Regression Models - Course Project

Arturo Equihua

Saturday, May 23rd, 2015

Executive Summary

The Motor Trend magazine is evaluating the relationship between transmission type (automatic and manual) vs. the mileage per gallon (mpg). From a 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 explained by using a model with other two predictorvariables: **Weight** and **Engine Displacement**.

Exploratory analysis

The data source is taken from the **mtcars** dataframe. A full description of it can be obtained by typing "?mtcars" in the R Studio session. A quick and dirty boxplot chart, shown in the **Appendix**, tests the specific relationship between transmission (am) and mileage (mpg), finding a tendency of automatic cars to have less mileage per gallon than manual cars Also, 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).

Regression Analysis

Transmission vs. mileage per gallon

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).

```
##
## 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). Just consider that the correlation coefficient is only 36%, which suggests there might be other variables that have stronger relationship with mpg than the transmission type.

Fit multiple variables vs. mileage per gallon

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). 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:

```
## 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. Once 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:

```
## 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:

```
##
## 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.

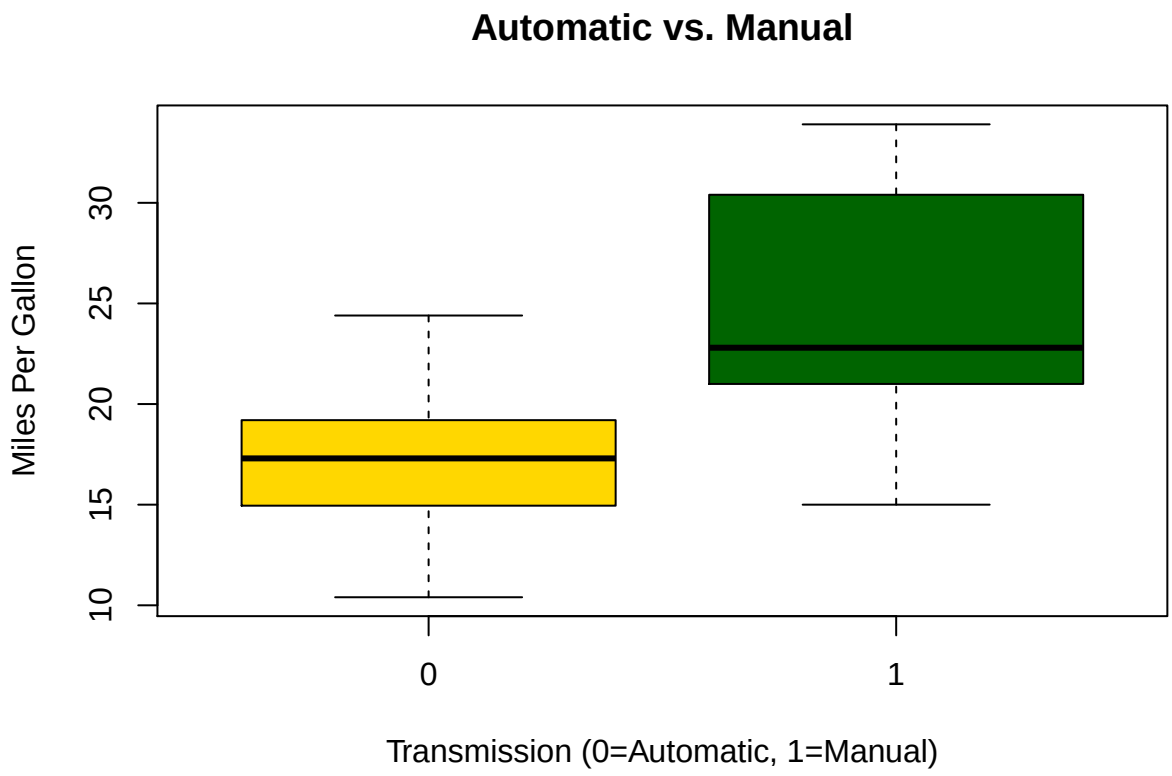
Some diagnostics of this model are reported in the Appendix. As shown there, the standardized residuals do not exhibit a cumulative distribution that matches the Normal distribution (there are some points clearly lower than the diagonal line). There also appear to be potential outliers in the various 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

Boxplot Automatic vs. Manual

The following is comparison of mileage per gallon between the manual vs. automatic transmission cars in the sample:

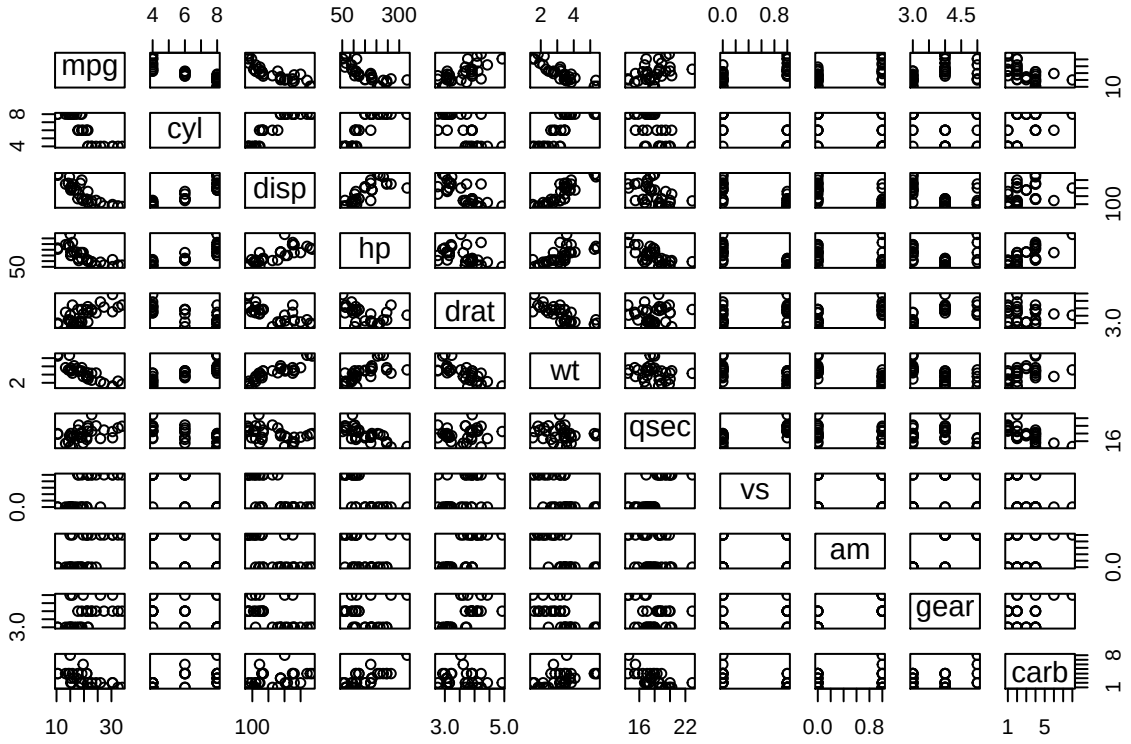
```
boxplot(mpg~factor(am),data=mtcars, main="Automatic vs. Manual",
  xlab="Transmission (0=Automatic, 1=Manual)", ylab="Miles Per Gallon",
  col=c("gold","darkgreen"))
```



Pair Chart

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

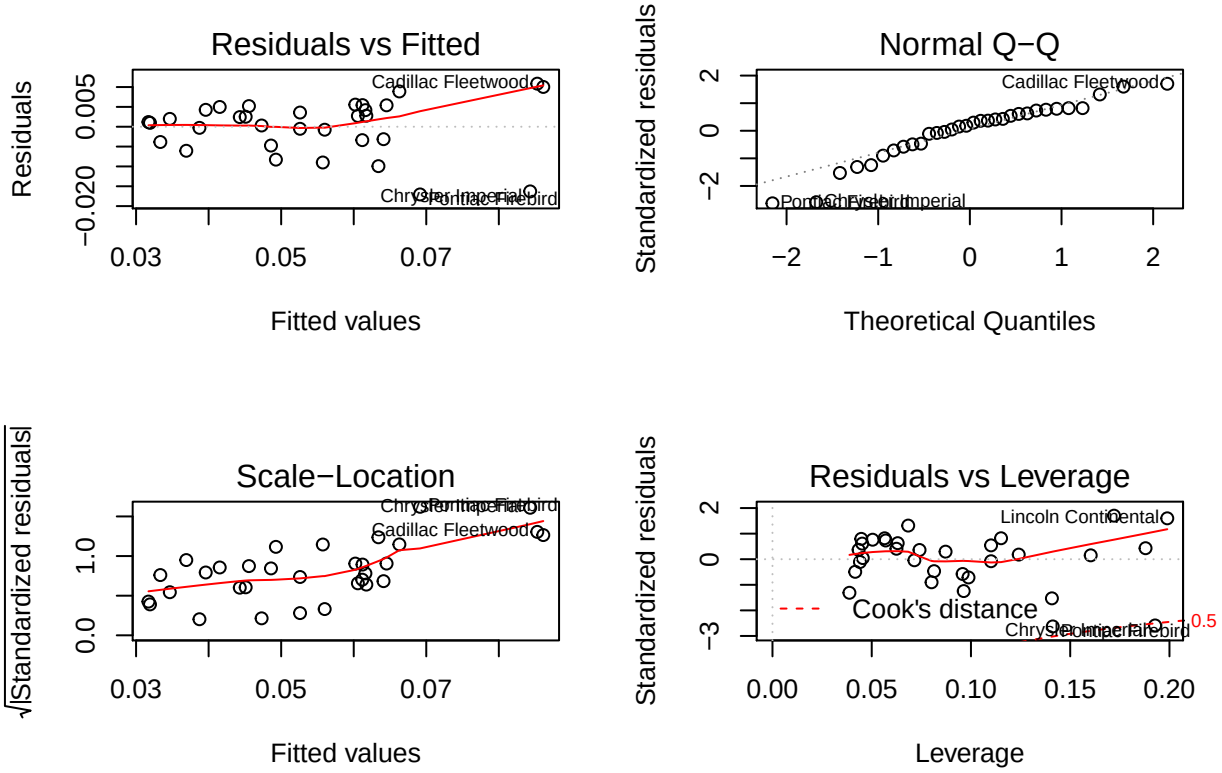
```
pairs(mtcars)
```



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:

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


Residuals vs Fitted
Residuals
Fitted values
Cadillac Fleetwood
Chrysler Imperial
Pontiac Firebird
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Cadillac Fleetwood
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
Cadillac Fleetwood
Residuals vs Leverage
Standardized residuals
Leverage
Lincoln Continental
Cook's distance
0.5