Does transmission type affect car fuel economy?

Michael Hunt

Monday, June 15, 2015

Executive Summary

We analyse the mtcars data set to answer the question of the title, using gallons per 100 miles (gpm) as an inverse metric of fuel economy. Simple physics suggests that many factors might affect fuel economy, besides transmission type. Correlation of several factors is indeed confirmed by an exploratory pairs plot. An initial regression of gpm on several explanatory variables suggests that among continuous variables, only weight is a significant predictor of fuel economy. A subsequent regression of fuel economy on weight and transmission type confirms this, and, further, finds evidence at the 95% confidence level that manual transmission cars require a mean value of 0.7 gallons more fuel per 100 miles than automatic transmission cars, all else being equal.

Pre-processing and exploratory analysis of the data

The data are loaded in from R. **The data for the Chrysler Imperial are removed from the set and the subsequent analysis, since a preliminary analysis showed this to be an outlier, with excessive influence on regression results.** To determine whether the transmission type of the cars has a significant effect on their fuel economy, we first express the latter as "gallons per 100 miles" variable, named gpm, where gpm $= \frac{100}{\text{mpg}}$. This is an inverse measure to mpg that is commonly used in many countries. It is more likely than mpg to have a linear relationship with factors affecting fuel economy, and so is a more suitable input to a linear regression analysis.

A pairs plot was created for those variables that on physical grounds we suspect might influence the fuel economy of the cars, that is disp (displacement), hp (gross horsepower), wt (weight), qsec (time to reach a quarter mile distance). This is shown in the appendix as Figure 1.

This suggests that all the above mentioned variables except qsec are correlated with fuel economy and could have a linear relation with it. In a first regression analysis, all will first be included as explanatory variables, together with the transmission type.

First regression analysis

```
for (i in c(2,8,9,10,11)){mtcarsno[,i]=factor(mtcarsno[,i])}
fitno<-lm(gpm~factor(am)+wt+disp+qsec+hp-1,data=mtcarsno)
round(cbind(summary(fitno)$coef,confint(fitno)),3)
```

```
##             Estimate Std. Error t value Pr(>|t|)  2.5 % 97.5 %
## factor(am)0    1.425      2.354   0.605    0.550 -3.423  6.272
## factor(am)1    1.719      2.162   0.795    0.434 -2.735  6.172
## wt             1.395      0.300   4.658    0.000  0.778  2.013
## disp           0.002      0.003   0.681    0.502 -0.003  0.007
## qsec          -0.078      0.114  -0.682    0.501 -0.312  0.157
## hp             0.003      0.003   0.871    0.392 -0.004  0.010
```

The intercept was excluded in the above analysis, so each of the estimate values in the summary above represent the amount by which the fuel economy, as expressed by gpm, will change, per unit change in the

variate with all the other variates included in the regression being held constant. The tiny p-value of the F-statistic for the model ($p = 2.2 \times 10^{-16}$) shows that the model as a whole has predictive power, but the p-values and confidence intervals for the continuous variables suggest that weight appears to be the only statistically significant factor among them in predicting fuel economy. It is the only one with a P value less than 0.05, and the only one for which the confidence interval does not straddle zero.

Calculate tolerances

The lack of significance of many of the variables indicated above may be becuase of multicollinearity. To further check this, we regress each one onto the other variables used above and calculate the tolerance ($T = 1 - R^2$) and and variable inflation factor $V = \frac{1}{T}$. Various threshold values are used for these in the literature, but we shall take $T < 0.2$ and hence $V > 5$ as a threshold indicator of collinearity

```
dispV<-1-summary(lm(disp~factor(am)+wt+hp+qsec,data=mtcarsno))$r.squared;
hpV<-1-summary(lm(hp~factor(am)+wt+disp+qsec,data=mtcarsno))$r.squared;
qsecV<-1-summary(lm(qsec~factor(am)+wt+disp+hp,data=mtcarsno))$r.squared;
tolerance<-data.frame(c("disp","hp","qsec"),round(c(dispV,hpV,qsecV),3))
colnames(tolerance)<-c("Predictor","VIF");tolerance
```

```
##   Predictor   VIF
## 1      disp 0.122
## 2        hp 0.202
## 3      qsec 0.264
```

We conclude that disp should not be used as a predictor since it is collinear with weight, but that inclusion of qsec and hp (just) will not affect regression results on weight. However, since the p-values for qsec and hp are high > 0.05, we leave them out anyway.

Second regression analysis

Given the result above, we perform a second regression, including only transmission type and weight as explanatory variables, with gpm as the response.

```
fitno<-lm(gpm~factor(am)+wt,data=mtcarsno)
round(cbind(summary(fitno)$coef,confint(fitno)),3)
```

```
##             Estimate Std. Error t value Pr(>|t|)  2.5 % 97.5 %
## (Intercept)   -0.957      0.678  -1.413    0.169 -2.345  0.431
## factor(am)1    0.698      0.327   2.137    0.042  0.029  1.366
## wt             1.919      0.179  10.699    0.000  1.552  2.287
```

This result suggests at the 95% confidence level (ie p<0.05) that besides weight, transmission type does affect fuel economy, with manual cars requiring a mean value of 0.7 gallons of fuel per 100 miles more than automatic cars, all else being equal. Figure 2 in the appendix illustrates this.

Check for validity of analysis

We check for independence, heteroskedacticity and normality of the data by plotting residuals vs fitted values (should be scattered evenly around the zero line, with no obvious pattern, and a qq plot - should be astraight line). See appendix Figure 3. The conditions for the validity of the analysis are found to be well met, for the data set in which one outlier - the Chrysler Imperial - was removed. This outlier can be seen in Figure 4.

Appendices

Figure 1: Pairs plot and correlation values

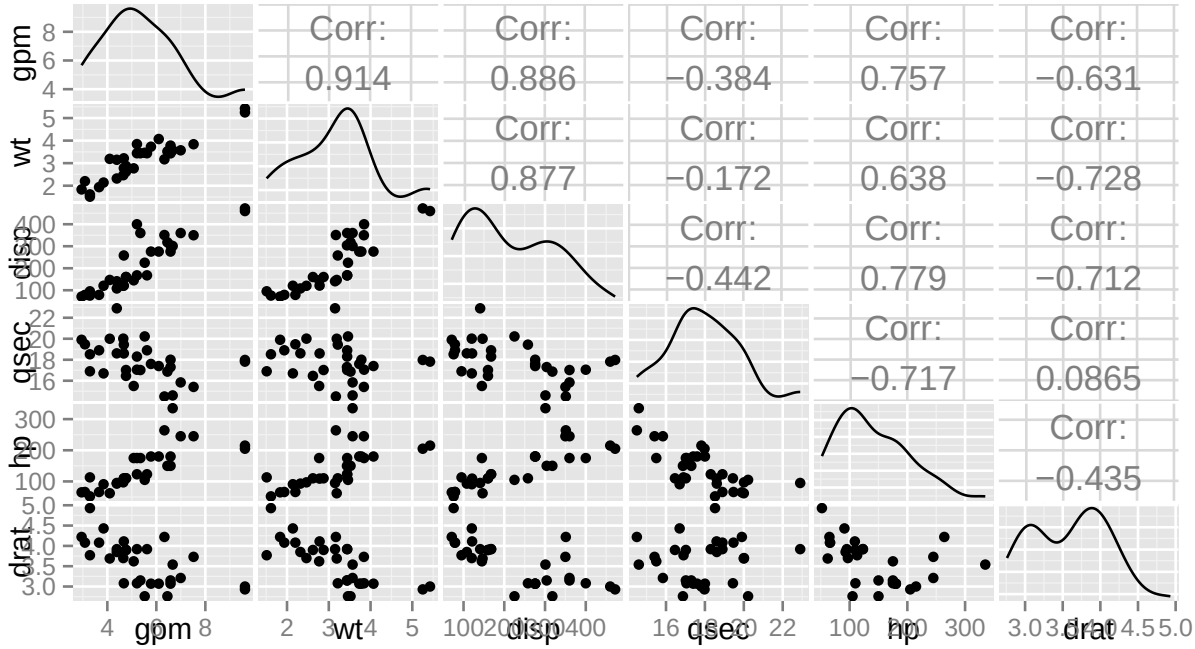

Figure 2: Illustration of difference that tranmsission type makes to fuel economy

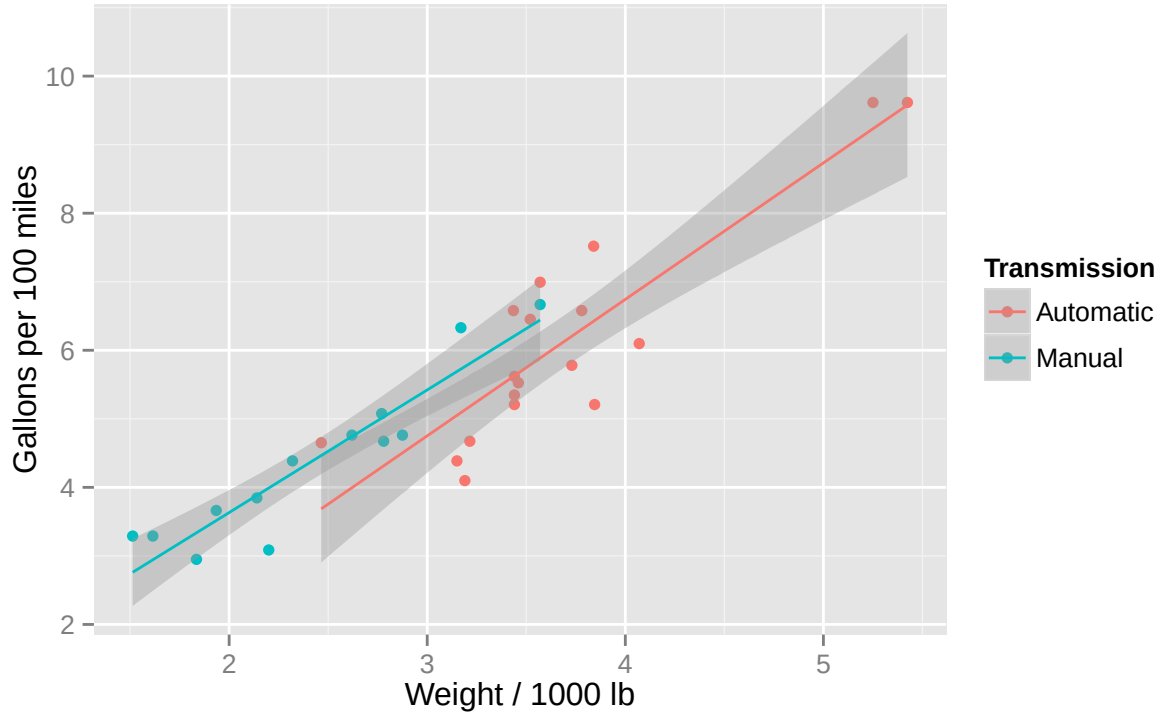

Figure 3: Check validity of analysis - outlier included

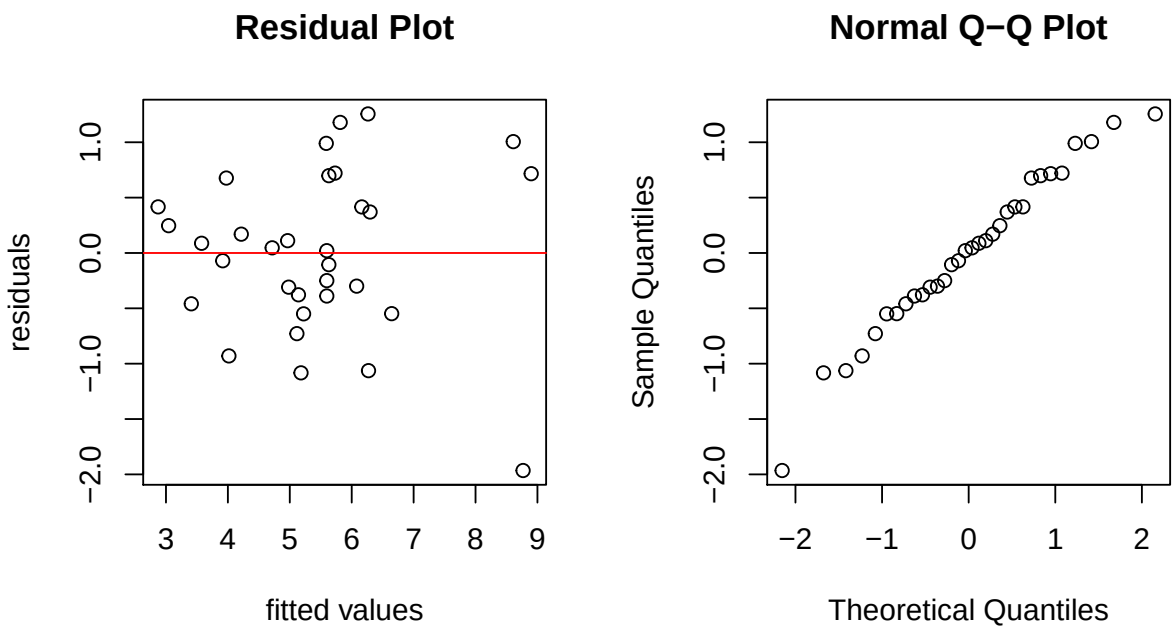

Figure 4: Check validity of analysis - outlier removed