Motor Trend - Transmission type does not affect MPG

Icaro C. Bombonato - May 2015

Summary

In this analysis, I explored the relationship between a set of variables and miles per gallon (MPG) (outcome). My goal was to answer these two questions:

- “Is an automatic or manual transmission better for MPG”
- “Quantify the MPG difference between automatic and manual transmissions”

The conclusion is that Transmission Type (am) does not influence Miles per Galon (mpg) at a significant level. What I found, was a cofounder in Weigth, that if ignored, can lead to a miss interpretation that Transmission Type has significant influence on Miles per Galon.

Analysis

First I make a scatterplotMatrix using the gclus package to look at the data, their correlations and their linear tendencies as **figure 1** shows.

After that, I change some variables to factor, am, cyl and vs and than I made a linear model of Miles per Galon (mpg) ~ TransmissionType (am) and take a look at the coefficients

```
coef(summary(lm(mpg ~ am, data = mtcars)))
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## ammanual     7.244939   1.764422  4.106127 2.850207e-04
```

Looking at the summary of that linear model, we can see that for automatic transmission, the coefficient is 17.147 mpg and that there is and estimated change of ~7.245 in mpg from Automatic to Manual transmission.
We also see that the diference is significant, since the p-value is very low: 0.000285.
So, looking ONLY for am and mpg, we can reject the null hipotesis that automatic and manual transmission has the same effect on mpg.

Now, we will try to prove it, using other models and variables from the dataset.

We will start with a model containing all variables, and then we will make it better.

```
lm(mpg ~ ., data = mtcars)
```

Looking at the summary of this model, there are no one significant value, p-value < 0.05. Which suggest that we can have some problem at the data. After performing a Bonferonni outlier test with the function outlierTest() in the car package, I removed two significant outliers from the dataset, “Chrysler Imperial” and “Fiat 128”, and then I fit the model again and take a look at the new summary:

```
# All other variables are ommited because they are not significant
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.56031   13.49486   1.524  0.14499
hp          -0.05309    0.02257  -2.352  0.03026 *
wt          -5.78433    1.64919  -3.507  0.00252 **
ammanual     1.08859    1.73087   0.629  0.53730
```

Surprisily, we can see that only Weight (wt) and Horsepower (hp) are significant to the model.
So, lets build a new model excluding all variables that are not important. We will keep Transmition, since it was the main variable for this analysis. And take a look at the coefficients again:

```
coef(summary(lm(mpg ~ wt + hp + am, data = mtcarsNew)))
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 36.3480274 2.43853756 14.9056664 2.986910e-14
## wt          -3.8793074 0.84371017 -4.5979147 9.715592e-05
## hp          -0.0305883 0.00843039 -3.6283376 1.222591e-03
## ammanual     0.8333006 1.22786055  0.6786606 5.033495e-01
```

Again, we see that Transmission is not significant for the model. Now, lets do a backward Akaike Information Criterion (AIC) to get the best model using our formula mpg ~ wt + hp + am.

```
stepAIC(fit02, direction="backward")
# Start:  AIC=50.07
# mpg ~ wt + hp + am
# Step:  AIC=48.6
# mpg ~ wt + hp
# Call:
# lm(formula = mpg ~ wt + hp, data = mtcarsNew)
```

Based on AIC value (lower is better) the correct model to call is **mpg ~ wt + hp** without the am variable.

After that, I investigate why transmission type alone seens to influence the mpg, but with other variables in the mix, it does not. And I found a cofounder variable, **Weight (wt)**. Automatic cars tends to have more weight, and weight causes variation in Miles per Galon. We show that in **figure 2**.

So, what we can infer taking a look at the summary of our best model?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.674426   1.443662  26.096  < 2e-16 ***
wt          -4.305250   0.558198  -7.713  2.7e-08 ***
hp          -0.028122   0.007531  -3.734  0.00089 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.144 on 27 degrees of freedom
Multiple R-squared:  0.8685,    Adjusted R-squared:  0.8588
F-statistic: 89.17 on 2 and 27 DF,  p-value: 1.273e-12
```

Conclusion

- Combined with others variables present in a car, transmission by itself does not have a signifcant impact in Miles per Galon
- For each unit increase in Weight (wt), we have a decrease of ~4.30 in Miles per Galon with 0.56 standard errors
- For each unit increase in Horse Power (hp), we have a decrease of 0.03 in Miles per Galon with 0.008 standard errors
- Our model cover ~86% of the variance explained in the data, has a significant p-value and also a significant F-statistic.

You can see the residual plot of our best model at **figure 3**.

Figure 1: Enhanced scatterplot matrix of mtcars data

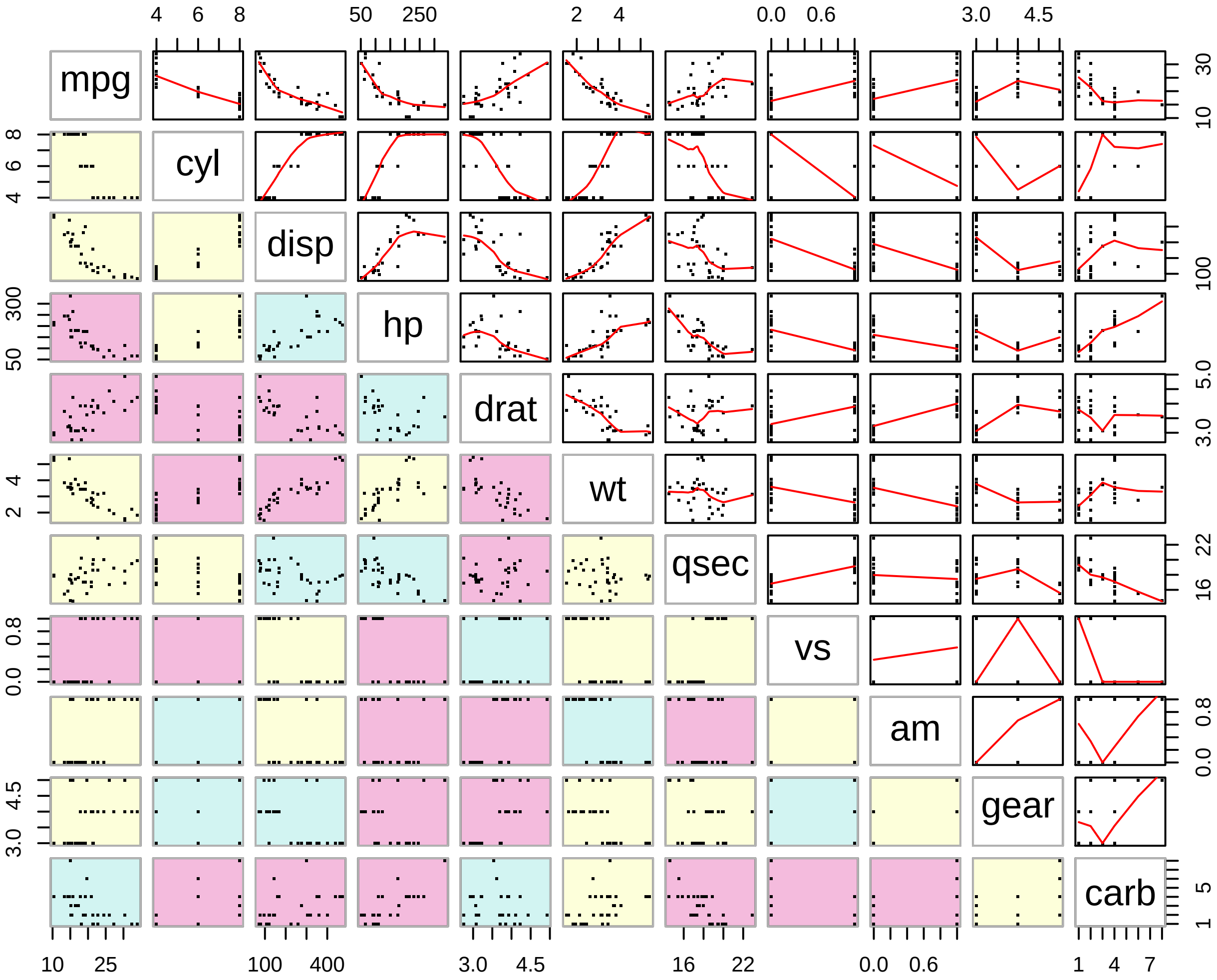


Figure 2: Transmission Type by Weight

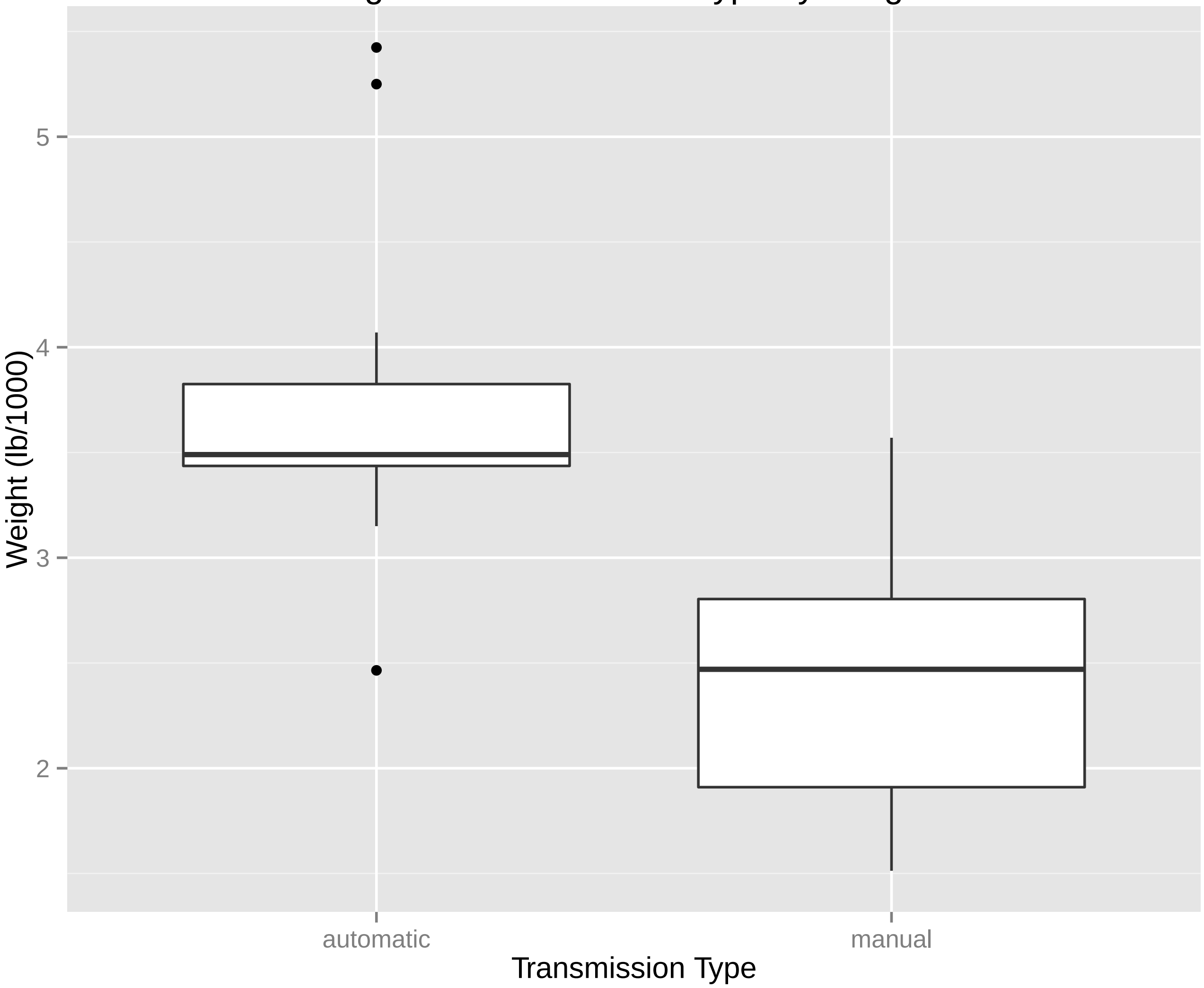
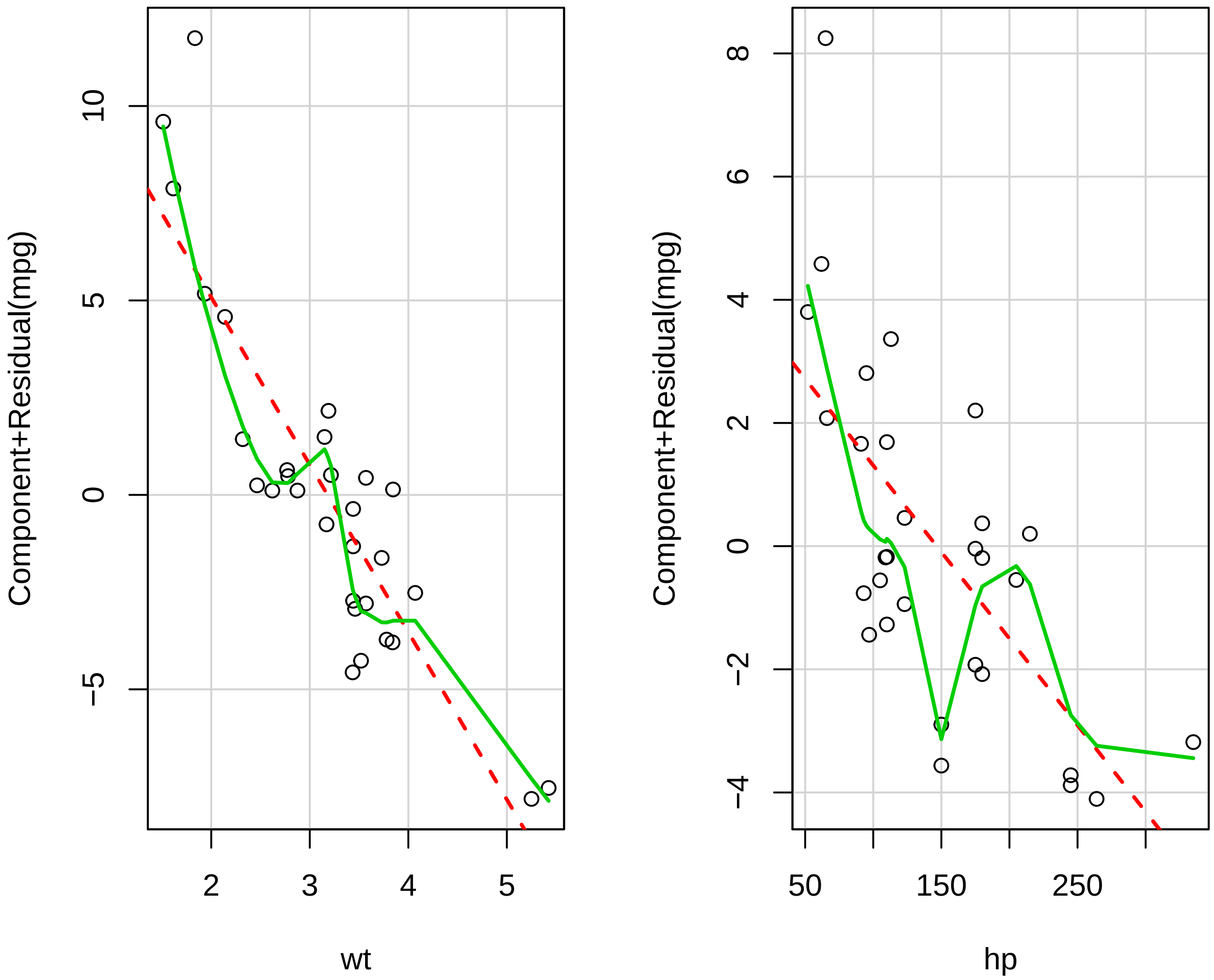


Figure 3: Component + Residuals of best model (mpg ~ wt + hp)


Component+Residual(mpg)
wt
Component+Residual(mpg)
hp