Does transmission type affect car fuel economy?

Michael Hunt

Monday, June 15, 2015

Executive Summary

We perform an analysis to answer the question of the title, using the mtcars data set, and 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 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 no evidence that transmission type makes any difference to fuel economy (P=0.2)

Pre-processing of the data

The data are loaded in from R. 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.

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)){mtcars[,i]=factor(mtcars[,i])}
fit<-lm(gpm~factor(am)+wt+disp+qsec+hp-1,data=mtcars)
round(cbind(summary(fit)$coef,confint(fit)),3)
```

```
##             Estimate Std. Error t value Pr(>|t|)  2.5 % 97.5 %
## factor(am)0    2.551      2.735   0.933    0.360 -3.071  8.172
## factor(am)1    2.587      2.521   1.026    0.314 -2.594  7.768
## wt             1.096      0.335   3.269    0.003  0.407  1.785
## disp           0.002      0.003   0.640    0.528 -0.004  0.008
## qsec          -0.092      0.133  -0.688    0.498 -0.366  0.183
## hp             0.004      0.004   0.889    0.382 -0.005  0.012
```

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 p-value of the F-statistic 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 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=mtcars))$r.squared;
hpV<-1-summary(lm(hp~factor(am)+wt+disp+qsec,data=mtcars))$r.squared;
qsecV<-1-summary(lm(qsec~factor(am)+wt+disp+hp,data=mtcars))$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 0.110
## 2        hp 0.192
## 3      qsec 0.264
```

We conclude that hp and disp should not be used as predictors since they are collinear with weight, but that inclusion of qsec will not affect regression results on weight. This is consistent with the low correlation between qsec and weight found earlier. However, since the p-value for qsec is high > 0.05 , we leave it 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.

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

```
##             Estimate Std. Error t value Pr(>|t|)  2.5 % 97.5 %
## (Intercept)   -0.128      0.743  -0.172    0.864 -1.648  1.391
## factor(am)1    0.483      0.376   1.285    0.209 -0.286  1.252
## wt             1.664      0.192   8.681    0.000  1.272  2.056
```

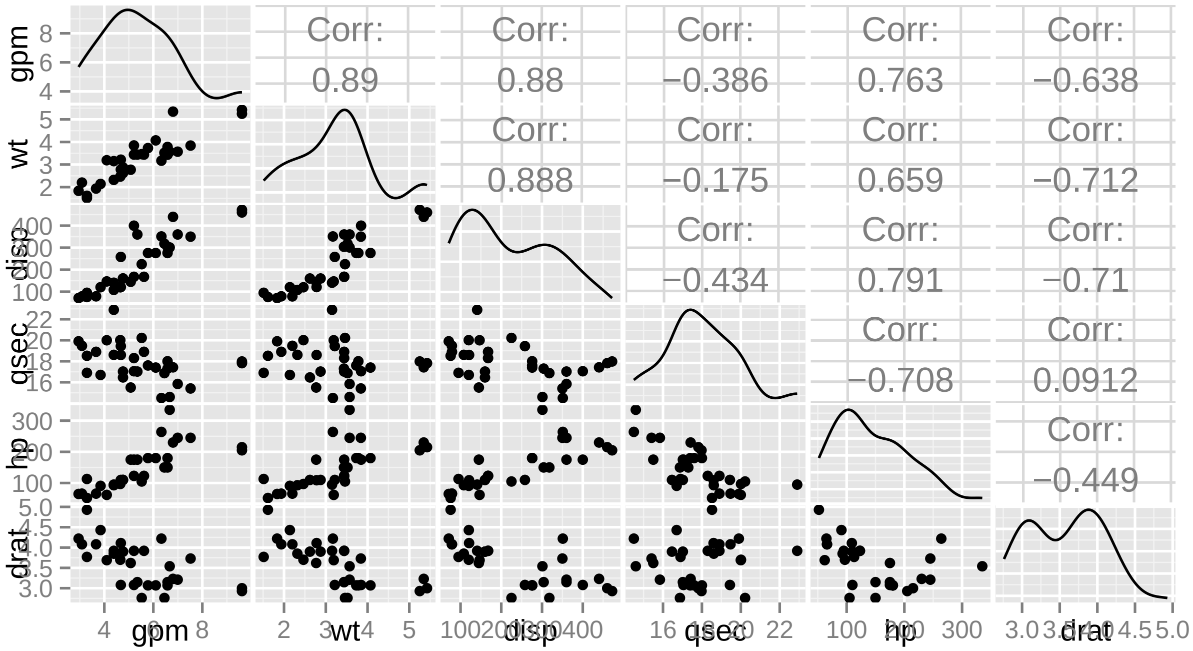
This result shows that we cannot infer from these data that transmission type affects fuel economy, while vehicle weight certainly does. The estimated gpm value for manual transmission at zero car weight is 0.483 more than for automatic transmissions, but with a P value of 0.2 we cannot exclude the possibility that this difference is zero. Indeed we see this also from the confidence interval for manual transmission. It straddles zero, meaning that we cannot be 95% confident that fuel economy is greater or less for manual transmision than it is for automatic transmission.

Check for validity of analysis

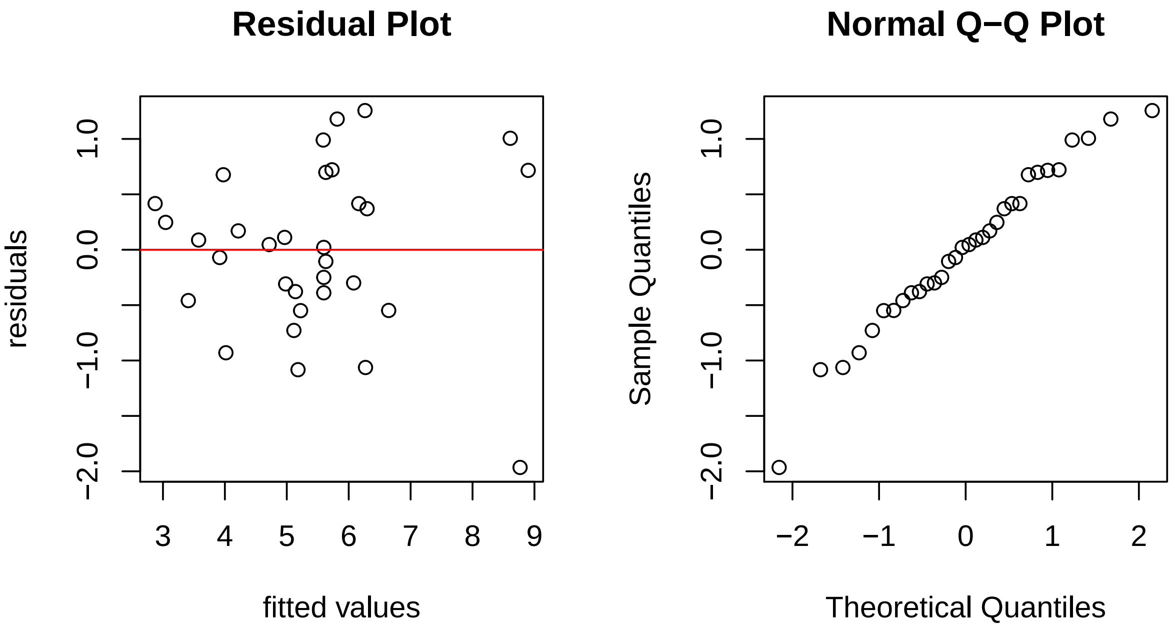
We check for independence, heteroskedacticity and normality of the data by plotting residuals vs fitted values (should be scattered evenly around the zero line, with no obvious pattern, and a qq plot - should be astraight line). See appendix.The conditions for the validity of the analysis are found to be well met, although there is one outlier - the Chrysler Imperial - which may have undue influence.

Appendices

Pairs plot and correlation values



Check validity of analysis


Residual Plot
residuals
fitted values
Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles