Car Fuel Efficiency Analysis

Yasser Gonzalez — http://yassergonzalez.com

July 2015

Executive Summary

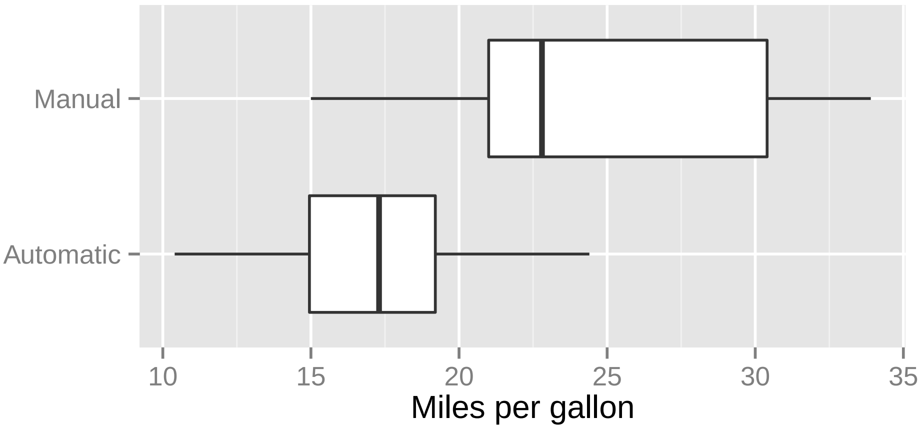
This report presents the results of an analysis of the fuel efficiency of a group of cars. Specifically, the study addresses the question of whether automatic or manual cars are more efficient, and attempts to quantify the difference between the two. The fuel efficiency is measured in miles per gallon (`mpg`). In addition to `mpg` and the transmission type (`am`), the information available about the cars include the number of cylinders (`cyl`), horsepower (`hp`), weight (`wt`), engine type (`vs`), among other relevant characteristics. The study shows that the transmission type alone does not provide enough information to make a well-supported claim about the fuel efficiency of a car. The remainder of the report provides the details of the statistical analysis that was carried out to arrive at that conclusion.

Exploratory Data Analysis

The data set in question was loaded into R and tidied (see the R code in the Appendix). The following table shows a number of rows selected from the resulting data set—it contains 32 observations in total.

car	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	V	Manual	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	V	Manual	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	Straight	Manual	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	Straight	Automatic	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	V	Automatic	3	2

Given that we are interested in comparing the fuel efficiency of cars with automatic or manual transmissions, we begin the analysis with a boxplot of the associated `mpg` distributions.



The plot above suggests that, among the group of studied cars, manual transmission cars have better fuel efficiency (as in higher `mpg` values) than cars with automatic transmission. This claim is studied in more detail with a regression analysis in the next section.

Regression Models Analysis

Motivated by what was observed in the boxplot, we construct first a linear model for explaining the `mpg` values in terms of the `am` variable (i.e. the transmission type) coded as a dummy variable as follows:

```r
only_am <- lm(mpg ~ am, data = mtcars)
```

The fitted coefficients table is shown in the Appendix. The coefficient corresponding to the `am` dummy variable associated with manual transmission in effect indicates that manual transmission cars have better fuel efficiency than automatic cars—roughly 7.2 miles per gallon better on average. The associated p-value also suggests that this result is statistically significant at the 1% level. However, the R^2 obtained is 0.36, indicating that the model leaves most of the variance of `mpg` unexplained. This is not a surprising result, considering that this model predicts the same `mpg` value for all manual transmission cars; and similarly, the same `mpg` value for all cars with automatic transmission. These results suggest that the model denoted as `only_am` is a poor fit for explaining the cars' fuel efficiency.

We construct another linear model next. It considers all the other variables in the data set as predictors (in addition to `am`) in order to understand if there are other factors that influence the `mpg` values. The second linear model can be fitted as follows:

```r
all_vars <- lm(mpg ~ am + ., data = select(mtcars, -car))
```

In this case, the obtained R^2 score is approximately 0.87, which means that the model explains roughly 87% of the variance of the `mpg` values. A likelihood ratio tests for nested models comparing the models `only_am` and `all_vars` indicates a statistically significant reduction at the 1% level of the residual sum of squares by including the other variables (see the details in the Appendix). Also, a plot included in the Appendix shows an appropriate distribution of the residuals.

These results evidence that the `all_vars` linear model is a better fit to explain the `mpg` values than the `only_am` model—i.e. that other variables influence the fuel efficiency of the cars besides the transmission type. In fact, the coefficient corresponding to the `am` dummy variable associated with manual transmission in the `all_vars` model is approximately 2.52 (i.e. the `mpg` difference between manual and automatic transmissions) and it is not statistically significantly different from zero at the 1% level.

Summarizing, the transmission type does not seem to have a significant impact on the fuel efficiency of the cars after having accounted for the effect of other variables such as the number of cylinders and the weight of the car.

Appendix

This appendix contains supplementary figures and the R code fragments not shown in the main sections to ease the reproducibility of the results.

Loading the necessary R packages

```r
library("knitr")
library("dplyr", warn.conflicts = FALSE)
library("ggplot2")
```

Loading and tidying the data set

```
library("datasets")
data("mtcars")

mtcars <- mtcars %>%
    add_rownames("car") %>%
    mutate(am = ifelse(am == 0, "Automatic", "Manual")) %>%
    mutate(vs = ifelse(vs == 0, "V", "Straight"))
```

Code to generate the `mpg` boxplot

```
ggplot(mtcars, aes(x = am, y = mpg)) +
    labs(x = NULL, y = "Miles per gallon") +
    geom_boxplot() +
    coord_flip()
```

Coefficients table for the `only_am` model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147368	1.124602	15.247492	0.000000
amManual	7.244939	1.764422	4.106127	0.000285

Coefficients table for the `all_vars` model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.6211370	19.0284151	0.6632784	0.5143680
amManual	2.5202269	2.0566506	1.2254035	0.2339897
cyl	-0.1114405	1.0450234	-0.1066392	0.9160874
disp	0.0133352	0.0178575	0.7467585	0.4634887
hp	-0.0214821	0.0217686	-0.9868407	0.3349553
drat	0.7871110	1.6353731	0.4813036	0.6352779
wt	-3.7153039	1.8944143	-1.9611887	0.0632522
qsec	0.8210407	0.7308448	1.1234133	0.2739413
vsV	-0.3177628	2.1045086	-0.1509915	0.8814235
gear	0.6554130	1.4932600	0.4389142	0.6652064
carb	-0.1994193	0.8287525	-0.2406258	0.8121787

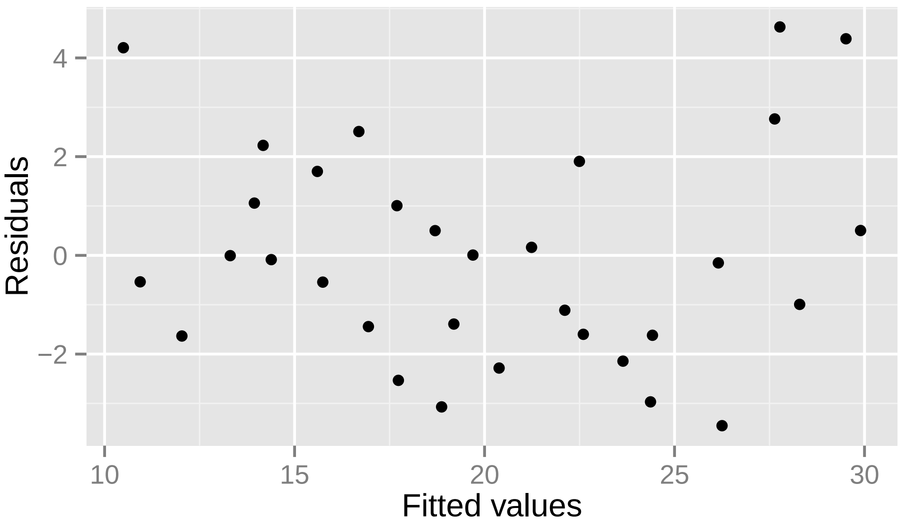
Comparison of the `only_am` and `all_vars` linear models

```r
anova(only_am, all_vars)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966				
21	147.4944	9	573.4022	9.071111	1.78e-05

Plot of the residuals of the `all_vars` linear model

```r
data <- data.frame(residuals = residuals(all_vars),
                   fitted_values = predict(all_vars))
ggplot(data, aes(x = fitted_values, y = residuals)) +
    labs(x = "Fitted values",  y = "Residuals") +
    geom_point()
```


Residuals
Fitted values