Regression Models Course Project: MPG Data Analysis

Morteza Taiebat

May 24, 2020

Executive Summary

In this report, we analyze the relationship between fuel efficiency in terms of miles per gallon (MPG) and transmission type (manual or automatic) using `mtcars` dataset. We determine how the transmission type impacts MPG and quantify the extent.

The analysis demonstrates the following observations: A t-test between transmission types reveals that a manual transmission yields a better miles per gallon, when compared with automatic (7.245 greater MPG). Fitting several linear regressions,it appears that the manual transmission contributes less significantly to MPG, only an improvement of 1.81 MPG. Further analysis shows a correlation between MPG and other variables, including weight, horsepower, and number of cylinders contributes more significantly to the overall MPG of vehicles. This may be sue to the confounding effects of some of these variables on both transmission type and MPG.

Load Required Packages and Data

Load the required packages (`ggplot2` & `GGally`) and the dataset. Convert categorical variables to factors.

```
data(mtcars)
mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual")) # 0=automatic, 1=manual
```

Exploratory Data Analysis

In the forst step, we have a closer look at the dimensions of dataset adn run a pairwise comparison to get a better understanding of the relationships between variables. *Appendix Figure 1* shows the pairwise relationship between all variables accounting for manual and automatic transmission. Several interesting observation emerges.

Appendix Figure 2 visually shows a significant increase in MPG for vehicles with a manual transmission compared to automatic, when no other variables is involved. Also, running a T-Test (with function `t.test()`) between transmission type and MPG rejects the null hypothesis that the difference between transmission types is insignificant. In fact, we observe a 7.24494 higher MPG for manual transmissions on average.

Regression Analysis

We start by running a regression where MPG is only explained by the transmission type. This is equivalent to a T-test. The results are now shown. The p-value of `single_model` is less than 0.0003, so we cannot reject the hypothesis that transmission type has an effect on MPG. However, the R-squared value for this regression is only 0.35, suggesting that nearly a third of variance in MPG can be attributed to transmission type alone.

```
single_model <- lm(mpg ~ am, data = mtcars); summary(single_model) # results not shown
```

The p-value is less than 0.0003, so we will not reject the hypothesis. However, the R-squared value for this test is only 0.35, suggesting that nearly a third of variance in MPG can be attributed to transmission type alone.

Next, we fit a model with all variables. We observe that all variables are statistically significant.

```
full_model <- lm(mpg ~ ., data = mtcars); summary(full_model) # results not shown
```

same can be done with anova test Now we use a Stepwise Algorithm `step()` function to determine which variables are most statistically significant using AIC. (the same process can be done by an `anova` test and manually selecting the significant variables)

```
reduced_model <- step(full_model) # results not shown
```

```
summary(reduced_model)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 33.70832390 2.60488618 12.940421 7.733392e-13
## cyl6        -3.03134449 1.40728351 -2.154040 4.068272e-02
## cyl8        -2.16367532 2.28425172 -0.947214 3.522509e-01
## hp          -0.03210943 0.01369257 -2.345025 2.693461e-02
## wt          -2.49682942 0.88558779 -2.819404 9.081408e-03
## amManual     1.80921138 1.39630450  1.295714 2.064597e-01
```

The new reduced model carries four variables where they minimize AIC criterion. These variables are *cylinders*, *horsepower*, *weight*, and *transmission*. The From R-squared, we can see that the reduced model explains nearly 87% of the variance in MPG. The estimated coefficients show that increasing the number of cylinders from 4 to 6 with decrease the MPG by 3.03. Further increasing the cylinders to 8 with decrease the MPG by 2.16. Increasing the horsepower decreases MPG by 3.21 for every 100 horsepower. Weight decreases the MPG by 2.5 for each 1000 lbs increase. A manual transmission improves the MPG by 1.81 but its p-value is 0.2 suggesting that the estimated coefficient is statistically insignificant. P-values for cyl and weight suggest that there is likely a confounding relationship between car Transmission Type and MPG.

Some diagnostics residual slots are shown in the **Appendix Figure 3**. We can observe that the heteroscedasticity is a minimal issue (the Scale-Location plot random distribution confirms the constant variance assumption). The residuals are normally distributed (shown in the normal Q-Q plot), and there are only a few outliers.

```
sum((abs(dfbetas(reduced_model)))>1)
```

```
## [1] 0
```

Conclusion

This report investigates the effect of transmission type on MPG. In the first glance, it appears that a manual transmissions deliver a higher MPG than automatic transmissions. However, when we control for weight, horsepower, & number of cylinders, the transmission type is no longer statistically significant. This might be explained by the confounding variables that contribute to both MPG and transmission type.

Appendix

Figure 1

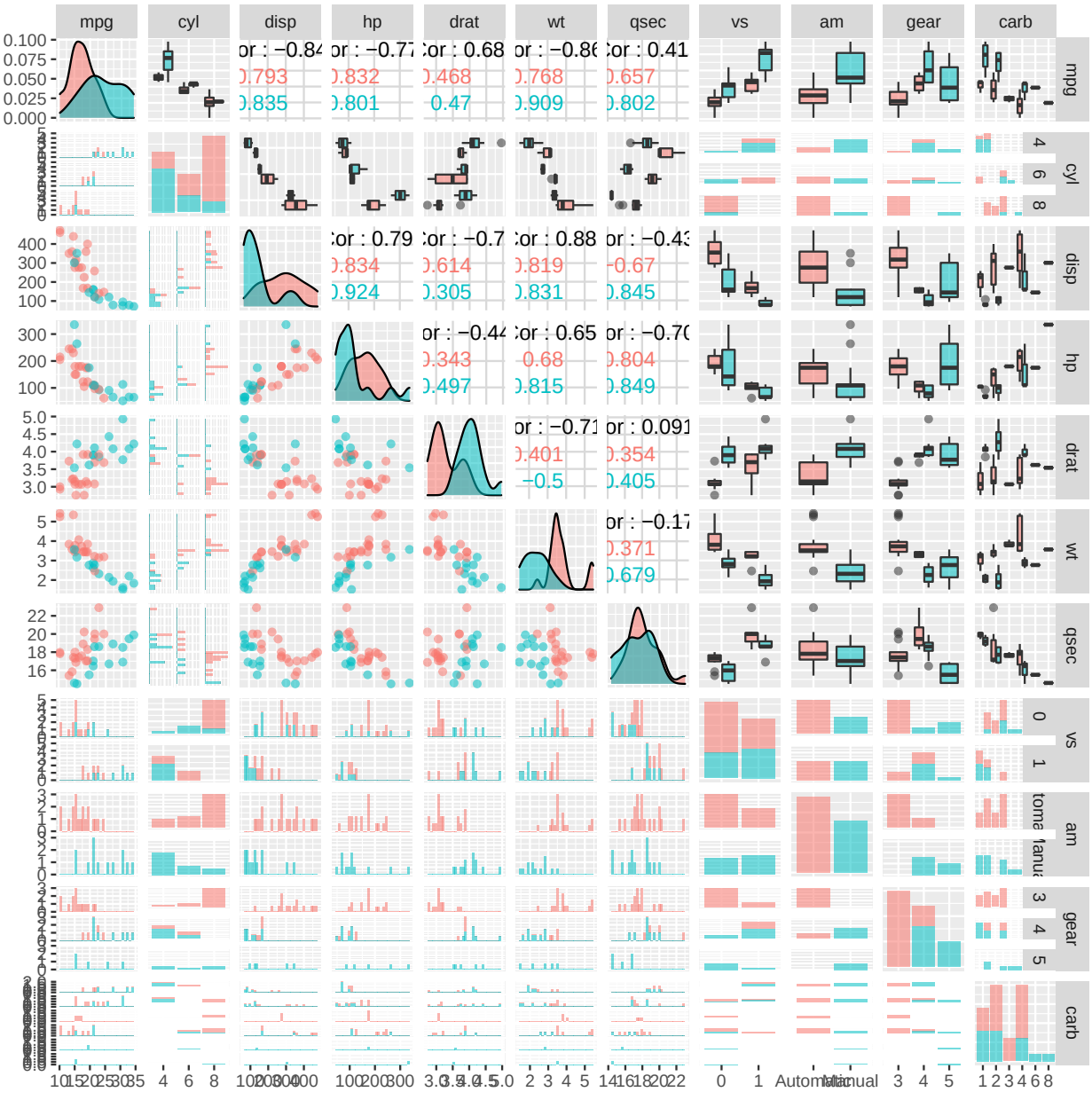

Figure 2

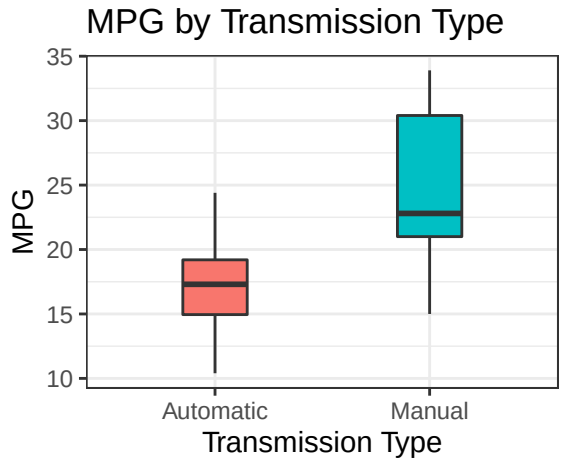

Figure 3