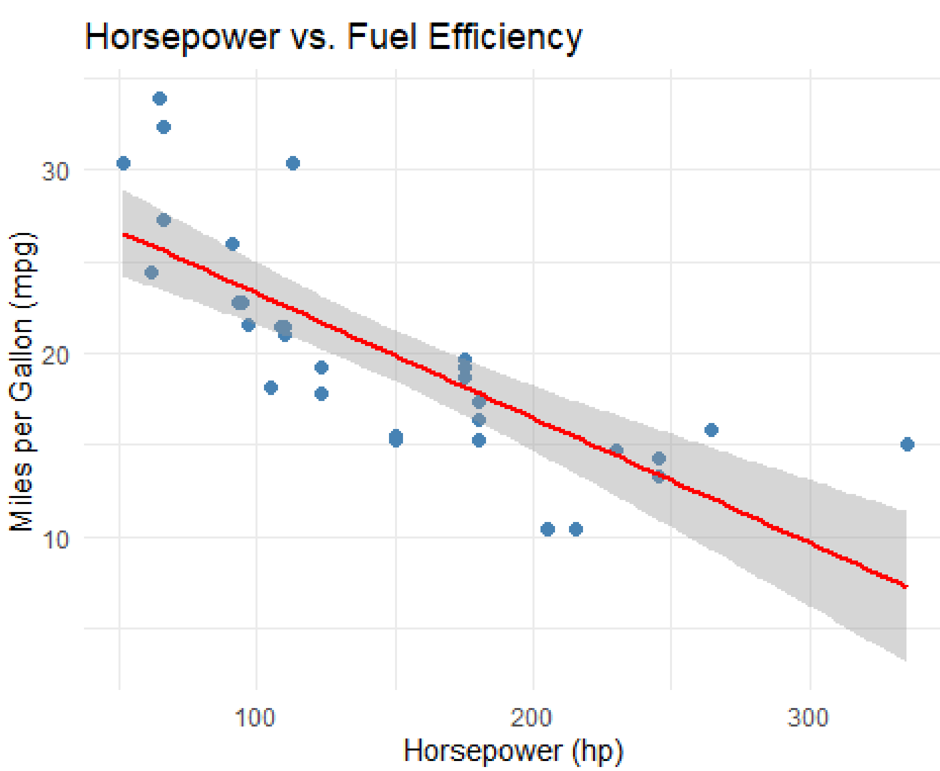
Tutorial 4 Memo

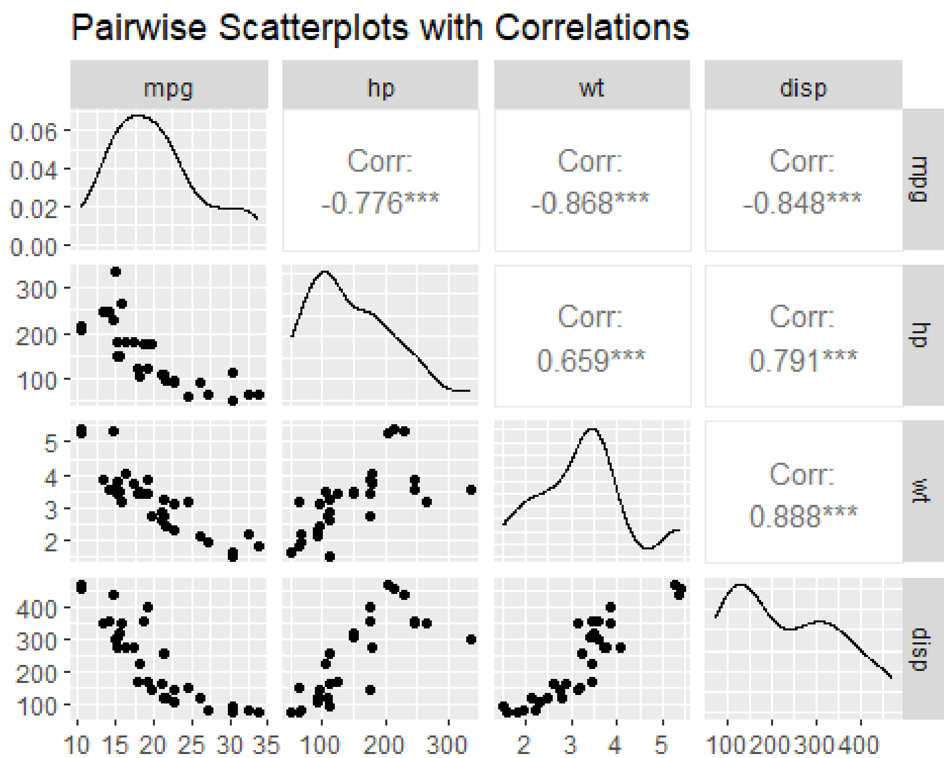
Dr. Niladri Chakraborty

2025-05-13

```r
# Load libraries
library(tidyverse)
library(broom)
library(ggfortify)
library(GGally)
library(ggpubr)

# Load data
data(mtcars)

# ------------------------------
# 1. Visualization: Scatterplot with linear smoothing
# ------------------------------
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point(color = "steelblue", size = 2) +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Horsepower vs. Fuel Efficiency",
       x = "Horsepower (hp)",
       y = "Miles per Gallon (mpg)") +
  theme_minimal()
```



```r
# -----------------------------------------
# 2. Histogram of Horsepower
# -----------------------------------------
p2 <- ggplot(mtcars, aes(x = hp)) +
  geom_histogram(binwidth = 20, fill = "orange", color = "black") +
  labs(title = "Histogram of Horsepower", x = "Horsepower", y = "Count") +
  theme_minimal()

# -----------------------------------------
# 3. Boxplot of MPG grouped by cylinder
# -----------------------------------------
p3 <- ggplot(mtcars, aes(x = factor(cyl), y = mpg, fill = factor(cyl))) +
  geom_boxplot() +
  labs(title = "Boxplot: MPG by Number of Cylinders", x = "Cylinders", y =
"Miles per Gallon") +
  theme_minimal()

# -----------------------------------------
# 4. Density plot of MPG
# -----------------------------------------
p4 <- ggplot(mtcars, aes(x = mpg)) +
  geom_density(fill = "lightblue") +
  labs(title = "Density Plot of MPG", x = "Miles per Gallon") +
  theme_minimal()

# ------------------------------
# 5. Correlation Analysis
# ------------------------------
p5 <- ggpairs(mtcars[, c("mpg", "hp", "wt", "disp")],
              title = "Pairwise Scatterplots with Correlations")

# Display pairwise plot separately
print(p5)
```




```r
# 6. Linear regression modeling
# ------------------------------------------
model <- lm(mpg ~ hp, data = mtcars)
summary(model)

##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

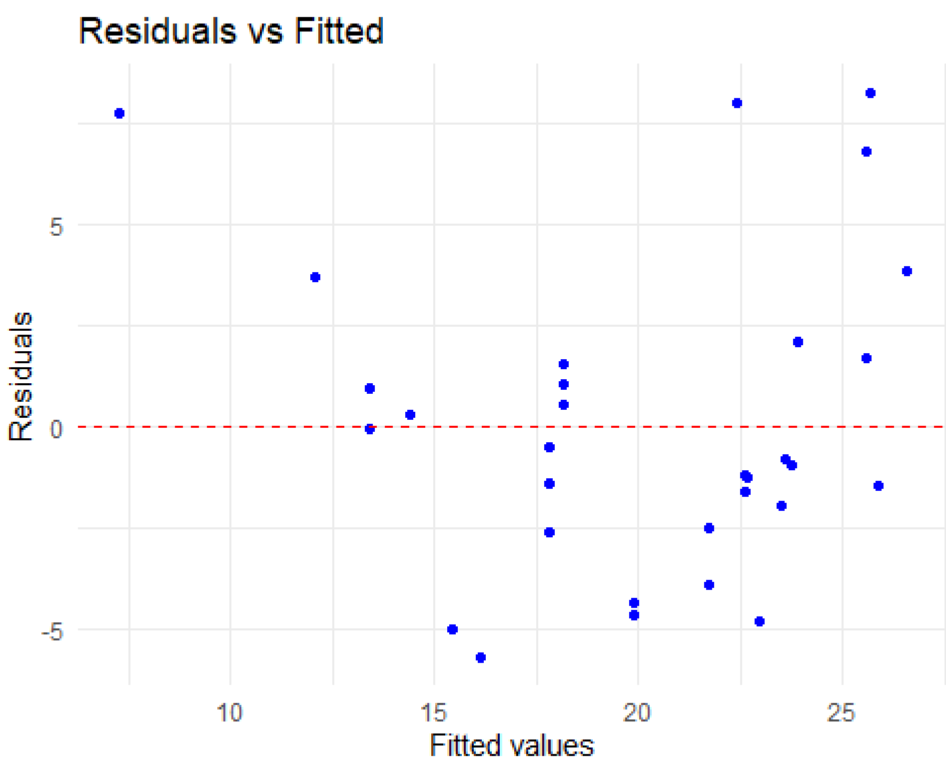
The output shows that, based on the p-values, the model is statistically significant for the dataset. However, R-squared and Adjusted R-squared values are not very high.

Let us perform some regression diagnostics as given below. We obtain the fitted (predicted) values of the 'mpg' variable and the corresponding residuals.

```
# Extract residuals and fitted values
residuals <- resid(model)
fitted <- fitted(model)
```

We create a plot of the residuals against the fitted values.

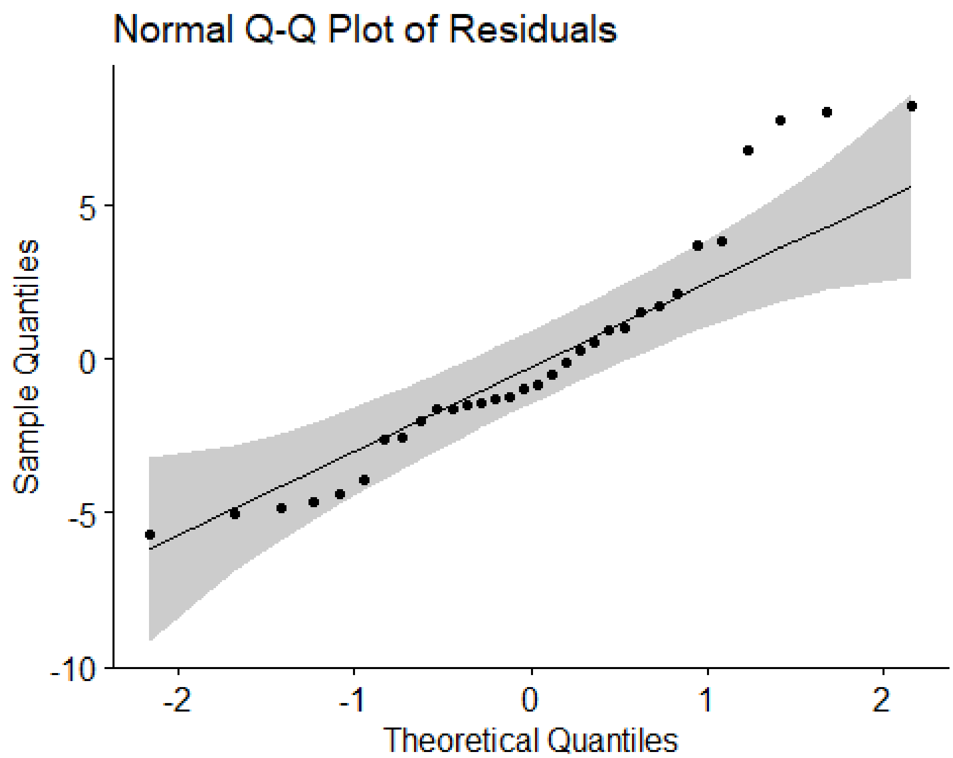
```
# ------------------------------------------
# 1. Residuals vs Fitted Values Plot
# ------------------------------------------
p1 <- ggplot(data = data.frame(fitted, residuals), aes(x = fitted, y =
residuals)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted", x = "Fitted values", y = "Residuals") +
  theme_minimal()
p1
```



The plot shows that the points follow a U-shape, i.e., the points are NOT scattered randomly.

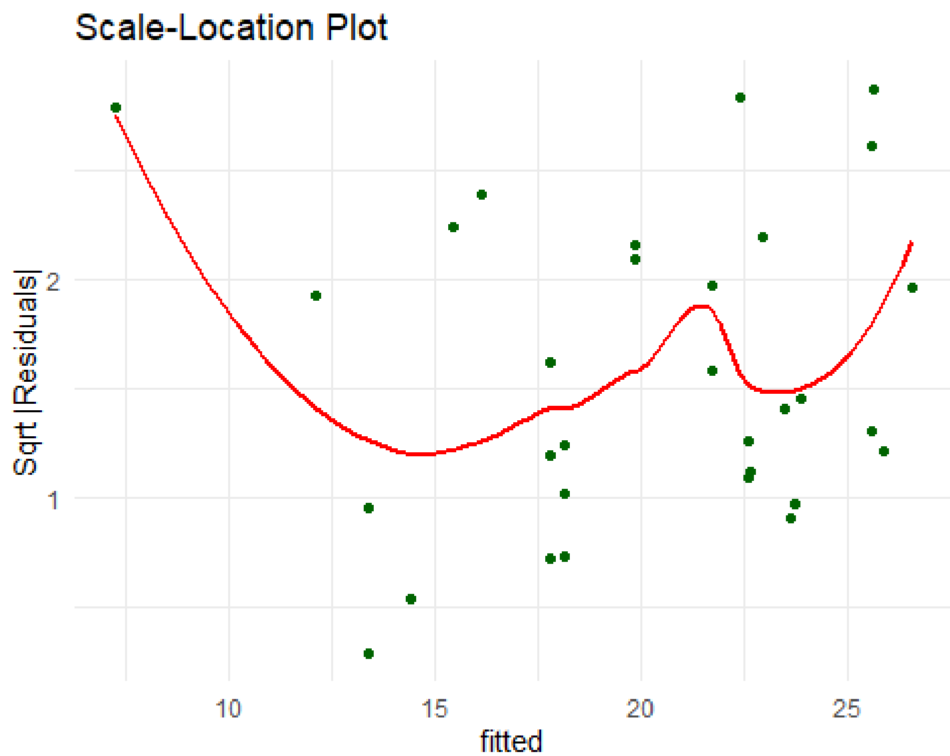
```
# ------------------------------------------
# 2. Normal Q-Q Plot
# ------------------------------------------
p2 <- ggqqplot(residuals,
```

```
                title = "Normal Q-Q Plot of Residuals",
                xlab = "Theoretical Quantiles", ylab = "Sample Quantiles")
p2
```



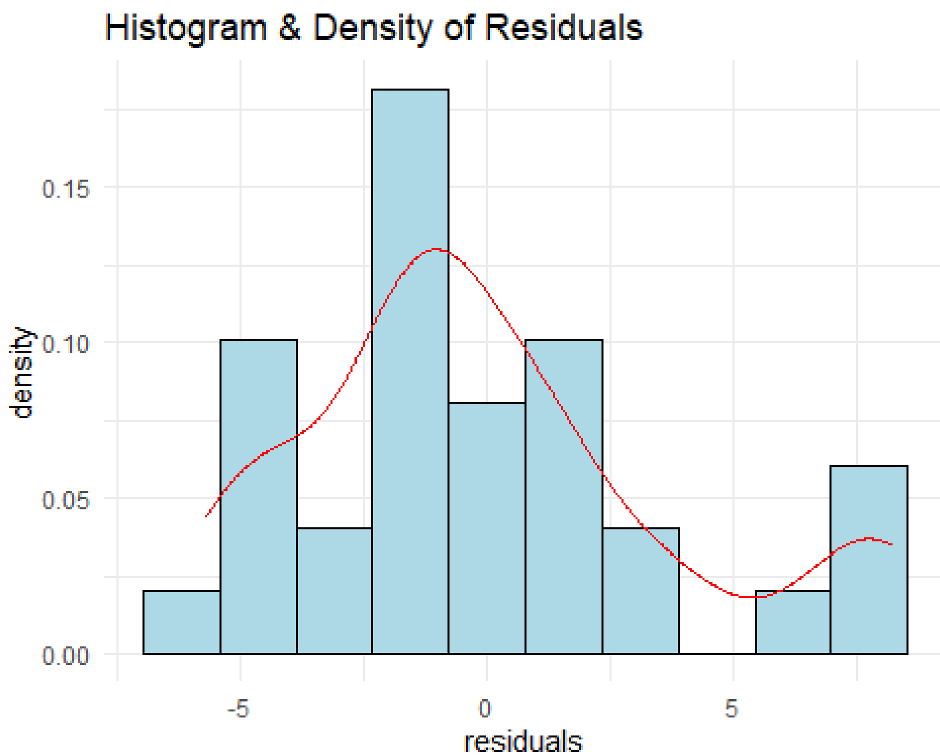
The Q-Q plot shows that the norma distribution assumption for the residuals is also not satisfied.

```
# 3. Scale-Location Plot (Spread vs Fitted)
# ---------------------------------------------
sqrt_resid <- sqrt(abs(residuals))
p3 <- ggplot(data = data.frame(fitted, sqrt_resid), aes(x = fitted, y =
sqrt_resid)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Scale-Location Plot", y = "Sqrt |Residuals|") +
  theme_minimal()
p3
```



The scale-location plot also shows a certain pattern, i.e., absense of randomness in the plot.

```
# ---------------------------------------------
# 4. Histogram and Density of Residuals
# ---------------------------------------------
p5 <- ggplot(data.frame(residuals), aes(x = residuals)) +
  geom_histogram(aes(y = ..density..), fill = "lightblue", bins = 10, color =
"black") +
  geom_density(color = "red") +
  labs(title = "Histogram & Density of Residuals") +
  theme_minimal()
p5
```



The histogram clearly shows that the residuals follow a bi-modal distribution, i.e., the normal distribution assumption for the residuals does not hold.

```
# -------------------------------------------
# 6. Shapiro-Wilk Normality Test
# -------------------------------------------
shapiro_test <- shapiro.test(residuals)
cat("Shapiro-Wilk p-value:", round(shapiro_test$p.value, 4), "\n")

## Shapiro-Wilk p-value: 0.0257
```

The p-value of the Shapiro-Wilks test suggests that the normality assumption is certainly invalid for the linear regression model that we fit for the data.

Next we apply a log-trnasformation on the 'mpg' values. We fit another linear regression model on the log(mpg) against the hp values. This is given below.

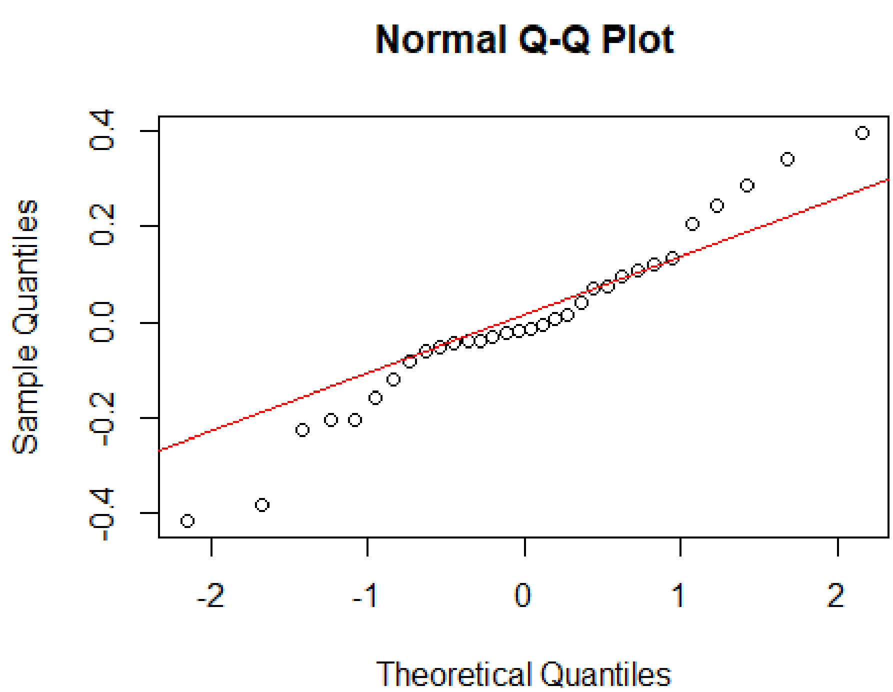
```
# Log-transform the response
model_log <- lm(log(mpg) ~ hp, data = mtcars)

# Check model summary
summary(model_log)

##
## Call:
## lm(formula = log(mpg) ~ hp, data = mtcars)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41577 -0.06583 -0.01737  0.09827  0.39621
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.4604669  0.0785838  44.035  < 2e-16 ***
## hp          -0.0034287  0.0004867  -7.045 7.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1858 on 30 degrees of freedom
## Multiple R-squared:  0.6233, Adjusted R-squared:  0.6107
## F-statistic: 49.63 on 1 and 30 DF,  p-value: 7.853e-08
```

```
# Recheck residual normality
shapiro.test(resid(model_log))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_log)
## W = 0.97261, p-value = 0.5744
```

```
# Q-Q plot for new residuals
qqnorm(resid(model_log))
qqline(resid(model_log), col = "red")
```



Now, after applying the transformation on the mpg values, we find that the model is significant and the normality assumption for the residuals is also valid. This is also observed by the Q-Q plot.

Based on the visualizations and statistical analysis, there is a clear negative relationship between horsepower and fuel efficiency. The scatterplot and linear regression model both suggest that cars with higher horsepower tend to have lower miles per gallon. The Pearson correlation coefficient is approximately -0.78, indicating a strong inverse relationship. Boxplots and density plots also highlight variation and skewness in mpg. Thus, the claim made by the data analyst is supported by the analytical outputs provided above.