Analyzing Motor Trends Data

Vanshaj S.

Executive Summary

Our goal in this analysis is to examine the **Motor Trends** data to answer the following questions:

1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.

The analysis below establishes the idea that `mpg` is barely, if at all, related to `am`, with a manual transmission giving a 0.177 mpg benefit over an automatic transmission, suggesting that it is entirely possible that there is in fact no relationship at all. However, the analysis also reveals that the results are confounded by the weight and number of cylinders, which have a much greater role in determining fuel efficiency than transmission type.

Exploratory Data Analysis

Before getting into the specific details of our analysis, let's explore some of the data.

```
data(mtcars)
```

Refer to the first figure in the appendix for a boxplot of fuel efficiency by transmission. The boxplot shows a clear difference in fuel efficiency based on transmission type, with manual transmissions giving greater fuel efficiency than automatic transmissions.

Now that we've explored some glaring relationships in the data, let's get into a more in-depth analysis.

T-Test

The simplest (and probably the least accurate, too, in this case) way to determine whether transmission affects fuel efficiency is to perform a t-test.

```
res.tt <- t.test(mpg ~ am, data = mtcars)
```

The p-value of 0.0013736 falls significantly under a 0.05 alpha level, suggesting that there is in fact a relationship between fuel efficiency and transmission type. However, there may be multiple other variables at play, so let's try some other things.

Fitting

Let's generate a few fit models to test against each other. We'll start with a couple of basic, unlikely linear models.

```
fit.full <- lm(mpg ~ ., data = mtcars)   # fit based on all other variables
fit.tran <- lm(mpg ~ am, data = mtcars)  # fit based solely on transmission
```

We'll go ahead and print out the p-values for each variable of these two fits.

```
summary(fit.full)$coef[, 4]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##  0.51812440  0.91608738  0.46348865  0.33495531  0.63527790  0.06325215
```

```
##        qsec          vs          am        gear        carb
##  0.27394127  0.88142347  0.23398971  0.66520643  0.81217871
```

```
summary(fit.tran)$coef[, 4]
```

```
##  (Intercept)           am
## 1.133983e-15 2.850207e-04
```

Assuming that the second model (which only relates transmission to fuel efficiency) is correct, the estimates suggest a 7.2449393 mpg increase in cars with manual transmissions. This number agrees with our t-test from before, which also gave a mean estimate difference of 7.2449393.

However, the r-squared value of this model is a mere 0.3597989, so it is likely that there are other variables at play. None of the p-values in the `fit.full` estimates are significant though, so we'll perform an analysis of variance to help us determine which variables matter.

```
summary(aov(mpg ~ ., data = mtcars))  # results not printed for brevity
```

From the generated results, it appears that the `cyl`, `disp`, and `wt` variables have the most significant effects on the variance of our data. We'll therefore use a model with those three variables, along with our variable of interest - `am`. For the sake of brevity, let's just look at the p-values.

```
fit.aov <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(fit.aov)$coef[,4]
```

```
##  (Intercept)          cyl         disp           wt           am
## 8.677574e-12 7.581533e-03 5.450930e-01 5.468412e-03 9.229196e-01
```

This model produces an r-squared value of 0.8326661, which is pretty high. The p-values for `cyl` and `wt` are under 0.05, suggesting that these variables confound the relationship between transmission type and fuel efficiency. Let's create a final model that uses these two variables and our variable of interest.

```
fit.final <- lm(mpg ~ cyl + wt + am, data = mtcars)
summary(fit.final)$coef[,1]
```

```
## (Intercept)         cyl          wt          am
##  39.4179334  -1.5102457  -3.1251422   0.1764932
```

Our final model then is `mpg = 39.418 - 1.510cyl - 3.125wt + 0.177am`. This suggests that the true difference in fuel efficiency is simply a 0.177 mpg benefit in manual transmission vehicles.

Uncertainty & Residuals

As far as uncertainty is concerned, below is a 95% confidence interval for the estimates in our final model.

```
t(confint(fit.final))
```

```
##        (Intercept)        cyl        wt        am
## 2.5 %     34.00715 -2.3752454 -4.991001 -2.495555
## 97.5 %    44.82871 -0.6452459 -1.259284  2.848541
```

The results presented do not include zero for the `cyl` and `wt` variables, meaning that it is highly likely that there's a relationship involved there. However, the `am` variable results **do** include zero, which makes it entirely within the realm of possibility that transmission is in fact unrelated to the fuel efficiency of a vehicle.

Refer to the last figure in the appendix below for residual diagnostic plots. The first one shows an essentially nonexistent trend in the residuals. The second shows a near-straight line, suggesting normality in the residuals. The other two plots essentially reinforce this information, which together establishes that the residuals suggest, for the most part, that this fit is appropriate.

Appendix

Below are a few useful figures to support the analysis.

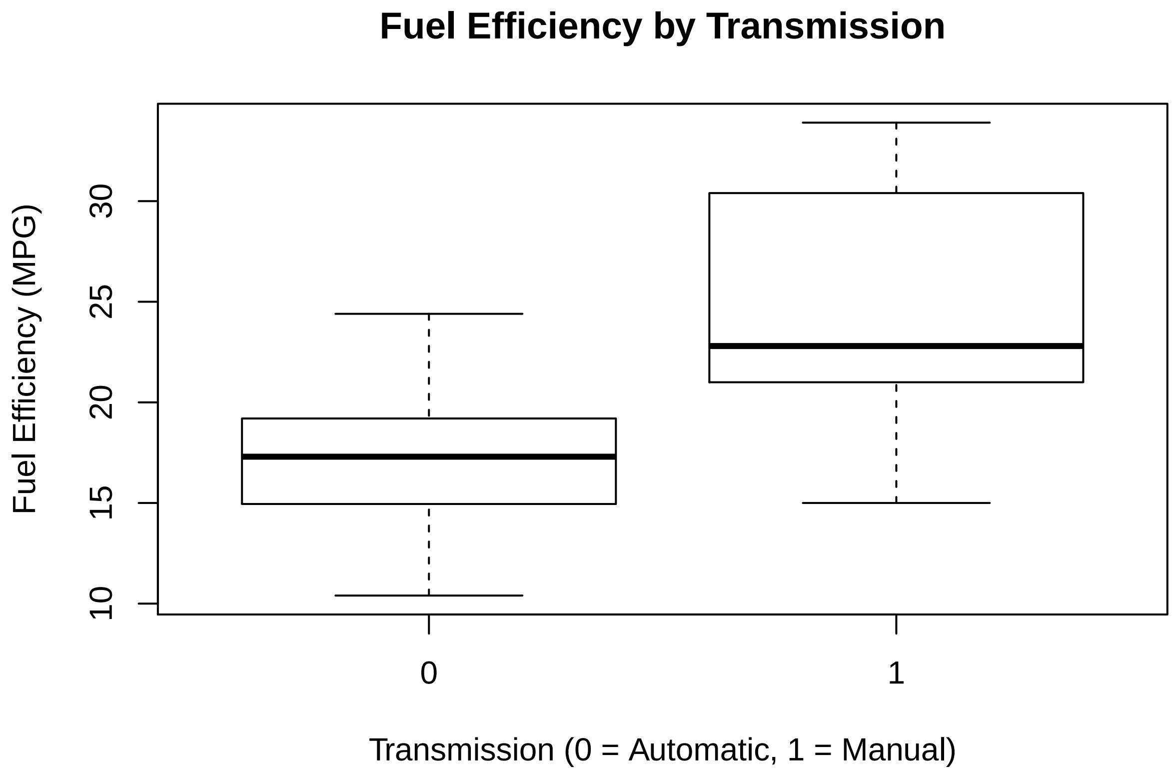

Figure 1. Boxplot of fuel efficiency by transmission.

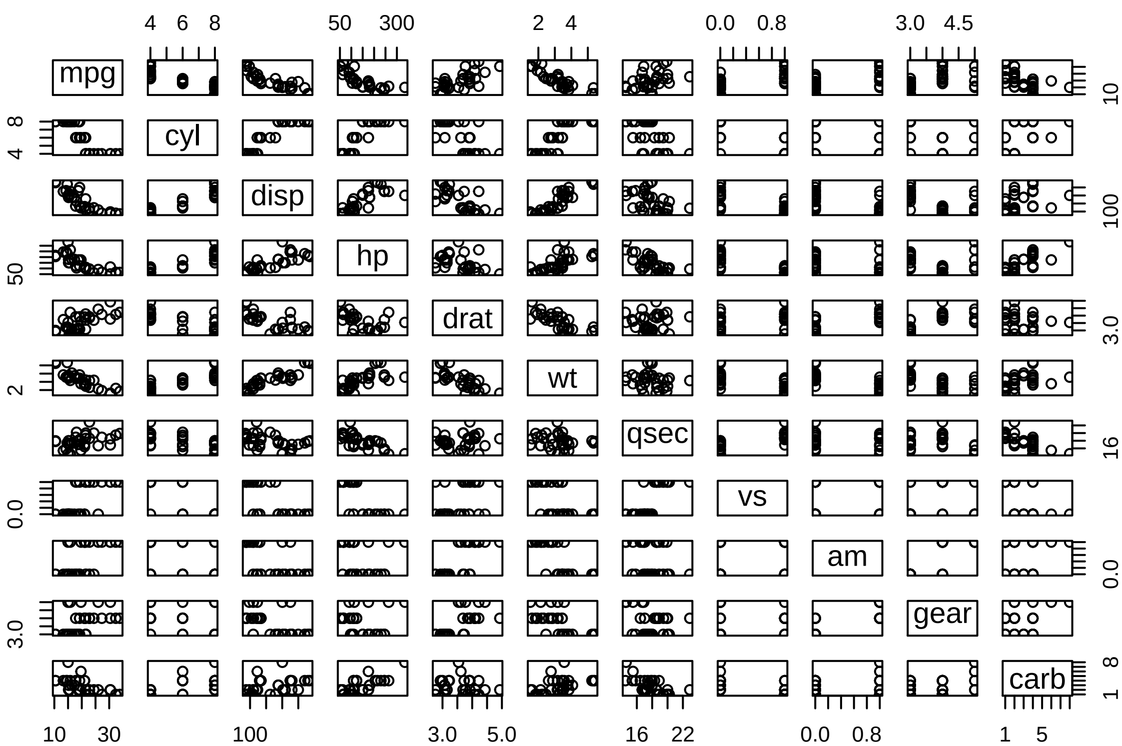

Figure 2. Pairs plot of mpg as related to all variables.

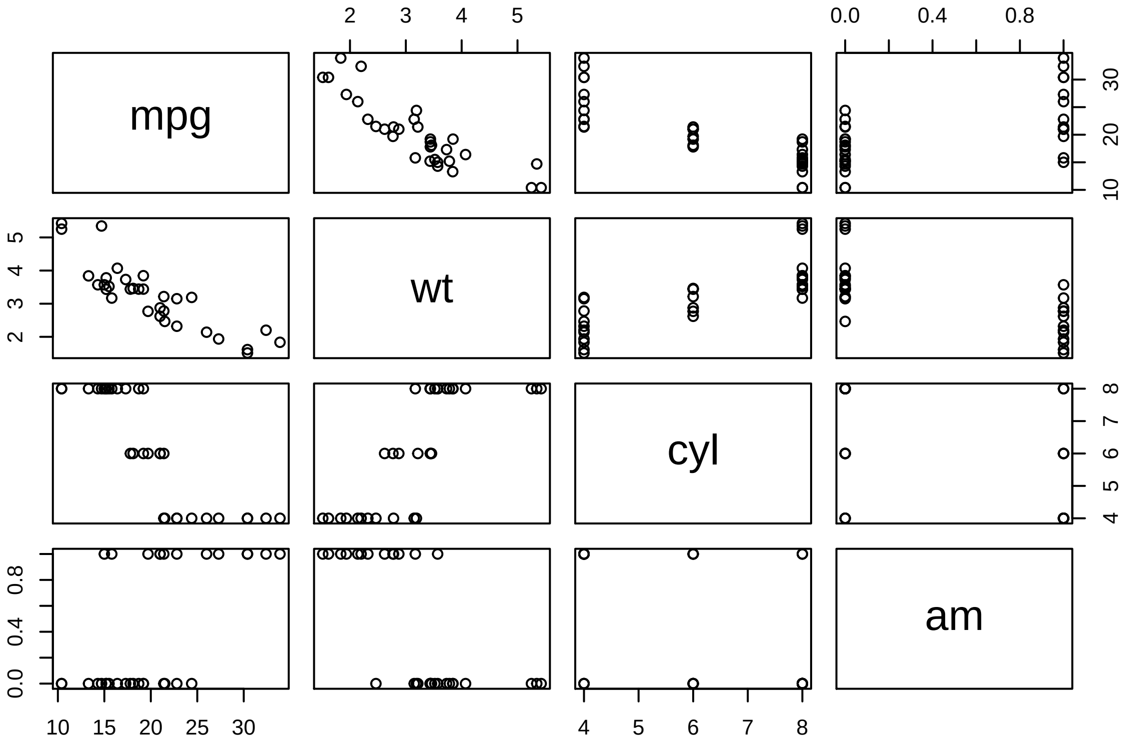

Figure 3. Pairs plot of mpg as related to the `fit.final` variables.

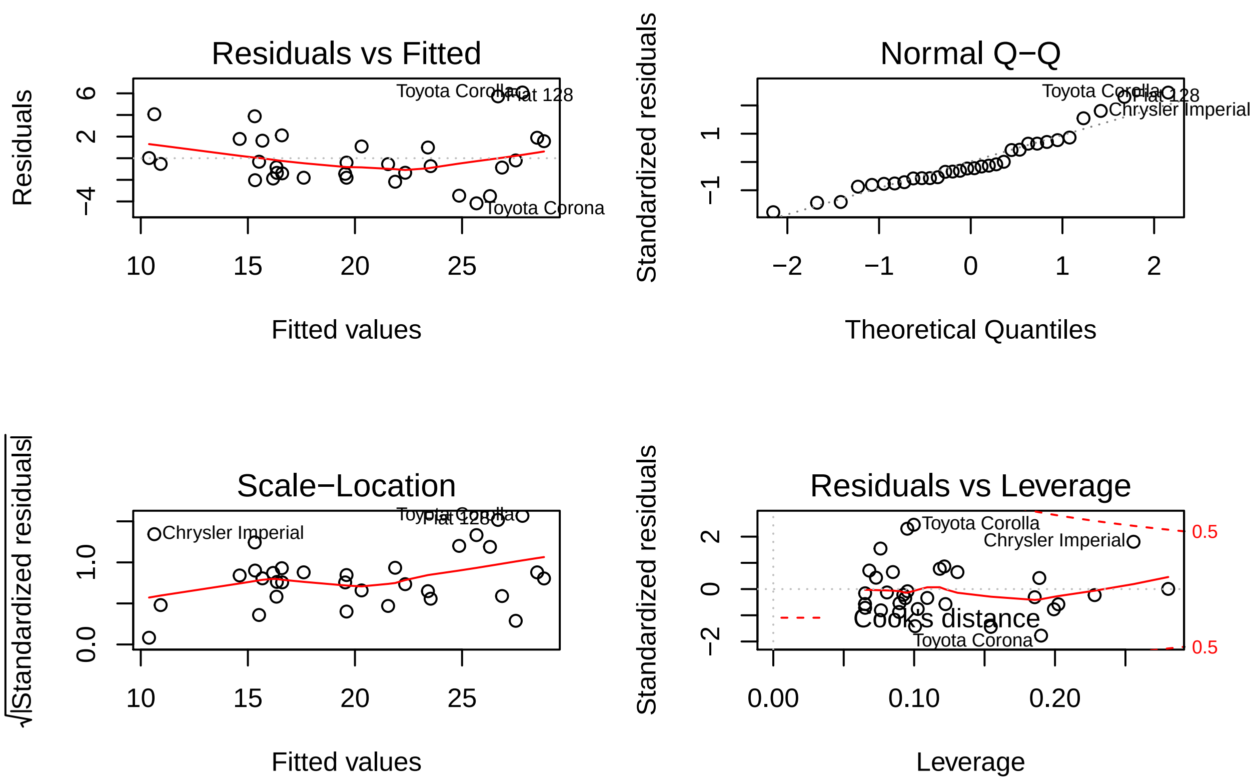

Figure 4. Residual diagnostics of `fit.final`.