STAT630 Homework 7

David Teng

Due Friday, November 22nd

Part 1: Data Cleaning

1.

```r
library(Hmisc)
library(readr)
library(dplyr)
library(ggplot2)

support <- read.table("data/support.txt", header = TRUE)
```

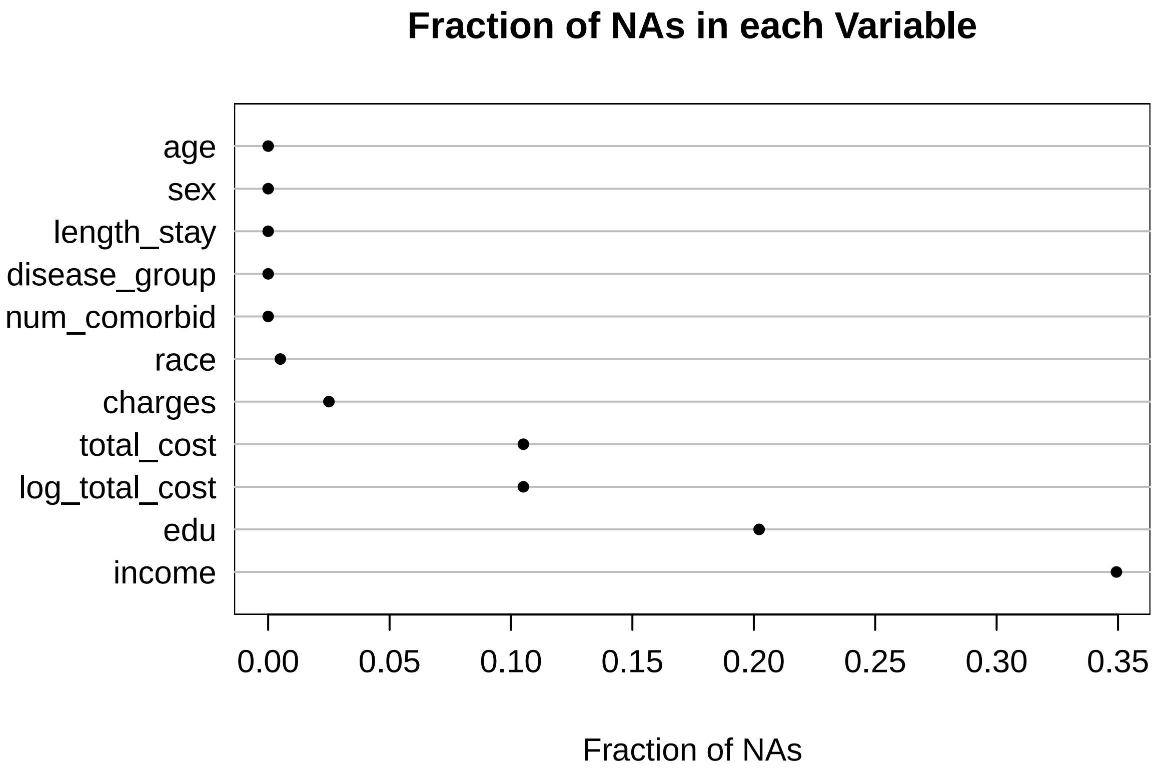
2.

```r
# a. Rename columns
support <- support %>%
  rename(
    length_stay = slos,
    disease_group = dzgroup,
    num_comorbid = num.co,
    total_cost = totcst
  )

# b.  Convert categorical variables to factors
support <-support%>%
  mutate(across(where(is.character),~ as.factor(capitalize(.))))

# c.  Add log_total_cost
support <-support %>%
  mutate(log_total_cost = log(total_cost))
```

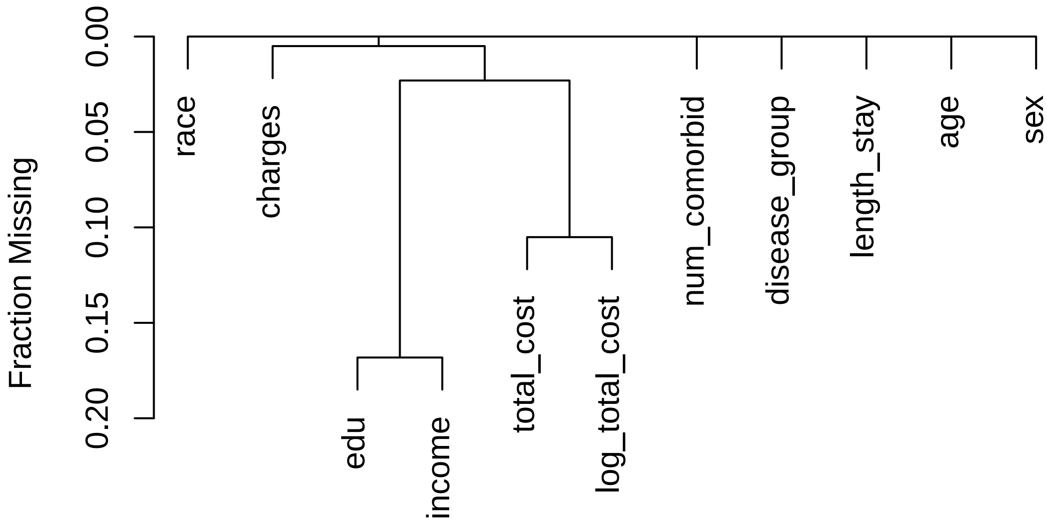
Part 2: Exploratory Data Analysis

3.

```r
na_patterns <- naclus(support)
naplot(na_patterns, "na per var")
```



```r
plot(na_patterns)
```

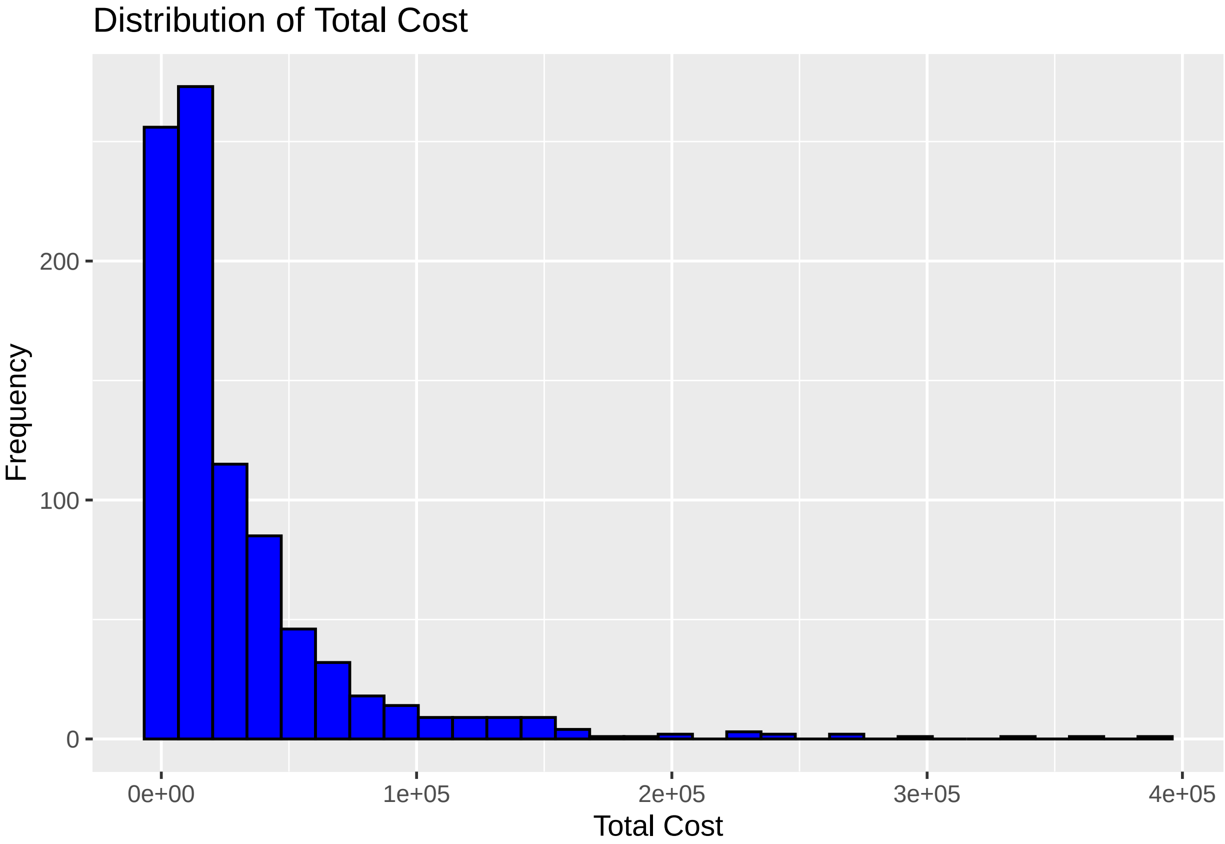


4.

The missing values for "income", "charges", and "total_cost" seem connected, indicating financial details weren't recorded for some patients. Education data ("edu") is also missing a lot, maybe because some patients didn't respond or data wasn't collected consistently. This could be due to patients not wanting to share private financial details like "total_cost" during data collection in medical care.

5.

```r
# Plot histogram
ggplot(support, aes(x = total_cost)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Distribution of Total Cost", x = "Total Cost", y = "Frequency")
```



6.

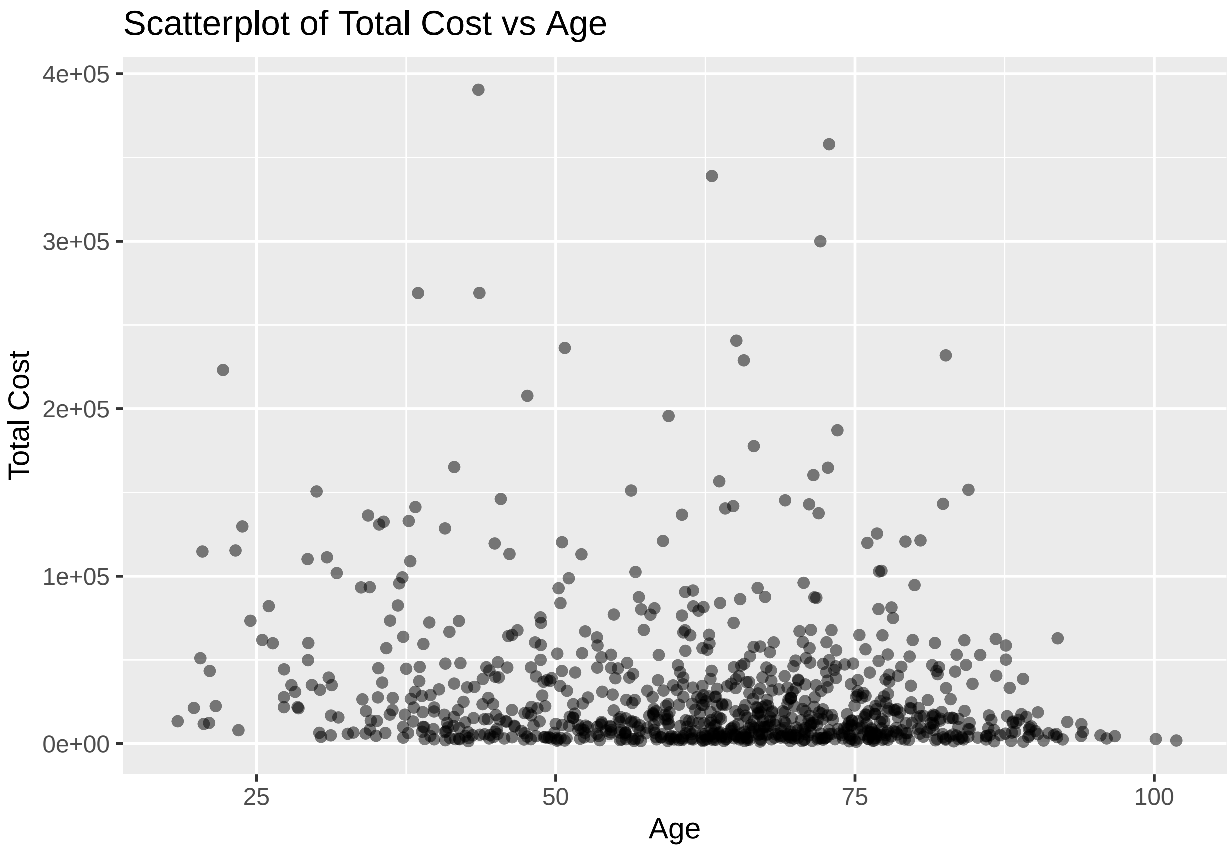
```
summary(support$total_cost)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1162    5908   15120   30524   37636  390460     105 
```

The blue histogram in Plot 1 shows a right-skewed distribution of total cost, with most values concentrated near zero. The center (median) is relatively low, and the spread is wide due to a few extremely high-cost outliers extending the range.

7.

```
# Scatterplot
ggplot(support, aes(x = age, y = total_cost)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scatterplot of Total Cost vs Age", x = "Age", y = "Total Cost")
```

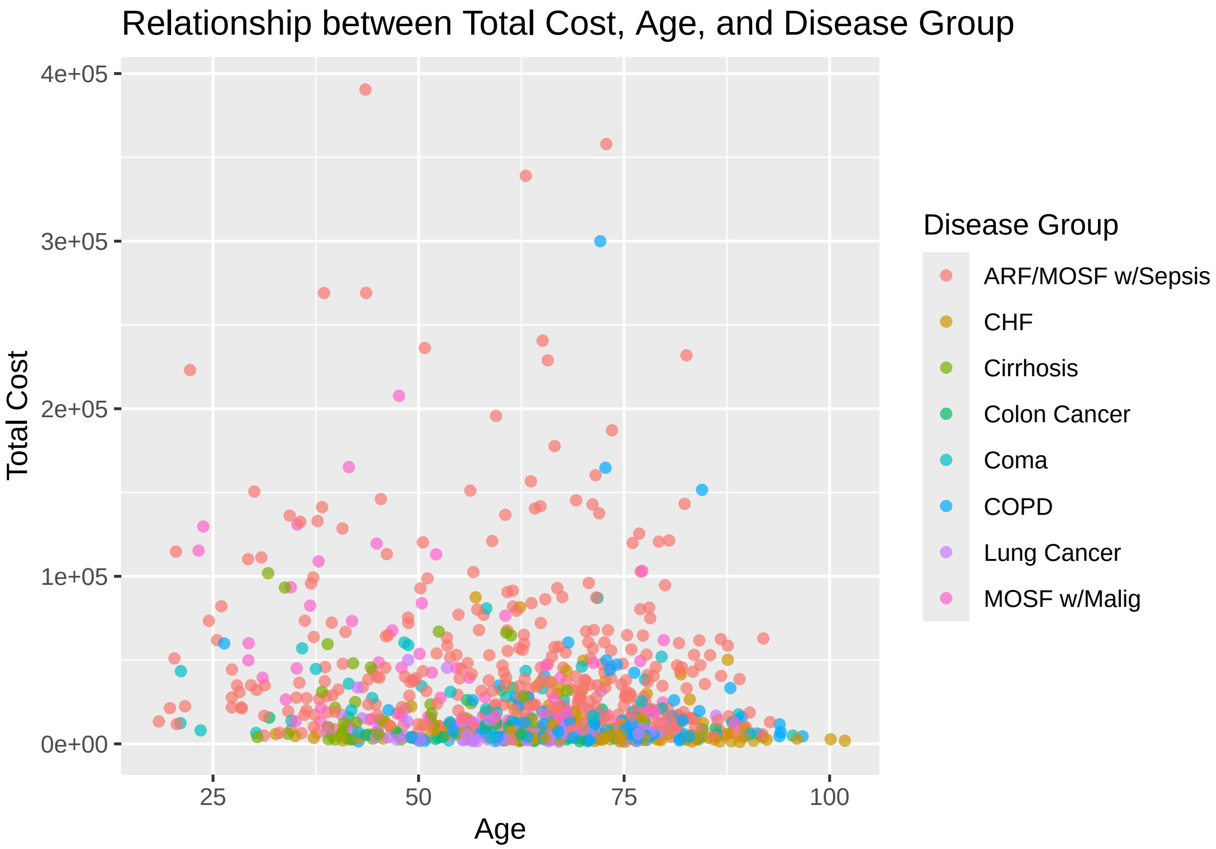


8.

A potential confounding variable is "disease_group". It is related to age because certain diseases are more prevalent in specific age groups. Additionally, it influences total cost as different diseases may require different levels of medical cares, significantly impacting healthcare expenditures.

Bonus:

```
ggplot(data = support, aes(x = age, y = total_cost, color = disease_group)) +
  geom_point(alpha = 0.7) +
  labs(
    title = "Relationship between Total Cost, Age, and Disease Group",
    x = "Age",
    y = "Total Cost",
    color = "Disease Group"
  )
```



Part 3: Data Analysis

9.

Step 1: Hypotheses

$H_0 : \beta_1 = 0$ (The slope of the regression line between total cost and age is zero. There is no linear relationship between total cost and age.)

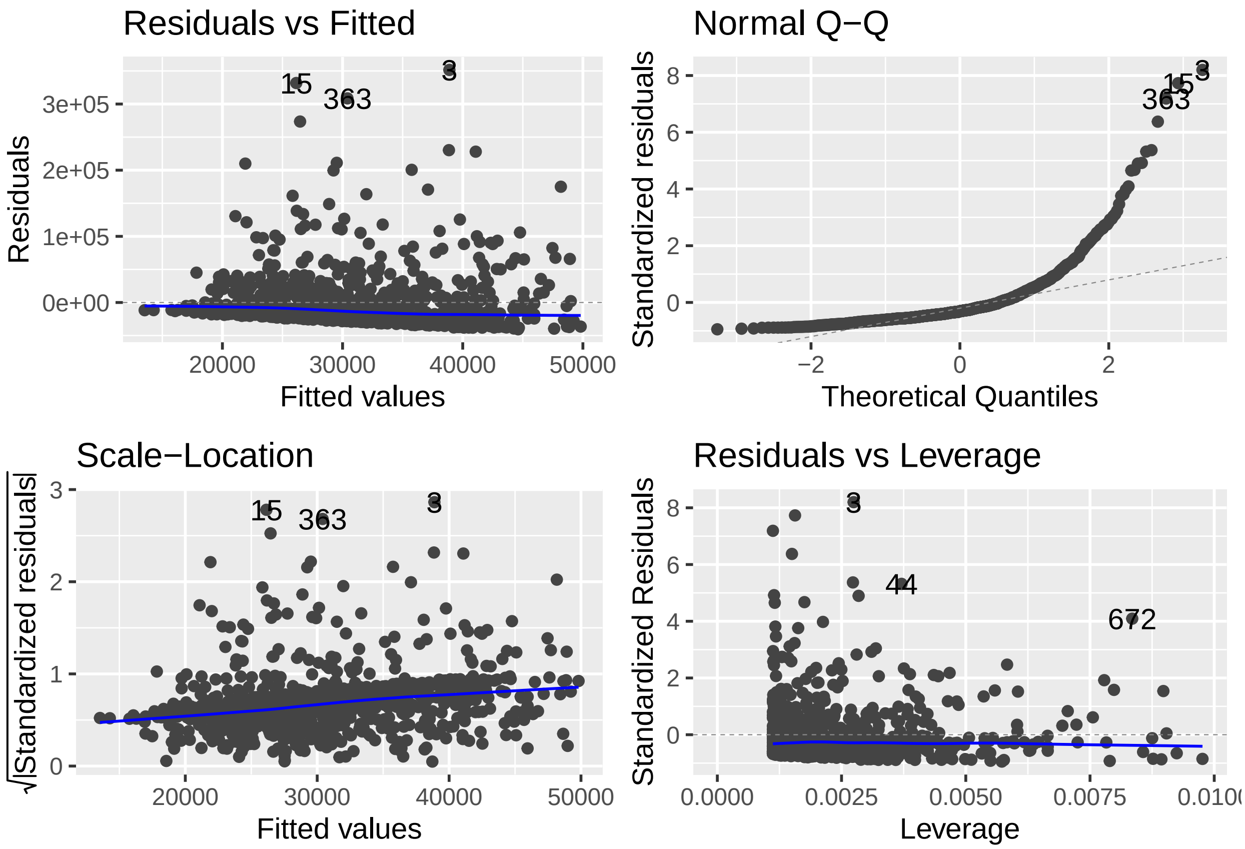
$H_A : \beta_1 \neq 0$ (The slope of the regression line between total cost and age is not zero. There is a linear relationship between total cost and age.)

10.

Step 2: Assumptions

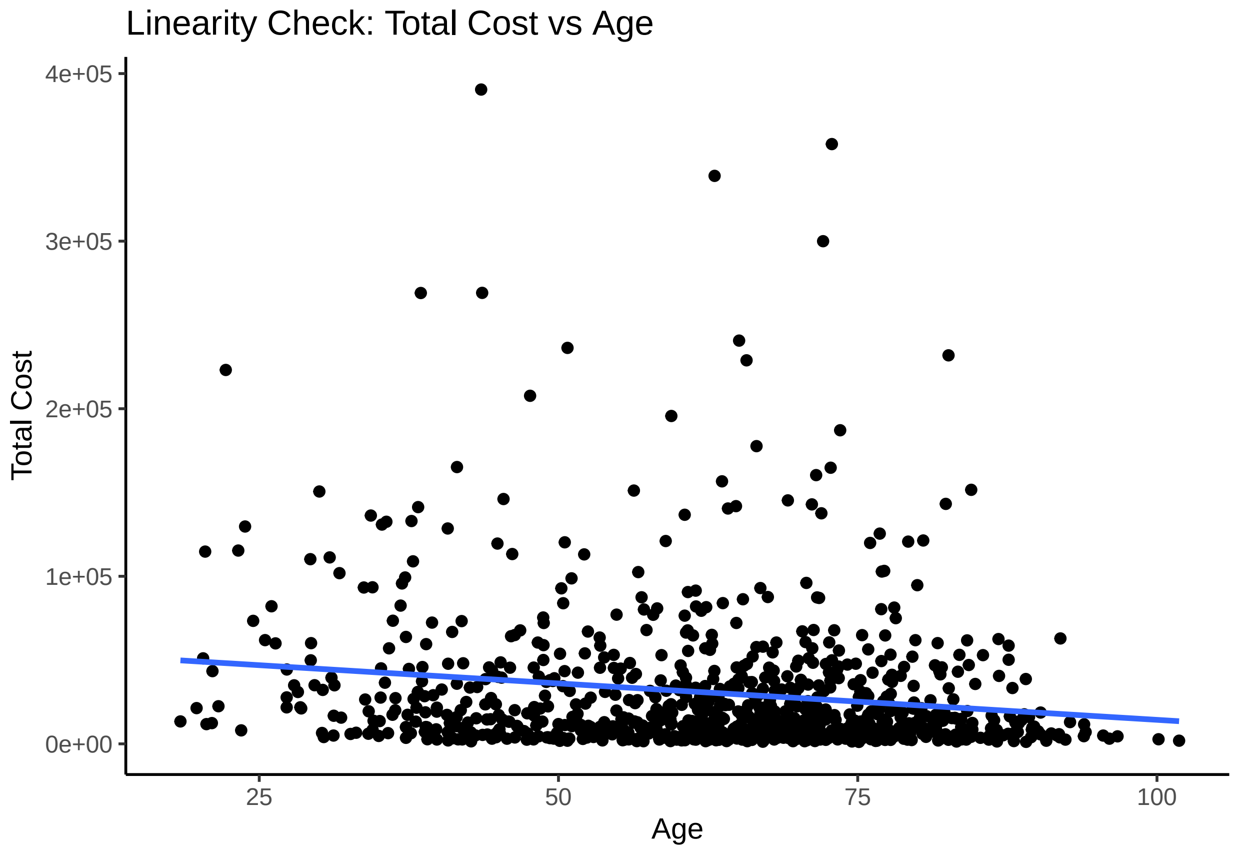
The ggplot2 way

```
mod <- lm(total_cost ~ age, data = support)
library(ggfortify)
autoplot(mod)
```



L - linearity

```
ggplot(support, aes(x = age, y = total_cost)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F) +
  labs(title = "Linearity Check: Total Cost vs Age",
       y = "Total Cost",
       x = "Age") +
  theme_classic()
```



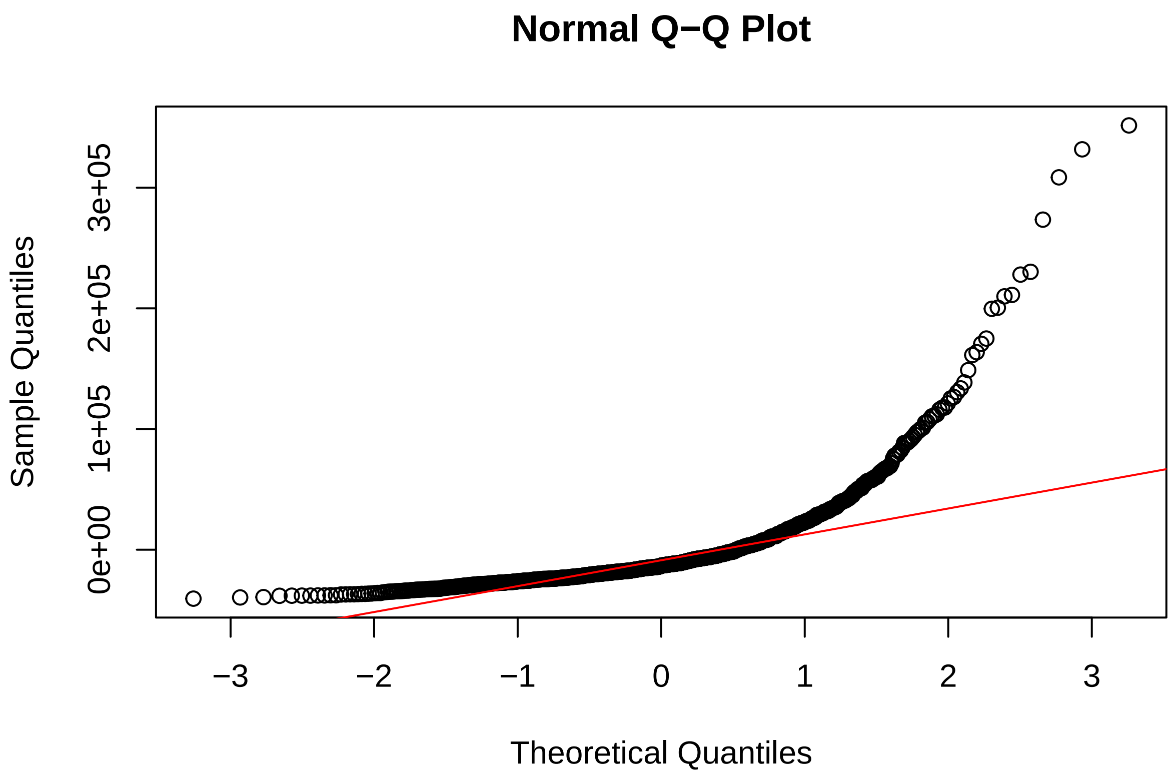
I - Independence

```r
library(lmtest)
dwtest(mod)
```

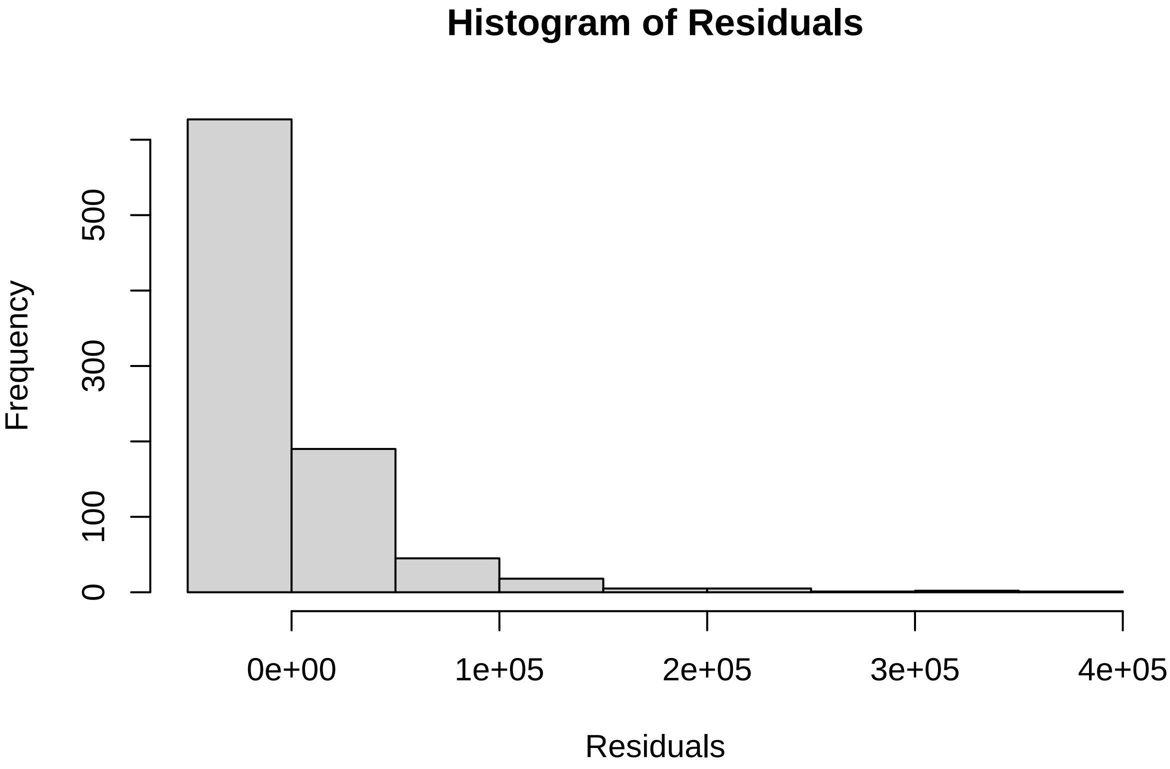
```
##
##  Durbin-Watson test
##
## data:  mod
## DW = 1.7727, p-value = 0.0003286
## alternative hypothesis: true autocorrelation is greater than 0
```

N - Normality

```r
qqnorm(resid(mod))
qqline(resid(mod), col = "red")
```

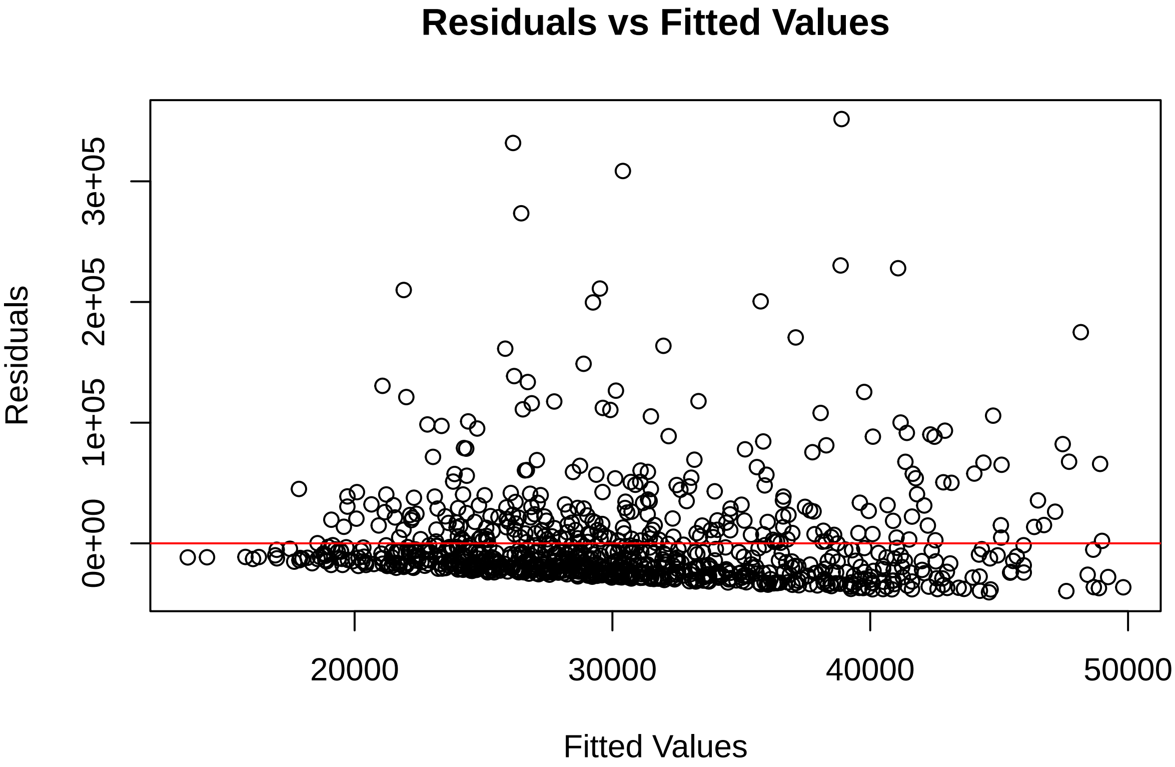


```r
hist(resid(mod),main = "Histogram of Residuals", xlab = "Residuals")
```



E - “Equal” (Constant) Variance

```r
plot(resid(mod) ~ fitted(mod),
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```



1. Linearity: Weak linear trend; partially satisfied.
2. Independence: DW = 1.7727: This is close to 2, which usually means no autocorrelation. P-value = 0.0003: Because this is very small, we reject the idea that there's no autocorrelation. Conclusion: Residuals are positively correlated, so independence is not satisfied.
3. Normality: Residuals are skewed; not satisfied.
4. Equal Variance: Variability increases with fitted values; not satisfied.

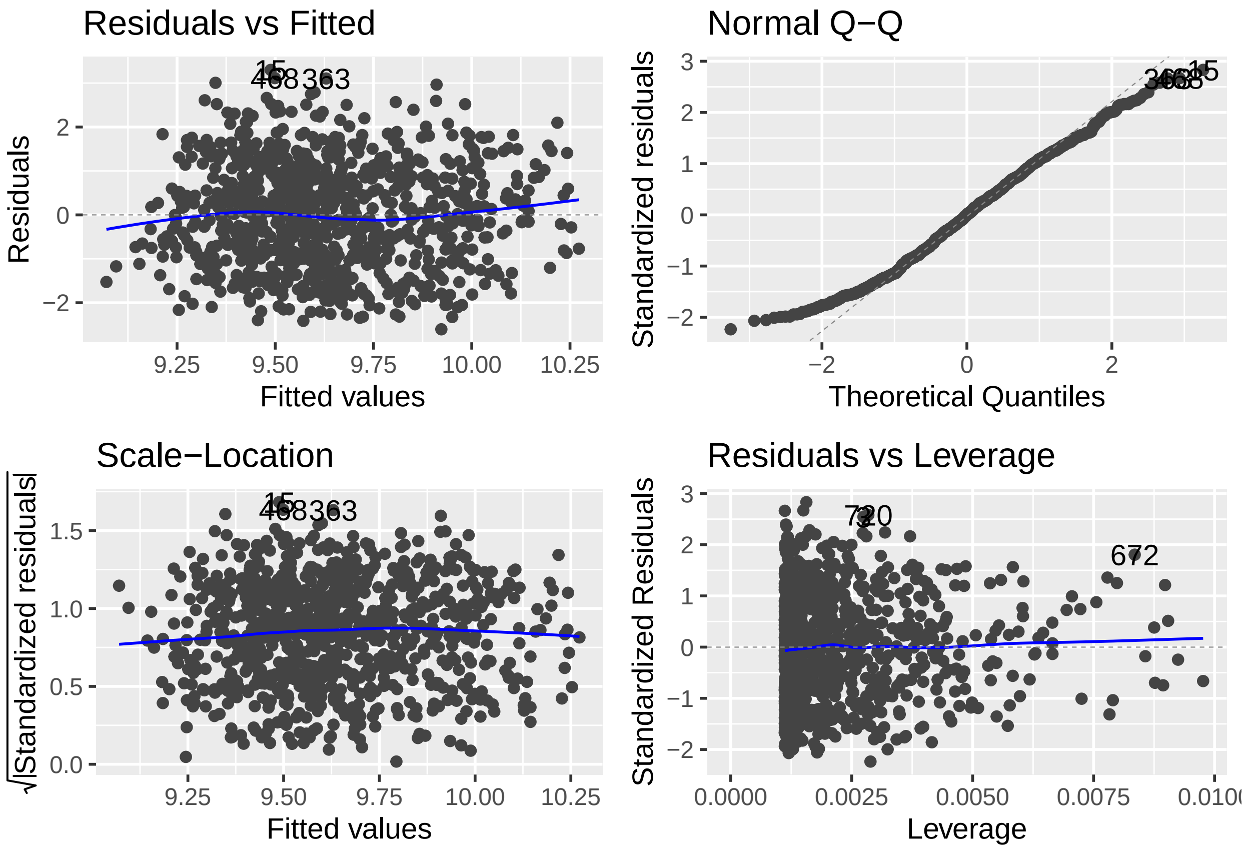
In summary, Linearity, Independence, Normality and Equal Variance assumptions are not satisfied.

11.

2. Check the necessary model assumptions again.

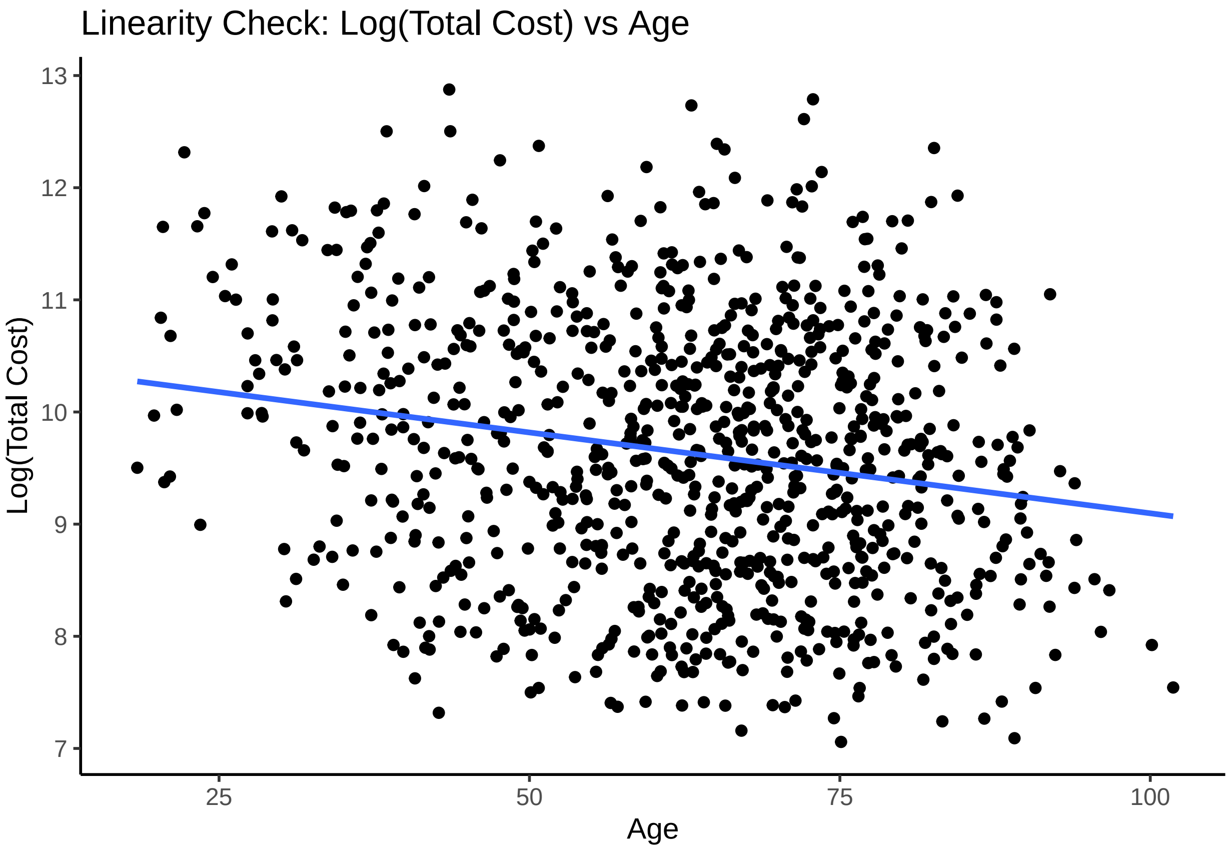
The ggplot2 way

```
mod2 <- lm(log_total_cost ~ age, data = support)
library(ggfortify)
autoplot(mod2)
```



L - linearity

```r
ggplot(support, aes(x = age, y = log_total_cost)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F) +
  labs(title = "Linearity Check: Log(Total Cost) vs Age",
       y = "Log(Total Cost)",
       x = "Age") +
  theme_classic()
```



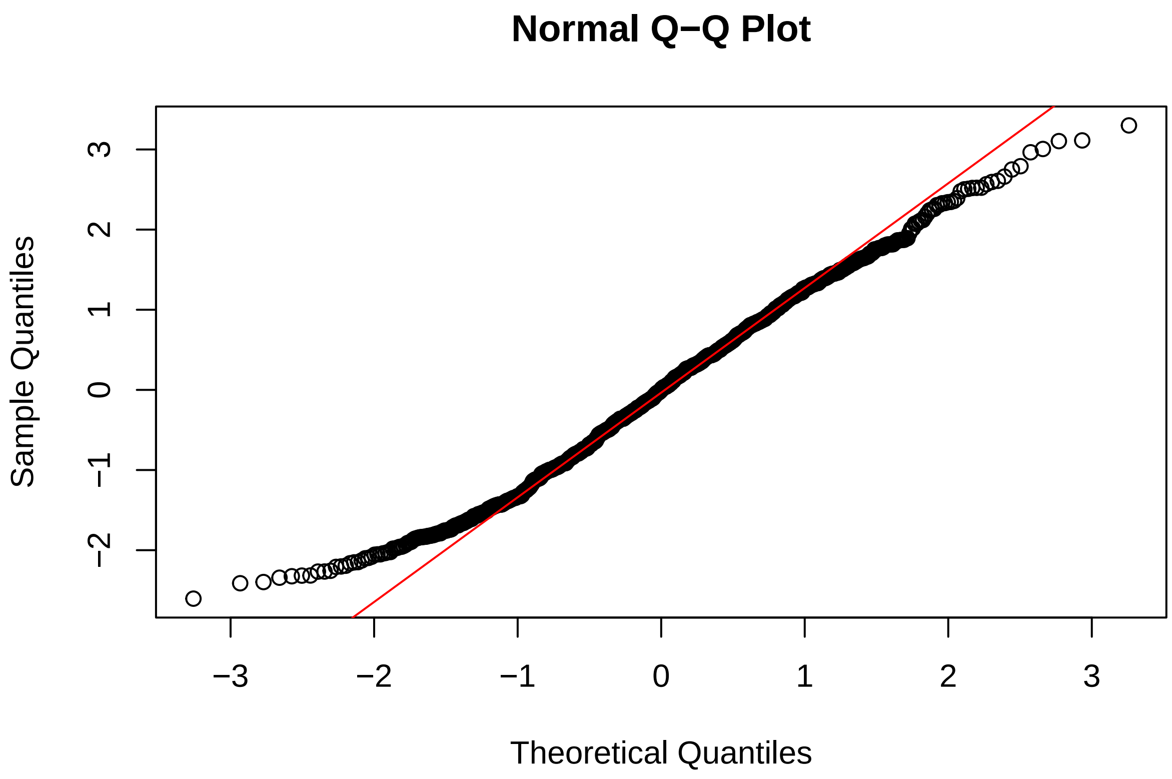
I - Independence

```r
library(lmtest)
dwtest(mod2)
```

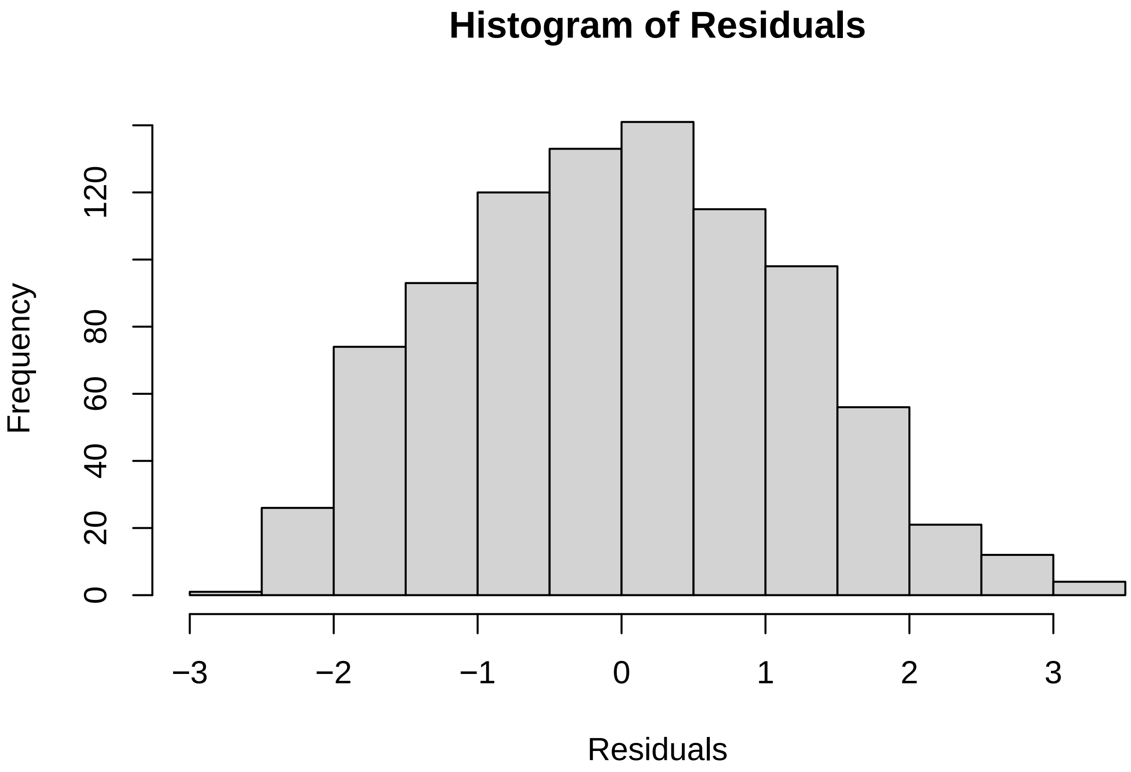
```
##
##  Durbin-Watson test
##
## data:  mod2
## DW = 1.9464, p-value = 0.2099
## alternative hypothesis: true autocorrelation is greater than 0
```

N - Normality

```r
# Q-Q Plot and Histogram for Residuals
qqnorm(resid(mod2))
qqline(resid(mod2), col = "red")
```

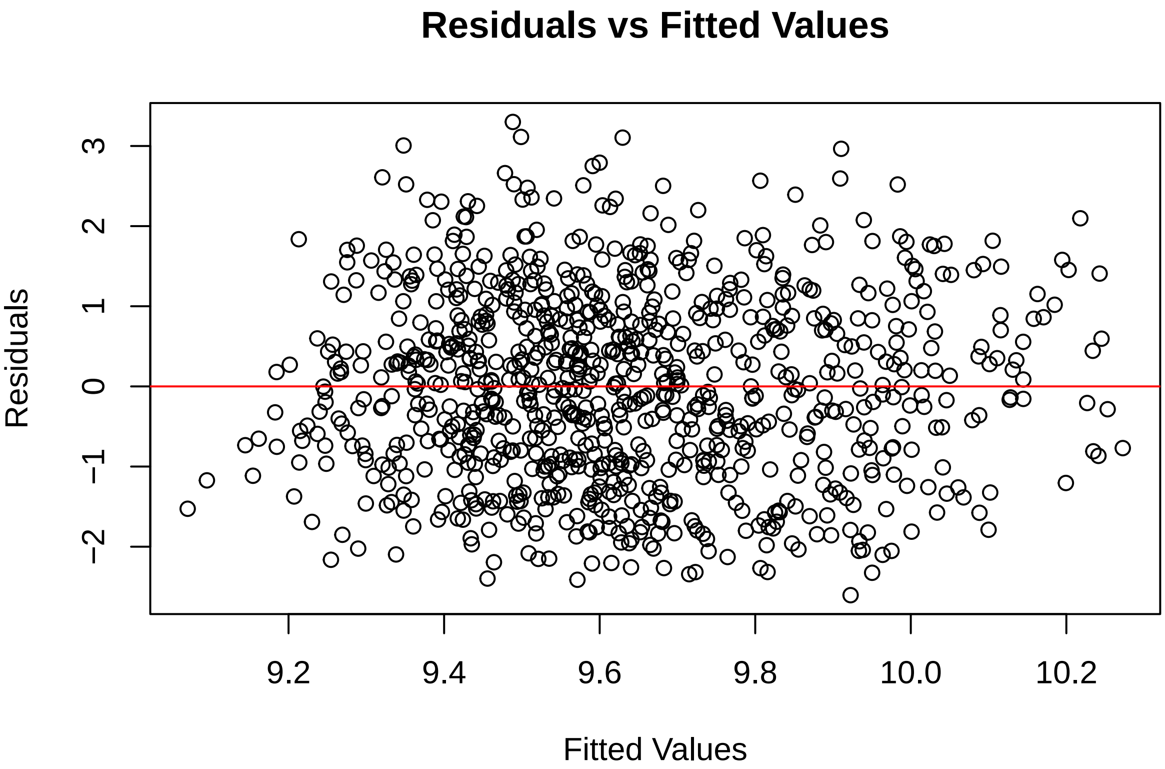


```
hist(resid(mod2), main = "Histogram of Residuals", xlab = "Residuals")
```



E - "Equal" (Constant) Variance

```
# Residuals vs Fitted Plot
plot(resid(mod2) ~ fitted(mod2),
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")
```



1. Linearity: The relationship between age and the log of total cost appears linear, based on the scatterplot with a smooth line.
2. Independence: Satisfied. The Durbin-Watson test for the log model (DW = 1.9464 and p-value = 0.2099). Because the p-value is greater than 0.05, we fail to reject the null hypothesis of no autocorrelation.
3. Normality: Satisfied. The residuals show a reasonably normal distribution in the Q-Q plot and histogram.
4. Equal Variance: Satisfied. The residuals have constant spread across fitted values in the residual vs. fitted plot.

In summary, the assumptions for this log-transformed model are better satisfied.

12.

Step 3: Test statistic

```
summary(mod2)$coefficients[2,3]
```

```
## [1] -5.892598
```

```
# Extract standard error
sqrt(diag(vcov(mod2)))
```

```
## (Intercept)         age
## 0.158379464 0.002445376
```

Step 4: p-value

```
p_value <- summary(mod2)$coefficients[2, 4]
p_value
```

```
## [1] 5.38548e-09
```

Step 5: Decision and Conclusion in context

```
alpha <- 0.05
if (p_value < alpha) {
  decision <- "Reject the H_0"
} else {
  decision <- "Fail to reject the H_0"
}
decision
```

```
## [1] "Reject the H_0"
```

Decision: Reject the H_0 Conclusion: We have enough evidence that there is a significant linear relationship between total cost and age.

13.

```
# compute the slope coefficient (beta)
beta <- summary(mod2)$coefficients[2, 1]
beta
```

```
## [1] -0.01440962
```

The slope coefficient of -0.0144 means that for each one-year increase in age, the log of total cost decreases by approximately 1.44%. In simpler terms, as age increases, the total cost tends to decrease, but the change is small.

Part 4: Discuss the Results

14.

Yes, we have enough evidence that age is linearly associated with the log of total cost. The hypothesis test in Question 12 showed a significant slope (p-value = 5.38548e-09), so we reject H_0, confirming a significant linear relationship. This confirms that age has a significant effect on the log of total cost. The negative slope of -0.0144 in Question 13 suggests that as age increases, total cost decreases slightly.

15.

No

16.

E - Excellent