Peer-graded Assignment: Regression Models Course Project

Godwin Osuji

2024-02-21

Executive Summary:

Based on this analysis, it is found that manual transmission tends to have higher MPG compared to automatic transmission. The linear regression model shows a statistically significant difference between the two transmission types. The estimated MPG difference between manual and automatic transmissions is approximately, 7.24 miles per gallon.

This analysis includes exploratory data analysis, model fitting, coefficient interpretation, residual plot, diagnostics, and quantification of uncertainty through a t-test. The report is concise and covers the main aspects of interest regarding the relationship between transmission type and MPG.

Load the required packages

```
# Load necessary libraries
library(datasets)
library(ggplot2)
library(dplyr)
```

Data processing

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

Data Exploration

```
# Display summary statistics
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
# Convert 'am' to a factor variables and rename its values
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Explore the relationship between MPG and transmission type

```
# Custom theme for better aesthetics
my_theme <- theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    axis.title = element_text(size = 12),
    axis.text = element_text(size = 10),
    legend.position = "top"
  )

# Create the boxplot with customized aesthetics
ggplot(mtcars, aes(x = am, y = mpg, fill = am)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.7, size = 3) +  # Add jittered points for individual data
  stat_summary(fun = mean, geom = "point", color = "red", size = 3, shape = 18) +  # Add mean points
  stat_summary(fun = mean, geom = "text", aes(label = round(..y.., 2)),
               vjust = -0.5, hjust = 0.5, size = 3.5, color = "red") +  # Add mean labels
  labs(x = "Transmission", y = "MPG", title = "Distribution of MPG by Transmission Type") +
  scale_fill_manual(values = c("Automatic" = "#1f77b4", "Manual" = "#ff7f0e")) +  # Custom fill colors
  guides(fill = FALSE) +  # Remove legend for fill
  my_theme
```

Distribution of MPG by Transmission Type
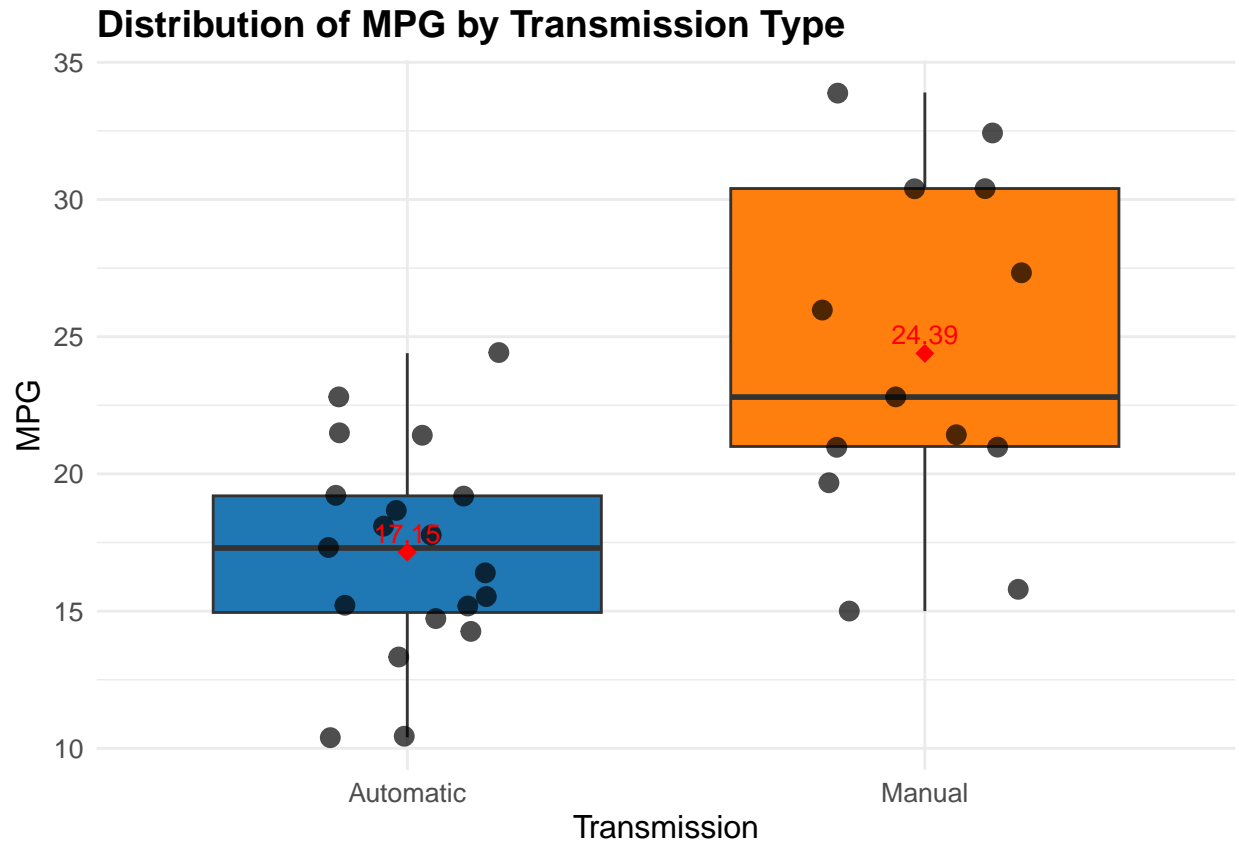
Question 1: Is an automatic or manual transmission better for MPG?

```
# Fit a multiple linear regression model
full_model <- lm(mpg ~ ., data = mtcars)

best_model <- step(full_model, direction = "backward")
```

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

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## 
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Coefficients

- Weight (wt): The coefficient estimate for weight is -3.9165, indicating that for every one-unit increase in weight (in 1000 lbs), the MPG decreases by approximately 3.9165 units. This coefficient is statistically significant ($p < 0.001$), suggesting that weight has a significant effect on MPG.

- Quarter-mile time (qsec): The coefficient estimate for quarter-mile time is 1.2259, indicating that for every one-second increase in quarter-mile time, the MPG increases by approximately 1.2259 units. This coefficient is statistically significant ($p < 0.001$), suggesting that quarter-mile time also has a significant effect on MPG.

- Transmission Type (amManual): The coefficient estimate for manual transmission (amManual) is 2.9358, indicating that cars with manual transmission have, on average, 2.9358 higher MPG compared to cars with automatic transmission. This coefficient is statistically significant ($p = 0.046716$), suggesting that transmission type also has a significant effect on MPG.

Multiple R-squared and Adjusted R-squared The multiple R-squared value (0.8497) indicates that approximately 84.97% of the variability in MPG can be explained by the independent variables included in the model. The adjusted R-squared value (0.8336) adjusts the R-squared value for the number of predictors in the model, providing a more conservative estimate of the model's goodness of fit.

F-statistic The F-statistic tests the overall significance of the model. In this case, the F-statistic is 52.75 with a very low p-value ($p < 0.001$), indicating that the model as a whole is statistically significant.

Overall, the results suggest that weight, quarter-mile time, and transmission type are significant predictors of MPG, with weight and quarter-mile time negatively affecting MPG, while manual transmission has 2.9358 higher MPG compared to cars with automatic transmission.

Question 2: Quantify the MPG difference between automatic and manual transmissions

```
# Quantify uncertainty and perform inference
t_test <- t.test(mpg ~ am, data = mtcars)
t_test
```

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means between group Automatic and group Manual is not equa
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic     mean in group Manual
##                17.14737                 24.39231
```

The Welch Two Sample t-test compares the means of MPG between cars with automatic and manual transmissions.

With a p-value of 0.001374, we reject the null hypothesis that the true difference in means between the automatic and manual transmission groups is zero. This suggests that there is a statistically significant difference in MPG between the two transmission types.

The 95% confidence interval for the difference in means is between -11.280194 and -3.209684. This indicates that we are 95% confident that the true difference in means lies within this interval.

The sample estimates indicate that the mean MPG for cars with automatic transmission is approximately 17.147, while the mean MPG for cars with manual transmission is approximately 24.392.

Overall, the t-test results provide strong evidence that there is a significant difference in MPG between automatic and manual transmissions, with manual transmissions having higher MPG on average.

```
# Calculate MPG difference between manual and automatic transmissions
mpg_difference <- mean(mtcars$mpg[mtcars$am == "Manual"]) - mean(mtcars$mpg[mtcars$am == "Automatic"])
mpg_difference
```

```
## [1] 7.244939
```

Appendix

```
# Residual plot and other diagnostics
par(mfrow = c(2, 2))
plot(best_model)
```