Fuel consumption comparison of automatic and manual transmission cars

Janos Brezniczky

27 December 2015

Executive Summary

The analysis of the given data did not suffice to prove a relationship between the transmission type (automatic/manual) and the miles per gallon (MPG). The final model suggests an increase in MPG of 0.276 (+- 2.75), which is equal to a random noise with a 84.2% chance. To conclude with more confidence, more data, potentially about high MPG/automatic transmission and low MPG/manual transmission cars is needed.

Exploratory Analysis

The analysis is based on the mtcars data set which is bundled with R. The R help contains a description of the data set. It consists of 11 fields as seen in figure 1. Except for the *V/S* field interpretation was provided there, for that one a little investigation was necessary (see appendix).

For this analysis the *am* variable is converted to a factor with levels *"automatic"* and *"manual"*, and *vs* is similarly dealt with.

Having a look at the correlation matrix (fig. 2, 3) suggests that most variables are intercorrelated and thus to ultimately answer and quantify which transmission type is more beneficial to the MPG of a car; confounder variables will need to be considered.

The MPG value seems to vary together with every other factor, given the correlation values, most weakly with the quarter mile time and the number of forward gears. The most significant correlations are experienced with the weight, displacement and the number of cylinders variables.

Impact of transmission type on consumption

Considering a simplistic model containing MPG as the outcome and only the transmission type as an explanatory variable is exhibited on figure 4. This chart both shows that the choice of automatic or manual transmission alone implies a big (around 5 MPG) change in fuel consumption and that there is little or no data to estimate from about low MPG manual transmission cars. The respective model (simple.model) accords with the first statement.

However, the aforementioned potential for confounders and the high correlation between MPG and other variables suggest to attempt adjusting for covariates. The highest correlation and reasonable thinking (heavier cars are likely to consume more fuel, especially on inclines) suggest weight (wt) is a good choice.

Augmented models

In the following, a series of nested models will attempt to refine the initial conclusion. When adjusting for further variables, the significance of the AM variable disappears and no further refinement brings it back.

1. Introducing weight into the model (fig. 5) removes the effect the transmission type formerly seemed to have. The p-value of AM increases to 0.988. This suggests the variable is unnecessary, contrary to weight and the intercept.

2. The next plot suggests that there is a "U"-shaped trend in the residuals. One possible explanation is the presence of non-linearity. Note that from the recent linear model one could have extrapolated as "heavy enough cars will have a negative MPG" due to coefficient signs, which is obvious nonsense. This is another way to reason for non-linearity in the response. While the data set is likely to be insufficient for identifying the particular non-linear relationship (investigating the domain could help), thinking of Taylor-polynomials suggests that the inclusion of higher order term(s) can easily be an adequate improvement. Including the square of weight among the predictors seems to confirm this (fig. 6), getting a coefficient of 1.3574 and a p-value well below 1%. The R-squared values also improve.

3. and 4. Looking for a pattern in the outliers (Fiat 128, etc.) shows (fig. 7) that the three cars are very similar in horsepower. This is a convenient covariate candidate both numerically (it is near-continuous) and semantically (increasing hp and keeping all other aspect equal is a technological challenge to a designer, other factors are likely to degrade when having to compromise). Including it and its second power further improves the models giving much smaller residuals (cca. -2 to +4) than before (cca. -4 to +6).

Justification of the model modifications

Analysis of variance (fig. 8) suggests that the nested improvements were correct, each iteration is acceptable on its own right.

According to the R-squared values, the explained part of the variation is above 85%, which, considering there may be known and unknown unknowns in explaining the MPG – such as aerodynamic properties of the chassis, bearing and suspension quality; potentially a good achievement.

Conclusion

With the given data, no reliable evidence of a relationship between transmission type and MPG has been found using even the most plausible refinement of the simplest straightforward model. The model building procedure is confirmed by R-squared values, ANOVA tests and p-values of variable-specific t-tests.

Any visible relationship of the type sought seems to be too weak compared to the prediction error, which – if there is a relationship at all, is probably down to having a dataset too small in size, or too much focused on automatic transmission cars with high MPG and manual ones with a low MPG.

Appendix: code and figures

Figure 1: Data interpretation

Field	Interpetation	Field	Interpretation
mpg	Miles/(US) gallon	qsec	1/4 mile time
cyl	Number of cylinders	vs	V/S
disp	Displacement	am	Transmission (0 = automatic, 1 = manual)
hp	Gross horsepower	gear	Number of forward gears
drat	Rear axle ratio	carb	Number of careburetors
wt	Weight (1000 lbs)		

For vs, "0 means a V-engine, and 1 straight engine". Source: http://stackoverflow.com/questions/18617174/r-mtcars-dataset-meaning-of-vs-variable

Figure 2: Examining relationships - correlation matrix plot

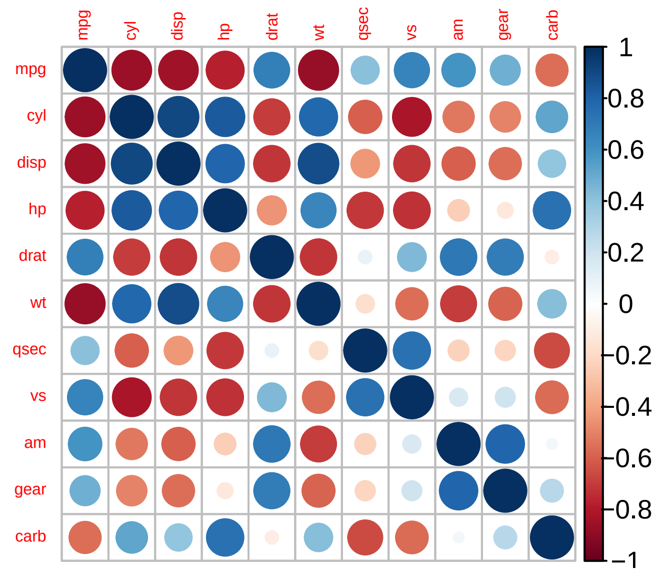
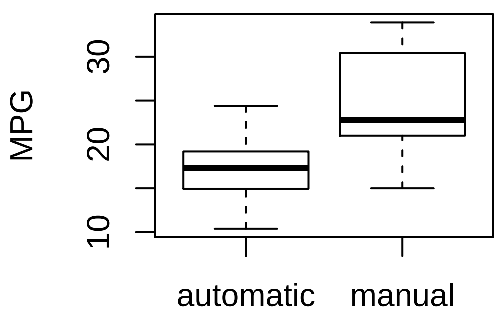


Figure 3: Correlations with MPG

```
##         mpg        cyl       disp         hp       drat         wt
##   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##        qsec         vs         am       gear       carb
##   0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Figure 4: Automatic vs. manual transmission fuel consumption




```
##              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

## adjusted R-squared:  0.3384589
```

Figure 5: Automatic vs. manual fuel consumption modelled with vehicle weight

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## ammanual    -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

```
## adjusted R-squared:  0.7357889
```

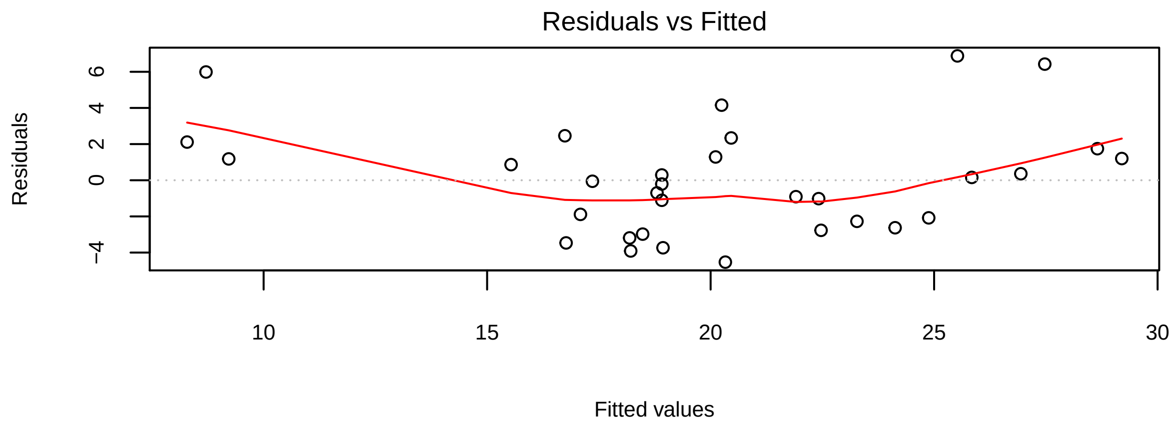


Figure 6: Further nested models - non-linearity

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   54.815037  5.5539045  9.869640 1.284838e-10
## ammanual      -1.862137  1.4040943 -1.326220 1.954841e-01
## wt           -15.316505  2.8793068 -5.319511 1.158486e-05
## I(wt^2)        1.357412  0.3816344  3.556838 1.359414e-03

## adjusted R-squared:  0.811515
```

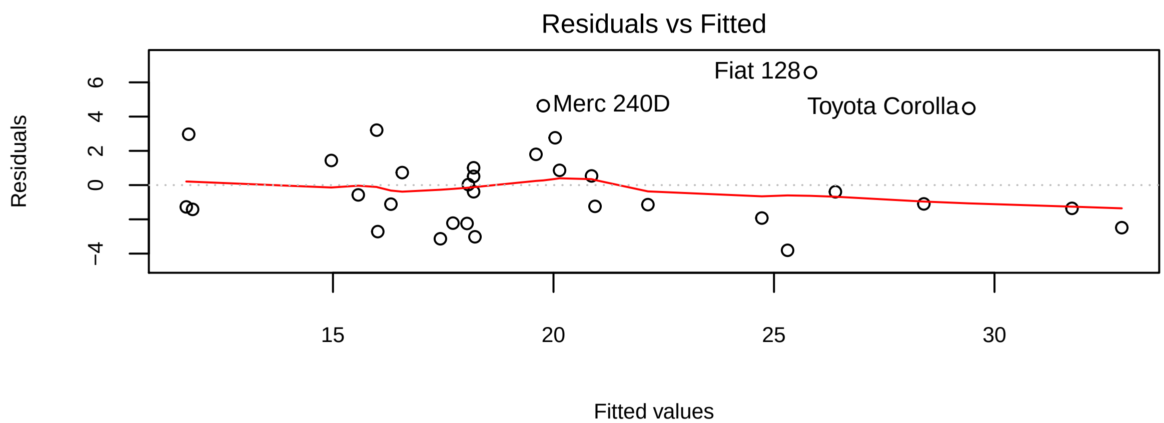


Figure 7: Further nested models - adding horsepower

```
##                 mpg cyl  disp hp drat    wt  qsec       vs        am gear
## Fiat 128       32.4   4  78.7 66 4.08 2.200 19.47 straight    manual    4
## Merc 240D      24.4   4 146.7 62 3.69 3.190 20.00 straight automatic    4
## Toyota Corolla 33.9   4  71.1 65 4.22 1.835 19.90 straight    manual    4
##                carb
## Fiat 128          1
## Merc 240D         2
## Toyota Corolla    1

## Nested model 3
```

```
##                  Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  47.03542258 5.552222031  8.4714592 4.388749e-09
## ammanual      0.27879319 1.431945437  0.1946954 8.470890e-01
## wt          -10.44601492 3.020099127 -3.4588318 1.815583e-03
## I(wt^2)       0.94753113 0.363808630  2.6044768 1.477934e-02
## hp           -0.02820296 0.009442122 -2.9869304 5.932641e-03

## adjusted R-squared:  0.8530812

## Nested model 4

##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 50.2752935296 5.443393e+00  9.2360209 1.080051e-09
## ammanual    -0.2755581882 1.372551e+00 -0.2007635 8.424457e-01
## wt          -9.5626829576 2.872194e+00 -3.3293998 2.609232e-03
## I(wt^2)      0.8814657114 3.437469e-01  2.5642867 1.646426e-02
## hp          -0.0946009898 3.255652e-02 -2.9057465 7.389472e-03
## I(hp^2)      0.0001775252 8.374073e-05  2.1199380 4.371580e-02

## adjusted R-squared:  0.8699158
```

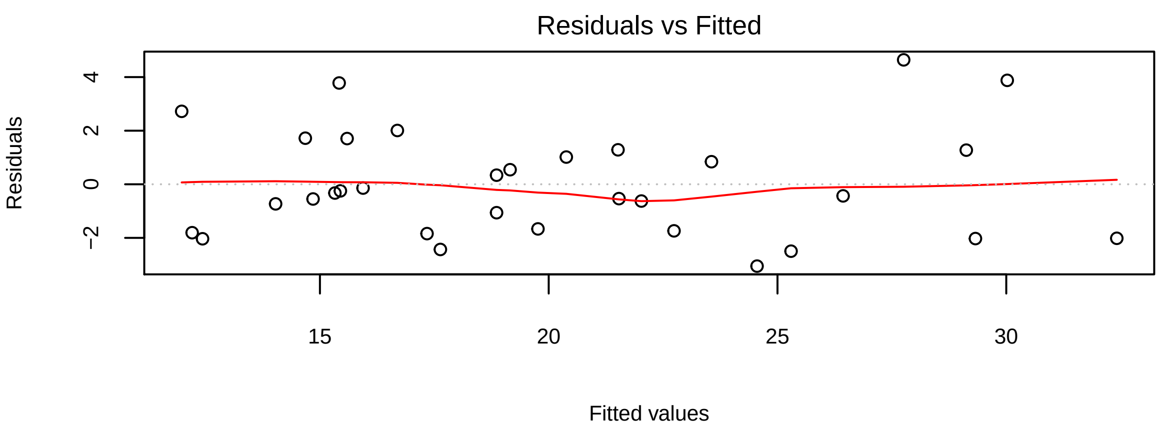


Figure 8: Analysis of variance

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + I(wt^2)
## Model 4: mpg ~ am + wt + I(wt^2) + hp
## Model 5: mpg ~ am + wt + I(wt^2) + hp + I(hp^2)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 93.6632 4.173e-10 ***
## 3     28 191.70  1     86.62 18.3308 0.0002237 ***
## 4     27 144.09  1     47.61 10.0763 0.0038392 **
## 5     26 122.86  1     21.24  4.4941 0.0437158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```