PH1820 Group Project

Yue Zhang

2025-04-04

```
getwd()
```

```
## [1] "/Users/yuezhang/Documents/Biostat/Biostatistics/PH1820/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)
library(ggpubr)
library(GGally)
library(knitr)
library(car)
library(MASS)
library(faraway)
library(corrplot)
library(ggcorrplot)
library(goftest)
```

#Research Question: #Predictive modeling for life expectancy - is there a significant difference in life expectancy between developed and developing countries?

#Load Data

```
life = read.csv("/Users/yuezhang/Documents/Biostat/Biostatistics/PH1820/Data/Raw/Life Expectancy Data.cs
```

#EDA

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

```
##   Country Year Status Life.expectancy Adult.Mortality infant.deaths Alcohol
## 1       0    0      0              10              10             0     194
##   percentage.expenditure Hepatitis.B Measles BMI under.five.deaths Polio
## 1                      0         553       0  34                 0    19
```

```
##   Total.expenditure Diphtheria HIV.AIDS GDP Population thinness..1.19.years
## 1               226         19        0 448        652                    34
##   thinness.5.9.years Income.composition.of.resources Schooling
## 1                 34                            167       163
```

```
str(life)
```

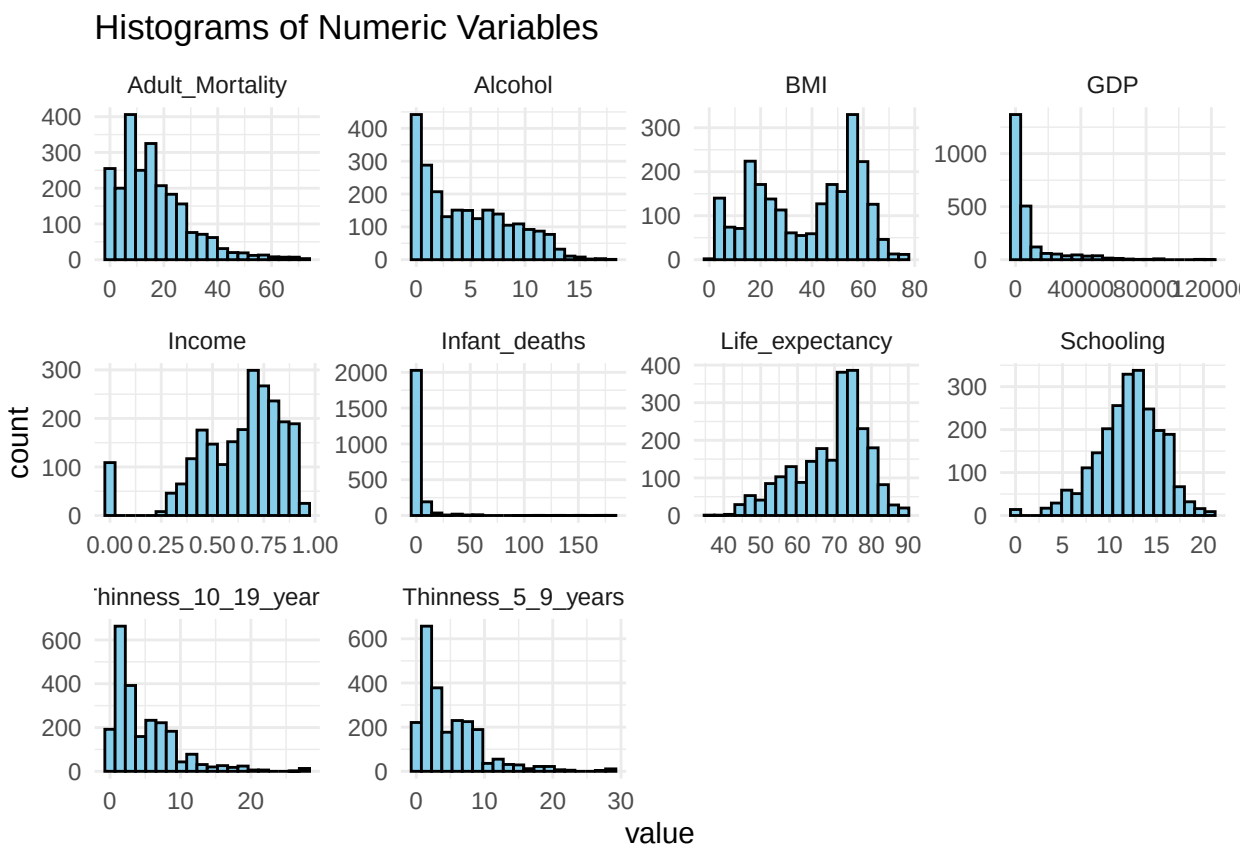
```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```
life = life[, c(
  "Life.expectancy",
  "Status",
  "Adult.Mortality",
  "infant.deaths",
  "Alcohol",
  "BMI",
  "thinness..1.19.years",
  "thinness.5.9.years",
  "Schooling",
  "GDP",
  "Income.composition.of.resources"
)]

life = life %>%
  drop_na() %>%
  rename(
    "Life_expectancy" = "Life.expectancy",
    "Adult_Mortality" = "Adult.Mortality",
    "Infant_deaths" = "infant.deaths",
    "Thinness_10_19_years" = "thinness..1.19.years",
    "Thinness_5_9_years" = "thinness.5.9.years",
    "Income" = "Income.composition.of.resources"
  )
```

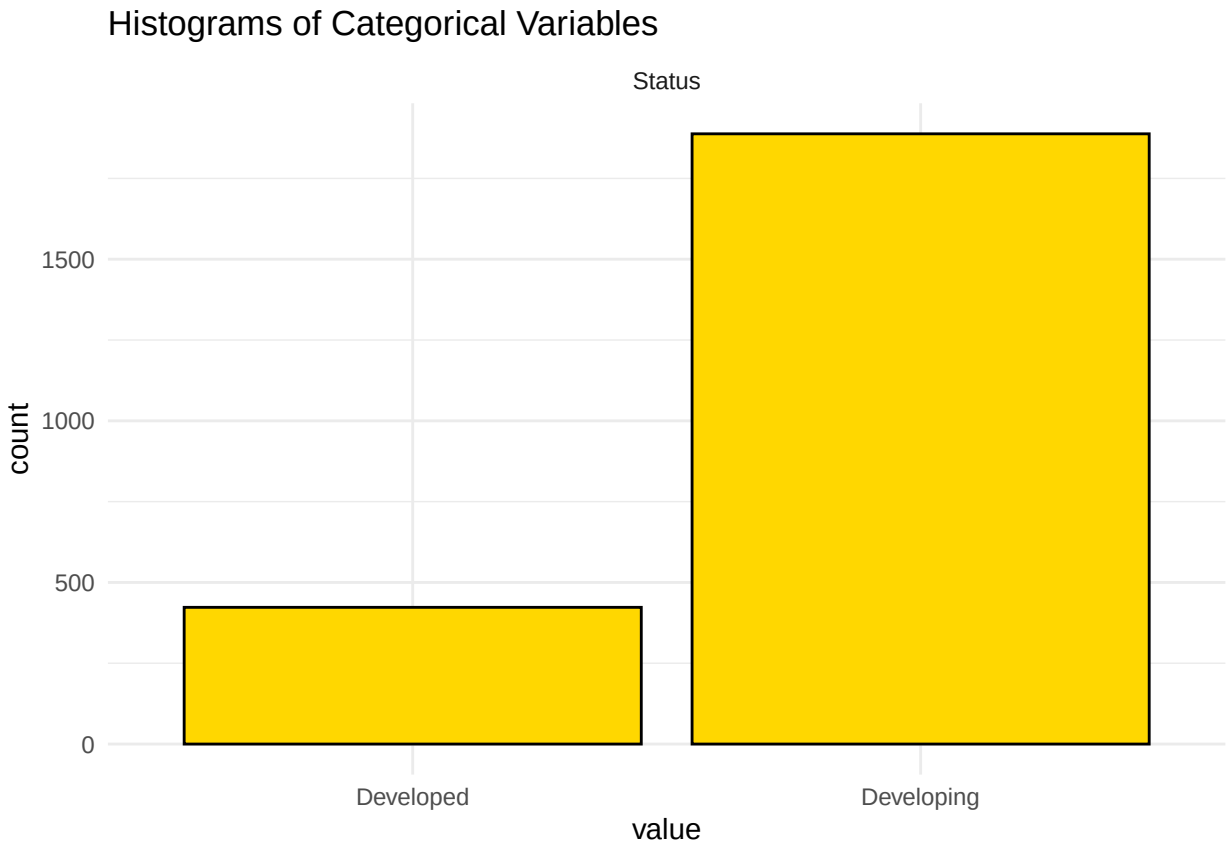
```r
#Convert the adult mortality, under five deaths, and infant deaths into % as in Kaggle they're deaths p
life = life %>% mutate(
  Adult_Mortality = Adult_Mortality / 10,
  Infant_deaths = Infant_deaths / 10
)

life$Status = as.factor(life$Status)

#Descriptive Statistics for variables
life %>%
  summarise(across(where(is.numeric), list(
    mean = ~mean(., na.rm = TRUE),
    sd   = ~sd(., na.rm = TRUE),
    min  = ~min(., na.rm = TRUE),
    max  = ~max(., na.rm = TRUE),
    missing = ~sum(is.na(.))
  ))) %>%
  pivot_longer(cols = everything(),
               names_to = c("Variable", "Statistic"),
               names_pattern = "(.*)_(.*)") %>%
  pivot_wider(names_from = Statistic, values_from = value)
```

```
## # A tibble: 10 x 6
##    Variable                mean        sd   min        max missing
##    <chr>                  <dbl>     <dbl> <dbl>      <dbl>   <dbl>
##  1 Life_expectancy        69.3       9.70  36.3      89           0
##  2 Adult_Mortality        16.2      12.8    0.1      72.3         0
##  3 Infant_deaths           3.16     12.9    0       180           0
##  4 Alcohol                 4.61      4.03   0.01     17.9         0
##  5 BMI                    38.1      19.8    1.4      77.1         0
##  6 Thinness_10_19_years    4.86      4.53   0.1      27.7         0
##  7 Thinness_5_9_years      4.90      4.62   0.1      28.6         0
##  8 Schooling              12.1       3.35   0        20.7         0
##  9 GDP                  7597.    14518.     1.68 119173.          0
## 10 Income                  0.630     0.215  0         0.948       0
```

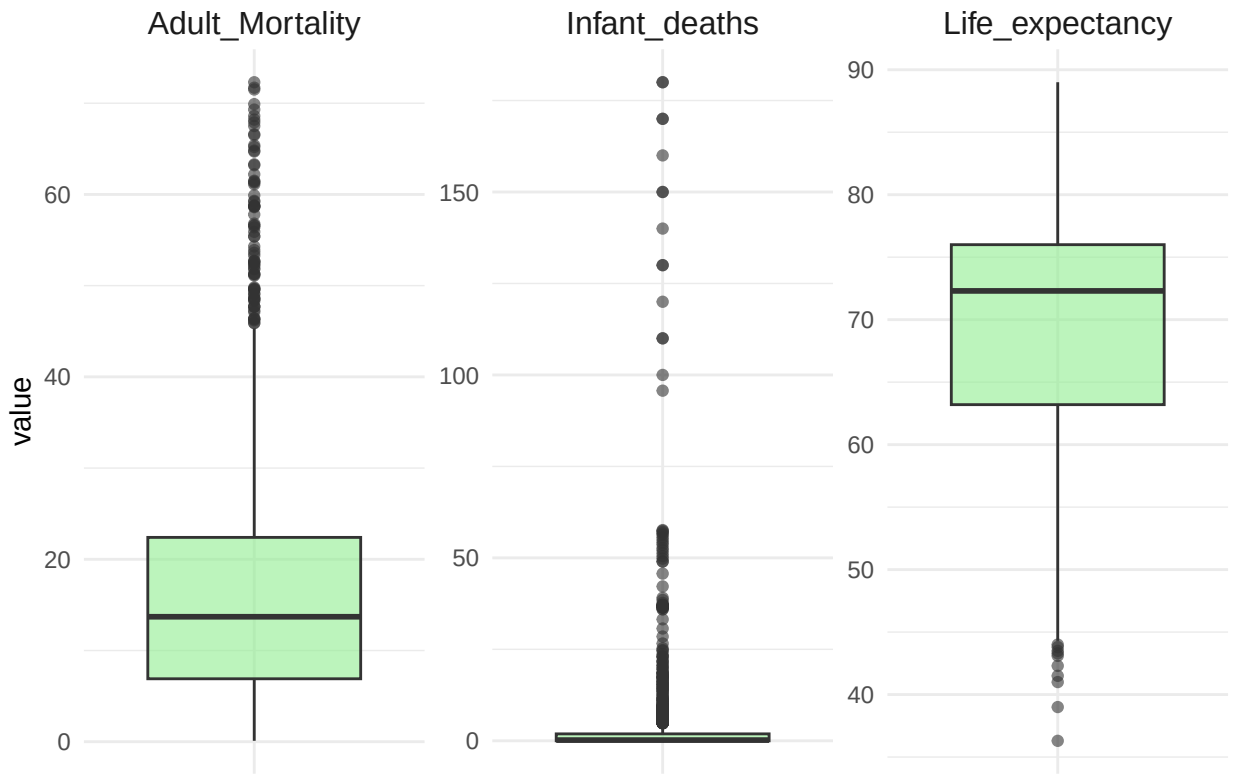
```r
#Histograms
life %>% select_if(is.numeric) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  facet_wrap(~ name, scales = "free") +
  theme_minimal() +
  ggtitle("Histograms of Numeric Variables")
```



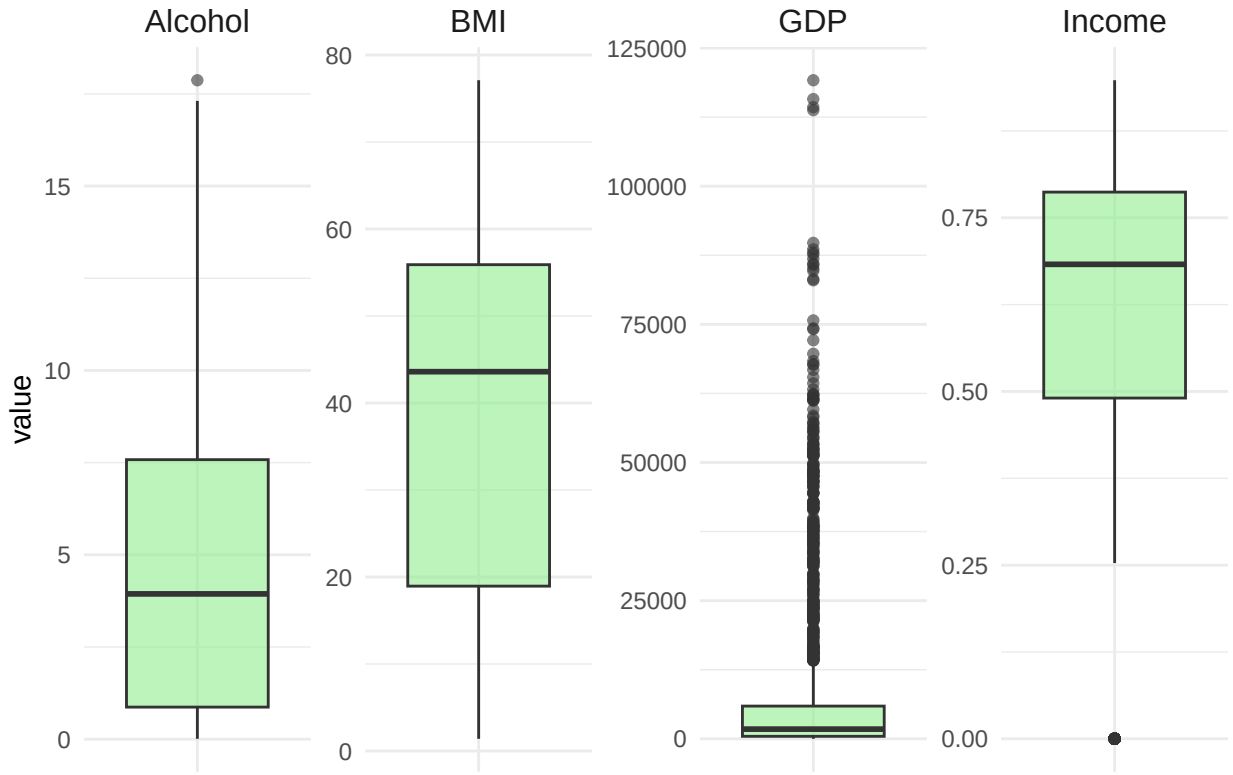

```
life %>% select_if(is.factor) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "gold", color = "black") +
  facet_wrap(~ name, scales = "free") +
  theme_minimal() +
  ggtitle("Histograms of Categorical Variables")
```



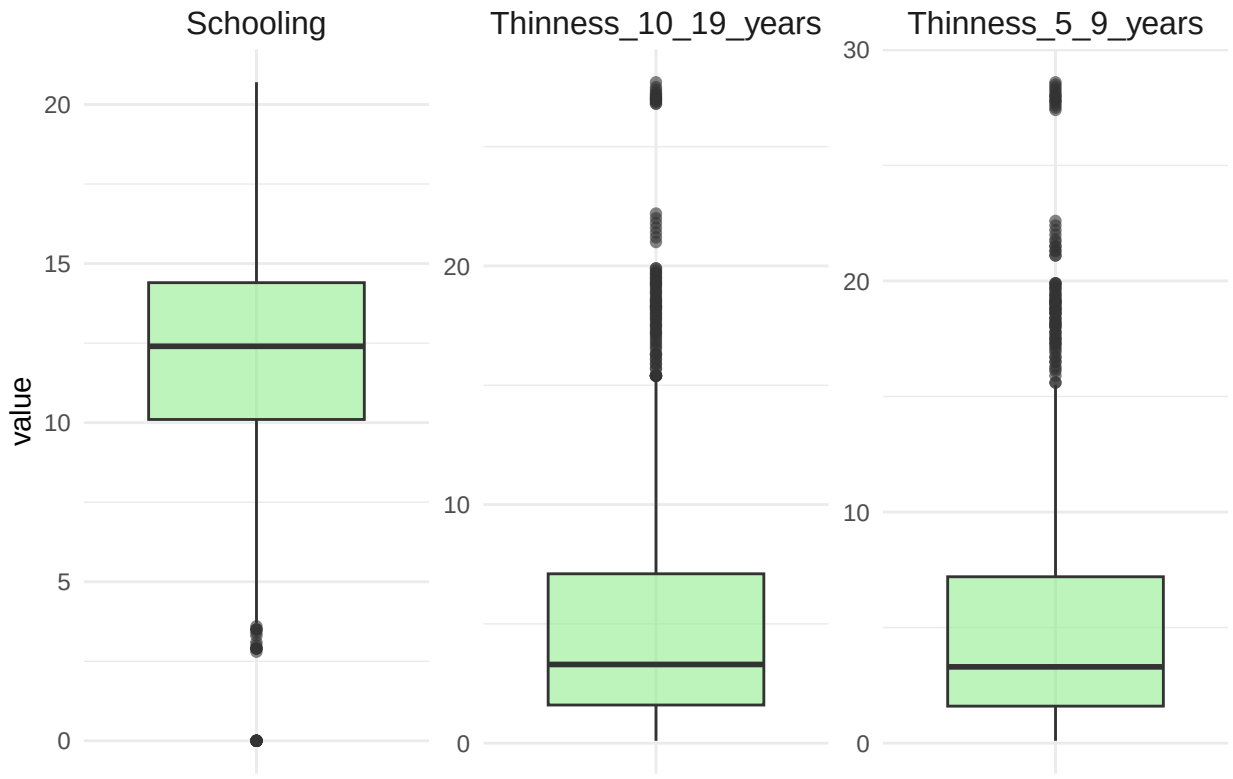

```
#Boxplots
life %>%
  dplyr::select(Life_expectancy, Adult_Mortality, Infant_deaths) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = "", y = value)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.6) +
  facet_wrap(~ name, scales = "free", nrow = 1) +
  theme_minimal() +
  ggtitle("Boxplots of Numeric Variables") +
  theme(
    axis.text.x = element_blank(),
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.text = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold")
  )
```



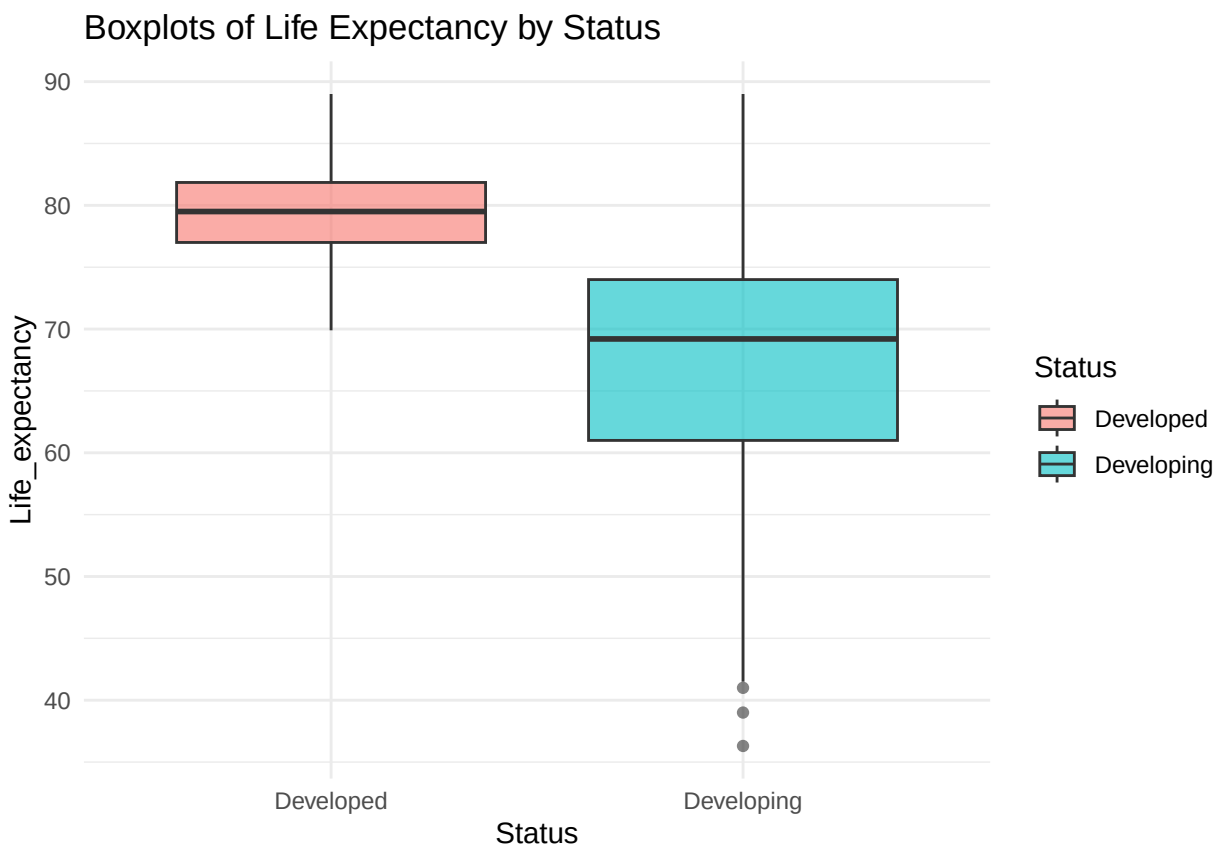

```
life %>%
  dplyr::select(Alcohol, BMI, GDP, Income) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = "", y = value)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.6) +
  facet_wrap(~ name, scales = "free", nrow = 1) +
  theme_minimal() +
  ggtitle("Boxplots of Numeric Variables") +
  theme(
    axis.text.x = element_blank(),
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.text = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold")
  )
```



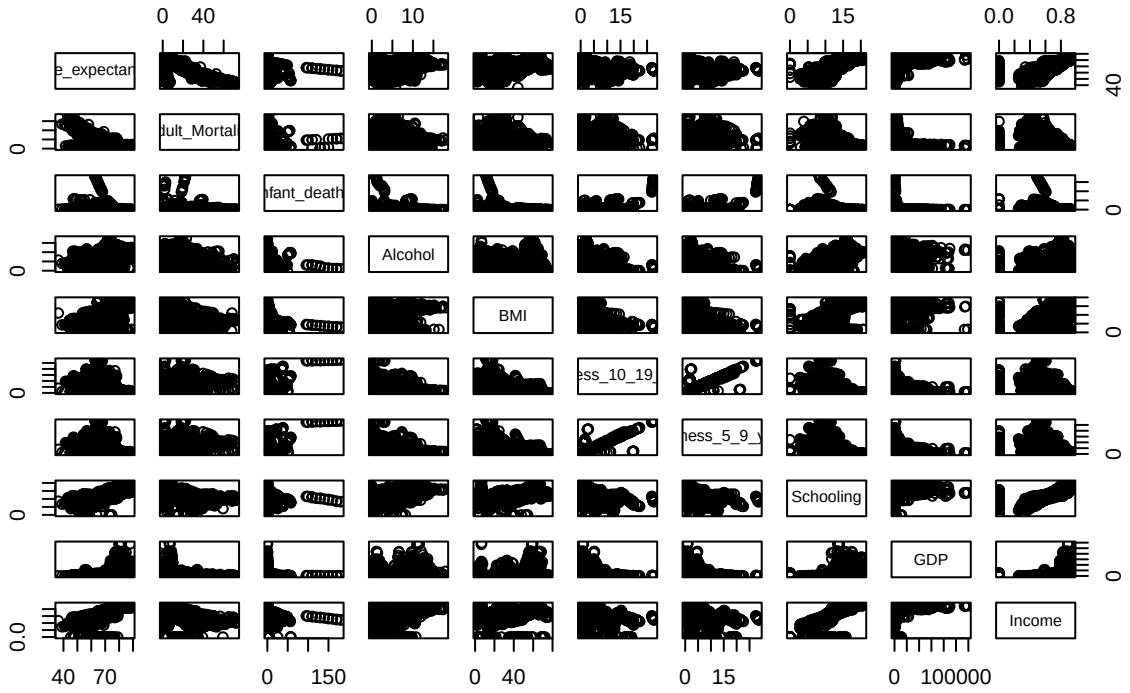

```r
life %>%
  dplyr::select(Thinness_5_9_years, Thinness_10_19_years, Schooling) %>%
  pivot_longer(cols = everything()) %>%
  ggplot(aes(x = "", y = value)) +
  geom_boxplot(fill = "lightgreen", alpha = 0.6) +
  facet_wrap(~ name, scales = "free", nrow = 1) +
  theme_minimal() +
  ggtitle("Boxplots of Numeric Variables") +
  theme(
    axis.text.x = element_blank(),
    axis.title.x = element_blank(),
    axis.ticks.x = element_blank(),
    strip.text = element_text(size = 12),
    plot.title = element_text(size = 14, face = "bold")
  )
```




```
ggplot(life, aes(x = Status, y = Life_expectancy, fill = Status)) +
  geom_boxplot(alpha = 0.6) +
  theme_minimal() +
  ggtitle("Boxplots of Life Expectancy by Status")
```



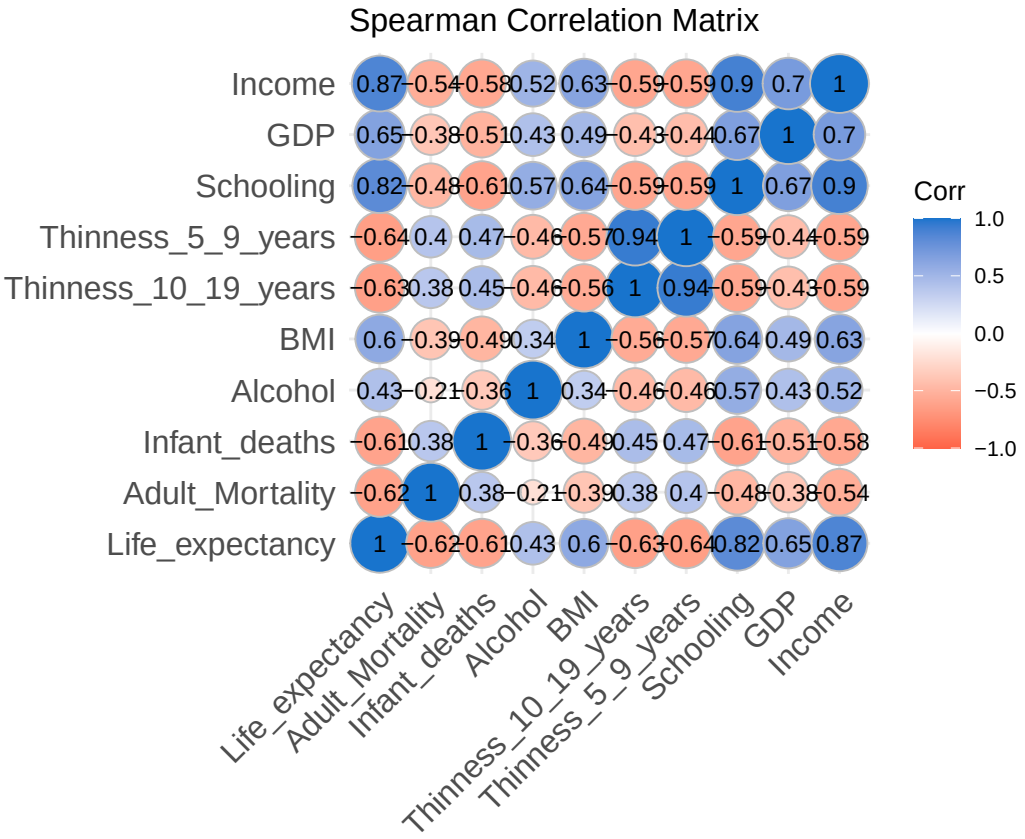

```
#Scatterplot
life_numeric = life %>% select_if(is.numeric)
pairs(life_numeric, main = "Scatterplot Matrix of Numeric Variables")
```



```r
#Correlation
cor_matrix = cor(life_numeric, method = "spearman", use = "pairwise.complete.obs")
print(cor_matrix)
```

```
##                       Life_expectancy Adult_Mortality Infant_deaths    Alcohol
## Life_expectancy             1.0000000      -0.6238504    -0.6145628  0.4301817
## Adult_Mortality            -0.6238504       1.0000000     0.3789025 -0.2053702
## Infant_deaths              -0.6145628       0.3789025     1.0000000 -0.3594502
## Alcohol                     0.4301817      -0.2053702    -0.3594502  1.0000000
## BMI                         0.6026058      -0.3901869    -0.4876205  0.3428487
## Thinness_10_19_years       -0.6257777       0.3823203     0.4516519 -0.4649118
## Thinness_5_9_years         -0.6356402       0.3990258     0.4692331 -0.4558256
## Schooling                   0.8158743      -0.4788217    -0.6130518  0.5665574
## GDP                         0.6473679      -0.3839143    -0.5110845  0.4271796
## Income                      0.8694348      -0.5358526    -0.5837963  0.5230530
##                                BMI Thinness_10_19_years Thinness_5_9_years
## Life_expectancy        0.6026058           -0.6257777         -0.6356402
## Adult_Mortality       -0.3901869            0.3823203          0.3990258
## Infant_deaths         -0.4876205            0.4516519          0.4692331
## Alcohol                0.3428487           -0.4649118         -0.4558256
## BMI                    1.0000000           -0.5606424         -0.5717655
## Thinness_10_19_years  -0.5606424            1.0000000          0.9398958
## Thinness_5_9_years    -0.5717655            0.9398958          1.0000000
## Schooling              0.6354777           -0.5930163         -0.5936351
## GDP                    0.4856409           -0.4257313         -0.4350441
## Income                 0.6281214           -0.5942579         -0.5932238
##                        Schooling        GDP     Income
## Life_expectancy        0.8158743  0.6473679  0.8694348
## Adult_Mortality       -0.4788217 -0.3839143 -0.5358526
## Infant_deaths         -0.6130518 -0.5110845 -0.5837963
## Alcohol                0.5665574  0.4271796  0.5230530
## BMI                    0.6354777  0.4856409  0.6281214
## Thinness_10_19_years  -0.5930163 -0.4257313 -0.5942579
## Thinness_5_9_years    -0.5936351 -0.4350441 -0.5932238
## Schooling              1.0000000  0.6723717  0.9018080
## GDP                    0.6723717  1.0000000  0.6979678
## Income                 0.9018080  0.6979678  1.0000000
```

```r
ggcorrplot(cor_matrix, method = "circle",
           lab = TRUE, lab_size = 3, colors = c("tomato", "white", "dodgerblue3"),
           title = "Spearman Correlation Matrix",
           ggtheme = theme_minimal()) +
  theme(text = element_text(size = 10))
```



#Fit preliminary model

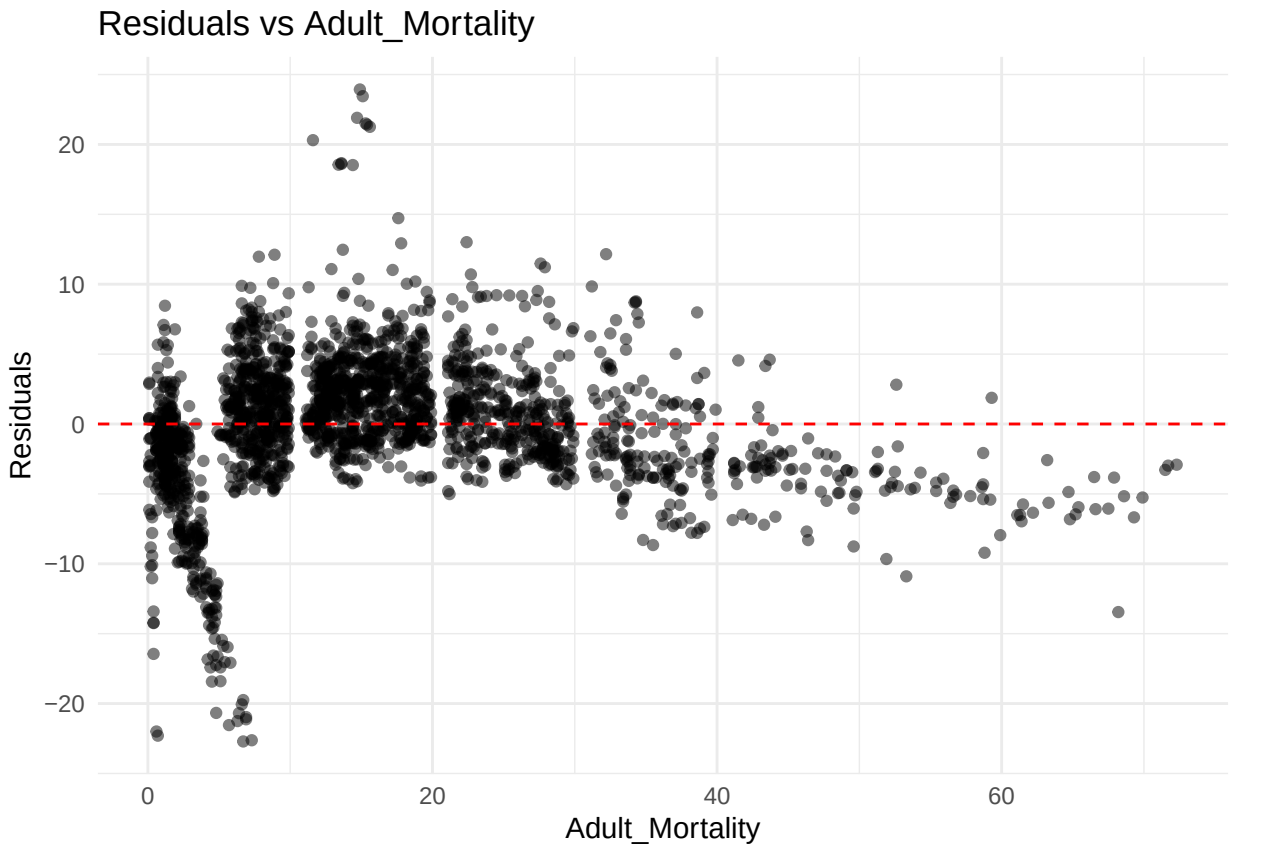
```
model_pre = lm(data = life, life$Life_expectancy ~ .)
summary(model_pre)
```

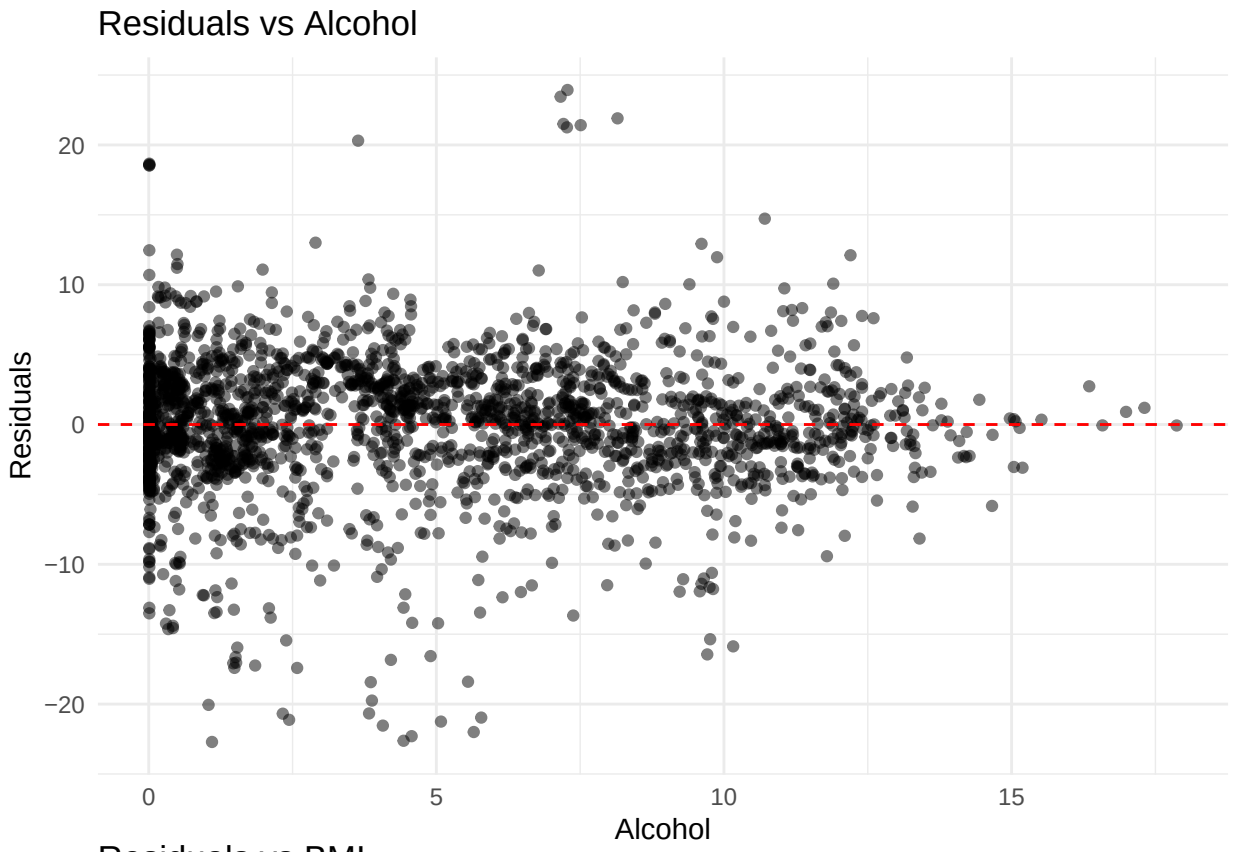
```
##
## Call:
## lm(formula = life$Life_expectancy ~ ., data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7096  -2.2343   0.2018   2.7638  23.9309
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.782e+01  7.323e-01  78.954  < 2e-16 ***
## StatusDeveloping     -1.708e+00  3.558e-01  -4.799 1.69e-06 ***
## Adult_Mortality      -2.805e-01  9.158e-03 -30.629  < 2e-16 ***
## Infant_deaths        -1.680e-03  9.003e-03  -0.187   0.8520
## Alcohol              -1.764e-01  3.406e-02  -5.179 2.43e-07 ***
## BMI                   5.200e-02  6.739e-03   7.717 1.77e-14 ***
## Thinness_10_19_years -1.244e-01  6.119e-02  -2.033   0.0421 *
## Thinness_5_9_years   -1.450e-02  6.028e-02  -0.240   0.8100
## Schooling             9.249e-01  5.578e-02  16.580  < 2e-16 ***
## GDP                   4.203e-05  8.238e-06   5.103 3.62e-07 ***
## Income                8.655e+00  7.993e-01  10.828  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

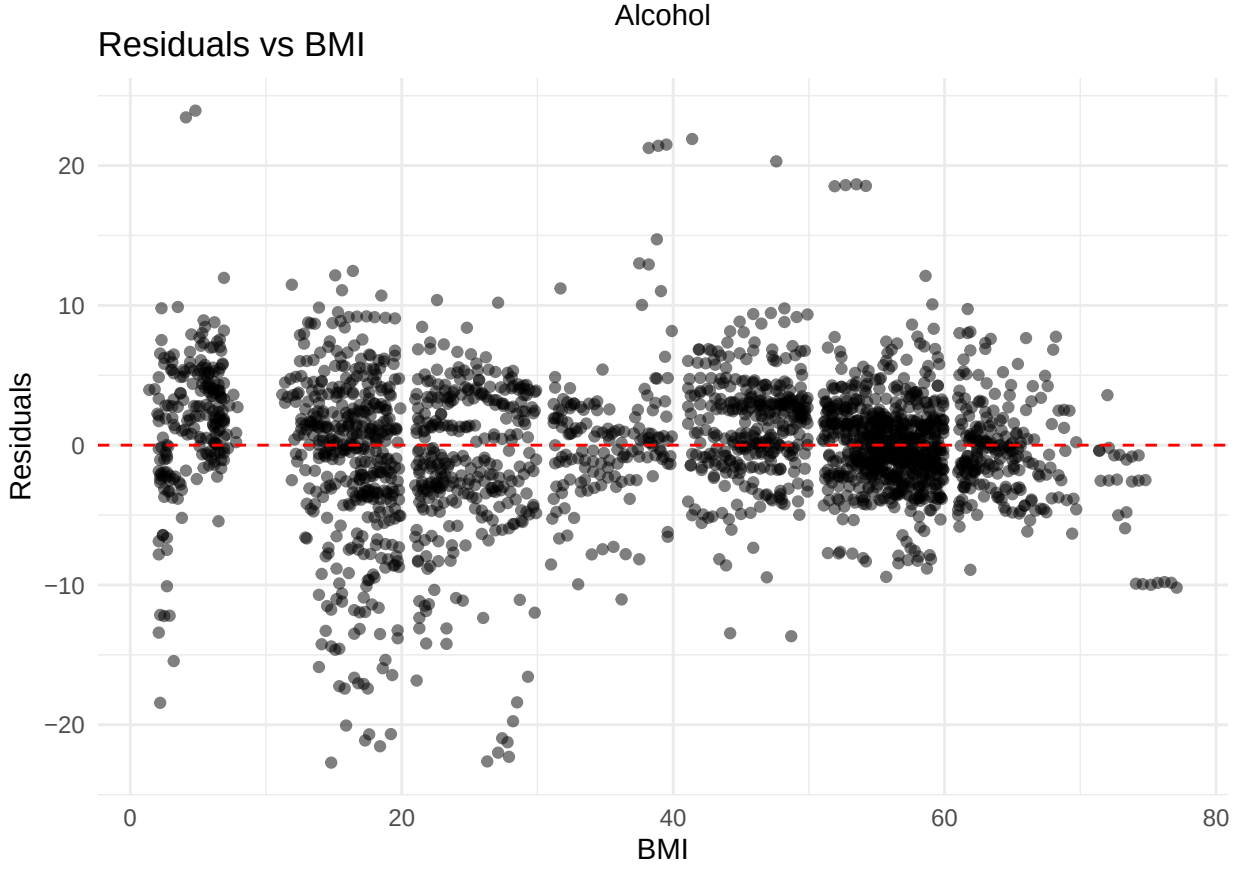
```
##
## Residual standard error: 4.781 on 2300 degrees of freedom
## Multiple R-squared:  0.7583, Adjusted R-squared:  0.7573
## F-statistic: 721.7 on 10 and 2300 DF,  p-value: < 2.2e-16
```

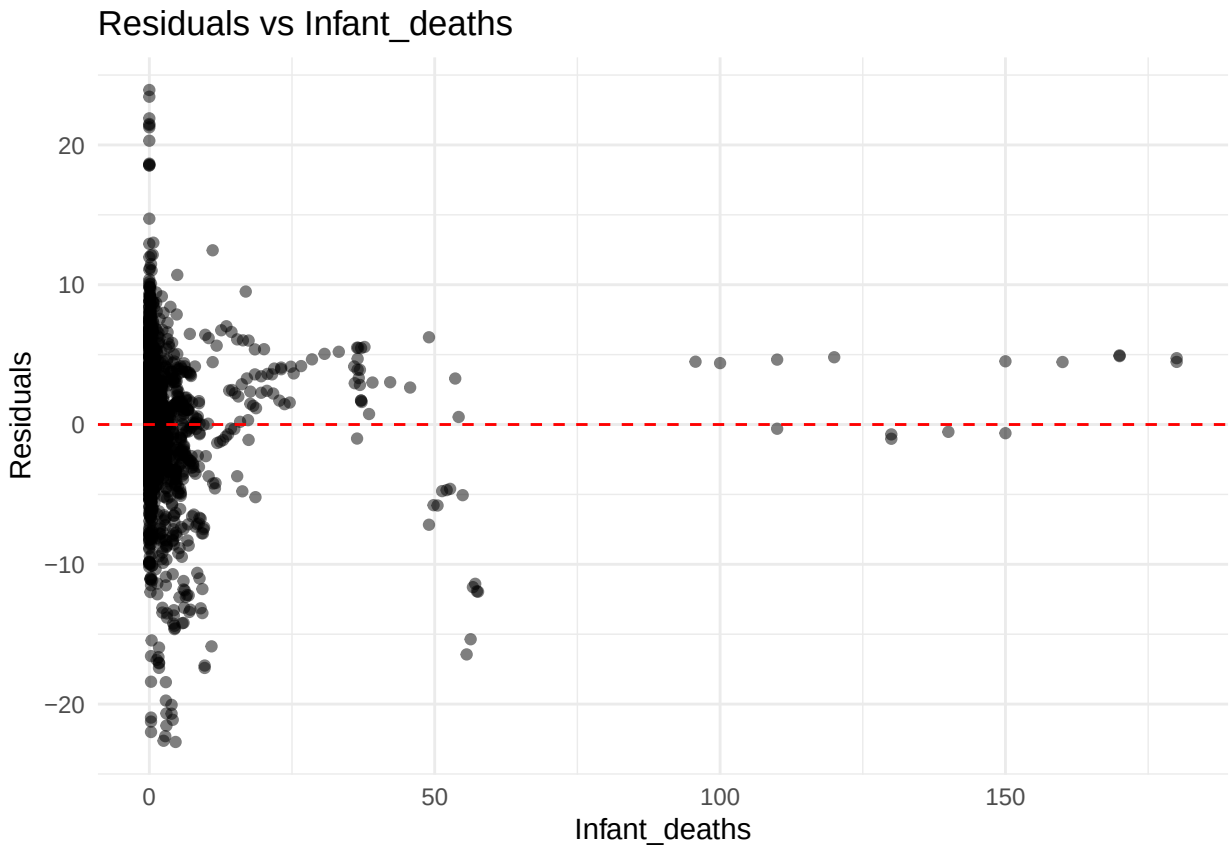
#Preliminary Diagnostics

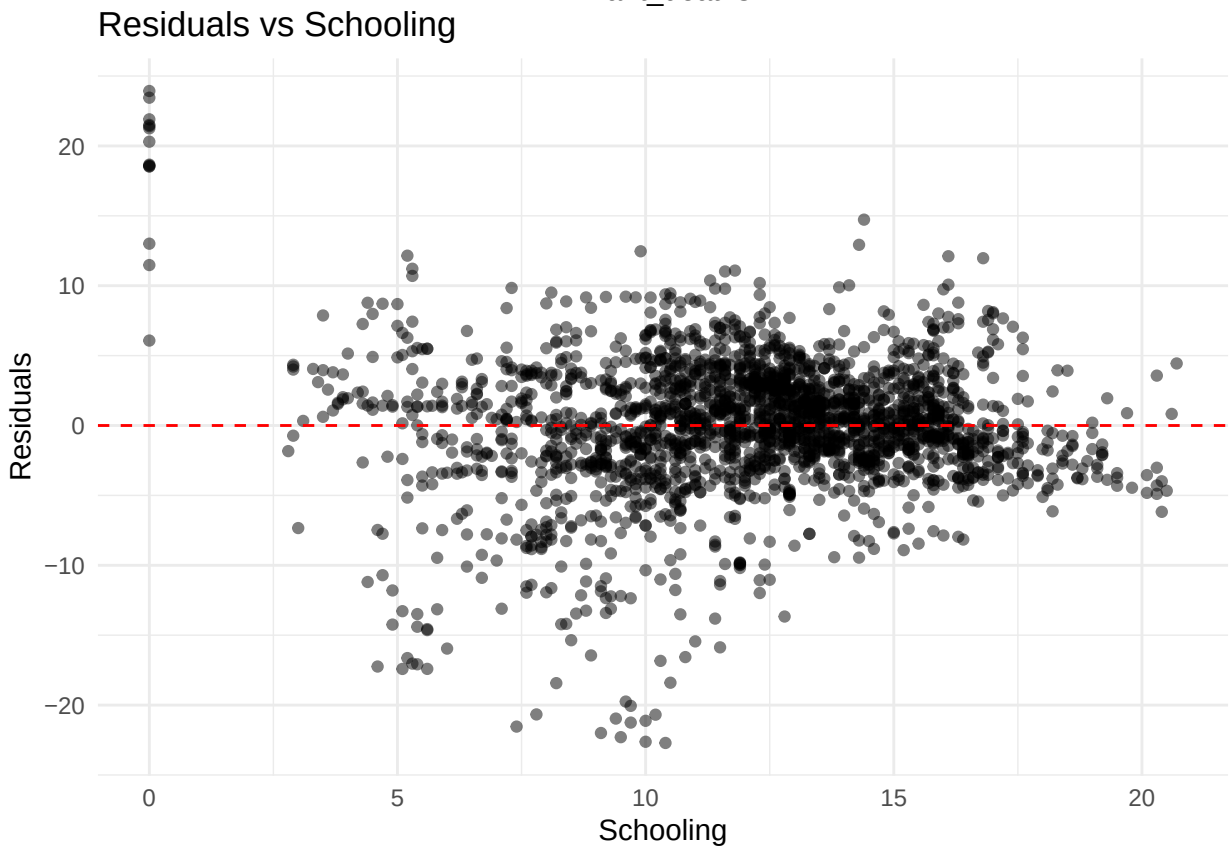
```
#Linearity
#Residuals vs. x
fitted_vals = fitted(model_pre)
residuals = resid(model_pre)

predictors = c("Adult_Mortality", "Alcohol", "BMI", "Infant_deaths",
               "Schooling", "Thinness_10_19_years", "Thinness_5_9_years", "GDP", "Income")

for (var in predictors) {
  p = ggplot(life, aes(x = .data[[var]], y = residuals)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
    ggtitle(paste("Residuals vs", var)) +
    theme_minimal() +
    ylab("Residuals") +
    xlab(var)

  print(p)
}
```

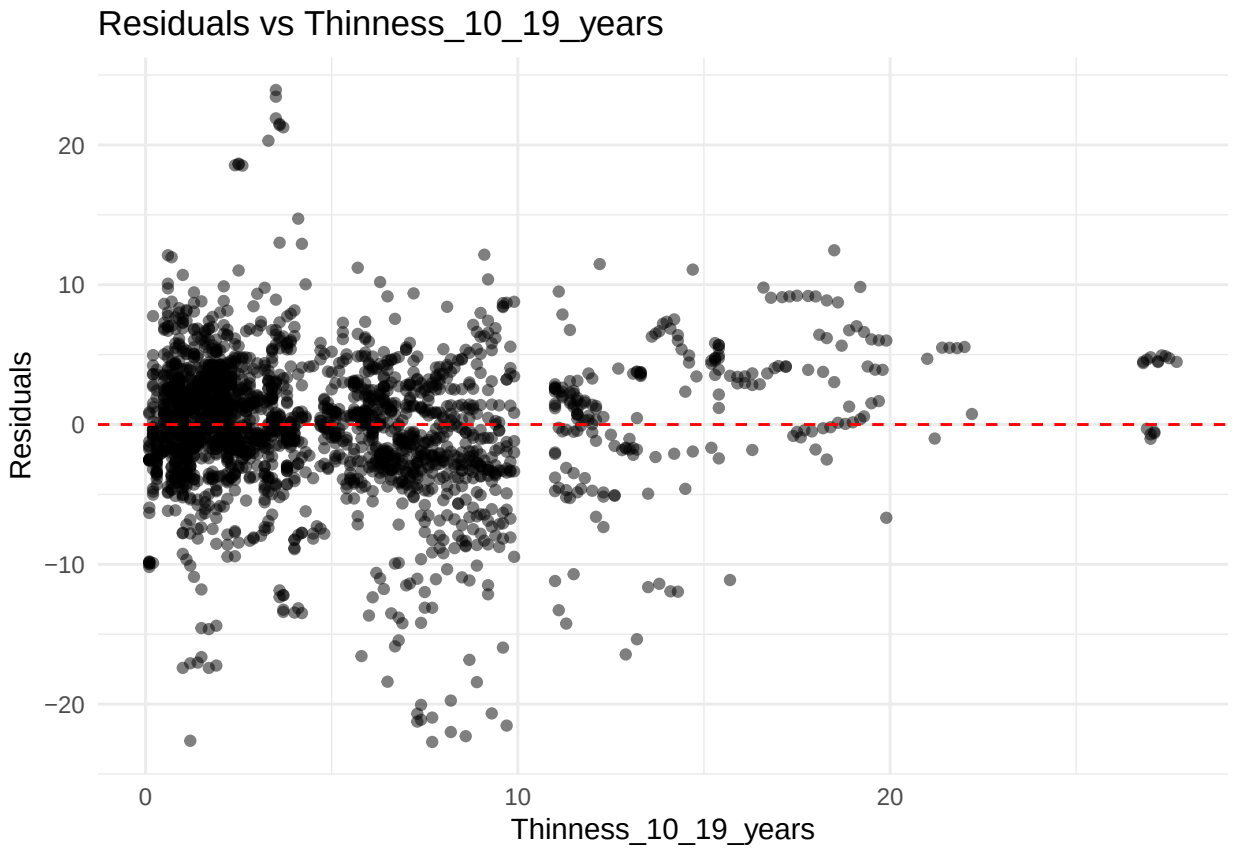


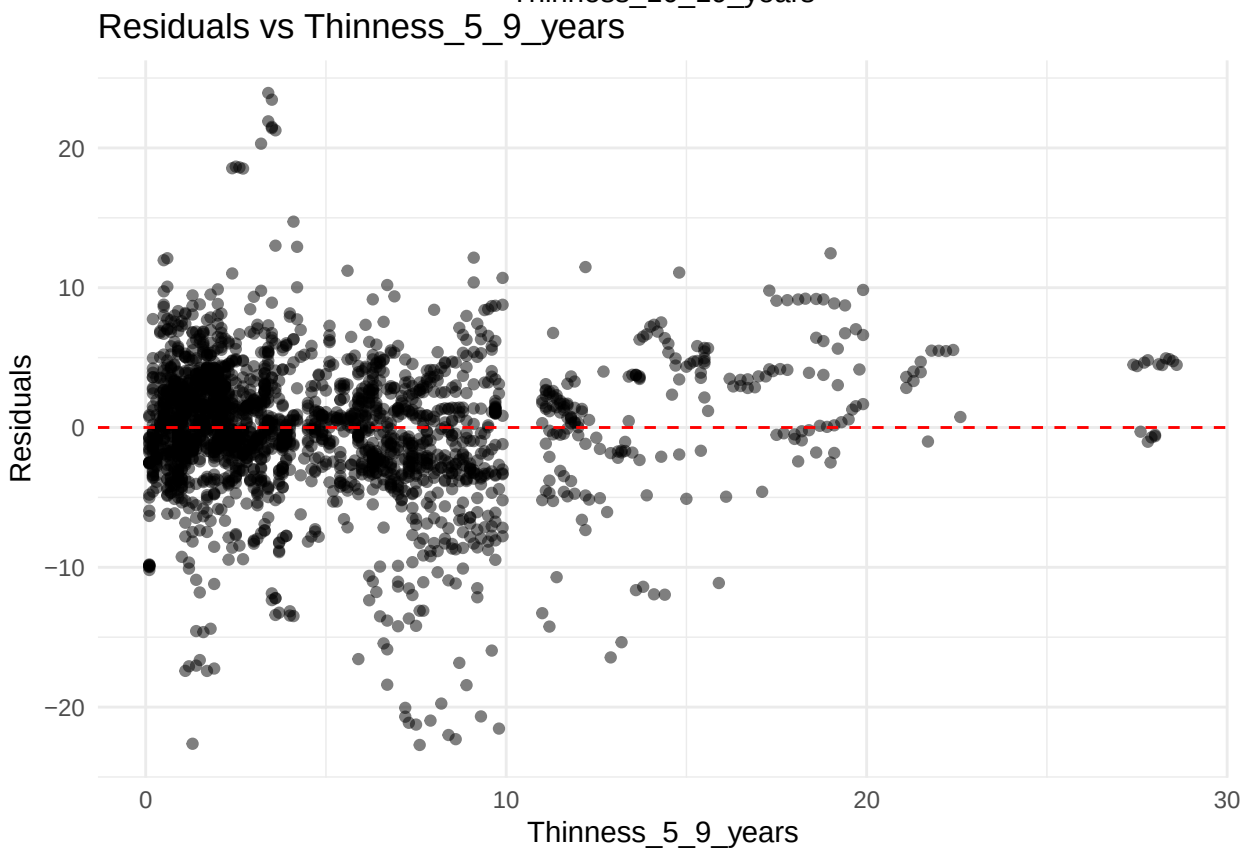

Residuals vs Alcohol
Residuals
Alcohol
20
10
0
−10
−20
0
5
10
15

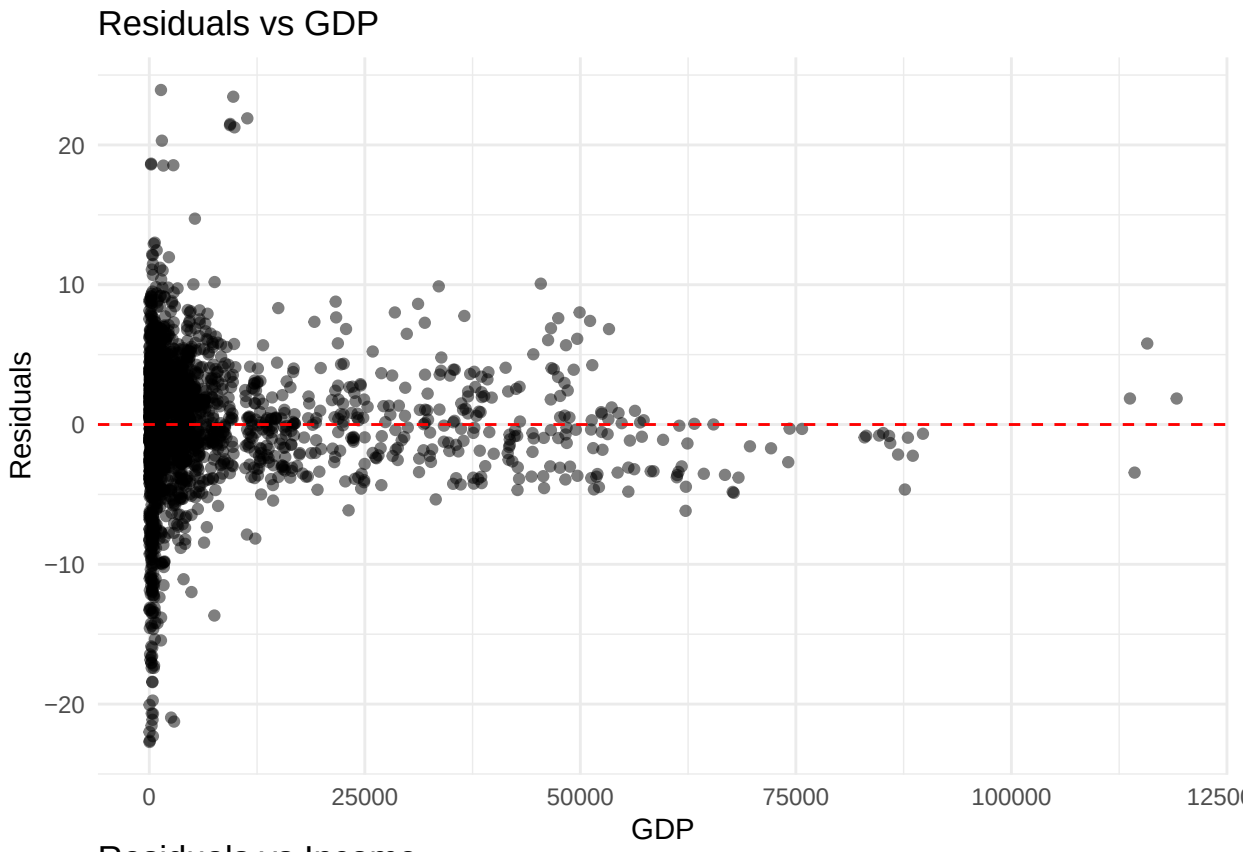

Residuals vs BMI
Residuals
BMI
20
10
0
−10
−20
0
20
40
60
80

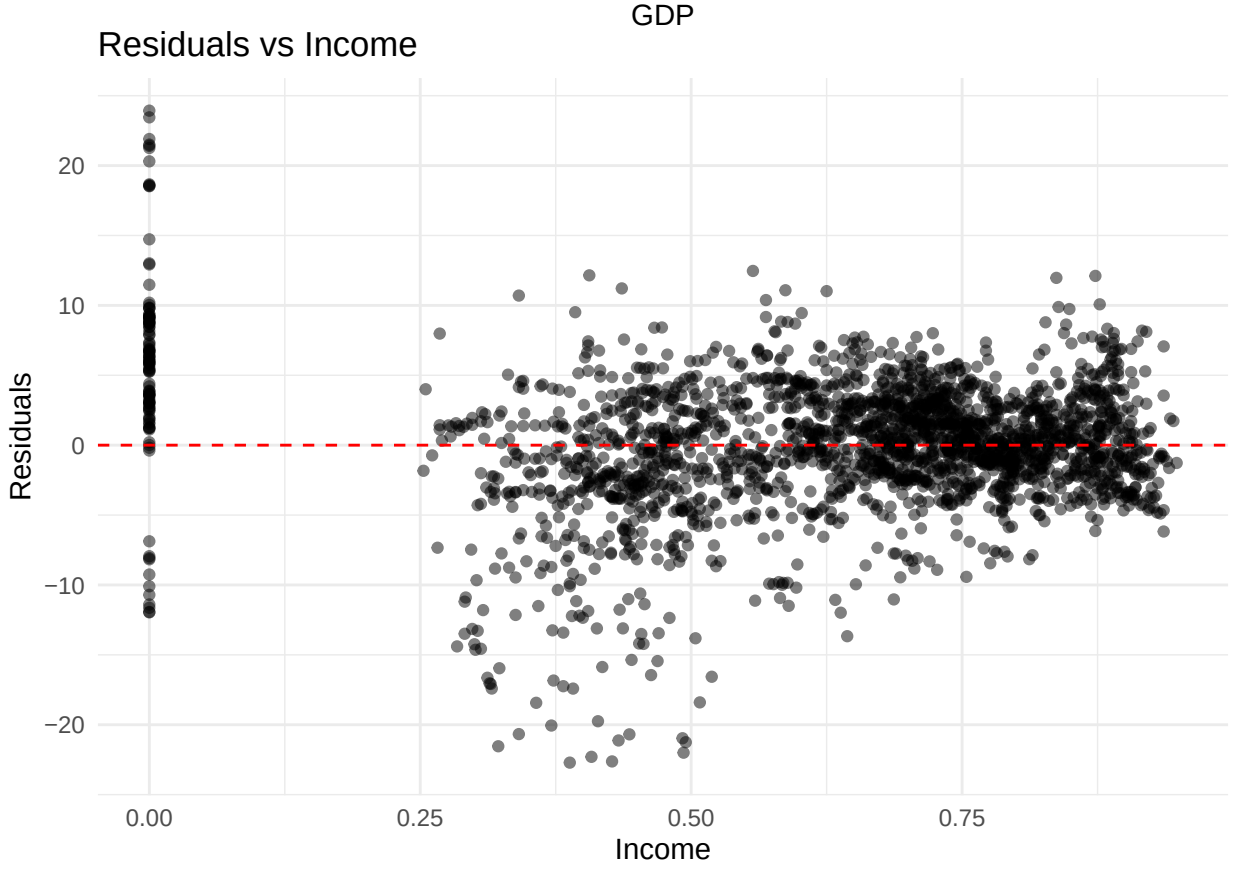

Residuals vs Infant_deaths
Residuals
Infant_deaths
20
10
0
−10
−20
0
50
100
150


Residuals vs Schooling
Residuals
Schooling
20
10
0
−10
−20
0
5
10
15
20

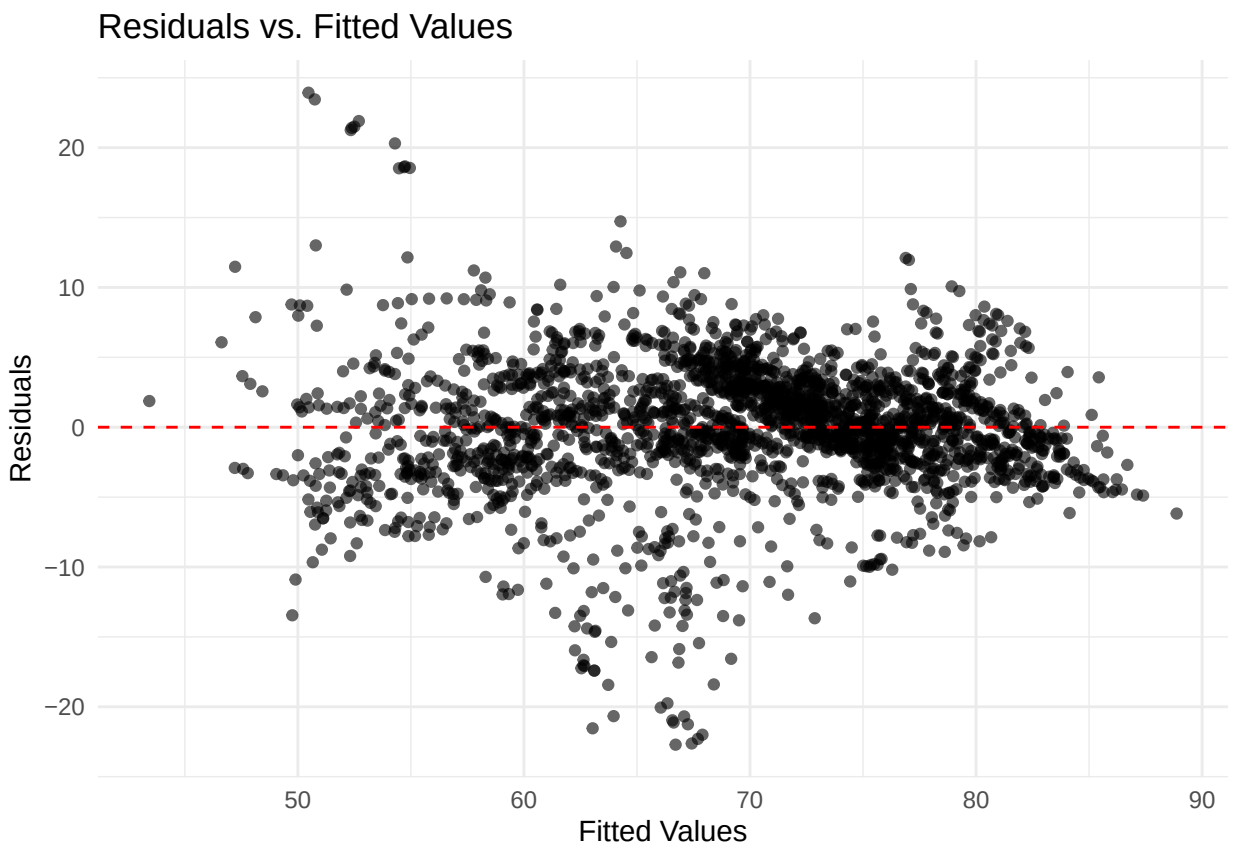

Residuals vs Thinness_10_19_years
Residuals
20
10
0
−10
−20
0
10
20
Thinness_10_19_years


Residuals vs Thinness_5_9_years
Residuals
20
10
0
−10
−20
0
10
20
30
Thinness_5_9_years


Residuals vs GDP
Residuals
GDP
20
10
0
−10
−20
0
25000
50000
75000
100000


Residuals vs Income
Residuals
Income
20
10
0
−10
−20
0.00
0.25
0.50
0.75

```r
# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals, y = residuals)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```r
#Linear

#Homoscedasticity
bptest(model_pre)
```

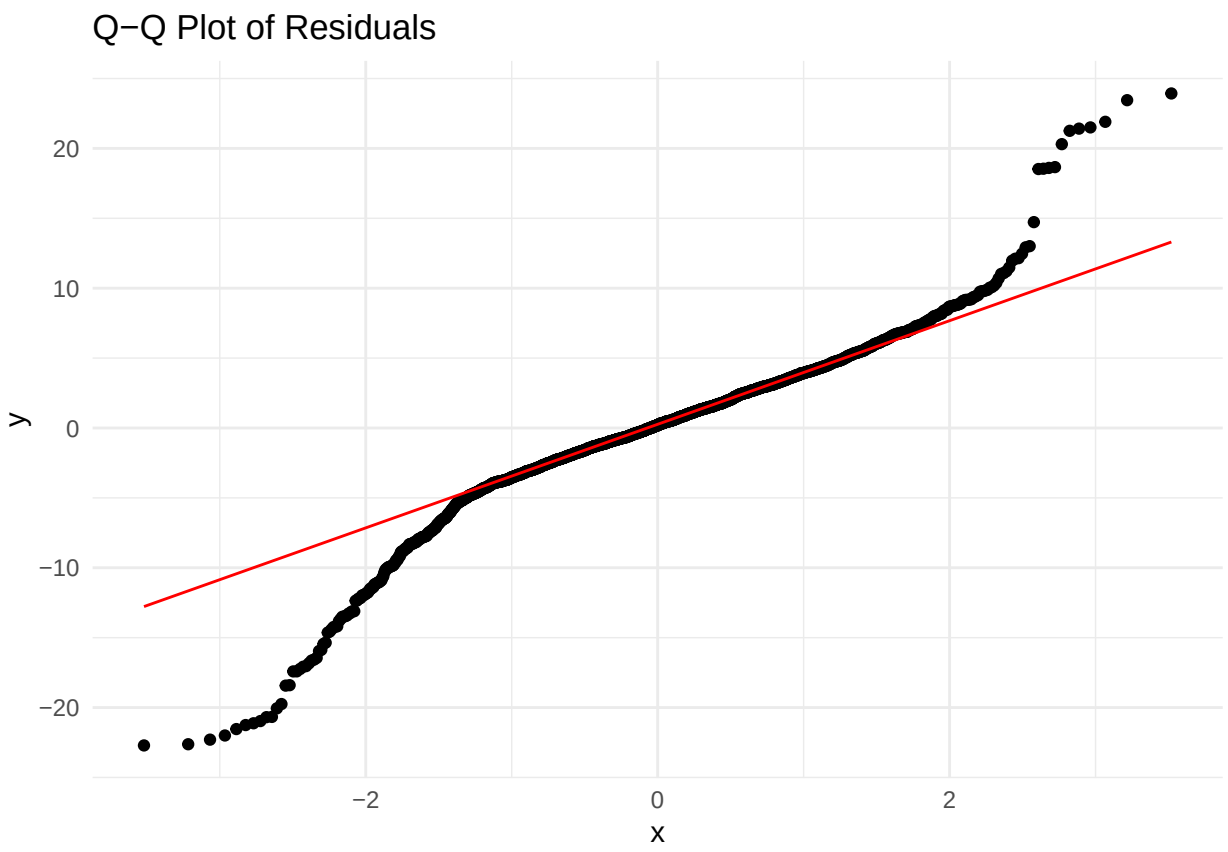
```
##
##  studentized Breusch-Pagan test
##
## data:  model_pre
## BP = 585.98, df = 10, p-value < 2.2e-16
```

```r
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality, since this is a large sample test, the Shapiro Wilk test will be sensitive, we will use And
goftest::ad.test(residuals)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
```

```
## data:  residuals
## An = Inf, p-value = 2.596e-07
```

```
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()
```

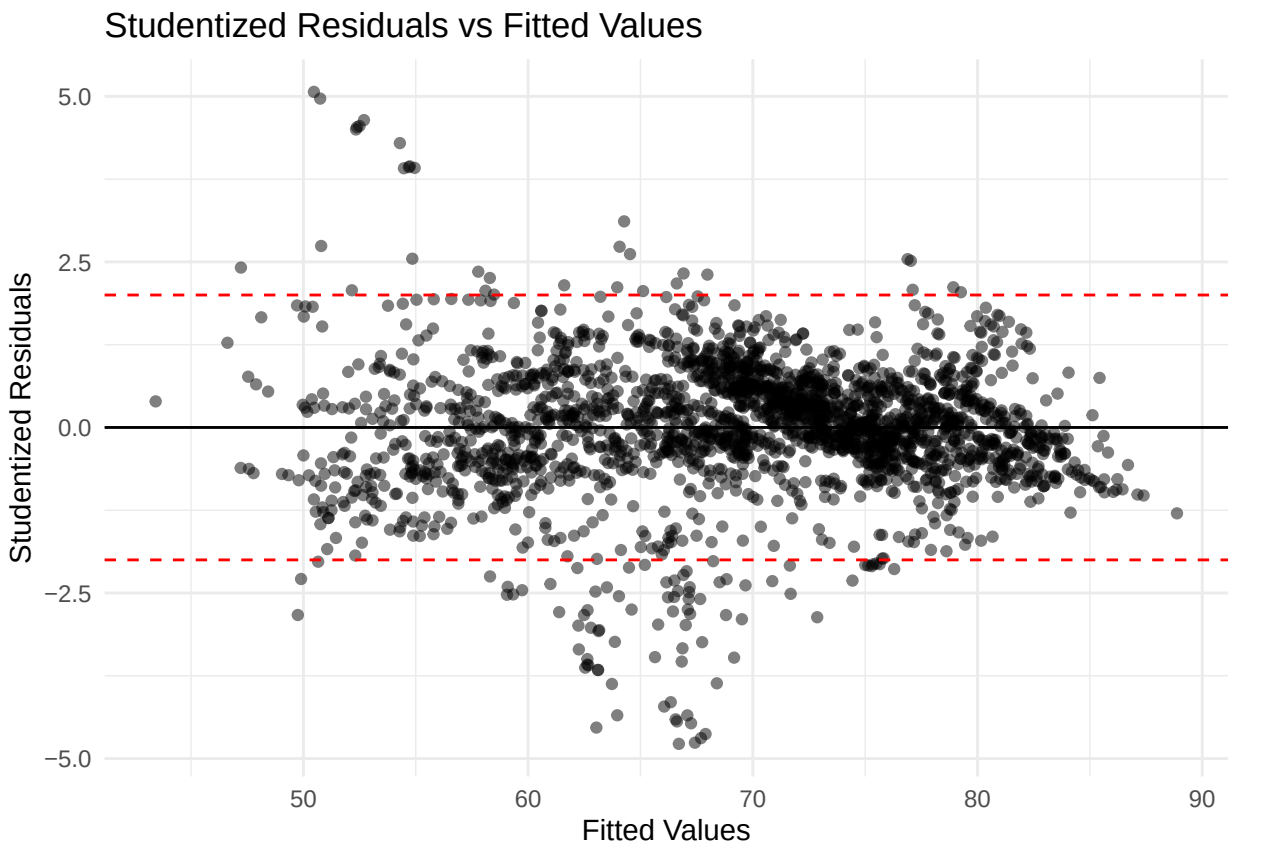



```
#Not normally distributed

#VIF
vif(model_pre)
```

```
##     StatusDeveloping      Adult_Mortality        Infant_deaths
##             1.914029             1.382652             1.369239
##              Alcohol                  BMI Thinness_10_19_years
##             1.902441             1.801401             7.757667
##   Thinness_5_9_years           Schooling                  GDP
##             7.840726             3.528343             1.445231
##               Income
##             2.981523
```

```
#Some Multicollinearity, as the thinness_5_9 and thinness_10_19 vif are larger than 5 and thinness_5_9

#Potential Outliers
resid_stud = rstudent(model_pre)
ggplot(data = NULL, aes(x = fitted_vals, y = resid_stud)) +
```

```
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = c(-2, 2), color = "red", linetype = "dashed") +
  geom_hline(yintercept = 0, color = "black") +
  labs(
    title = "Studentized Residuals vs Fitted Values",
    x = "Fitted Values",
    y = "Studentized Residuals"
  ) +
  theme_minimal()
```




```
potential_outliers = which(abs(resid_stud) > 2)
print(potential_outliers)
```

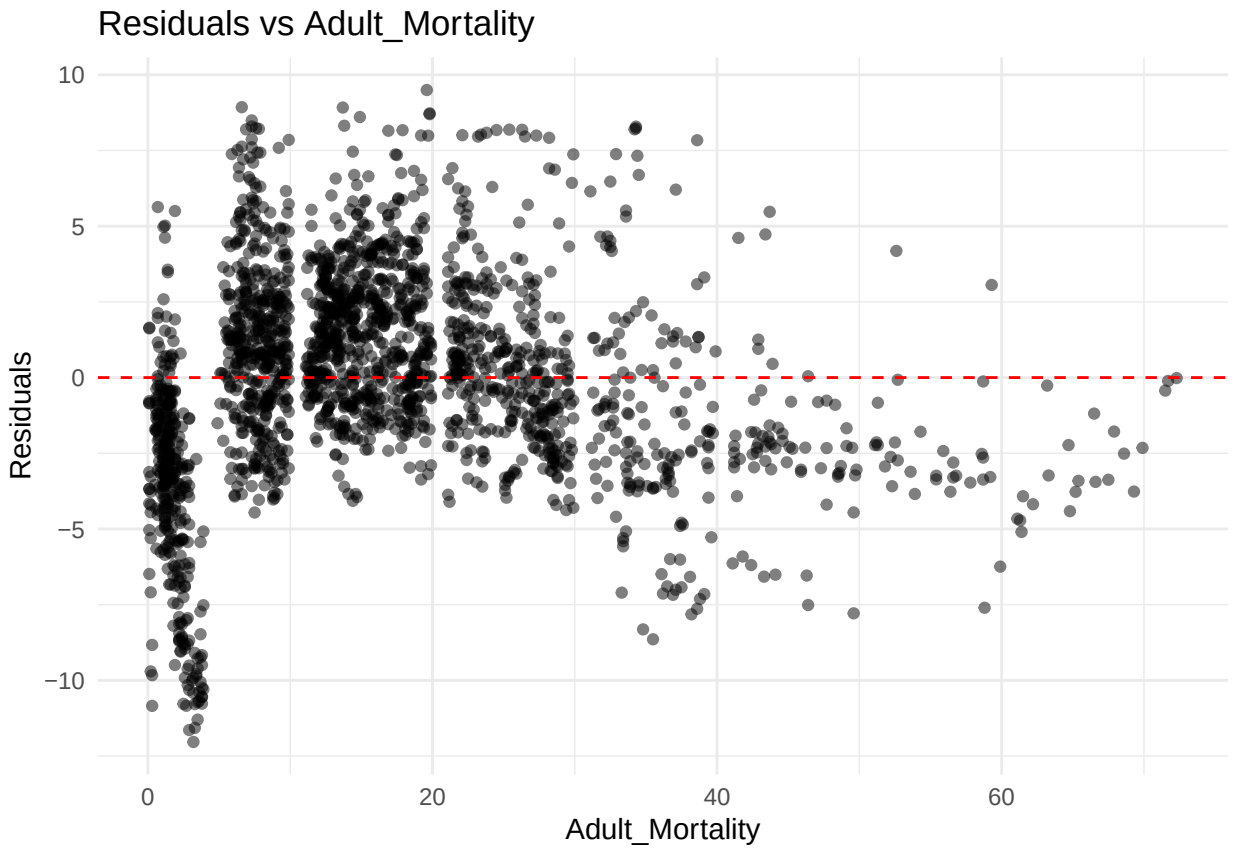
```
##   61   62   72   73   74   75   76   77  170  262  272  286  287  290  364  390
##   61   62   72   73   74   75   76   77  170  262  272  286  287  290  364  390
##  412  417  420  421  446  449  450  451  452  462  463  464  465  466  467  608
##  412  417  420  421  446  449  450  451  452  462  463  464  465  466  467  608
##  663  668  681  710  756  757  770  776  778  852  853  854  892  893  899  919
##  663  668  681  710  756  757  770  776  778  852  853  854  892  893  899  919
## 1136 1137 1138 1139 1140 1141 1142 1199 1219 1222 1225 1285 1292 1295 1326 1340
## 1136 1137 1138 1139 1140 1141 1142 1199 1219 1222 1225 1285 1292 1295 1326 1340
## 1342 1428 1429 1430 1431 1448 1449 1450 1451 1452 1453 1460 1461 1491 1554 1577
## 1342 1428 1429 1430 1431 1448 1449 1450 1451 1452 1453 1460 1461 1491 1554 1577
## 1578 1579 1580 1581 1582 1673 1719 1788 1790 1882 1884 1890 1892 1894 1896 1897
## 1578 1579 1580 1581 1582 1673 1719 1788 1790 1882 1884 1890 1892 1894 1896 1897
## 1907 1946 1966 2009 2011 2013 2014 2015 2101 2105 2191 2198 2199 2200 2203 2288
## 1907 1946 1966 2009 2011 2013 2014 2015 2101 2105 2191 2198 2199 2200 2203 2288
```

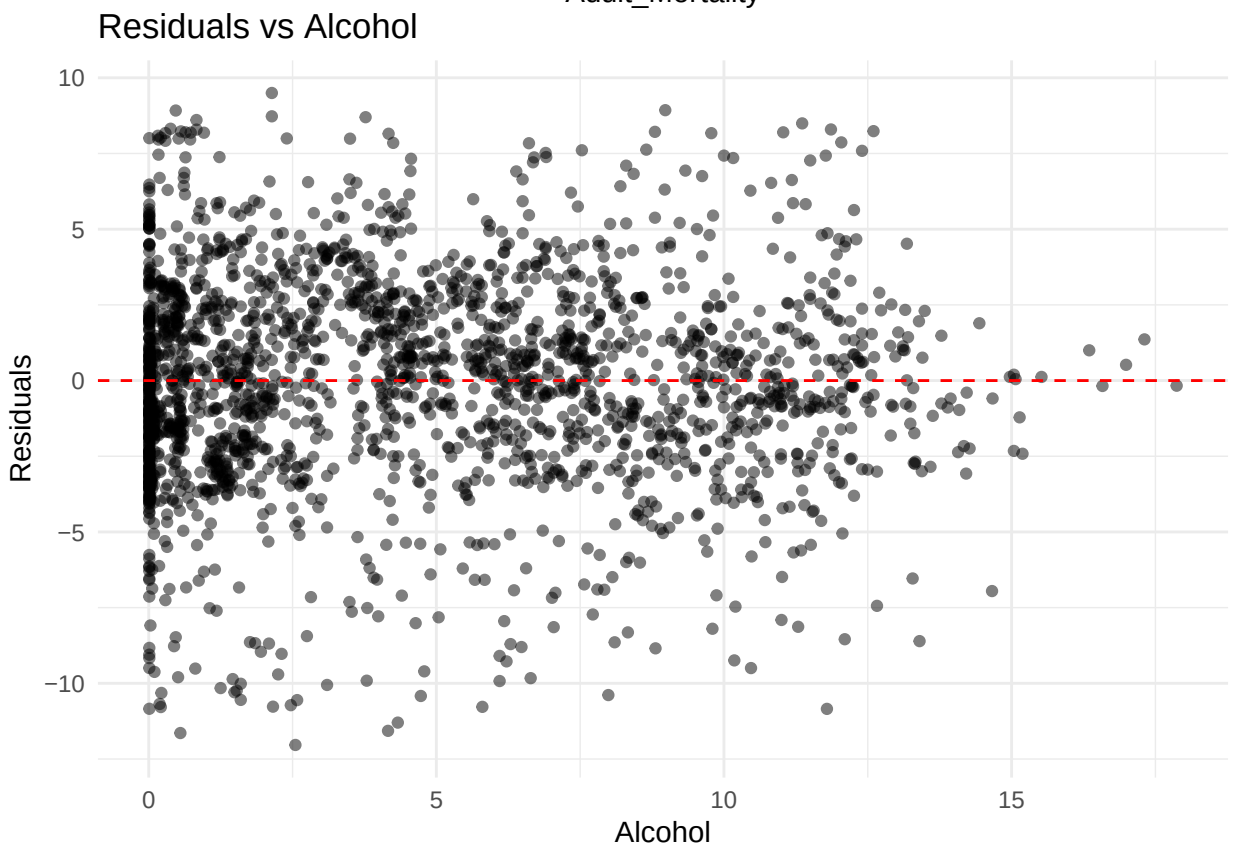
```
## 2293 2294 2304 2305 2309
## 2293 2294 2304 2305 2309
```

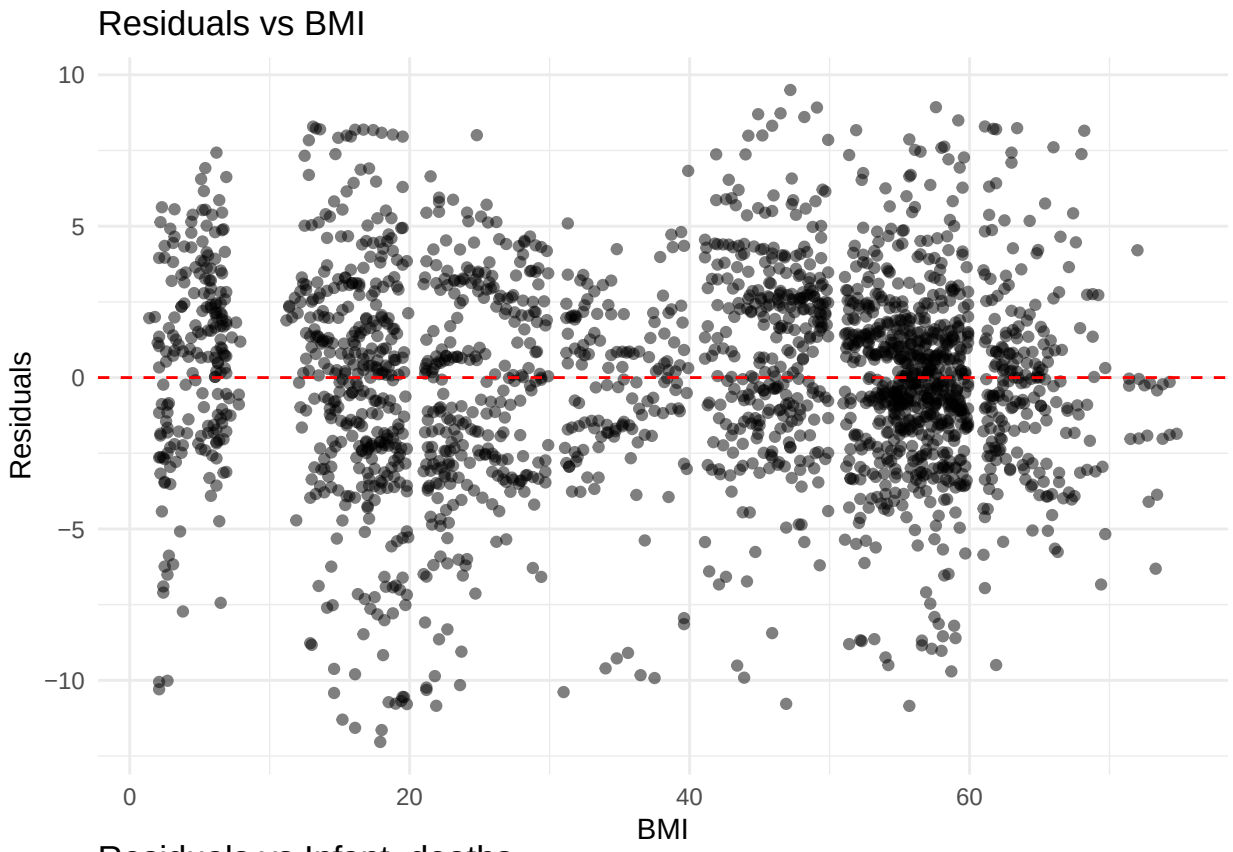
```
life = life[-potential_outliers,]
#Remove thinness_5_9 due to high vif and non significant p value
life$Thinness_5_9_years = NULL
model_2 = lm(data = life, life$Life_expectancy ~ .)
summary(model_2)
```

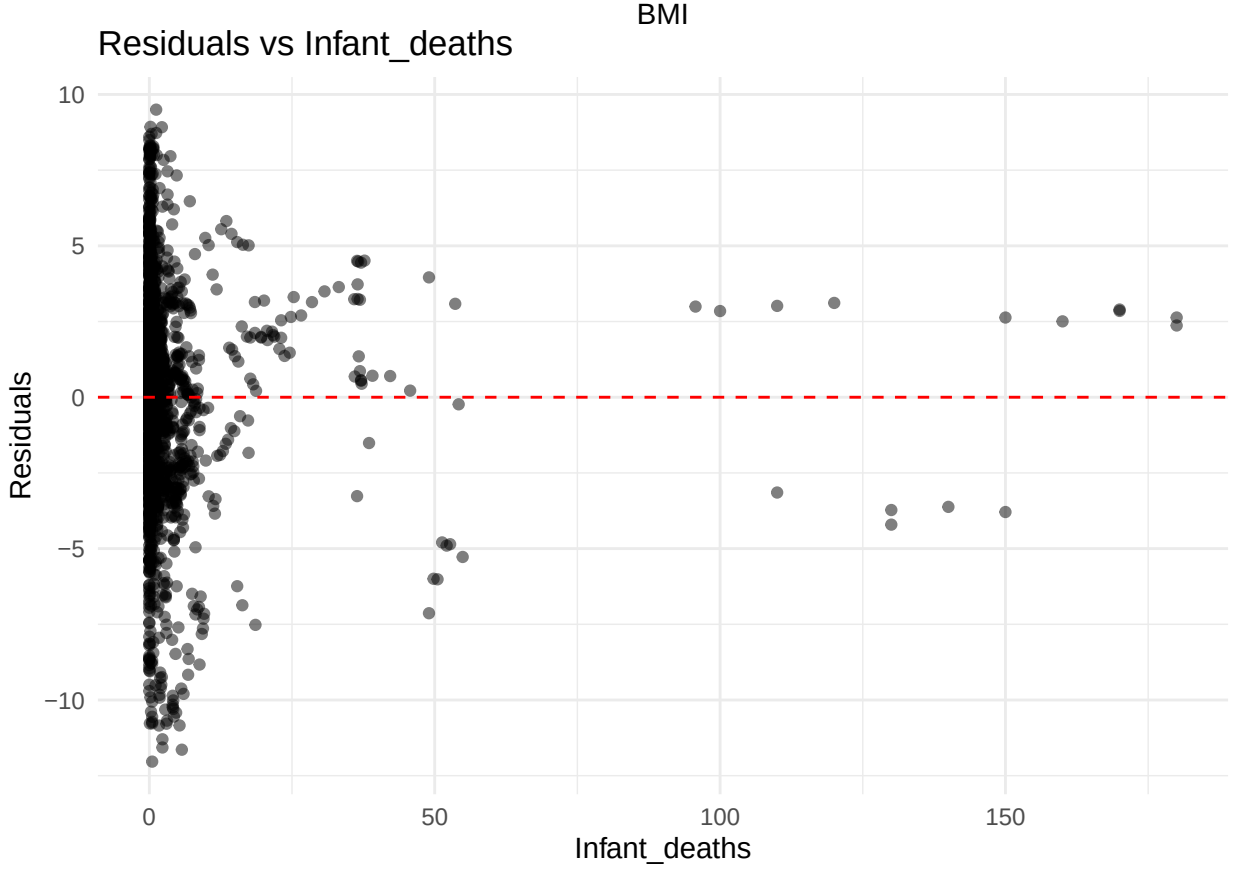
```
##
## Call:
## lm(formula = life$Life_expectancy ~ ., data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0356  -2.0510   0.0906   2.2237   9.4981
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.052e+01  5.867e-01 103.160  < 2e-16 ***
## StatusDeveloping      -1.104e+00  2.575e-01  -4.287 1.89e-05 ***
## Adult_Mortality       -3.517e-01  7.147e-03 -49.202  < 2e-16 ***
## Infant_deaths          6.538e-03  6.524e-03   1.002    0.316
## Alcohol               -1.281e-01  2.566e-02  -4.993 6.41e-07 ***
## BMI                    2.185e-02  4.951e-03   4.414 1.06e-05 ***
## Thinness_10_19_years -1.498e-01  2.288e-02  -6.545 7.37e-11 ***
## Schooling              9.438e-01  4.335e-02  21.770  < 2e-16 ***
## GDP                    2.913e-05  5.930e-06   4.913 9.62e-07 ***
## Income                 7.331e+00  6.132e-01  11.955  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.414 on 2184 degrees of freedom
## Multiple R-squared:  0.862,  Adjusted R-squared:  0.8614
## F-statistic:  1516 on 9 and 2184 DF,  p-value: < 2.2e-16
```

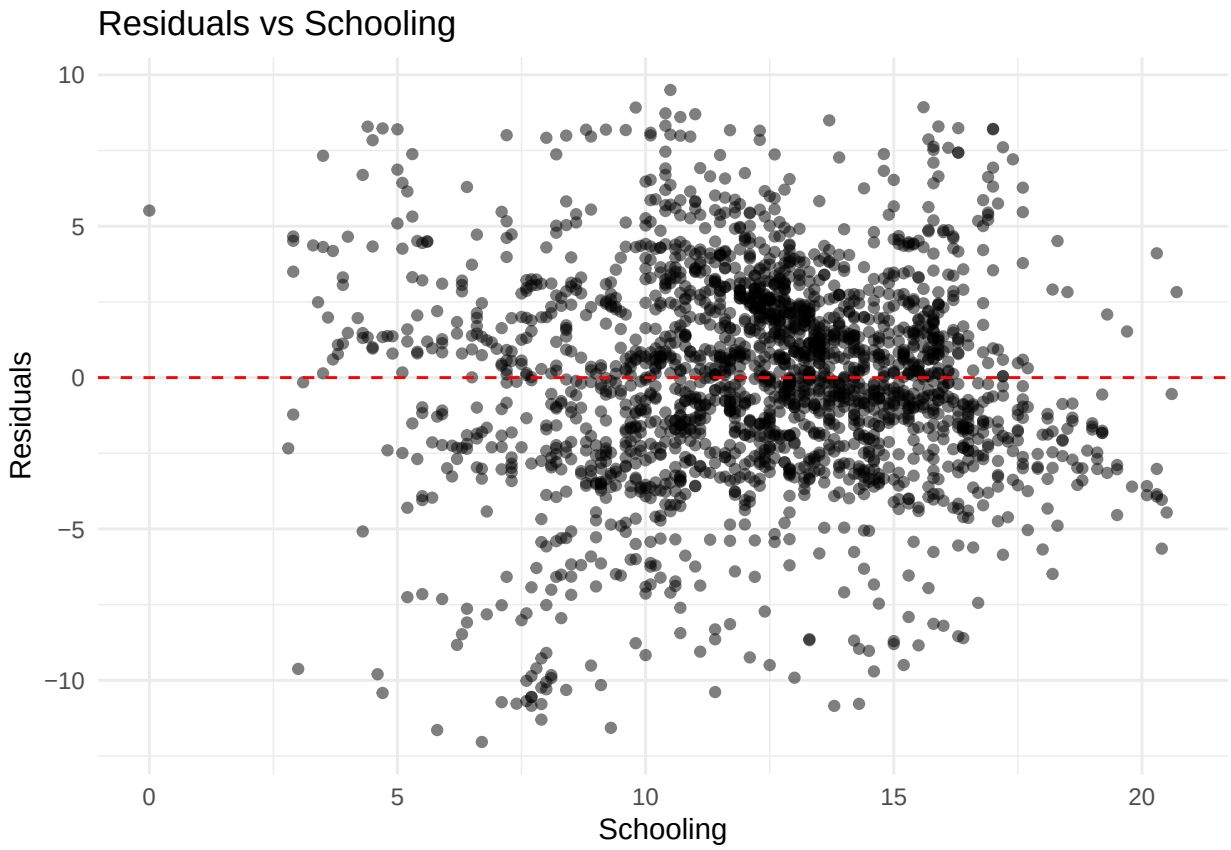
```
#Residuals vs. x
fitted_vals2 = fitted(model_2)
residuals2 = resid(model_2)

predictors2 = c("Adult_Mortality", "Alcohol", "BMI", "Infant_deaths",
                "Schooling", "Thinness_10_19_years", "GDP", "Income")

for (var in predictors2) {
  p = ggplot(life, aes(x = .data[[var]], y = residuals2)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
    ggtitle(paste("Residuals vs", var)) +
    theme_minimal() +
    ylab("Residuals") +
    xlab(var)

  print(p)
}
```

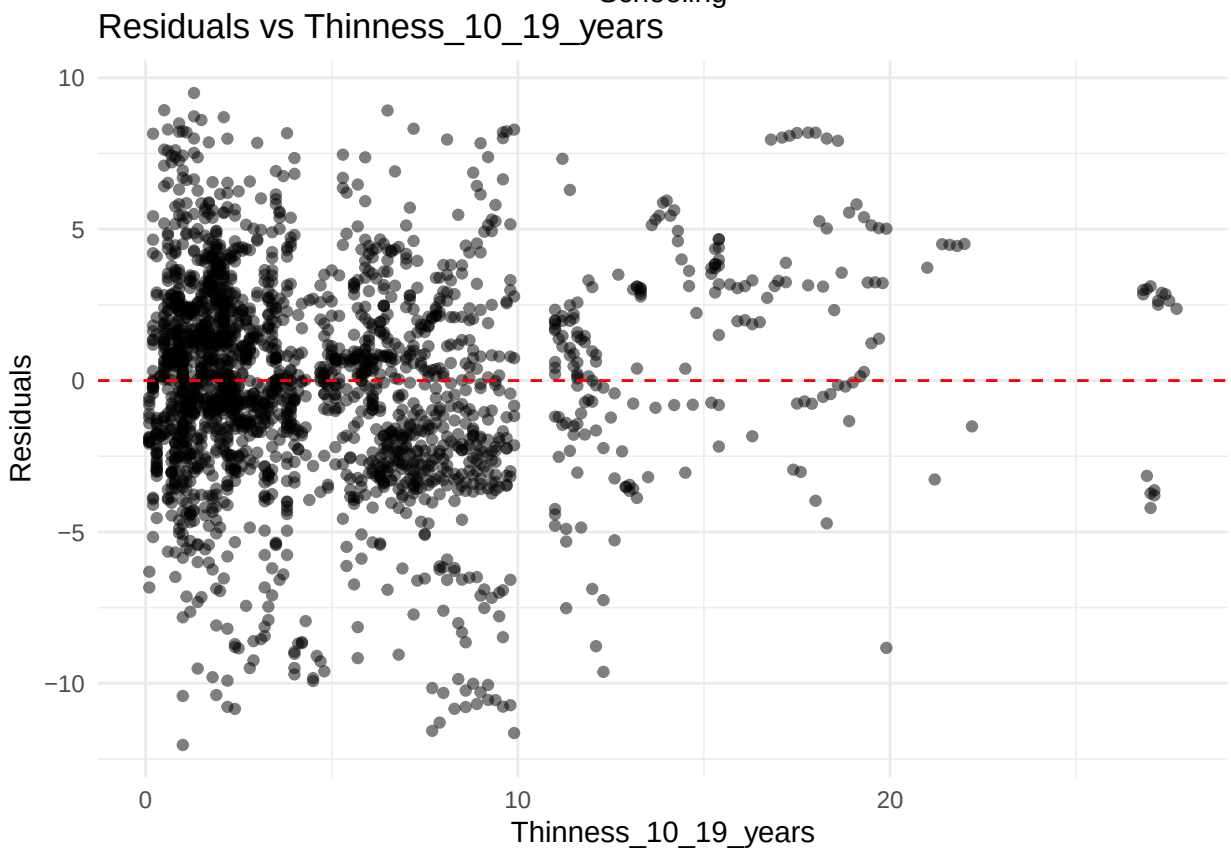

Residuals vs Adult_Mortality
Residuals
Adult_Mortality
10
5
0
−5
−10
0
20
40
60

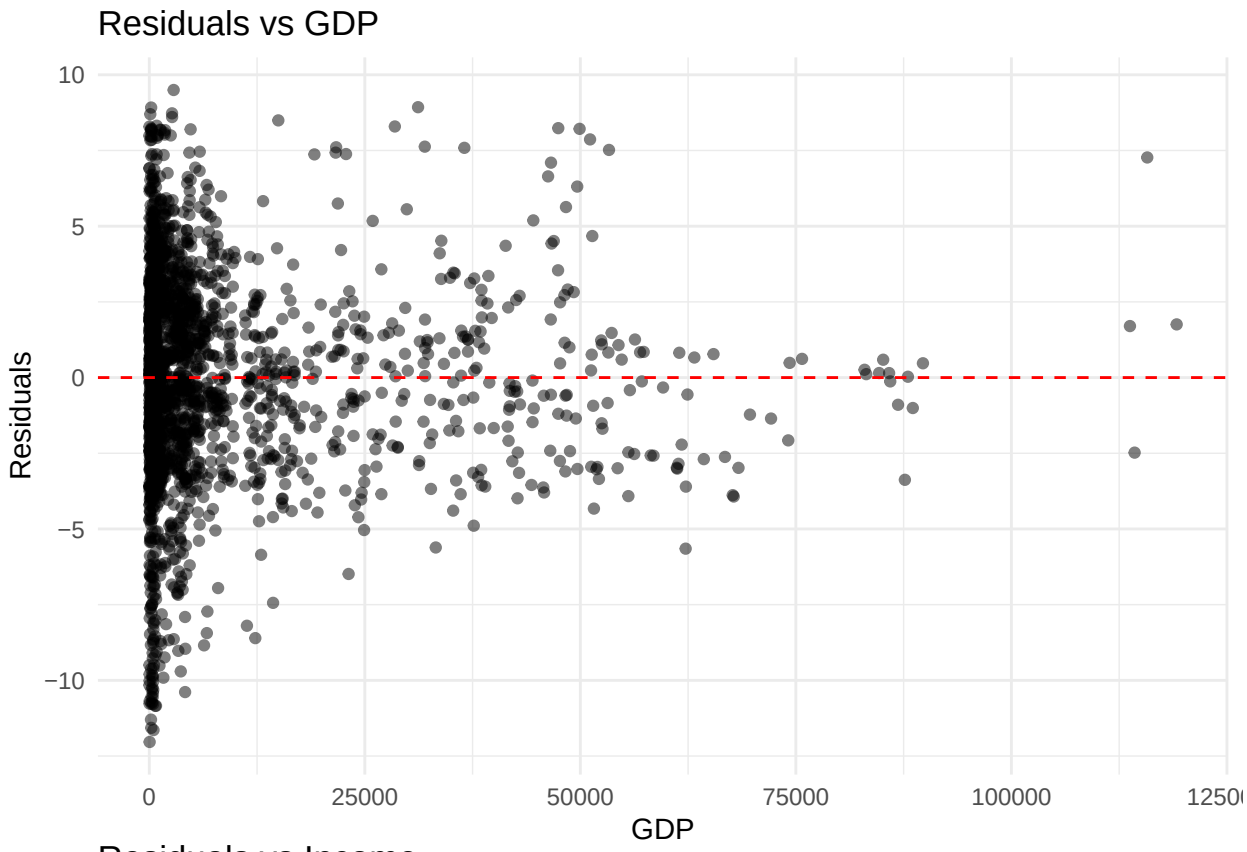

Residuals vs Alcohol
Residuals
Alcohol
10
5
0
−5
−10
0
5
10
15

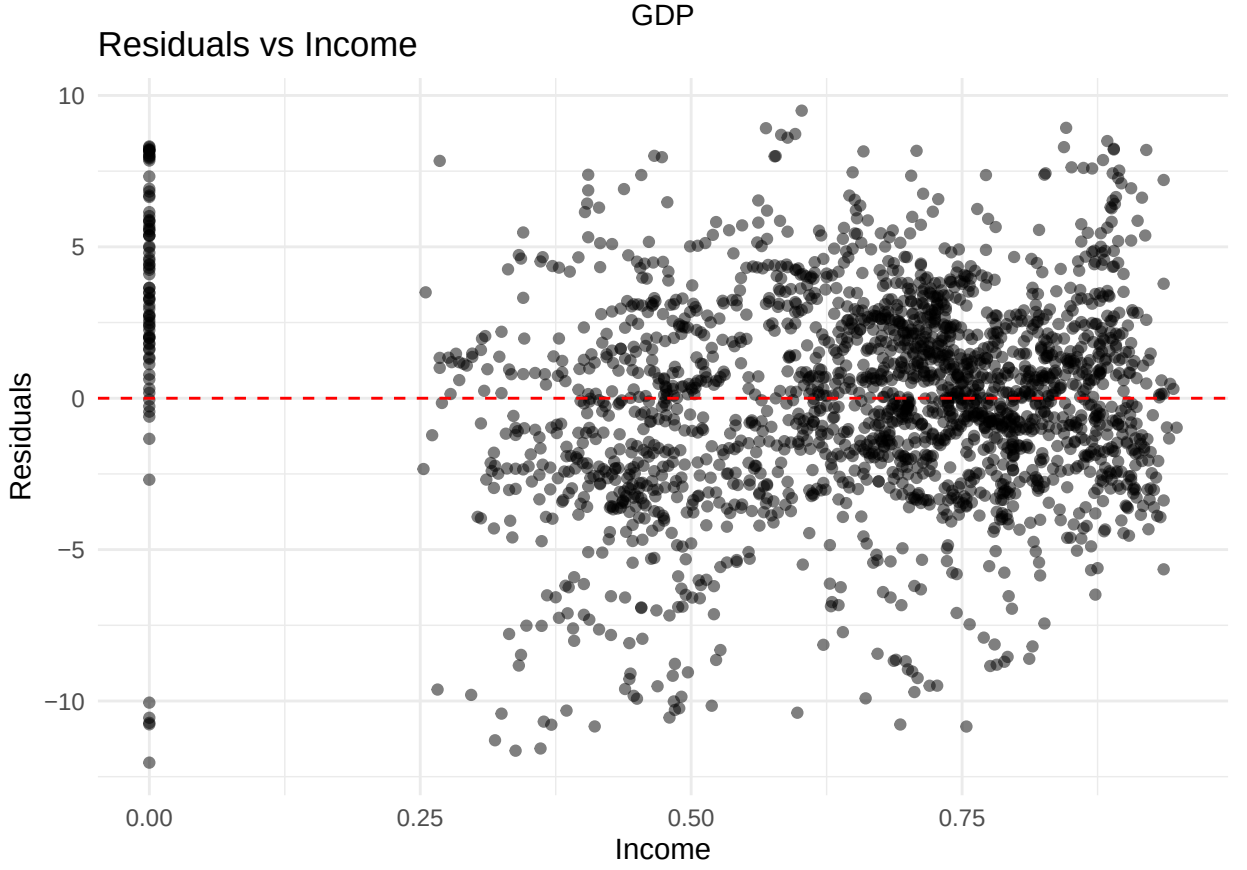

Residuals vs BMI
Residuals
BMI
10
5
0
−5
−10
0
20
40
60


Residuals vs Infant_deaths
Residuals
Infant_deaths
10
5
0
−5
−10
0
50
100
150

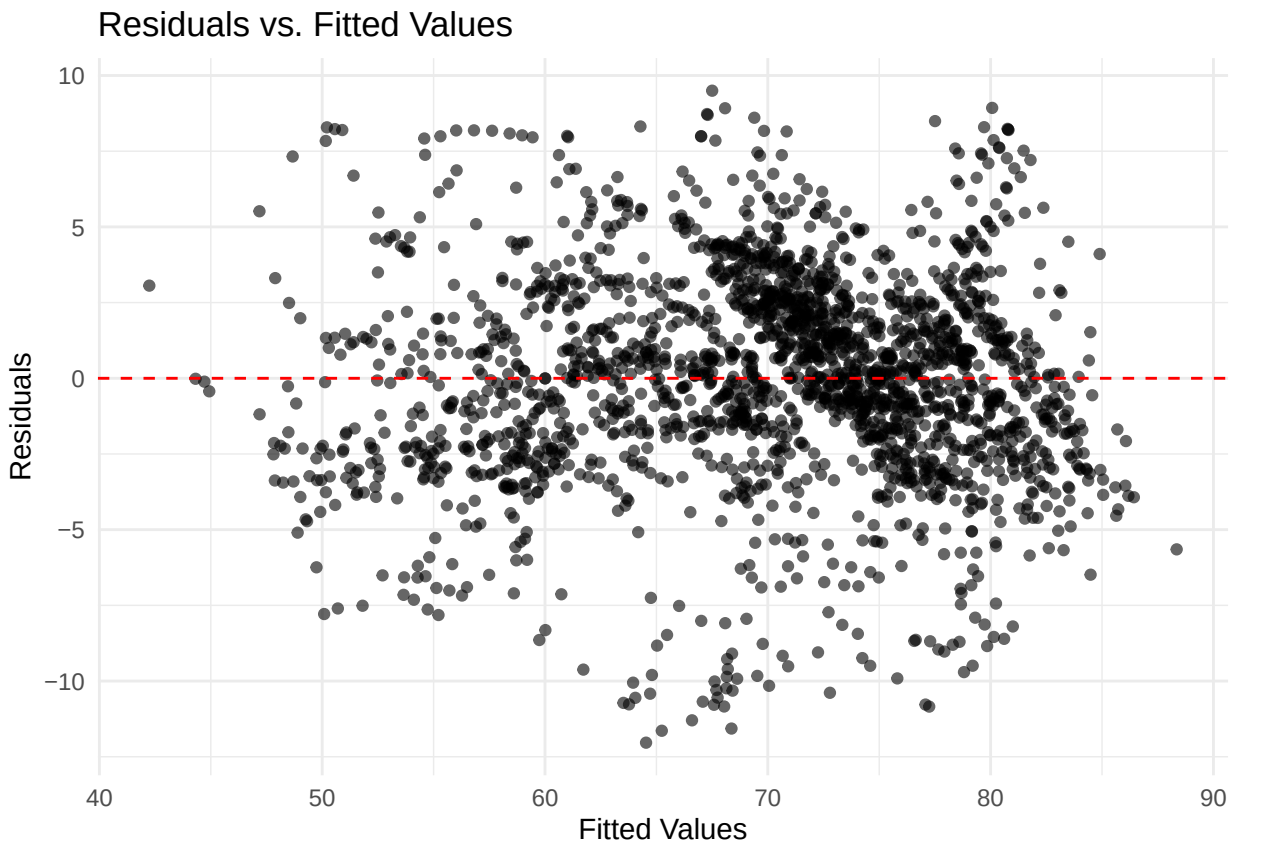

Residuals vs Schooling
Residuals
Schooling
10
5
0
−5
−10
0
5
10
15
20


Residuals vs Thinness_10_19_years
Residuals
Thinness_10_19_years
10
5
0
−5
−10
0
10
20


Residuals vs GDP
Residuals
GDP
10
5
0
−5
−10
0
25000
50000
75000
100000


Residuals vs Income
Residuals
Income
10
5
0
−5
−10
0.00
0.25
0.50
0.75

```r
# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals2, y = residuals2)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```r
#Linear

#Homoscedasticity
bptest(model_2)
```

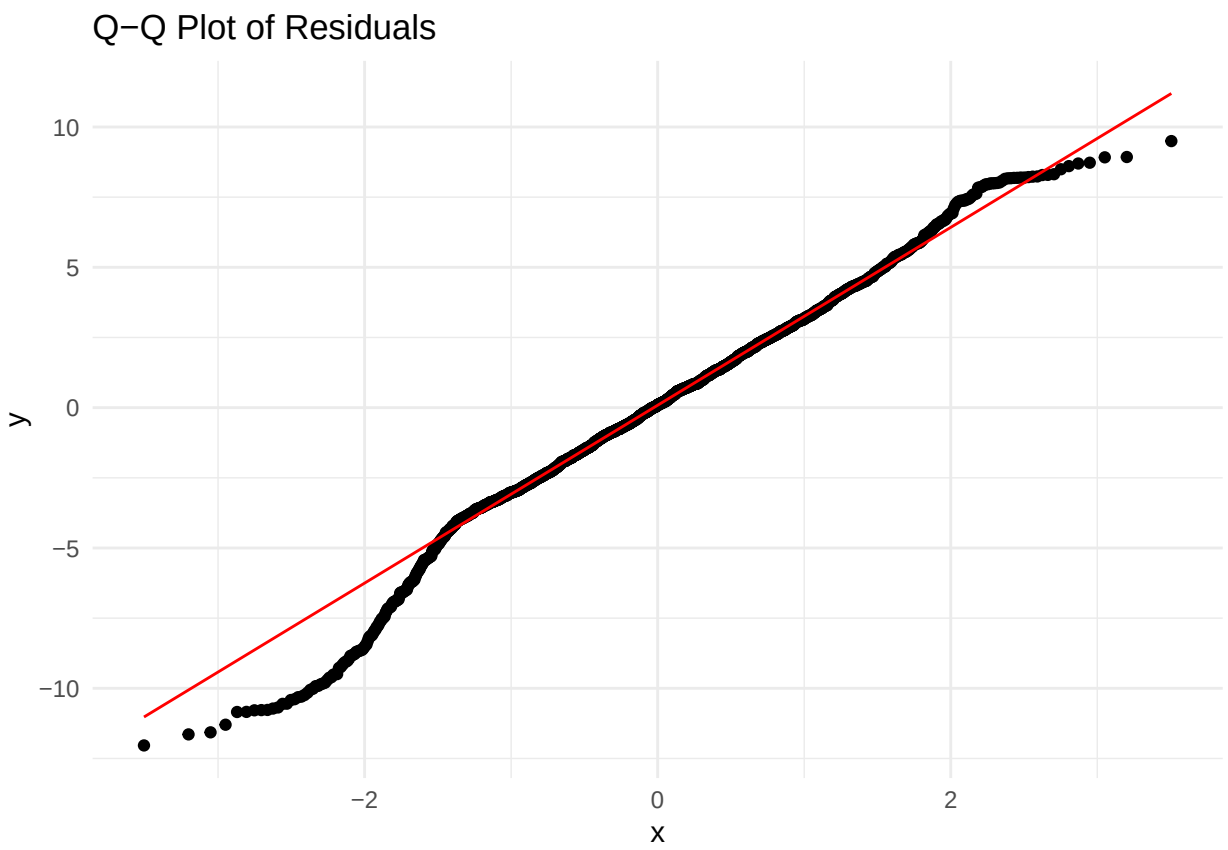
```
##
##  studentized Breusch-Pagan test
##
## data:  model_2
## BP = 337.65, df = 9, p-value < 2.2e-16
```

```r
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality, since this is a large sample test, the Shapiro Wilk test will be sensitive, we will use And
goftest::ad.test(residuals2)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
```

```
## data:  residuals2
## An = Inf, p-value = 2.735e-07
```

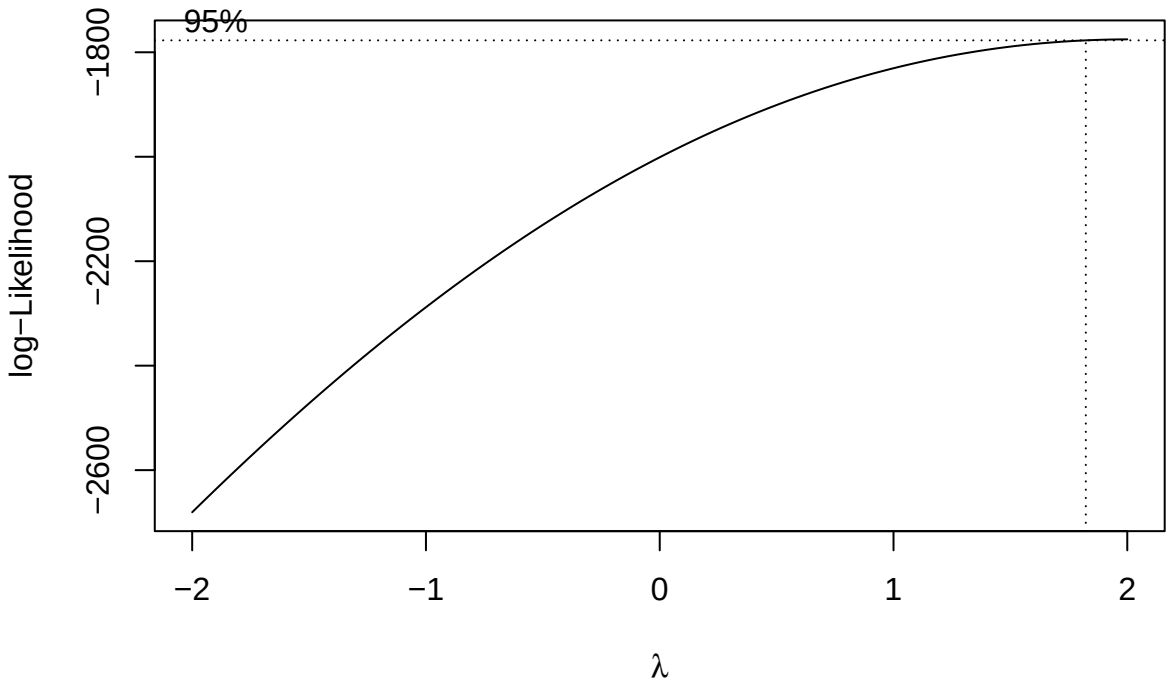
```
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals2)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()
```




```
#Not normally distributed
```

#Transformation

```
#Boxcox
b = boxcox(model_2)
```




```
lambda = b$x[which.max(b$y)]
lambda
```

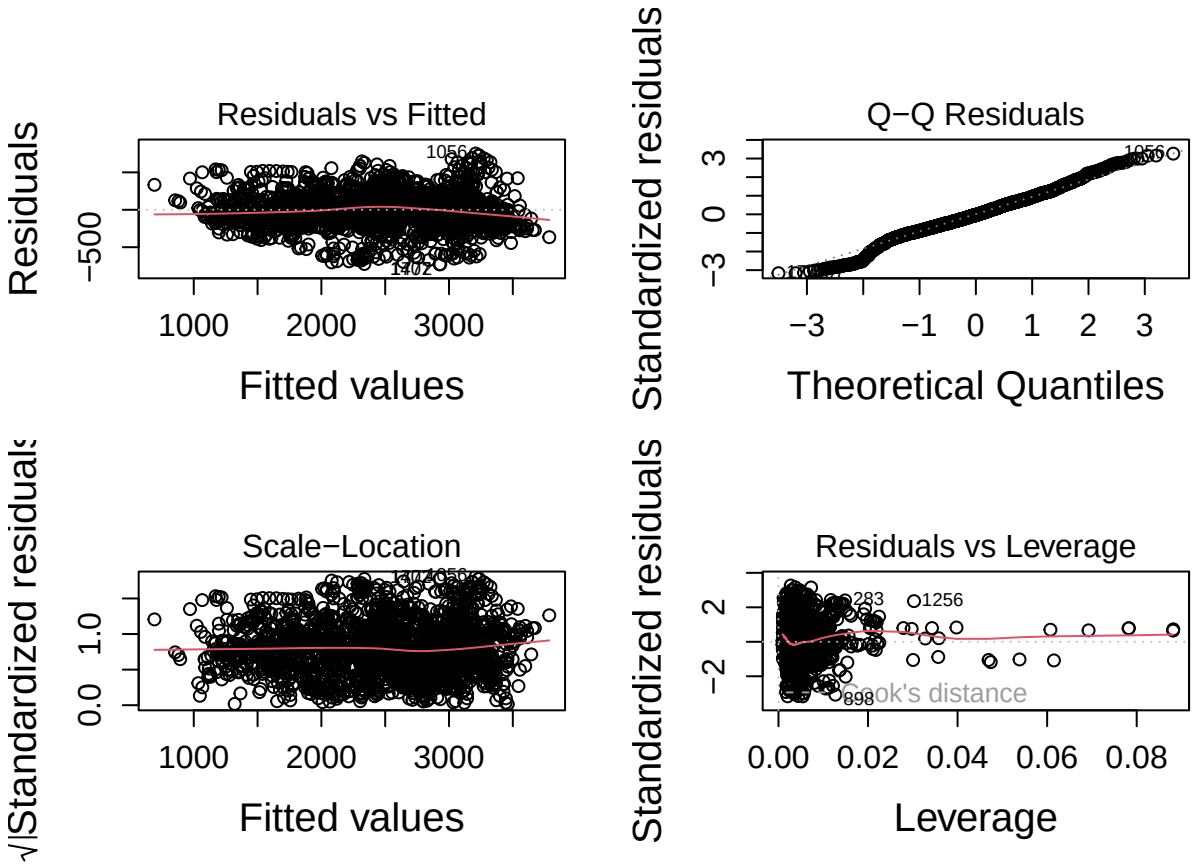
```
## [1] 2
```

```
life$Life_expectancy_bc = (life$Life_expectancy^2-1)/2

model_bc = lm(data = life, Life_expectancy_bc ~ .-Life_expectancy )
summary(model_bc)
```

```
##
## Call:
## lm(formula = Life_expectancy_bc ~ . - Life_expectancy, data = life)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -728.90 -141.55   -7.96  147.44  751.16
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.846e+03  3.963e+01  46.585  < 2e-16 ***
## StatusDeveloping       -1.157e+02  1.739e+01  -6.651 3.67e-11 ***
## Adult_Mortality        -2.166e+01  4.828e-01 -44.868  < 2e-16 ***
## Infant_deaths           4.661e-01  4.407e-01   1.058     0.29
## Alcohol                -8.742e+00  1.733e+00  -5.044 4.94e-07 ***
## BMI                     1.358e+00  3.344e-01   4.060 5.08e-05 ***
## Thinness_10_19_years -1.093e+01  1.546e+00  -7.072 2.05e-12 ***
## Schooling               6.387e+01  2.929e+00  21.810  < 2e-16 ***
## GDP                     2.937e-03  4.006e-04   7.332 3.18e-13 ***
## Income                  5.004e+02  4.142e+01  12.079  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

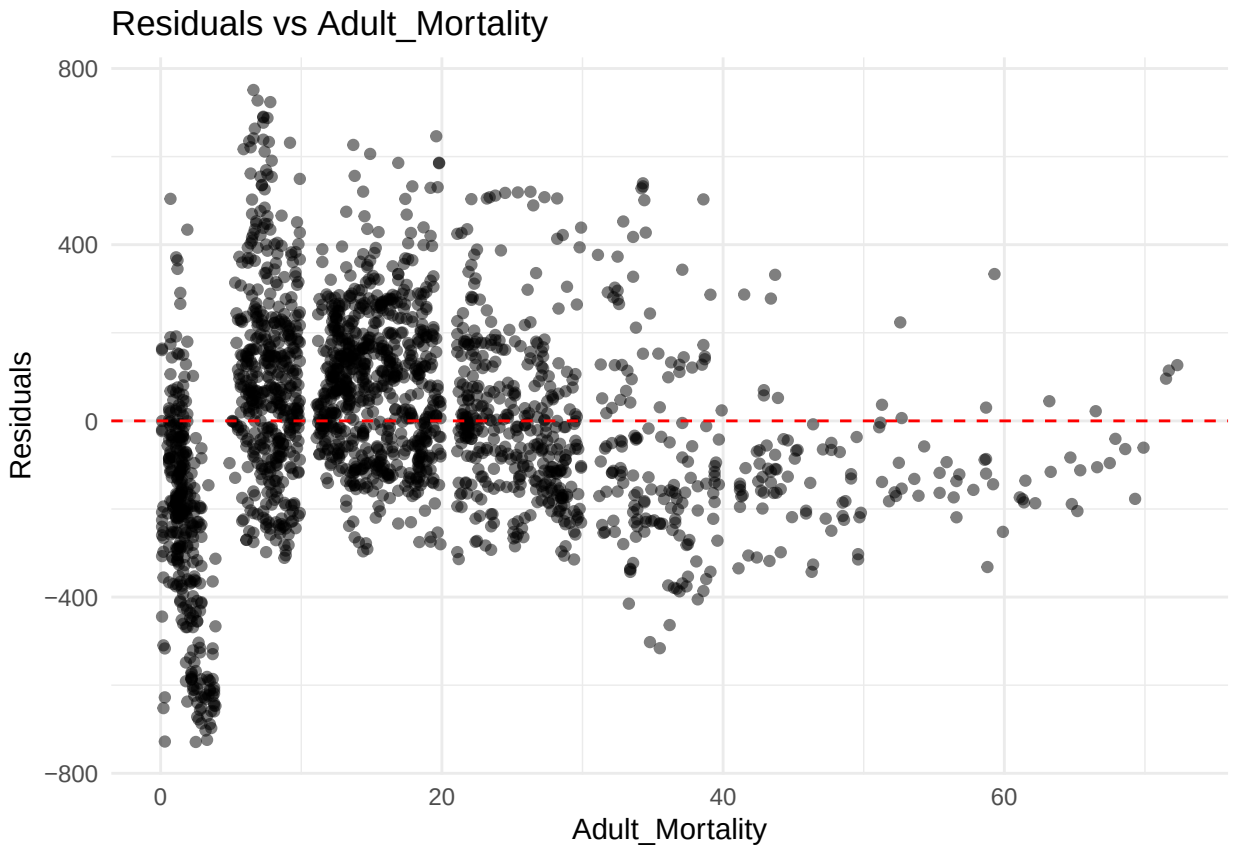
```
## Residual standard error: 230.6 on 2184 degrees of freedom
## Multiple R-squared:  0.8599, Adjusted R-squared:  0.8593
## F-statistic:  1489 on 9 and 2184 DF,  p-value: < 2.2e-16
```

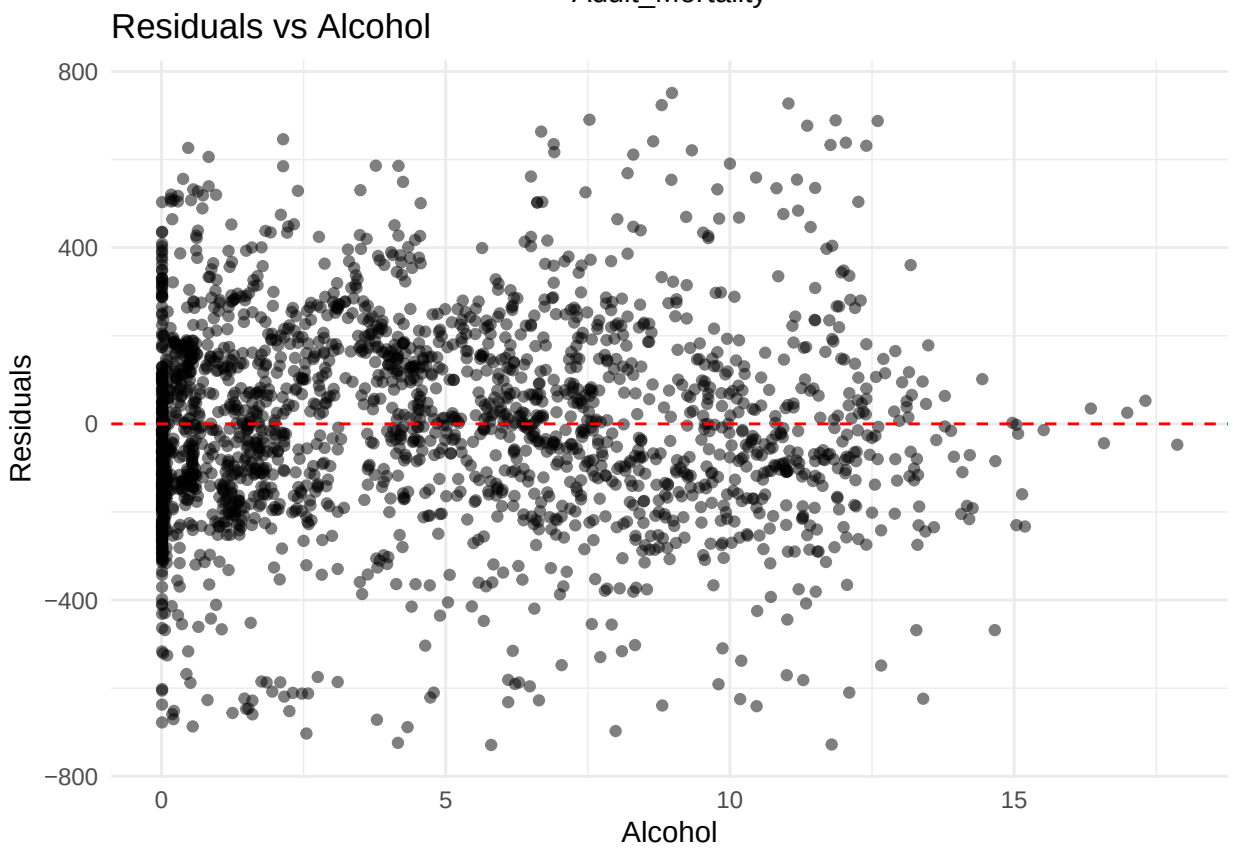
```r
par(mfrow = c(2,2))
plot(model_bc, cex.axis = 1.2, cex.lab = 1.5)
```

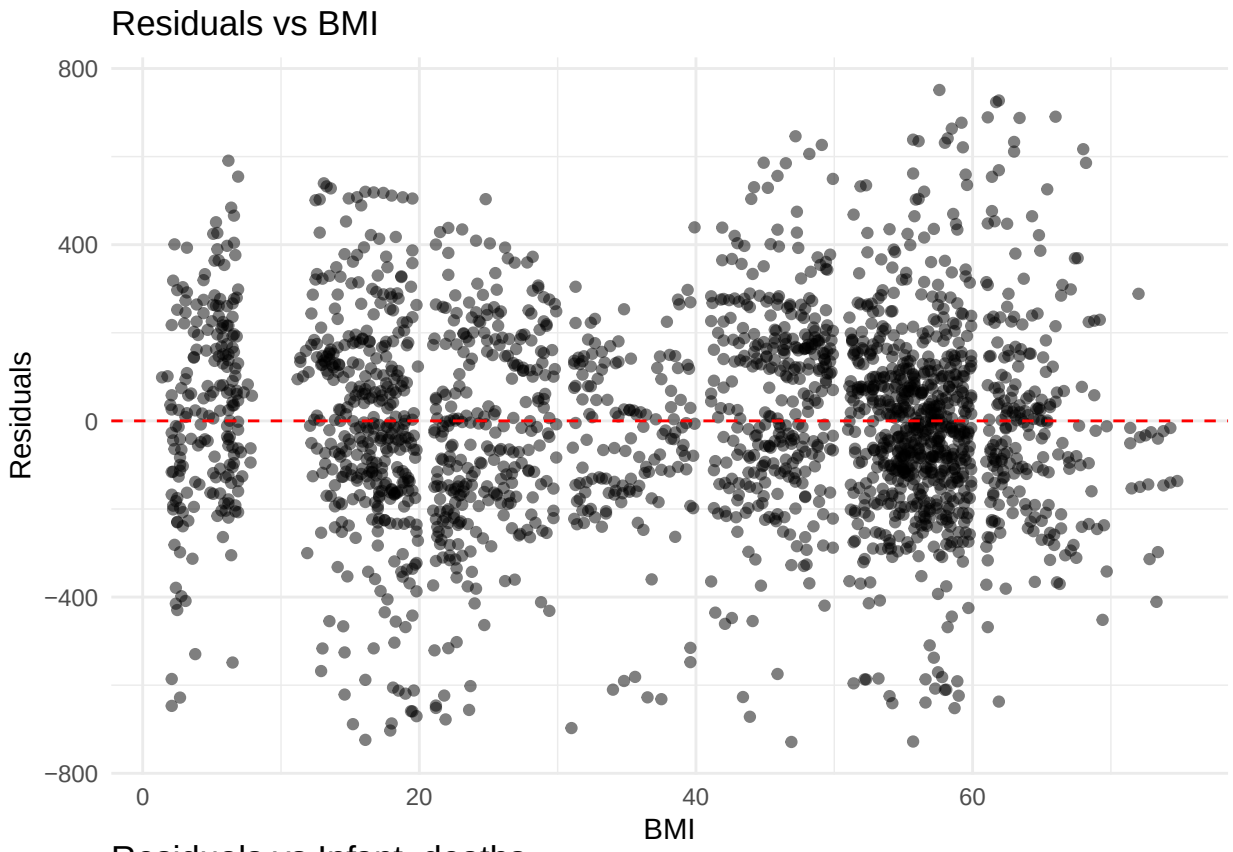


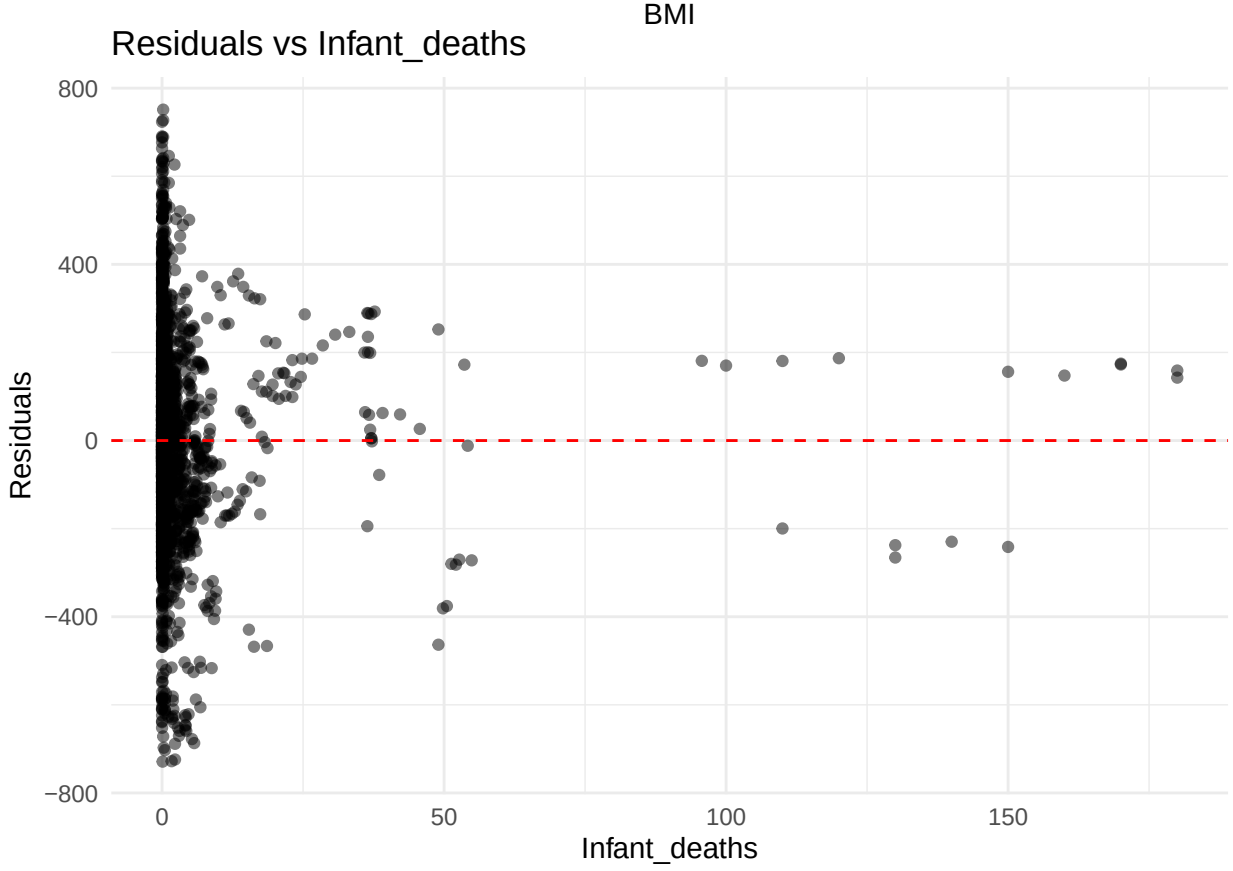
#Diagnostics

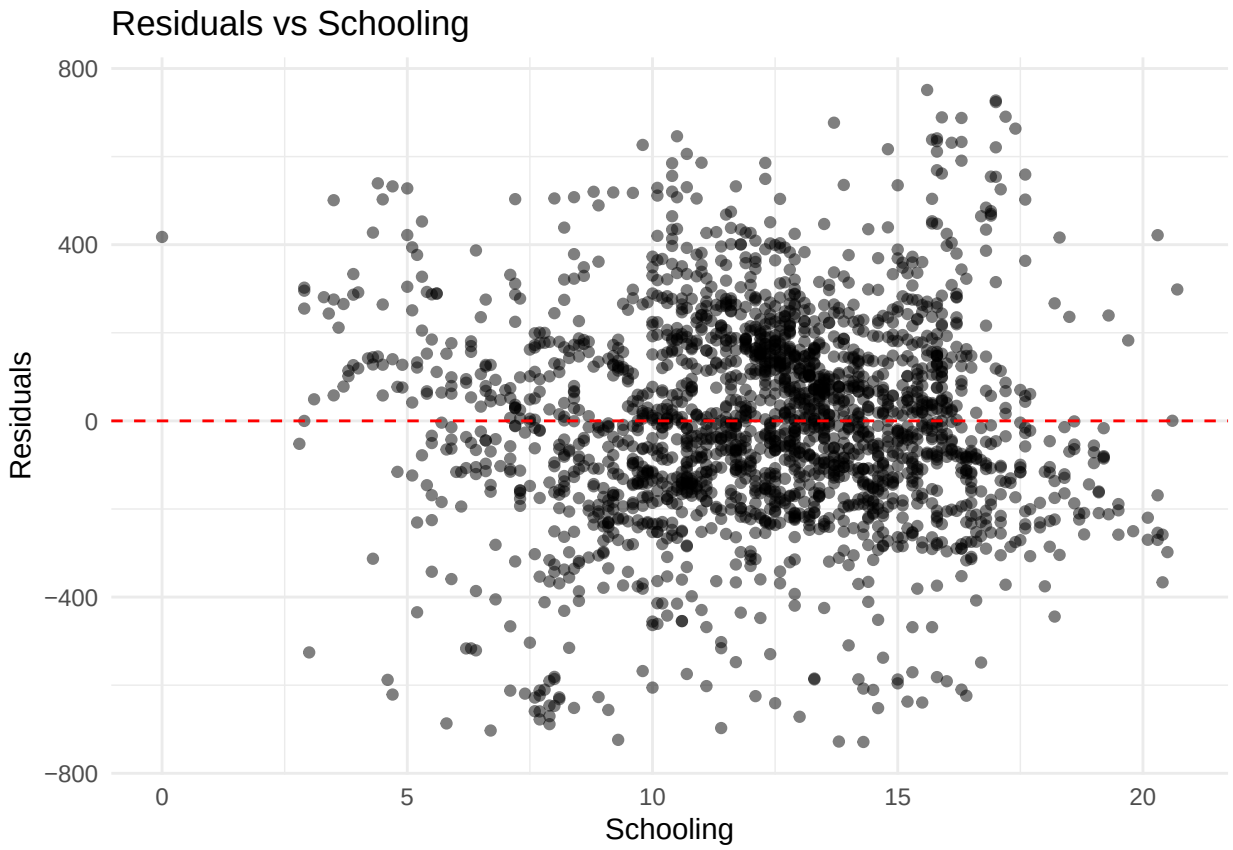
```r
#Linearity
fitted_vals_bc = fitted(model_bc)
residuals_bc = resid(model_bc)

for (var in predictors2) {
  p = ggplot(life, aes(x = .data[[var]], y = residuals_bc)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red")+
    ggtitle(paste("Residuals vs", var)) +
    theme_minimal() +
    ylab("Residuals") +
    xlab(var)

  print(p)
}
```

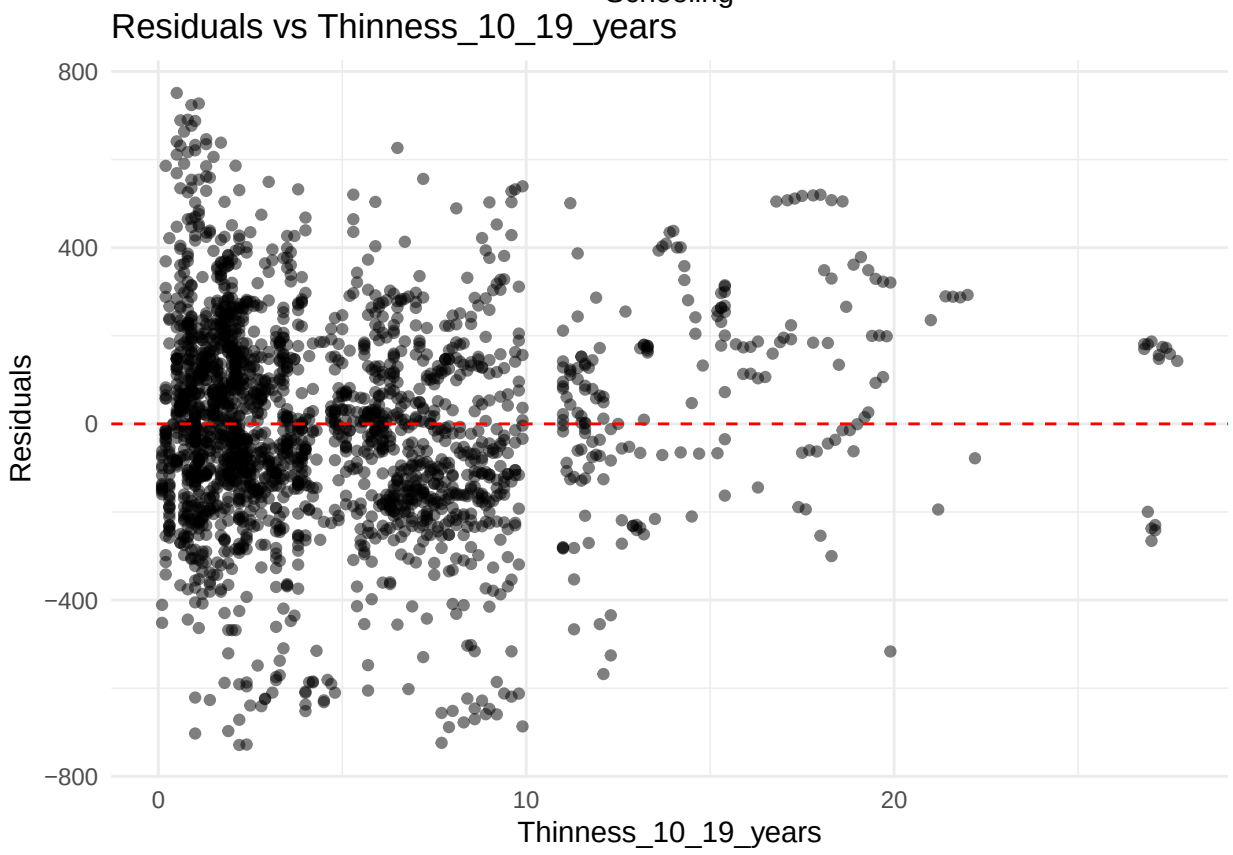

Residuals vs Adult_Mortality
800
400
0
−400
−800
Residuals
0
20
40
60
Adult_Mortality

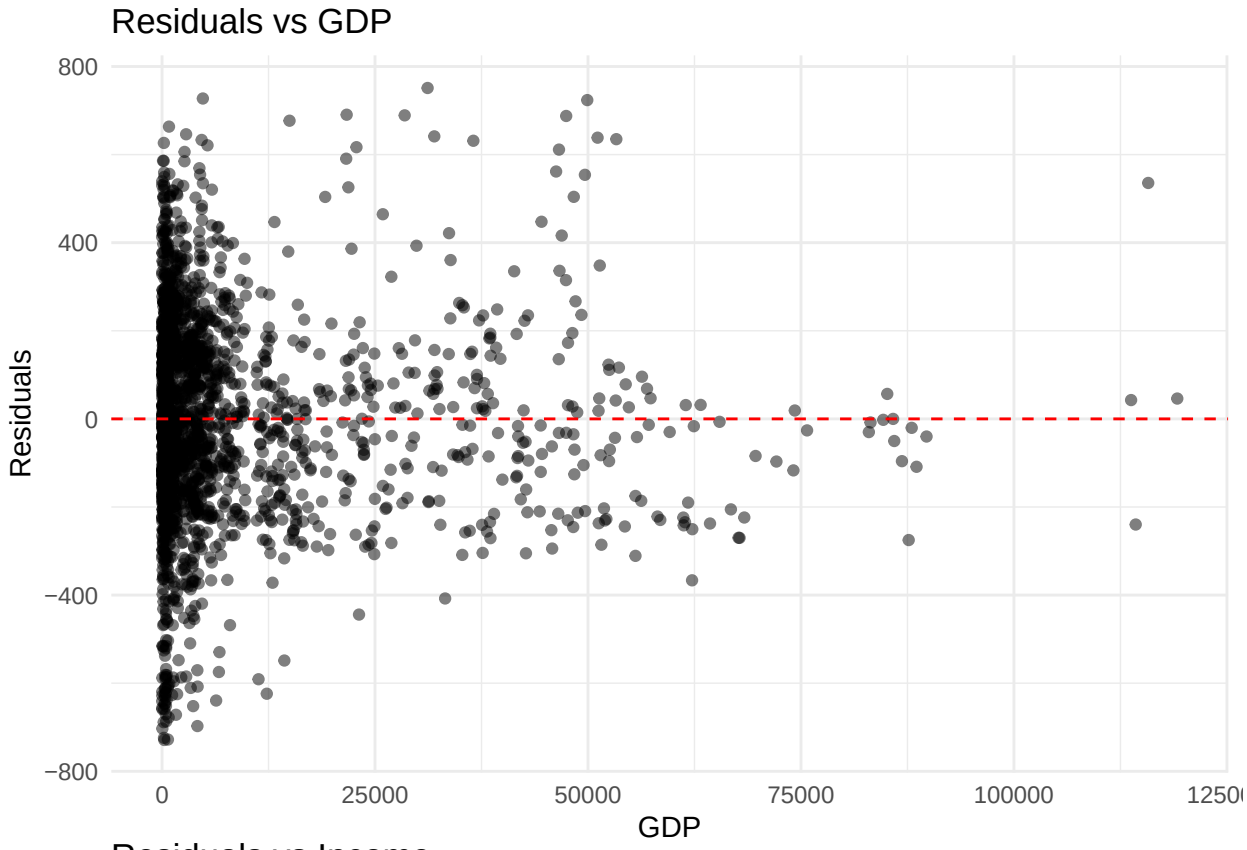

Residuals vs Alcohol
800
400
0
−400
−800
Residuals
0
5
10
15
Alcohol

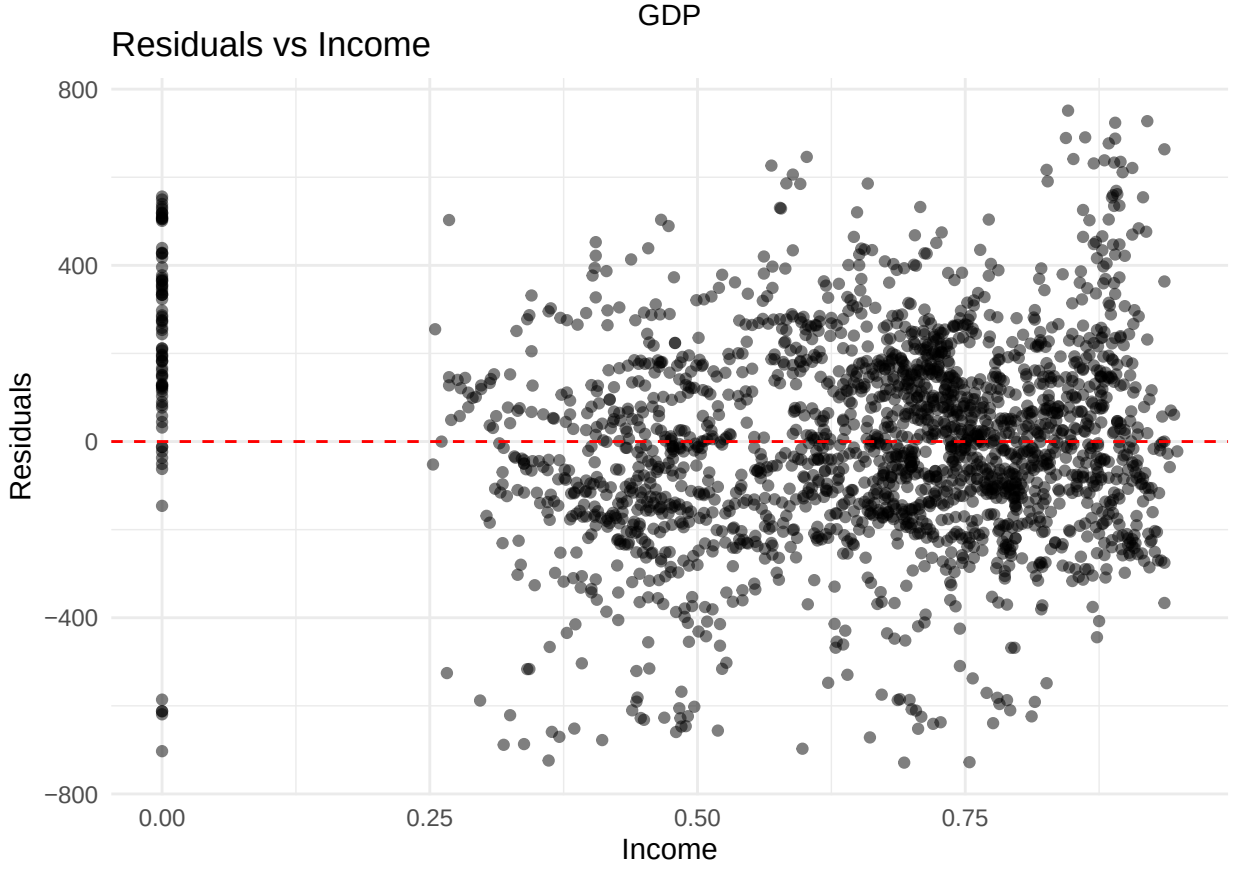

Residuals vs BMI
800
400
0
−400
−800
Residuals
0
20
40
60
BMI


Residuals vs Infant_deaths
800
400
0
−400
−800
Residuals
0
50
100
150
Infant_deaths

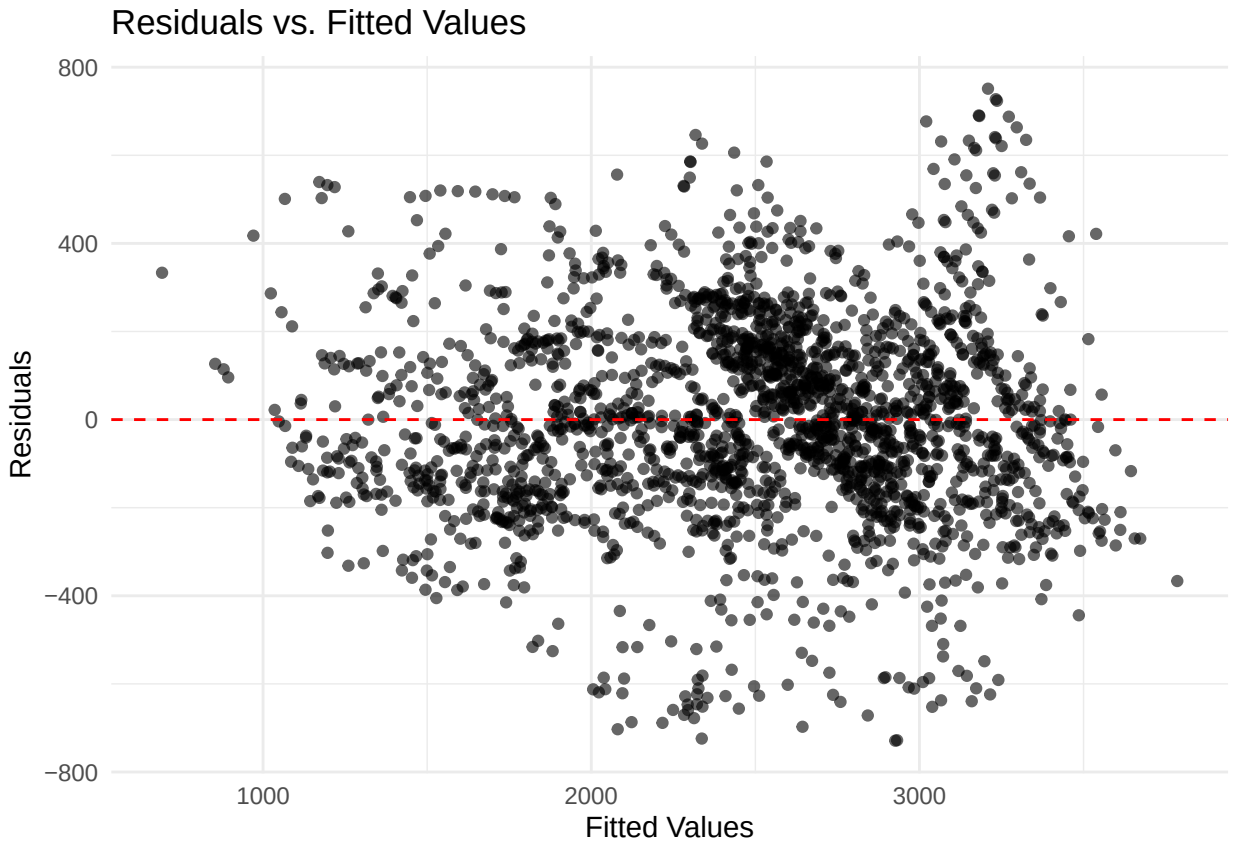

Residuals vs Schooling
800
400
0
−400
−800
Residuals
0
5
10
15
20
Schooling


Residuals vs Thinness_10_19_years
800
400
0
−400
−800
Residuals
0
10
20
Thinness_10_19_years


Residuals vs GDP
800
400
0
−400
−800
Residuals
0
25000
50000
75000
100000
GDP


Residuals vs Income
800
400
0
−400
−800
Residuals
0.00
0.25
0.50
0.75
Income

```r
# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals_bc, y = residuals_bc)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```r
#Linear

#Homoscedasticity
bptest(model_bc)
```

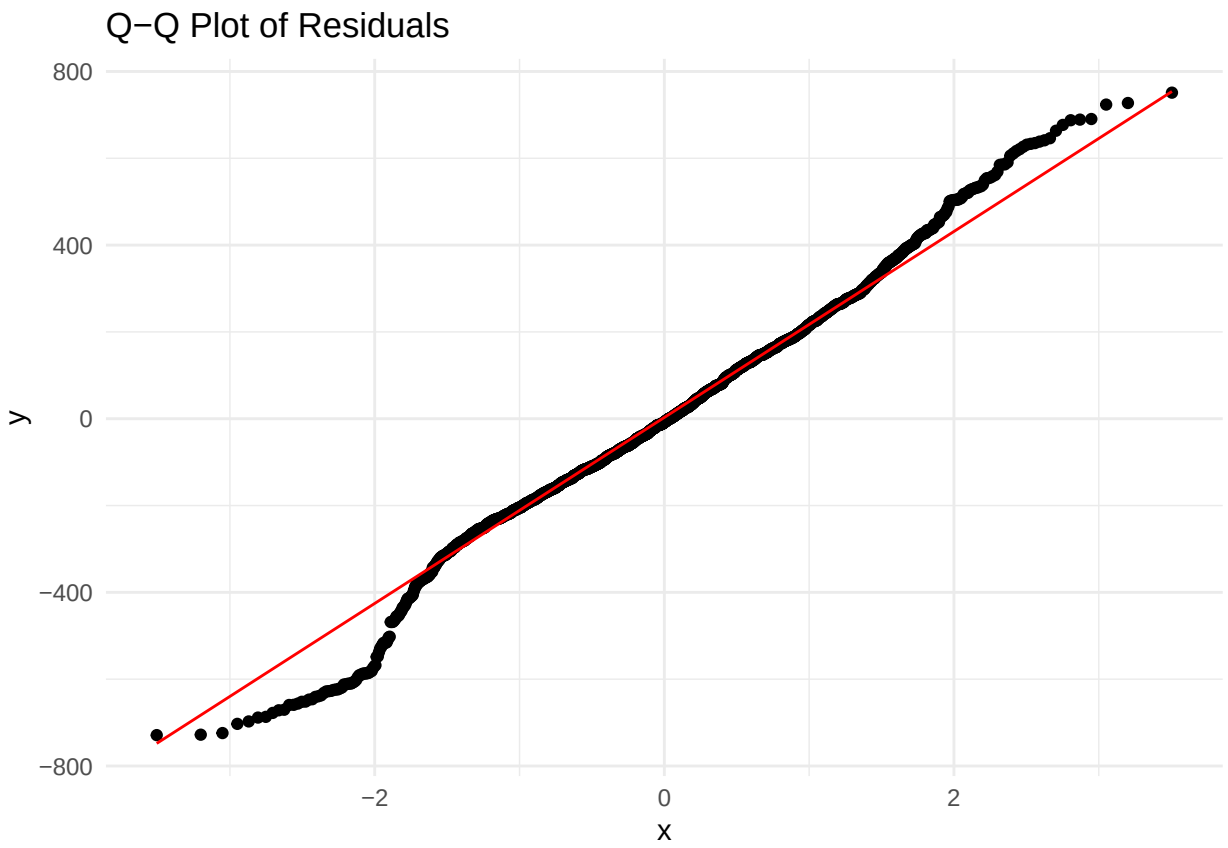
```
##
##  studentized Breusch-Pagan test
##
## data:  model_bc
## BP = 253.05, df = 9, p-value < 2.2e-16
```

```r
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality
goftest::ad.test(residuals_bc)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
```

```
## data:  residuals_bc
## An = Inf, p-value = 2.735e-07
```

```
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals_bc)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()
```




```
#Not normally distributed

#VIF
vif(model_bc)
```

```
##     StatusDeveloping      Adult_Mortality        Infant_deaths
##             1.931741             1.566780             1.342180
##              Alcohol                  BMI Thinness_10_19_years
##             2.030717             1.788878             2.022034
##            Schooling                  GDP               Income
##             3.597825             1.446324             2.952699
```

```
#No Multicollinearity
```

#Transformation again

```
life$Adult_Mortality_c = scale(life$Adult_Mortality, center = TRUE, scale = FALSE)
life$Adult_Mortality_sq = life$Adult_Mortality_c^2

model_3 = lm(
```

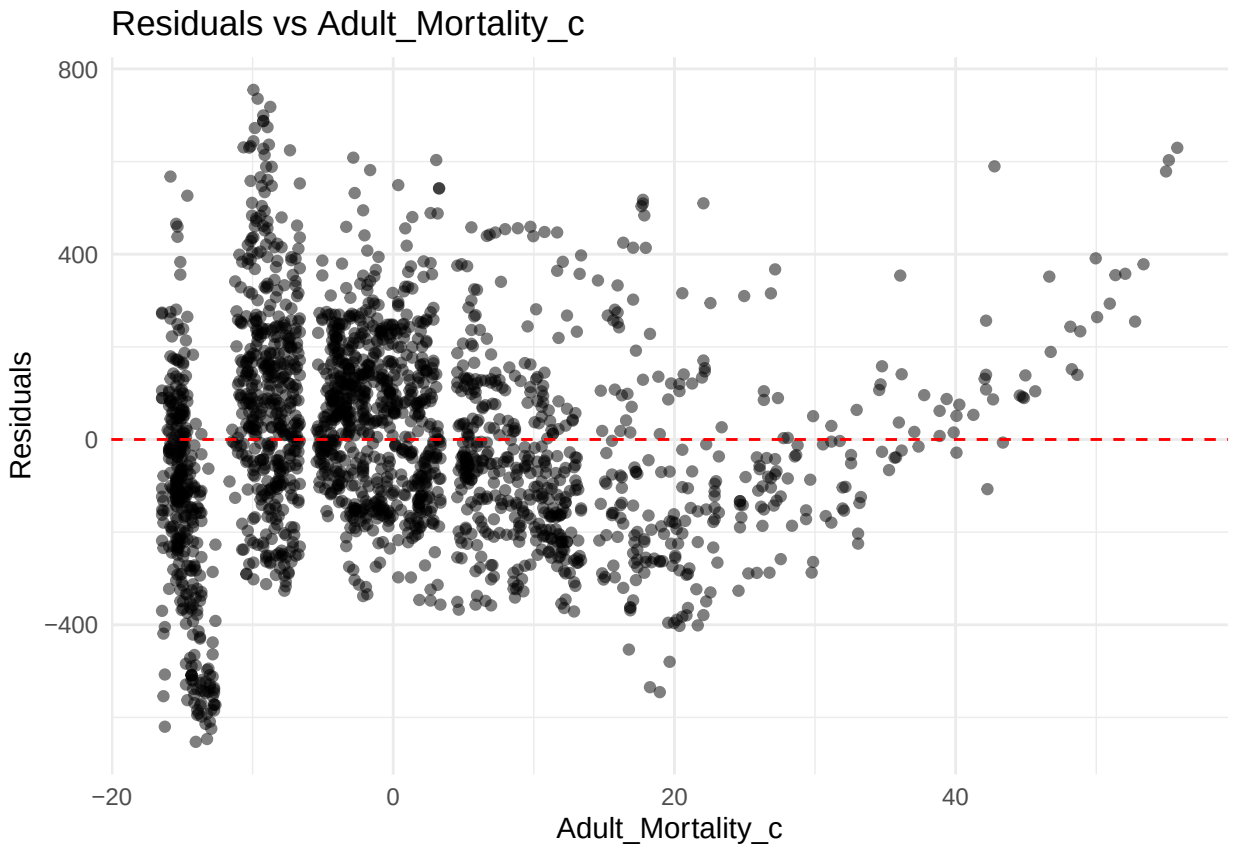
```
  data = life,
  Life_expectancy_bc ~
    Status +
    Adult_Mortality_c + Adult_Mortality_sq +
    BMI +
    Schooling +
    Infant_deaths +
    Alcohol +
    GDP +
    Income +
    Thinness_10_19_years
)
summary(model_3)
```

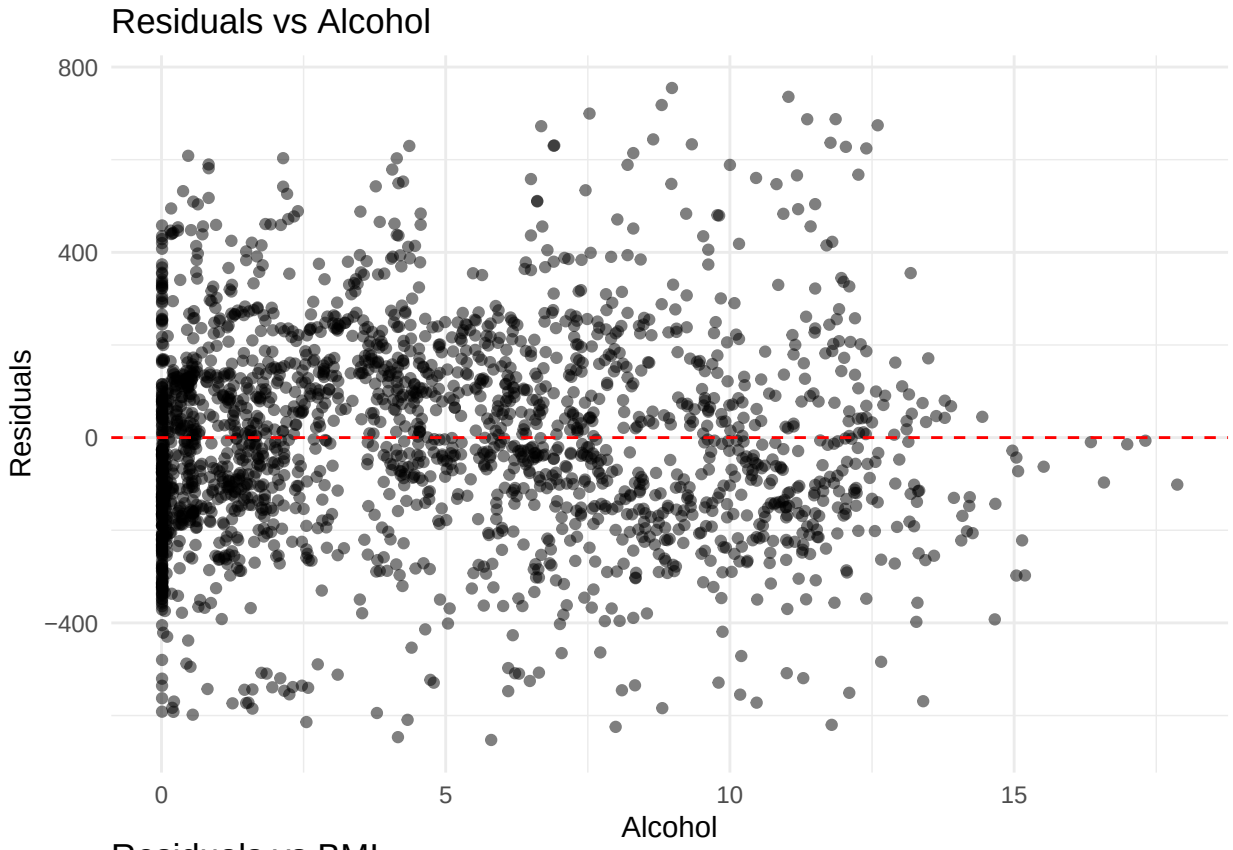
```
##
## Call:
## lm(formula = Life_expectancy_bc ~ Status + Adult_Mortality_c +
##     Adult_Mortality_sq + BMI + Schooling + Infant_deaths + Alcohol +
##     GDP + Income + Thinness_10_19_years, data = life)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -652.7 -146.2  -10.2  136.5  754.8
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.501e+03  3.549e+01  42.278  < 2e-16 ***
## StatusDeveloping       -1.303e+02  1.678e+01  -7.770 1.20e-14 ***
## Adult_Mortality_c      -1.700e+01  5.826e-01 -29.187  < 2e-16 ***
## Adult_Mortality_sq     -2.552e-01  1.925e-02 -13.252  < 2e-16 ***
## BMI                     1.443e+00  3.219e-01   4.482 7.76e-06 ***
## Schooling               6.683e+01  2.827e+00  23.641  < 2e-16 ***
## Infant_deaths           3.448e-01  4.241e-01   0.813    0.416
## Alcohol                -9.330e+00  1.668e+00  -5.593 2.52e-08 ***
## GDP                     3.357e-03  3.867e-04   8.681  < 2e-16 ***
## Income                  4.956e+02  3.986e+01  12.432  < 2e-16 ***
## Thinness_10_19_years -1.011e+01  1.489e+00  -6.792 1.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 221.9 on 2183 degrees of freedom
## Multiple R-squared:  0.8703, Adjusted R-squared:  0.8697
## F-statistic:  1465 on 10 and 2183 DF,  p-value: < 2.2e-16
```

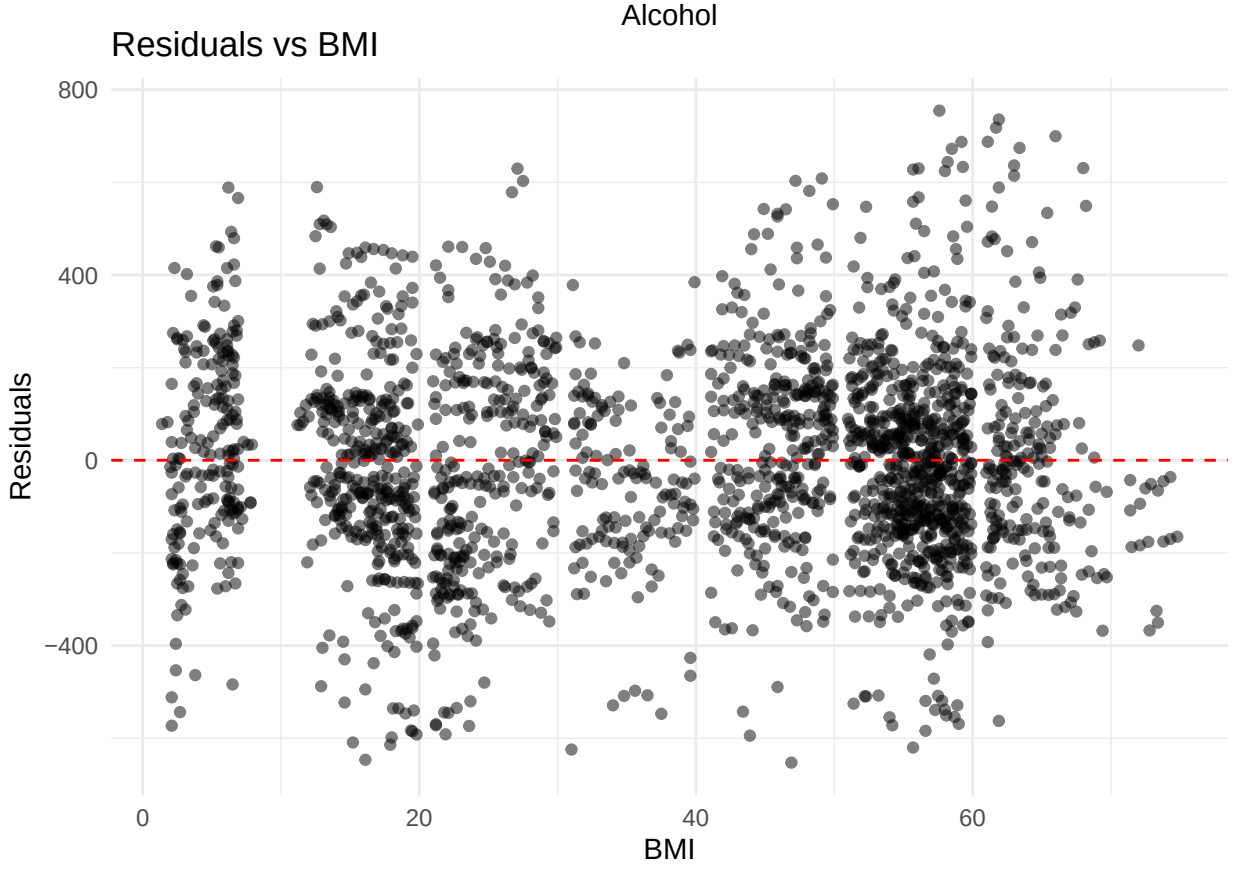
#Diagnostics

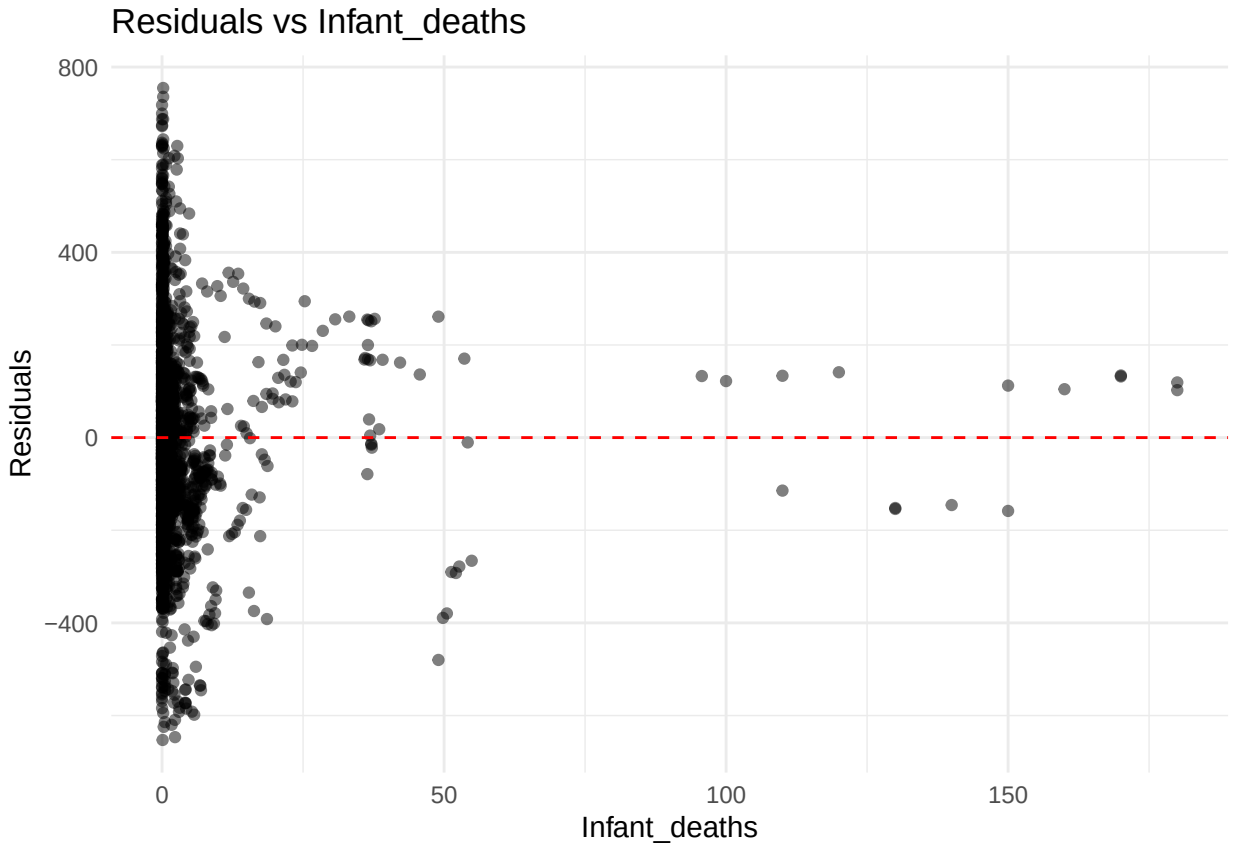
```
#Linearity
fitted_vals3 = fitted(model_3)
residuals3 = resid(model_3)

predictors3 = c("Adult_Mortality_c", "Alcohol", "BMI", "Infant_deaths",
                "Schooling", "Thinness_10_19_years", "GDP", "Income")

for (var in predictors3) {
  p = ggplot(life, aes(x = .data[[var]], y = residuals3)) +
```

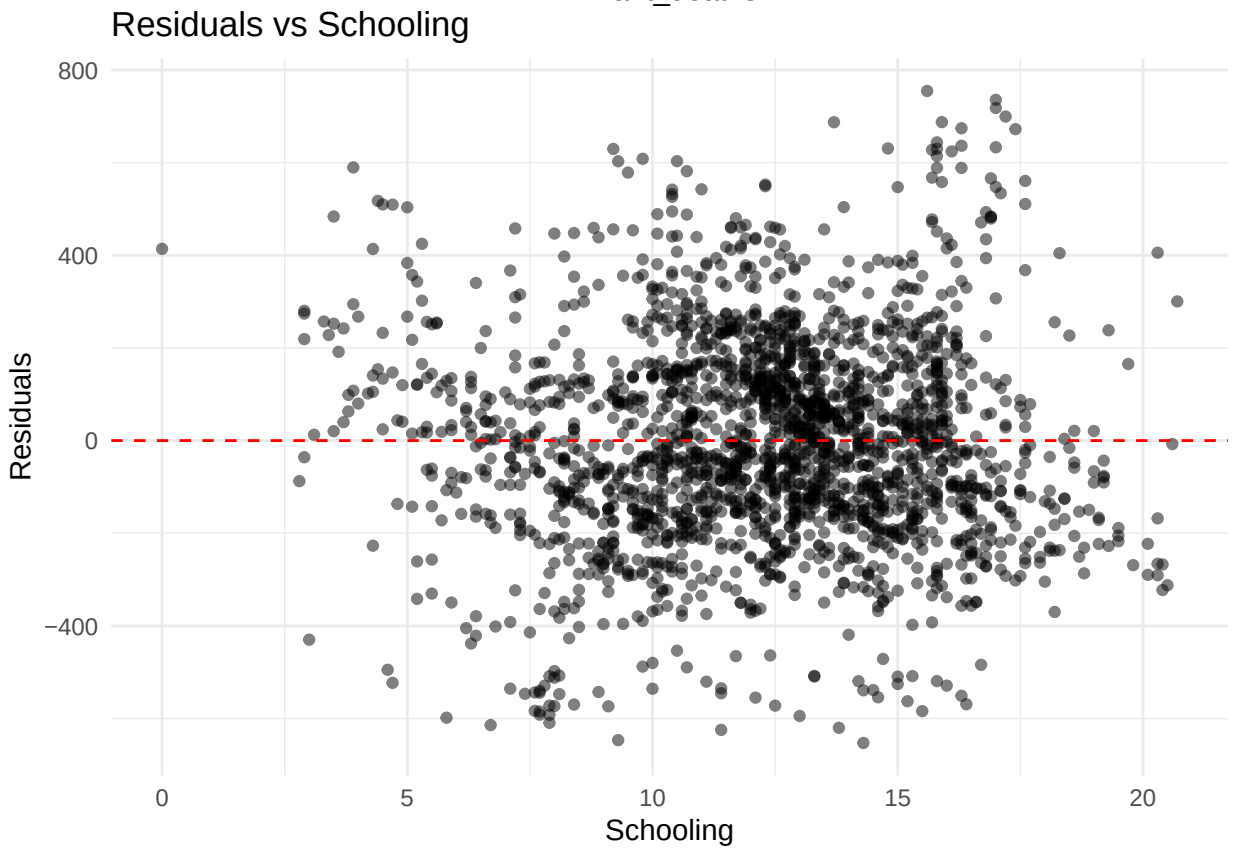
```
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red")+
    ggtitle(paste("Residuals vs", var)) +
    theme_minimal() +
    ylab("Residuals") +
    xlab(var)

  print(p)
}
```

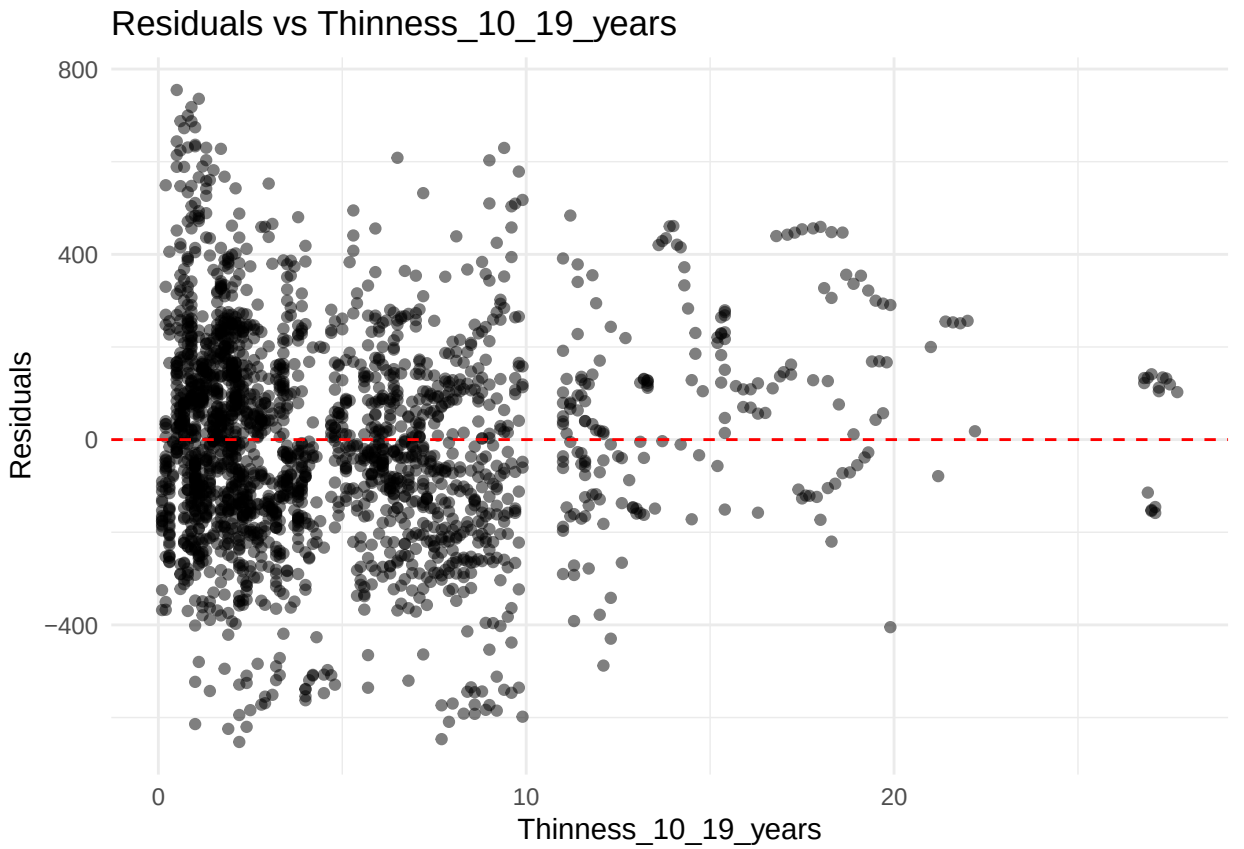


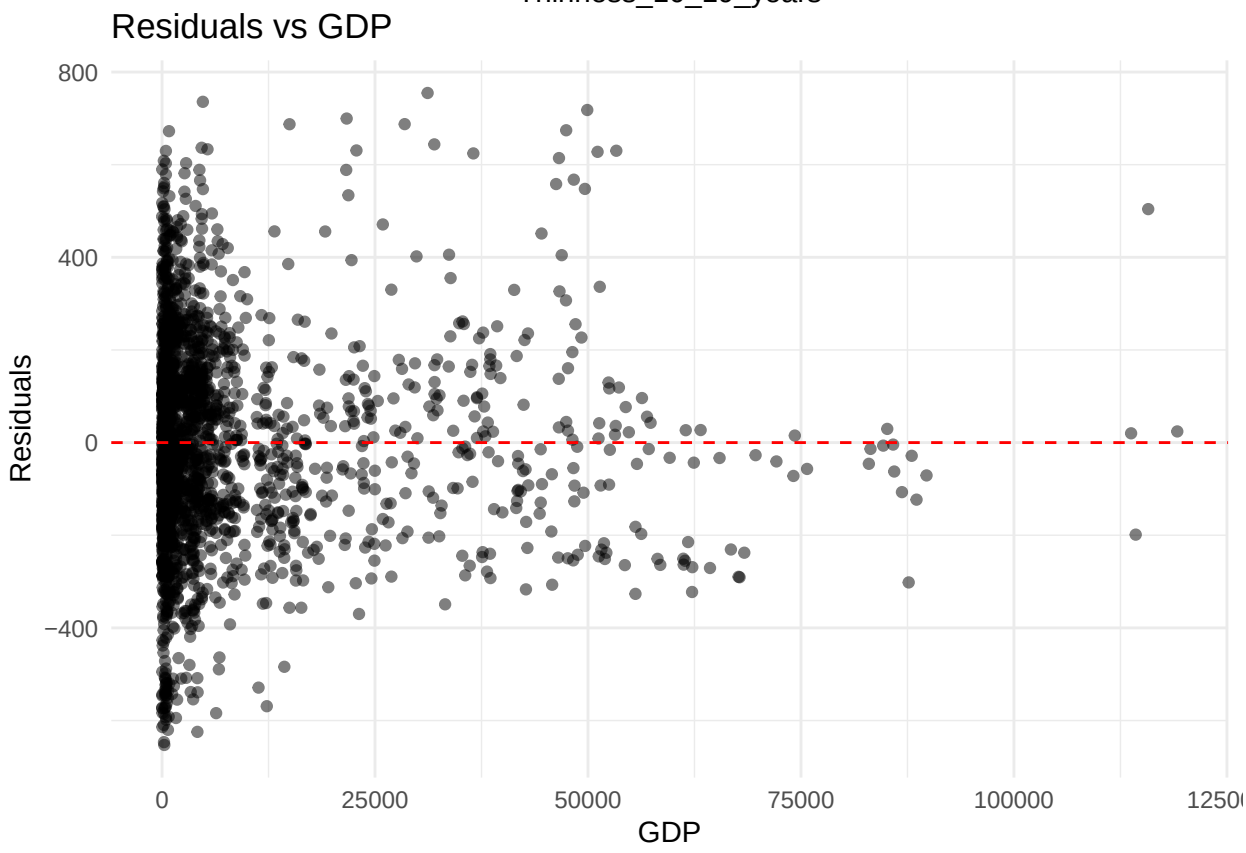

Residuals vs Alcohol
800
400
0
−400
Residuals
0
5
10
15
Alcohol

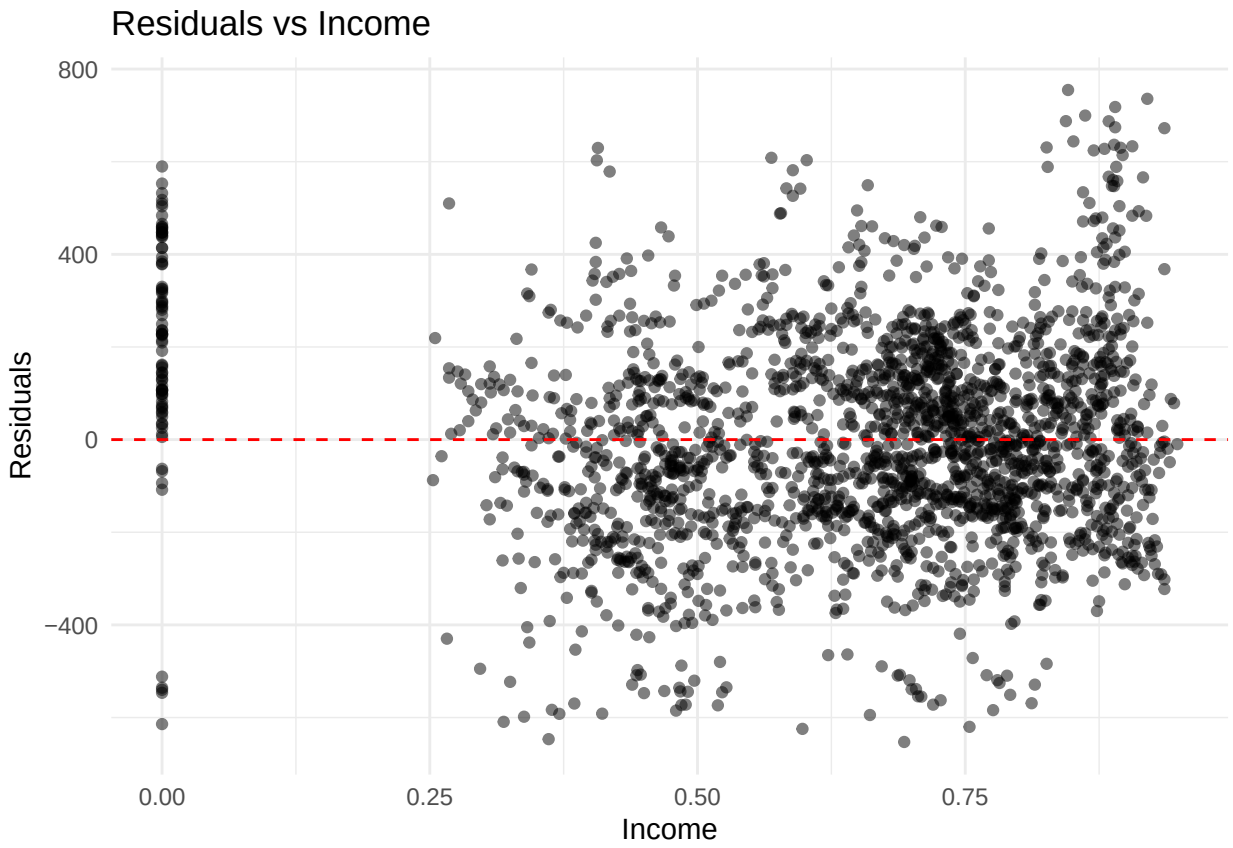

Residuals vs BMI
800
400
0
−400
Residuals
0
20
40
60
BMI


Residuals vs Infant_deaths
800
400
0
−400
Residuals
0
50
100
150
Infant_deaths

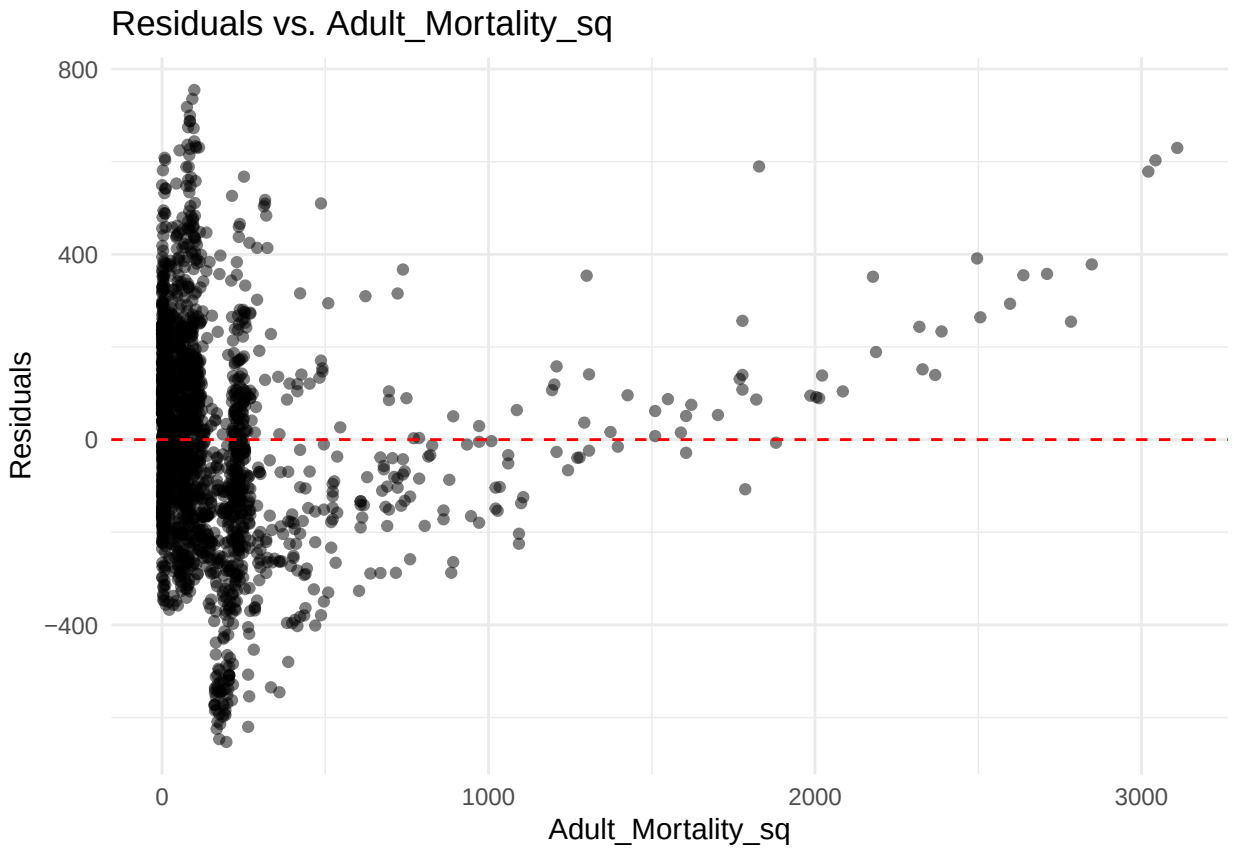

Residuals vs Schooling
800
400
0
−400
Residuals
0
5
10
15
20
Schooling


Residuals vs Thinness_10_19_years
800
400
0
−400
Residuals
0
10
20
Thinness_10_19_years

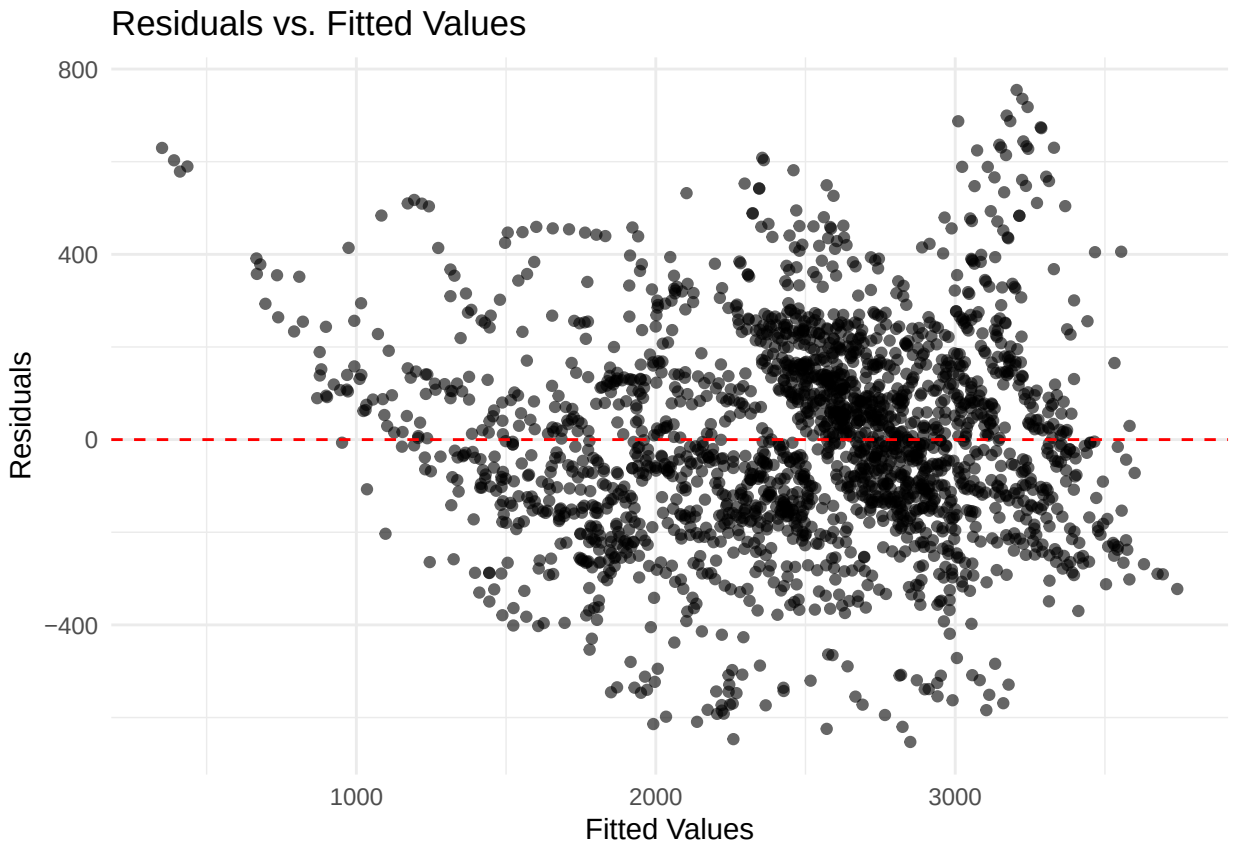

Residuals vs GDP
800
400
0
−400
Residuals
0
25000
50000
75000
100000
GDP




```
ggplot(data = NULL, aes(x = life$Adult_Mortality_sq, y = residuals3)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +
  ggtitle("Residuals vs. Adult_Mortality_sq") +
  xlab("Adult_Mortality_sq") +
  ylab("Residuals")
```




```r
# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals3, y = residuals3)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```
#Linear

#Homoscedasticity
bptest(model_3)
```

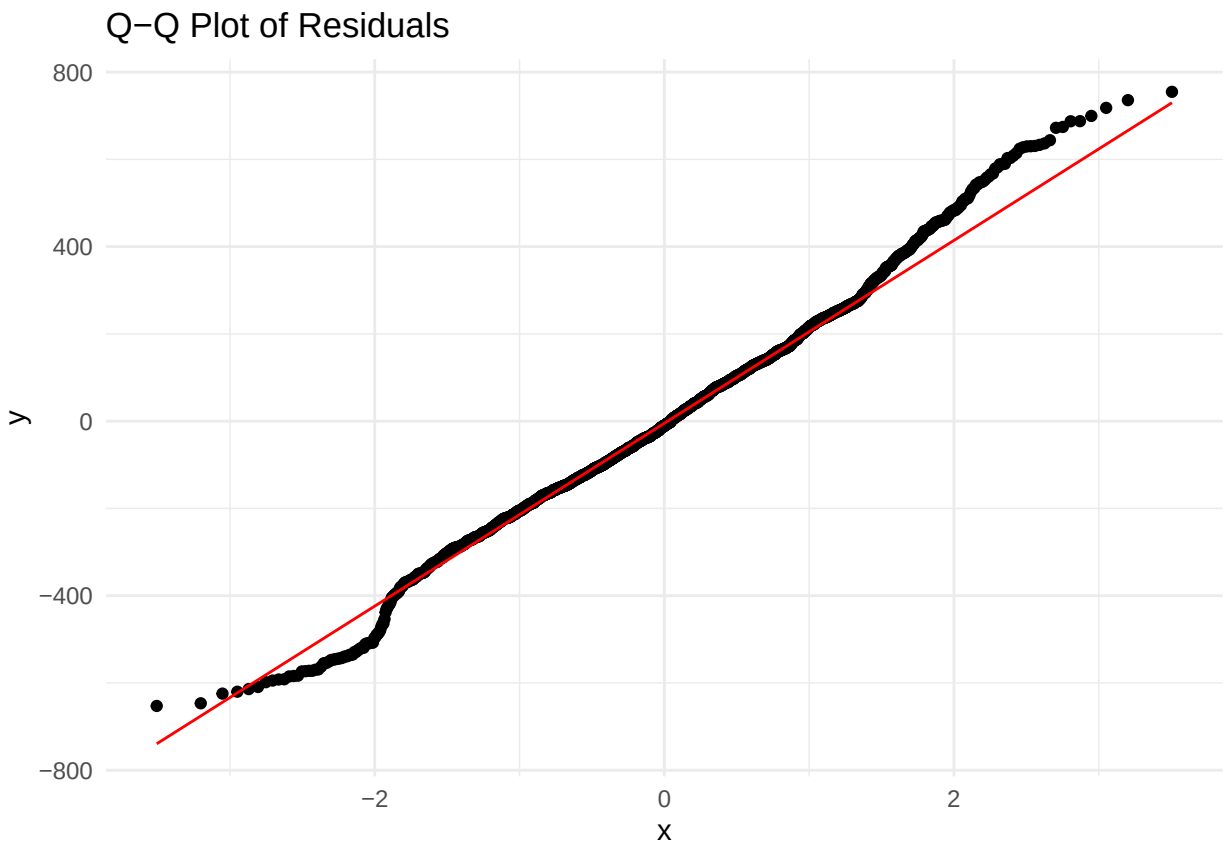
```
##
##  studentized Breusch-Pagan test
##
## data:  model_3
## BP = 186.46, df = 10, p-value < 2.2e-16
```

```
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality
goftest::ad.test(residuals3)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
## data:  residuals3
## An = Inf, p-value = 2.735e-07
```

```
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals3)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
```

```
  theme_minimal()
```




```
#Not normally distributed
```

#Weighted Leasted Squares adjusting for non-constant variance

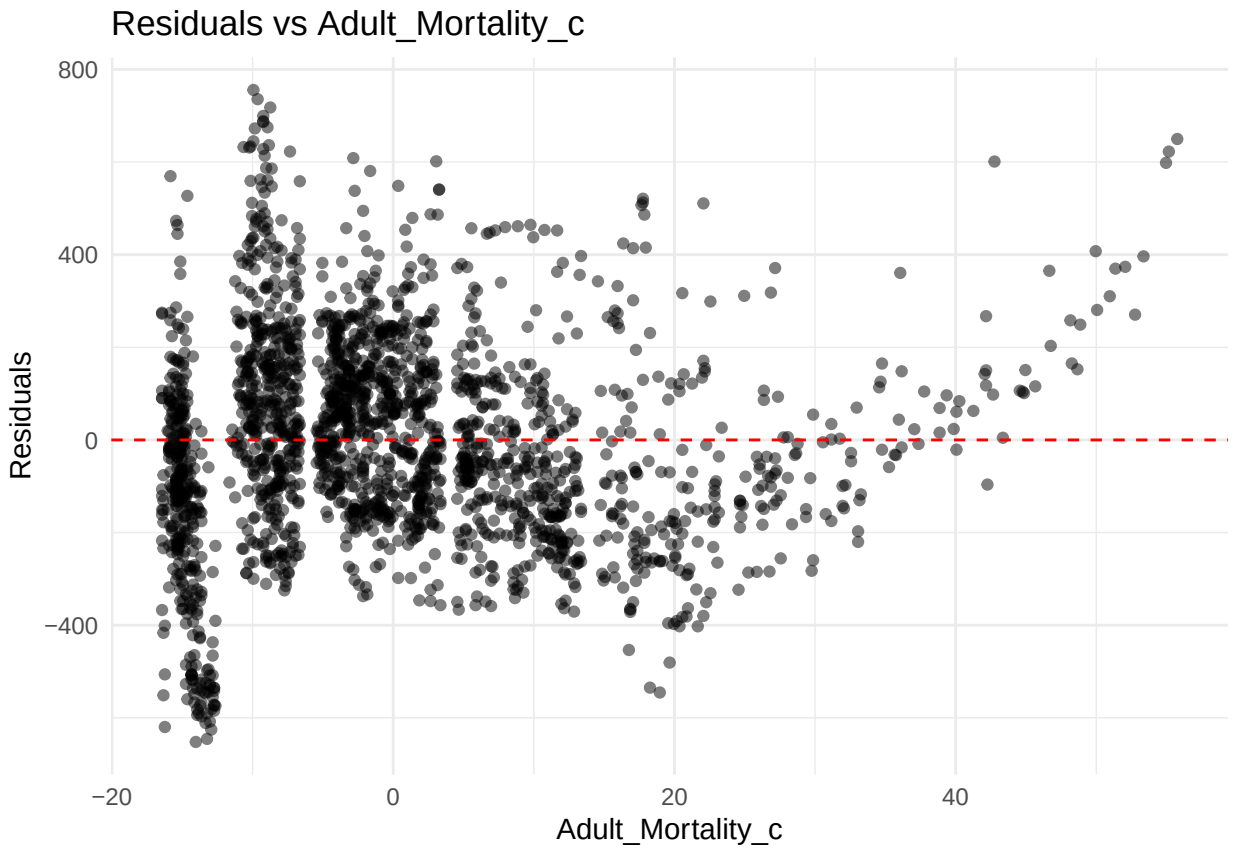
```
weights = 1/fitted(lm(abs(residuals3) ~ fitted_vals3))^2
model_wls = lm(
  data = life,
  Life_expectancy_bc ~
    Status +
    Adult_Mortality_c + Adult_Mortality_sq +
    BMI +
    Schooling +
    Infant_deaths +
    Alcohol +
    GDP +
    Income +
    Thinness_10_19_years,
  weights = weights
)
summary(model_wls)
```

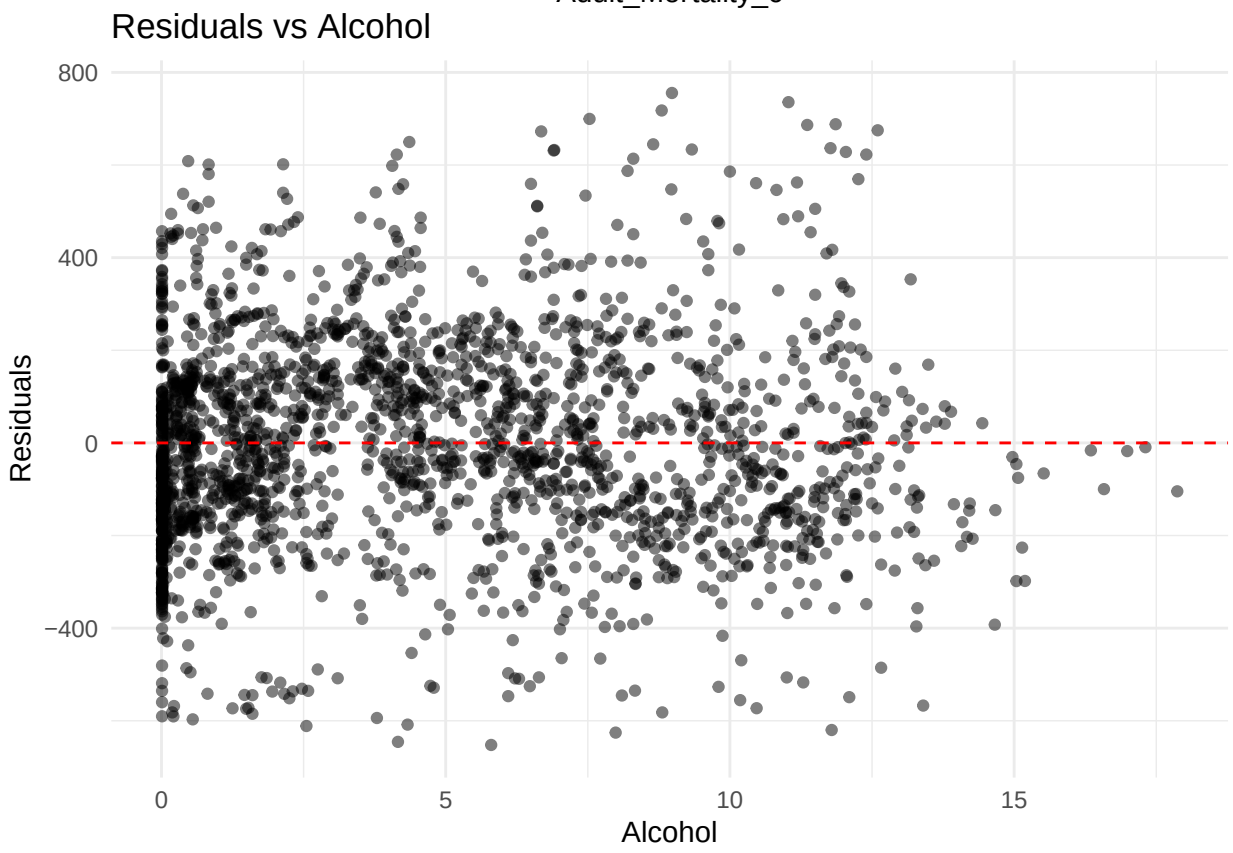
```
##
## Call:
## lm(formula = Life_expectancy_bc ~ Status + Adult_Mortality_c +
##     Adult_Mortality_sq + BMI + Schooling + Infant_deaths + Alcohol +
##     GDP + Income + Thinness_10_19_years, data = life, weights = weights)
```

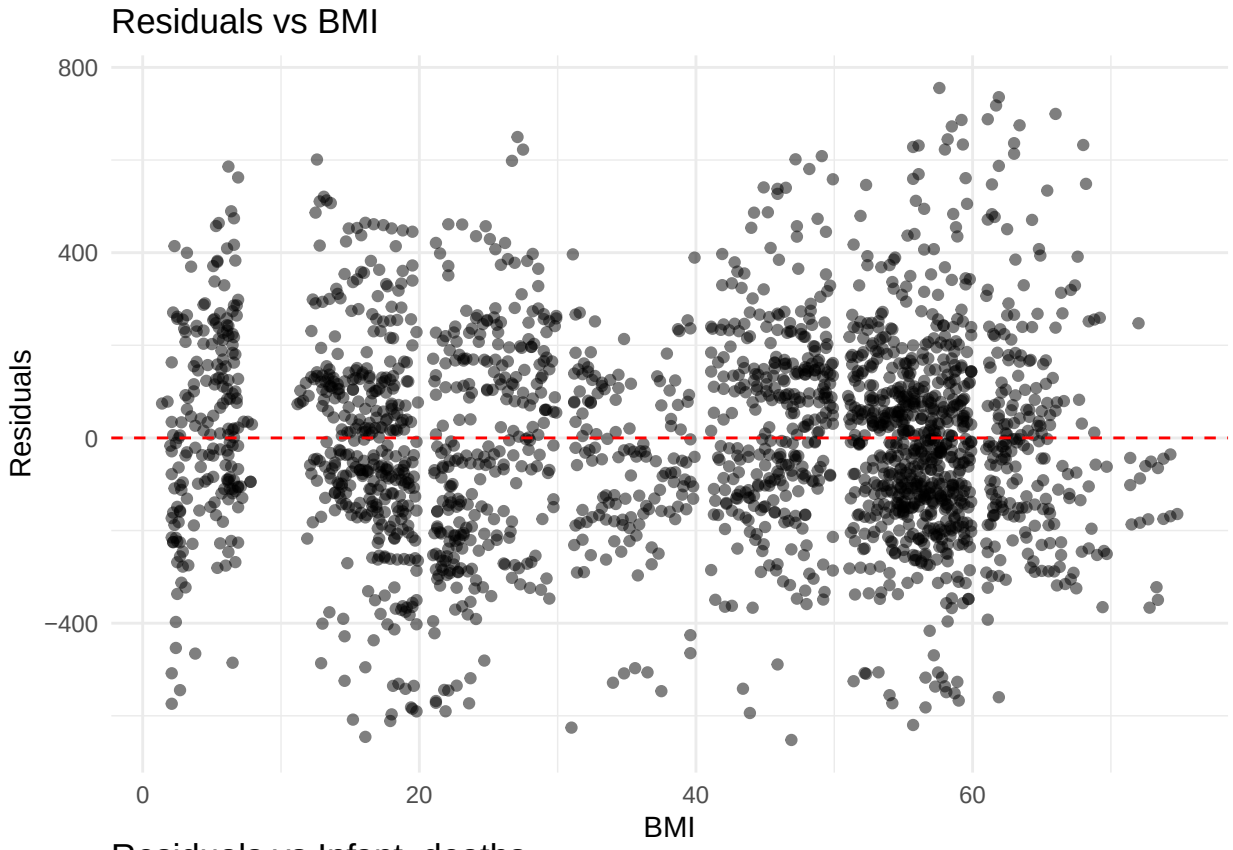
```
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7940 -0.8385 -0.0562  0.7860  4.4386
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.501e+03  3.571e+01  42.038  < 2e-16 ***
## StatusDeveloping      -1.287e+02  1.660e+01  -7.754 1.35e-14 ***
## Adult_Mortality_c     -1.696e+01  5.834e-01 -29.065  < 2e-16 ***
## Adult_Mortality_sq    -2.625e-01  1.965e-02 -13.363  < 2e-16 ***
## BMI                    1.375e+00  3.205e-01   4.291 1.86e-05 ***
## Schooling              6.651e+01  2.839e+00  23.426  < 2e-16 ***
## Infant_deaths          3.563e-01  4.292e-01   0.830    0.407
## Alcohol               -9.195e+00  1.665e+00  -5.524 3.71e-08 ***
## GDP                    3.337e-03  3.799e-04   8.784  < 2e-16 ***
## Income                 5.052e+02  4.017e+01  12.575  < 2e-16 ***
## Thinness_10_19_years -1.031e+01  1.502e+00  -6.863 8.78e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.279 on 2183 degrees of freedom
## Multiple R-squared:  0.8684, Adjusted R-squared:  0.8678
## F-statistic:  1441 on 10 and 2183 DF,  p-value: < 2.2e-16
```

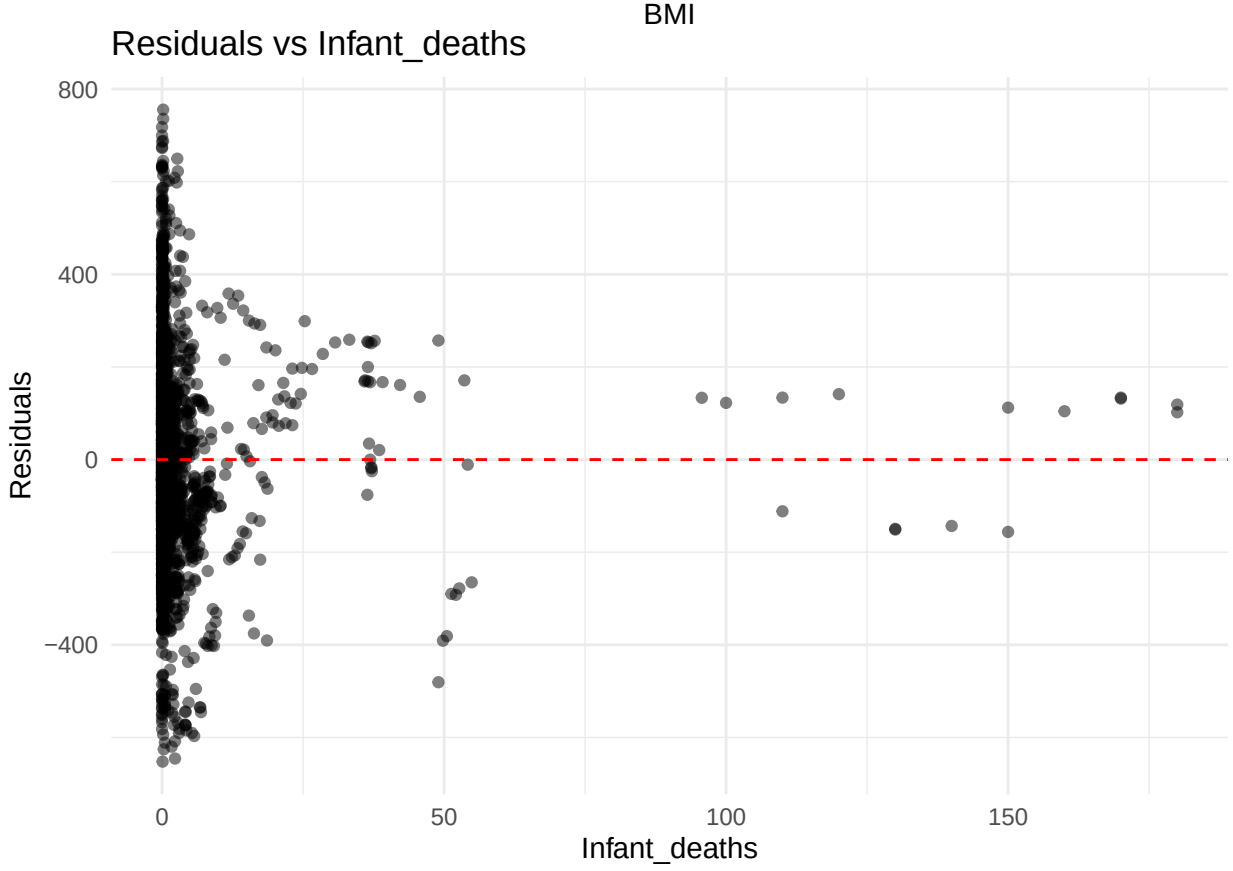
#Diagnostics

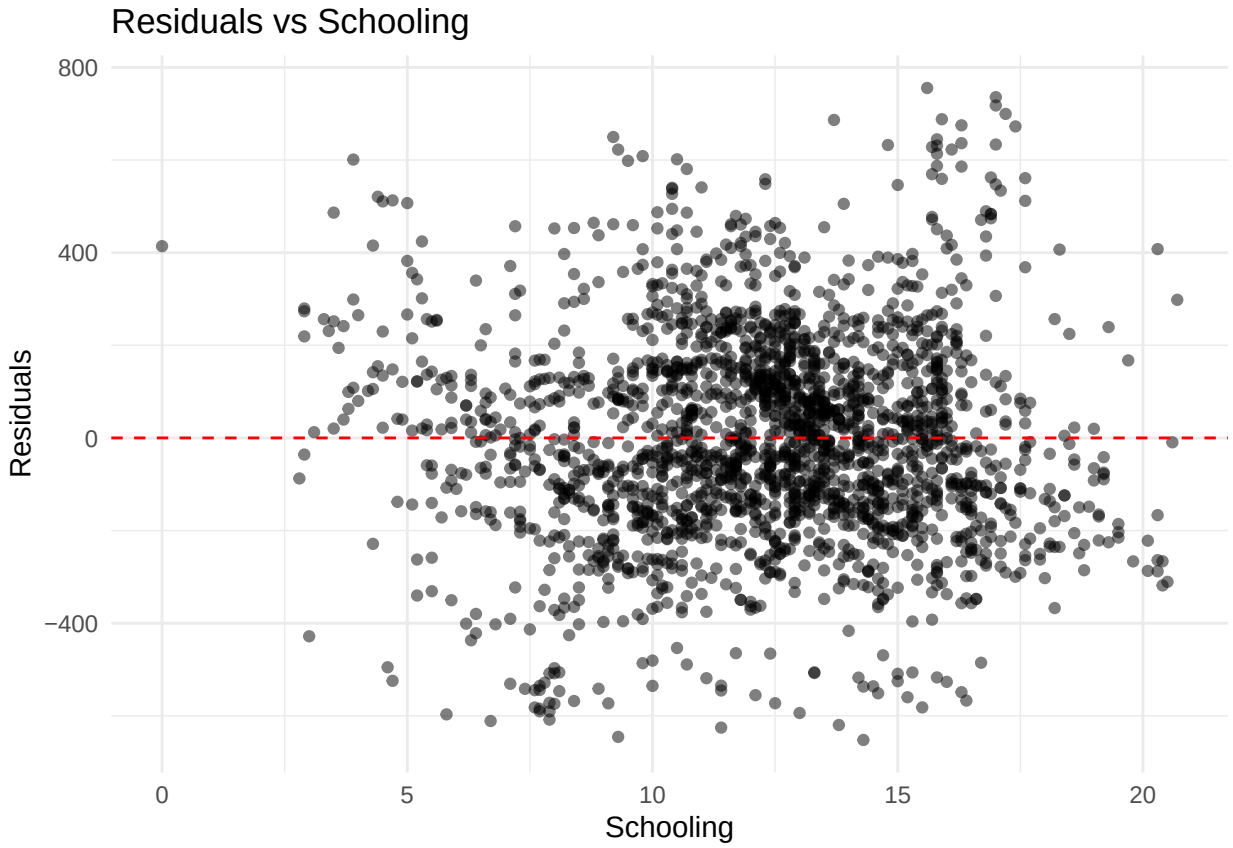
```
#Linearity
fitted_vals_wls = fitted(model_wls)
residuals_wls = resid(model_wls)

for (var in predictors3) {
  p = ggplot(life, aes(x = .data[[var]], y = residuals_wls)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red")+
    ggtitle(paste("Residuals vs", var)) +
    theme_minimal() +
    ylab("Residuals") +
    xlab(var)

  print(p)
}
```

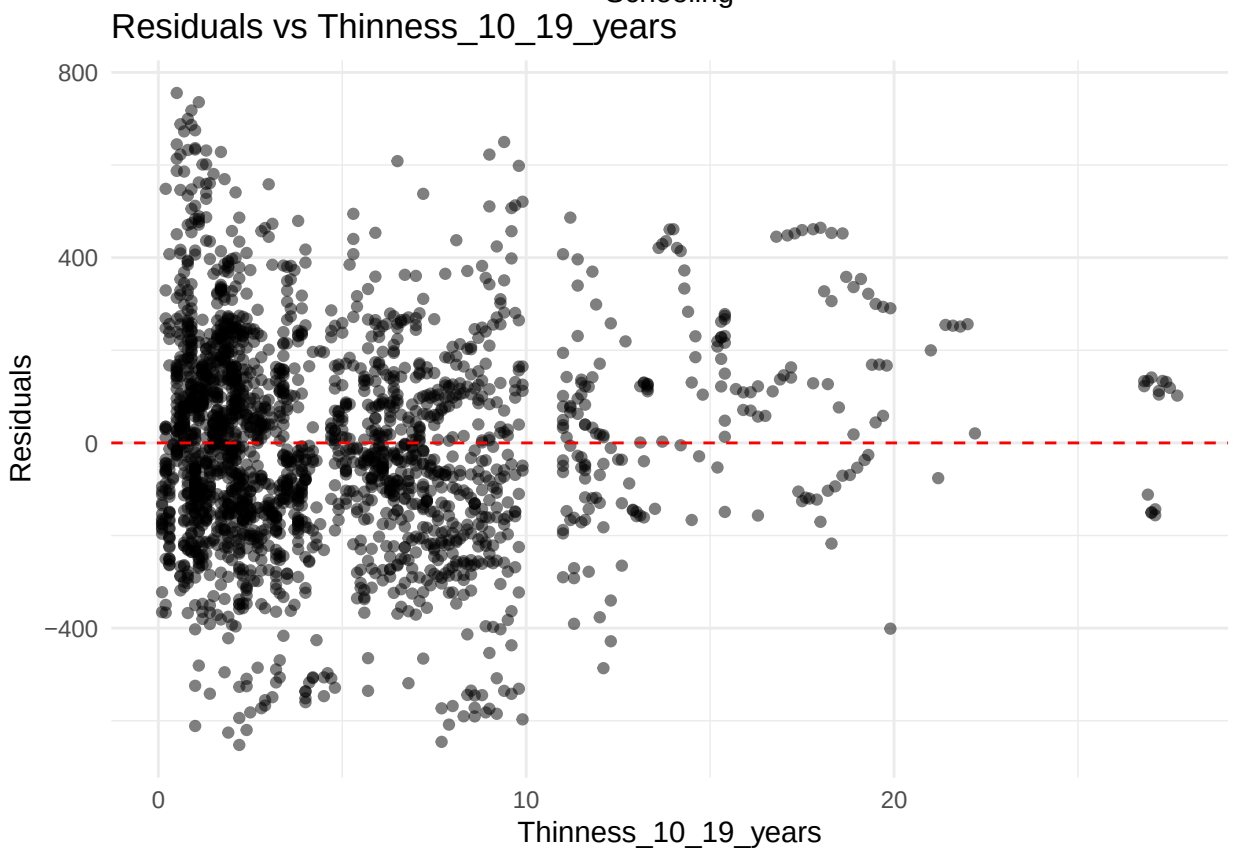

Residuals vs Adult_Mortality_c
800
400
0
−400
Residuals
−20
0
20
40
Adult_Mortality_c

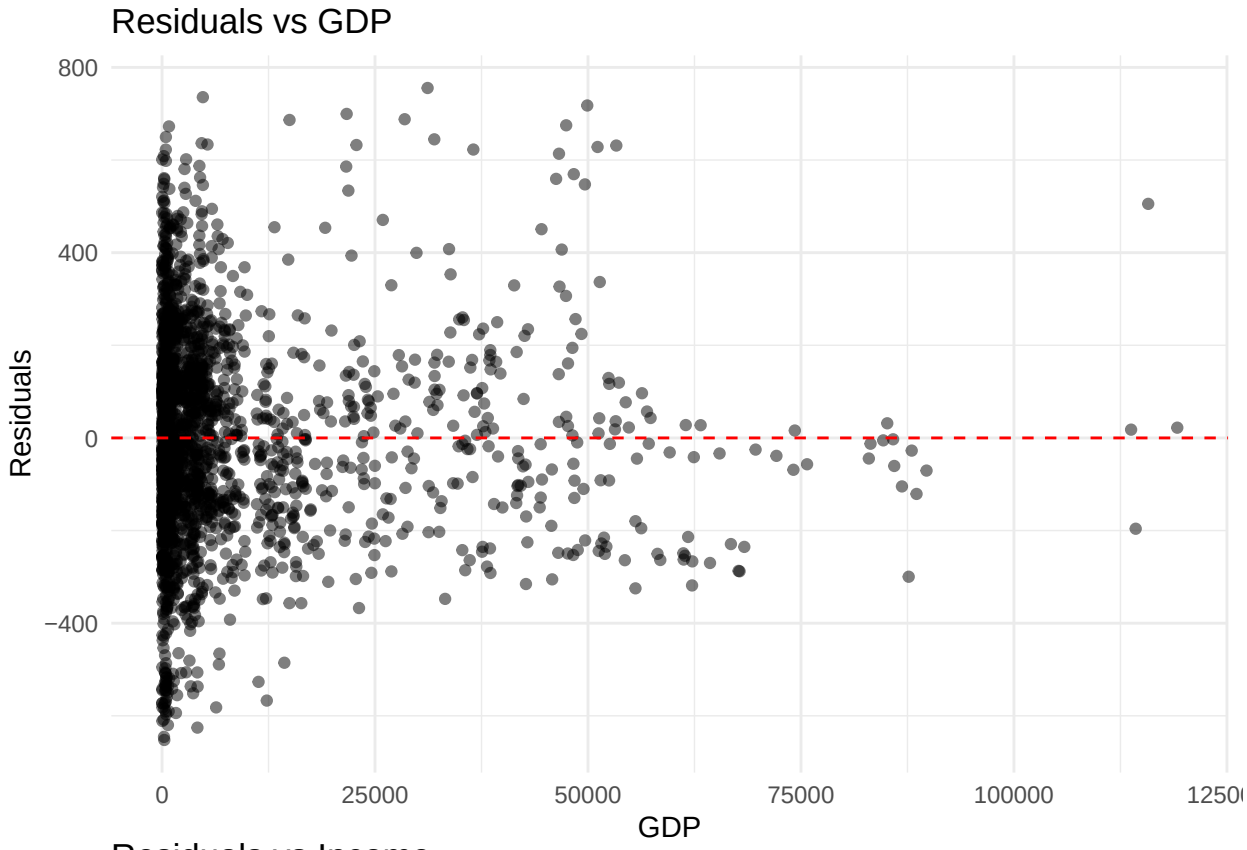

Residuals vs Alcohol
800
400
0
−400
Residuals
0
5
10
15
Alcohol

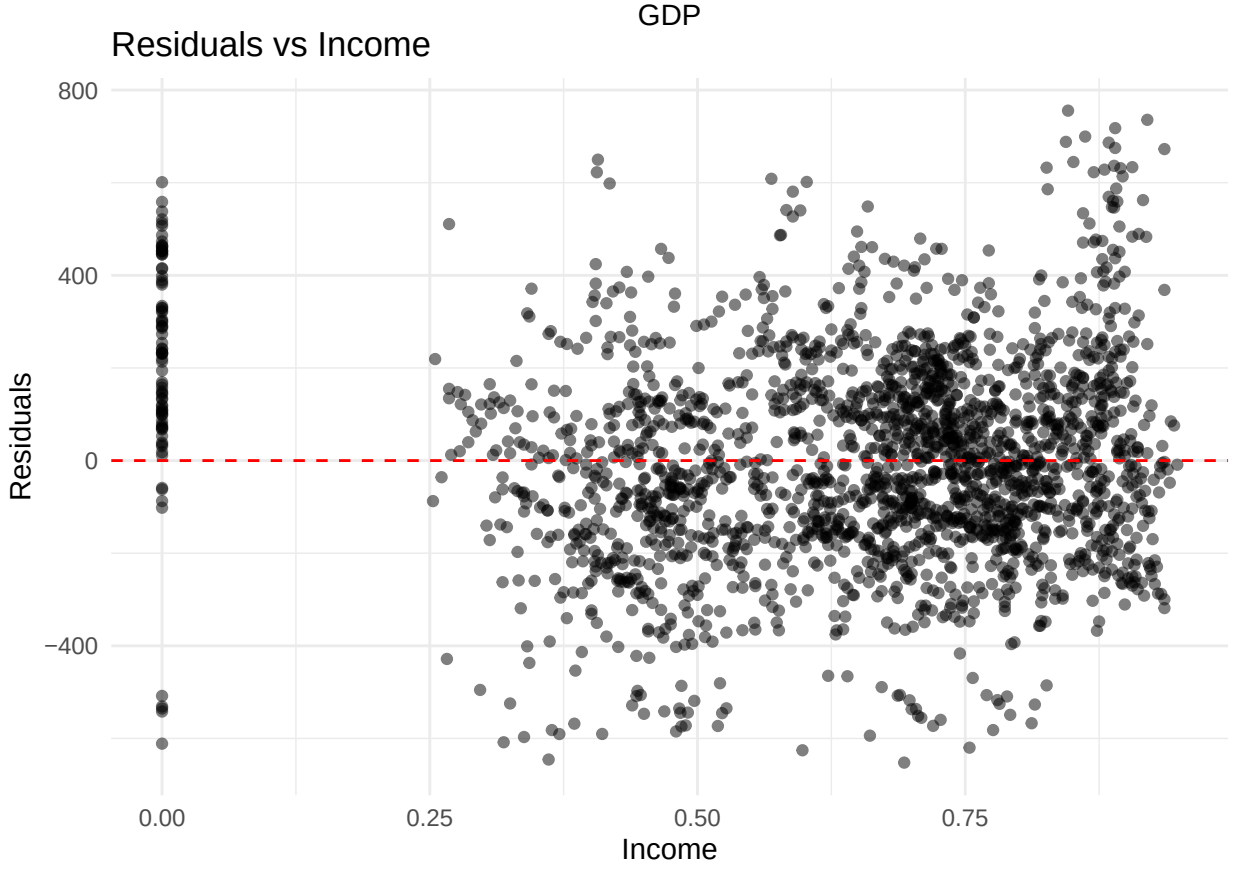

Residuals vs BMI
800
400
0
−400
Residuals
0
20
40
60
BMI


Residuals vs Infant_deaths
800
400
0
−400
Residuals
0
50
100
150
Infant_deaths

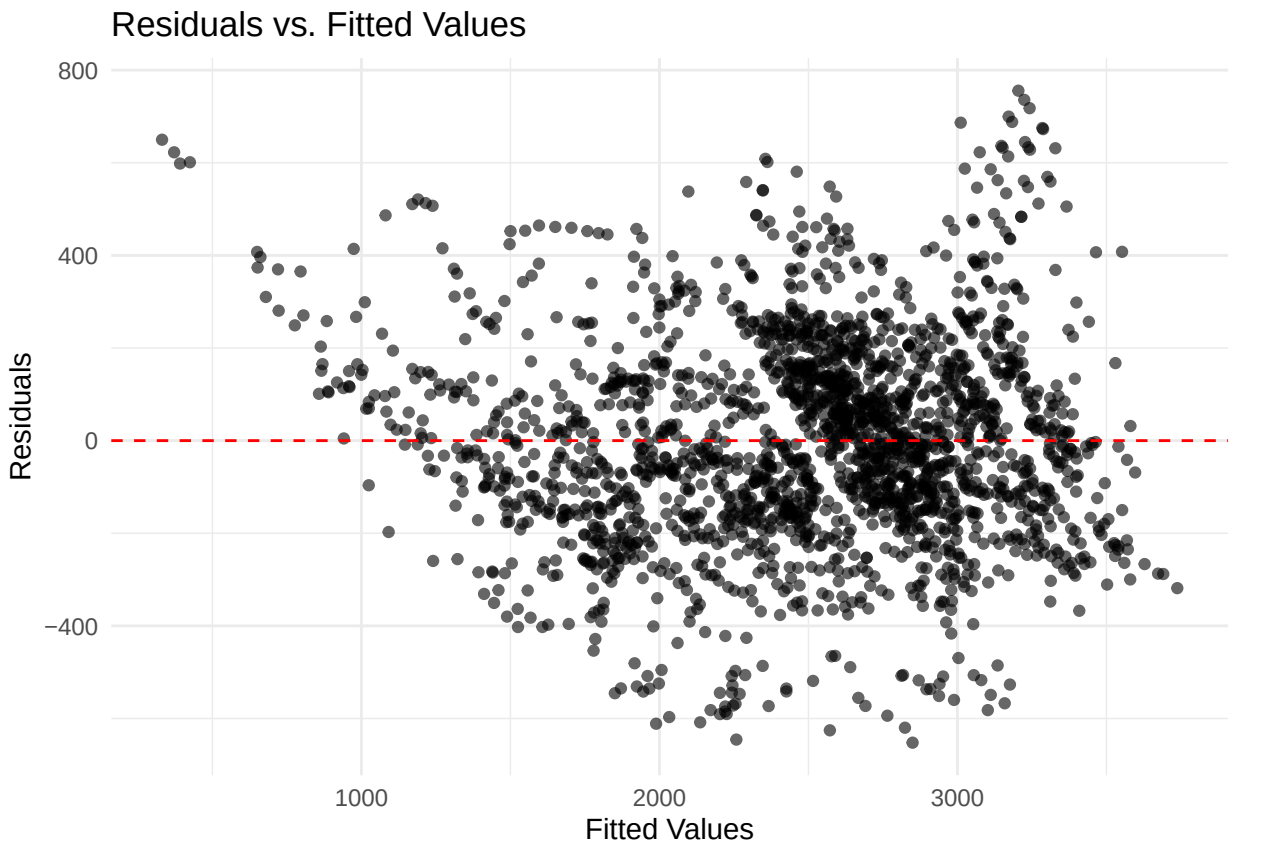

Residuals vs Schooling
800
400
0
−400
Residuals
0
5
10
15
20
Schooling


Residuals vs Thinness_10_19_years
800
400
0
−400
Residuals
0
10
20
Thinness_10_19_years


Residuals vs GDP
800
400
0
−400
Residuals
0
25000
50000
75000
100000
GDP


Residuals vs Income
800
400
0
−400
Residuals
0.00
0.25
0.50
0.75
Income

```r
# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals_wls, y = residuals_wls)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```r
#Linear

#Homoscedasticity
bptest(model_wls)
```

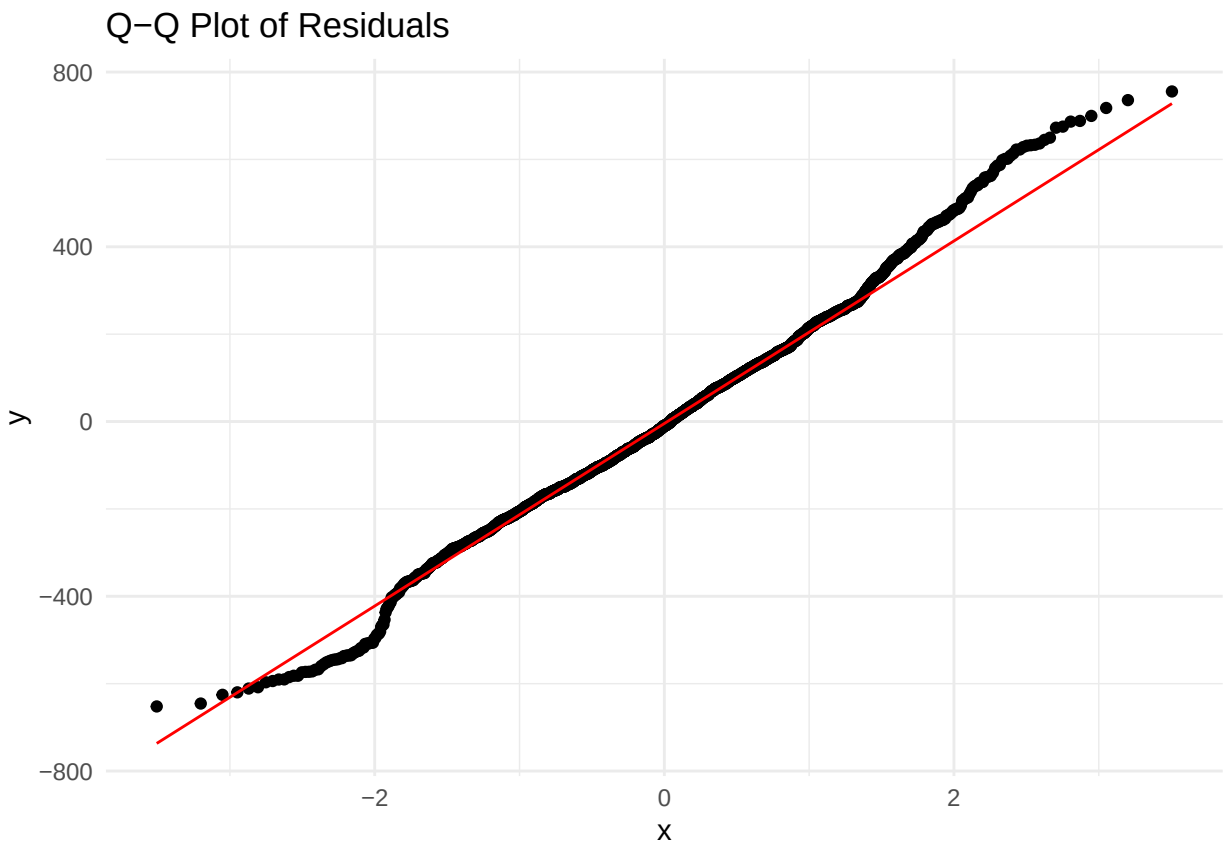
```
##
##  studentized Breusch-Pagan test
##
## data:  model_wls
## BP = 0.025449, df = 10, p-value = 1
```

```r
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality
goftest::ad.test(residuals_wls)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
```

```
## data:  residuals_wls
## An = Inf, p-value = 2.735e-07
```

```r
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals_wls)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()
```

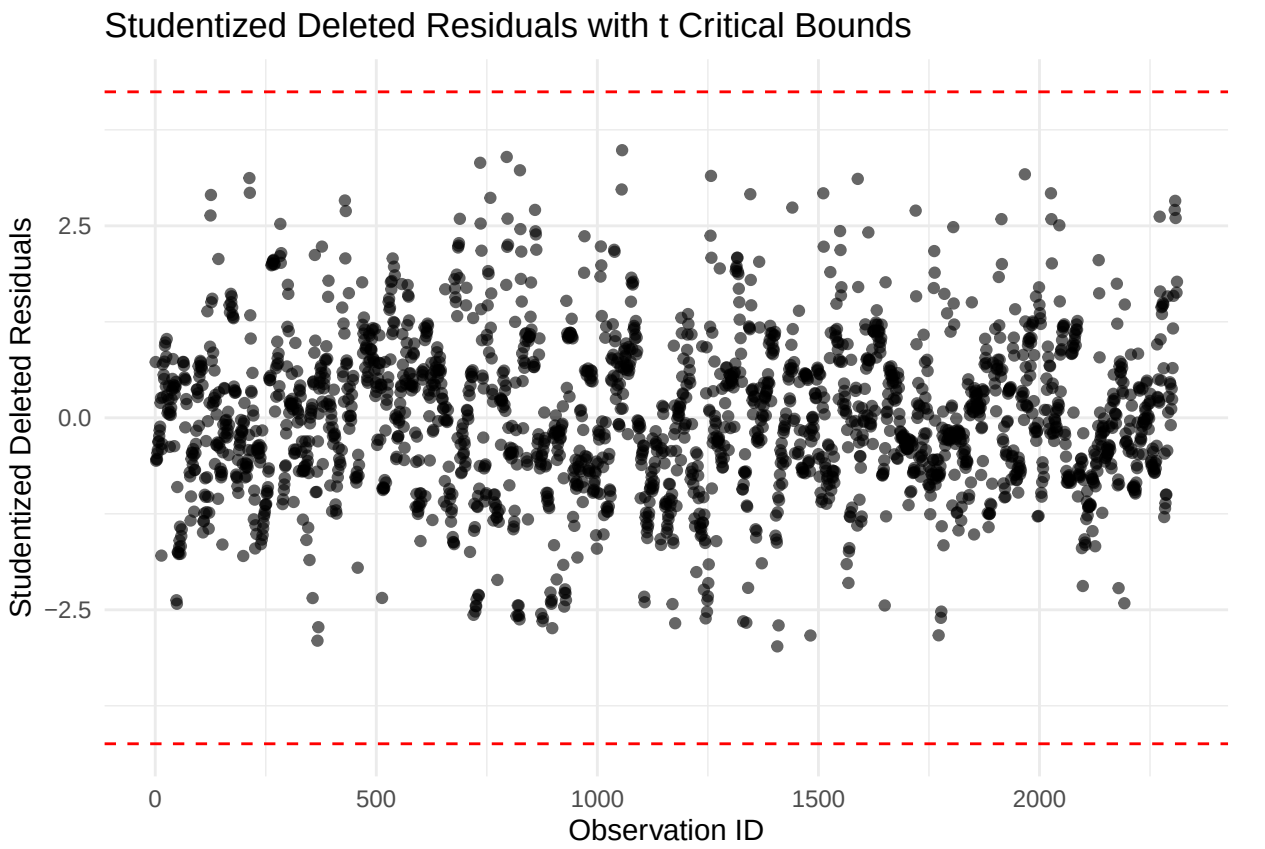



```r
#Normality applies
```

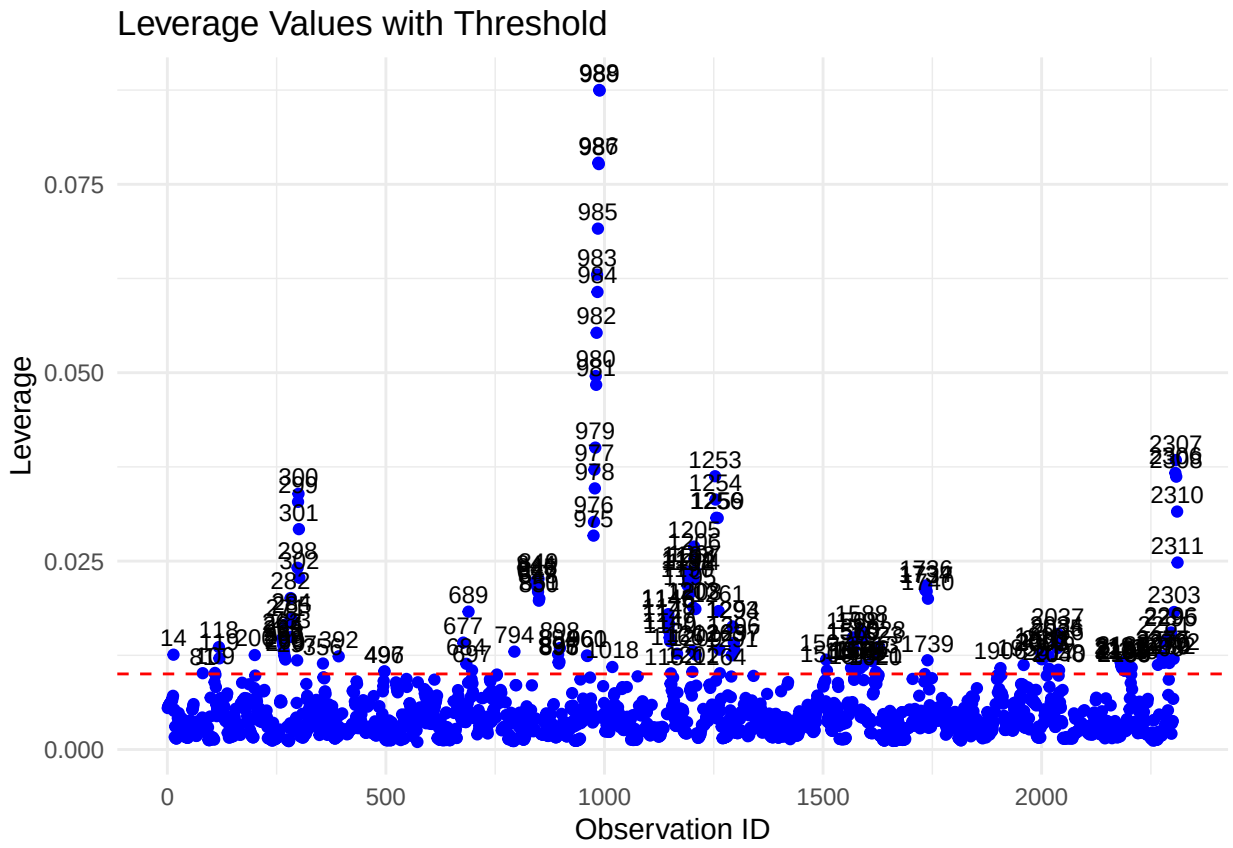
#Check outliers and influential points

```r
#Studentized deleted residuals. Threshold using Bonferroni-adjusted p-values
stu_del_resid = rstudent(model_wls)
n = nrow(life)
alpha = 0.05
t_critical = qt(1-alpha/(2*n), df = n-length(coefficients(model_wls)))
stu_del_resid_df = data.frame(ID = as.numeric(rownames(model_wls$model)), rstudent = stu_del_resid)
out_stud = which(abs(stu_del_resid) > t_critical)
ggplot(stu_del_resid_df, aes(x = ID, y = rstudent)) +
  geom_point(alpha = 0.6) +
  geom_hline(yintercept = c(-t_critical, t_critical), color = "red", linetype = "dashed") +
  geom_text(aes(label = ifelse(abs(stu_del_resid) > t_critical, ID, "")),
            vjust = -0.5, size = 3, color = "black") +
  theme_minimal() +
  ggtitle("Studentized Deleted Residuals with t Critical Bounds") +
  ylab("Studentized Deleted Residuals") +
```

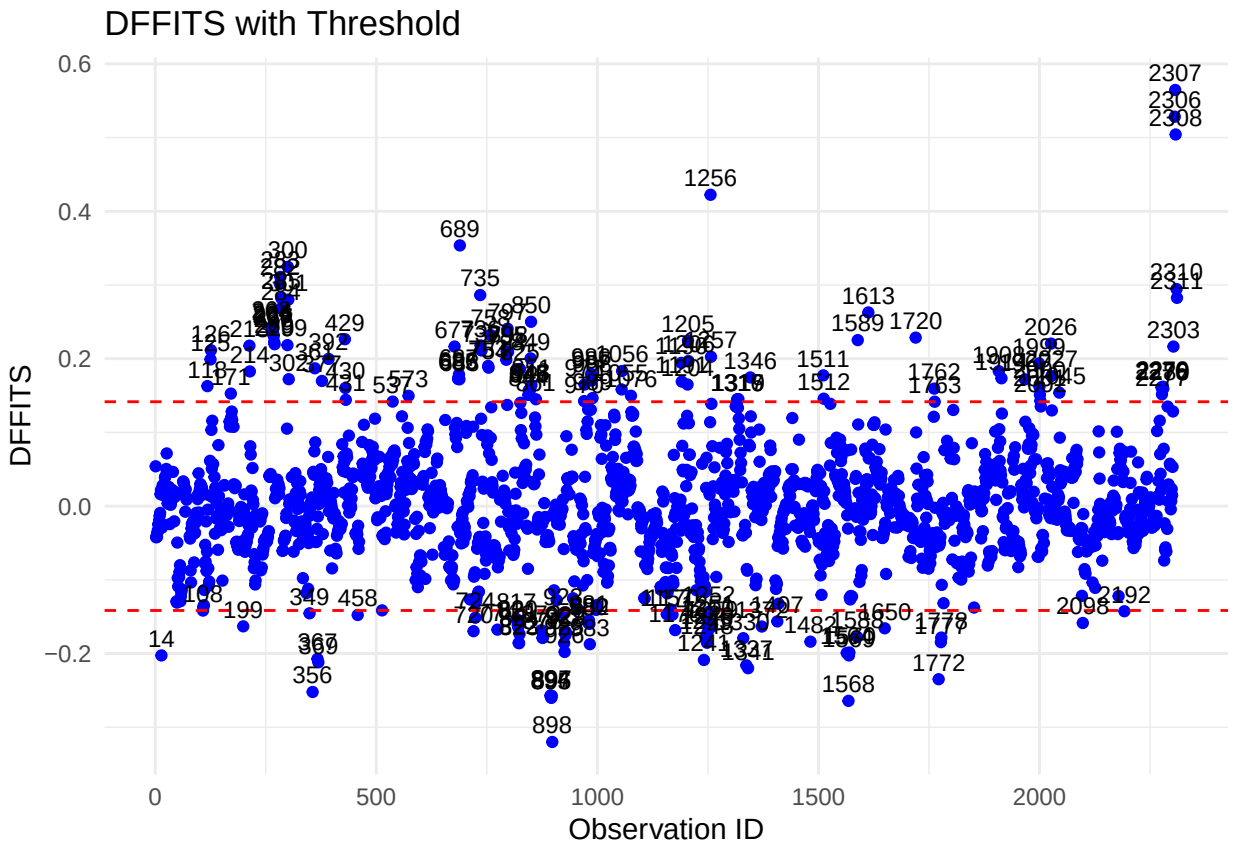
```
  xlab("Observation ID")
```



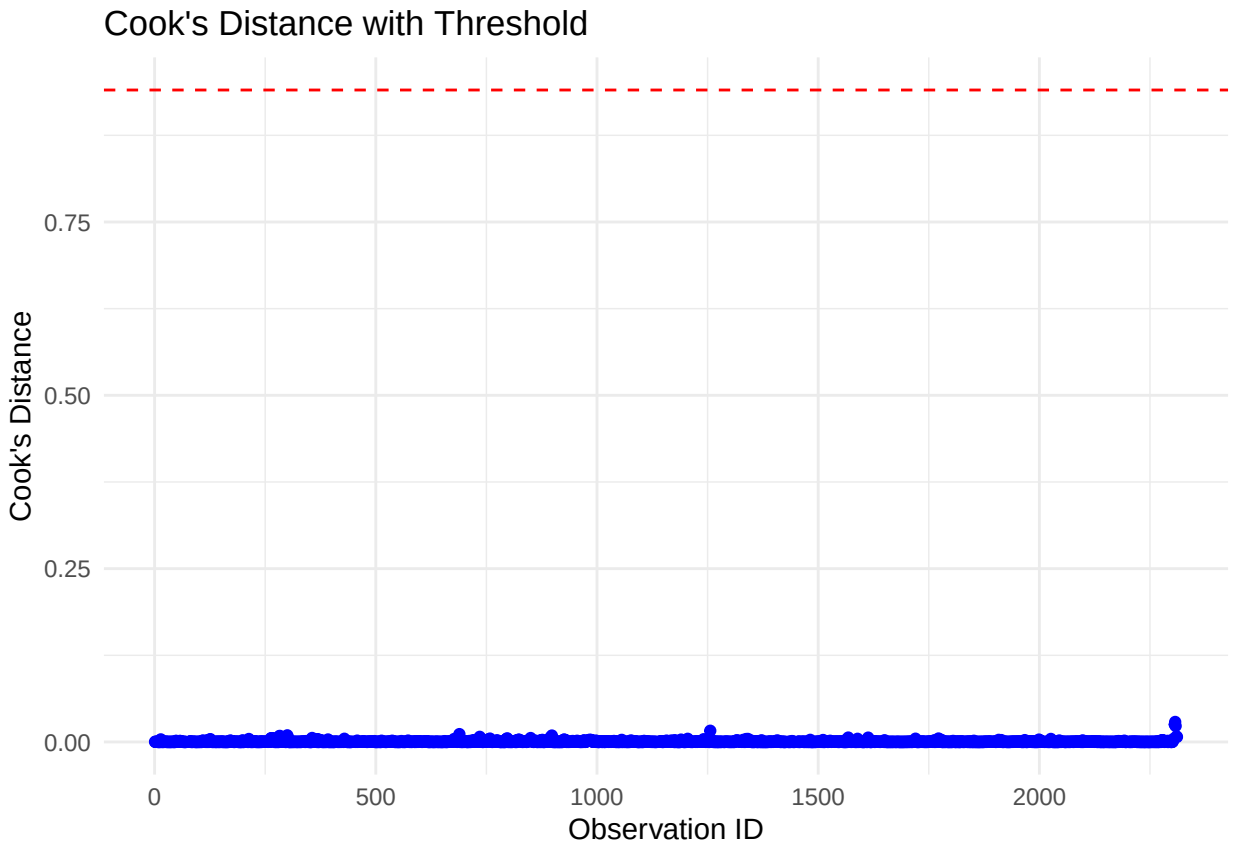

```
#High leverage
h = hatvalues(model_wls)
p = length(coef(model_wls))
avgLeverage = 2*p/n
highLeverage = which(h > avgLeverage)

leverage_df = data.frame(
  ID = as.numeric(rownames(model_wls$model)),
  leverage = h
)

ggplot(leverage_df, aes(x = ID, y = leverage)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = avgLeverage, color = "red", linetype = "dashed") +
  geom_text(aes(label = ifelse(leverage > avgLeverage, ID, "")),
            vjust = -0.5, size = 3, color = "black") +
  theme_minimal() +
  labs(
    title = "Leverage Values with Threshold",
    x = "Observation ID",
    y = "Leverage"
  )
```




```
#Influential points for prediction
dffits_val = dffits(model_wls)
cd_val = cooks.distance(model_wls)
dffits_threshold = 2*sqrt(p/n)
cooks_threshold = qf(0.5, df1 = p, df2 = n-p)

out_dffits = which(abs(dffits_val) > dffits_threshold)
dffits_df = data.frame(
  ID  = as.numeric(rownames(model_wls$model)),
  DFFITS = dffits_val
)
ggplot(dffits_df, aes(x = ID, y = DFFITS)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = c(-dffits_threshold, dffits_threshold), linetype = "dashed", color = "red") +
  geom_text(aes(label = ifelse(abs(DFFITS) > dffits_threshold, ID, "")),
            vjust = -0.5, size = 3) +
  theme_minimal() +
  labs(
    title = "DFFITS with Threshold",
    x = "Observation ID",
    y = "DFFITS"
  )
```




```
out_cd = which(cd_val > cooks_threshold)
cooks_df = data.frame(
  CooksD = cd_val,
  ID = as.numeric(rownames(model_wls$model))
)
ggplot(cooks_df, aes(x = ID, y = CooksD)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = cooks_threshold, linetype = "dashed", color = "red") +
  geom_text(aes(label = ifelse(CooksD > cooks_threshold, ID, "")),
            vjust = -0.5, size = 3) +
  theme_minimal() +
  labs(
    title = "Cook's Distance with Threshold",
    x = "Observation ID",
    y = "Cook's Distance"
  )
```



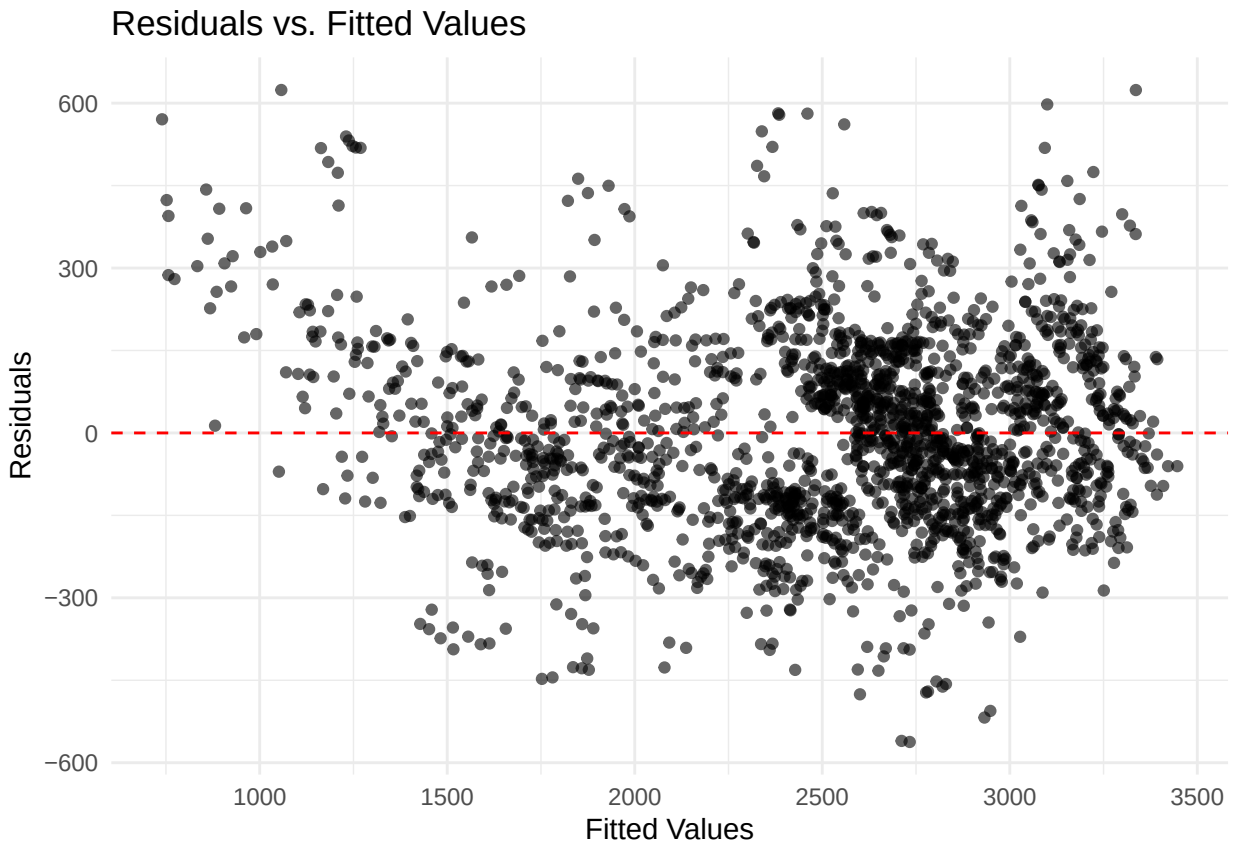
#Removing influential points and outliers

```
outliers = unique(c(out_dffits, out_cd, out_stud, highLeverage))
life_clean = life[-outliers, ]

model_3c = lm(
  data = life_clean,
  Life_expectancy_bc ~
    Status +
    Adult_Mortality_c + Adult_Mortality_sq +
    BMI +
    Schooling +
    Infant_deaths +
    Alcohol +
    GDP +
    Income +
    Thinness_10_19_years
)
fitted_vals_3c = fitted(model_3c)
residuals_3c = resid(model_3c)

weights_c = 1/fitted(lm(abs(residuals_3c) ~ fitted_vals_3c))^2
model_wls_c = lm(
  data = life_clean,
  Life_expectancy_bc ~
    Status +
    Adult_Mortality_c + Adult_Mortality_sq +
    BMI +
```

```
    Schooling +
    Infant_deaths +
    Alcohol +
    GDP +
    Income +
    Thinness_10_19_years,
  weights = weights_c
)
summary(model_wls_c)
```

```
##
## Call:
## lm(formula = Life_expectancy_bc ~ Status + Adult_Mortality_c +
##     Adult_Mortality_sq + BMI + Schooling + Infant_deaths + Alcohol +
##     GDP + Income + Thinness_10_19_years, data = life_clean, weights = weights_c)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2020 -0.8713 -0.0750  0.7997  5.0461
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.456e+03  3.560e+01  40.909  < 2e-16 ***
## StatusDeveloping      -9.641e+01  1.352e+01  -7.130 1.41e-12 ***
## Adult_Mortality_c     -1.804e+01  5.496e-01 -32.828  < 2e-16 ***
## Adult_Mortality_sq    -3.785e-01  2.482e-02 -15.246  < 2e-16 ***
## BMI                    5.598e-01  2.792e-01   2.005 0.045100 *
## Schooling              2.061e+01  3.328e+00   6.193 7.18e-10 ***
## Infant_deaths          2.768e-01  9.027e-01   0.307 0.759160
## Alcohol               -5.381e+00  1.420e+00  -3.790 0.000155 ***
## GDP                    1.906e-03  3.744e-04   5.090 3.92e-07 ***
## Income                 1.423e+03  7.230e+01  19.676  < 2e-16 ***
## Thinness_10_19_years -9.277e+00  1.334e+00  -6.956 4.77e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.26 on 1921 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.9012
## F-statistic:  1763 on 10 and 1921 DF,  p-value: < 2.2e-16
```

```
fitted_vals_wls_c = fitted(model_wls_c)
residuals_wls_c = resid(model_wls_c)

# Residuals vs Fitted Plot
ggplot(data = NULL, aes(x = fitted_vals_wls_c, y = residuals_wls_c)) +
  geom_point(alpha = 0.6, color = "black") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Fitted Values", y = "Residuals", title = "Residuals vs. Fitted Values") +
  theme_minimal()
```




```
#Linear

#Homoscedasticity
bptest(model_wls_c)
```

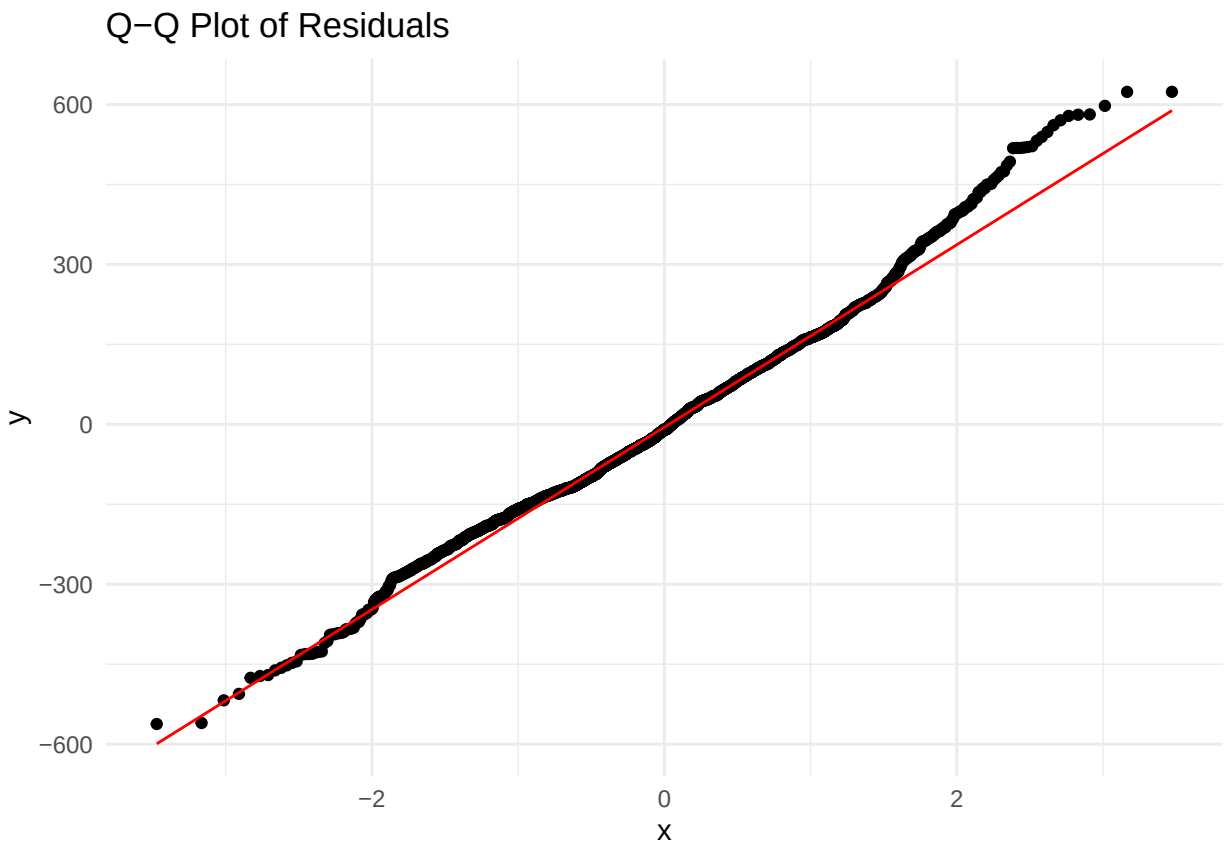
```
##
##  studentized Breusch-Pagan test
##
## data:  model_wls_c
## BP = 0.031076, df = 10, p-value = 1
```

```
#Non constant variance. Also the residuals vs. fitted has a funnel shape

#Normality
goftest::ad.test(residuals_wls_c)
```

```
##
##  Anderson-Darling test of goodness-of-fit
##  Null hypothesis: uniform distribution
##  Parameters assumed to be fixed
##
## data:  residuals_wls_c
## An = Inf, p-value = 3.106e-07
```

```
# Q-Q Plot of residuals
ggplot(data = NULL, aes(sample = residuals_wls_c)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
```

```
theme_minimal()
```




```
#Normality applies
```

#The qq plot improves a lot so we decided to remove all the influential points and outliers