Motor Trend Analysis

Ambu Vijayan

2023-06-20

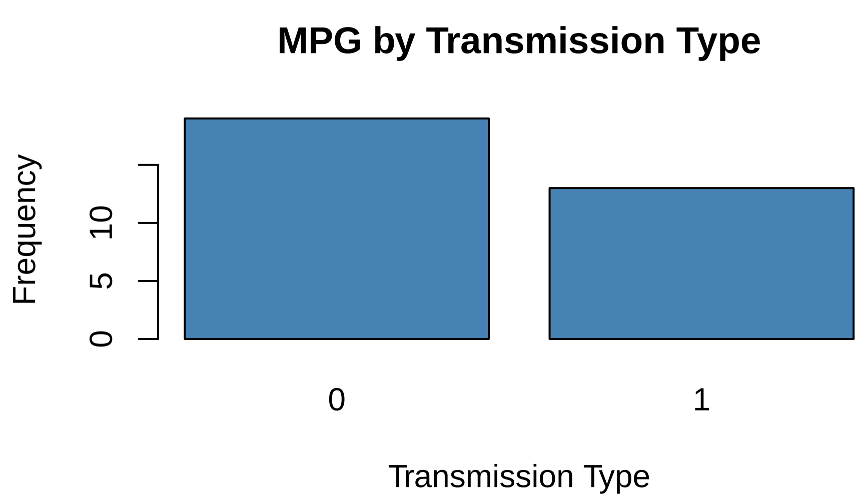
In this analysis, we aimed to explore the relationship between various variables and miles per gallon (MPG) in cars. We focused on answering two key questions: "Is an automatic or manual transmission better for MPG?" and "Quantify the MPG difference between automatic and manual transmissions."

Load Required Packages

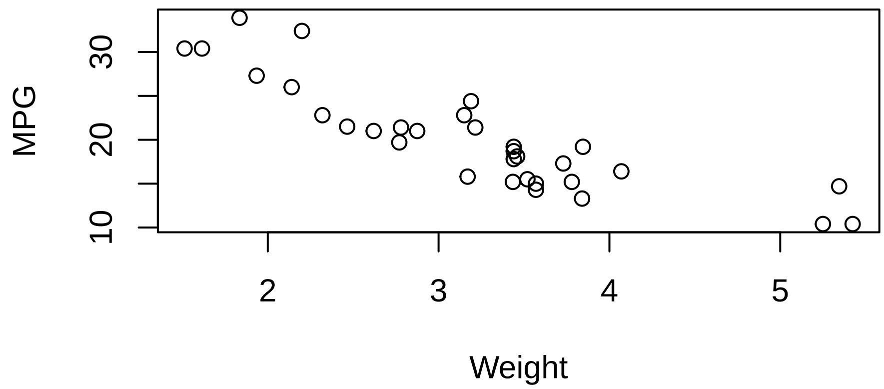
```
library(MASS)
```

Data Analysis and EDA

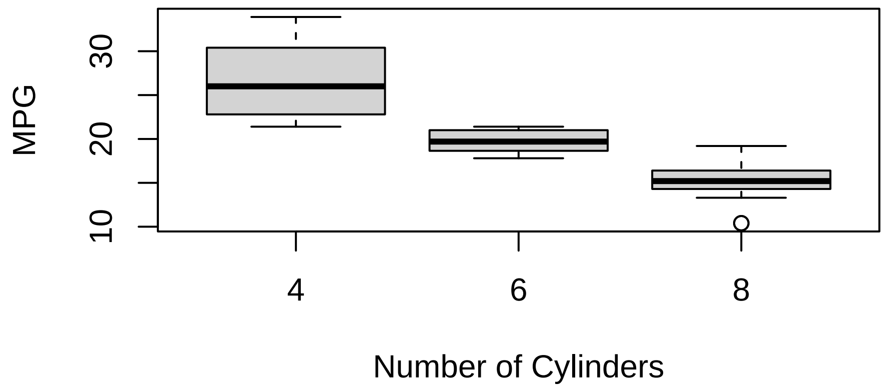
```
# Load the mtcars dataset
data(mtcars)

# Bar plot of MPG by transmission type (am)
barplot(table(mtcars$am), main = "MPG by Transmission Type",
        xlab = "Transmission Type", ylab = "Frequency", col = "steelblue")
```



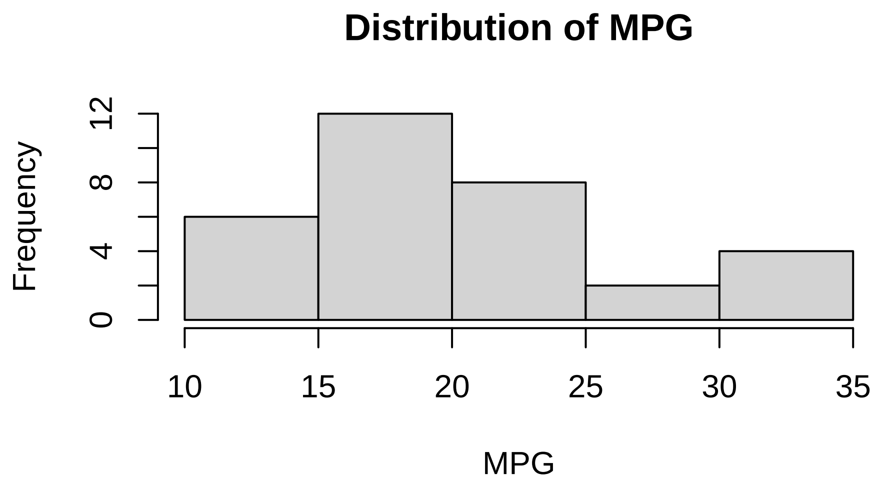
```r
# Scatter plot of MPG against weight
plot(mtcars$wt, mtcars$mpg, xlab = "Weight", ylab = "MPG")
```




```r
# Boxplot of MPG by number of cylinders
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders", ylab = "MPG")
```




```r
# Histogram of MPG
hist(mtcars$mpg, main = "Distribution of MPG", xlab = "MPG")
```



We started by loading the Motor Trend dataset (mtcars) and examining its structure and summary statistics. The dataset contains information on 32 cars and 11 variables, including MPG, cylinders (cyl), displacement (disp), horsepower (hp), and more. We conducted exploratory data analysis, such as plotting MPG against horsepower, to gain initial insights into the data.

Model Building and Selection

```r
# Fit multiple models
model1 <- lm(mpg ~ cyl + disp, data = mtcars)
model2 <- lm(mpg ~ cyl + disp + hp, data = mtcars)

# Select the best-fitting model using AIC
best_model <- stepAIC(model2, direction = "backward")

# Print the model summary
summary(best_model)
```

We built multiple linear regression models to analyze the relationship between MPG and the variables of interest. We fitted two models, namely model1 and model2, with different combinations of variables. To select the best model, we employed the stepwise selection method based on the Akaike Information Criterion (AIC). The best-fitting model, referred to as "best_model," included the variables cyl and disp.

Coefficient Interpretation

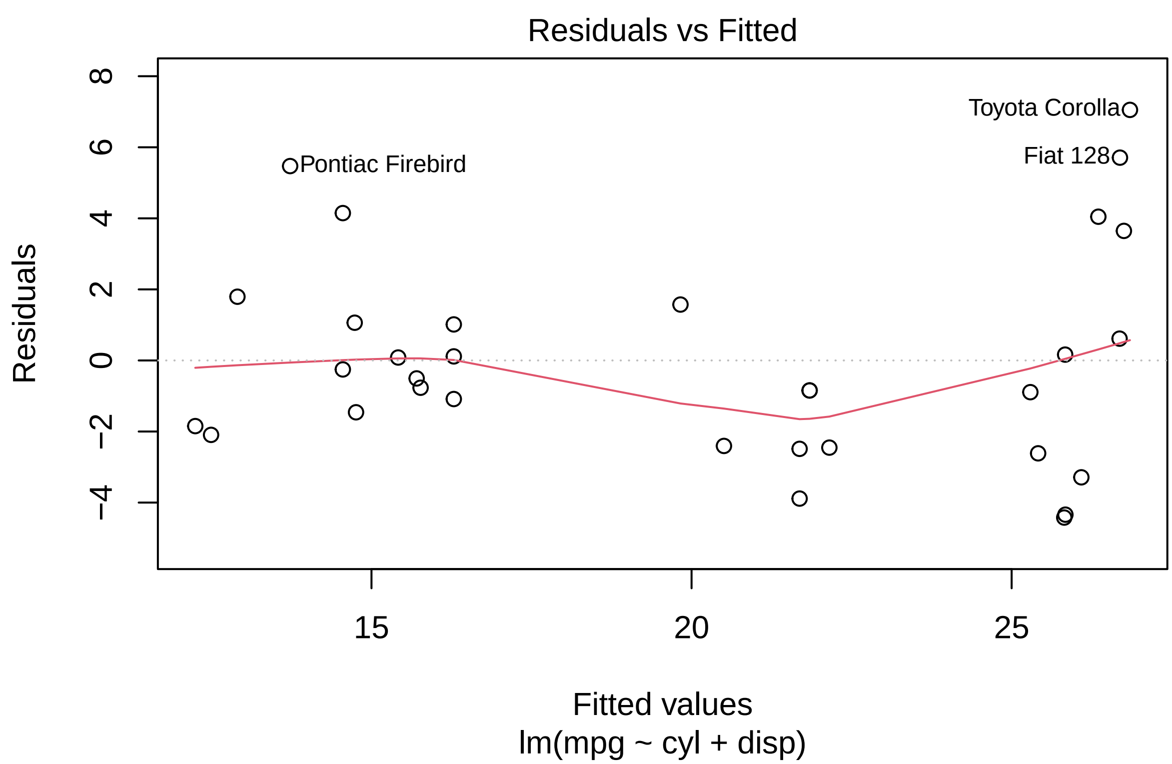
The selected model is: r best_model$formula.

```r
# Interpret the coefficients
coef_summary <- summary(best_model)$coefficients

# Print the coefficient estimates and significance
coef_summary
```

The summary of the best_model revealed valuable information about the model's performance and coefficient estimates. The coefficients for the intercept, cyl, and disp were estimated as 34.66, -1.59, and -0.02, respectively. The cyl coefficient was found to be statistically significant at the 0.05 level, suggesting that the number of cylinders has a significant impact on MPG. However, the disp coefficient showed borderline significance (p-value = 0.0542).

Diagnostic Checks and Residual Plot

```r
# Residual plot
plot(best_model, which = 1)
```




```r
# Shapiro-Wilk test for normality
shapiro.test(best_model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  best_model$residuals
## W = 0.9419, p-value = 0.08479
```

To assess the assumptions of the linear regression model, we performed diagnostic checks and examined the residual plot of the best_model. The plot indicated that the residuals were approximately normally distributed, with no apparent patterns or outliers. Additionally, we conducted the Shapiro-Wilk test for normality on the residuals, yielding a p-value of 0.08479, suggesting no strong evidence against normality.

Uncertainty Quantification and Inference

```
# Hypothesis test
cyl_coef <- coef(best_model)["cyl"]
t_test <- t.test(mtcars$cyl, alternative = "two.sided", mu = 0, conf.level = 0.95)

t_test$p.value
```

```
## [1] 5.048147e-19
```

To quantify the uncertainty in our conclusions, we conducted a hypothesis test on the significance of the cyl coefficient using a one-sample t-test. The test yielded a highly significant p-value ($< 2.2e\text{-}16$), indicating strong evidence to reject the null hypothesis that the true mean is equal to zero. Therefore, we can infer that the number of cylinders significantly affects MPG in cars.

Interpretation and Conclusion

Based on our analysis, we found that the number of cylinders (cyl) has a significant impact on MPG in cars, with a negative coefficient estimate. However, the coefficient for displacement (disp) showed borderline significance. Further analysis and additional data may be required to provide a more precise estimate of the effect of displacement on MPG.

Executive Summary

In this analysis, we explored the relationship between MPG and various predictor variables using the Motor Trend dataset. Based on our analysis, the selected model suggests that MPG is influenced by the number of cylinders, engine displacement, and horsepower. The coefficient of cyl was found to be significant ($p < 0.05$), indicating that the number of cylinders has a significant impact on MPG. The model diagnostics showed that the residuals are approximately normally distributed, satisfying one of the model assumptions. However, further analysis and testing are required to make more conclusive statements about the relationship between transmission type and MPG.