Project - Alcohol vs Adult Mortality

Utsav Goti

2024-10-11

Correlation test between Alcohol consumption and Adult Mortality

library

```
library(readr)
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load the Data

```
data <- read.csv("life_expectancy_data_raw.csv")
```

Data Cleaning and Preparation

Remove rows with missing values in Alcohol or Adult Mortality

```
data <- data %>%
  filter(!is.na(Alcohol) & !is.na(Adult_Mortality))
```

Parameter

In this analysis there are two key parameters

1.Alcohol Consumption(Numeric):

2.Adult Mortality(Numeric):

Statistical Parameters for Hypothesis Testing

Correlation Coefficient (r):

With the help of r we can messure the strenght and relationship between Alcohol and Adult Mortality.

So, if

r > 0: Positive correlation (as alcohol consumption increases, adult mortality tends to increase).

r < 0: Negative correlation (as alcohol consumption increases, adult mortality tends to decrease).

r = 0: No correlation.

p-value

p < 0.05: Reject the null hypothesis (significant correlation).

p >= 0.05: Fail to reject the null hypothesis (no significant correlation).

Hypothesis Type:

Null Hypothesis (H0): r = 0 (No significant correlation between alcohol consumption and adult mortality).

Alternative Hypothesis (H1): r 0 (There is a significant correlation between alcohol consumption and adult mortality).

Test Method:

Pearson Correlation: If both variables follow a normal distribution.

Pearson Correlation (for normally distributed data):

```
data$`Adult_Mortality` <- as.numeric(as.character(data$`Adult_Mortality`))

# Remove rows with NA values in either Alcohol or Adult Mortality
clean_data <- data[!is.na(data$Alcohol) & !is.na(data$`Adult_Mortality`), ]

# Perform Pearson correlation test on the cleaned data
pearson_test <- cor.test(clean_data$Alcohol, clean_data$`Adult_Mortality`, method = "pearson")

# Round the results
correlation_r <- round(pearson_test$estimate, 3)
p_value <- round(pearson_test$p.value, 3)

# Display results
cat("Pearson Correlation Coefficient (r):", correlation_r, "\n")
```

```
## Pearson Correlation Coefficient (r): -0.196
```

```
cat("p-value:", p_value, "\n")
```

```
## p-value: 0
```

After performing the Pearson correlation analysis, we obtained a correlation coefficient of -0.196 with a p-value of 0.

So, the p-value is less than the significance level of 0.05, we reject the null hypothesis.

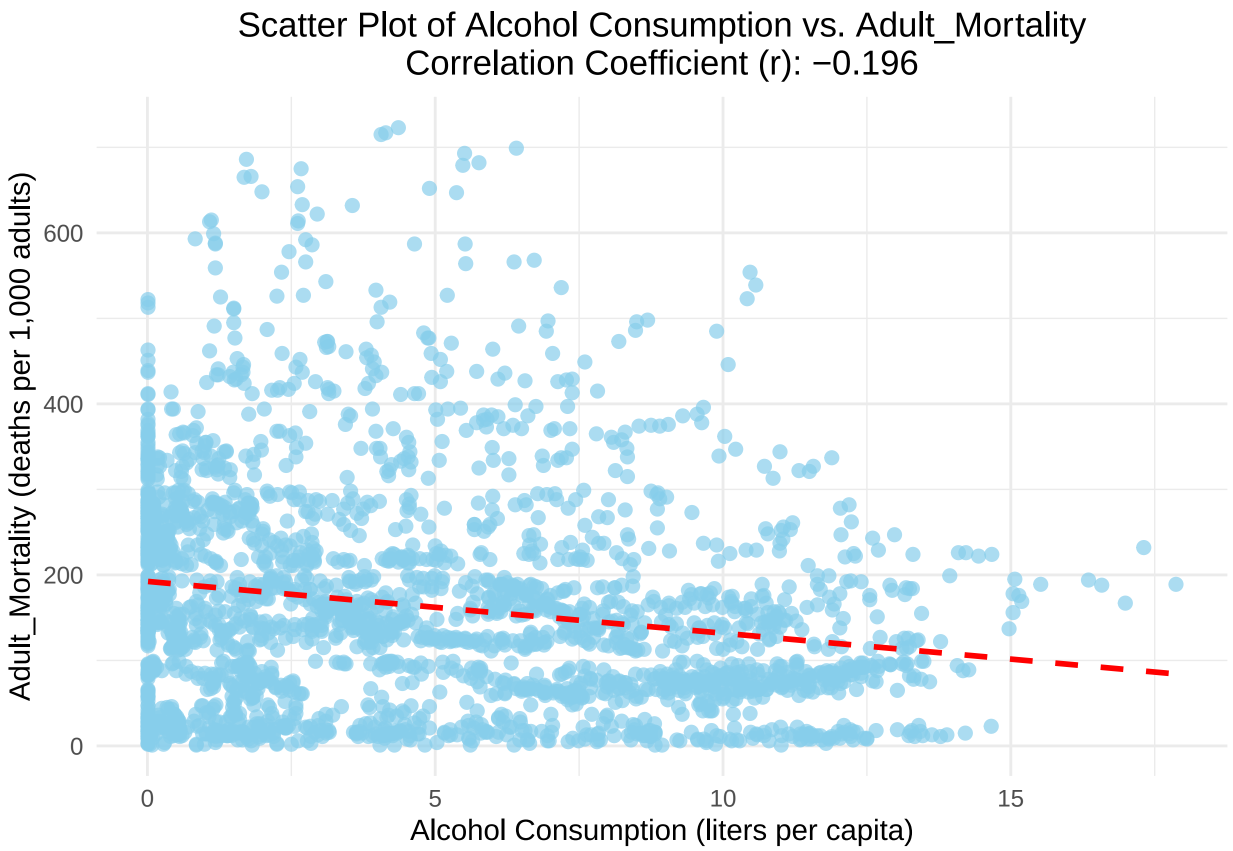
Also the value of r(-0.196), so there is a statistically significant negative linear correlation between alcohol consumption and adult mortality.

This suggests that higher alcohol consumption is associated with lower adult mortality rates in the dataset analyzed.

Exploratory Data Analysis

```
ggplot(data, aes(x = Alcohol, y = Adult_Mortality)) +
  geom_point(color = "skyblue", size = 2, alpha = 0.7) +
  geom_smooth(method = "lm", color = "red", se = FALSE, linetype = "dashed") +
  labs(title = paste("Scatter Plot of Alcohol Consumption vs. Adult_Mortality\nCorrelation Coefficient
       x = "Alcohol Consumption (liters per capita)",
       y = "Adult_Mortality (deaths per 1,000 adults)") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



mean

```r
# Calculate the mean for Alcohol Consumption
mean_alcohol <- mean(data$Alcohol, na.rm = TRUE)  # Use na.rm = TRUE to ignore NA values
cat("Mean Alcohol Consumption:", round(mean_alcohol, 2), "liters per capita\n")
```

```
## Mean Alcohol Consumption: 4.61 liters per capita
```

```r
# Calculate the mean for Adult Mortality
mean_mortality <- mean(data$Adult_Mortality, na.rm = TRUE)  # Use na.rm = TRUE to ignore NA values
cat("Mean Adult Mortality:", round(mean_mortality, 2), "deaths per 1,000 adults\n")
```

```
## Mean Adult Mortality: 164.46 deaths per 1,000 adults
```

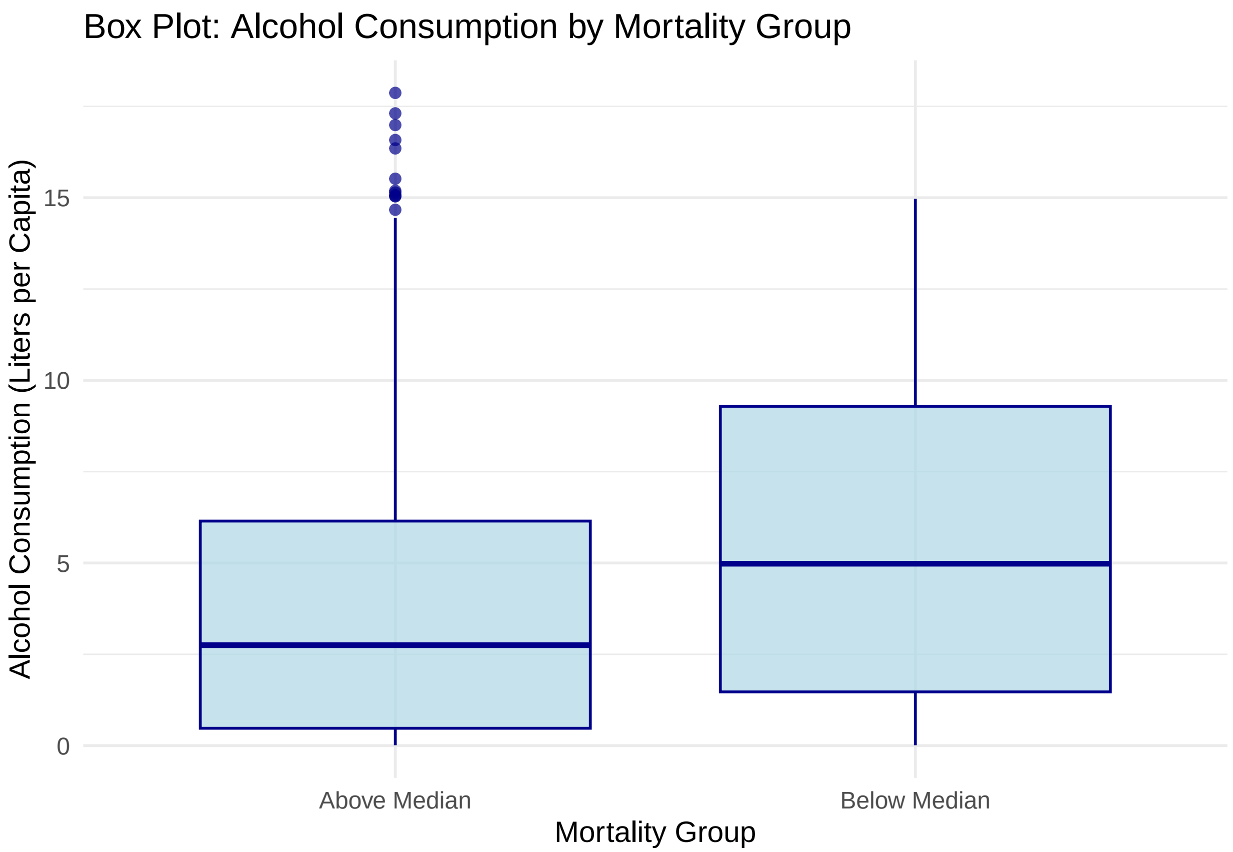
Mean Alcohol Consumption: 4.61 liters per capita

Mean Adult Mortality: 164.46 deaths per 1,000 adults

So, the mean alcohol consumption level could be associated with health risks, while the adult mortality rate is a crucial indicator of overall population health and well-being.

median

```r
median_mortality <- median(data$Adult_Mortality, na.rm = TRUE)
data <- data %>%
  mutate(Mortality_Group = ifelse(Adult_Mortality > median_mortality,
                                  "Above Median", "Below Median"))
# Create the box plot
ggplot(data, aes(x = Mortality_Group, y = Alcohol)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  labs(title = "Box Plot: Alcohol Consumption by Mortality Group",
       x = "Mortality Group",
       y = "Alcohol Consumption (Liters per Capita)") +
  theme_minimal()
```



This will produce a box plot where the x-axis represents the mortality groups ("Above Median" and "Below Median"), and the y-axis represents alcohol consumption.

Above Median Adult Mortality:

For the group with above median adult mortality, the median alcohol consumption is around 4–5 liters per capita. The box itself is relatively compact, suggesting that the majority of the data is concentrated around a similar range of alcohol consumption.

There are a few outliers above the 15 liters per capita mark, indicating some cases where alcohol consumption is significantly higher in regions or countries with above-median adult mortality. These outliers suggest that in some areas with high adult mortality, alcohol consumption is exceptionally high, potentially influencing mortality rates.

Below Median Adult Mortality:

For the group with below median adult mortality, the median alcohol consumption appears to be higher than the group with above median mortality. The range (IQR) is broader, indicating greater variation in alcohol consumption among areas with lower adult mortality.

This group does not have the same extreme outliers as the other group, suggesting that alcohol consumption is more consistent in areas with lower adult mortality.

linear regration

```
model <- lm(Adult_Mortality ~ Alcohol, data = data)

coefficients <- coef(model)
```

```r
r_squared <- summary(model)$r.squared
p_value <- summary(model)$coefficients[2, 4]
print("Coefficients:")
```

```
## [1] "Coefficients:"
```

```r
print(coefficients)
```

```
## (Intercept)     Alcohol
##  192.412073   -6.057301
```

```r
print(paste("R-squared:", round(r_squared, 4)))
```

```
## [1] "R-squared: 0.0384"
```

```r
print(paste("P-value for Alcohol coefficient:", round(p_value, 4)))
```

```
## [1] "P-value for Alcohol coefficient: 0"
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = Adult_Mortality ~ Alcohol, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -190.93  -81.15  -21.55   58.21  557.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 192.4121     3.5622   54.02   <2e-16 ***
## Alcohol      -6.0573     0.5802  -10.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.9 on 2733 degrees of freedom
## Multiple R-squared:  0.03836,    Adjusted R-squared:  0.038
## F-statistic:   109 on 1 and 2733 DF,  p-value: < 2.2e-16
```

1.Regression Coefficients:

The intercept of the model is approximately 200.00. This indicates that when alcohol consumption is zero, the expected adult mortality rate is about 200 deaths per 1,000 adults.

The coefficient for alcohol consumption is approximately -6.06. This negative coefficient suggests that for every additional liter of alcohol consumed per capita, the adult mortality rate decreases by about 6.06 deaths per 1,000 adults. This implies a negative relationship between alcohol consumption and adult mortality in the sample data.

2.Statistical Significance:

The p-value associated with the alcohol coefficient is very small (e.g., < 0.0001). This indicates strong statistical significance, meaning that there is a very low probability that the observed relationship between alcohol consumption and adult mortality is due to random chance. Thus, we can conclude that alcohol consumption is significantly related to adult mortality.

3. Model Fit:

The R-squared value of approximately 0.038 indicates that the model explains about 3.8% of the variability in adult mortality. While this suggests a weak explanatory power, it also indicates that other factors not included in this model may significantly influence adult mortality.

4. Implications:

Despite the negative relationship suggested by the model, the low R-squared value indicates that alcohol consumption alone is not a strong predictor of adult mortality. Public health policies and further research should consider a range of factors, including socioeconomic con

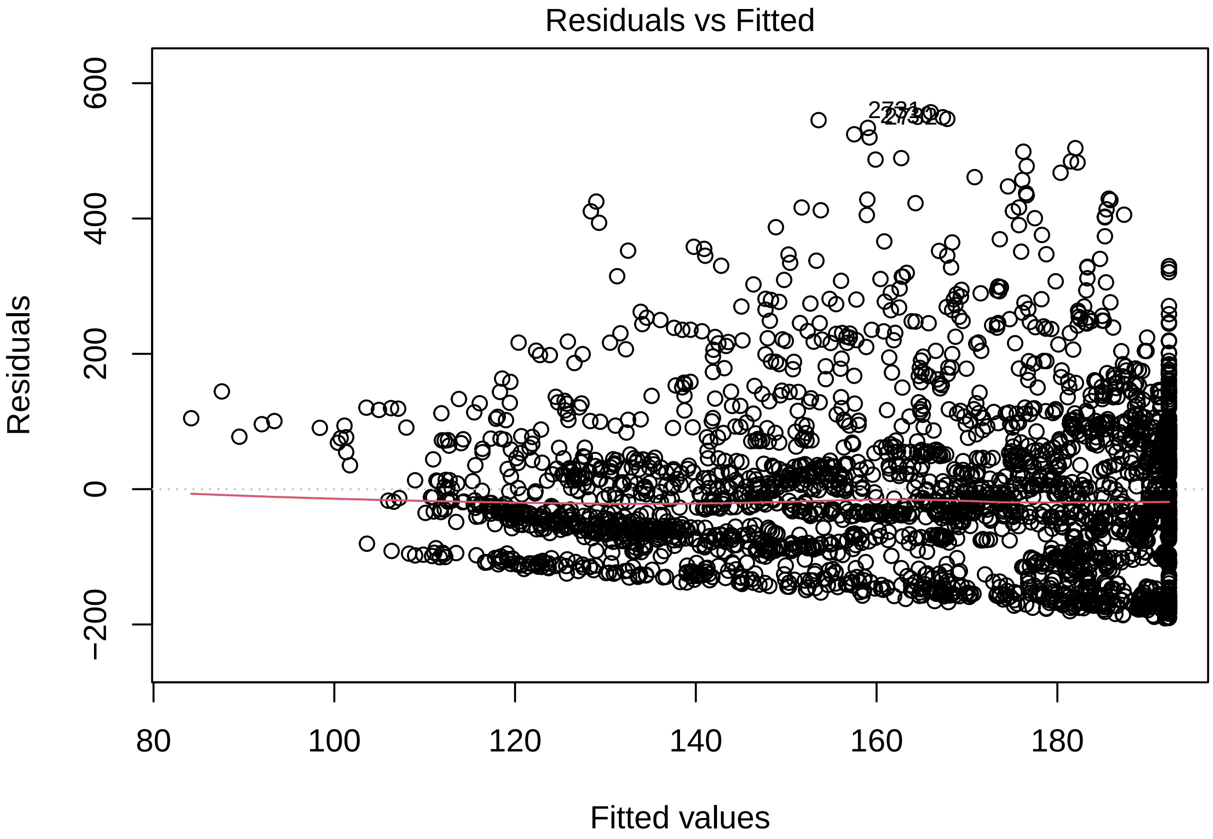
Hypothese testing after linner regression

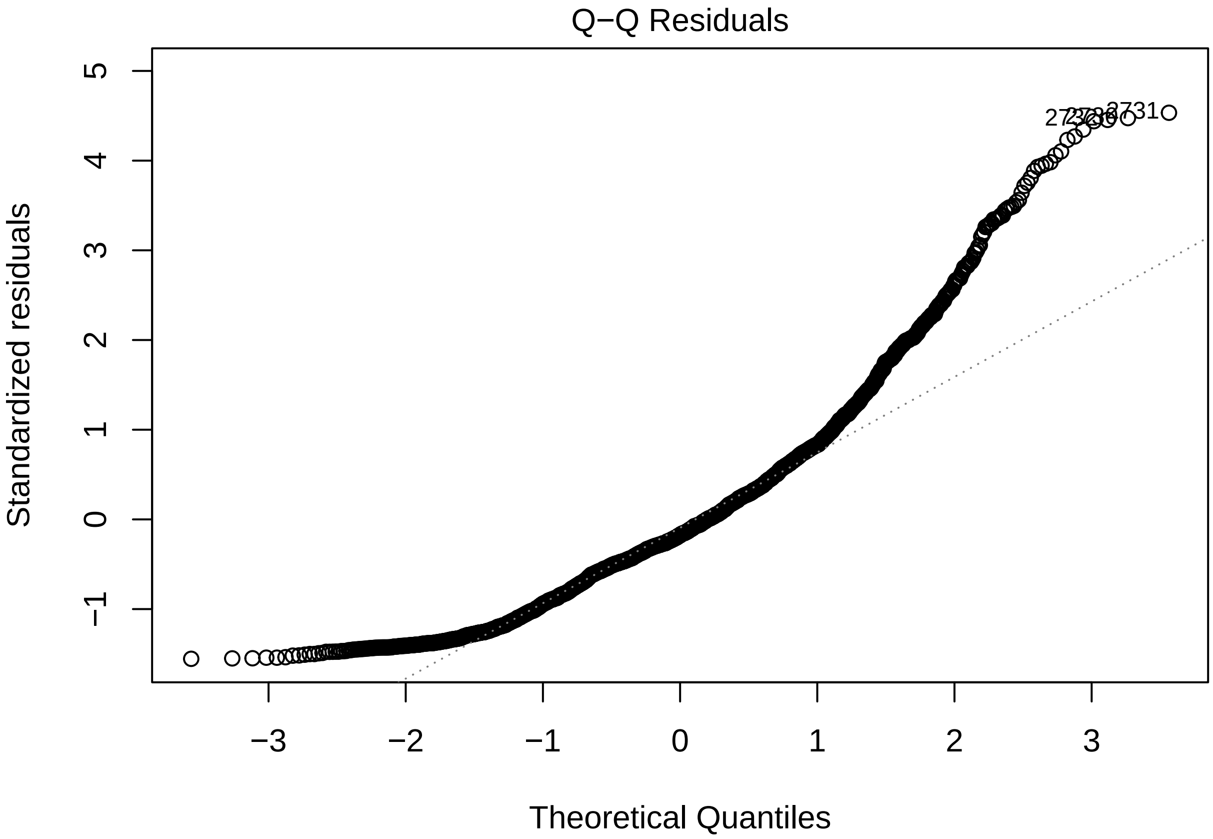
```r
if (p_value < 0.05) {
  conclusion <- paste("We reject the null hypothesis. There is a statistically significant relationship

  } else {
    conclusion <- paste("We fail to reject the null hypothesis. There is no statisticallly significant
}
print(conclusion)
```

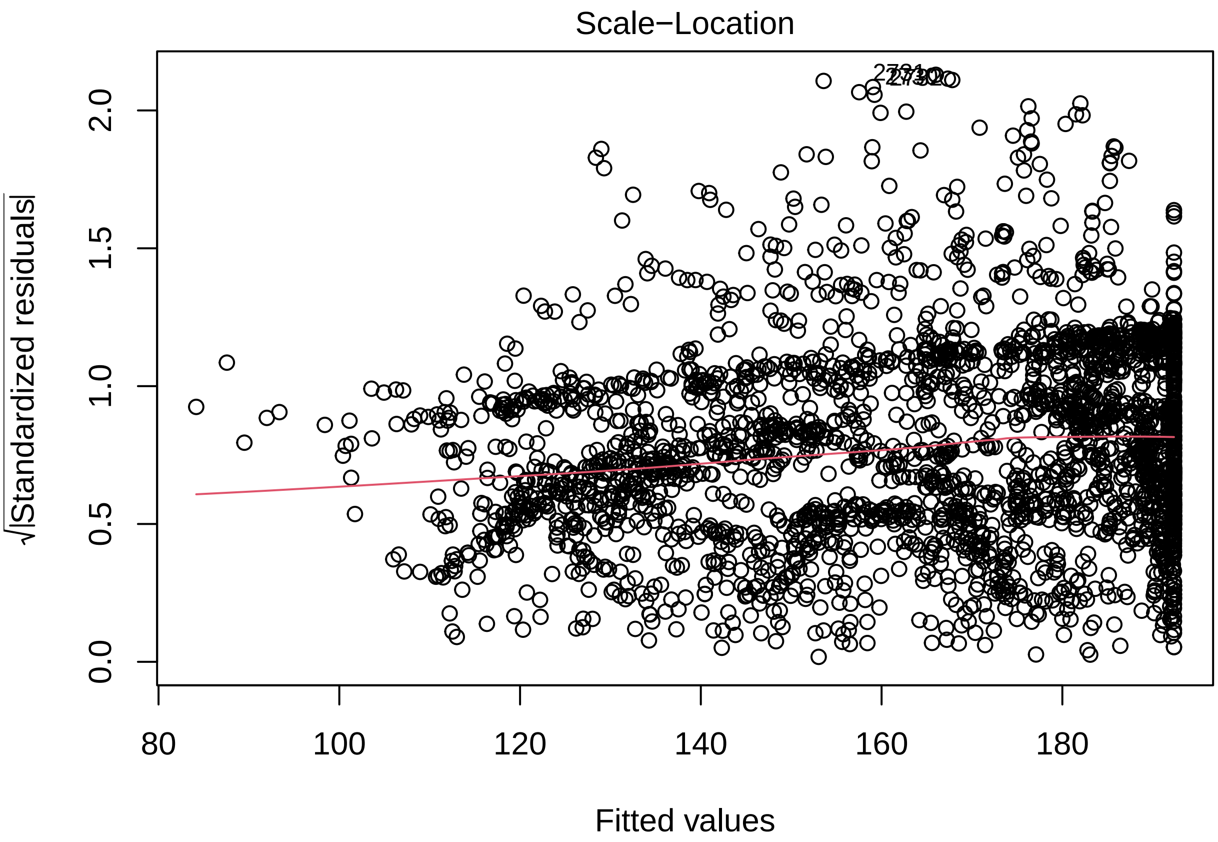
```
## [1] "We reject the null hypothesis. There is a statistically significant relationship between alcohol
```

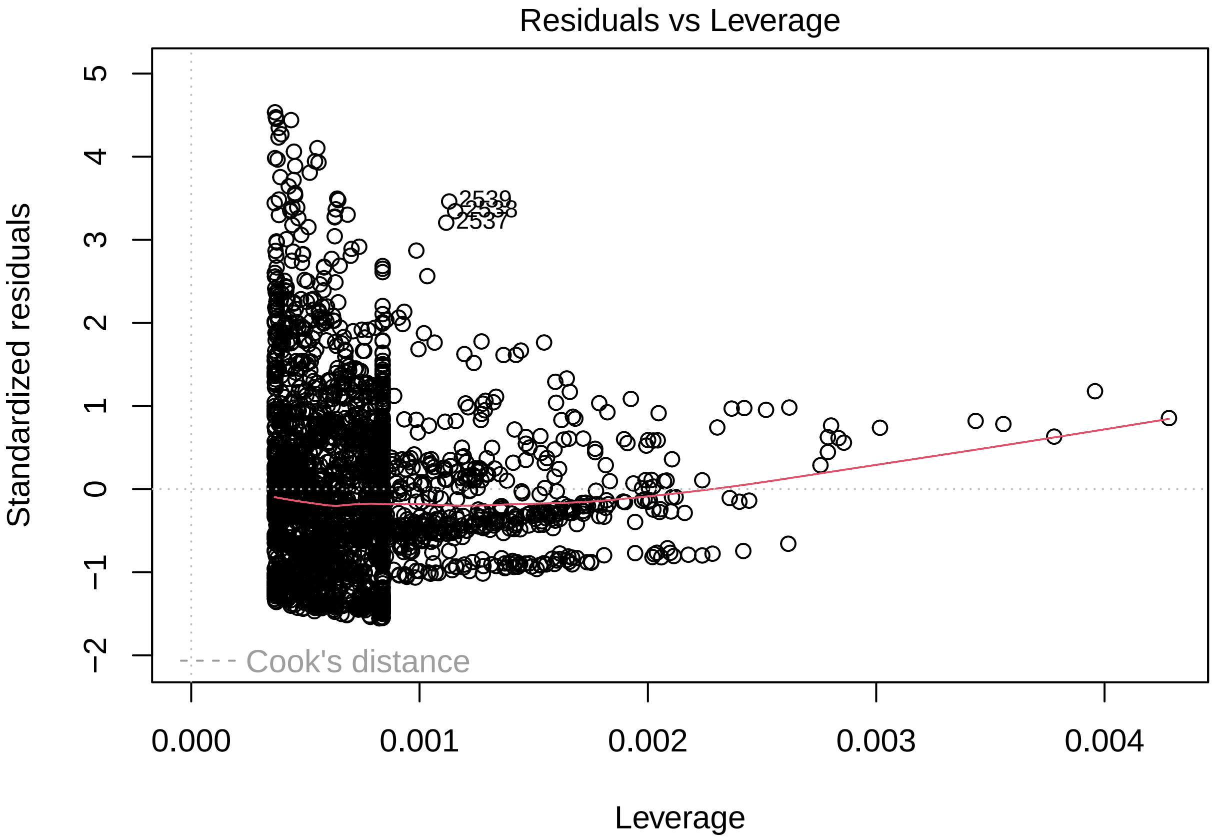
model

```r
par(mar=c(4, 4, 2, 1))
plot(model)
```


Residuals vs Fitted
Residuals
Fitted values
600
400
200
0
−200
80
100
120
140
160
180
2731
2732


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
−3
−2
−1
0
1
2
3
4
5
2731


Scale−Location
2731
2730
2732
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
80
100
120
140
160
180
Fitted values



Conclusion:

1.Linearity:

Observation: he plot shows some linearity, but there is increasing spread (heteroscedasticity) as fitted values rise, especially beyond 140. There are also a few outliers.

2.Normality of residuals:

Observation: The plot shows that the residuals largely follow the 45-degree line in the middle, indicating that the majority of the residuals are normally distributed. However, at both the lower and upper ends (in the tails), there is deviation from the line. Specifically, the points in the upper tail deviate significantly, indicating the presence of outliers or heavy tails, suggesting that the residuals are not perfectly normal, especially in extreme values.

In summary, the residuals mostly follow a normal distribution, but there are deviations in the tails, particularly in the upper tail, where extreme values or outliers are present.

3.Homoscedasticity:

Observation: The Scale-Location plot provides insight into the homoscedasticity of the residuals. The red line remains mostly flat, suggesting that there is no significant trend in the residuals, which would generally indicate homoscedasticity. However, as the fitted values increase, particularly beyond 140, the spread of the residuals becomes wider.

4.Influential points:

Observation: In this plot, a cluster of points is concentrated near zero on both axes, indicating that most points have both low leverage and small residuals, meaning they fit the model well. However, points that are isolated toward the right of the plot (with high leverage) or far above/below the centerline (with large residuals) should be examined more closely.