Regression Models - Course Project

Arturo Equihua

Saturday, May 23rd, 2015

Executive Summary

The Motor Trend magazine is evaluating the relationship between several variables and the mileage per gallon (mpg) in a sample of major car brands and models, in order to establish whether or not a manual vs. automatic transmission is better than the other. By taking a sample dataset of car model statistics, and doing regression analysis, it is found that, even though manual transmission cars perform better in fuel economy than automatic transmission cars (avg 7.2 mpg of difference), the mileage per gallon can be much better predicted by using a model with other two variables: Weight and Engine Displacement.

Source Data and Research Questions

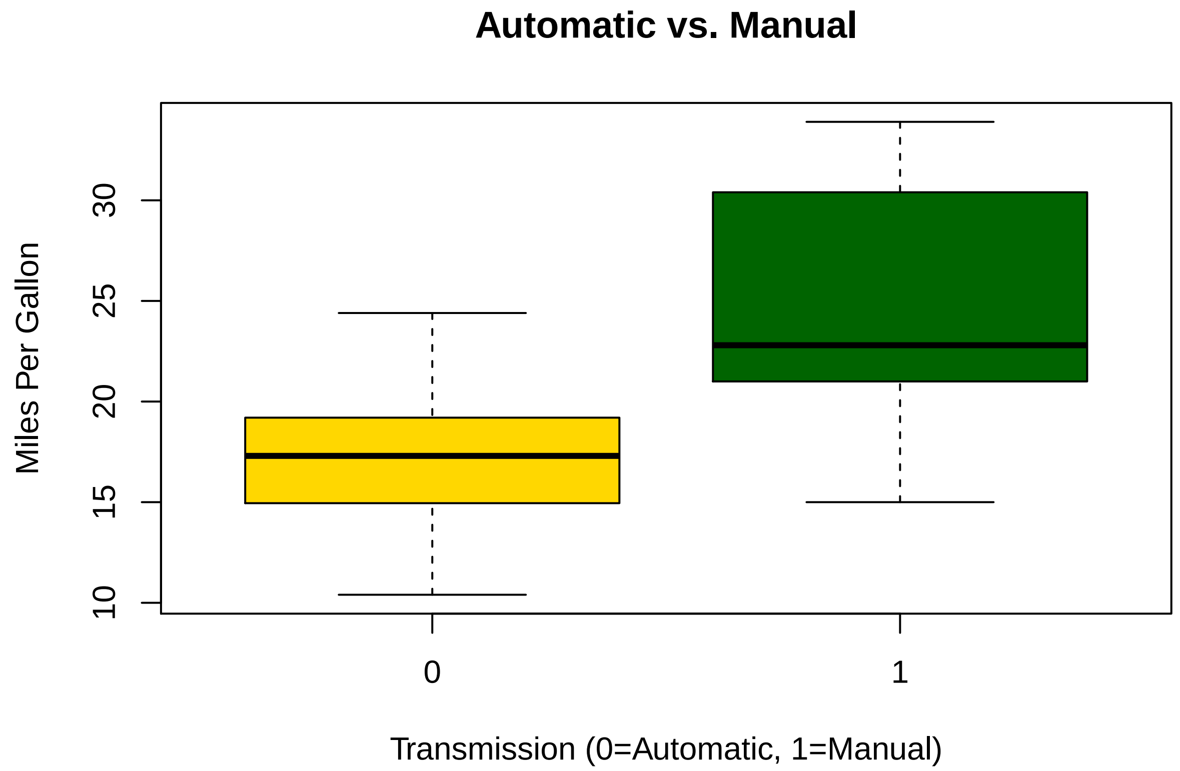
The data source is taken from the **mtcars** dataframe, that has the following variables: mpg = Miles/(US) gallon cyl = Number of cylinders disp = Displacement (cu.in.) hp = Gross horsepower drat = Rear axle ratio wt = Weight (lb/1000) qsec = 1/4 mile time vs = V/S am= Transmission (0 = automatic, 1 = manual) gear = Number of forward gears carb = Number of carburetors.

This document seeks to address the following questions:

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

Exploratory analysis

The following is a quick and dirty chart that tests the specific relationship between transmission (am) and mileage (mpg):



From the above there seems to be a tendency of automatic cars to have less mileage per gallon than manual cars.

In the **Appendix**, a pairs chart is shown to look at the other variables and discover potential strong relationships. From that chart there seems to be a strong relationship between mpg and other variables such as engine displacement (disp), HorsePower (hp) or weight (wt).

The following section seeks to identify those relationships in a more formal way.

Regression Analysis

Fit transmission vs. mileage per gallon

By applying some statistical inference, a 95% t-test to compare the average mpg for automatic vs. manual transmission (as two separate groups) is done below:

```
## [1] 15.29946 18.99528
## attr(,"conf.level")
## [1] 0.95
```

```
x2$conf.int
```

```
## [1] 20.66593 28.11869
## attr(,"conf.level")
## [1] 0.95
```

The calculation above suggests that, with 95% of confidence, automatic transmission cars range between 15.3 and 19.0 mpg, whereas manual transmission cars range between 20.67 and 28.12 mpg. That is, manual transmission cars clearly provide higher fuel economy.

To quantify that more formally, a simple regression model is constructed, with transmission type as predictor (coded as 1 or 0, where 0 is automatic and 1 is manual), and fuel economy as outcome (miles per gallon, or mpg).

```r
mtcarsfit1 <- lm(mpg ~ am, data = mtcars)
summary(mtcarsfit1)
```

```
## 
## Call:
## lm(formula = mpg ~ am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.3923 -3.0923 -0.2974  3.2439  9.5077 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385 
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The intercept and slope coefficients can be interpreted as follows: In average, automatic transmission achieves 17.147 mpg, whereas manual transmissions (that is, when the variable am = 1), provides additional 7.245 mpg (that is, 17.147 + 7.245 = 24.39 mpg). That is, our simple model could say that, in average, the manual transmission is 7.245 mpg better than automatic transmission.

However, it cannot be ignored that the correlation coefficient (36%) suggest that transmission type only accounts for about 36% of the variance in fuel consumption. That is, other predictor variables seem to be needed in order to obtain a more significant model for the mileage per gallon.

Fit multiple variables vs. mileage per gallon

Testing predictors with variance Analysis

The next step is to find additional variables that might account for the difference observed between manual and automatic transmissions. To that end, and as a first approach, it can be assumed that automatic transmissions may be found in cars that are heavier in weight, or that have greater engine displacement, so these two variables could be tested in terms of variance inflation. So the next step is to test the relationship between transmission and fuel economy by constructing two baseline regression models with weight or engine displacement as a predictor, and then an augmented model that adds automatic/manual transmission as a variable.

```r
mtcarsbasemodel1 <- lm(1/mpg ~ wt, data = mtcars)
mtcarsaugmodel1 <- lm(1/mpg ~ wt + am, data = mtcars)
anova(mtcarsbasemodel1, mtcarsaugmodel1)
```

```
## Analysis of Variance Table
##
## Model 1: 1/mpg ~ wt
## Model 2: 1/mpg ~ wt + am
##   Res.Df       RSS Df  Sum of Sq      F Pr(>F)
## 1     30 0.0017402
## 2     29 0.0016465  1 9.3699e-05 1.6504 0.2091
```

```
mtcarsbasemodel2 <- lm(1/mpg ~ disp, data = mtcars)
mtcarsaugmodel2 <- lm(1/mpg ~ disp + am, data = mtcars)
anova(mtcarsbasemodel2, mtcarsaugmodel2)
```

```
## Analysis of Variance Table
##
## Model 1: 1/mpg ~ disp
## Model 2: 1/mpg ~ disp + am
##   Res.Df       RSS Df  Sum of Sq      F Pr(>F)
## 1     30 0.0018892
## 2     29 0.0018842  1 4.9936e-06 0.0769 0.7836
```

Each of these variance analyses suggests that whether a car has automatic or manual transmission does not affect fuel consumption AFTER one controls for either weight or engine displacement. Specifically, if weight is set as a predictor as a preliminary model, and then create a second model that has both weight of car and automatic/manual transmission as a second predictor. The second model does not improve (F = 1.65, df=1,30, p > .20) A similar analysis for engine displacement and transmission shows that adding the transmission variable does not improve the model (F < 1).

Finding a comprehensive model

The final question to address is what "best" mpg prediction model could be built with the existing variables in the mtcars dataset. Since some variables can be assumed related to each other (e.g. cylinders and engine displacement so only engine displacement is chosen. Similarly, Horsepower is likely to be a function of engine displacement, weight, and number of carburetors, so number of carburetors is added but not horsepower. A similar argument can be made to reject quarter of mile, qsec.). So it is only needed to test whether there is evidence to include both weight and displacement in the model, with a test of model difference in a hierarchical sequence:

```
mtcarsbasemodel3 <- lm(1/mpg ~ wt, data = mtcars)
mtcarsaugmodel3 <- lm(1/mpg ~ wt + disp, data = mtcars)
anova(mtcarsbasemodel3, mtcarsaugmodel3)
```

```
## Analysis of Variance Table
##
## Model 1: 1/mpg ~ wt
## Model 2: 1/mpg ~ wt + disp
##   Res.Df       RSS Df  Sum of Sq      F  Pr(>F)
## 1     30 0.0017402
## 2     29 0.0014226  1 0.00031754 6.4731 0.01654 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This ANOVA result confirms that weight and displacement can both be predictors, as they each contribute uniquely to the model.

Therefore, having weight and engine displacement as confirmed predictors, the rest of variables is added to the ANOVA, to identify further variables to add to the model:

```
mtcarsbasemodel4 <- lm(1/mpg ~ wt + disp, data = mtcars)
mtcarsaugmodel4 <- lm(1/mpg ~ wt + disp + drat + vs + gear + carb, data = mtcars)
anova(mtcarsbasemodel4, mtcarsaugmodel4)
```

```
## Analysis of Variance Table
##
## Model 1: 1/mpg ~ wt + disp
## Model 2: 1/mpg ~ wt + disp + drat + vs + gear + carb
##   Res.Df       RSS Df  Sum of Sq     F Pr(>F)
## 1     29 0.0014226
## 2     25 0.0011914  4 0.00023122 1.213 0.3302
```

The augmented model does not improve the degree of fit (F = 1.21, df = 4,25, p > .33). It is then accepted that the baseline model with weight and displacement as predictors is sufficient. The following analysis shows that:

```
summary(mtcarsbasemodel4)
```

```
##
## Call:
## lm(formula = 1/mpg ~ wt + disp, data = mtcars)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.017062 -0.003494  0.001582  0.004440  0.010852
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.353e-02  5.198e-03   2.604  0.01439 *
## wt          8.622e-03  2.796e-03   3.084  0.00445 **
## disp        5.615e-05  2.207e-05   2.544  0.01654 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.007004 on 29 degrees of freedom
## Multiple R-squared:  0.8299, Adjusted R-squared:  0.8181
## F-statistic: 70.73 on 2 and 29 DF,  p-value: 7.013e-12
```

The model with weight and engine displacement explains 83% of the variance, and each of the coefficients for weight and engine displacement is statistically significantly greater than 0.

Some diagnostics of this model are reported in the Appendix. As shown there, the standardized residuals do not exhibit a cumulative distribution that matchesthe Normal distribution (there are some points clearly lower than the diagonal line). There also appear to be potential outliers in the various residual plots. However, the dataset is very small (only 32 cases), and it does not seem justified to drop any of the cases for this analysis.

Appendix

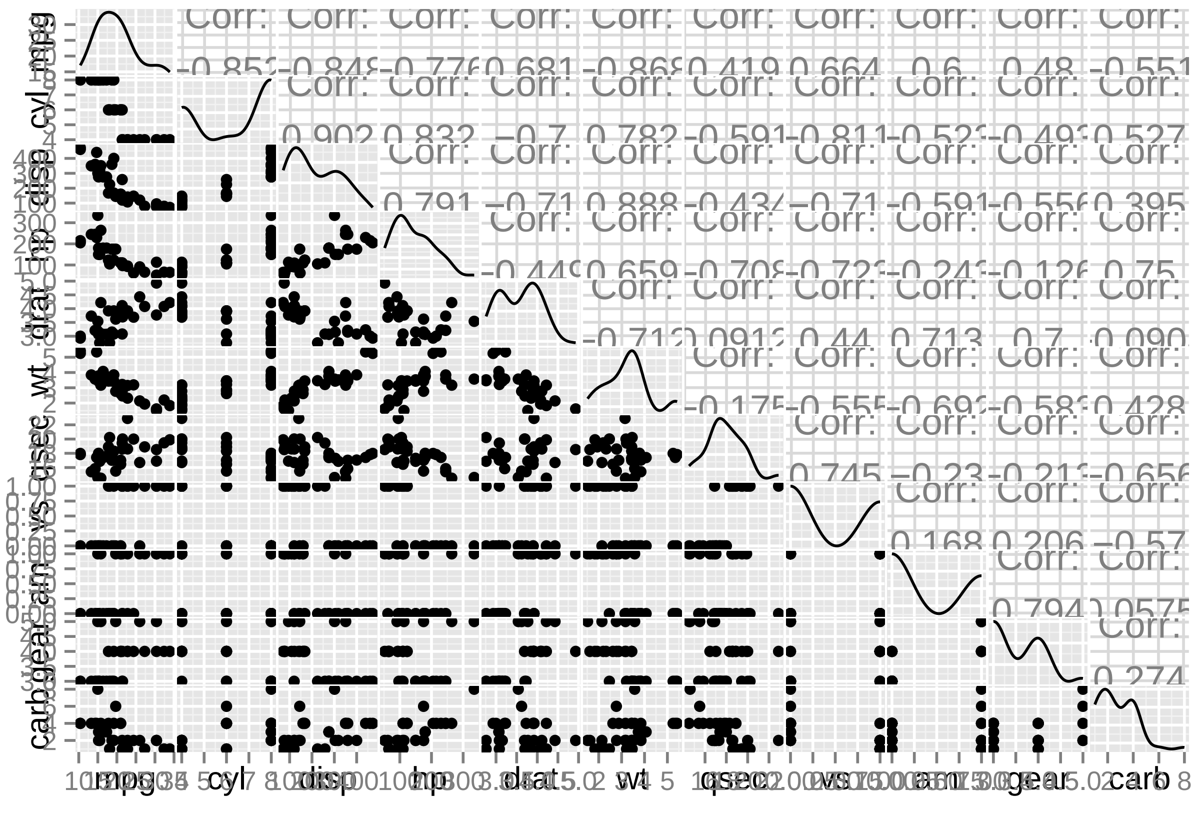
Model analysis

The following is a pair chart that looks for potential relationships between all the variables collected:

```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
##
## The following object is masked from 'mtcars':
##
##     mpg
```

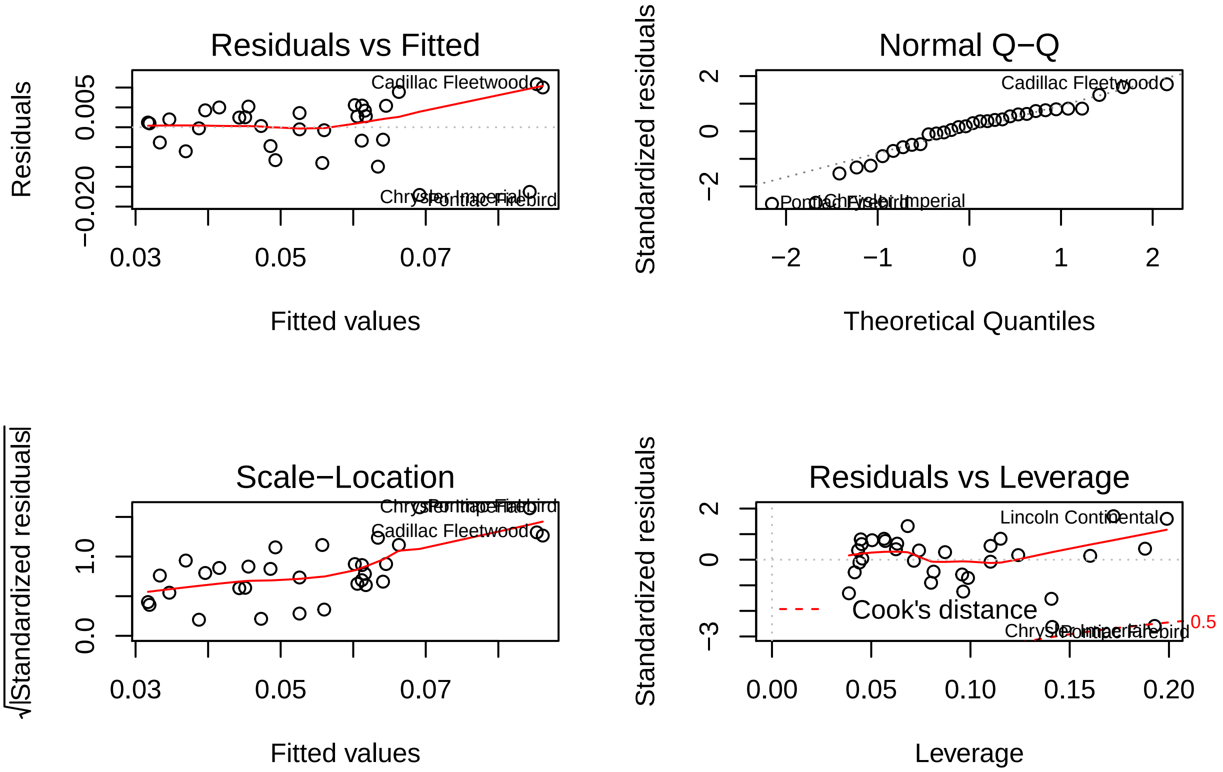
```r
library(GGally)
g2 <- ggpairs(mtcars, diag=list(continuous="density", discrete="bar"), axisLabels="show")
print(g2)
```



Model analysis

The mpg regression model with weight and engine displacement found in the main report is plotted below, to graphically assess the quality of it:

```r
par(mfrow = c(2, 2))
plot(mtcarsbasemodel4)
```



Residuals vs Fitted
Residuals
0.005
0.000
−0.020
0.03
0.05
0.07
Fitted values
Cadillac Fleetwood
Chrysler Imperial
Pontiac Firebird
Normal Q−Q
Standardized residuals
2
0
−2
−1
1
Theoretical Quantiles
Cadillac Fleetwood
Pontiac Firebird
Chrysler Imperial
Scale−Location
√|Standardized residuals|
1.0
0.0
Fitted values
Chrysler Imperial
Pontiac Firebird
Cadillac Fleetwood
Residuals vs Leverage
Standardized residuals
−3
0.00
0.10
0.15
0.20
Leverage
Cook's distance
0.5
Lincoln Continental
Chrysler Imperial
Pontiac Firebird