Motor Trend: The Effect Of Transmission Type On Fuel Economy

Luis Sandino

Created on 2016-04-12 10:35:22 using R version 3.2.4

Executive Summary

This report studies the effect of transmission type (automatic or manual) on fuel economy in Miles Per Gallon (MPG). More precisely, the MPG difference between automatic and manual transmissions is quantified with single and multi-variable linear regression models using the `mtcars` data set available in R library `datasets`, in order to determine which type of transmission is better for MPG.

The analysis showed that **no significant difference in MPG can be satated between the automatic and manual transmission** with the given data. However, it seems that car weight and 6 cylinder engines have greater impact on MPG. With 95% confidence, a **1000lb** increase in weight results in a **0.68 to 4.32 decrease** in MPG, and a change from **4** cylinders to **6** results in a **0.14 to 5.92** decrease in MPG.

Exploratory Analysis

Load data and examine variables:

```r
data(mtcars)
head(mtcars,3)
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

The `mtcars` data set has 11 variables for 32 different car models (observations). These variables are: **mpg**: Miles/(US) gallon; **cyl**: Number of cylinders; **disp**: Displacement (cu.in.); **hp**: Gross horsepower; **drat**: Rear axle ratio; **wt**: Weight (1000 lbs); **qsec**: 1/4 mile time; **vs**: Engine (0 = V, 1 = straight); **am**: Transmission (0 = automatic, 1 = manual); **gear**: Number of forward gears; **carb**: Number of carburetors. Variables **`cyl`**, **`vs`**, **`gear`**, **`carb`** and **`am`** are treated as factor variables in the analysis. Fig 1 (See Appendix) shows a pairs plot for this data set.

Single variable linear regression

Firstly, the relationship between MPG and transmission type is studied with a single variable linear regression:

```r
fit1 <- lm(mpg ~ am,data=mtcars)
summary1 <- summary(fit1)
print(summary1$coefficients)
```

```
##              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
```

The R^2 value obtained with the model above is **0.3598**, which means that only the **36%** of total variability is explained by this simple regression model.

Multivariable linear regression

In order to improve the model, several variables should be included in the linear regression. The candidates are those variables with high correlation with MPG. As can be seen in Fig 1 (See Appendix), these are **wt**, **disp** and **hp** with correlation values -0.8677, -0.8476 and -0.7762, respectively. Regarding factor variables, it seems reasonable to also include **cyl** in the model, as the number of cylinders is expected to be directly related to fuel consumption. After trying several combinations, the best fit is obtained with **am**, **cyl**, **wt** and **hp**:

```
fit2 <- lm(mpg ~ am + cyl + wt + hp,data=mtcars)
summary2 <- summary(fit2)
print(summary2$coefficients)
```

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

The adjusted R^2 value obtained with the model above is **0.8401**, which means that **84%** of total variability is explained by this model.

Residuals and Diagnostics

As can be seen in Fig 2 (See Appendix), the Residuals vs Fitted and the Scale-Location plots seem to be randomly scattered, which indicates there are no patterns for missing variables and heteroskedasticity. The Normal Q-Q plot indicates the residuals are normally distributed. Finally, the Residuals vs Leverage plot tells that data from *Toyota Corolla* and *Chrysler Imperial* have substantial influence on the regression model.

Inference and Conclusion

Confidence intervals for the model parameters are calculated as follows:

```
sum <- summary2$coefficients
round(sum[i,1] + c(-1,1)*qt(.975,df=fit2$df)*sum[i,2],2)
```

Table 1: Confidence intervals for model parameters

Variable	Conf. Int. low	Conf. Int. high
mean MPG (4 cyl, Auto)	28.35	39.06
am Manual	-1.06	4.68
cyl 6	-5.92	-0.14
cyl 8	-6.86	2.53
wt	-4.32	-0.68
hp	-0.06	0.00

As can be seen, with 95% confidence, the mean MPG for the reference (4 cylinder, automatic transmission) is in the interval [**28.35 39.06**]. Regarding p-values (look multi-variable linear regression model), they are less than 0.05 for `cyl6`, `wt` and `hp` only. Therefore, in this cases the null hypothesis (difference in the mean equal to 0) is rejected . There is strong evidence in favor of the alternative hypothesis: *true difference in means is not equal to 0.* On the contrary, p-value is grater than 0.05 for `cyl8` and `am`, the principal variable of interest in this analysis. Thus, in this two cases null hypothesis rejection fails. This study concludes that *no difference in MPG can be satated between the automatic and manual transmission* with the given data. Nevertheless, it seems variables `wt` and `cyl6` have greater impact on MPG. With 95% confidence, a **1000lb** increase in weight results in a **0.68 to 4.32 decrease** in MPG, and a change from **4** cylinders to **6** results in a **0.14 to 5.92** decrease in MPG.

Appendix

Fig. 1 Pairs plot for `mtcars` data set. Diagonal: *Density plot* for Continuous vs. Continuous and *Bar plot* for Continuous vs. Discrete. Upper: *Scatter plot* with regression line for Continuous vs. Continuous and *Box plot* for Continuous vs. Discrete. Lower: *Correlation* for Continuous vs. Continuous only.

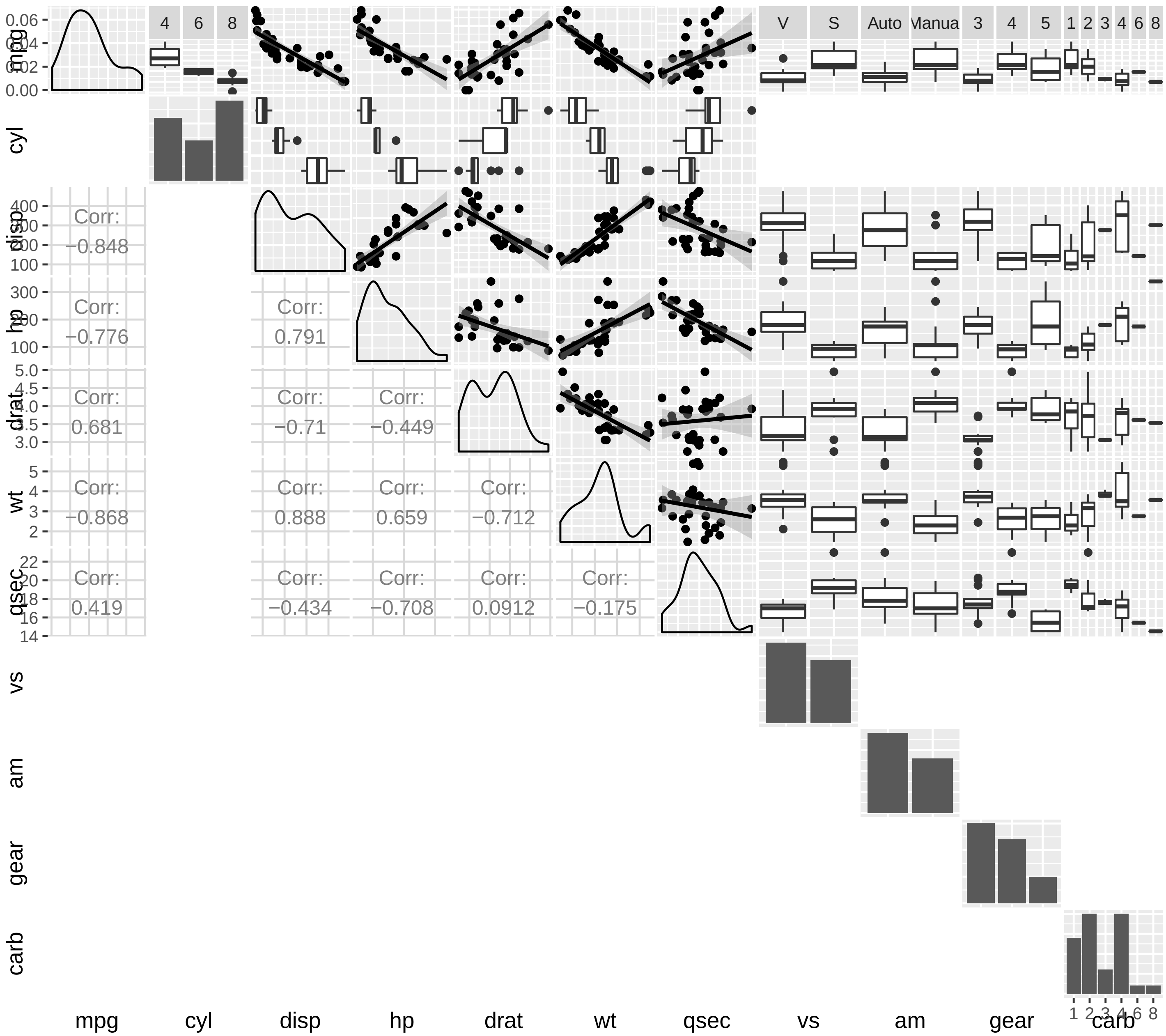

Fig. 2 Residuals and Diagnostics plot