Project - Alcohol vs Life expectancy

Utsav Goti

2024-10-12

Correlation test between Alcohol consumption and Life expectancy.

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 Life expectancy.

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

Parameter

In this analysis there are two key parameters:

1.Alcohol Consumption (Numeric):

2.Life Expectancy:

Statistical Parameters for Hypothesis Testing

Correlation Coefficient (r):

With the help of r, we can measure the strength and relationship between Alcohol Consumption and Life Expectancy.

So, if:

r > 0: Positive correlation (as alcohol consumption increase, life expectancy tends to increase).

r < 0: Negative correlation (as alcohol consumption increases, life expectancy 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 life expectancy).

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

Test Method:

Pearson Correlation: If both variables follow a normal distribution.

Pearson Correlation (for normally distributed data):

```
pearson_test <- cor.test(data$Alcohol, data$Life.expectancy, method = "pearson")

correlation_r <- round(pearson_test$estimate, 3)
p_value <- round(pearson_test$p.value, 3)

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

```
## Pearson Correlation Coefficient (r): 0.405
```

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

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

After performing the Pearson correlation analysis, we obtained a correlation coefficient of 0.405 with a p-value of 0.

An r value of 0.405 indicates a moderate positive correlation between Alcohol Consumption and Life Expectancy. This suggests that, as alcohol consumption increases, life expectancy tends to also increase, although this relationship is not very strong.

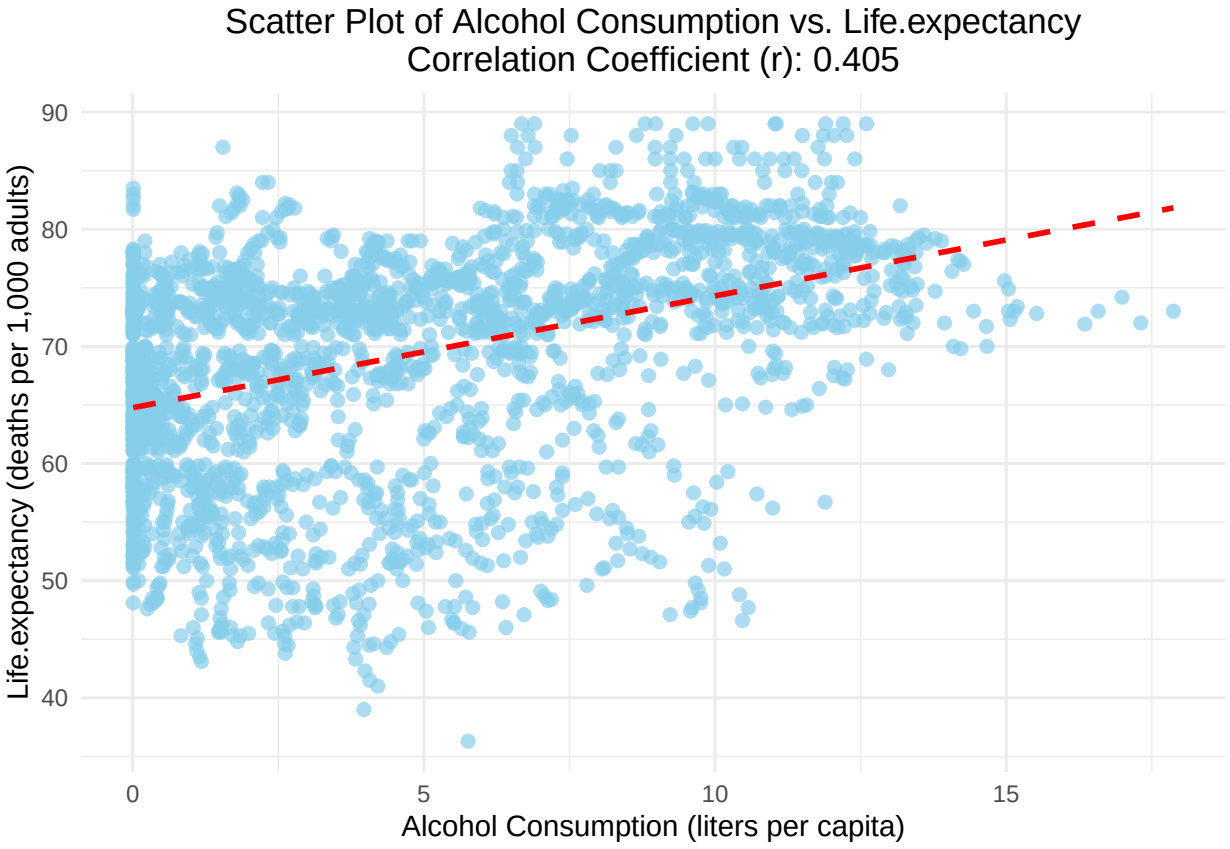
Since the p-value is less than the significance level of 0.05, we reject the null hypothesis.

This suggests that higher alcohol consumption is associated with higher life expectancy in the dataset analyzed.

Exploratory Data Analysis

```
ggplot(data, aes(x = Alcohol, y = Life.expectancy)) +
  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. Life.expectancy\nCorrelation Coefficient
       x = "Alcohol Consumption (liters per capita)",
       y = "Life.expectancy (deaths per 1,000 adults)") +
  theme_minimal()+
  theme(plot.title = element_text(hjust = 0.5))
```

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



Mean

```r
#mean of Alcohol Consumption
mean_alcohol <- mean(data$Alcohol, na.rm = TRUE)
cat("Mean Alcohol Consumption:", round(mean_alcohol, 2), "liters per capita\n")
```

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

```r
#mean of Life Expectancy
mean_life_expectancy <- mean(data$Life.expectancy, na.rm = TRUE)
cat("Mean Life Expectancy:", round(mean_life_expectancy, 2), "years\n")
```

```
## Mean Life Expectancy: 69.17 years
```

Mean Alcohol Consumption: 4.61 liters per capita

Mean Life Expectancy: 69.17 years

The average life expectancy is 69.17 years, suggesting that people in this population tend to live for a long time and are generally healthy

Median

```r
median_life_expectancy <- median(data$Life.expectancy, na.rm = TRUE)

data <- data %>%
  mutate(Life_Expectancy_Group = ifelse(Life.expectancy> median_life_expectancy,
                                        "Above Median", "Below Median"))

ggplot(data, aes(x = Life_Expectancy_Group, y = Alcohol)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", alpha = 0.7) +
  labs(title = "Box Plot: Alcohol Consumption by Life Expectancy Group",
       x = "Life Expectancy Group",
       y = "Alcohol Consumption (Liters per Capita)") +
  theme_minimal()
```

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

Above Median Life Expectancy:

For the group with above median life expectancy, the median alcohol consumption is around 6 liters per capita. The box is more extended, suggesting that there is a wider spread of alcohol consumption within this group. Despite the variability, most of the values fall within a fairly moderate range of alcohol consumption, implying that in regions with higher life expectancy, alcohol consumption is generally moderate.

There are a few extreme outliers, particularly those consuming more than 15 liters per capita. These indicate that even in areas where life expectancy is above the median, some regions or countries have very high levels of alcohol consumption. However, the majority of these regions maintain consumption within a moderate range.

Below Median Life Expectancy:

In contrast, for the group with below median life expectancy, the median alcohol consumption is noticeably lower than that of the group with above median life expectancy. The box plot reveals that this group has more variation and a greater number of outliers, with values extending beyond 10 liters per capita, showing a wider range of alcohol consumption in regions with lower life expectancy.

The presence of more outliers suggests that in areas with lower life expectancy, alcohol consumption levels vary significantly, with some regions exhibiting relatively high consumption, which may be associated with the lower life expectancy. This variation could reflect social, economic, or health factors influencing both alcohol consumption and life expectancy.

linear regration

```
model <- lm(Life.expectancy ~ Alcohol, data = data)

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

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

```
print(coefficients)
```

```
## (Intercept)     Alcohol
##  64.7633398   0.9546442
```

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

```
## [1] "R-squared: 0.1639"
```

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

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

```
summary(model)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Alcohol, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.962  -4.722   1.622   6.408  20.757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.76334    0.25321  255.76   <2e-16 ***
## Alcohol      0.95464    0.04124   23.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.735 on 2733 degrees of freedom
## Multiple R-squared:  0.1639, Adjusted R-squared:  0.1636
## F-statistic: 535.8 on 1 and 2733 DF,  p-value: < 2.2e-16
```

1.Regression Coefficients:

The intercept of the model is approximately 64.76. This indicates that when alcohol consumption is zero, the expected life expectancy is about 64.76 years.

The coefficient for alcohol consumption is approximately 0.95. This positive coefficient suggests that for every additional liter of alcohol consumed per capita, life expectancy increases by about 0.95 years. This implies a positive relationship between alcohol consumption and life expectancy 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 there is a very low probability that the observed relationship between alcohol consumption and life expectancy is due to random chance. Thus, we can conclude that alcohol consumption is significantly related to life expectancy.

3.Model Fit:

The R-squared value of approximately 0.1639 indicates that the model explains about 16.39% of the variability in life expectancy. While this suggests a moderate explanatory power, it also indicates that other factors not included in this model may significantly influence life expectancy.

4.Implications:

Despite the positive relationship suggested by the model, the R-squared value indicates that alcohol consumption alone is not a strong predictor of life expectancy. Other variables may also play a crucial role in determining life expectancy outcomes in the population studied.

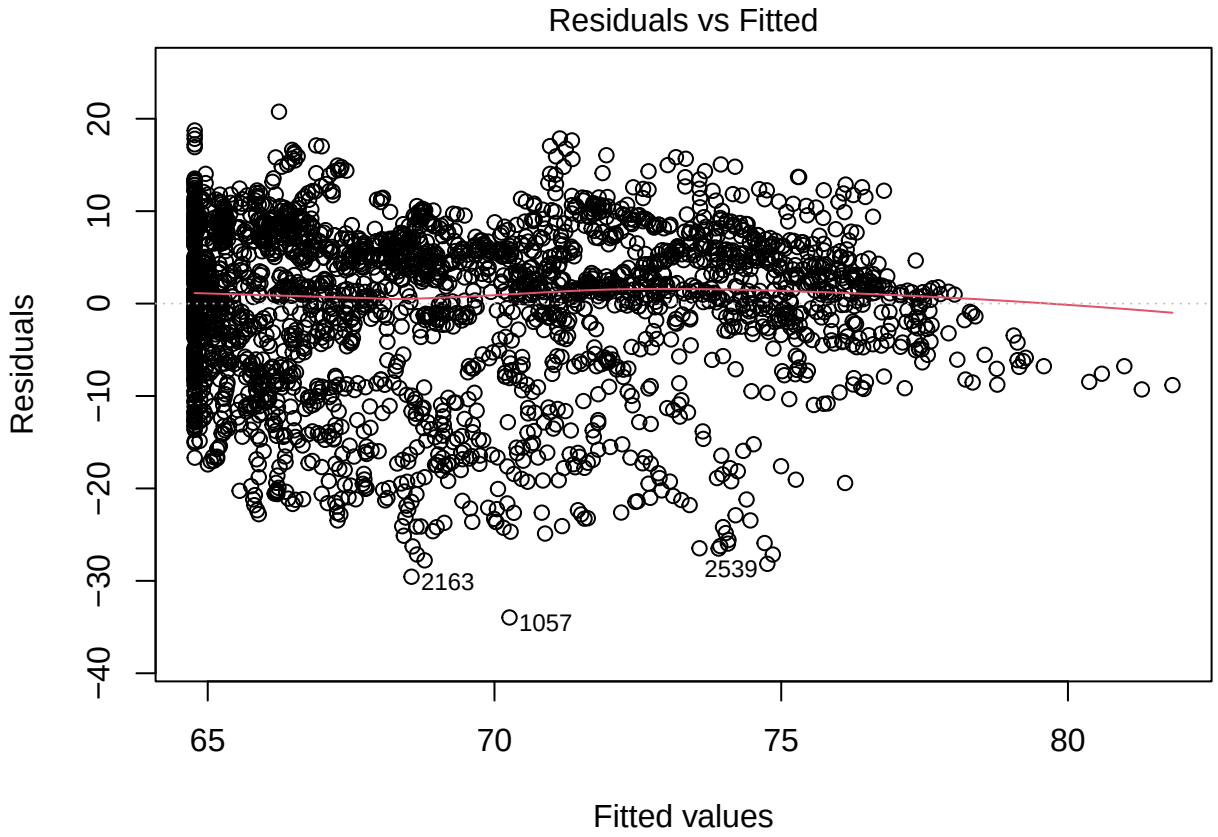
Hypothese testing after linner regression

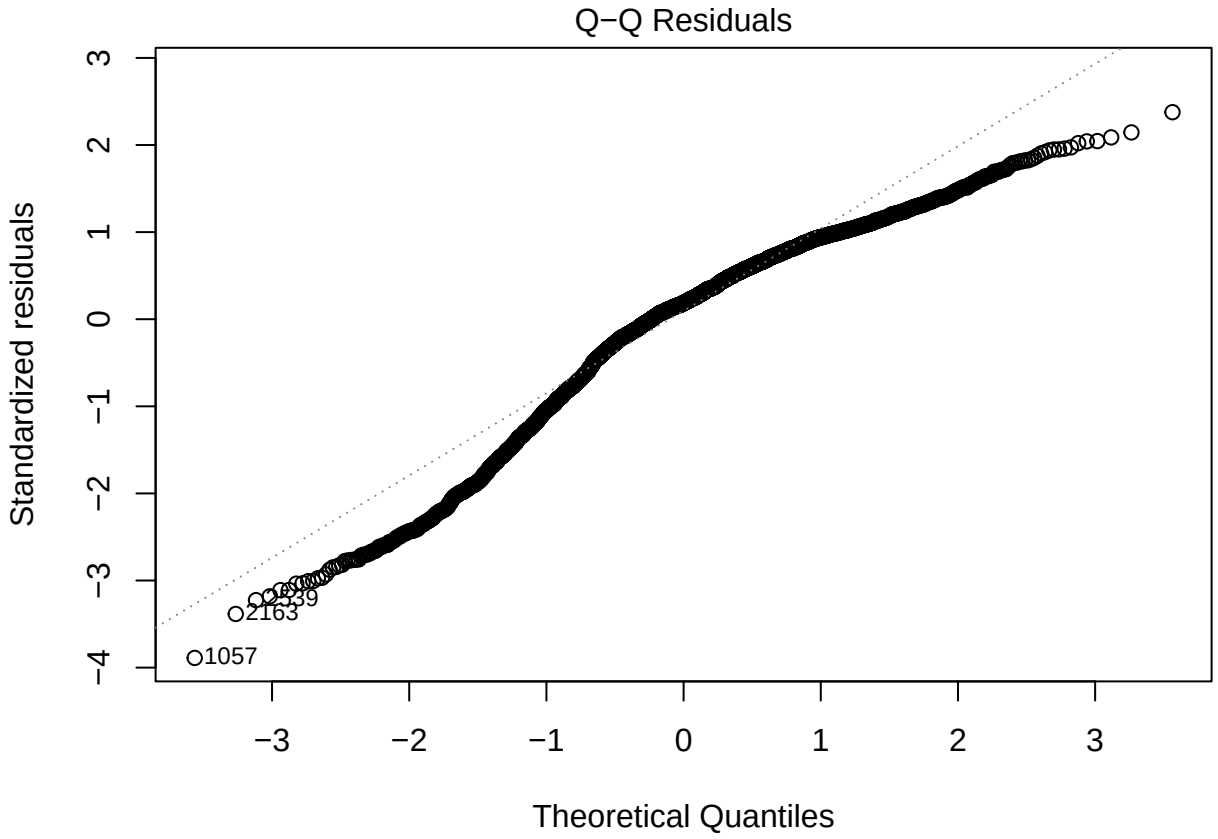
```
if (p_value < 0.05) {
  conclusion <- paste("We reject the null hypothesis. There is a statistically significant relationship

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

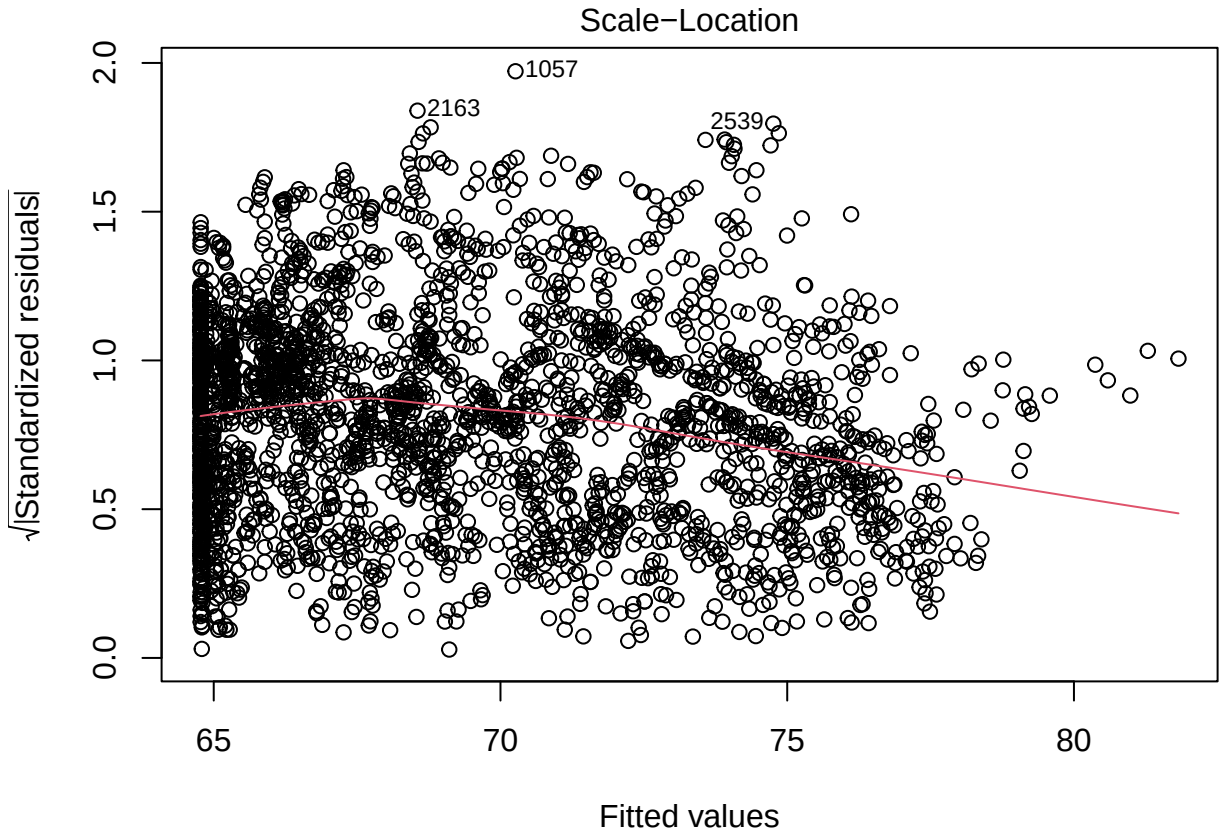
```
## [1] "We reject the null hypothesis. There is a statistically significant relationship between alcohol
```

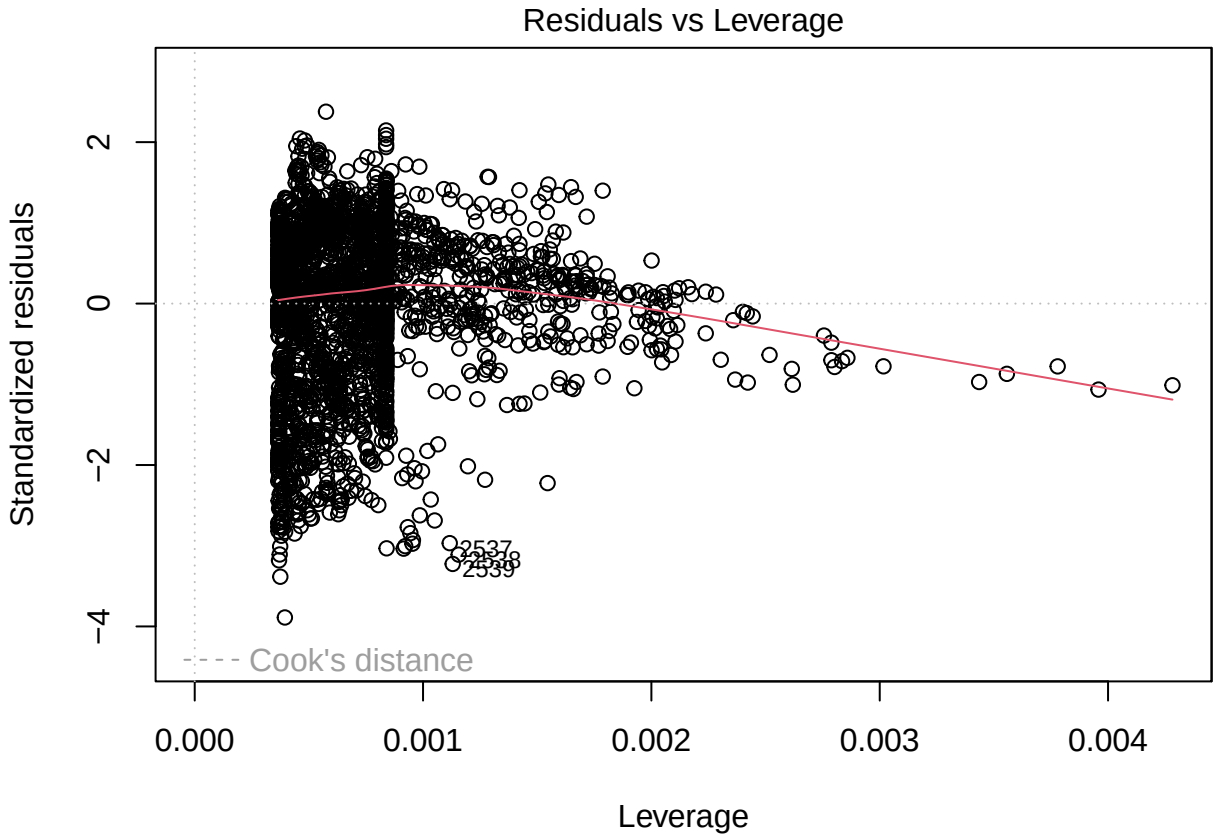
Model

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


Residuals vs Fitted
Residuals
Fitted values
20
10
0
−10
−20
−30
−40
65
70
75
80
2163
1057
2539


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
−4
−3
−2
−1
0
1
2
3
2163
1057


Scale−Location
√|Standardized residuals|
Fitted values
1057
2163
2539
0.0
0.5
1.0
1.5
2.0
65
70
75
80



Conclusion:

1.Linearity:

Observation: The plot are distributed symmetrically around the zero line for a large portion of the fitted values. This indicates that the relationship between the predictors and the outcome variable may be adequately captured by the linear model in certain ranges.

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 center line (with large residuals) should be examined more closely.