Coursera Regression Modelling Project

Christopher Jones

December 18, 2017

Motor Trend Data Analysis:
The Effect Of Transmission Type On Fuel Economy

Executive Summary

This report examines the impact on gas mileage attributable to car transmission type (automatic vs manual), controlling for a variety of other variables. Our data will be R's built-in `mtcars` dataset, and we will use regression modelling techniques in the analysis.

Despite at first glance there appearing to be a large (42.7%) increase in mileage from automatic to manual, taking account of the correlations among other variables shows this influence to be somewhat less. What remains after accounting for the primary variables is a more modest (17.2%) increase going from automatic to manual.

Data Loading & Exploatory Analysis

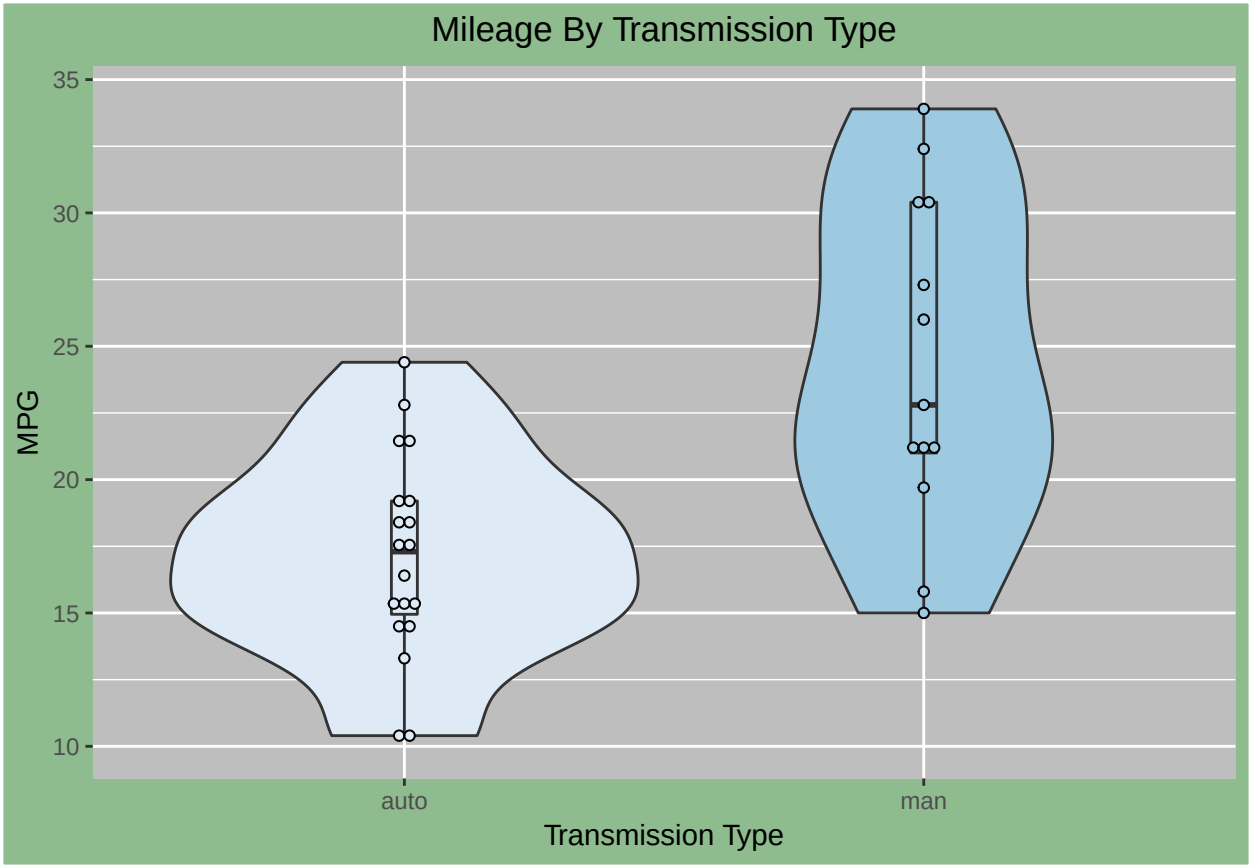
First we load libaries we'll use, along with the data to be analyzed:

```
##                   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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Next we perform some light data processing for easy later use.

Now we take a quick look at our data with violin plots, to get an initial idea of the data and their distributions.

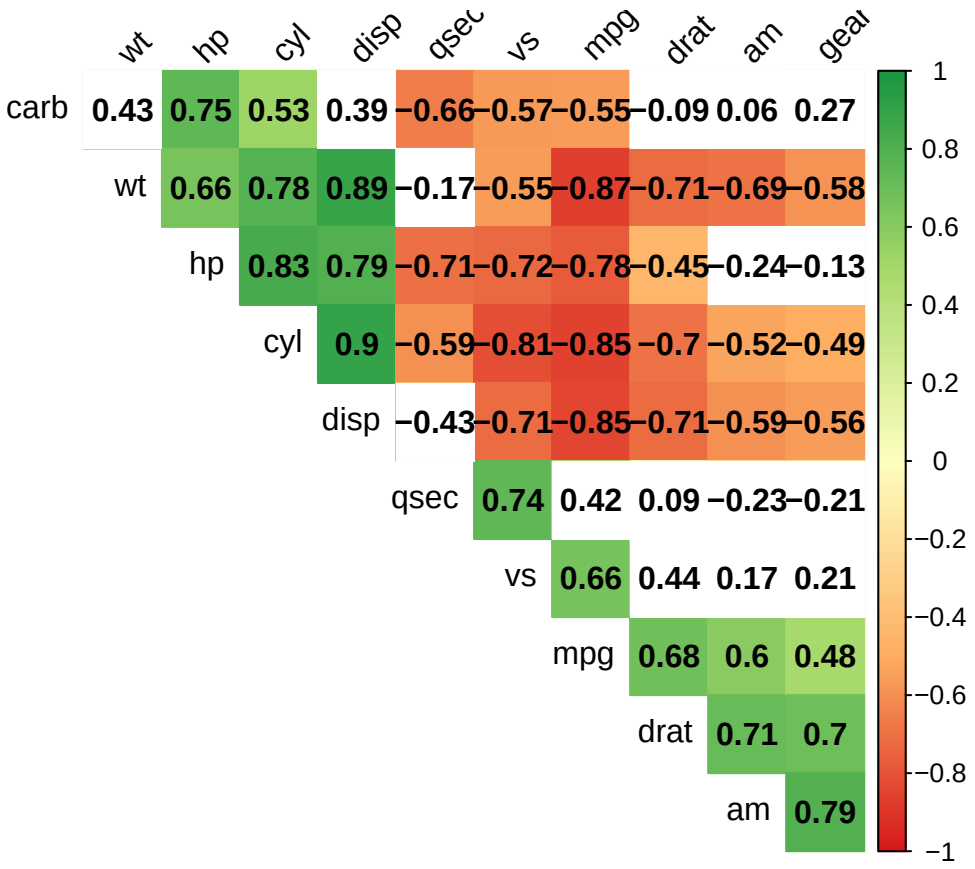
The first plot is mileage vs transmission type, the predictor and outcome we are interested in here. We see a clear difference in their means and in their overall distributions. The mileages for automatics are concentrated near the group's mean of $\mu_A = 17.1$ mpg, while those for manuals are much more uniformly distributed along the category's range, with a mean of $\mu_M = 24.4$ mpg.



–>

Model Selection

We look at the variable correlation plot to begin model selection:

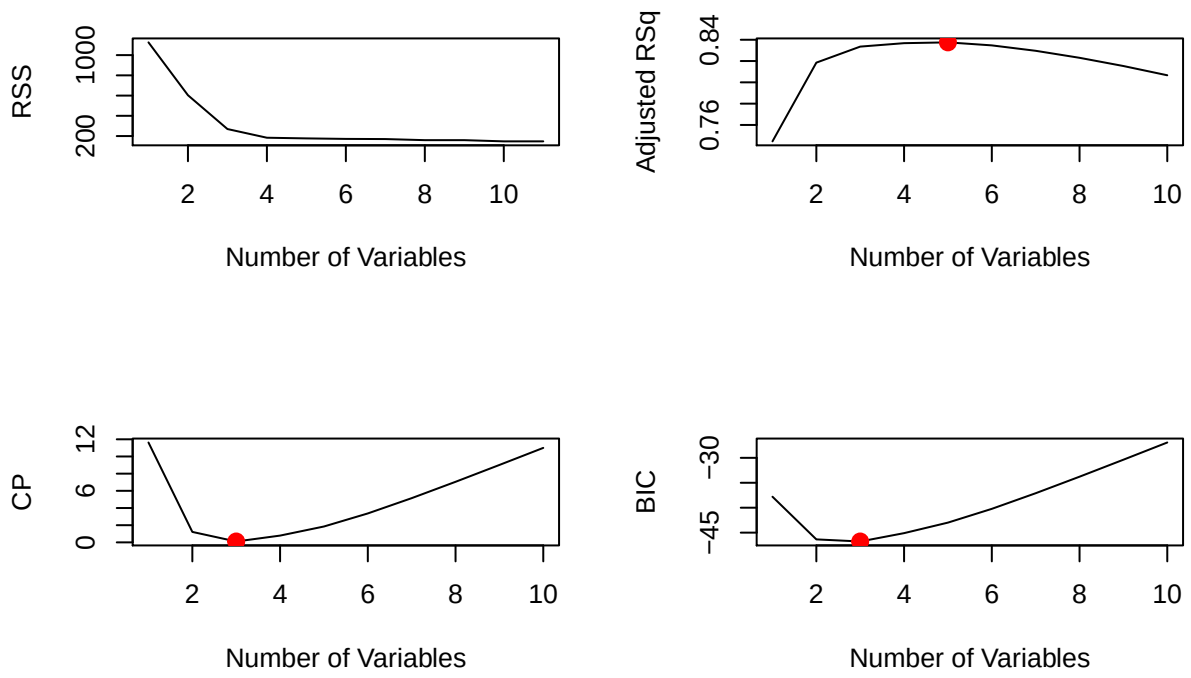


With so much correlation, we wish to remove as many variables as possible to avoid model overfit (weak predictive ability) but no more (to avoid bias from underfit). To assist in determining which variables to remove, we use the best-subsets technique from the`leaps` package.

```
##          cyl disp hp  drat wt  qsec vs  am  gear carb
## 1  ( 1 ) " " " "  " " " " "*" " "  " " " " " " " "
## 2  ( 1 ) "*" " "  " " " " "*" " "  " " " " " " " "
## 3  ( 1 ) " " " "  " " " " "*" "*"  " " "*" " " " "
## 4  ( 1 ) " " " "  "*" " " "*" "*"  " " "*" " " " "
## 5  ( 1 ) " " "*"  "*" " " "*" "*"  " " "*" " " " "
## 6  ( 1 ) " " "*"  "*" "*" "*" "*"  " " "*" " " " "
## 7  ( 1 ) " " "*"  "*" "*" "*" "*"  " " "*" "*" " "
## 8  ( 1 ) " " "*"  "*" "*" "*" "*"  " " "*" "*" "*"
## 9  ( 1 ) " " "*"  "*" "*" "*" "*"  "*" "*" "*" "*"
## 10  ( 1 ) "*" "*"  "*" "*" "*" "*"  "*" "*" "*" "*"
```

Because the purpose of this report is to determine the influence of transmission type (variable `am`), we reject the 2 most parsimonious models as they don't include the `am` variable.

To decide among the remaining model possibilities, we examine some of the key model performance indicators:



So we see that while maximum variability explained is attained at the 5-variable model, there isn't much improvement beyond the 3-variable model. Also, desired minimums of CP (precision) and BIC (informativeness) are attained at the 3-variable model. Therefore we select the 3-variable model consisting of predictors `wt`, `qsec`, and `am`.

Linear Regression & Diagnostics

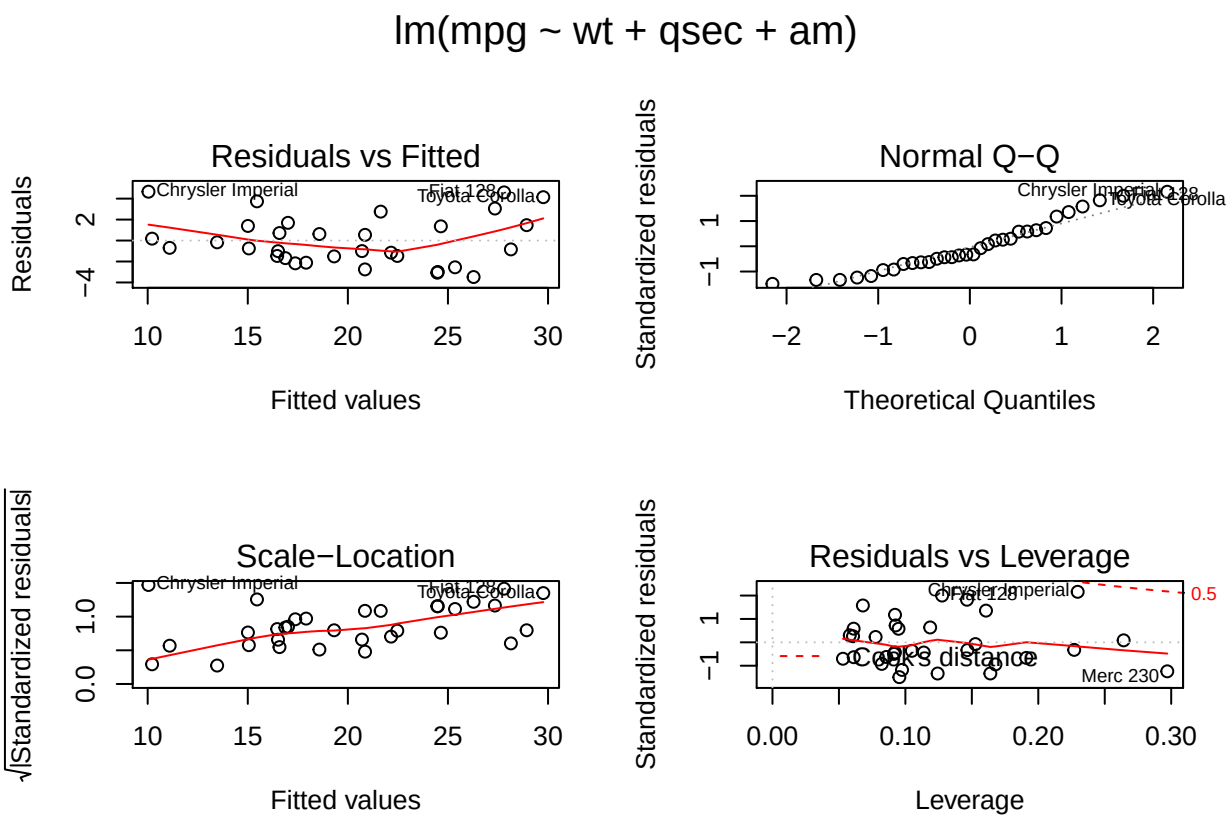
Running the 3-variable model selected in the previous section, `mpg` ~ `wt` + `qsec` + `am`, gives us:

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Our model appears to explain about 83% of the variability in the data (close to the maximum). According to the model, when weight and quarter second time are held constant, going from automatic to manual transmission changes mileage by about 2.92 mpg.

Now we run some diagnostics to verify the goodness of our selected model.



With this diagnostics plot, we observe the following:

1. On the residuals vs fitted plot there appears to be no particular pattern to the residuals.
2. The residual quantile-quantile plot falls approximately on the x=y line, supporting the condition that errors are normally distributed.
3. The scale-location plot is similar to the residuals plot - the points appear randomly spread.
4. There appear to a a couple of candidates for outliers in the data: the Chrysler Imperial and the Mercedes 230. Further investigation may determine these observations are better off excluded.

Conclusions

We have seen in this report that a good linear model indicates an increase of approximately 2.92 mpg when switching from automatic to manual transmission (holding primary predictors constant). Other candidate models were considered, being rejected either as not being informative/parsimonious, or for not containing the predictor we are focussed on in this report. Robustness of the chosen model was verified via standard residual plots.