The Relationship of Transmission on Fuel Economy

zedleb

29 september 2016

Summary

This analysis of the mtcars data set explores the relationships between certain variables and the reported miles-per-gallon (MPG) of various vehicles. We are specifically interested in the following two questions:

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

Executive Conclusions

In general, manual transmission is considered more economical in terms of fuel consumption than automatic transmission. When Simple Linear Regression is employed with **transmission** as the *only* explanatory variable, fuel consumption is increased by about 7.25 mpg for automatic transmission. However, transmission alone explains only 34% of the variance in mpg. The addition of confounders **weight** and **cylinder number** (selected via ANOVA), into the final model explains 84% of the variance in mpg. Therefore, we can conclude that transmission alone is not sufficient to explain the difference in fuel economy of vehicles - the vehicle weight and number of cylinders must also be considered.

Exploratory Analysis

Best variables

See Appendix 1: Exploratory Box Plot of MPG against transmission for a quick comparison of the variable of interest - there does appear to be a difference in fuel economy for the two types of transmission.

However, since `mtcars` is a data frame with 11 variables. We will run analysis of variance to guage the best variables to include in a model.

```
bestVars <- aov(mpg ~ ., data = mtcars)
summary(bestVars)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## cyl          2  824.8   412.4  51.377 1.94e-07 ***
## disp         1   57.6    57.6   7.181   0.0171 *
## hp           1   18.5    18.5   2.305   0.1497
## drat         1   11.9    11.9   1.484   0.2419
## wt           1   55.8    55.8   6.950   0.0187 *
## qsec         1    1.5     1.5   0.190   0.6692
## vs           1    0.3     0.3   0.038   0.8488
## am           1   16.6    16.6   2.064   0.1714
## gear         2    5.0     2.5   0.313   0.7361
## carb         5   13.6     2.7   0.339   0.8814
## Residuals   15  120.4     8.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking for variables with p value $<$ **0.05** gives us:

- cyl
- disp
- wt

The relations between these variable and mpg can be seen in the pairs plot (Appendix 2: Exploratory pairs). Both `wt` and `disp` have **strong negative** linear relationships with `mpg`

- Correlation wt to mpg: -0.868
- Correlation disp to mpg: -0.848
- Mean MPG for Automatic vehicles: 17.147
- Mean MPG for Manual vehicles: 24.392

Regression Analysis

Our first model is the base model - using the variable `am` (transmission) as the explanatory variable.

```
base_mdl  <- lm(mpg ~ am, data = mtcars)
coef(base_mdl)
```

```
## (Intercept)         am1
##   17.147368    7.244939
```

This simple model would suggest that manual transmission increases `mpg` by 7.25.

However, the Adjusted R-squared value is only 0.3385, which means the base model explains about 34% of the variance of the MPG variable. This low value indicates that we need to add other variables to the model.

Keeping the best variables selected in the exploratory analysis in mind, we will begin multiple model building with the full data set using all the variables available and try to find the best model fit. For this we use the stepwise method - eliminating variables by both backward and forward selection methods by the AIC (Aikake Information Criterion) algorithm.

We will then compare the best model with the base model using anova. The full print out of results is suppressed.

```
full_mdl   <- lm(mpg ~ ., data = mtcars)
best_mdl <- step(full_mdl, direction = "both")
```

```
coef(best_mdl)
```

```
## (Intercept)        cyl6        cyl8          hp          wt         am1
## 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

Now the effect of manual transmission `am1` is to raise fuel economy by only 1.8 mpg (if all other variables are held constant) - but the p-value is too high to be significant.

The Adjusted R-squared value is 0.8401, which means about 84% of the variance is now explained by the selected variables in the best model.

We now compare the two models to see if adding the confounder variables make any difference. Our null hypothesis is that they do not contribute to the accuracy of the model.

```
anova(base_mdl, best_mdl)[2,6]
```

```
## [1] 1.688435e-08
```

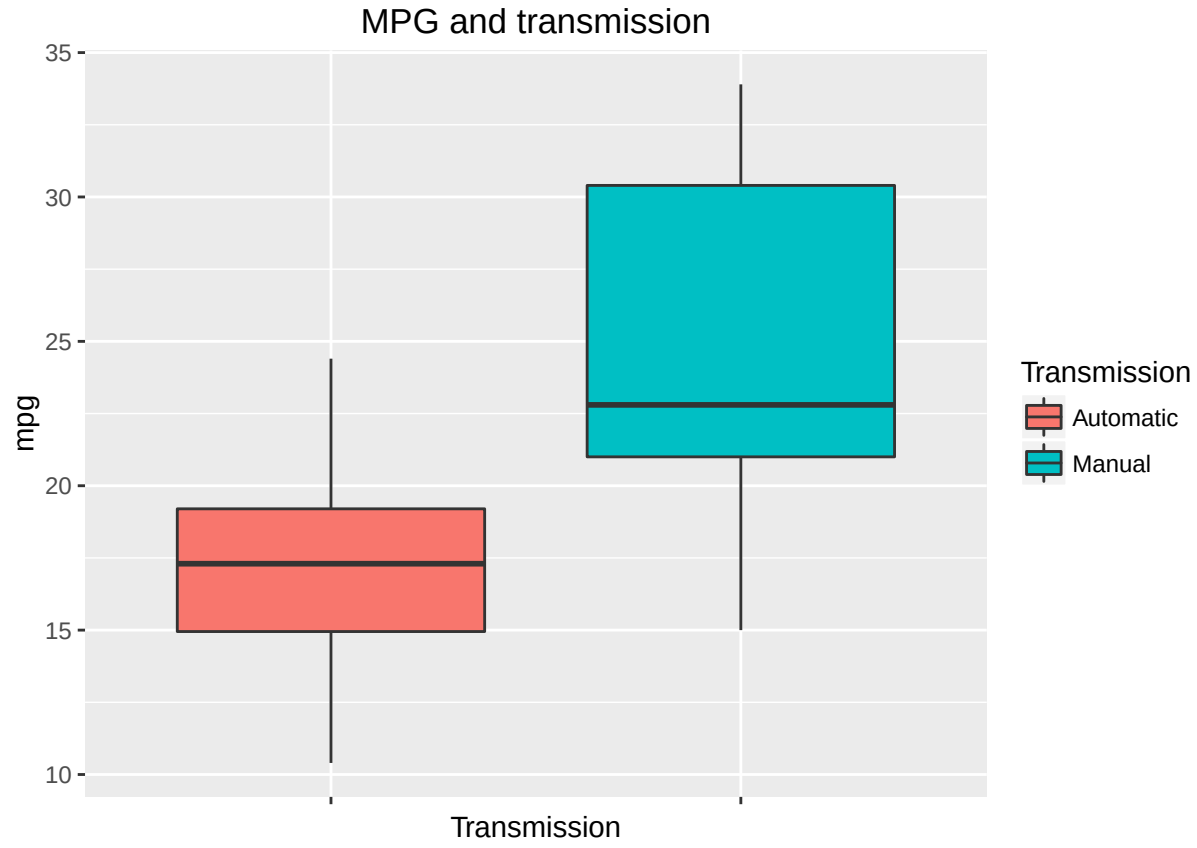
With the highly significant p-value we can **reject null hypothesis** that the confounder variables `cyl`, `wt` and `hp` do not contribute to the accuracy of the model.

Appendices

Full markdown code available on Github
https://github.com/ZedLeb/RegressionModelling

Appendix 1: Exploratory box

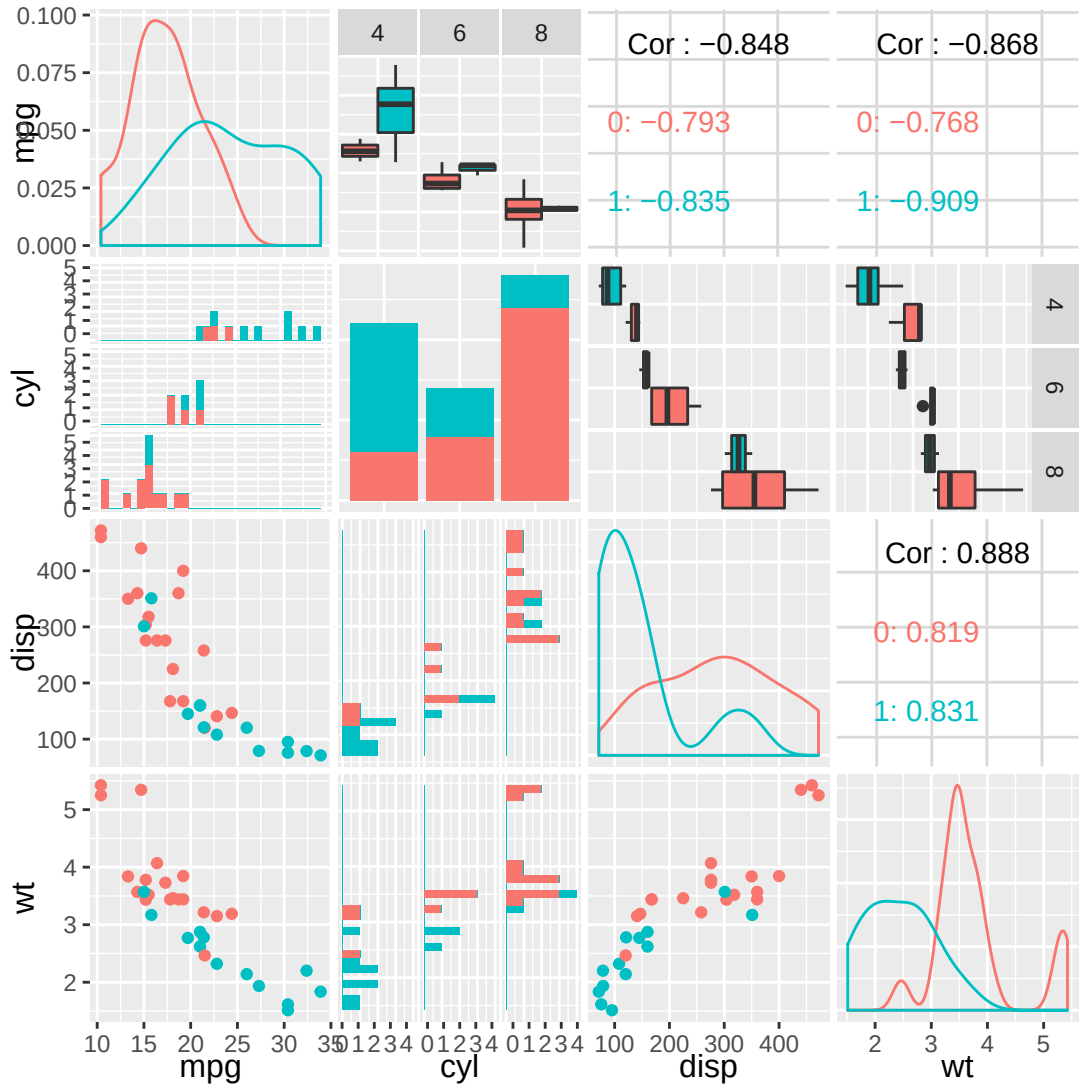
```
p <- ggplot(mtcars, aes(x=am, y=mpg, fill = am)) +
  geom_boxplot() +
  labs(x = "Transmission", title = "MPG and transmission") +
  scale_x_discrete(labels=element_blank()) +
  scale_fill_discrete(name="Transmission",
                      breaks=c(0,1),
                      labels=c("Automatic","Manual"))

p
```



Appendix 2: Exploratory pairs

NB Legends avoided for clarity - but automatic transmission is **blue** and manual is **red**

```
library(GGally)
ggpairs(mtcars, mapping = aes(color = am),
                columns = c("mpg", "cyl", "disp", "wt"),
                  #upper = "blank",
                  #diag = NULL,
              legends=F)
```



Appendix 3: Residual plots

The residuals appear randomly distributed. The normal Q-Q plot indicates the model meets normal distribution. Scale-Location graph shows constant variance.

```
par(mfrow=c(2,2))
plot(best_mdl,pch=16)
```