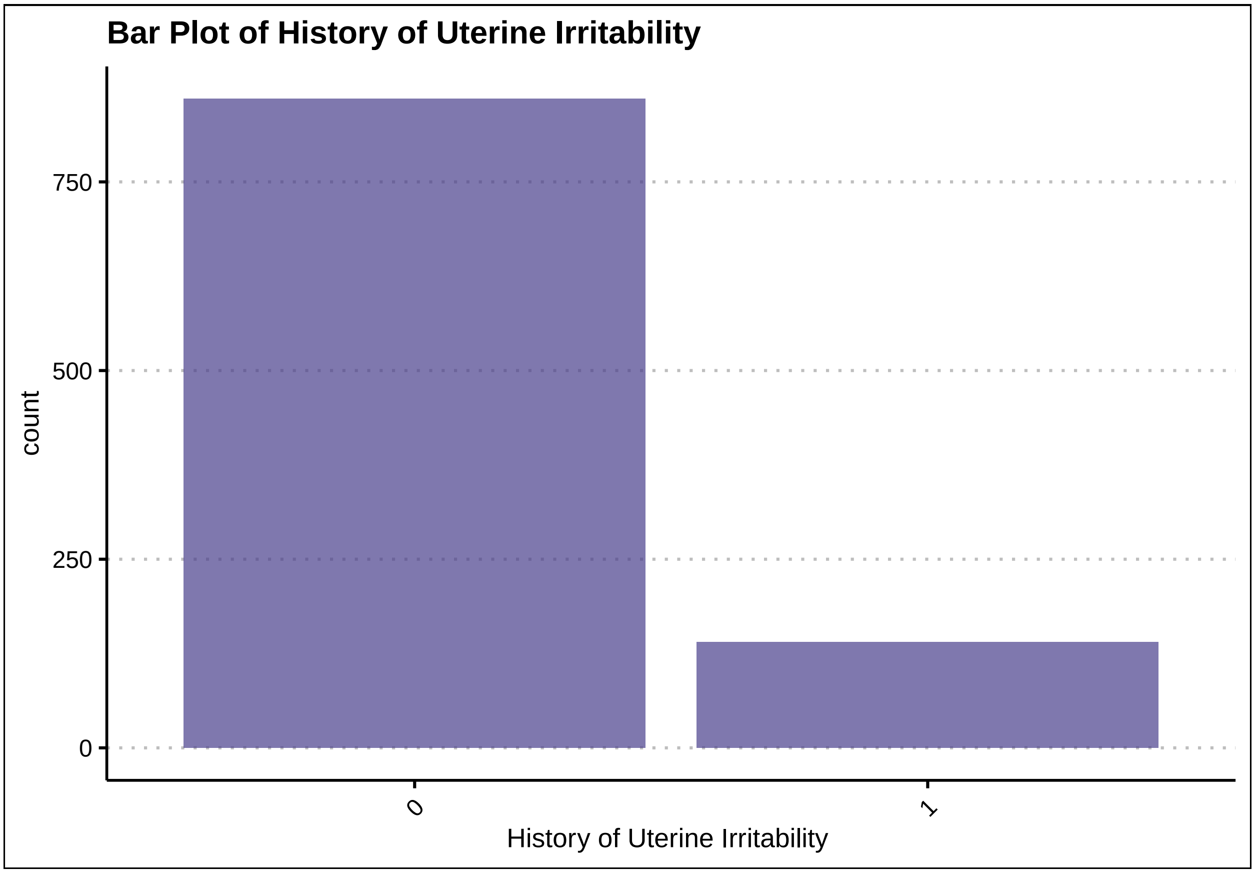
```
#Bar Plot of History of Preterm Labor
ggplot(bwt, aes(x = factor(ptl))) +
  geom_bar(fill = "tomato1", alpha = 0.7) +
  xlab("History of Preterm Labor") +
  ggtitle("Bar Plot of History of Preterm Labor")
```



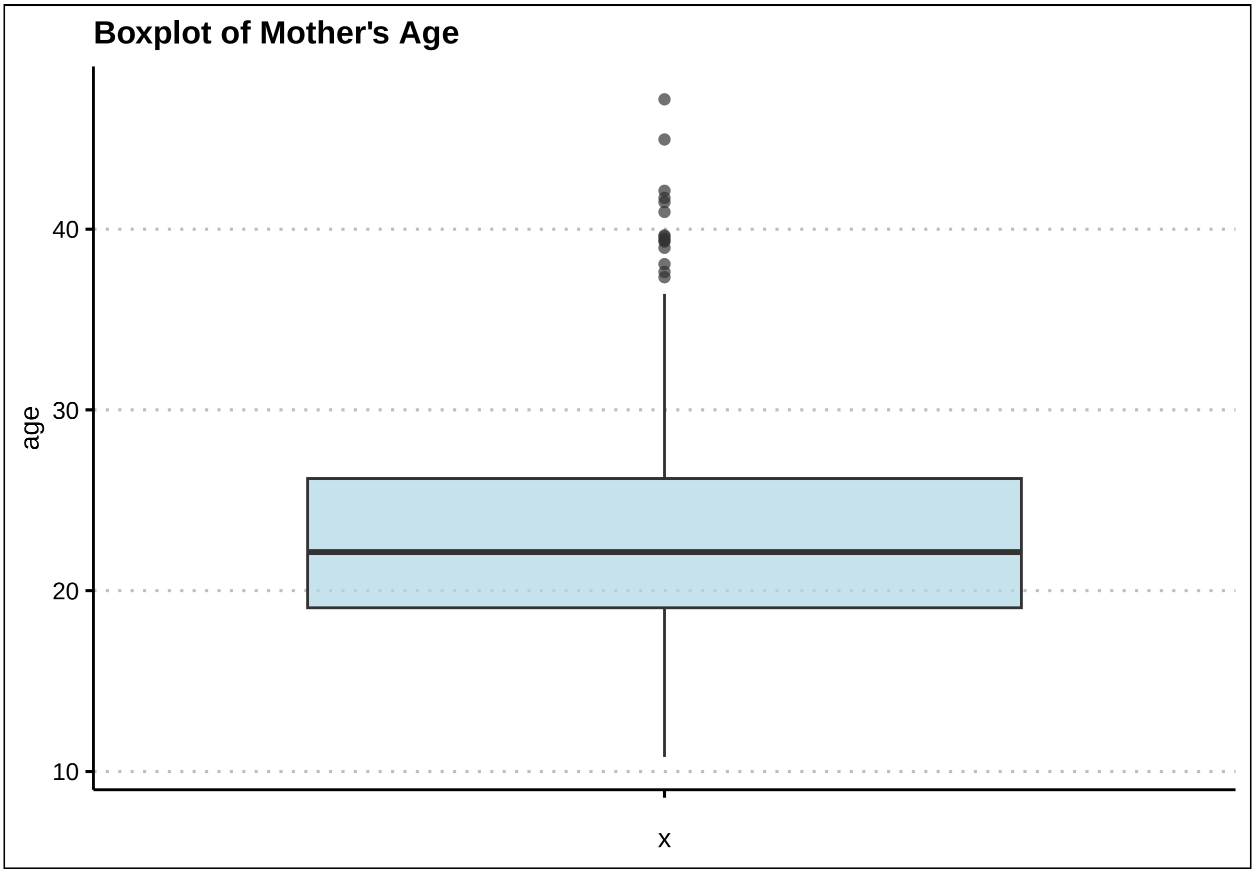
```
#Bar Plot of History of Hypertension
ggplot(bwt, aes(x = factor(ht))) +
  geom_bar(fill = "deeppink", alpha = 0.7) +
  xlab("History of Hypertension") +
  ggtitle("Bar Plot of History of Hypertension")
```



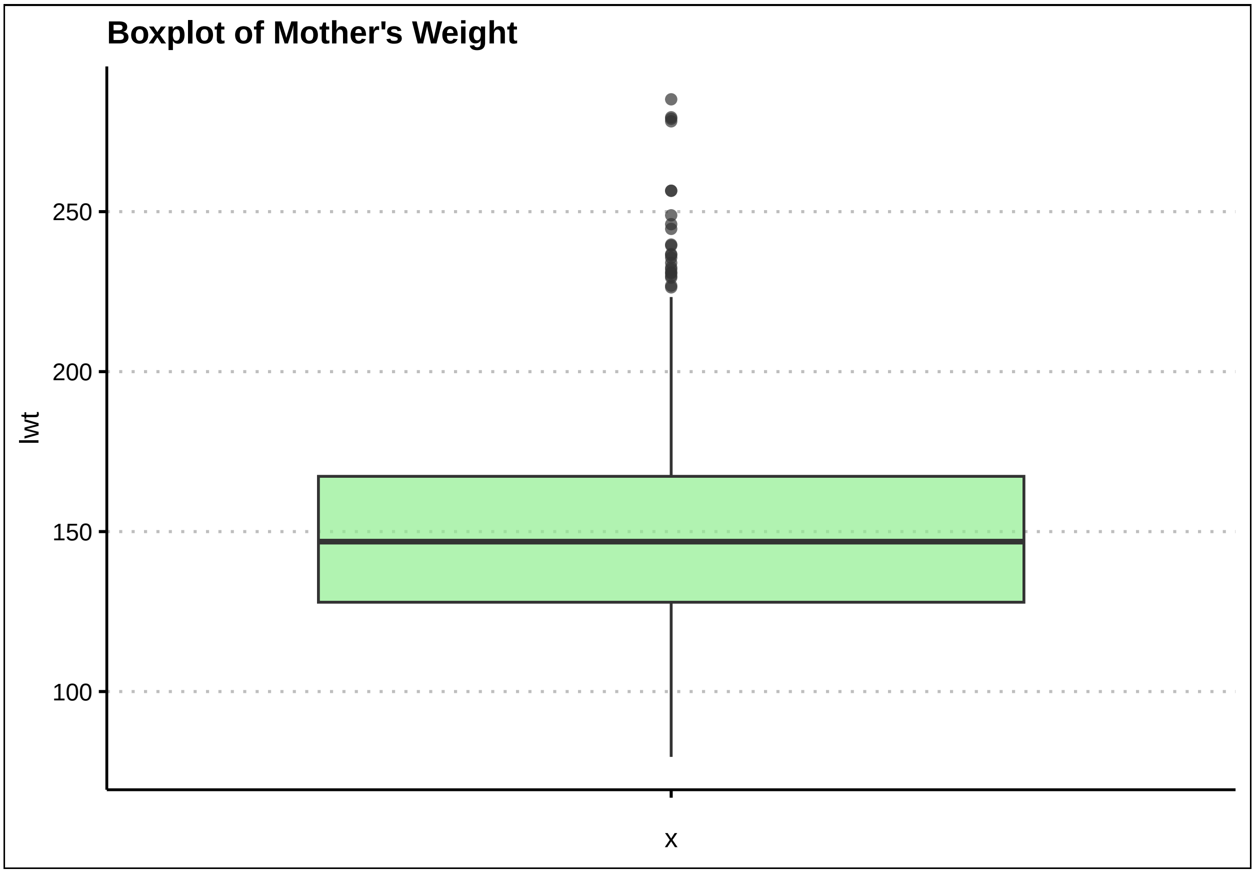
```
#Bar Plot of History of Uterine Irritability
ggplot(bwt, aes(x = factor(ui))) +
  geom_bar(fill = "darkslateblue", alpha = 0.7) +
  xlab("History of Uterine Irritability ") +
  ggtitle("Bar Plot of History of Uterine Irritability ")
```



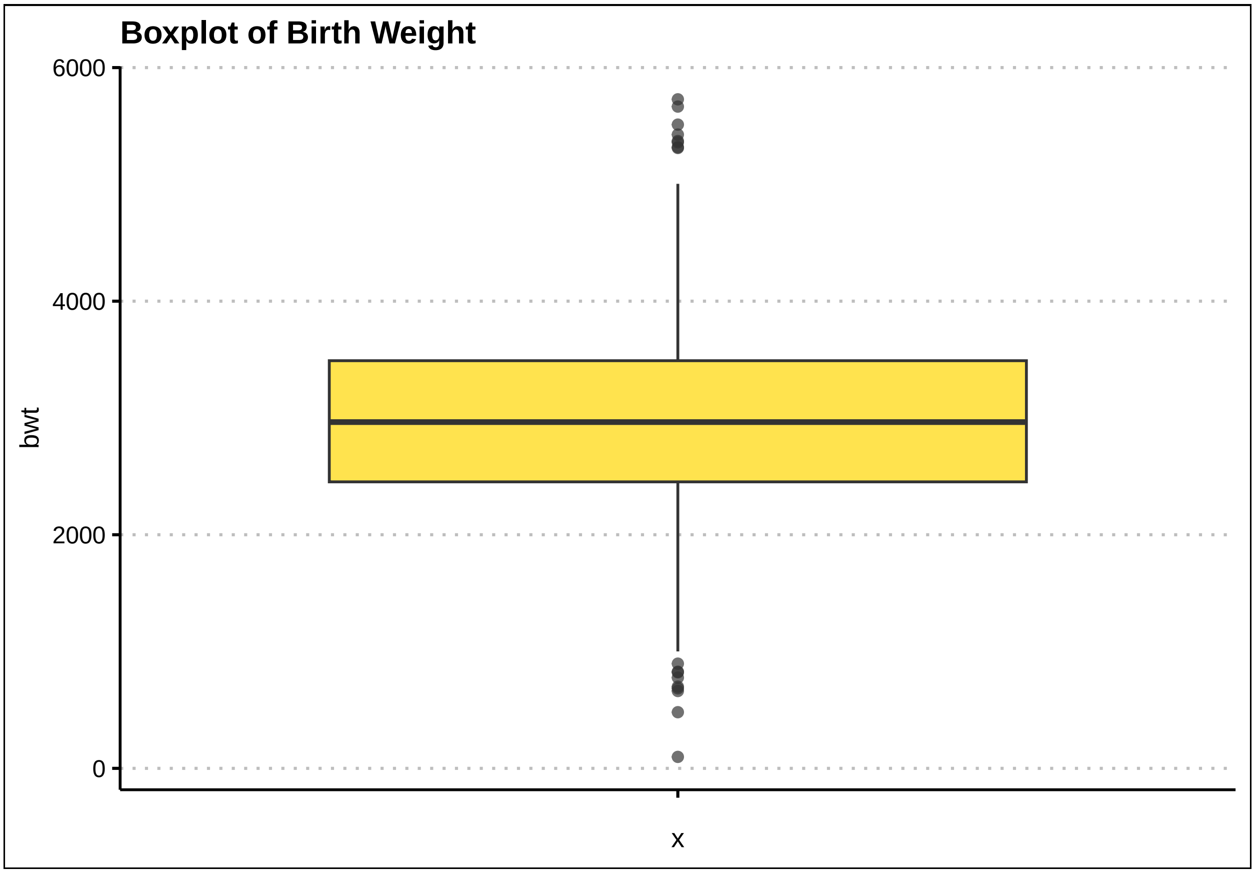
```
#Identify Outliers
ggplot(bwt, aes(x = "", y = age)) +
  geom_boxplot(fill = "lightblue", alpha = 0.7) +
  ggtitle("Boxplot of Mother's Age")
```



```
ggplot(bwt, aes(x = "", y = lwt)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.7) +
  ggtitle("Boxplot of Mother's Weight")
```



```r
ggplot(bwt, aes(x = "", y = bwt)) +
  geom_boxplot(fill = "gold", alpha = 0.7) +
  ggtitle("Boxplot of Birth Weight")
```



#Testing Simple Association

```
#Perform a t-test comparing age between low birth weight (low = 1) and normal birth weight (low = 0)
t.test(age ~ low, data = bwt)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by low
## t = 1.6587, df = 455.86, p-value = 0.09787
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1251195  1.4792737
## sample estimates:
## mean in group 0 mean in group 1
##        23.04547        22.36839
```

```
#Perform a t-test comparing weight between low birth weight (low = 1) and normal birth weight (low = 0)
t.test(lwt ~ low, data = bwt)
```

```
##
##  Welch Two Sample t-test
##
## data:  lwt by low
## t = 4.6413, df = 533.15, p-value = 4.364e-06
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   5.665601 13.981135
## sample estimates:
## mean in group 0 mean in group 1
##        153.1519        143.3286
```

```
#Chi-square test for association between smoking status and birth weight
chisq.test(table(bwt$smoke, bwt$low))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(bwt$smoke, bwt$low)
## X-squared = 12.344, df = 1, p-value = 0.0004425
```

```
#Chi-square test for association between race and birth weight
chisq.test(table(bwt$race, bwt$low))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(bwt$race, bwt$low)
## X-squared = 8.3753, df = 2, p-value = 0.01518
```

```
bwt$is_caucasian = ifelse(bwt$race == 1, 1, 0)

caucasian_low_table = table(bwt$is_caucasian, bwt$low)

print(caucasian_low_table)
```

```
##
##       0   1
##   0 311 136
##   1 420 133
```

```
chisq_result = chisq.test(caucasian_low_table)
chisq_result
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  caucasian_low_table
## X-squared = 4.7889, df = 1, p-value = 0.02864
```

```
bwt$is_aa = ifelse(bwt$race == 2, 1, 0)

aa_low_table = table(bwt$is_aa, bwt$low)

print(aa_low_table)
```

```
##
##       0   1
##   0 640 240
##   1  91  29
```

```
chisq_result2 = chisq.test(aa_low_table)
chisq_result2
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
## 
## data:  aa_low_table
## X-squared = 0.37218, df = 1, p-value = 0.5418
```

```
bwt$is_o = ifelse(bwt$race == 3, 1, 0)

o_low_table = table(bwt$is_o, bwt$low)

print(o_low_table)
```

```
##    
##       0   1
##   0 511 162
##   1 220 107
```

```
chisq_result3 = chisq.test(o_low_table)
chisq_result3
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  o_low_table
## X-squared = 7.9405, df = 1, p-value = 0.004834
```

```
#Chi-square test for association between preterm labor and birth weight
chisq.test(table(bwt$ptl, bwt$low))
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  table(bwt$ptl, bwt$low)
## X-squared = 1.6519, df = 1, p-value = 0.1987
```

```
#Chi-square test for association between history of hypertension and birth weight
chisq.test(table(bwt$ht, bwt$low))
```

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  table(bwt$ht, bwt$low)
## X-squared = 4.1178, df = 1, p-value = 0.04244
```

```
#Chi-square test for association between history of uterine irritability and birth weight
chisq.test(table(bwt$ui, bwt$low))
```

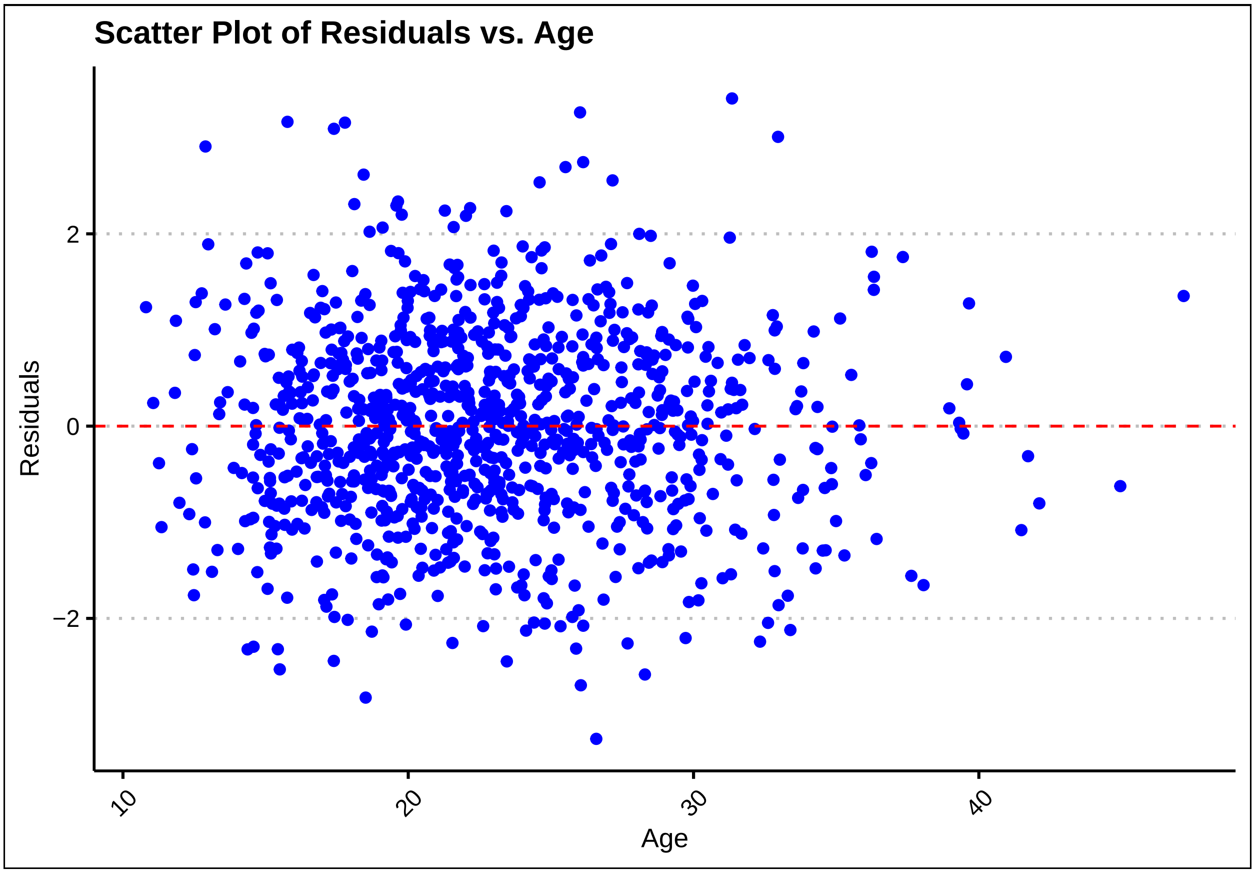
```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  table(bwt$ui, bwt$low)
## X-squared = 30.428, df = 1, p-value = 3.465e-08
```

#Linear Regression

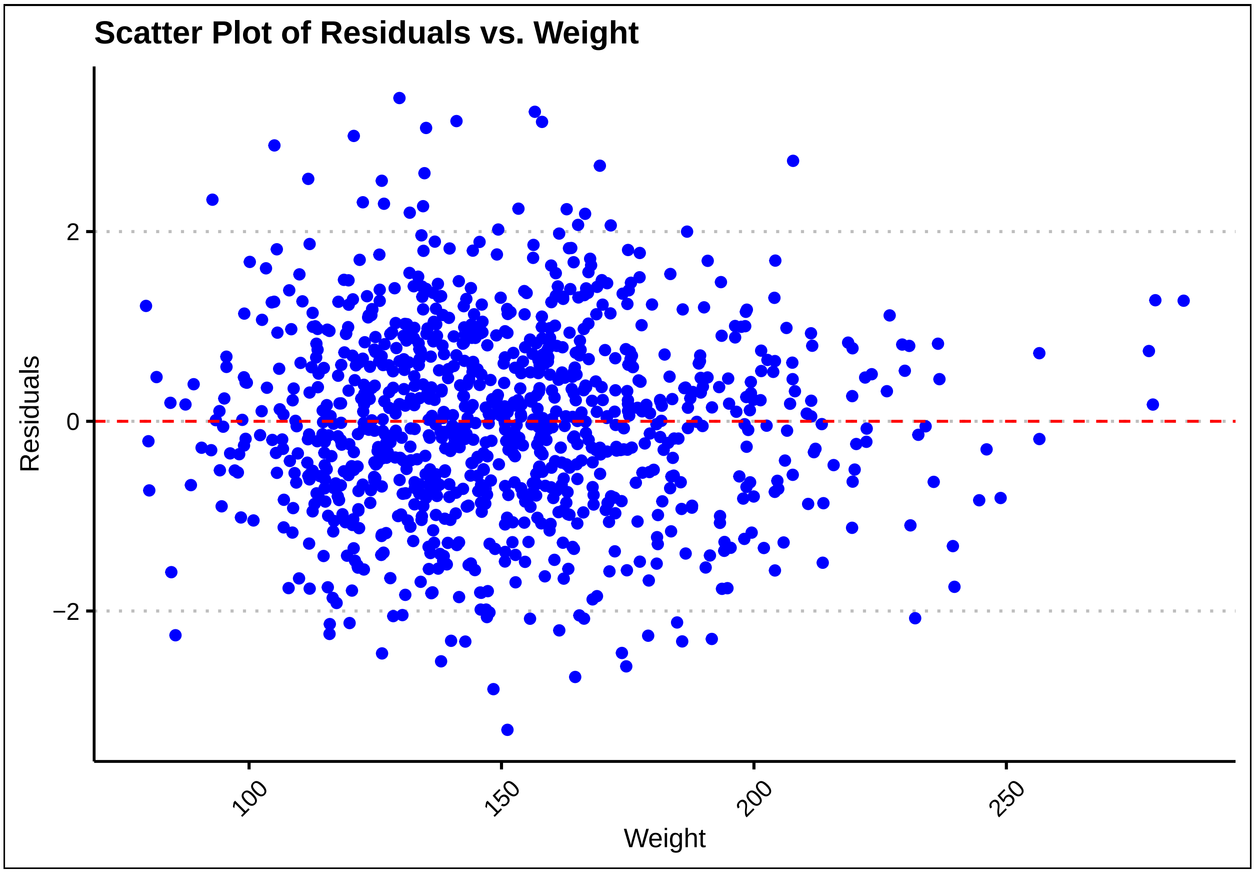
```
bwt_model = lm(bwt ~ age + lwt + race + smoke + ptl + ht + ui, data = bwt)
summary(bwt_model)
```

```
## 
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui, 
##     data = bwt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2404.64  -498.20    -5.41   495.22  2518.72
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2747.2149   163.8710  16.765  < 2e-16 ***
## age           -1.2518     4.3744  -0.286 0.774821
## lwt            4.6710     0.7806   5.984 3.04e-09 ***
## race        -135.4816    27.0320  -5.012 6.38e-07 ***
## smoke       -275.5430    50.2206  -5.487 5.20e-08 ***
## ptl           57.7048    64.3120   0.897 0.369797
## ht          -330.3174    97.3277  -3.394 0.000716 ***
## ui          -560.3181    68.6249  -8.165 9.71e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 745.7 on 992 degrees of freedom
## Multiple R-squared:  0.1538, Adjusted R-squared:  0.1478
## F-statistic: 25.76 on 7 and 992 DF,  p-value: < 2.2e-16
```

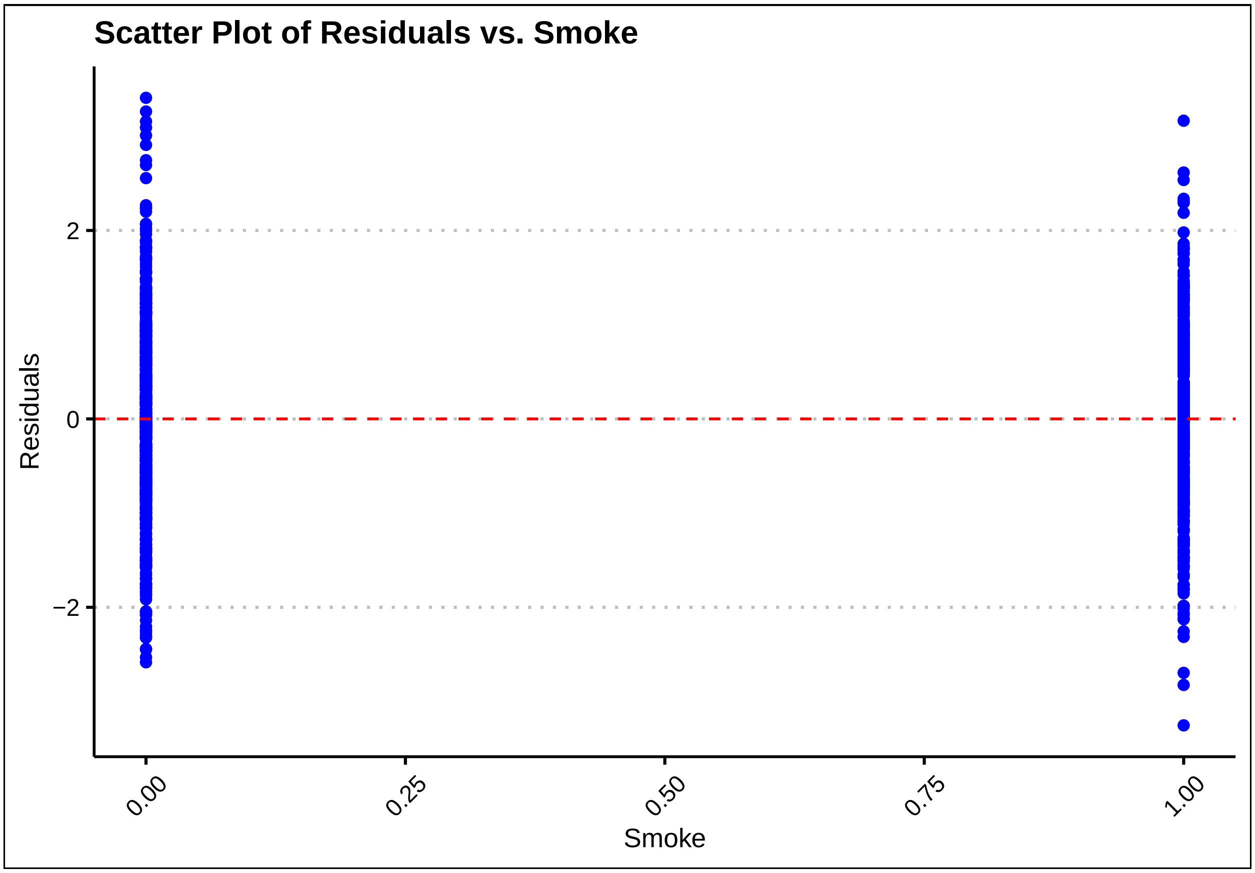
```r
sr = rstudent(bwt_model)

#Scatter Plot of Studentized Residuals vs. Age
ggplot(bwt, aes(x = age, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Age",
       x = "Age",
       y = "Residuals")
```



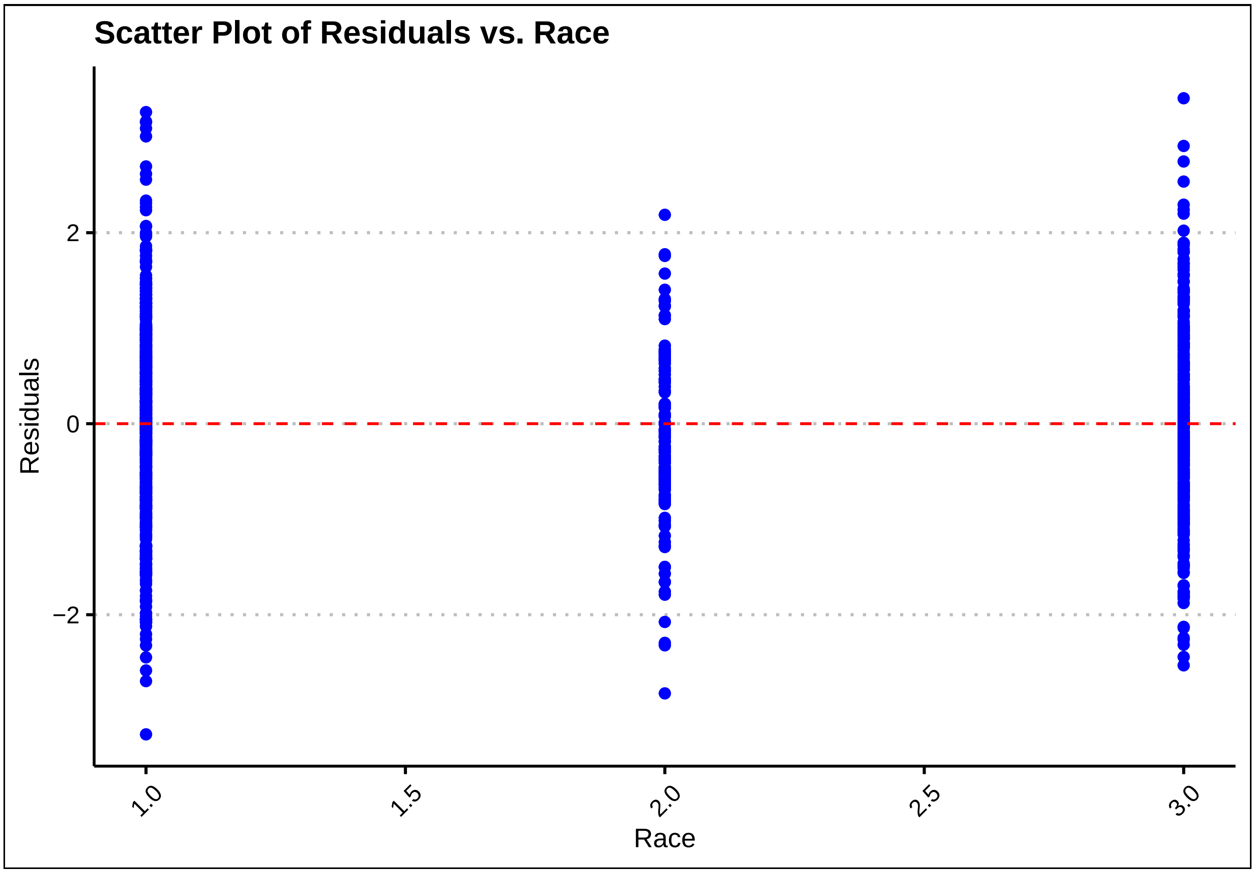
```
#Scatter Plot of Studentized Residuals vs. Weight
ggplot(bwt, aes(x = lwt, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Weight",
       x = "Weight",
       y = "Residuals")
```



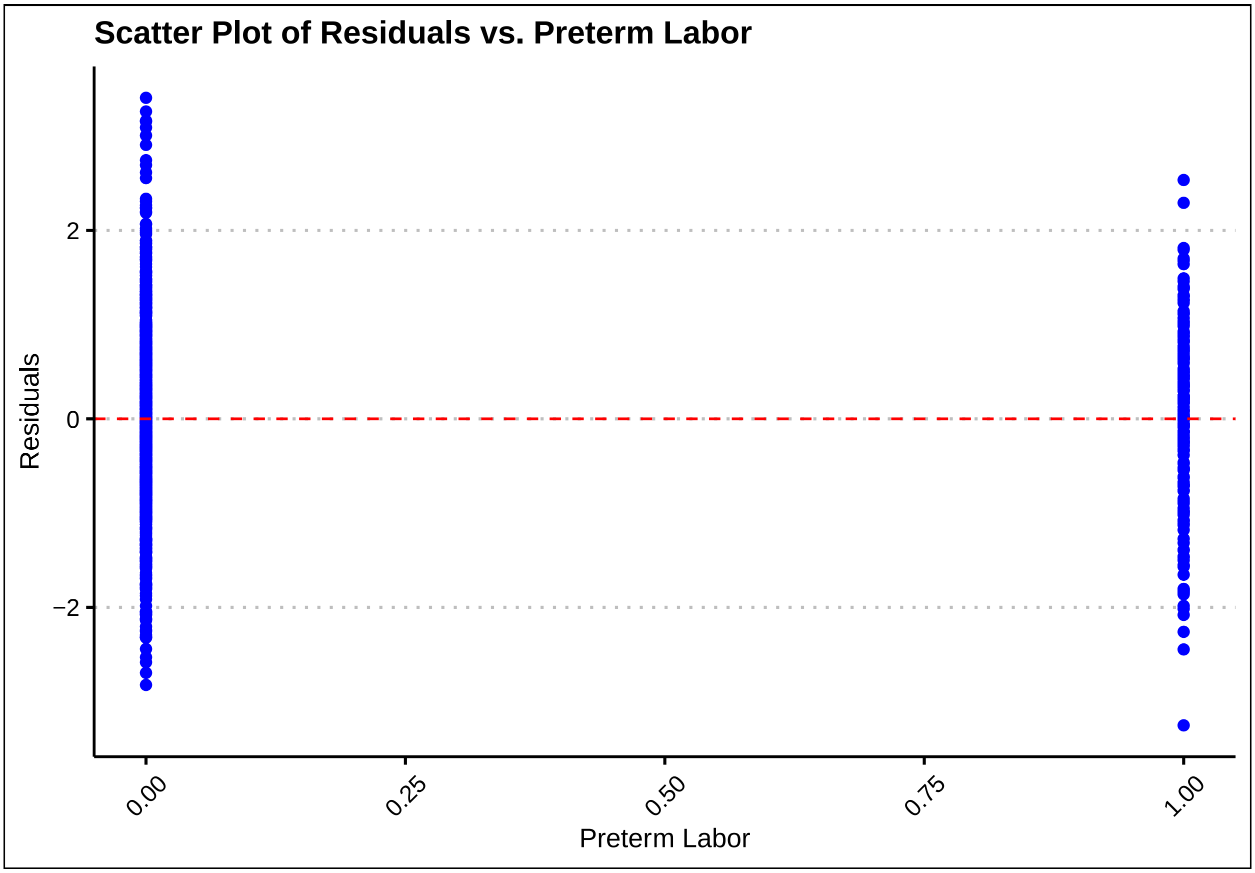
```
#Scatter Plot of Studentized Residuals vs. Smoke
ggplot(bwt, aes(x = smoke, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Smoke",
       x = "Smoke",
       y = "Residuals")
```



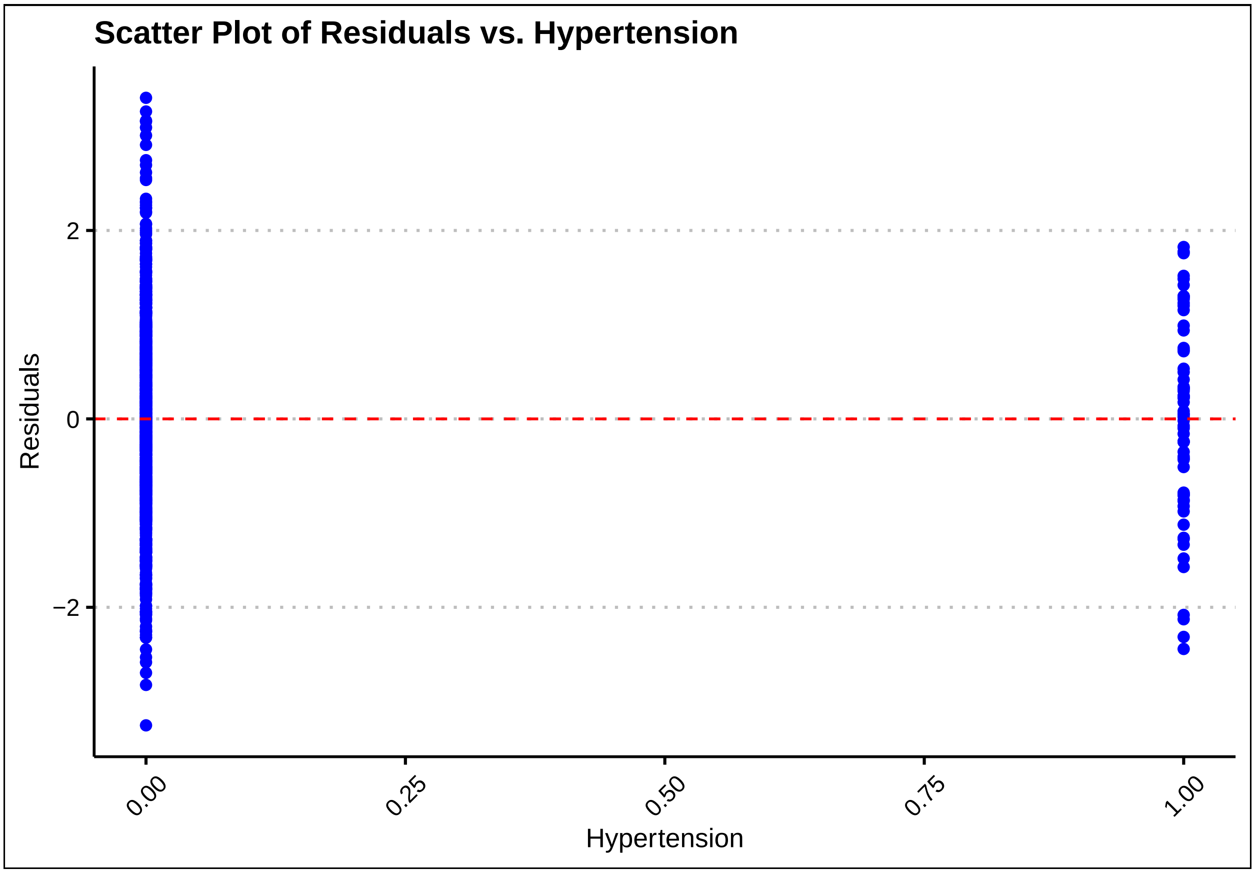
```
#Scatter Plot of Studentized Residuals vs. Race
ggplot(bwt, aes(x = race, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Race",
       x = "Race",
       y = "Residuals")
```



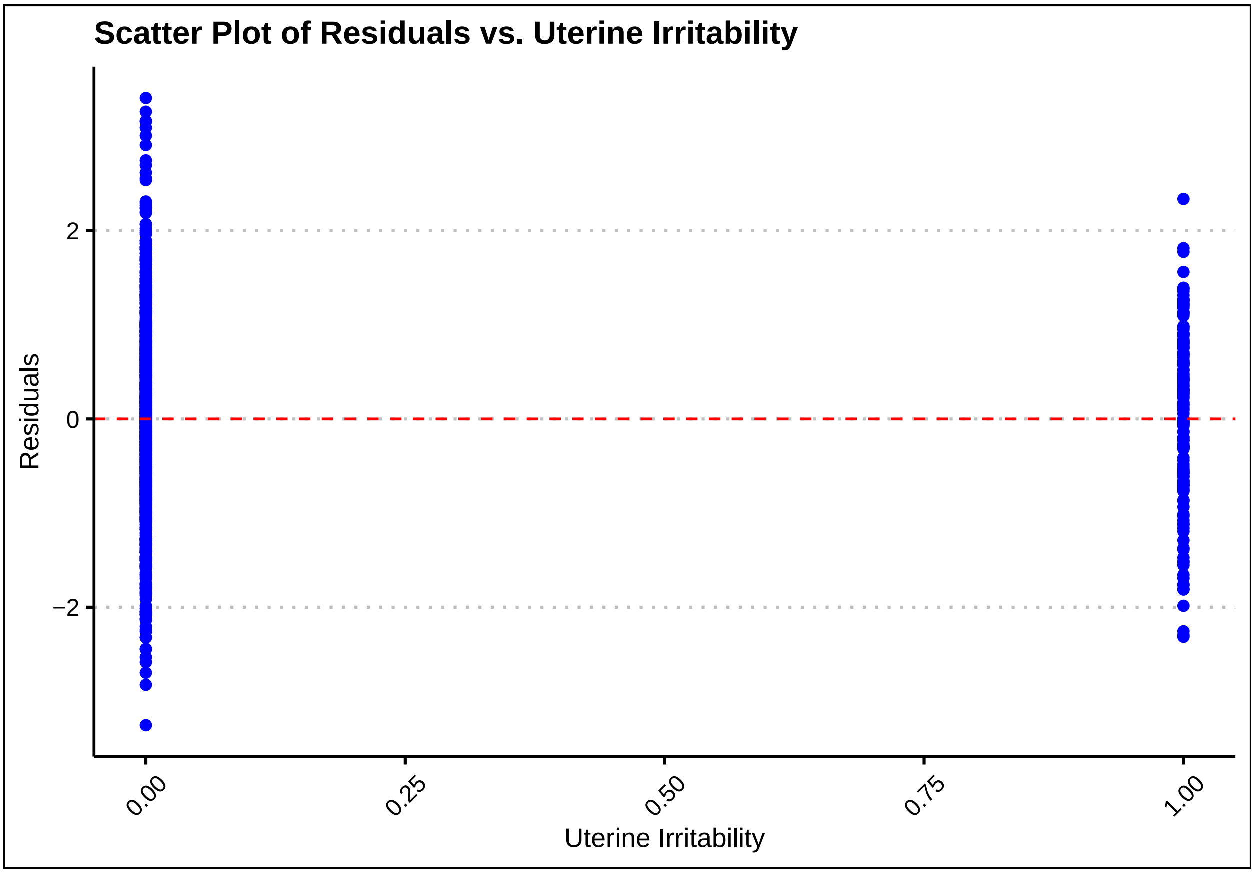
```
#Scatter Plot of Studentized Residuals vs. Preterm Labor
ggplot(bwt, aes(x = ptl, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Preterm Labor",
       x = "Preterm Labor",
       y = "Residuals")
```



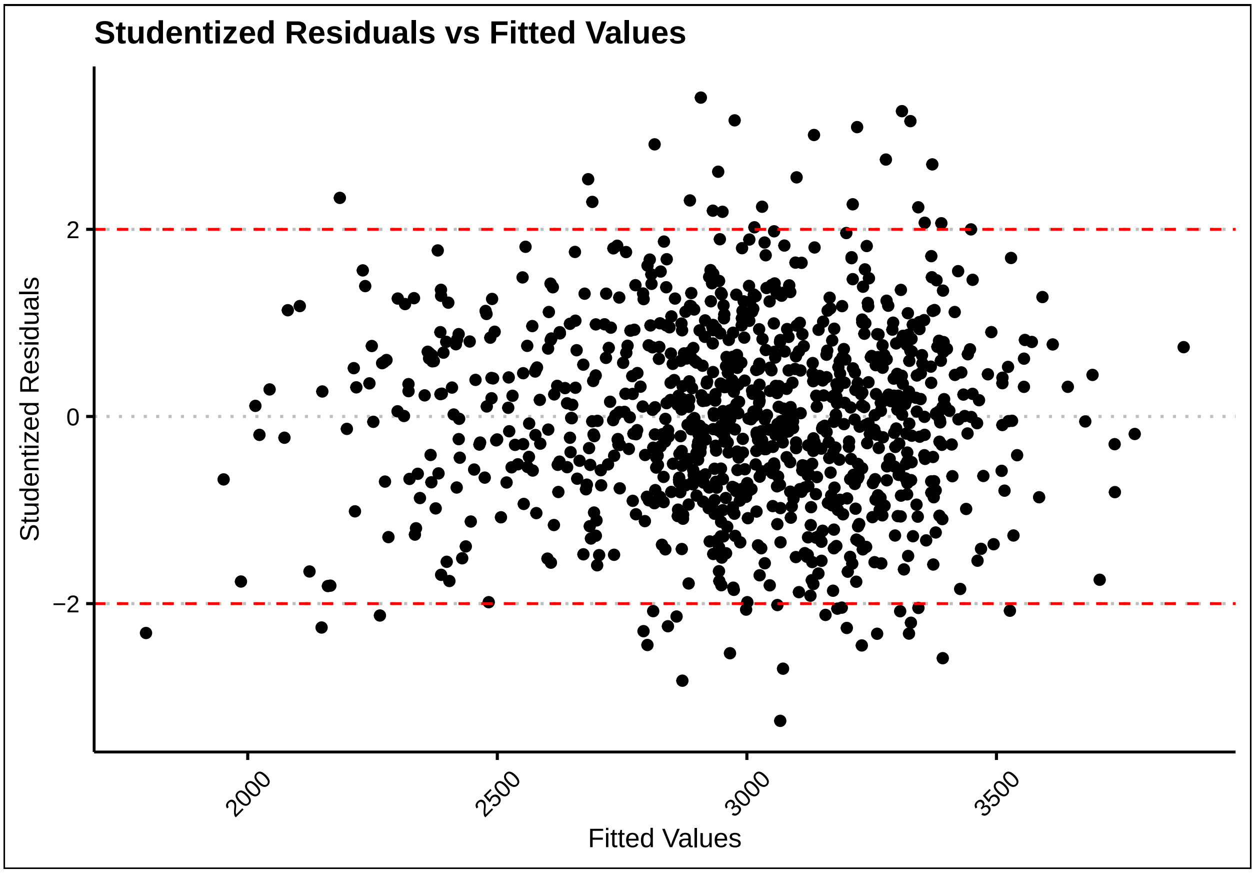
```
#Scatter Plot of Studentized Residuals vs. Hypertension
ggplot(bwt, aes(x = ht, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Hypertension",
       x = "Hypertension",
       y = "Residuals")
```



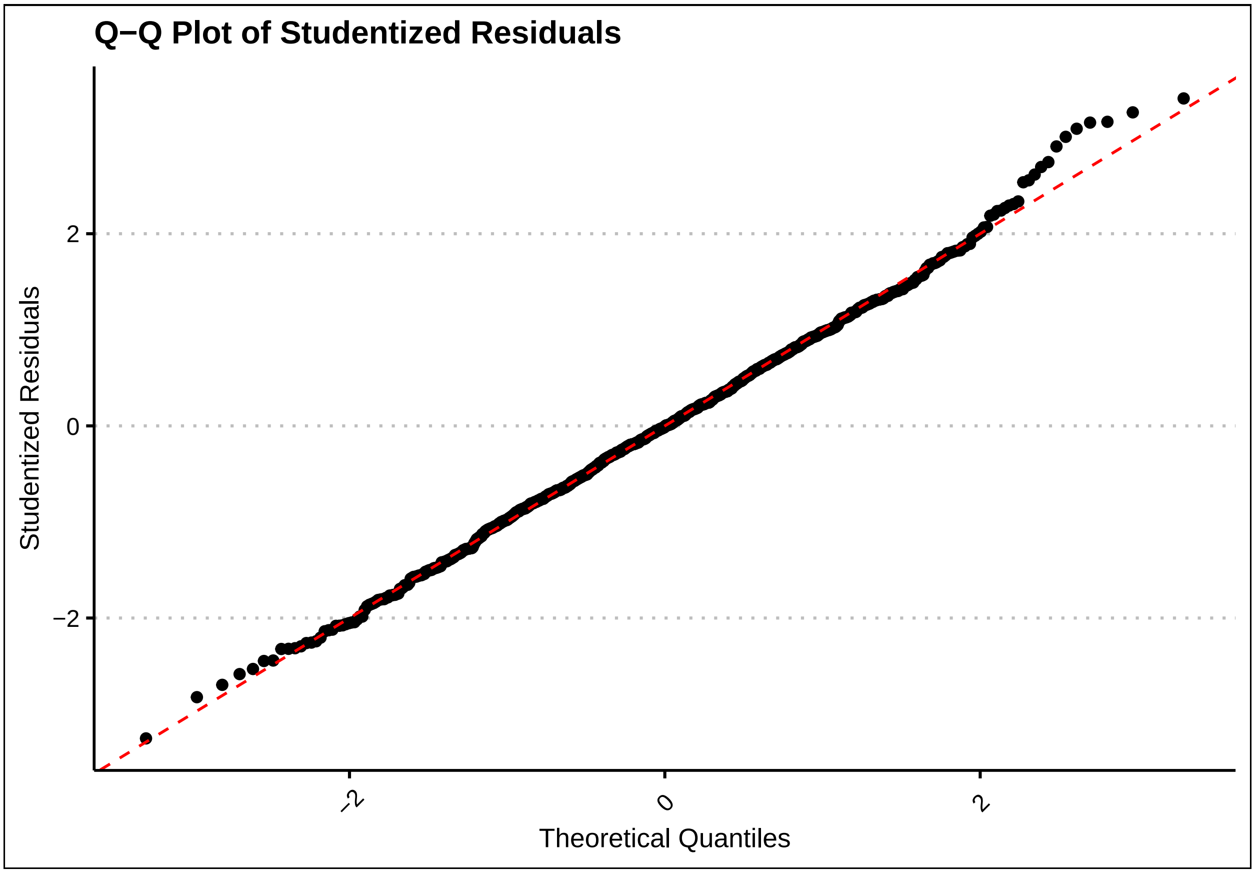
```r
#Scatter Plot of Studentized Residuals vs. Uterine Irritability
ggplot(bwt, aes(x = ui, y = sr)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Uterine Irritability",
       x = "Uterine Irritability",
       y = "Residuals")
```



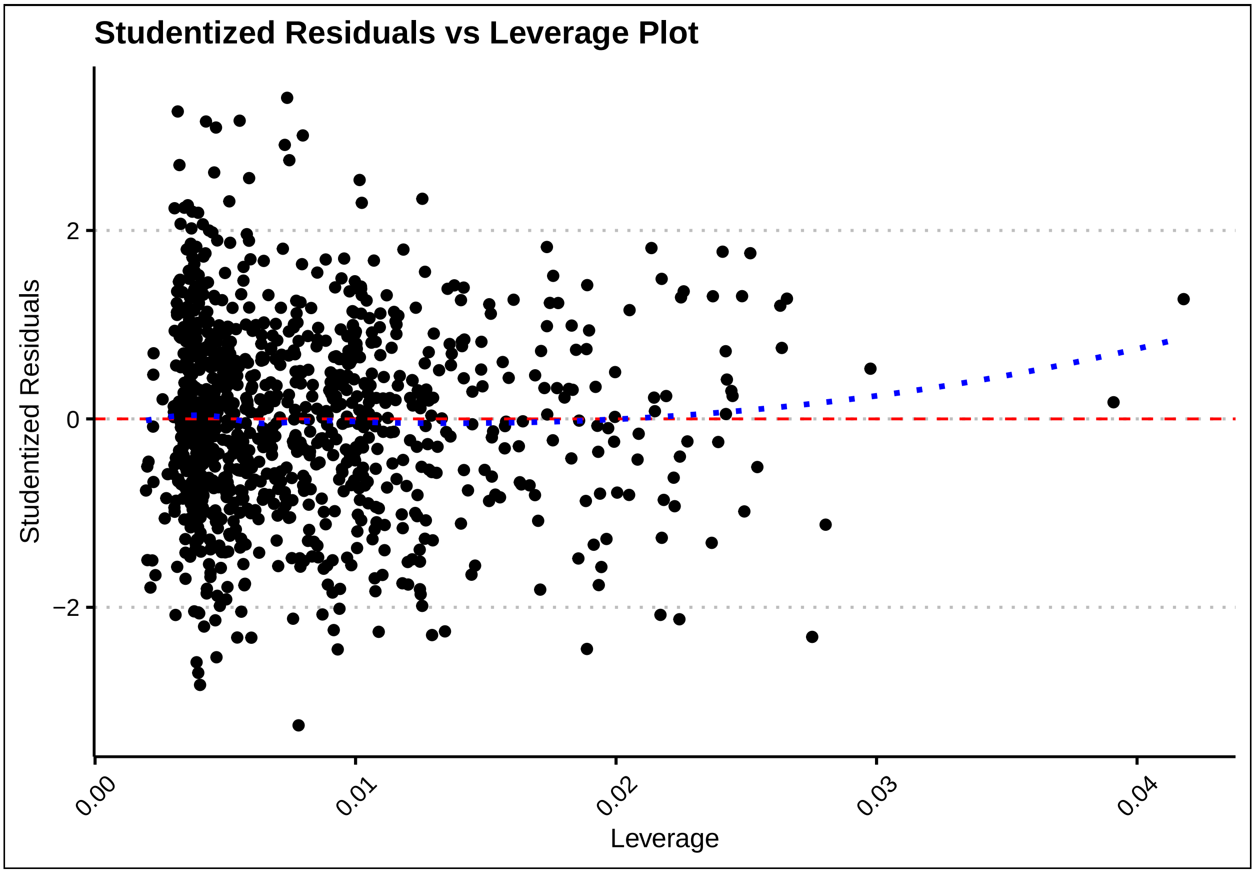
```
#Studentized Residuals vs. Fitted Value
ggplot(data = data.frame(Fitted = fitted(bwt_model), Residuals = sr),
       aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = c(-2, 2), linetype = "dashed", color = "red") +
  labs(title = "Studentized Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Studentized Residuals")
```



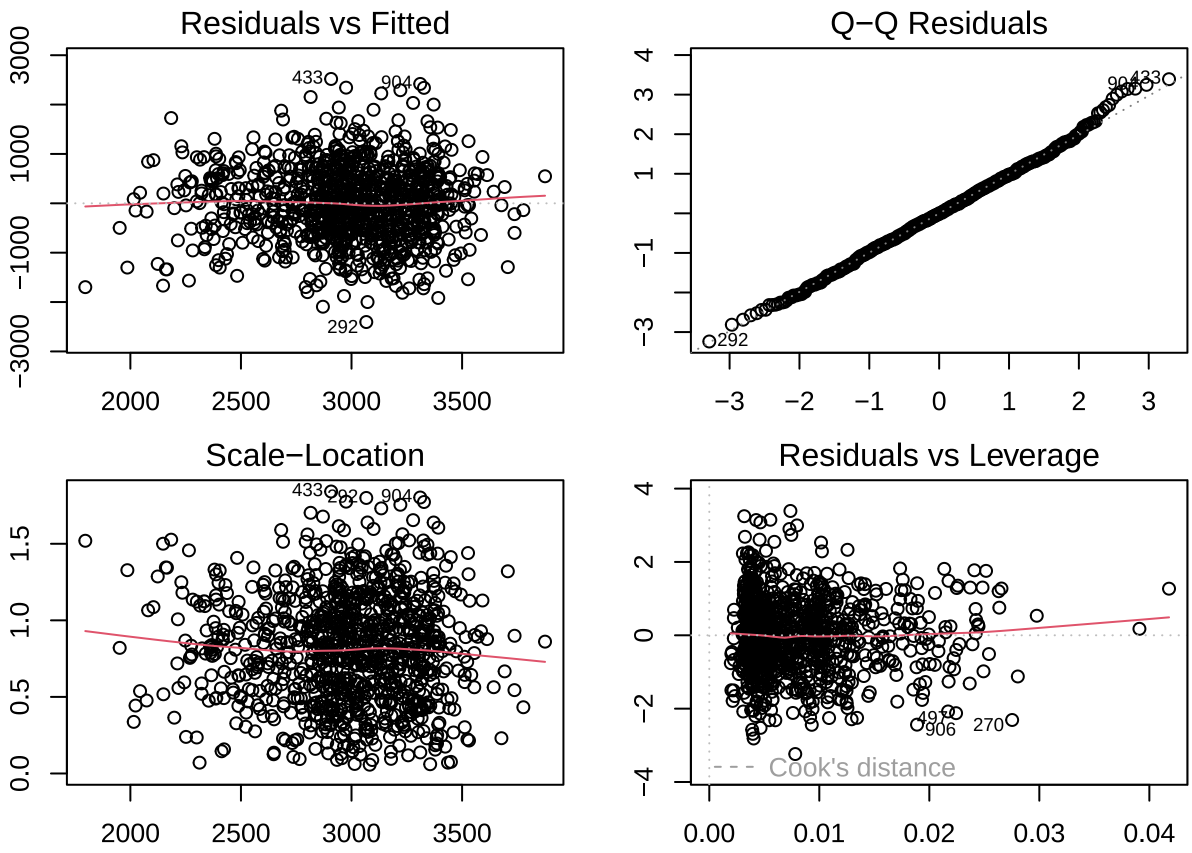
```
#Q-Q Plot for Studentized Residuals
qq_data = data.frame(
  Theoretical = qqnorm(sr, plot.it = FALSE)$x,
  Sample = qqnorm(sr, plot.it = FALSE)$y
)

ggplot(qq_data, aes(x = Theoretical, y = Sample)) +
  geom_point() +
  geom_abline(
    intercept = 0,
    slope = 1,
    color = "red",
    linetype = "dashed"
  ) +
  labs(title = "Q-Q Plot of Studentized Residuals", x = "Theoretical Quantiles", y = "Studentized Resid
```



```r
#Studentized Residuals vs. Leverage Plot
h = hatvalues(bwt_model)
leverage_data = data.frame(Leverage = h, StudentizedResiduals = sr)


ggplot(leverage_data, aes(x = Leverage, y = StudentizedResiduals)) +
  geom_point() +
  geom_hline(yintercept = 0,
             color = "red",
             linetype = "dashed") +
  geom_smooth(
    method = "loess",
    se = FALSE,
    color = "blue",
    linetype = "dotted"
  ) +
  labs(title = "Studentized Residuals vs Leverage Plot", x = "Leverage", y = "Studentized Residuals")
```



```r
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(bwt_model, cex.axis = 1, cex.lab = 1)
```



#Diagnostic Checks

```
df = dffits(bwt_model)
observations = names(sr)
n = nrow(bwt)
for (observation in observations){
  obs_sr = sr[observation]
  p_value = 2*pt(obs_sr, (n-8))
  if(p_value < 0.05){
    print(paste(observation, "is an outlier"))
  }
}
```

```
## [1] "28 is an outlier"
## [1] "68 is an outlier"
## [1] "106 is an outlier"
## [1] "161 is an outlier"
## [1] "201 is an outlier"
## [1] "226 is an outlier"
## [1] "232 is an outlier"
## [1] "270 is an outlier"
## [1] "274 is an outlier"
## [1] "292 is an outlier"
## [1] "299 is an outlier"
## [1] "364 is an outlier"
## [1] "382 is an outlier"
## [1] "390 is an outlier"
## [1] "411 is an outlier"
## [1] "436 is an outlier"
## [1] "457 is an outlier"
## [1] "497 is an outlier"
## [1] "499 is an outlier"
## [1] "513 is an outlier"
## [1] "519 is an outlier"
## [1] "549 is an outlier"
## [1] "608 is an outlier"
## [1] "710 is an outlier"
## [1] "887 is an outlier"
## [1] "892 is an outlier"
## [1] "906 is an outlier"
## [1] "946 is an outlier"
```

```
p = length(coef(bwt_model))
avgLeverage = 2*p/n
highLeverage = which(h > avgLeverage)
influential = which(df > 2*sqrt(p/n))
print(paste("High Leverage Point:", toString(highLeverage)))
```

```
## [1] "High Leverage Point: 3, 9, 13, 29, 33, 48, 60, 64, 69, 78, 82, 116, 121, 146, 148, 163, 178, 183
```

```
print(paste("Influential:", toString(influential)))
```

```
## [1] "Influential: 48, 60, 62, 69, 148, 183, 316, 383, 433, 536, 550, 556, 632, 693, 704, 706, 709, 82
```

As results shown, we will remove all the outliers and influential points.Then do the linear regression again to see whether the model improved.

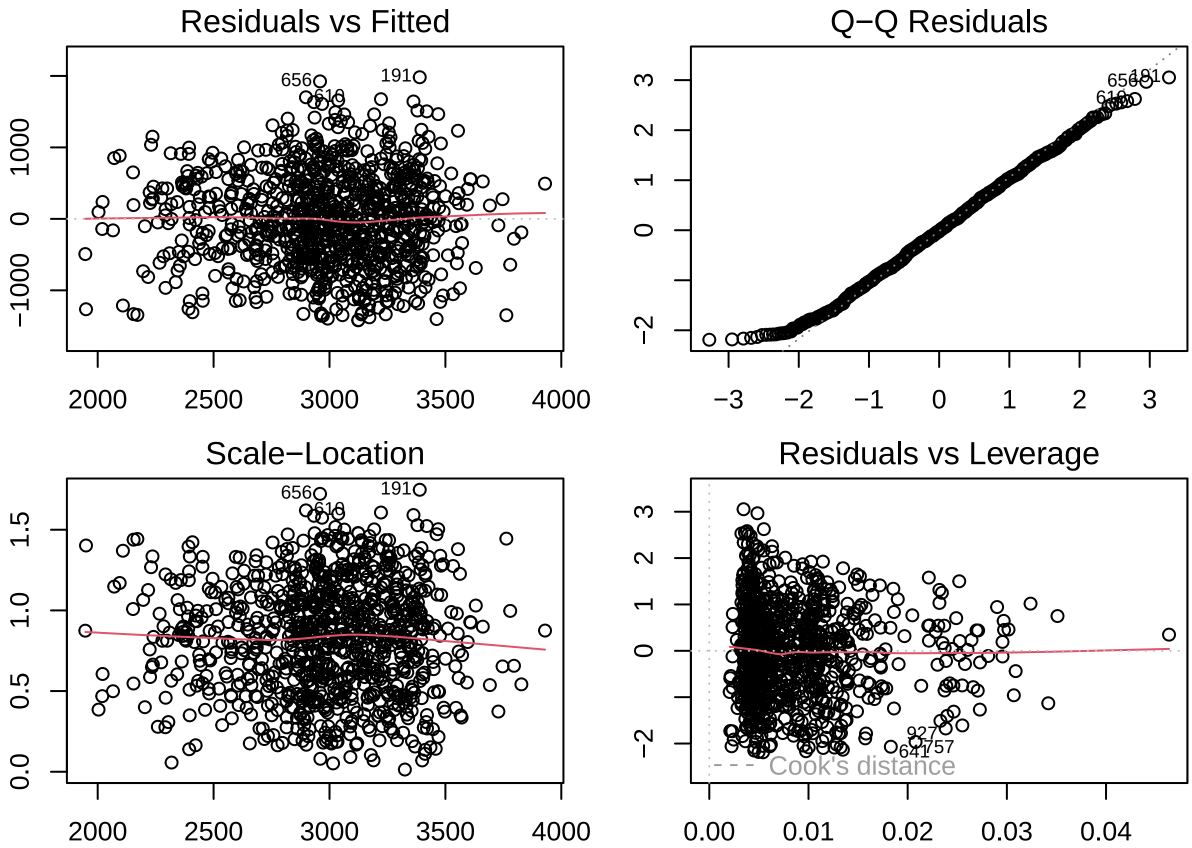
#Refit the Model

```
bwt_new = bwt[-c(28, 48, 60, 62, 68, 69, 106, 148, 161, 183, 201, 226, 232, 270, 274, 292, 299, 316, 364

bwt_model2 = lm(bwt ~ age + lwt + race + smoke + ptl + ht + ui, data = bwt_new)
summary(bwt_model2)
```

```
##
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui,
##     data = bwt_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1419.32  -480.62   -13.09   457.75  1981.77
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2749.3465   149.6874  18.367  < 2e-16 ***
## age           -3.0358     4.0035  -0.758    0.448
## lwt            5.0675     0.7088   7.150 1.75e-12 ***
## race        -141.4258    24.2364  -5.835 7.39e-09 ***
## smoke       -275.6111    45.0514  -6.118 1.39e-09 ***
## ptl           73.6399    58.0798   1.268    0.205
## ht          -457.4925    95.8887  -4.771 2.12e-06 ***
## ui          -566.2879    61.9126  -9.147  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 650.1 on 937 degrees of freedom
## Multiple R-squared:  0.2022, Adjusted R-squared:  0.1962
## F-statistic: 33.93 on 7 and 937 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(bwt_model2, cex.axis = 1, cex.lab = 1)
```



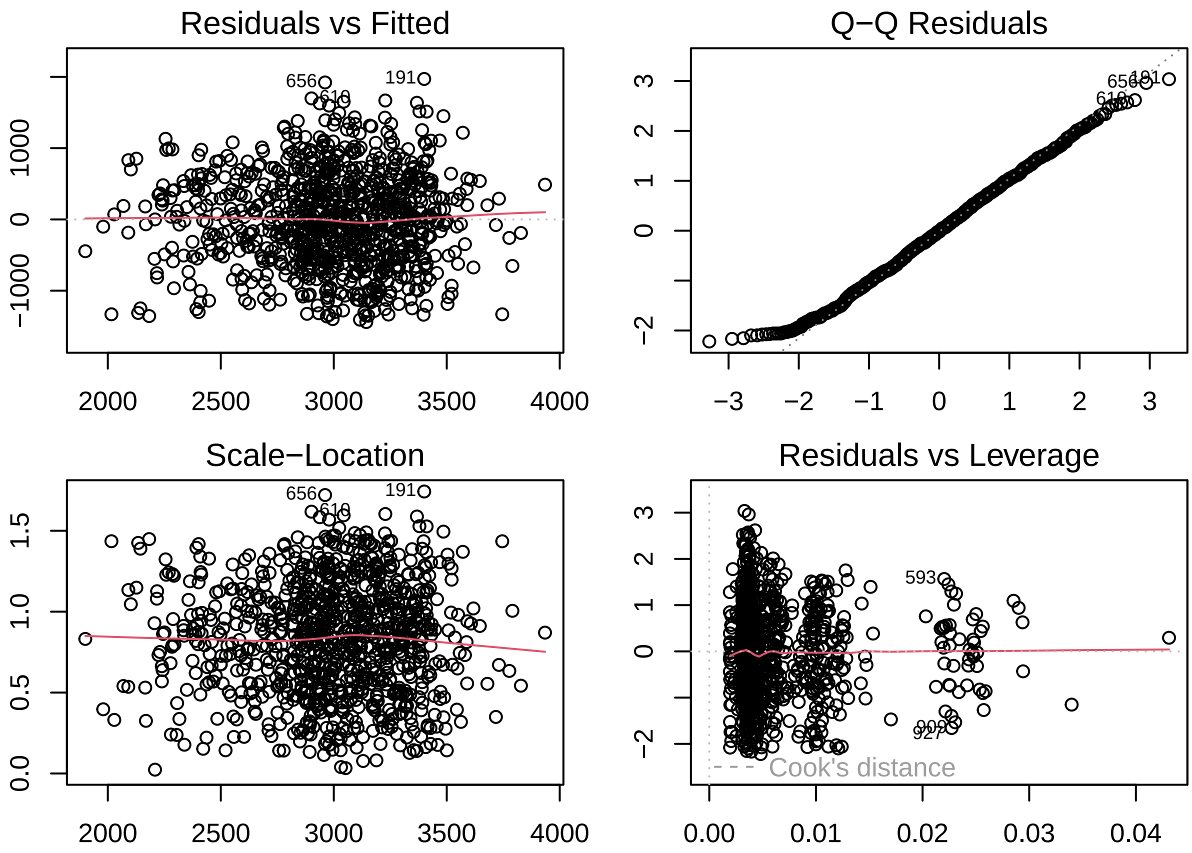
```r
olsrr::ols_step_best_subset(bwt_model2)
```

```
##            Best Subsets Regression
## -------------------------------------------
## Model Index    Predictors
## -------------------------------------------
##      1         ui
##      2         lwt ui
##      3         lwt smoke ui
##      4         lwt race smoke ui
##      5         lwt race smoke ht ui
##      6         lwt race smoke ptl ht ui
##      7         age lwt race smoke ptl ht ui
## -------------------------------------------
##
##                                                                 Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC           SBIC          SBC
## ----------------------------------------------------------------------------------------------------
##   1        0.0830      0.0820      0.0791    136.0585    15053.2647    12370.9504    15067.8183    45
##   2        0.1310      0.1292      0.1257     81.6224    15004.4039    12322.1545    15023.8086    43
##   3        0.1529      0.1502       0.146     57.8495    14982.2341    12300.0344    15006.4900    42
##   4        0.1813      0.1778      0.1729     26.5286    14952.0407    12270.0743    14981.1478    40
##   5        0.2005      0.1963      0.1911      5.9890    14931.6247    12249.9076    14965.5830    39
##   6        0.2017      0.1966      0.1906      6.5750    14932.2006    12250.5175    14971.0101    39
##   7        0.2022      0.1962      0.1894      8.0000    14933.6209    12251.9635    14977.2815    39
## ----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
#Reduced model: bwt ~ lwt + race + smoke + ht + ui
bwt_model3 = lm(bwt ~ lwt + race + smoke + ht + ui, data = bwt_new)
summary(bwt_model3)
```

```
##
## Call:
## lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = bwt_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1439.62  -476.60   -12.95   451.58  1970.75
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2705.7584   124.4408  21.743  < 2e-16 ***
## lwt            4.9160     0.6982   7.041 3.67e-12 ***
## race        -138.5982    23.9583  -5.785 9.87e-09 ***
## smoke       -268.0582    44.6746  -6.000 2.81e-09 ***
## ht          -455.1259    95.8639  -4.748 2.38e-06 ***
## ui          -556.5304    61.5186  -9.047  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 650.1 on 939 degrees of freedom
## Multiple R-squared:  0.2005, Adjusted R-squared:  0.1963
## F-statistic:  47.1 on 5 and 939 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(bwt_model3, cex.axis = 1, cex.lab = 1)
```

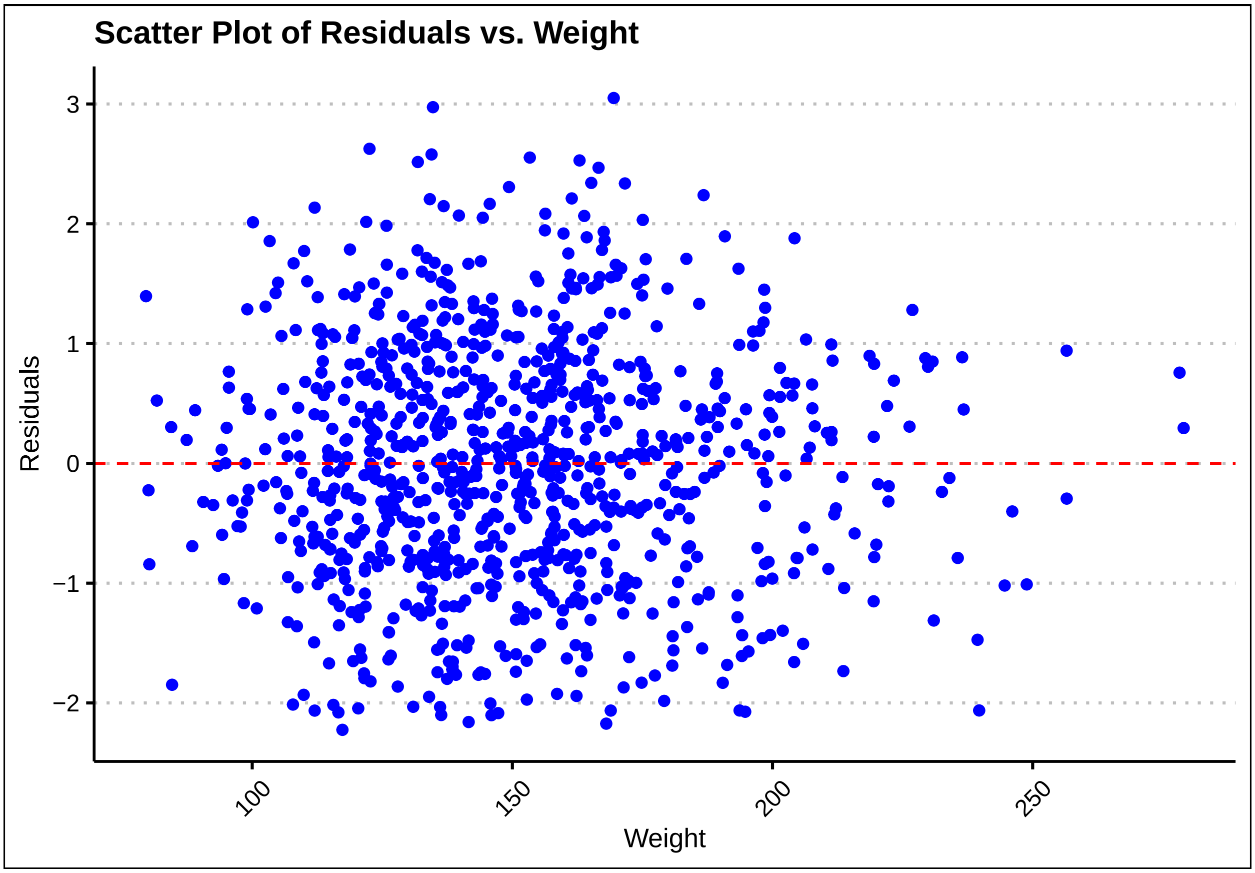

Residuals vs Fitted

Q−Q Residuals

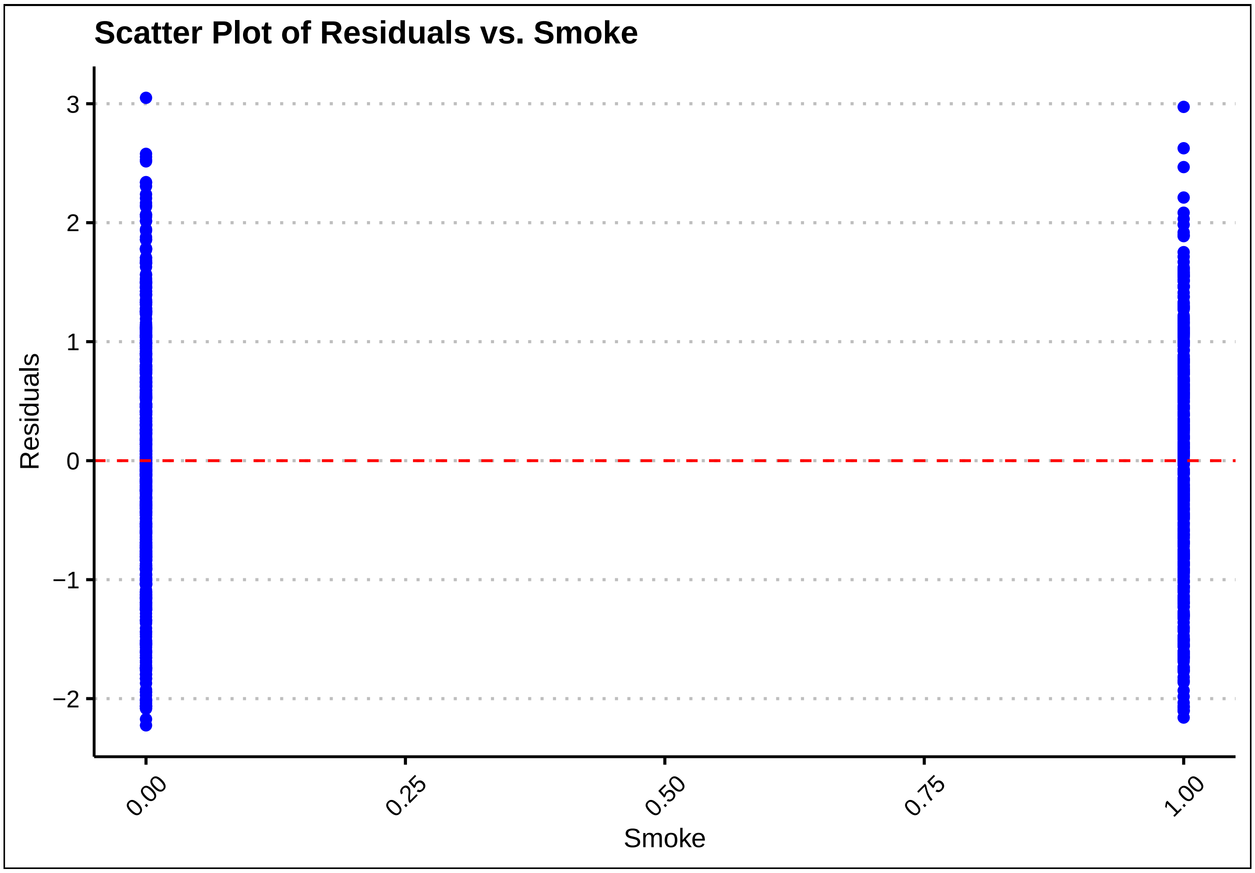
Scale−Location

Residuals vs Leverage

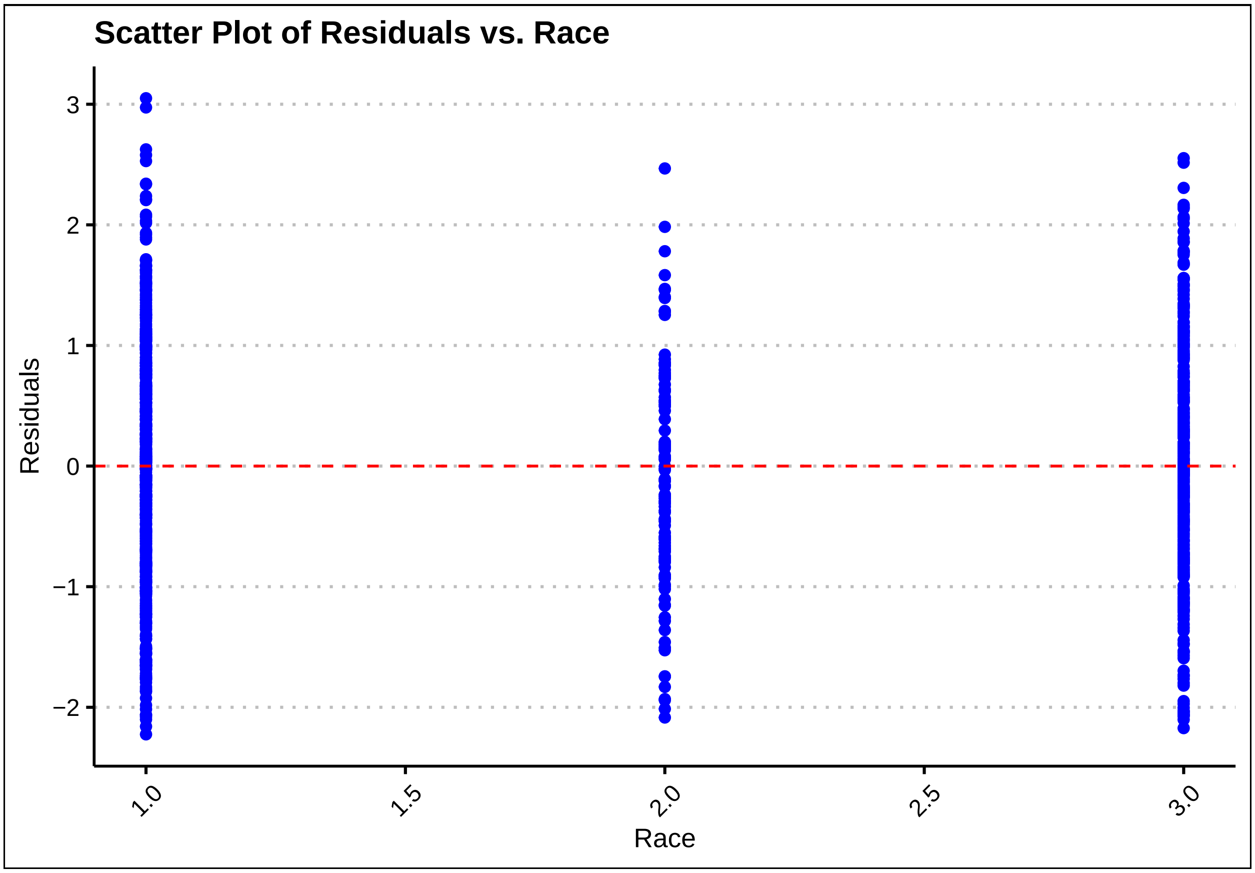
```
sr2 = rstudent(bwt_model3)

#Scatter Plot of Studentized Residuals vs. Weight
ggplot(bwt_new, aes(x = lwt, y = sr2)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Weight",
       x = "Weight",
       y = "Residuals")
```



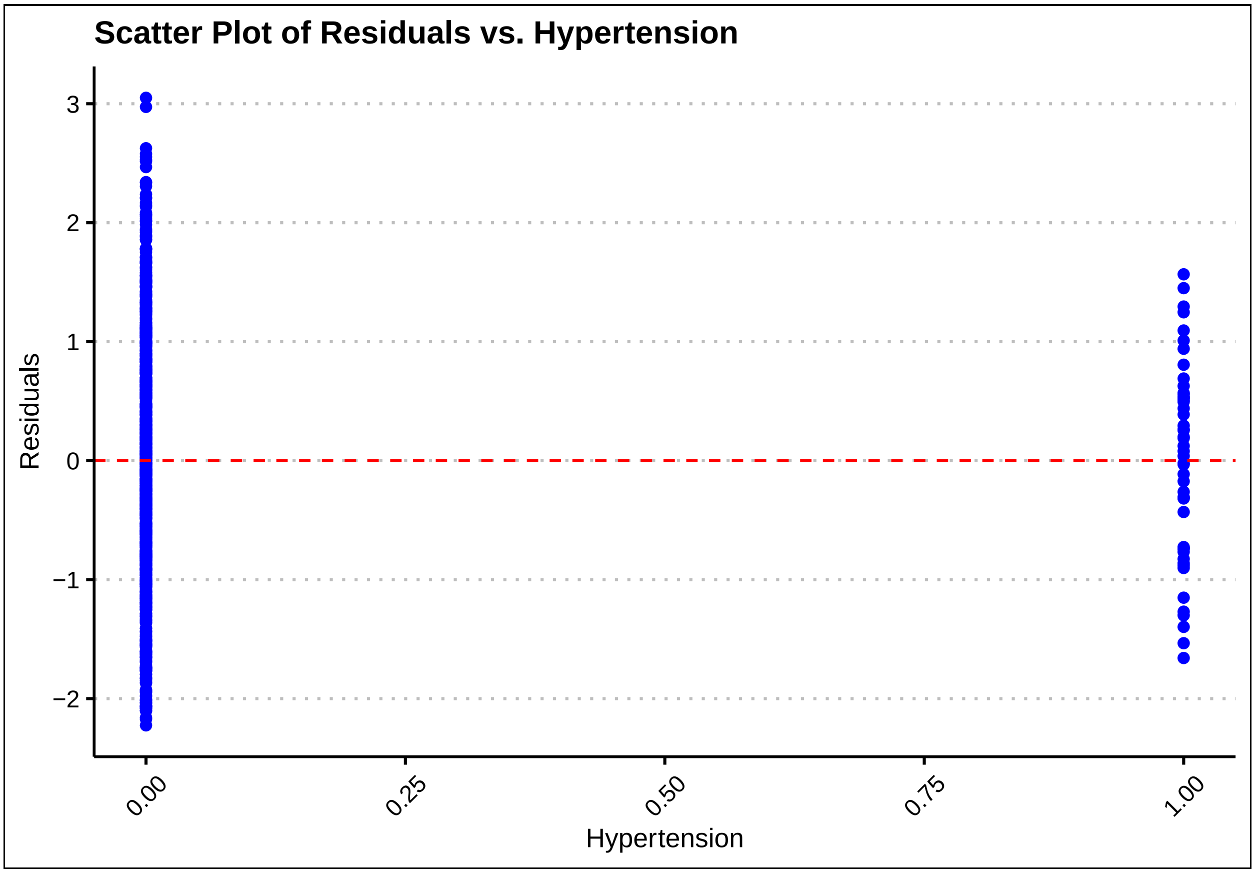
```
#Scatter Plot of Studentized Residuals vs. Smoke
ggplot(bwt_new, aes(x = smoke, y = sr2)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Smoke",
       x = "Smoke",
       y = "Residuals")
```



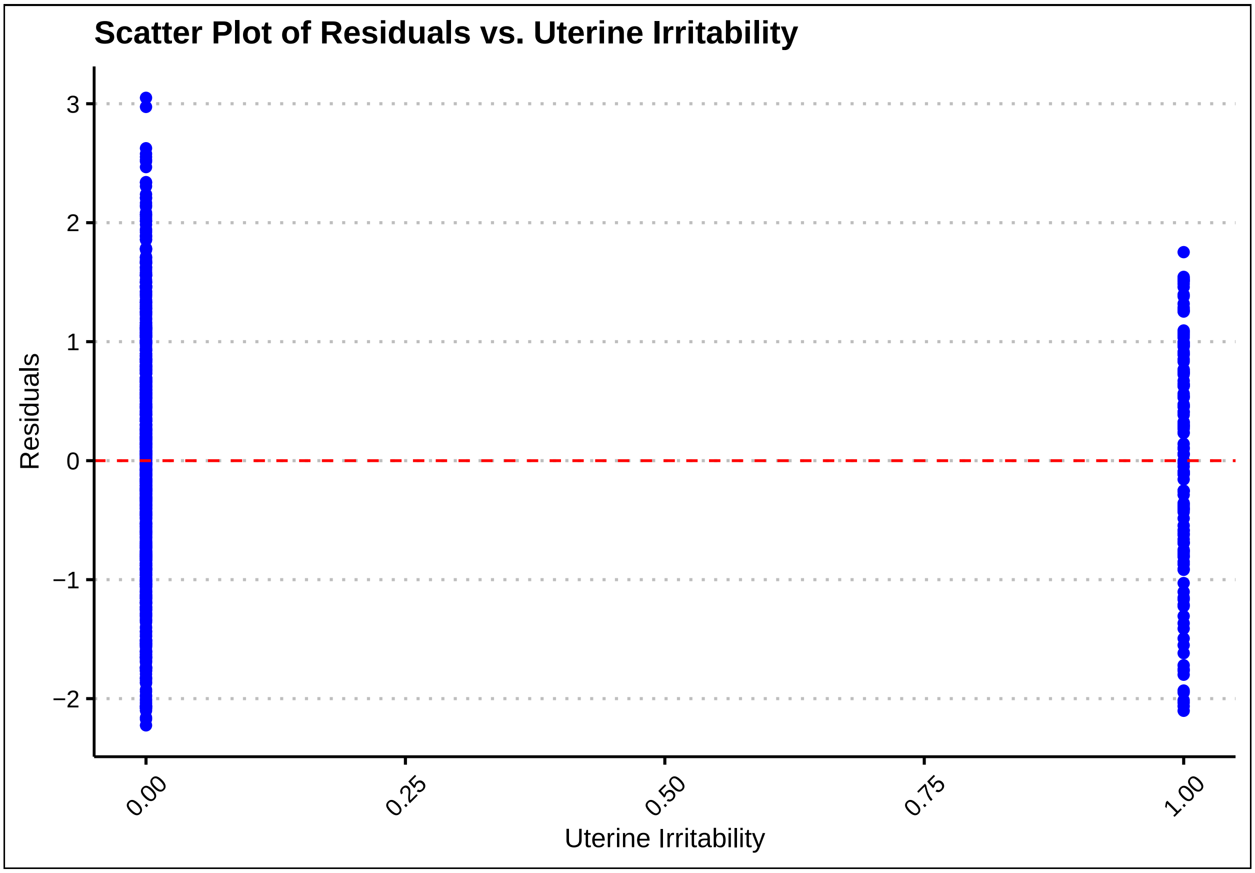
```
#Scatter Plot of Studentized Residuals vs. Race
ggplot(bwt_new, aes(x = race, y = sr2)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Race",
       x = "Race",
       y = "Residuals")
```



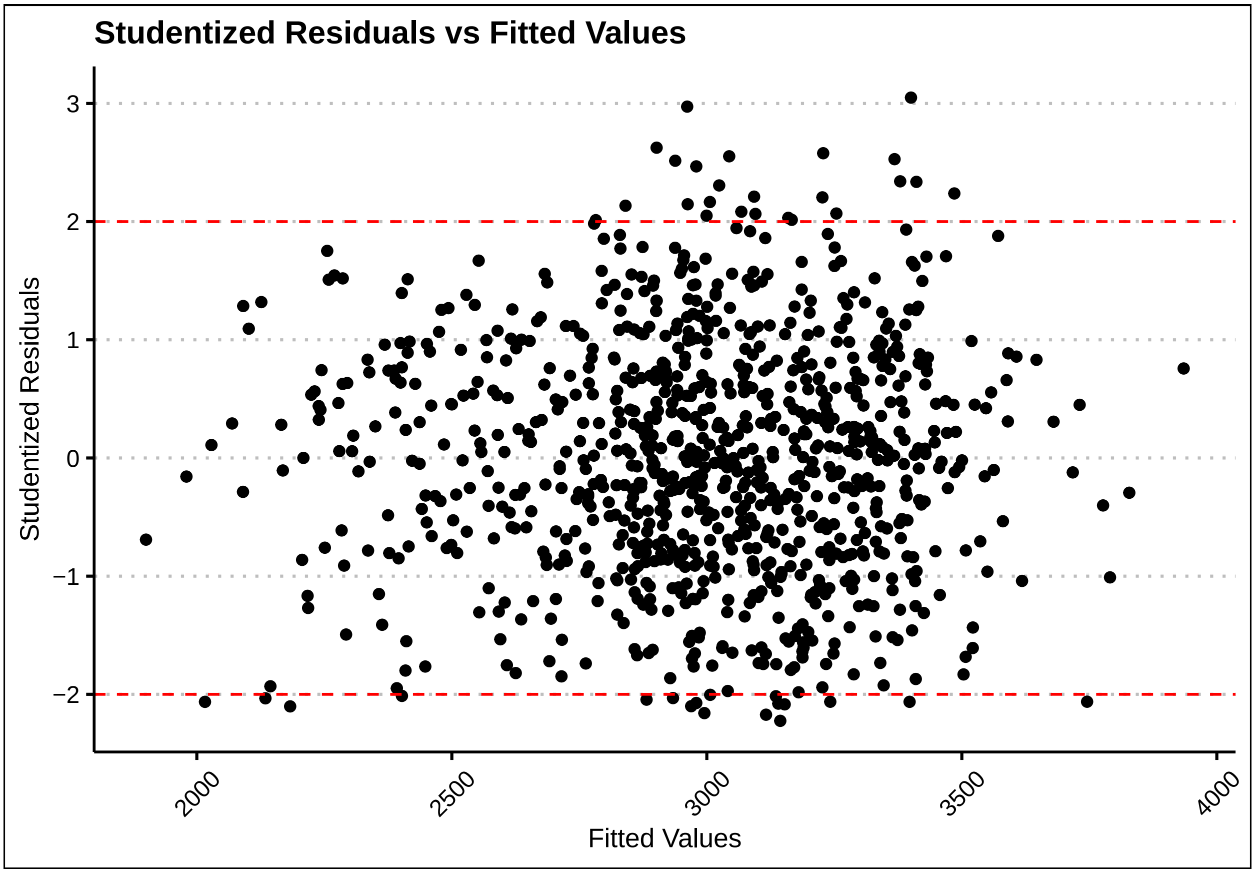
```r
#Scatter Plot of Studentized Residuals vs. Hypertension
ggplot(bwt_new, aes(x = ht, y = sr2)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Hypertension",
       x = "Hypertension",
       y = "Residuals")
```



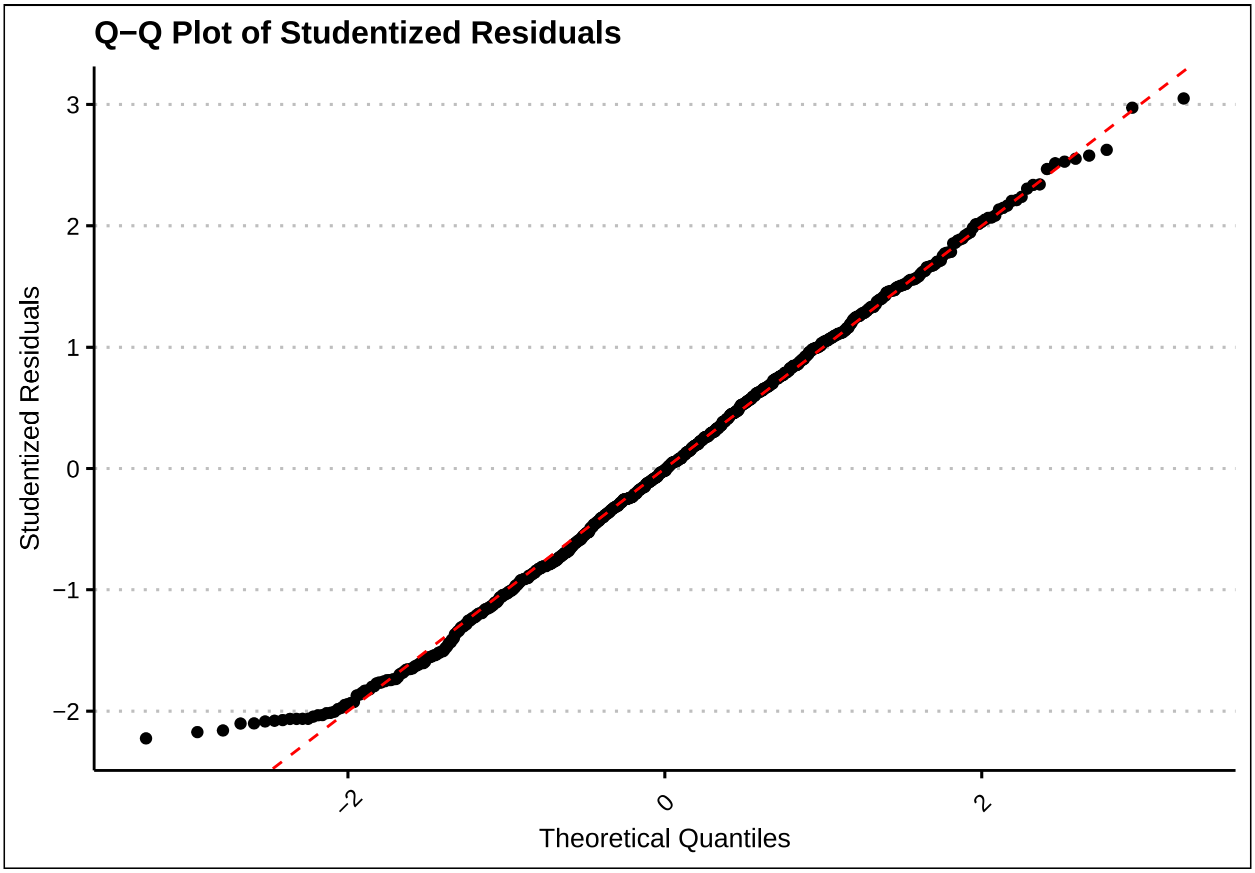
```r
#Scatter Plot of Studentized Residuals vs. Uterine Irritability
ggplot(bwt_new, aes(x = ui, y = sr2)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Uterine Irritability",
       x = "Uterine Irritability",
       y = "Residuals")
```



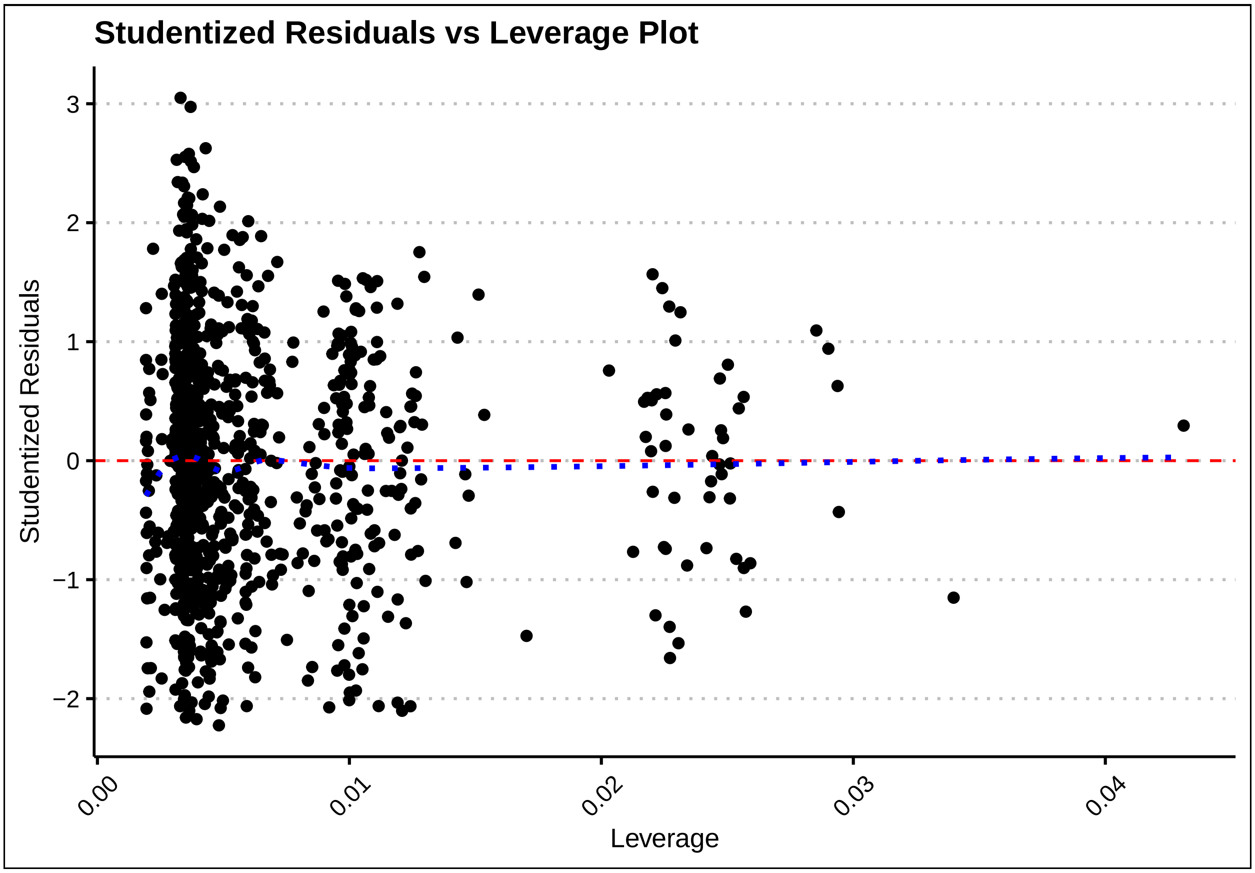
```r
#Studentized Residuals vs. Fitted Value
ggplot(data = data.frame(Fitted = fitted(bwt_model3), Residuals = sr2),
       aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = c(-2, 2), linetype = "dashed", color = "red") +
  labs(title = "Studentized Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Studentized Residuals")
```



```r
#Q-Q Plot for Studentized Residuals
qq_data2 = data.frame(
  Theoretical = qqnorm(sr2, plot.it = FALSE)$x,
  Sample = qqnorm(sr2, plot.it = FALSE)$y
)

ggplot(qq_data2, aes(x = Theoretical, y = Sample)) +
  geom_point() +
  geom_abline(
    intercept = 0,
    slope = 1,
    color = "red",
    linetype = "dashed"
  ) +
  labs(title = "Q-Q Plot of Studentized Residuals", x = "Theoretical Quantiles", y = "Studentized Resid
```

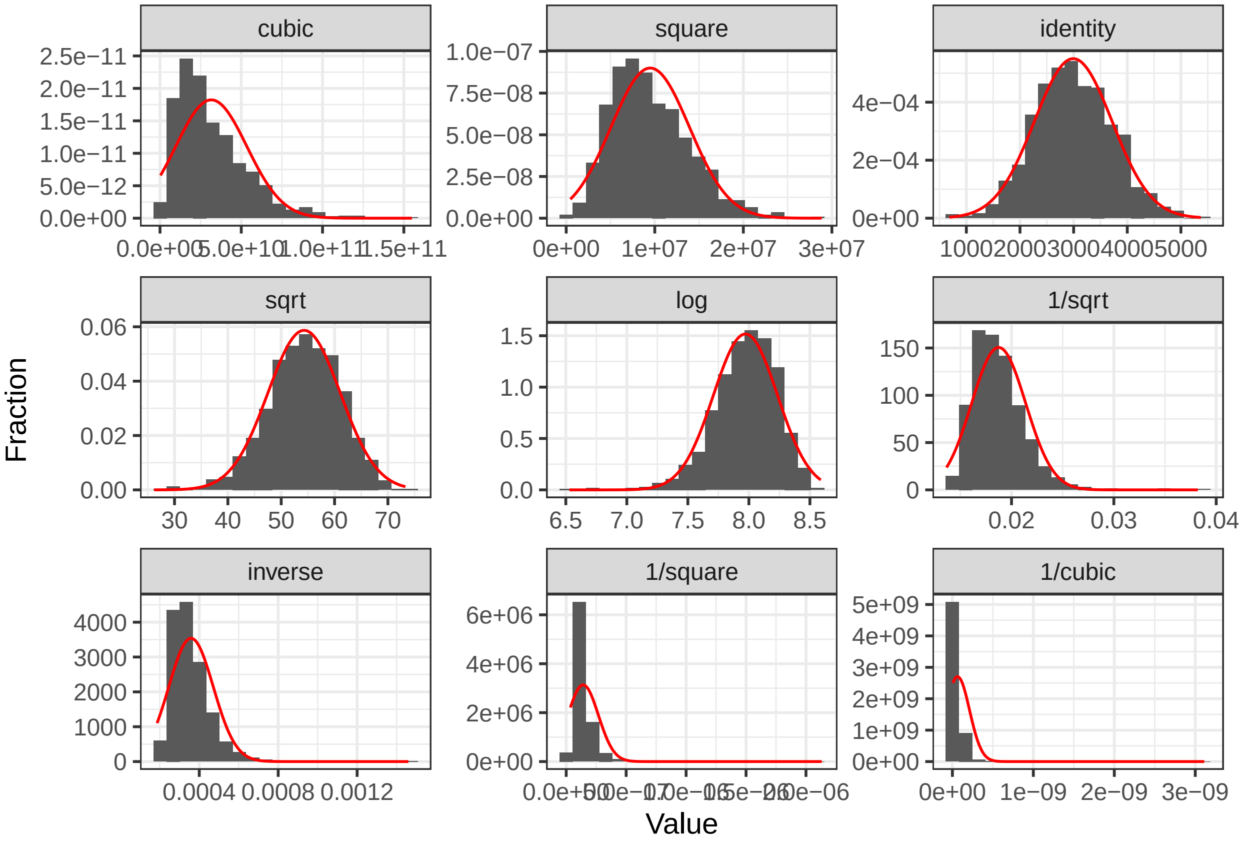


```
#Studentized Residuals vs. Leverage Plot
h2 = hatvalues(bwt_model3)
leverage_data2 = data.frame(Leverage = h2, StudentizedResiduals = sr2)


ggplot(leverage_data2, aes(x = Leverage, y = StudentizedResiduals)) +
  geom_point() +
  geom_hline(yintercept = 0,
             color = "red",
             linetype = "dashed") +
  geom_smooth(
    method = "loess",
    se = FALSE,
    color = "blue",
    linetype = "dotted"
  ) +
  labs(title = "Studentized Residuals vs Leverage Plot", x = "Leverage", y = "Studentized Residuals")
```



#Variable Transformation

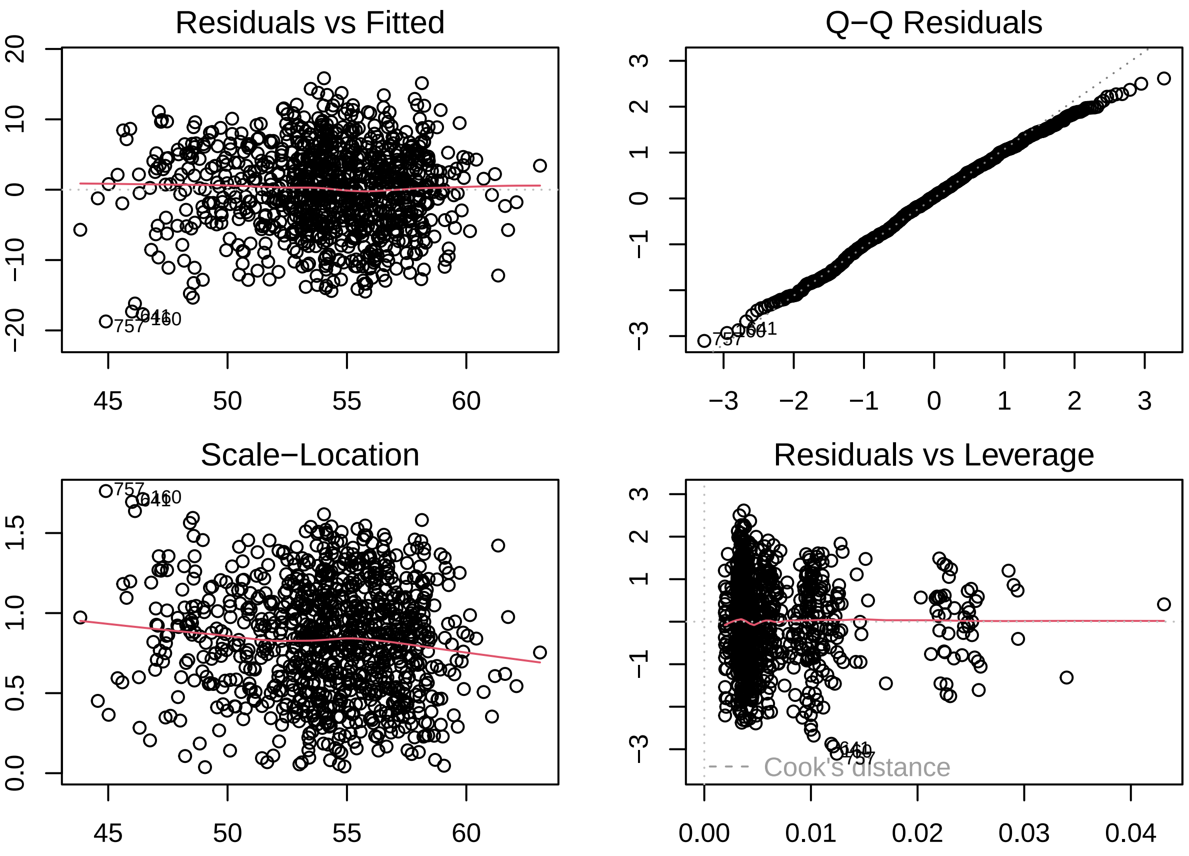
```
#Define whether the independent variable needs transformation
gladder(bwt_new$bwt)
```



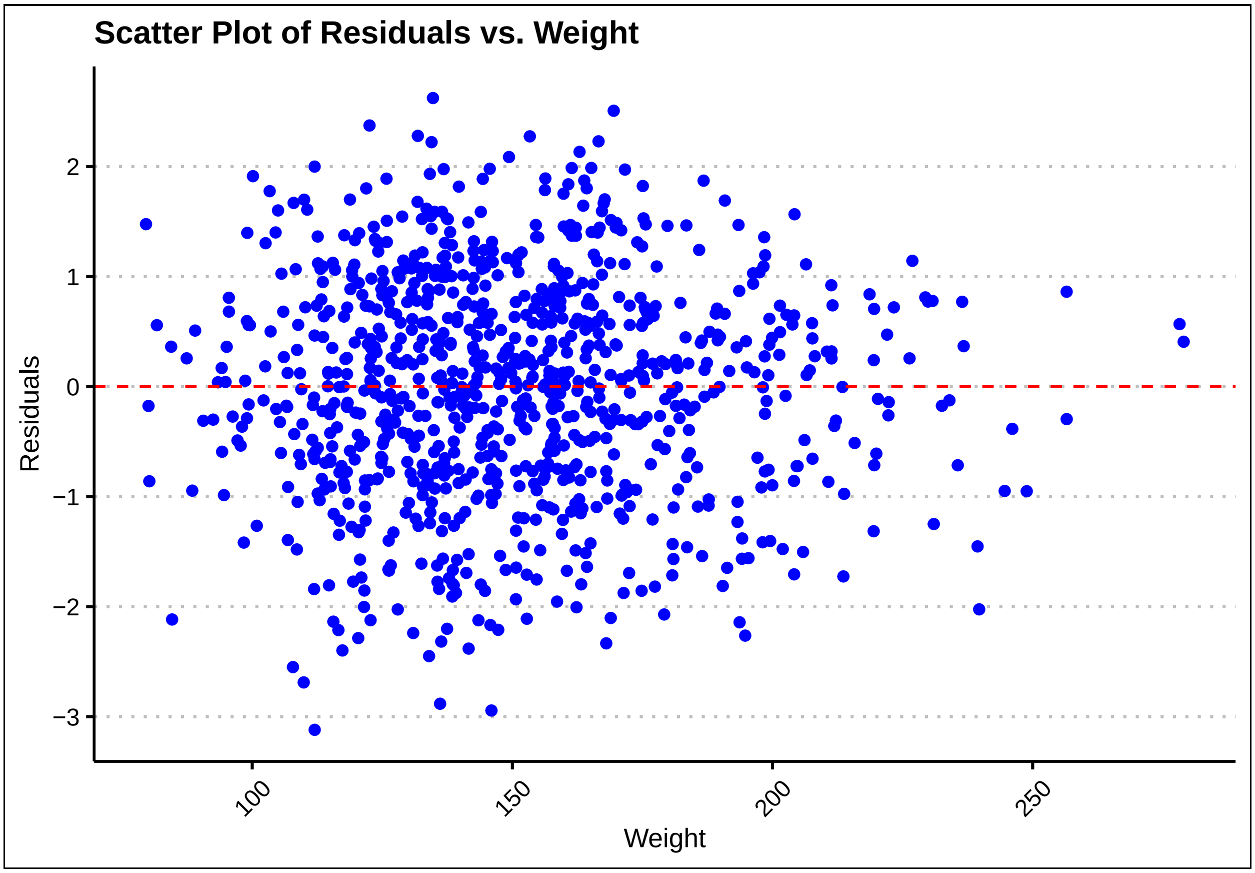
```r
#Based on the plot above, we would choose sqrt transformation for bwt
bwt_model4 = lm(sqrt(bwt) ~ lwt + race + smoke + ht + ui, data = bwt_new)
summary(bwt_model4)
```

```
##
## Call:
## lm(formula = sqrt(bwt) ~ lwt + race + smoke + ht + ui, data = bwt_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.7283  -4.2849   0.2529   4.3780  15.8393
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.741819   1.161544  44.546  < 2e-16 ***
## lwt          0.045504   0.006517   6.983 5.48e-12 ***
## race        -1.314708   0.223629  -5.879 5.73e-09 ***
## smoke       -2.519664   0.416997  -6.042 2.19e-09 ***
## ht          -4.204080   0.894804  -4.698 3.01e-06 ***
## ui          -5.473923   0.574221  -9.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.068 on 939 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.2031
## F-statistic: 49.12 on 5 and 939 DF,  p-value: < 2.2e-16
```

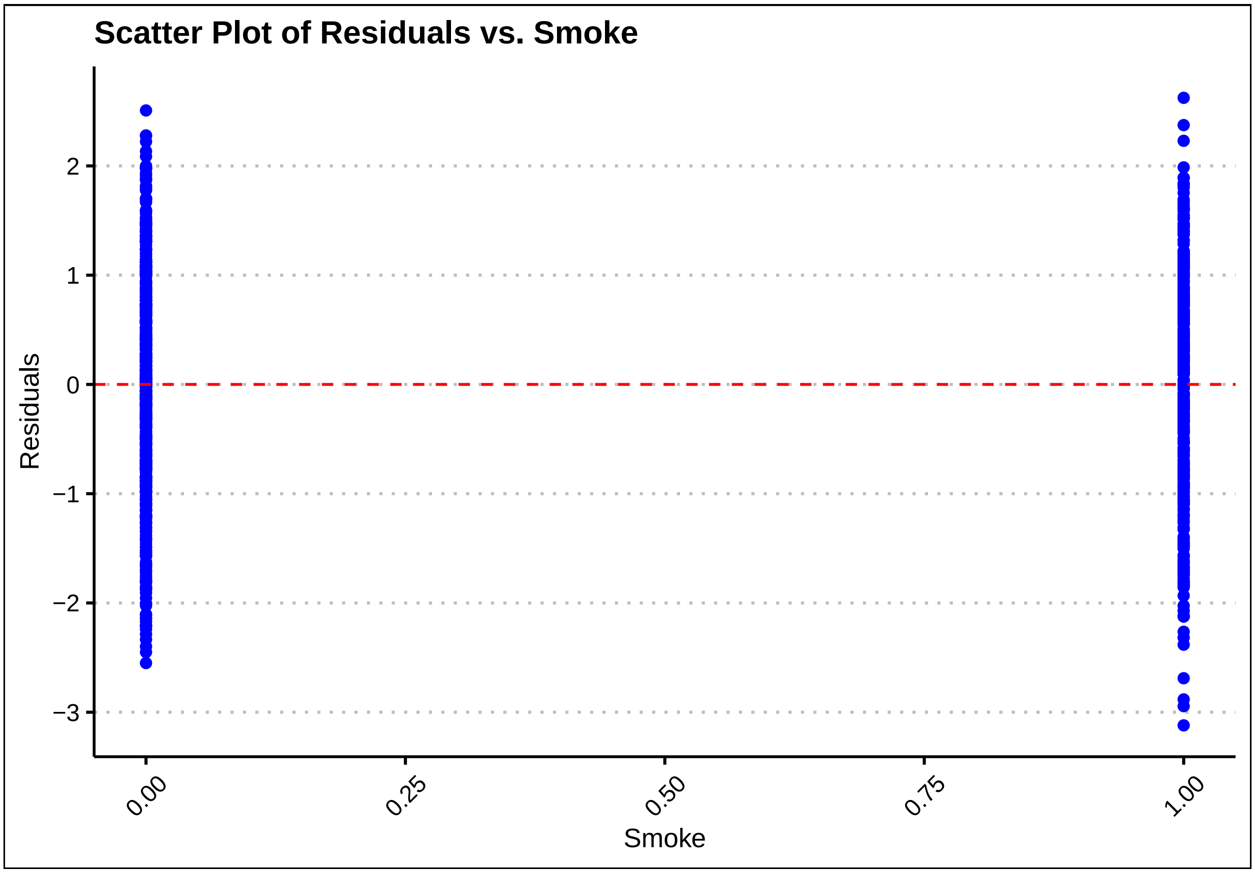
```r
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(bwt_model4, cex.axis = 1, cex.lab = 1)
```



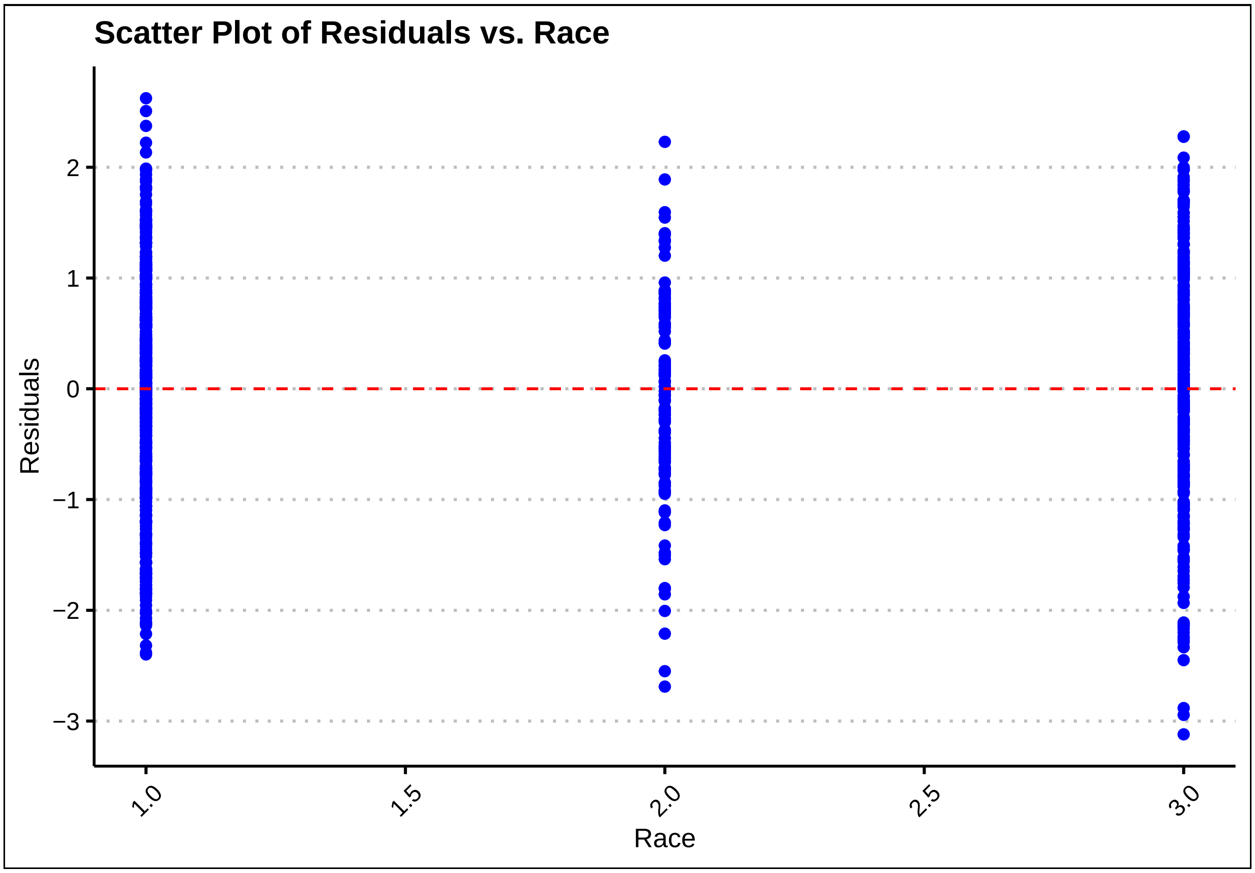
```r
sr3 = rstudent(bwt_model4)

#Scatter Plot of Studentized Residuals vs. Weight
ggplot(bwt_new, aes(x = lwt, y = sr3)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Weight",
       x = "Weight",
       y = "Residuals")
```



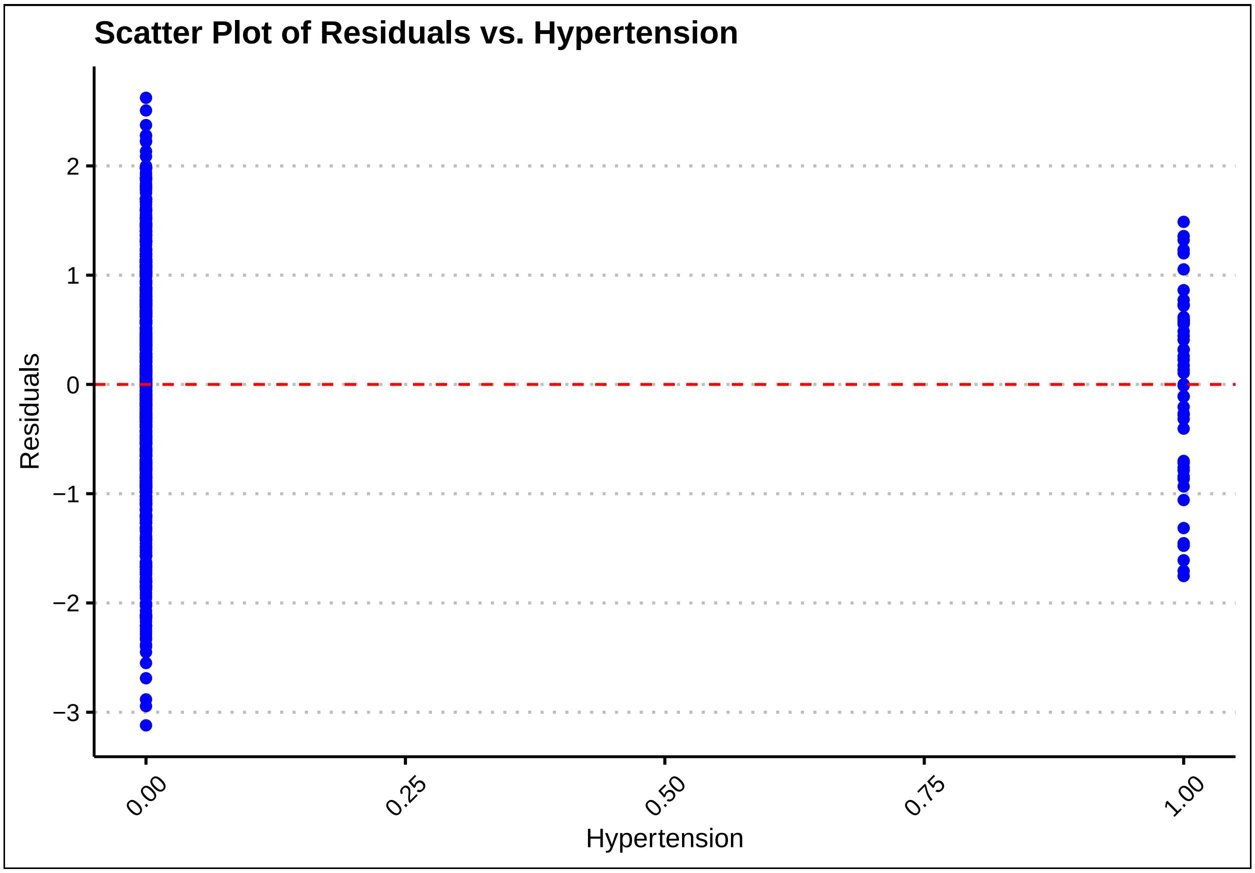
```
#Scatter Plot of Studentized Residuals vs. Smoke
ggplot(bwt_new, aes(x = smoke, y = sr3)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Smoke",
       x = "Smoke",
       y = "Residuals")
```



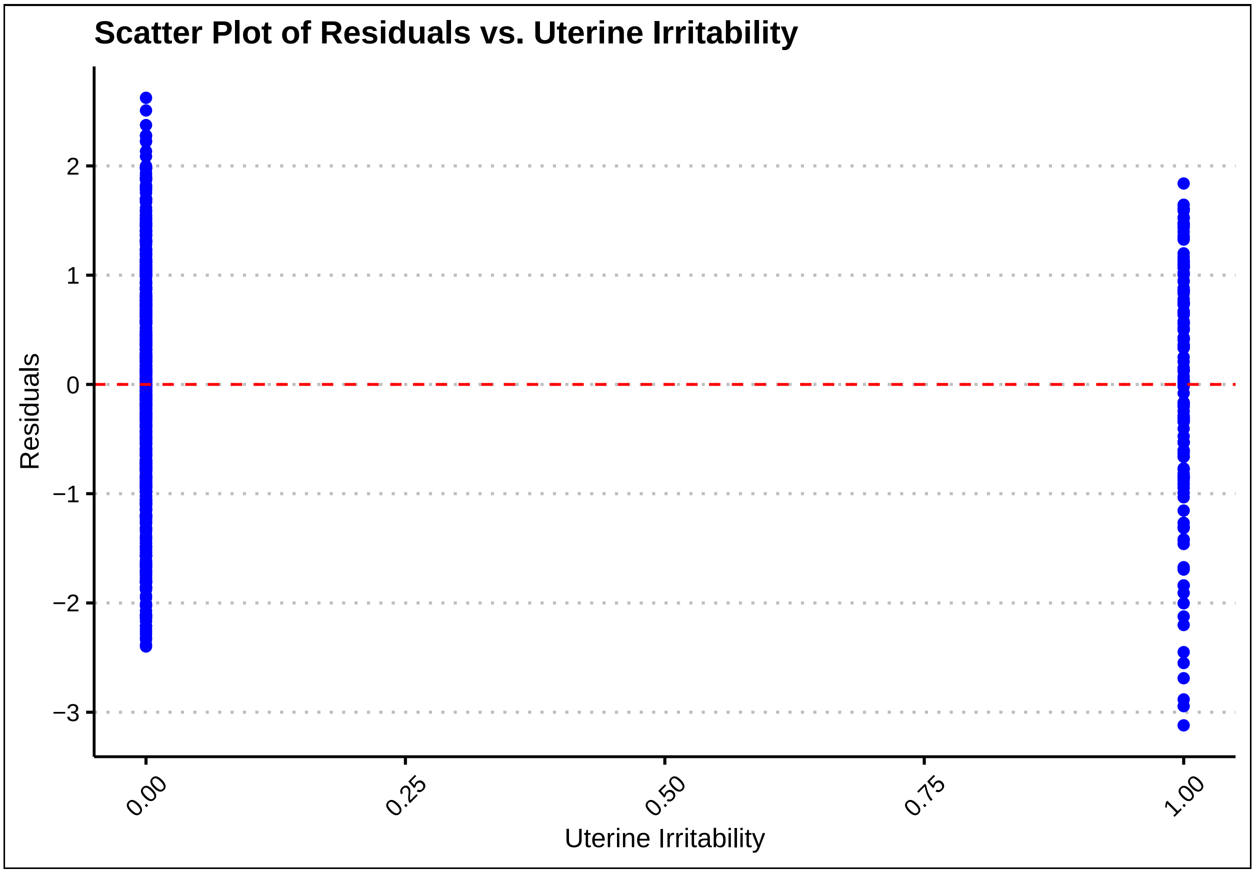
```
#Scatter Plot of Studentized Residuals vs. Race
ggplot(bwt_new, aes(x = race, y = sr3)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Race",
       x = "Race",
       y = "Residuals")
```



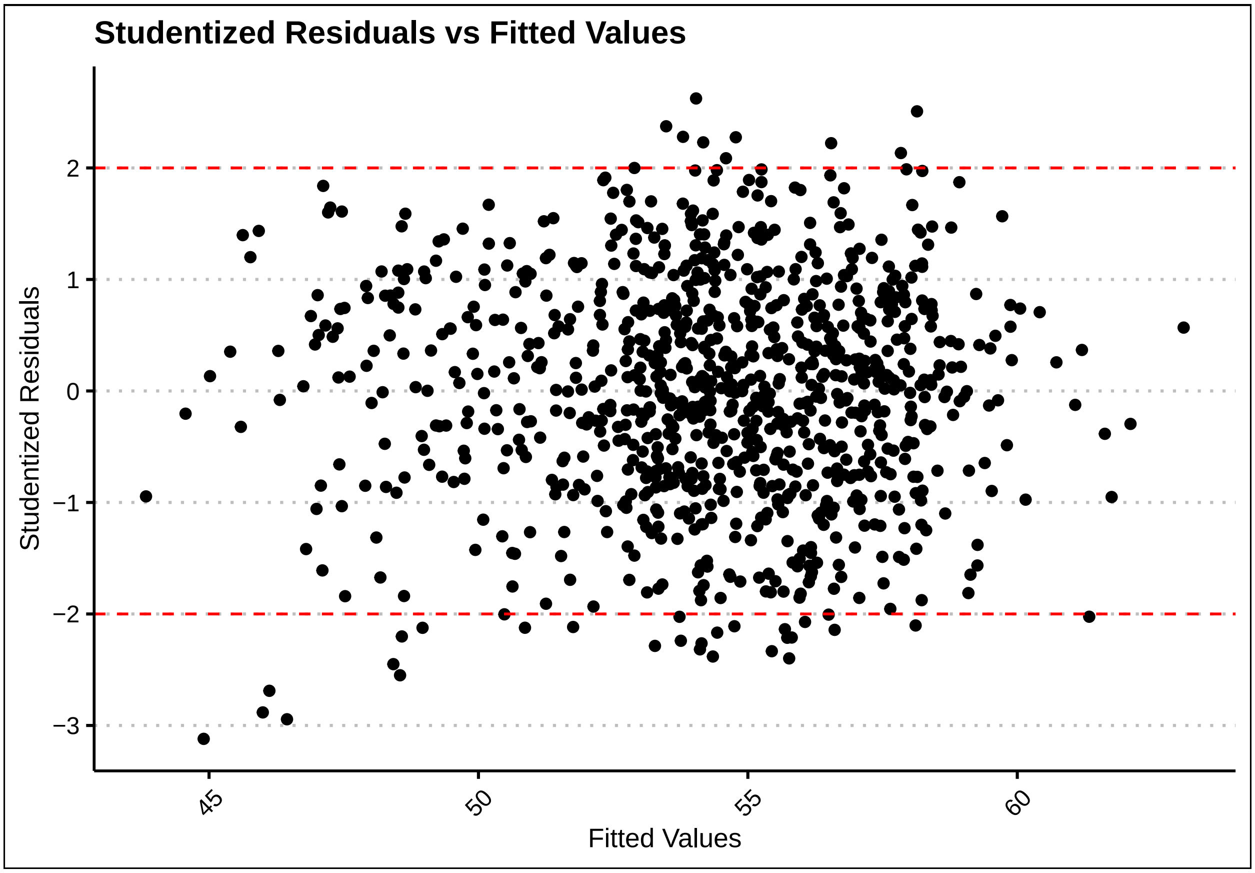
```r
#Scatter Plot of Studentized Residuals vs. Hypertension
ggplot(bwt_new, aes(x = ht, y = sr3)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Hypertension",
       x = "Hypertension",
       y = "Residuals")
```



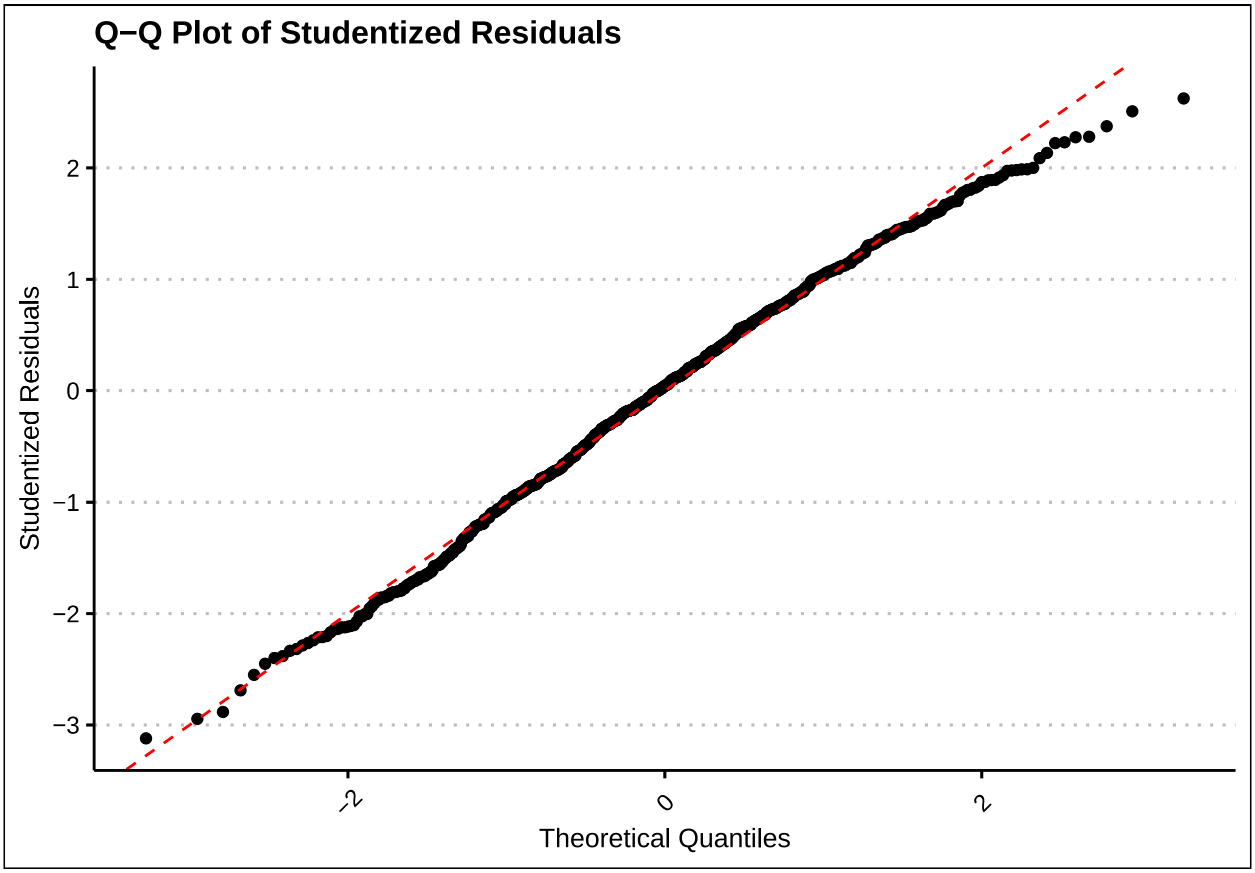
```
#Scatter Plot of Studentized Residuals vs. Uterine Irritability
ggplot(bwt_new, aes(x = ui, y = sr3)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Scatter Plot of Residuals vs. Uterine Irritability",
       x = "Uterine Irritability",
       y = "Residuals")
```



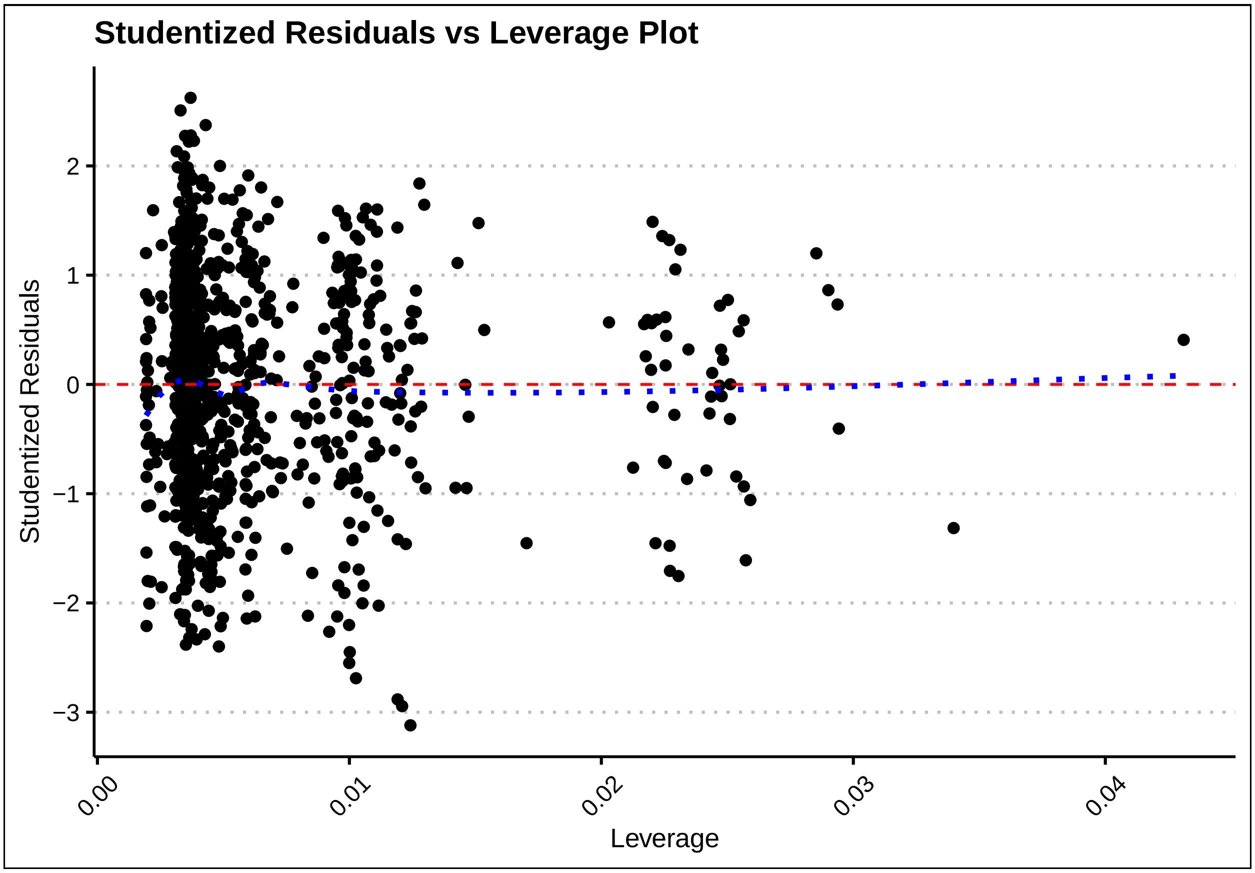
```
#Studentized Residuals vs. Fitted Value
ggplot(data = data.frame(Fitted = fitted(bwt_model4), Residuals = sr3),
       aes(x = Fitted, y = Residuals)) +
  geom_point() +
  geom_hline(yintercept = c(-2, 2), linetype = "dashed", color = "red") +
  labs(title = "Studentized Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Studentized Residuals")
```



```
#Q-Q Plot for Studentized Residuals
qq_data3 = data.frame(
  Theoretical = qqnorm(sr3, plot.it = FALSE)$x,
  Sample = qqnorm(sr3, plot.it = FALSE)$y
)

ggplot(qq_data3, aes(x = Theoretical, y = Sample)) +
  geom_point() +
  geom_abline(
    intercept = 0,
    slope = 1,
    color = "red",
    linetype = "dashed"
  ) +
  labs(title = "Q-Q Plot of Studentized Residuals", x = "Theoretical Quantiles", y = "Studentized Residu
```



```r
#Studentized Residuals vs. Leverage Plot
h3 = hatvalues(bwt_model4)
leverage_data3 = data.frame(Leverage = h3, StudentizedResiduals = sr3)


ggplot(leverage_data3, aes(x = Leverage, y = StudentizedResiduals)) +
  geom_point() +
  geom_hline(yintercept = 0,
             color = "red",
             linetype = "dashed") +
  geom_smooth(
    method = "loess",
    se = FALSE,
    color = "blue",
    linetype = "dotted"
  ) +
  labs(title = "Studentized Residuals vs Leverage Plot", x = "Leverage", y = "Studentized Residuals")
```


Studentized Residuals vs Leverage Plot
2
1
0
−1
−2
−3
Studentized Residuals
0.00
0.01
0.02
0.03
0.04
Leverage