Draft 4

Vikas_Reddy_Bodireddy

2023-12-04

Dataset Link

```
life_exp <- read.csv("/Users/vikasreddybodireddy/Desktop/630 drafts/Life Expectancy Data.csv")
n_1<- nrow(life_exp)
n_2 <- nrow(na.omit(life_exp))
# Total rows with NA
n_1
```

```
## [1] 2938
```

```
#total rows without NA
n_2
```

```
## [1] 1649
```

Here there are more than 40% data missing in our dataset, so we have taken mean of data and add those to the Missing values into the data.

Imputing the missing values:

```
numerical_columns <- c("Life.expectancy","Adult.Mortality", "infant.deaths", "Alcohol", "percentage.exp

for (col in numerical_columns) {
  life_exp[[col]][is.na(life_exp[[col]])] <- mean(life_exp[[col]], na.rm = TRUE)
}
categorical_columns <- c("Country", "Status")

getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

for (col in categorical_columns) {
  mode_value <- getmode(life_exp[[col]][!is.na(life_exp[[col]])])
  life_exp[[col]][is.na(life_exp[[col]])] <- mode_value
}
# rows with NA values after mean imputation.
sum(is.na(life_exp))
```

```
## [1] 0
```

```r
life_exp <- life_exp
summary(life_exp)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.20
##  Mode  :character   Median :2008   Mode  :character   Median :72.00
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.60
##                     Max.   :2015                      Max.   :89.00
##  Adult.Mortality infant.deaths       Alcohol       percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.010   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 1.093   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 4.160   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.603   Mean   :  738.251
##  3rd Qu.:227.0   3rd Qu.:  22.0   3rd Qu.: 7.390   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.870   Max.   :19479.912
##   Hepatitis.B       Measles              BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:80.94   1st Qu.:     0.0   1st Qu.:19.40   1st Qu.:   0.00
##  Median :87.00   Median :    17.0   Median :43.00   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:96.00   3rd Qu.:   360.2   3rd Qu.:56.10   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.370    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.938    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.330    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   580.49   1st Qu.:4.189e+05   1st Qu.: 1.60
##  Median :  3116.56   Median :3.676e+06   Median : 3.40
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  7483.16   3rd Qu.:1.275e+07   3rd Qu.: 7.10
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.60      1st Qu.:0.5042                  1st Qu.:10.30
##  Median : 3.40      Median :0.6620                  Median :12.10
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7720                  3rd Qu.:14.10
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
```

```r
dim(life_exp)
```

```
## [1] 2938   22
```

SLR - for Individual Variables:

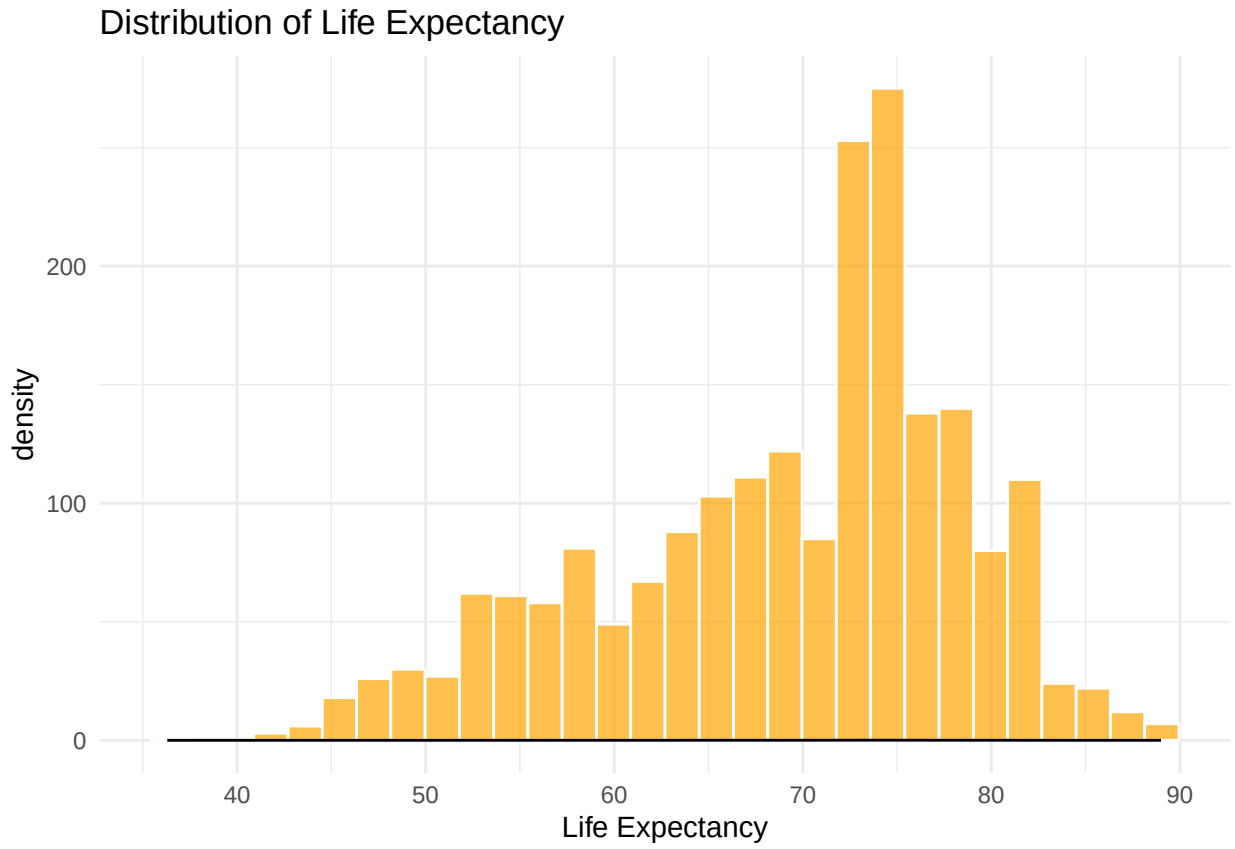
```
create_life_exp_subset <- function(data, selected_year) {
  data_subset <- data %>%
 # filter(Year == selected_year) %>%
  select(Life.expectancy, Adult.Mortality, Alcohol,
         Income.composition.of.resources, Schooling, Country,
         Status, GDP) %>%
  na.omit() %>%
  mutate(

    life_exp = Life.expectancy,
    adult_mortality = Adult.Mortality,
    alcohol = Alcohol,
    income_composition = Income.composition.of.resources,
    schooling = Schooling, gdp = GDP
  ) %>%
  select(life_exp, adult_mortality, alcohol, income_composition, schooling,
          gdp)
}#user_input <- readline(prompt = "Enter the year from 2001 to 2015: ")

#selected_year <- as.integer(user_input = 2004)

selected_year <- as.integer(2004)

life_exp2 <- create_life_exp_subset(life_exp, selected_year)

# Continue with other operations on the life_exp_subset
numeric_data <- select_if(life_exp2, is.numeric)
summary(life_exp2)
```

```
##     life_exp     adult_mortality    alcohol       income_composition
##  Min.   :36.30   Min.   :  1.0   Min.   : 0.010   Min.   :0.0000
##  1st Qu.:63.20   1st Qu.: 74.0   1st Qu.: 1.093   1st Qu.:0.5042
##  Median :72.00   Median :144.0   Median : 4.160   Median :0.6620
##  Mean   :69.22   Mean   :164.8   Mean   : 4.603   Mean   :0.6276
##  3rd Qu.:75.60   3rd Qu.:227.0   3rd Qu.: 7.390   3rd Qu.:0.7720
##  Max.   :89.00   Max.   :723.0   Max.   :17.870   Max.   :0.9480
##    schooling          gdp
##  Min.   : 0.00   Min.   :     1.68
##  1st Qu.:10.30   1st Qu.:   580.49
##  Median :12.10   Median :  3116.56
##  Mean   :11.99   Mean   :  7483.16
##  3rd Qu.:14.10   3rd Qu.:  7483.16
##  Max.   :20.70   Max.   :119172.74
```

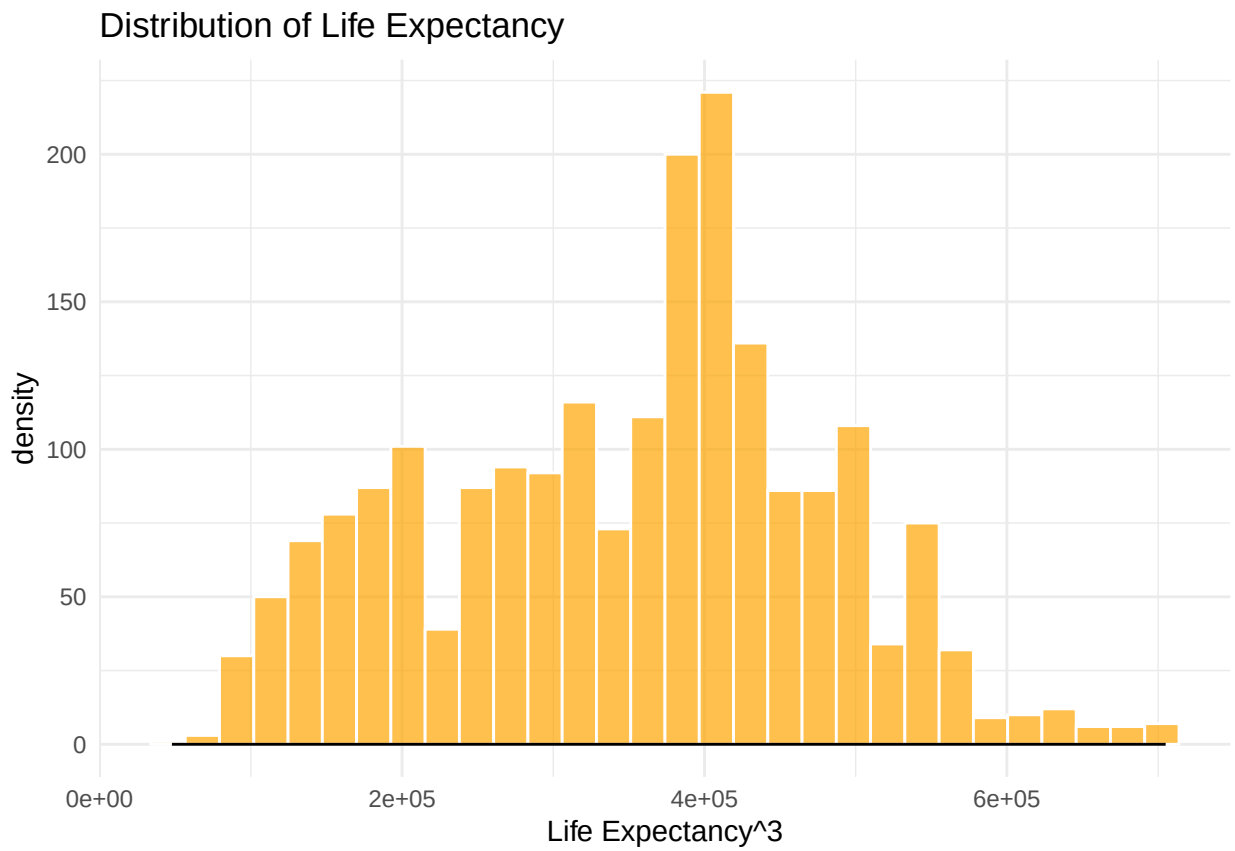
```
print(dim(life_exp2))
```

```
## [1] 2938    6
```

```r
set.seed(999)
splitIndex <- createDataPartition(life_exp2$life_exp, p = 0.7, list = FALSE)
train_data <- life_exp2[splitIndex, ]
test_data <- life_exp2[-splitIndex, ]

ggplot(data = train_data, aes(x = (life_exp))) +
  geom_histogram( fill = 'orange', color = 'white', alpha = 0.7) +
  geom_density(alpha = 0.2, fill = 'blue') +
  labs(title = 'Distribution of Life Expectancy', x = 'Life Expectancy') +
  theme_minimal()
```




```r
ggplot(data = train_data, aes(x = (life_exp)^3)) +
  geom_histogram( fill = 'orange', color = 'white', alpha = 0.7) +
  geom_density(alpha = 0.2, fill = 'blue') +
  labs(title = 'Distribution of Life Expectancy', x = 'Life Expectancy^3') +
  theme_minimal()
```

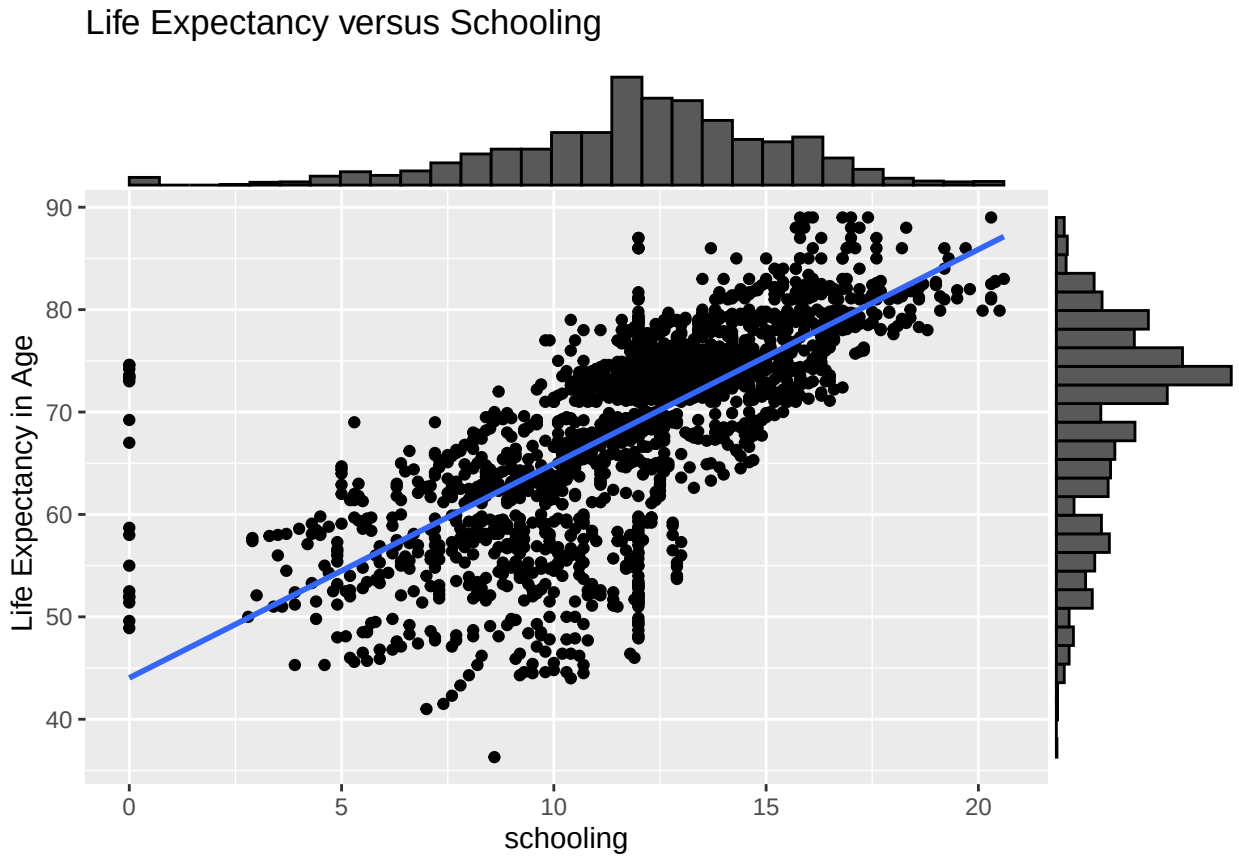


SLR Based on our Predictor of Interest: Predictors of Interest: Schooling, income composition, gdp, alcohol Assumptions: Independence: All participant data is independent of each other, as each value for country is also determined by the year and doesnot depend on its subsequent years.

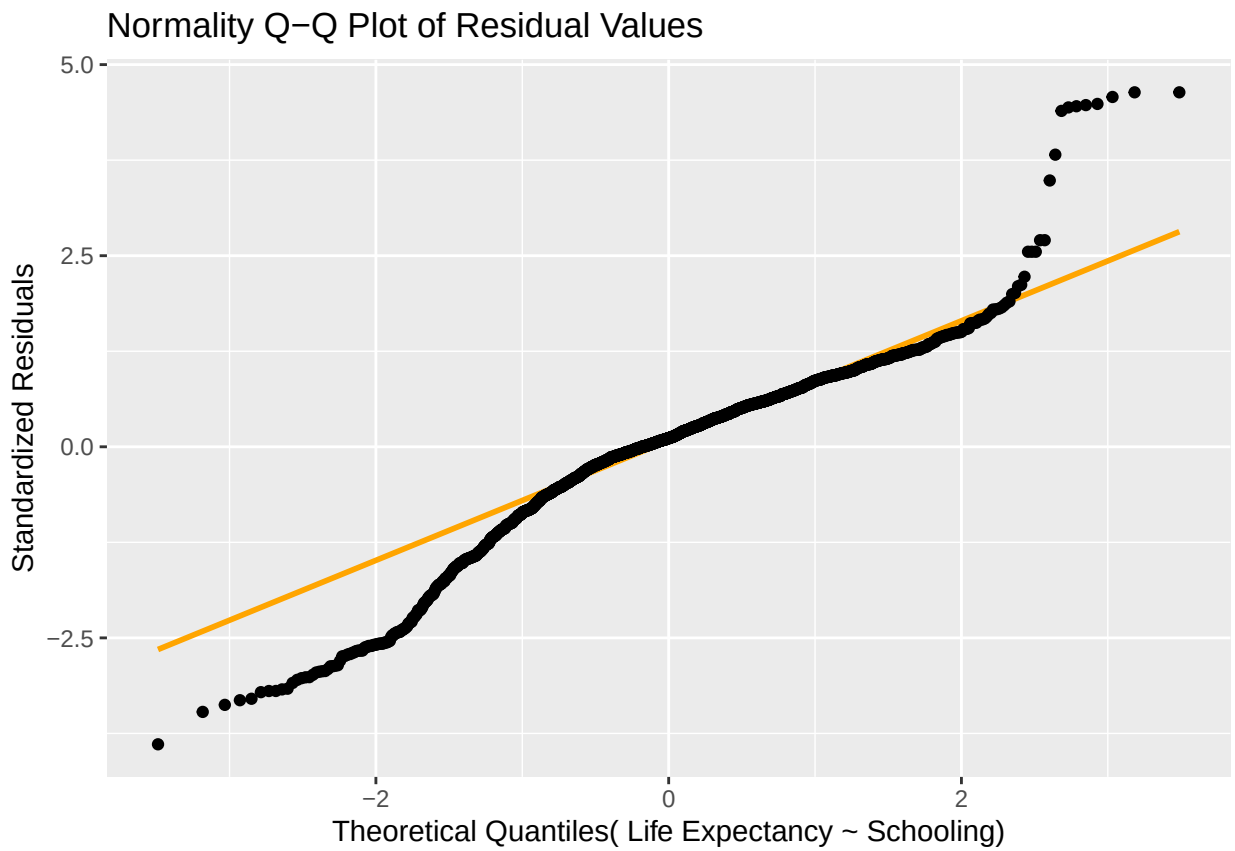
```
lm_scl <- lm(life_exp ~ schooling, data = train_data)
summary(lm_scl)
```

```
##
## Call:
## lm(formula = life_exp ~ schooling, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.7303  -2.9584   0.7479   4.0291  30.5613
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.03871    0.55688   79.08   <2e-16 ***
## schooling    2.09205    0.04468   46.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.614 on 2057 degrees of freedom
## Multiple R-squared:  0.516,  Adjusted R-squared:  0.5157
## F-statistic:  2193 on 1 and 2057 DF,  p-value: < 2.2e-16
```

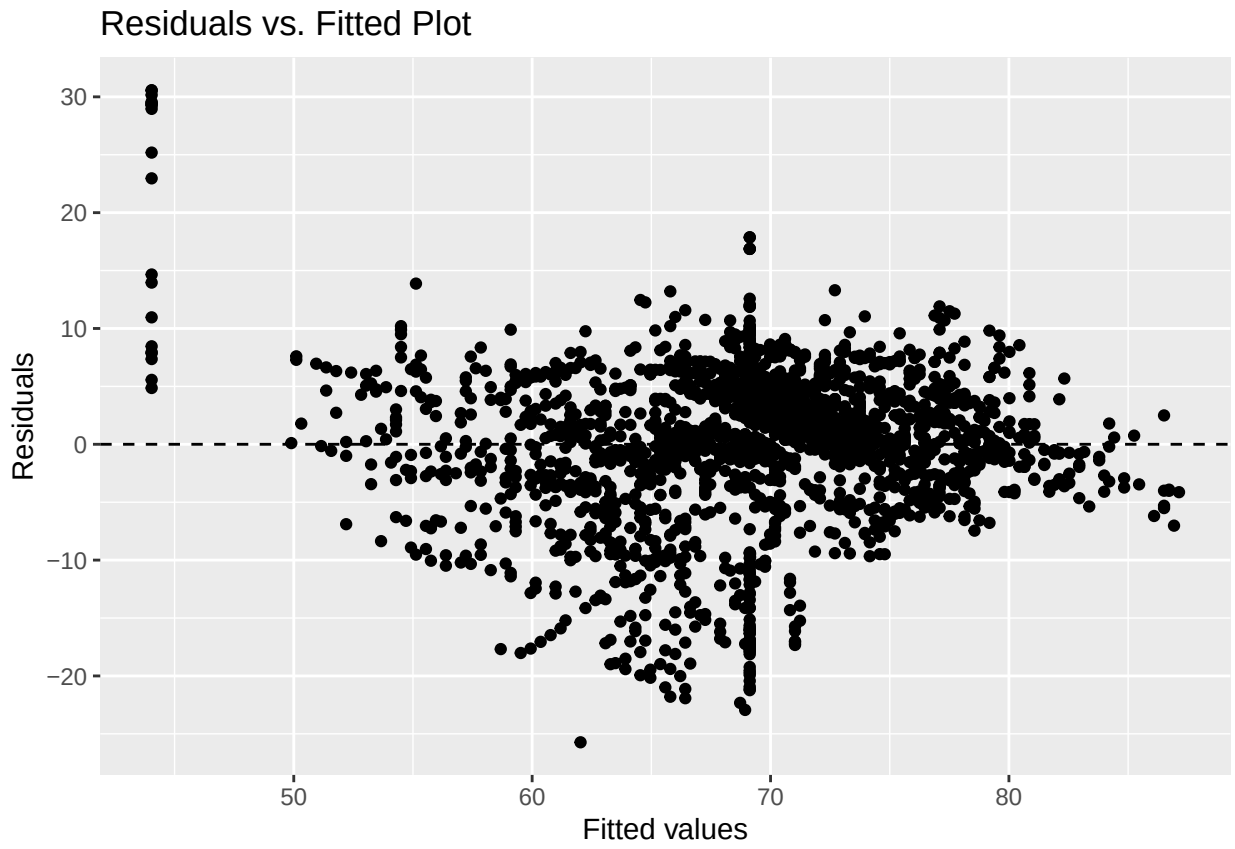
```r
# Linearity: From Scatter plot linearity is satisfied.
plot_1 <- ggplot(data = train_data, aes(y = life_exp, x = schooling)) +
          geom_point() +
          geom_smooth(method = "lm", se = FALSE)+
  ggtitle("Life Expectancy versus Schooling")+
  labs(X = "Schooling In years", y = " Life Expectancy in Age")
ggExtra::ggMarginal(plot_1, type = "histogram")
```




```r
#Normality: Though the plot seems to be good, there are many points that pull the lower tail down.
residuals <- rstandard(lm_scl)
residuals_df <- data.frame(std_residuals = residuals)

# Then use it directly in the ggplot call
ggplot(data = residuals_df, aes(sample = residuals)) +
  stat_qq_line(linewidth = 1, color = "orange") +
  stat_qq() +
  ggtitle("Normality Q-Q Plot of Residual Values")+
  labs(y = "Standardized Residuals", x = "Theoretical Quantiles( Life Expectancy ~ Schooling)")
```




```
# from The plot we can see that Life expectancy values are left skewed(from
#bottom to top), and voilate normal distribution, similarly we can see that schooling is also left skew
# so we may nee

#Homoscedacity: Variance is not distributed clearly.
# Residuals vs. Fitted Plot (Homoscedasticity)
  fitted_values <- fitted(lm_scl)
  resid_values <- resid(lm_scl)
  ggplot(data.frame(fitted_values, resid_values),
                                aes(x = fitted_values, y = resid_values)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    ggtitle("Residuals vs. Fitted Plot")+
    labs(x = "Fitted values", y = "Residuals")
```

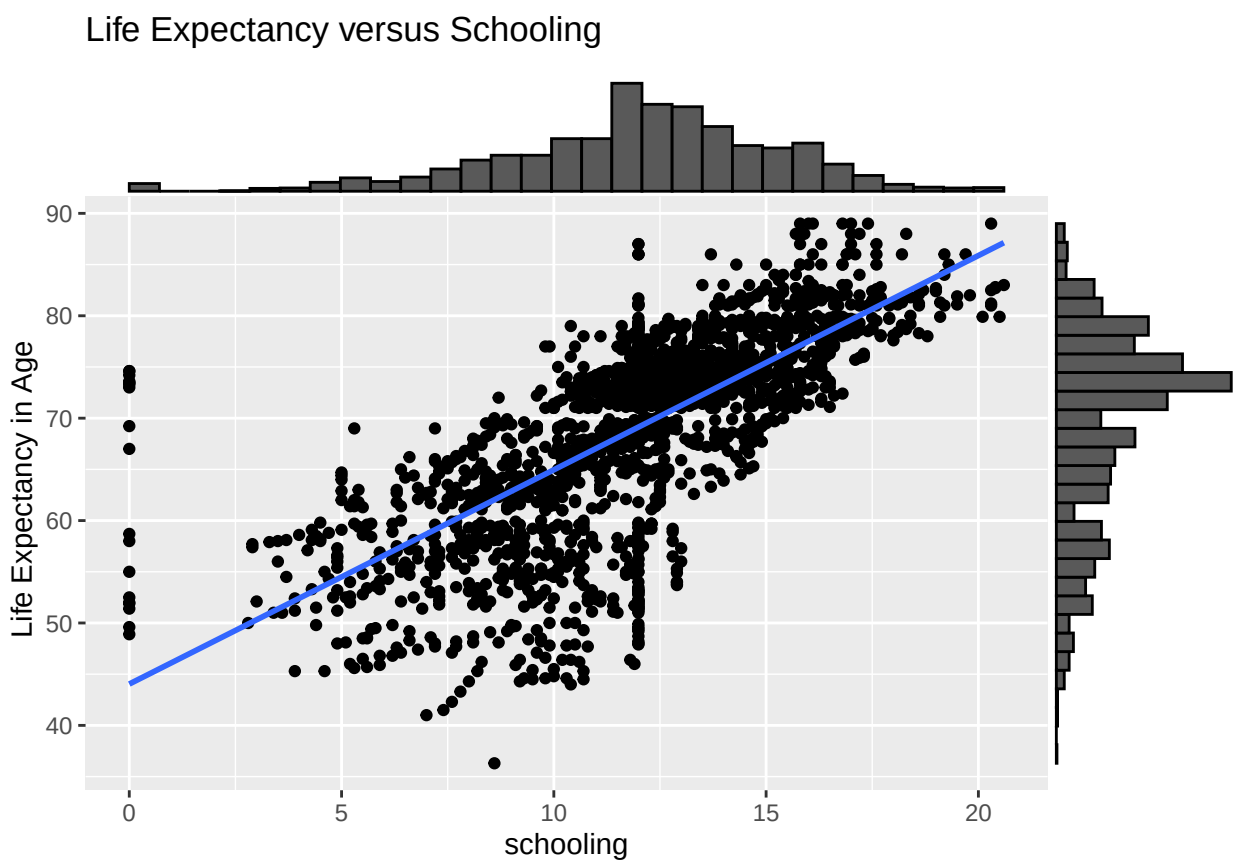



```
lm_scl_2 <- lm(life_exp^3 ~ (schooling)^3, data = train_data)
summary(lm_scl_2)
```

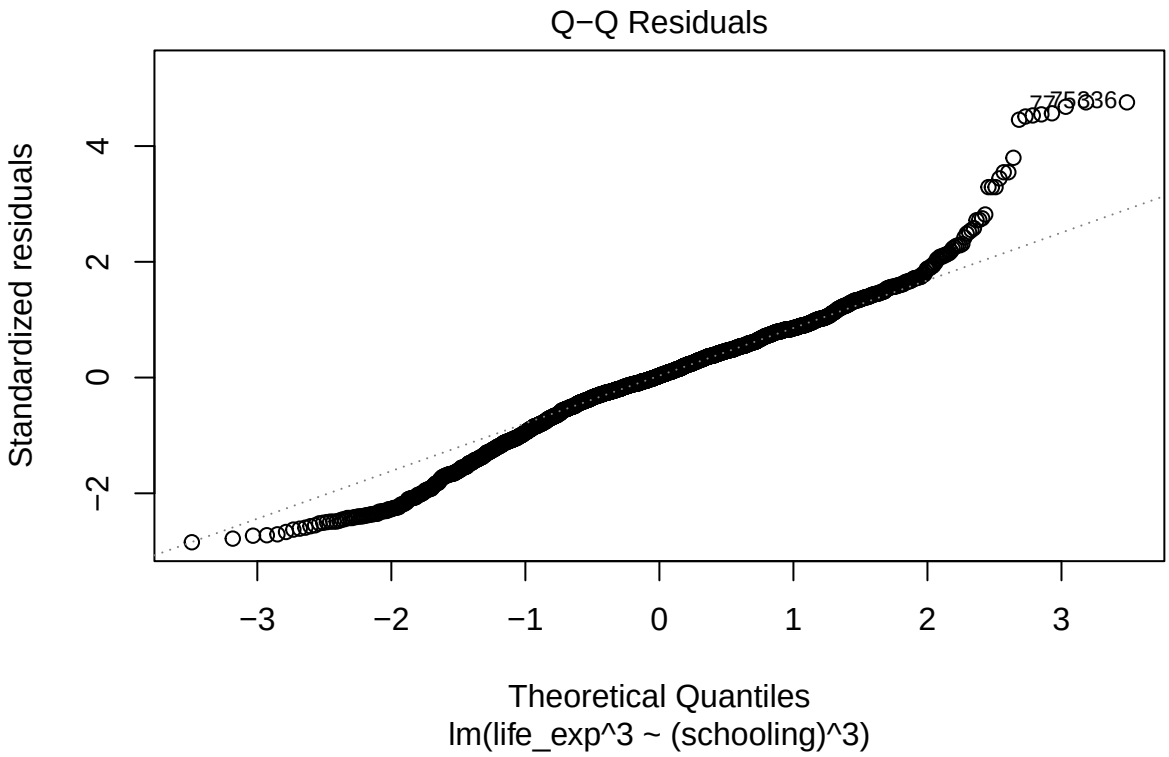
```
##
## Call:
## lm(formula = life_exp^3 ~ (schooling)^3, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -248560  -45648    2663   51406  414056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1105.1     7357.5    0.15    0.881
## schooling    28974.0      590.3   49.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87380 on 2057 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5392
## F-statistic:  2409 on 1 and 2057 DF,  p-value: < 2.2e-16
```

```
# Linearity: From Scatter plot linearity is satisfied.
plot_2 <- ggplot(data = train_data, aes(y = life_exp^3, x = (schooling)^3) )+
          geom_point() +
```

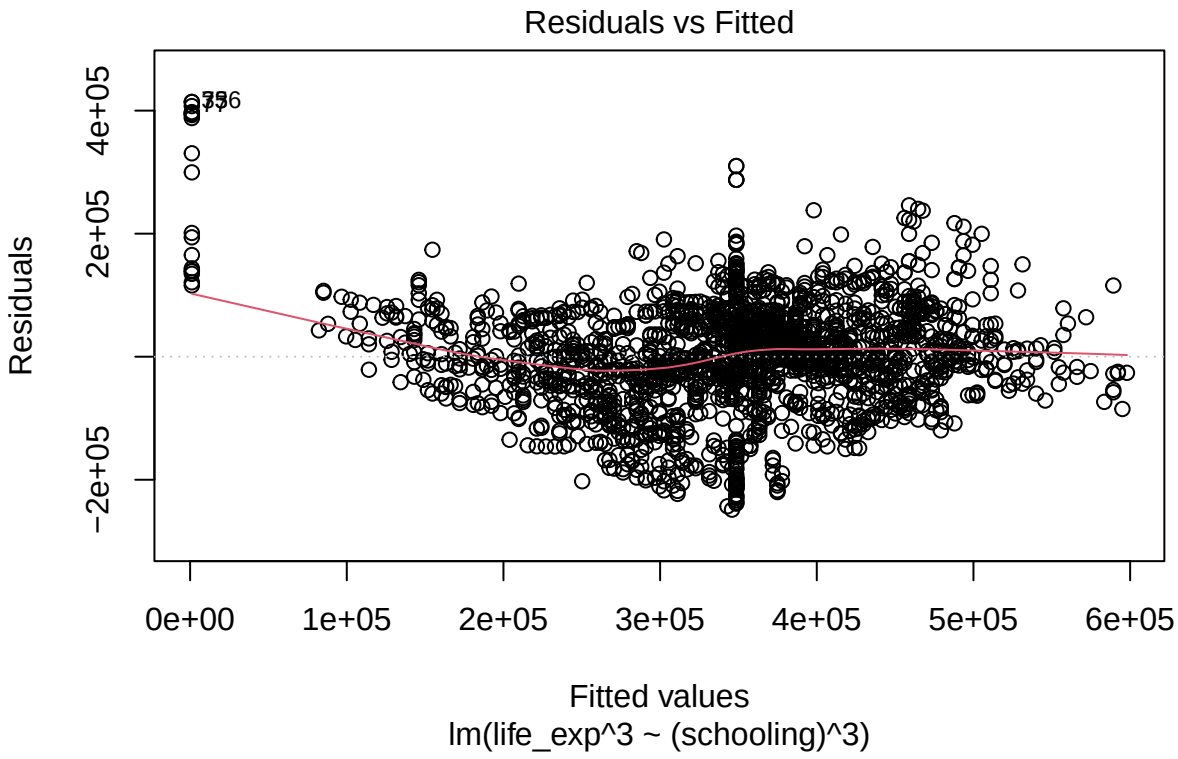
```
          geom_smooth(method = "lm", se = FALSE)
ggExtra::ggMarginal(plot_1, type = "histogram")
```




```
#Normality: Though the plot seems to be good, there are many points that pull the lower tail down.
plot(lm_scl_2, which = 2)
```




```
#Homoscedacity: Variance is not distributed clearly.
plot(lm_scl_2, which = 1)
```



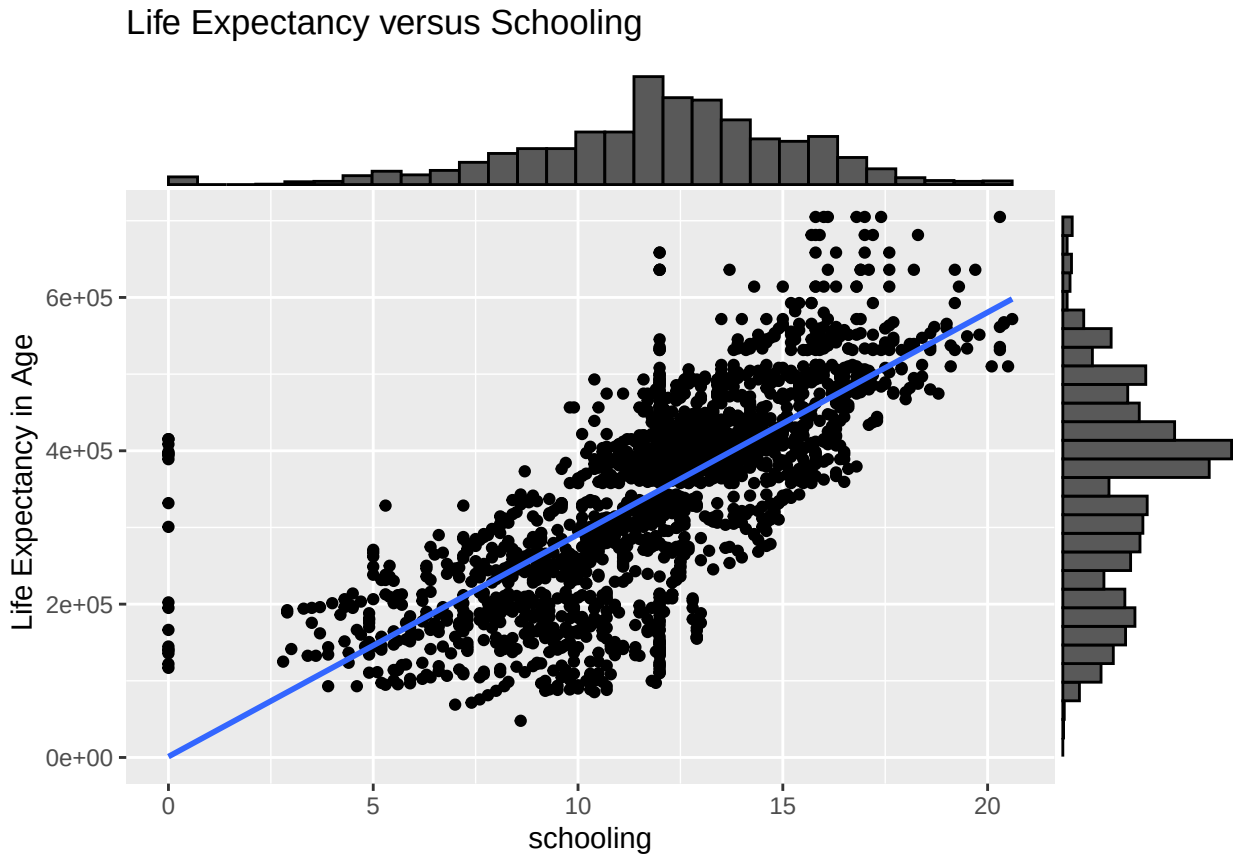
Final Model

```
set.seed(999)



remove_high_leverage <- function(model, data) {
  n <- nrow(data)
  p <- length(coef(model))  # Number of predictors including intercept
  leverage_threshold <- 2 * (p + 1) / n

  leverage_points <- which(hatvalues(model) > leverage_threshold)
  data_cleaned <- data[-leverage_points, ]

  return(data_cleaned)
}
life_exp2_cleaned <- remove_high_leverage(lm_scl_2, train_data)

lm_scl_2 <- lm(life_exp^3 ~ schooling, data = life_exp2_cleaned)
summary(lm_scl_2)
```

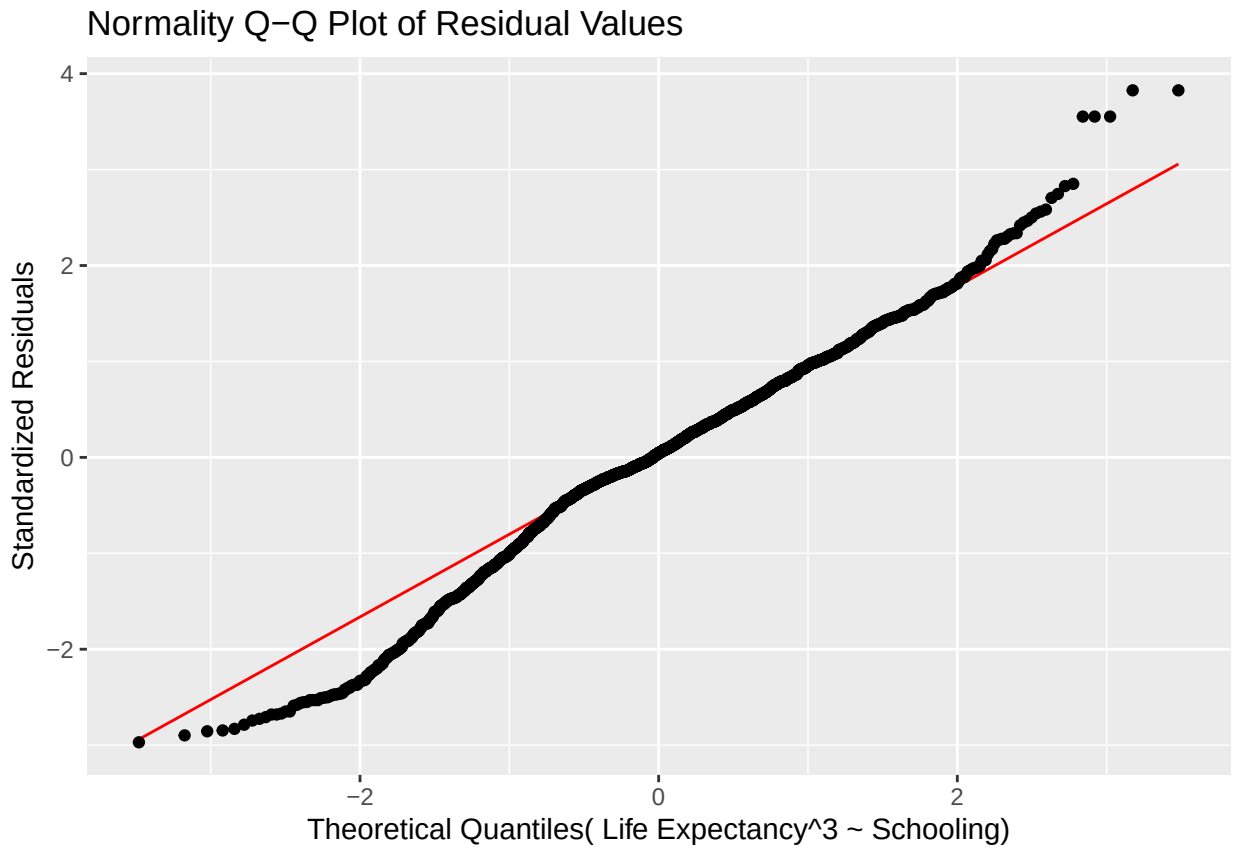
```
##
## Call:
```

```
## lm(formula = life_exp^3 ~ schooling, data = life_exp2_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243876  -42809    3762   52605  314159
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -60418.1     8056.1   -7.50 9.57e-14 ***
## schooling    33750.5      642.5   52.53  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82130 on 1996 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5801
## F-statistic:  2760 on 1 and 1996 DF,  p-value: < 2.2e-16
```

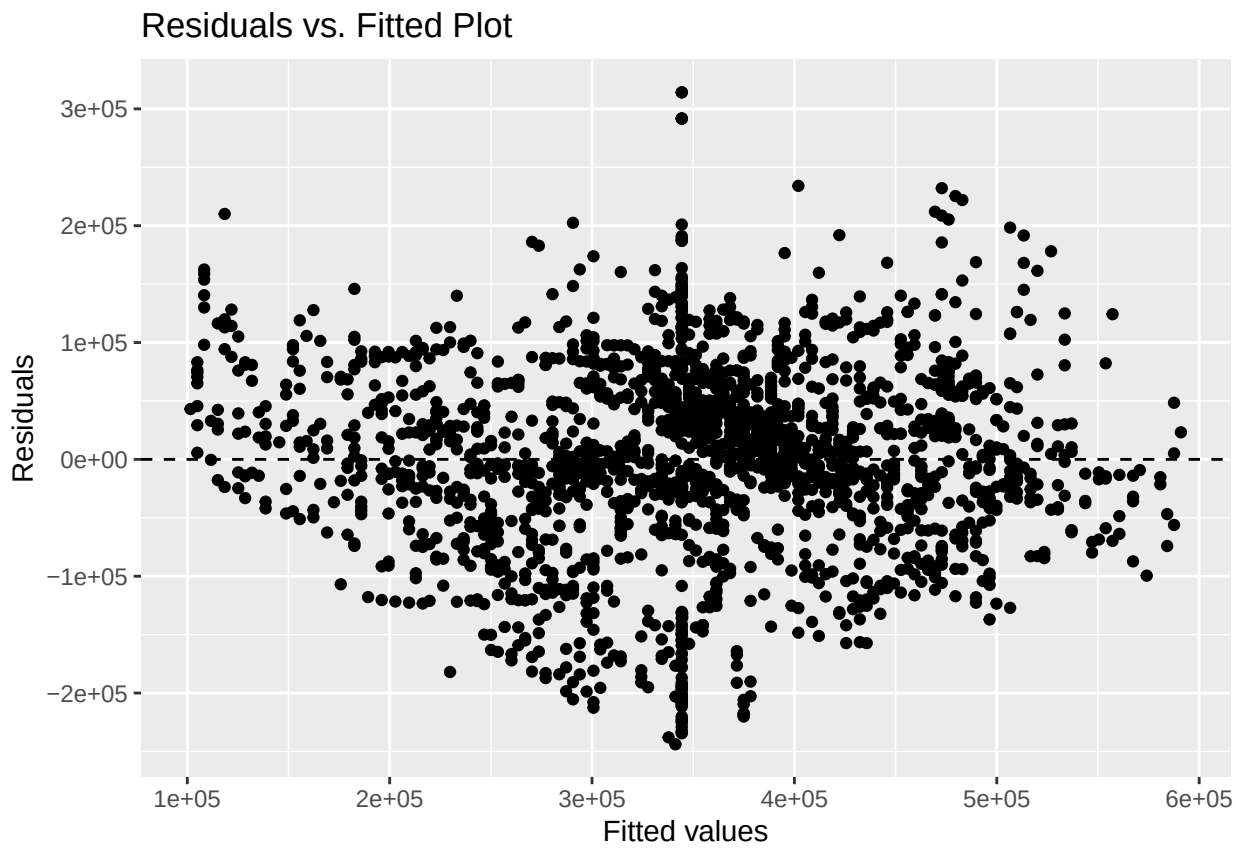
```
create_plots <- function(model, data) {
  # Linearity Plot
  summary_model <- summary(model)
  plot_linearity <- ggplot(data, aes(x = schooling, y = (life_exp^3))) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)+
     ggtitle("Life Expectancy versus Schooling")+
  labs(X = "Schooling In years", y = " Life Expectancy in Age")
  linearity_plot <- ggExtra::ggMarginal(plot_linearity, type = "histogram")

  # Q-Q Plot (Normality)

 residuals <- rstandard(model)
residuals_df <- data.frame(std_residuals = residuals)

# Then use it directly in the ggplot call
plot_qq <- ggplot(data = residuals_df, aes(sample = residuals)) +
  stat_qq_line(color = "red") +
  stat_qq()+
  ggtitle("Normality Q-Q Plot of Residual Values")+
  labs(y = "Standardized Residuals", x = "Theoretical Quantiles( Life Expectancy^3 ~ Schooling)")
  # Residuals vs. Fitted Plot (Homoscedasticity)
  fitted_values <- fitted(model)
  resid_values <- resid(model)
  plot_resid_vs_fitted <- ggplot(data.frame(fitted_values, resid_values),
                                 aes(x = fitted_values, y = resid_values)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    ggtitle("Residuals vs. Fitted Plot")+
    labs(x = "Fitted values", y = "Residuals")
  return(list(summary = summary_model,linearity = linearity_plot, qq = plot_qq, resid_vs_fitted = plot_
}

# Example usage

create_plots(lm_scl_2, train_data)
```




```
## $summary
##
## Call:
## lm(formula = life_exp^3 ~ schooling, data = life_exp2_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243876  -42809    3762   52605  314159
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -60418.1     8056.1   -7.50 9.57e-14 ***
## schooling    33750.5      642.5   52.53  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82130 on 1996 degrees of freedom
## Multiple R-squared:  0.5803, Adjusted R-squared:  0.5801
## F-statistic:  2760 on 1 and 1996 DF,  p-value: < 2.2e-16
##
##
## $linearity
##
## $qq
```




```
##
## $resid_vs_fitted
```



Accuracy of the Model:

Life.Expectancy VS Alcohol

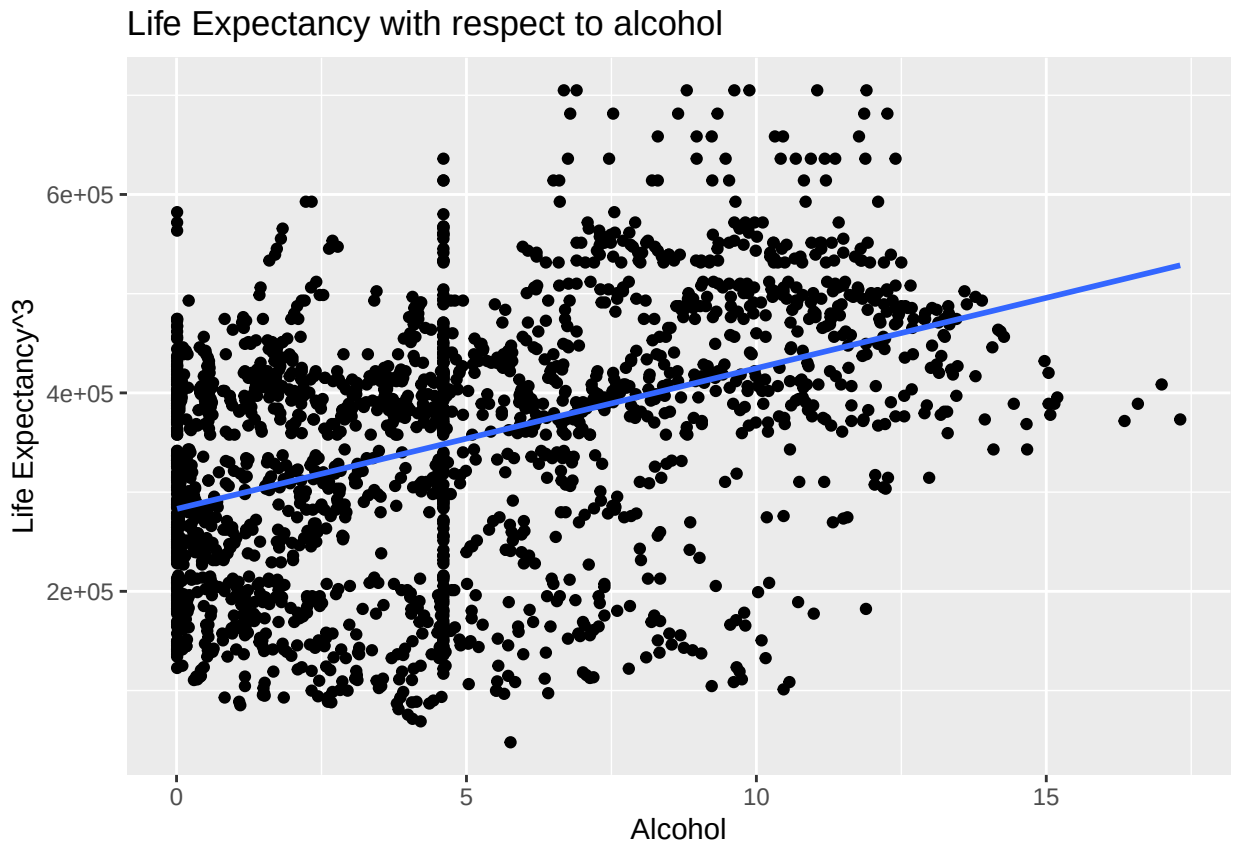
Model 1:

Linearity: Scatter plot - Not satisfied

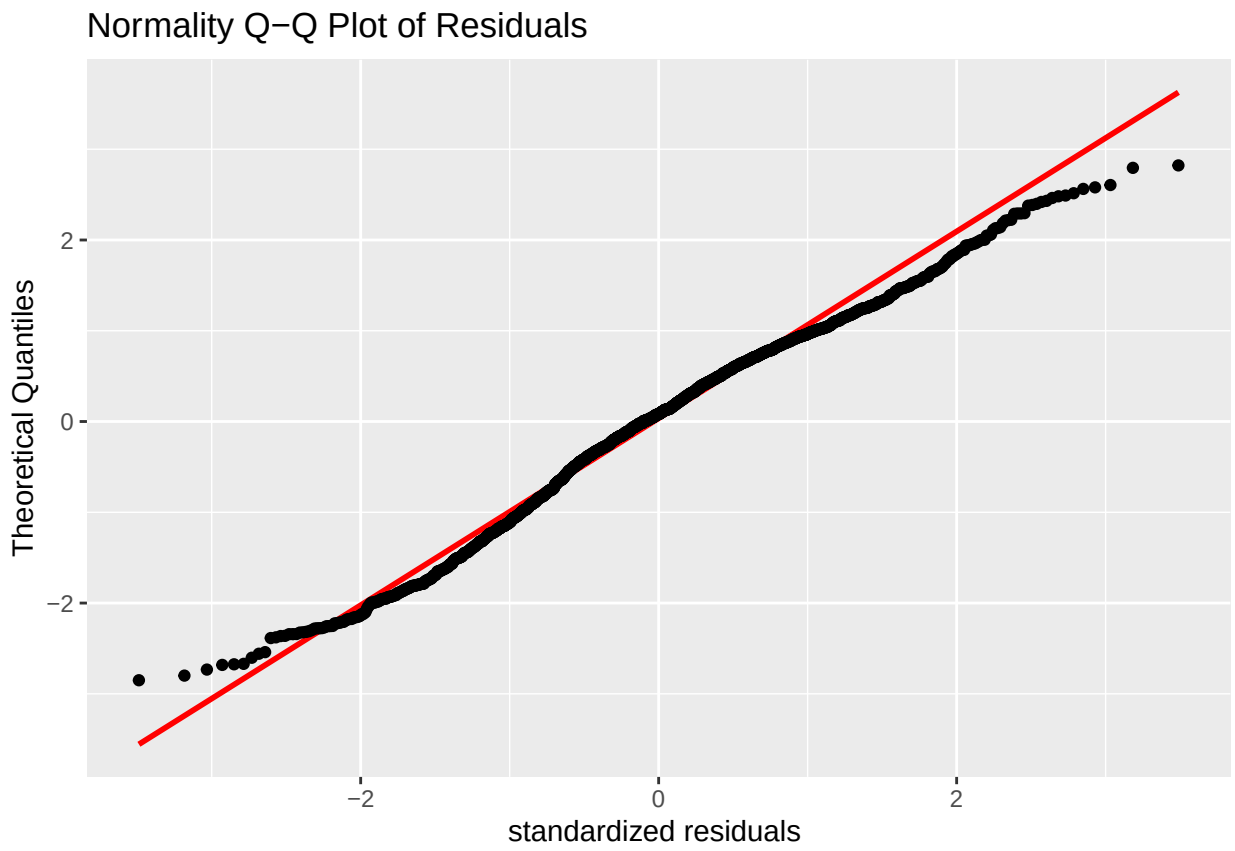
```
lm_ach <- lm(data = train_data, life_exp^3 ~ (alcohol))
create_plots <- function(model, data) {
# Linearity Plot
  summary_model <- summary(model)
  plot_linearity <- ggplot(data, aes(x = alcohol, y = life_exp^3)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)+
    ggtitle("Life Expectancy with respect to alcohol")+
    labs(x = "Alcohol", y = "Life Expectancy^3")
 # linearity_plot <- ggExtra::ggMarginal(plot_linearity, type = "histogram")

# Q-Q Plot (Normality)

plot_qq <- ggplot(data = data, aes(sample = rstandard(model))) +
  stat_qq_line(linewidth = 1, col = "red") +
  stat_qq()+
  ggtitle("Normality Q-Q Plot of Residuals")+
  labs(x = "standardized residuals ", y = "Theoretical Quantiles")
```

```r
# Residuals vs. Fitted Plot (Homoscedasticity)
  fitted_values <- fitted(model)
  resid_values <- resid(model)
  plot_resid_vs_fitted <- ggplot(data.frame(fitted_values, resid_values),
                                 aes(x = fitted_values, y = resid_values)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    ggtitle("Residuals vs. Fitted values Plot")+
    labs(x = "Fitted Values", y = "Residuals")

  return(list(summary = summary_model,linearity = plot_linearity, qq = plot_qq, resid_vs_fitted = plot_
}

# Example usage

create_plots(lm_ach, train_data)
```

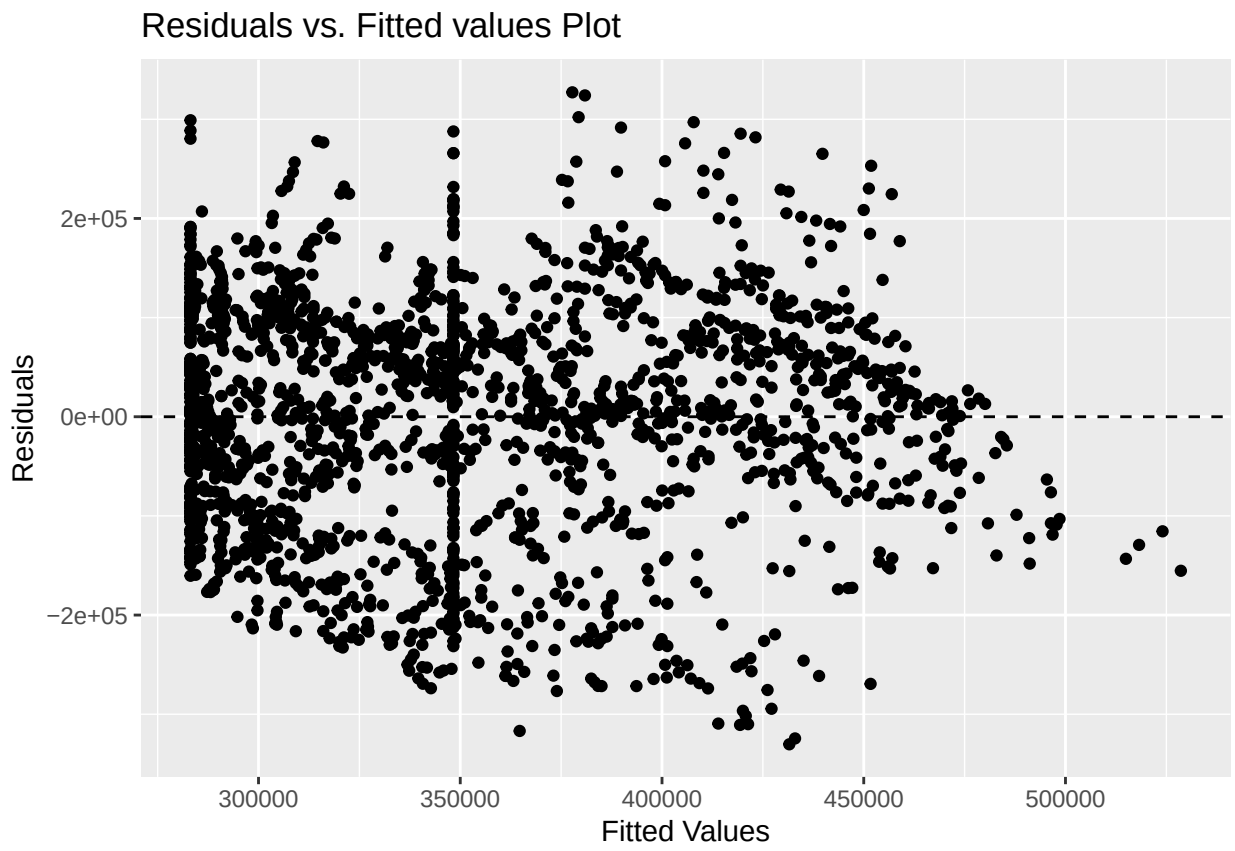
```
## $summary
##
## Call:
## lm(formula = life_exp^3 ~ (alcohol), data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -330357  -76382    9612   84616  327191
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 283000.7     3979.5   71.11   <2e-16 ***
## alcohol      14188.2      649.2   21.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 116000 on 2057 degrees of freedom
## Multiple R-squared:  0.1885, Adjusted R-squared:  0.1881
## F-statistic: 477.7 on 1 and 2057 DF,  p-value: < 2.2e-16
##
##
## $linearity
```


Life Expectancy with respect to alcohol
Life Expectancy^3
6e+05
4e+05
2e+05
0
5
10
15
Alcohol

```
##
## $qq
```



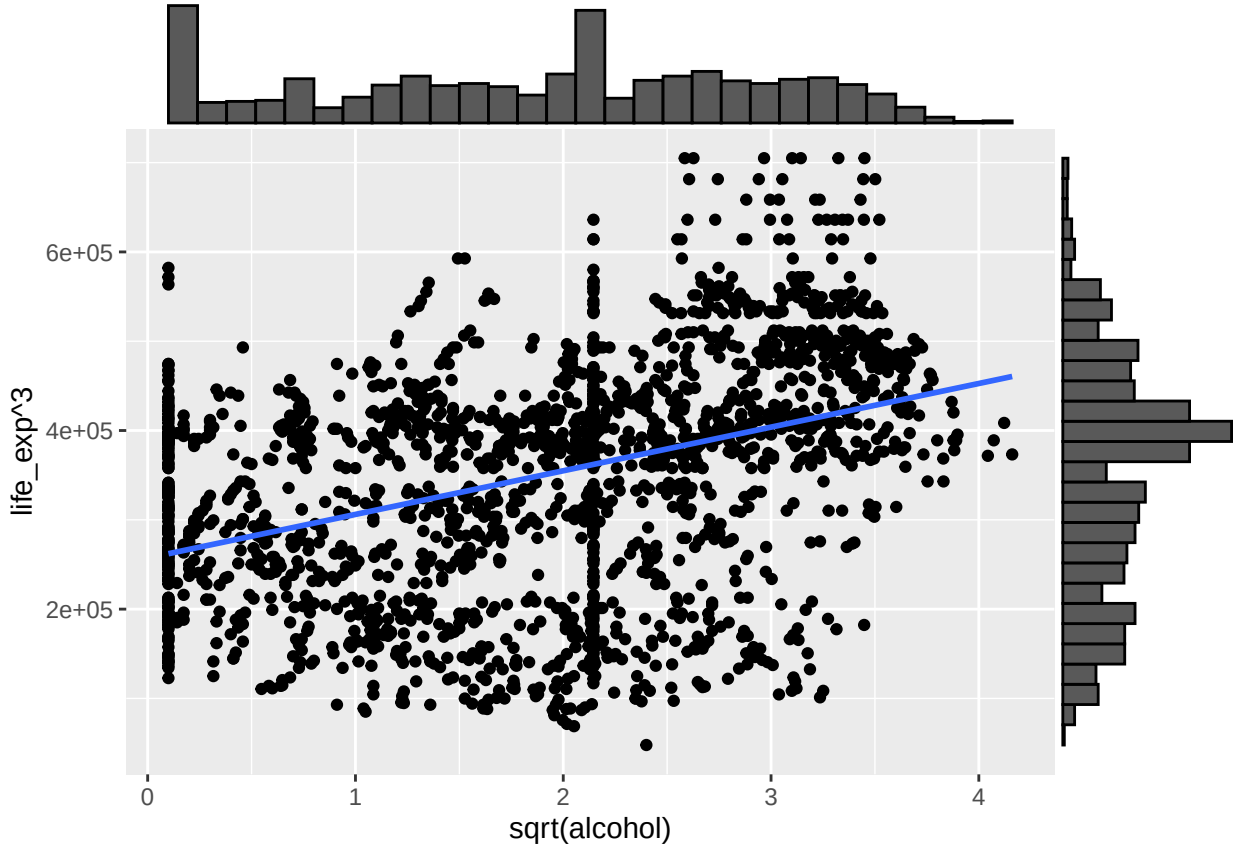

```
##
## $resid_vs_fitted
```



Final Model :

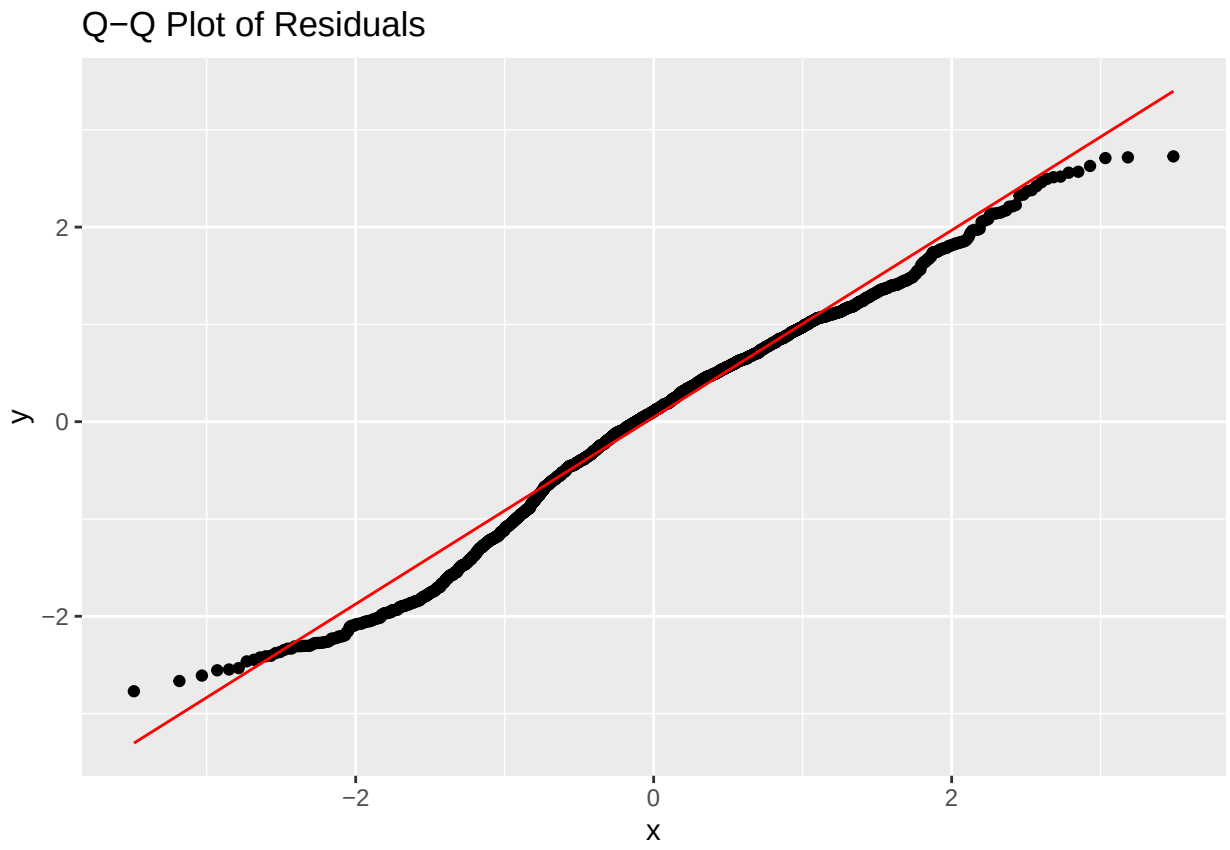
```
set.seed(999)
#train_data$cube_alcohol<- train_data$alcohol
lm_ach_reduced <- lm(data = train_data, life_exp^3 ~ sqrt(alcohol))

create_plots <- function(model, data) {
# Linearity Plot
  summary_model <- summary(model)
  plot_linearity <- ggplot(data, aes(x =sqrt(alcohol), y = life_exp^3)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE)
  linearity_plot <- ggExtra::ggMarginal(plot_linearity, type = "histogram")

# Q-Q Plot (Normality)
residuals <- rstandard(model)
residuals_df <- data.frame(std_residuals = residuals)

# Then use it directly in the ggplot call
plot_qq <- ggplot(data = residuals_df, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  ggtitle("Q-Q Plot of Residuals")
  #histogram
gg_hist<- ggplot(data = data, aes(x = sqrt(alcohol))) +
  geom_histogram( fill = 'orange', color = 'white', alpha = 0.7) +
```

```
  geom_density(alpha = 0.2, fill = 'blue') +
  labs(title = 'Distribution of Life Expectancy', x = 'Alcohol') +
  theme_minimal()

# Residuals vs. Fitted Plot (Homoscedasticity)
  fitted_values <- fitted(model)
  resid_values <- resid(model)
  plot_resid_vs_fitted <- ggplot(data.frame(fitted_values, resid_values),
                                 aes(x = fitted_values, y = resid_values)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    ggtitle("Residuals vs. Fitted Plot")

  return(list(summary = summary_model,linearity = linearity_plot, qq = plot_qq, histogram = gg_hist, res
}
create_plots(lm_ach_reduced, train_data)
```

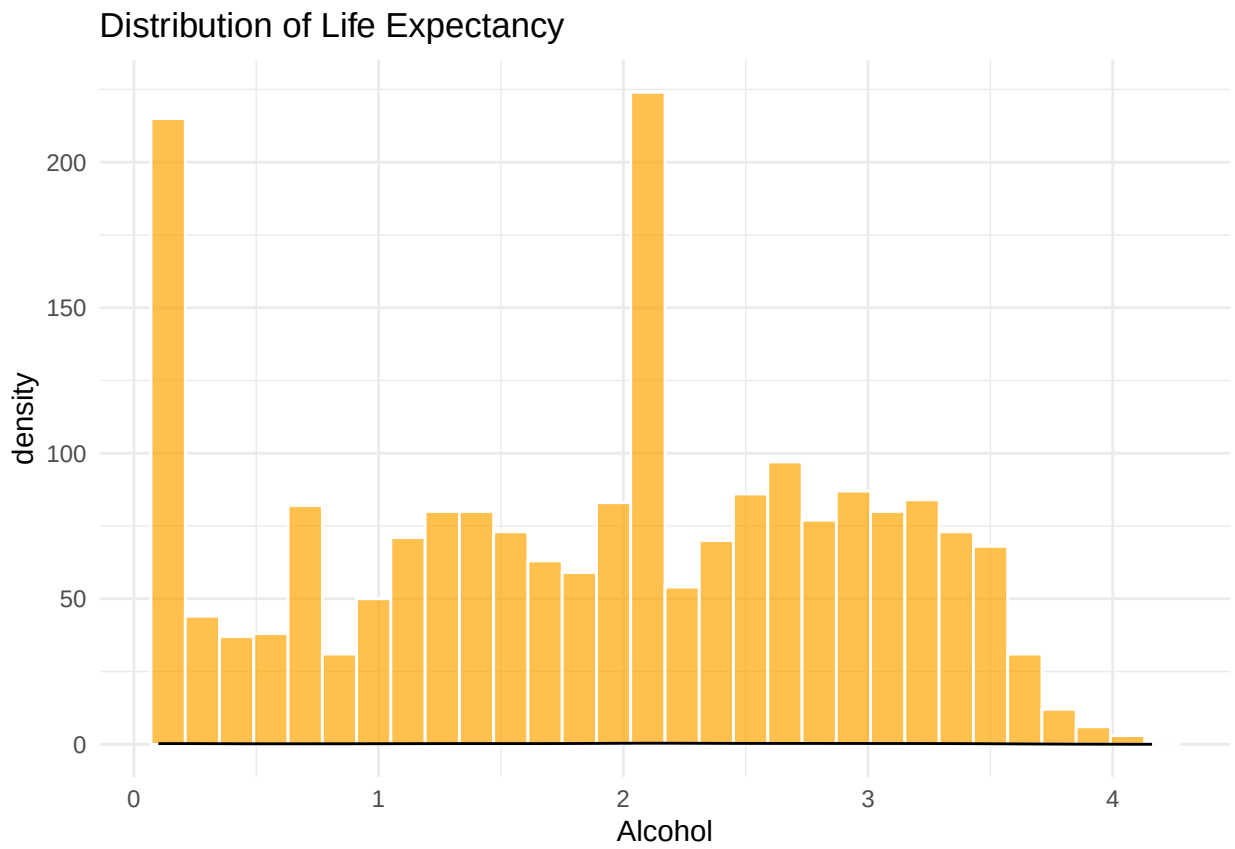



```
## $summary
##
## Call:
## lm(formula = life_exp^3 ~ sqrt(alcohol), data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -326651  -70856   12536   81852  321465
```

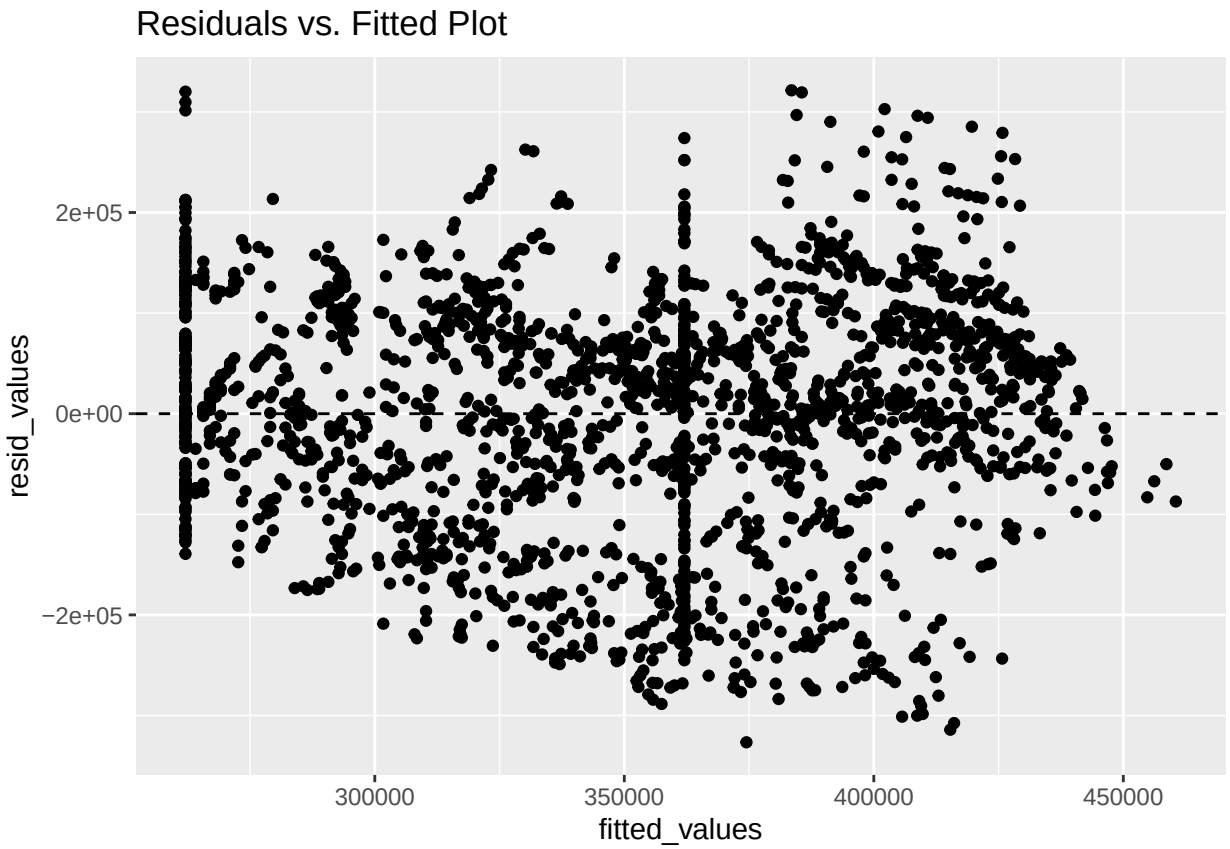
```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     257179       5327   48.27   <2e-16 ***
## sqrt(alcohol)    48876       2458   19.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117900 on 2057 degrees of freedom
## Multiple R-squared:  0.1612, Adjusted R-squared:  0.1608
## F-statistic: 395.5 on 1 and 2057 DF,  p-value: < 2.2e-16
##
##
## $linearity
##
## $qq
```




```
##
## $histogram
```




```
##
## $resid_vs_fitted
```



Conclusion:

Schooling:

Hypothesis:Null Hypothesis: H_0 there is no relationship between Life Expectancy and Schooling. i.e $\beta_1 = 0$,

and the model equation is $Y = \beta_0 + \epsilon$

Alternate Hypothesis: H_A There is a relationship between life Expectancy and Schooling. i.e $\beta_1 \neq 0$,

and the model equation is $Y = \beta_0 + \beta_1 * X_1 + \epsilon$

Linearity: satisfied.

Independence: all the data is independent of each other.

Normality: From the qqplot we can see the plot is almost normal, eventhough we have the tails are little distorted.

T-test statistic: From above the test statistic value for life expectancy to schooling is 62.27, and

P value is $2.2{*}10^{-16}$ as p value is nearer to 0 we reject the Null Hypothesis H_0 : and conclude that there is relationship between Life Expectancy and Schooling.

Conclusion: The final estimated model equation is $\hat{lifeexpectancy}^3 = 56\hat{6}8 + 28\hat{6}62 * Schooling + \epsilon$

Interpretation:

Intercept (939.58) : the estimated value of squared life expectancy value when Schooling is 0.

Slope(325.81) : The estimated change in squared value of life expectancy for one unit change in value of schooling. in general, for every year increase in schooling the squared life expectancy value is expected to increase by 325.81.

Alcohol: Hypothesis:Null Hypothesis: H_0 there is no relationship between Life Expectancy and Schooling. i.e $\beta_2 = 0$,

and the model equation is $Y = \beta_0 + \epsilon$

Alternate Hypothesis: H_A There is a relationship between life Expectancy and Schooling. i.e $\beta_2 \neq 0$,

and the model equation is $Y = \beta_0 + \beta_2 * X_2 + \epsilon$

Linearity: satisfied.

Independence: all the data is independent of each other.

Normality: From the qqplot we can see the plot is almost normal, eventhough we have the tails are little distorted.

T-test statistic: From above the test statistic value for life expectancy to schooling is 19.86, and

P value is 2.2*10^{-16} as p value is nearer to 0 we reject the Null Hypothesis H_0 : and conclude that there is relationship between Life Expectancy and Schooling.so we reject the Null Hypothesis and conclude that there is a relationship between Alcohol and Life_expectancy.

and the relationship is positive.

Conclusion: The final estimated model equation is $\hat{lifeexpectancy}^3 = \hat{257179} + \hat{48876} * Alcohol + \epsilon$

Interpretation:

intercept ($\hat{257179}$): This coefficient is the estimated cube of life expectancy when the square root of alcohol consumption is zero. It represents the starting point of the relationship between the cubic life expectancy and the square root of alcohol consumption according to the model's fit to the data.

Slope($\hat{48876}$): This coefficient indicates the amount of change in the cube of life expectancy for each one-unit increase in the square root of alcohol consumption. It suggests that if the square root of alcohol consumption increases by one unit (which corresponds to alcohol consumption itself increasing by the square of that amount), the model predicts an increase of 48876 in the cube of life expectancy.

MLR(Multi Linear regression for all the traits included).

MLR : Here the model equation for Multi linear regression is

$$Y = \beta_0 + \beta_1 * X_1 + \beta_3 * X_2 + \beta_4 * X_3 + \beta_5 * X_4 + \beta_6 * X_5 + \beta_7 * X_6 + \epsilon$$

Our Model Equation at begining of the Multi linear regression is :

$$LifeExpectation = \beta_0 + \beta_1 * AdultMortality + \beta_3 * Alcohol + \beta_4 * IncomeComposition + \beta_5 * Schooling + \beta_6 * Status + \beta_7 * GDP + \epsilon$$

```
#Linearityby pairs plot and lm model:
lm_multi <- lm(data = train_data, life_exp ~ adult_mortality + alcohol+ income_composition + schooling+g
summary(lm_multi)
```

```
##
## Call:
## lm(formula = life_exp ~ adult_mortality + alcohol + income_composition +
```

```
##     schooling + gdp, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4403  -2.0163   0.5219   2.7224  21.0651
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.688e+01  5.878e-01  96.770  < 2e-16 ***
## adult_mortality    -3.350e-02  1.034e-03 -32.406  < 2e-16 ***
## alcohol             1.348e-01  3.356e-02   4.017 6.10e-05 ***
## income_composition  9.517e+00  9.236e-01  10.304  < 2e-16 ***
## schooling           9.019e-01  6.117e-02  14.745  < 2e-16 ***
## gdp                 4.536e-05  9.874e-06   4.594 4.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.11 on 2053 degrees of freedom
## Multiple R-squared:  0.7116, Adjusted R-squared:  0.7109
## F-statistic:  1013 on 5 and 2053 DF,  p-value: < 2.2e-16
```

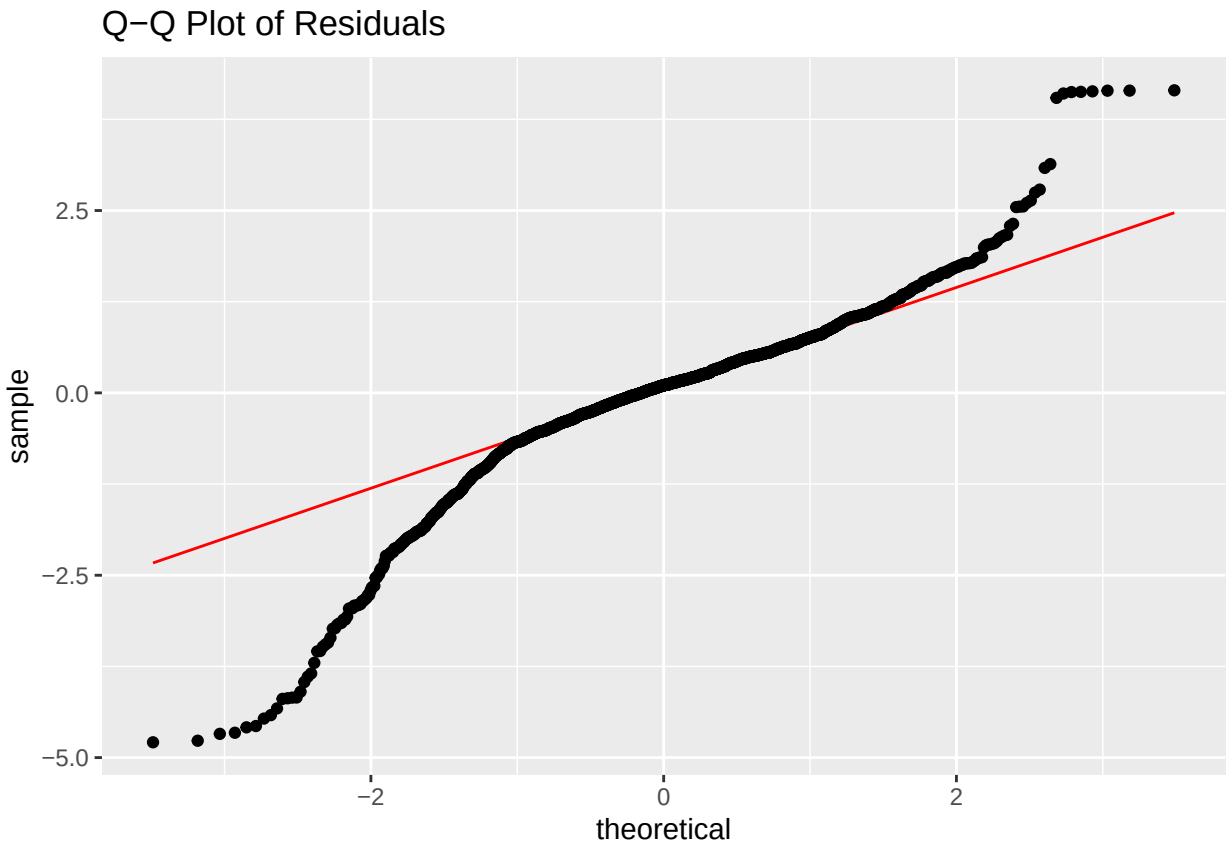
```
vif(lm_multi)
```

```
##    adult_mortality            alcohol income_composition          schooling
##           1.280764           1.376865           2.875645           3.140446
##                gdp
##           1.292302
```

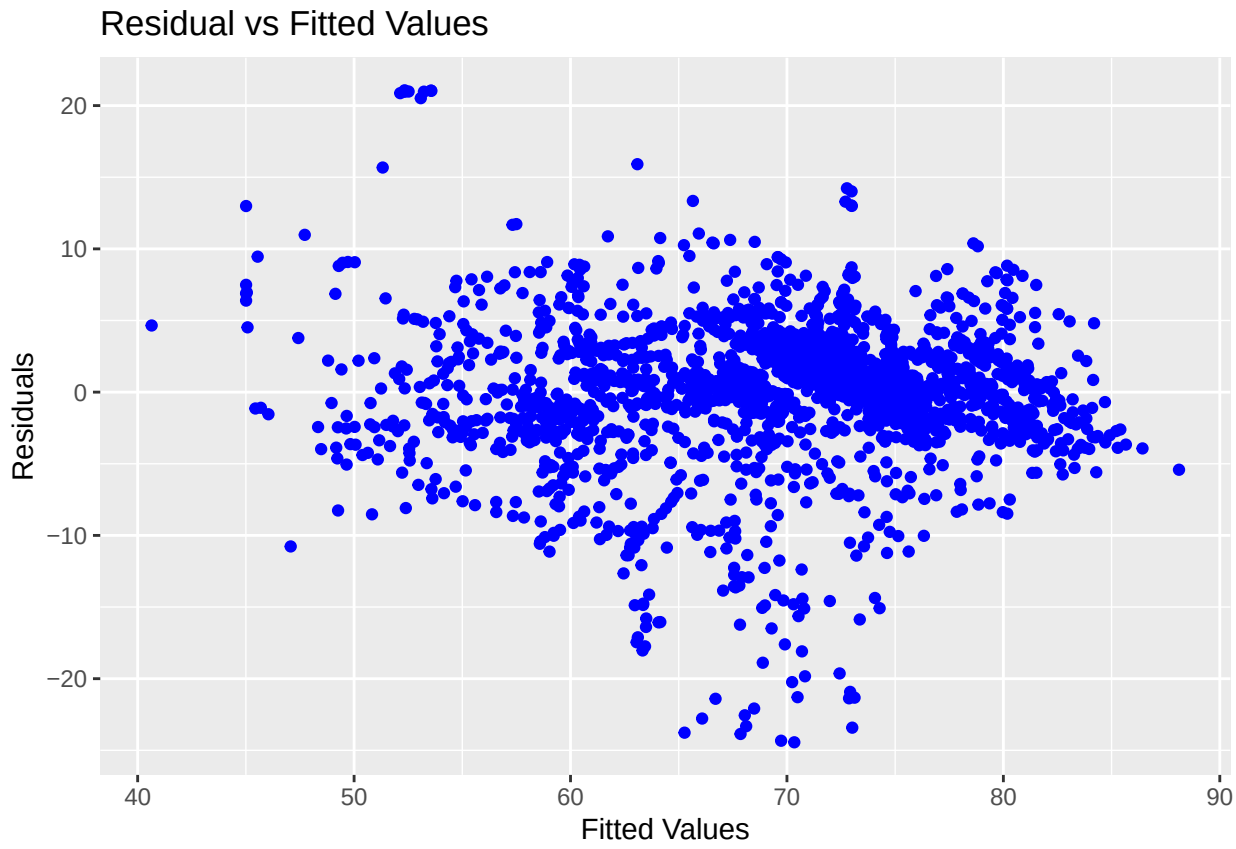
```
AIC(lm_multi)
```

```
## [1] 12568.42
```

```
#Normality
residuals <- rstandard(lm_multi)
residuals_df <- data.frame(std_residuals = residuals)

# Then use it directly in the ggplot call
ggplot(data = residuals_df, aes(sample = residuals)) +
  stat_qq_line(color = "red") +
  stat_qq() +
  ggtitle("Q-Q Plot of Residuals")
```




```
#
ggplot(data = train_data, aes(x = lm_multi$fitted.values, y = lm_multi$residuals)) +
geom_point(shape = 19, col = "blue") +
xlab("Fitted Values") +
ylab("Residuals") +
ggtitle("Residual vs Fitted Values")
```



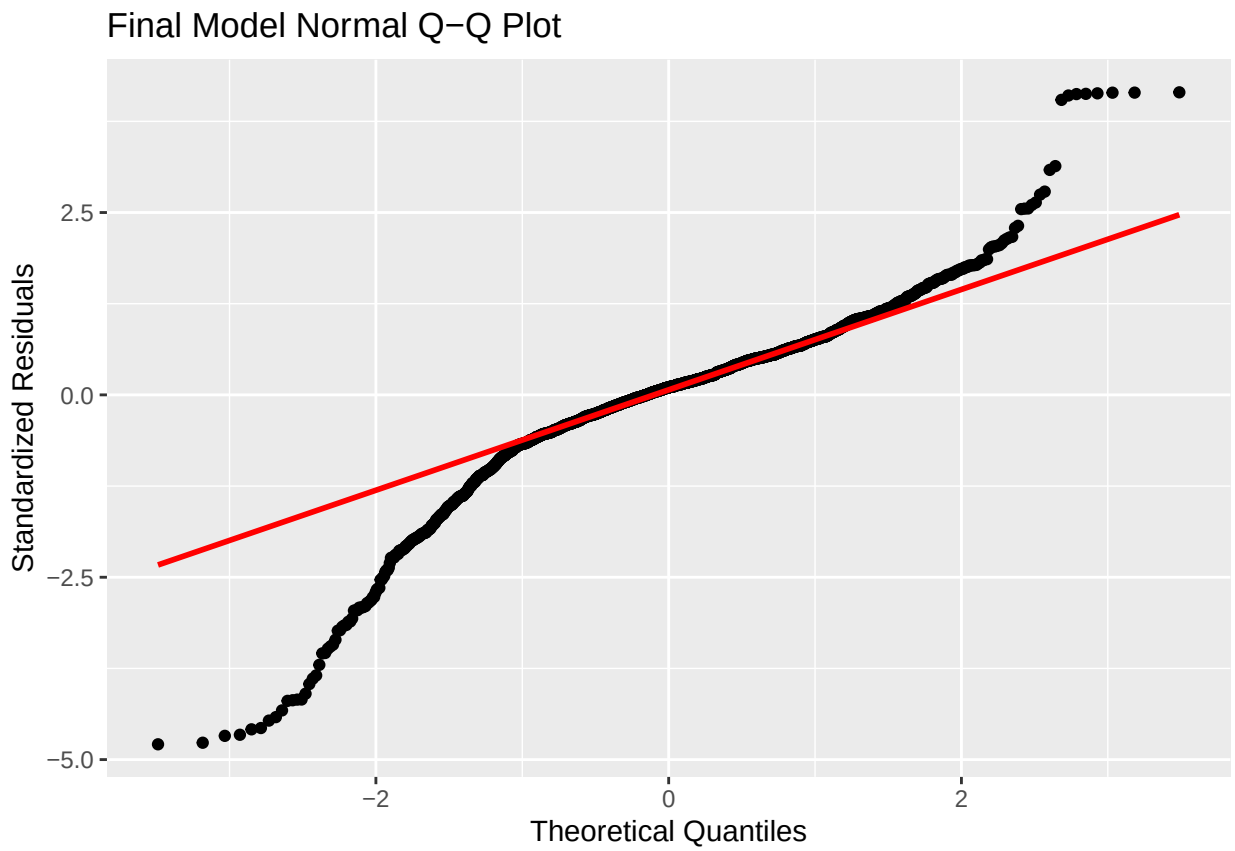

```
#Best Fit equation:
step(lm_multi, direction = "both")
```

```
## Start:  AIC=6723.23
## life_exp ~ adult_mortality + alcohol + income_composition + schooling +
##     gdp
##
##                       Df Sum of Sq   RSS    AIC
## <none>                             53607 6723.2
## - alcohol              1     421.4 54029 6737.4
## - gdp                  1     551.0 54158 6742.3
## - income_composition   1    2772.5 56380 6825.1
## - schooling            1    5676.8 59284 6928.5
## - adult_mortality      1   27421.1 81028 7571.8

##
## Call:
## lm(formula = life_exp ~ adult_mortality + alcohol + income_composition +
##     schooling + gdp, data = train_data)
##
## Coefficients:
##         (Intercept)      adult_mortality              alcohol  income_composition
##           5.688e+01           -3.350e-02            1.348e-01            9.517e+00
##           schooling                  gdp
##           9.019e-01            4.536e-05
```

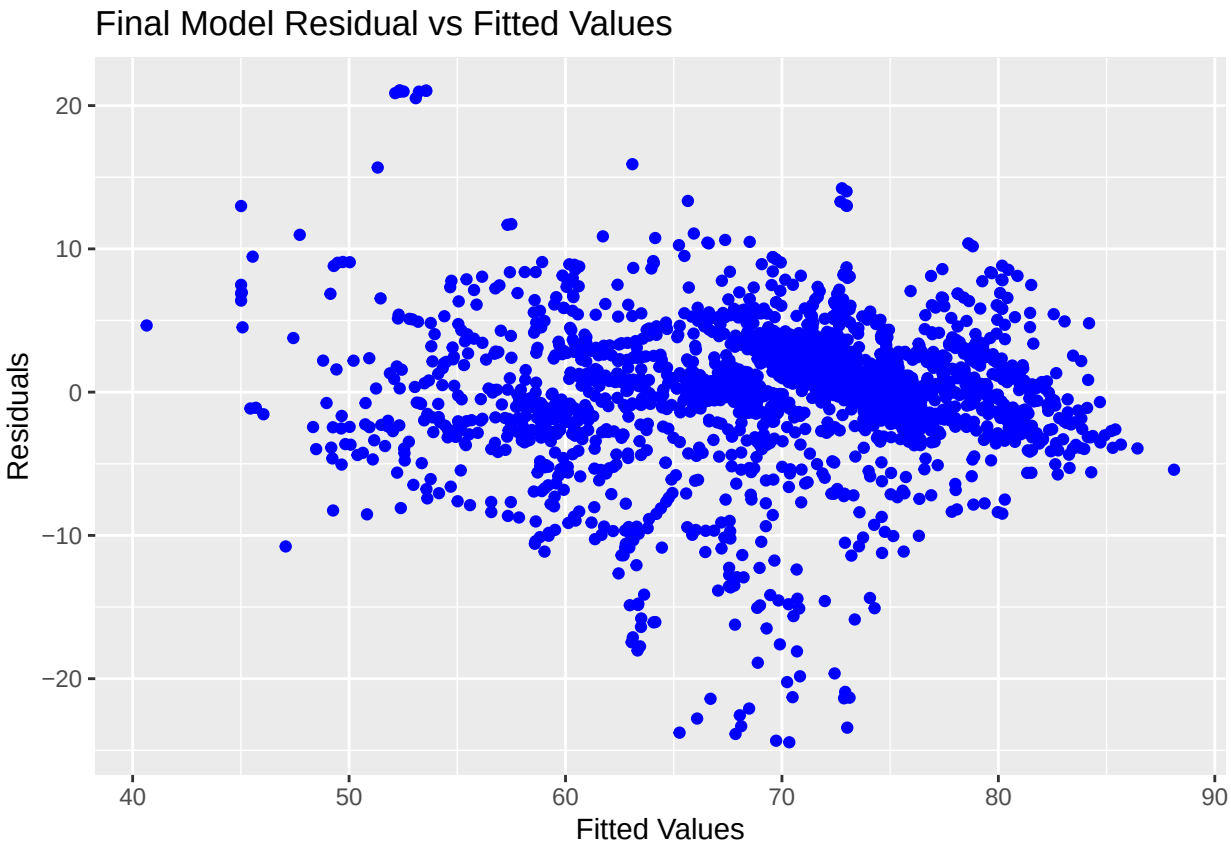
```r
lm_best_fit <-lm(formula = life_exp ~ adult_mortality + income_composition
                 + alcohol+
    schooling + gdp, data = train_data)
summary(lm_best_fit)
```

```
##
## Call:
## lm(formula = life_exp ~ adult_mortality + income_composition +
##     alcohol + schooling + gdp, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4403  -2.0163   0.5219   2.7224  21.0651
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.688e+01  5.878e-01  96.770  < 2e-16 ***
## adult_mortality    -3.350e-02  1.034e-03 -32.406  < 2e-16 ***
## income_composition  9.517e+00  9.236e-01  10.304  < 2e-16 ***
## alcohol             1.348e-01  3.356e-02   4.017 6.10e-05 ***
## schooling           9.019e-01  6.117e-02  14.745  < 2e-16 ***
## gdp                 4.536e-05  9.874e-06   4.594 4.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.11 on 2053 degrees of freedom
## Multiple R-squared:  0.7116, Adjusted R-squared:  0.7109
## F-statistic:  1013 on 5 and 2053 DF,  p-value: < 2.2e-16
```

```r
#Normality
ggplot(data = train_data, aes(sample = rstandard(lm_best_fit))) +
stat_qq() +
stat_qq_line(linewidth = 1, col ="red") +
xlab("Theoretical Quantiles") +
ylab("Standardized Residuals") +
ggtitle("Final Model Normal Q-Q Plot")
```




```
#
ggplot(data = train_data, aes(x = lm_best_fit$fitted.values, y = lm_best_fit$residuals)) +
geom_point(shape = 19, col = "blue") +
xlab("Fitted Values") +
ylab("Residuals") +
ggtitle("Final Model Residual vs Fitted Values")
```


Final Model Residual vs Fitted Values
Residuals
20
10
0
−10
−20
40
50
60
70
80
90
Fitted Values