Automatic vs Manual Transmission: Which is More Fuel Efficient?

Tamer Koksal

Executive Summary

This study purports to establish what factors make automobiles more fuel efficient. Regression analysis suggest that automobiles with longer quarter mile time tend to be more fuel efficient. There is an inteaction effect between transmission type and weight on fuel efficiency. For cars lighter than 3.40 thousands lbs. manual transmission cars are more fuel efficient than automatic transmission cars, otherwise the opposite is the case.

1. Introduction

This project is intended to explore what automobile characteristics affect fuel consumption/efficiency. Particularly, we are interested in whether automatic or manual transmission is more fuel efficient and how different is fuel consumption between these two transmissions. To study this research question `mtcars` data set found in `datasets` R-package is used. This data comprises information on 32 automobiles and is extracted from the 1974 Motor Trend US magazine[1]. To get to know the data I print below the the first 3 cases/rows of the data along with the variable names/columns.

```
library(printr)
head(mtcars, 3)
```

	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

Description of the variable names is provided in the Appendix(1). Descriptions are based on the `mtcars` documentation and Hocking (1976)[2]

2. Exploratory Analysis

First of all we begin by exploring the types of the variables (see Appendix(2)). Although some of the variables are of categoric nature, all of them are defined as numeric. So, firstly we make a copy of the `mtcars` and call it `mcars`, and then convert `am`, `vs`, and `cyl` (of oredered levels) into factor variables.

```
library(dplyr)
mcars <- mtcars
mcars <- mutate(mcars, cyl = factor(cyl, ordered = TRUE),
                vs = factor(vs, labels = c("V", "S")),
                am = factor(am, labels = c("Auto", "Manual")))
```

[1] Henderson and Velleman (1981), Building multiple regression models interactively. *Biometrics*, 37, 391–411.
[2] Hocking R.R. (1976). The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32, 1-49.

Fuel consumption is measured in miles per US gallon (`mpg`). As a preliminary exploration of what variables might affect `mpg`, we plot pairwise scatter plots across all variables in the data (see Appendix(3)). When these plots are examined, one can see that almost all of the variables appear to have a bivariate relationship with fuel consumption (`mpg`). However, when a regression of `mpg` is run on all of the other variables in the dataset, one can see that none of the variables are significant at $p \leq .05$ (Appendix(4)). So, we need to decide about the correct model with fewer variables.

3. Model Selection and Analysis of fuel efficiency (mpg)

In explaining `mpg` variation we have to come up with the correct regression model, that is, we need to determine all relevant variables having effect on `mpg`. Otherwise, we would have omitted variable(s) bias. Furthermore, we also need to be careful not to have redundant variables or highly correlated variables as regressors. Ohterwise, we would suffer from variance inflation, which in turn would mask significant variables from turning out to be significant. In determining the correct model, the appropriate approach would be to rely on theory. However, since I don't have a domain expertise, I use a statistical method for coming up with the correct model, which is the `stepwise backward regression`. This method starts with a pre-specified full model and iteratively excludes variables possessing the highest `p-value` one at a time. Therefore, after performing stepwise backward regression (Appendix (5)) we come up with the following model `mpg ~ am + wt + qsec`.

```
fitbackward <- lm(mpg ~ am + wt + qsec, data=mcars)
summary(fitbackward)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
amManual	2.935837	1.4109045	2.080819	0.0467155
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162

When we check for possibility of an interaction effect, we see that adding an interaction term significantly improves the model (see Appendix (5)). Therefore, our final true model is `mpg ~ am + wt + qsec + am*wt`.

Regression diagnostics checks (Appendix (6)) suggest no problems with non-linearity, non-constant variance of error terms (heteroscedasticity), and outliers.

```
fitint <- update(fitbackward, mpg ~ am + wt + qsec + am*wt)

summary(fitint)$coefficients
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.723053	5.8990407	1.648243	0.1108925
amManual	14.079428	3.4352512	4.098515	0.0003409
wt	-2.936531	0.6660253	-4.409038	0.0001489
qsec	1.016974	0.2520152	4.035366	0.0004030
amManual:wt	-4.141376	1.1968119	-3.460340	0.0018086

Regression results suggest that all of the variables `am`, `wt`, and `qsec`, as well as the interaction term `am*wt` are significantly associated with `mpg`. When there is no interaction term in the model it is straightforward to interpret the effect of each individual variable on the dependent variable, in this case `mpg`, just by looking to the sign and magnitude of the respective coefficient estimates.

For example when we calculate the predicted `mpg`s using the `fitbackward`, which is the model without interaction, by holding constant `wt` and `qsec` at their means, for manual cars and automatic cars respectively, and take their difference, we see that the difference is nothing but the coefficient estimate of `am` in the regression model (`fitbackward`). See below:

```
dataAuto <- data.frame(am="Auto", wt=mean(mcars$wt), qsec=mean(mcars$qsec))
dataManual <- data.frame(am="Manual", wt=mean(mcars$wt), qsec=mean(mcars$qsec))

predict(fitbackward, dataManual) - predict(fitbackward, dataAuto)
```

```
##        1
## 2.935837
```

In our case, `qsec` is not in the interaction term, so its coefficient estimate tell us that the longer the quarter mile time (`qsec`) the higher the `mpg`. In other words, a second increase in quarter mile time is associated with 1.02 increase in `mpg`. This can aslo be interpreted as the slower the accelaration, the more fuel efficient automobiles are.

As for the interpretation of the effect of `am` and `wt` on `mpg`, the effect of each individual variable on `mpg` depends on the value of the other individual variable in the interaction. So, if our model is as follows:

$$mpg = \beta_0 + \beta_1 am + \beta_2 wt + \beta_3 qsec + \beta_4 am \times wt$$

For a unit increase in `am`, that is, manual transmission cars compared to automatic ones can drive $\beta_1 + \beta_4 \times wt$ more or less depending on the value of the `wt` and thus the value and the sign of this expression. So let's compute the value of this expression at the mean of `wt`:

```
fitint$coef[2] + fitint$coef[5]*mean(mcars$wt)
```

```
##  amManual
## 0.7555845
```

That is, on average, manual transmission cars can drive 0.76 more miles compared to automatic transmission cars for the mean value 3.22 of weight.

This computation can be cross validated using the difference of `predict()` functions similarly as above:

```
predict(fitint, dataManual) - predict(fitint, dataAuto)
```

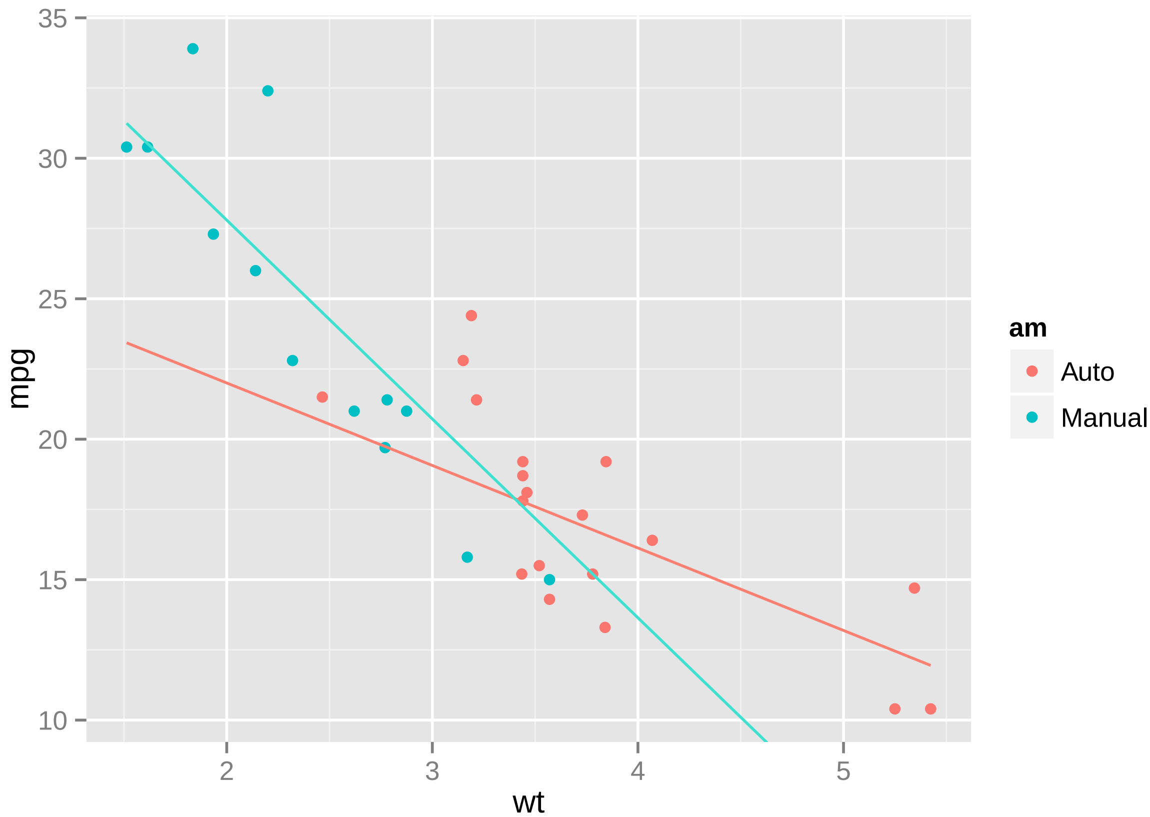
```
##         1
## 0.7555845
```

A unit increase in `wt` (weight) in cars is associated with $\beta_2 + \beta_4 \times am$ increase or decrease in `mpg` depending on the value and sign of this expression as well as the transmission type. For manual cars for instance, a unit increase in weight is associated with -7.08 change in `mpg`.

The effect of change in weight (`wt`) on `mpg` both for automatic and manual transmission is graphically illustrated below.

```r
dataAuto <- data.frame(am="Auto", wt=mcars$wt, qsec=mean(mcars$qsec))
dataManual <- data.frame(am="Manual", wt=mcars$wt, qsec=mean(mcars$qsec))
fita <- predict(fitint, dataAuto)
fitm <- predict(fitint, dataManual)

library(ggplot2)
ggplot(mcars, aes(x=wt, y=mpg)) + geom_point(aes(color = am)) +
    geom_line(x=mcars$wt, y=fita, color="salmon") +
    geom_line(x=mcars$wt, y=fitm, color="turquoise")
```



The value of the `wt` where the two lines intersect is: $-\frac{\beta_1}{\beta_4}$

```r
-fitint$coef[2]/fitint$coef[5]
```

```
## amManual
## 3.399698
```

As can be seen from the graph, for vehicles lighter than 3.40 thousands lbs. the manual transmission ones (the line in turquoise) compared to automatic transmission ones (the line in salmon) have greater `mpg`, that is manual cars are more fuel efficient than automatic ones. For vehicles havier than 3.40 thousands lbs., the salmon line gets higher than the turquoise line, that is automatic cars have greater mpg (are more fuel efficient) compared to manual cars.

As for the uncertainty of the results, this model explains 0.896 (the R^2-value) of the variation in the `mpg`.