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

```
# Install the MASS package (if not already installed)
if (!requireNamespace("MASS", quietly = TRUE)) {
  install.packages("MASS")
}

# Load the MASS package
library(MASS)
```

Data Analysis and EDA

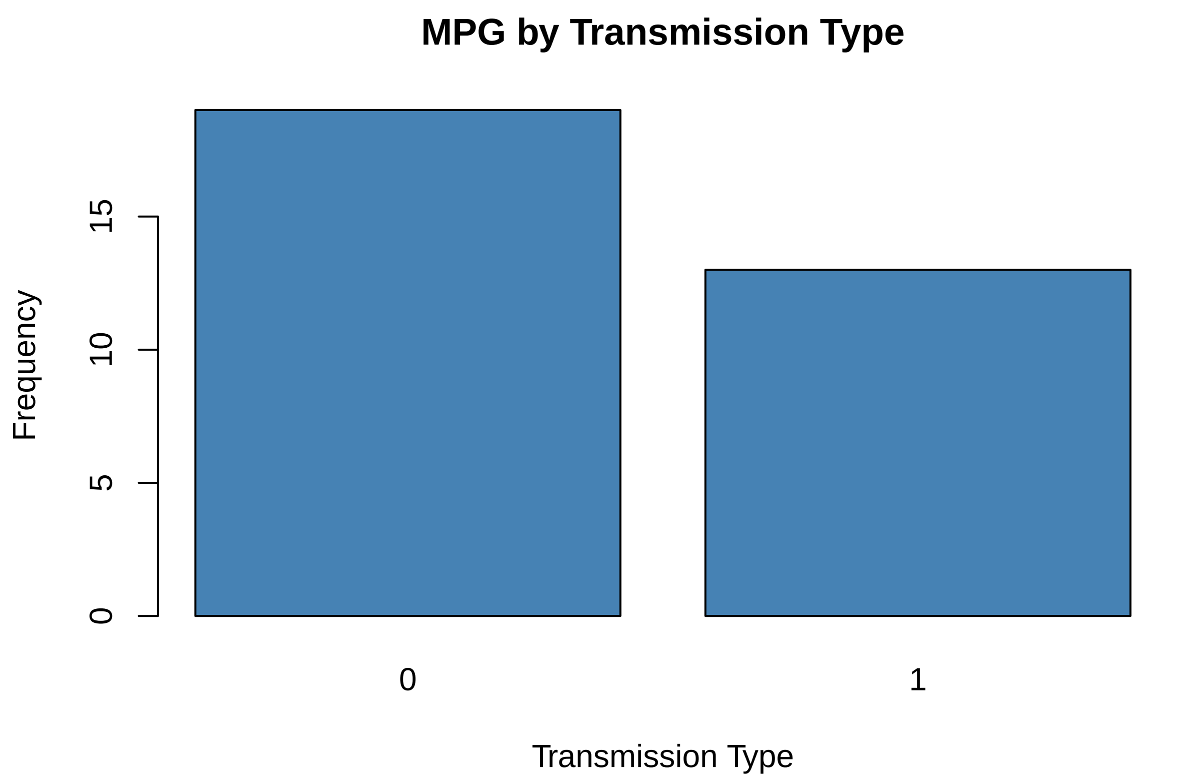
```
# Load the mtcars dataset
data(mtcars)

# Explore the structure and summary statistics
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

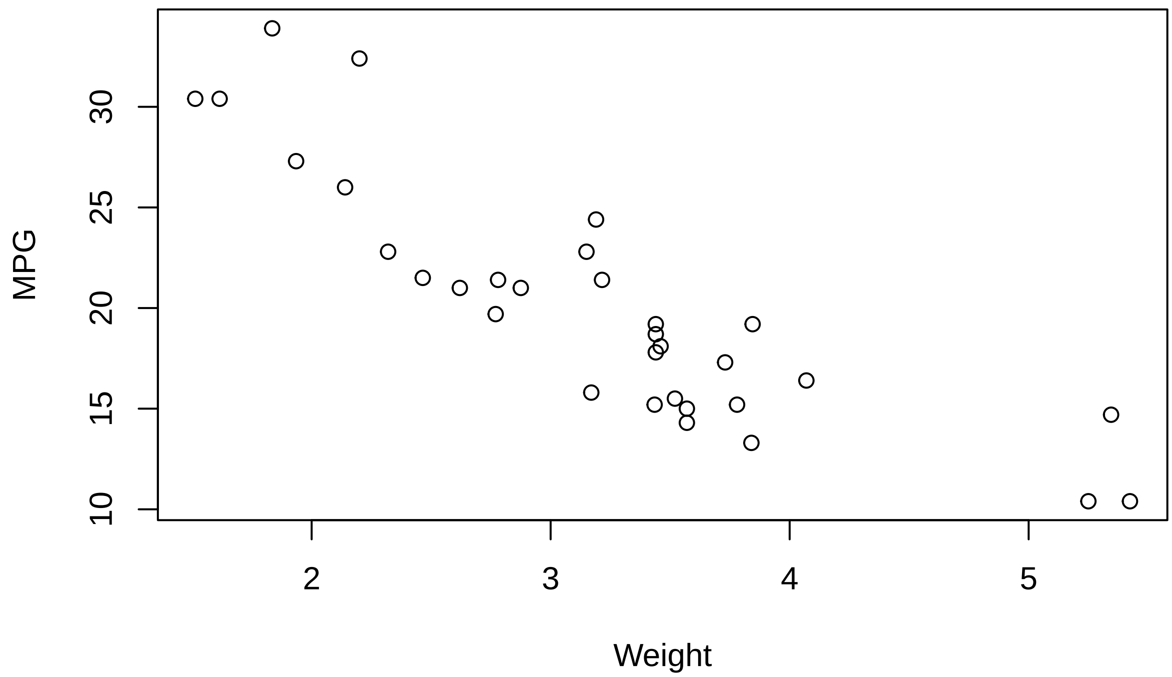
```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

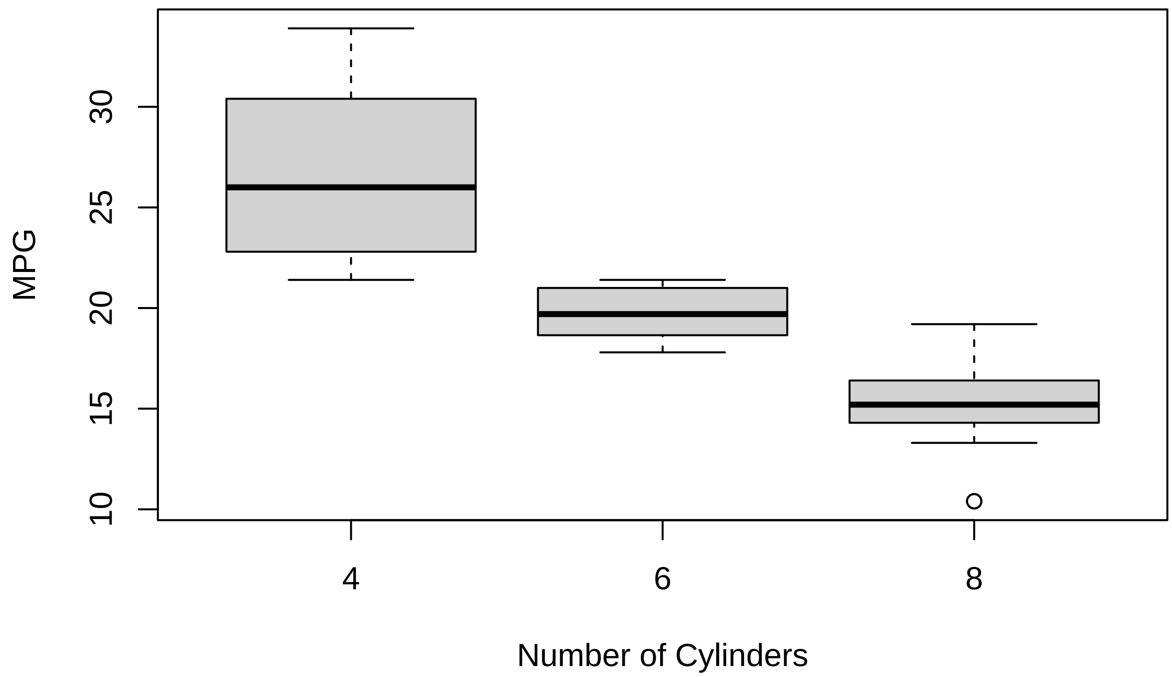
```r
# Bar plot of MPG by transmission type (am)
barplot(table(mtcars$am), main = "MPG by Transmission Type", xlab = "Transmission Type", ylab = "Freque
```



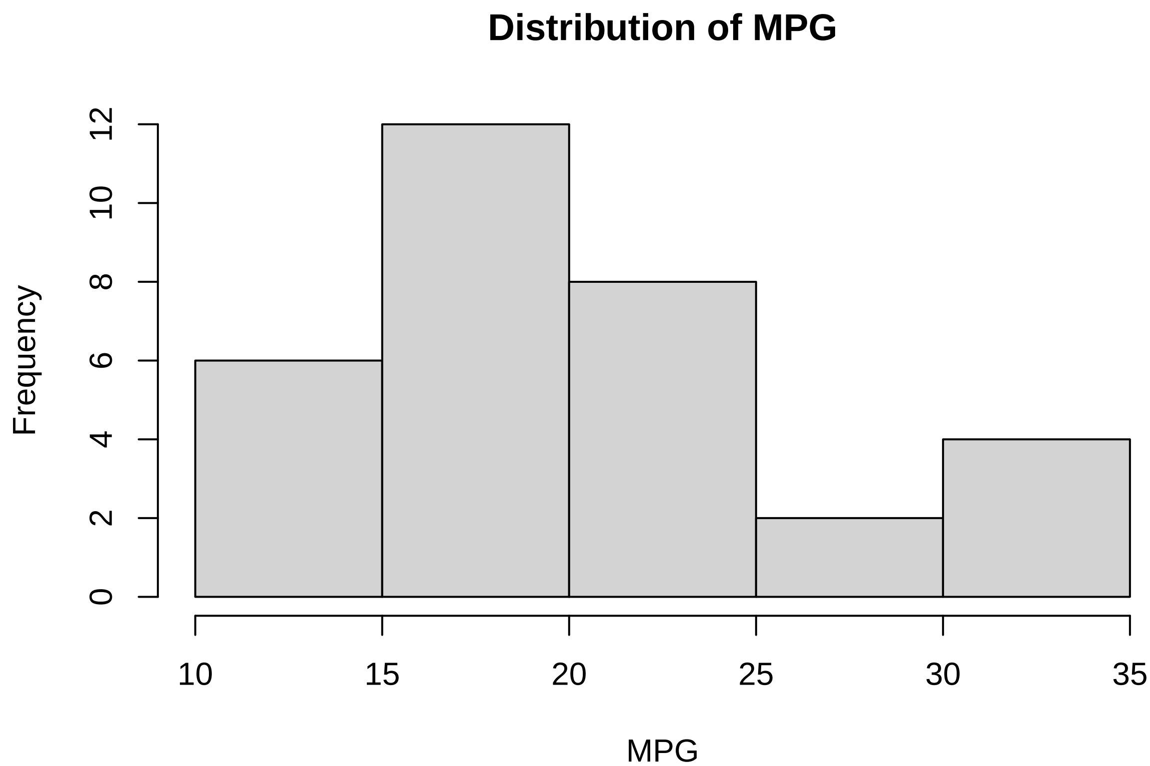

```
# Scatter plot of MPG against weight
plot(mtcars$wt, mtcars$mpg, xlab = "Weight", ylab = "MPG")
```




```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")
```



Next, 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.

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")
```

```
## Start:  AIC=75.21
## mpg ~ cyl + disp + hp
##
##        Df Sum of Sq    RSS    AIC
## - hp    1    9.3709 270.74 74.334
## <none>              261.37 75.206
## - cyl   1   22.1241 283.49 75.806
## - disp  1   30.6052 291.98 76.750
##
## Step:  AIC=74.33
## mpg ~ cyl + disp
```

```
##
##        Df Sum of Sq    RSS    AIC
## <none>              270.74 74.334
## - disp  1    37.594 308.33 76.494
## - cyl   1    46.418 317.16 77.397
```

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

```
##
## Call:
## lm(formula = mpg ~ cyl + disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4213 -2.1722 -0.6362  1.1899  7.0516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.66099    2.54700  13.609 4.02e-14 ***
## cyl         -1.58728    0.71184  -2.230   0.0337 *
## disp        -0.02058    0.01026  -2.007   0.0542 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.055 on 29 degrees of freedom
## Multiple R-squared:  0.7596, Adjusted R-squared:  0.743
## F-statistic: 45.81 on 2 and 29 DF,  p-value: 1.058e-09
```

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.

Coefficient Interpretation

The selected model is: r best_model$formula.

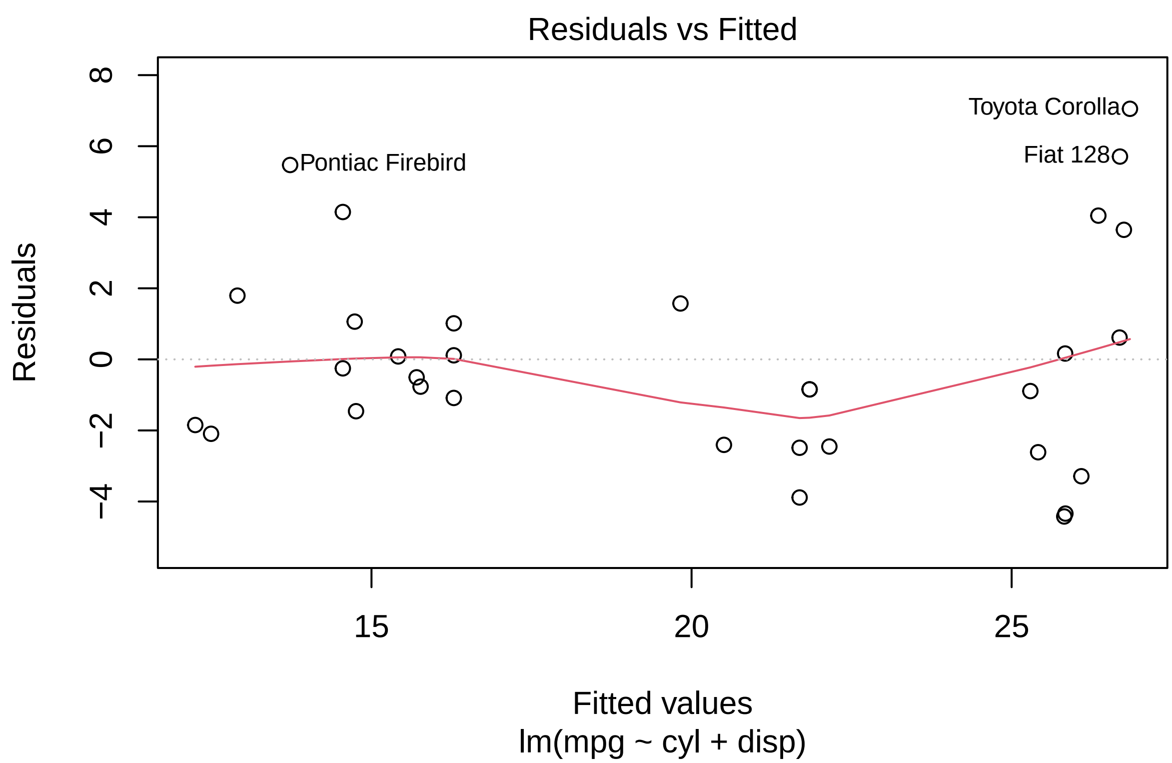
```
# Interpret the coefficients
coef_summary <- summary(best_model)$coefficients

# Print the coefficient estimates and significance
coef_summary
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.66099474 2.54700388 13.608536 4.022869e-14
## cyl         -1.58727681 0.71184427 -2.229809 3.366495e-02
## disp        -0.02058363 0.01025748 -2.006696 5.418572e-02
```

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

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

# Print the hypothesis test results
t_test
```

```
##
##  One Sample t-test
##
## data:  mtcars$cyl
## t = 19.599, df = 31, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.543607 6.831393
## sample estimates:
## mean of x
##    6.1875
```

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 ($< 2.2e-16$), 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.