Regression Analysis Coursework

Jonathan Bourne

Saturday, October 25, 2014

Exectutive Summary

The report reviews 5 different models and selects a model without any interaction, the modal has an adjusted R^2 of 84% and shows a increase of 1.8mpg when driving a manual car This document has been compiled as a PDF using Knitr and LaTeX. The source code can be found on github here

Introduction

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

Data gathering and exploration

The data set comes from R datasets it includes 32 observations and 11 variables. No observations were removed from the data set. The follwing variables were converted from numeric to factor variables am, vs, cyl, gear, carb, no transformations were performed on the data set no new variables were created. A plot showing the relationship between all variables can be found in the appendix

	Type	P.values
1	All Cars	0.12
2	Manual	0.54
3	Automatic	0.90

The above table shows that the results of the Shapiro Wilks test which suggest that the distribution of the mpg is reasonably normal. this can be shown visually graph form (see appendix)

Analysis

after exploring and reclassing the data an intitial look at the difference between the mean mpg for Automatic and manual vehicles was performed, this showed adifference between the means of 7.2449 mpg with a t-test returning a p-value of 0.0014 meaning the null hypothesis was rejected at the 0.05 confidence level and that the manual cars get more miles per gallon of travel. A bar plot of this result can be seen in the appendix

However the above result does not control for the other variables in the dataset, in order to provide a more meaningful measure of the difference between the two types of car multiple linear regression modelling was performed looking at several different model structures. As there were many variables backwards stepwise regression was performed in order to select the variables to be used in the model. R uses the AIC value when performing a stepwise regression.

The different models investigated where as follows

- modal 0 : mpg = am+carb+cyl+disp+drat+gear+hp+qsec+vs+wt+Intercept
- modal 0 step: mpg = am+cyl+hp+wt+Intercept
- modal 1 : mpg = am+am:carb+am:cyl+am:disp+am:drat+am:gear+am:hp+am:qsec+am:vs+am:wt+carb+cyl+disp+dr
- modal 1 step : mpg = am+am:cyl+am:disp+am:drat+am:hp+am:qsec+am:wt+carb+cyl+disp+drat+gear+hp+qsec+vs
- modal 2 : mpg = am+Intercept

After the final modal was selected based on it's total adjusted R^2 value and the ease of interpreting the coefficients.

	modals	norm.test	adj.r.squares	coeff
1	mod0	0.30	0.78	16.00
2	mod0step	0.45	0.84	5.00
3	mod1	0.00	0.96	23.00
4	mod1step	0.00	0.96	23.00
5	mod2	0.86	0.34	1.00

The above table of the diagnostics of the 5 modals generated, has the folling columns

- modals: The modal identifier
- norm.test: The results of the shapiroWilk test on the normality of the residuals of that modal, values larger than 0.05 can be considered normal
- adj.r.squares: The adjusted R^2 of the model
- coeff: the number of coefficients included in the modal

Reviewing the modals, mod2 can be rejected as it is essentially the same as the two means that were explored at the beginning of this document. Although the Mod1 varients have extremly high R^2 values due to the interaction terms, these interaction terms make interpreting the model more difficult. To add to this although the modals overall have a very high R^2 the R^2 of the actual coefficients are often very low adding to the confusion of interpreting the results. It is possible to do a montecarlo analysis on the Automatic/Manual variable but this needs to be carefullt controlled otherwise unrealistic results can be obtained (see appendix for an example) Due to these considerations the Modal "mod0step" was chosen to be used in the final evaluation.

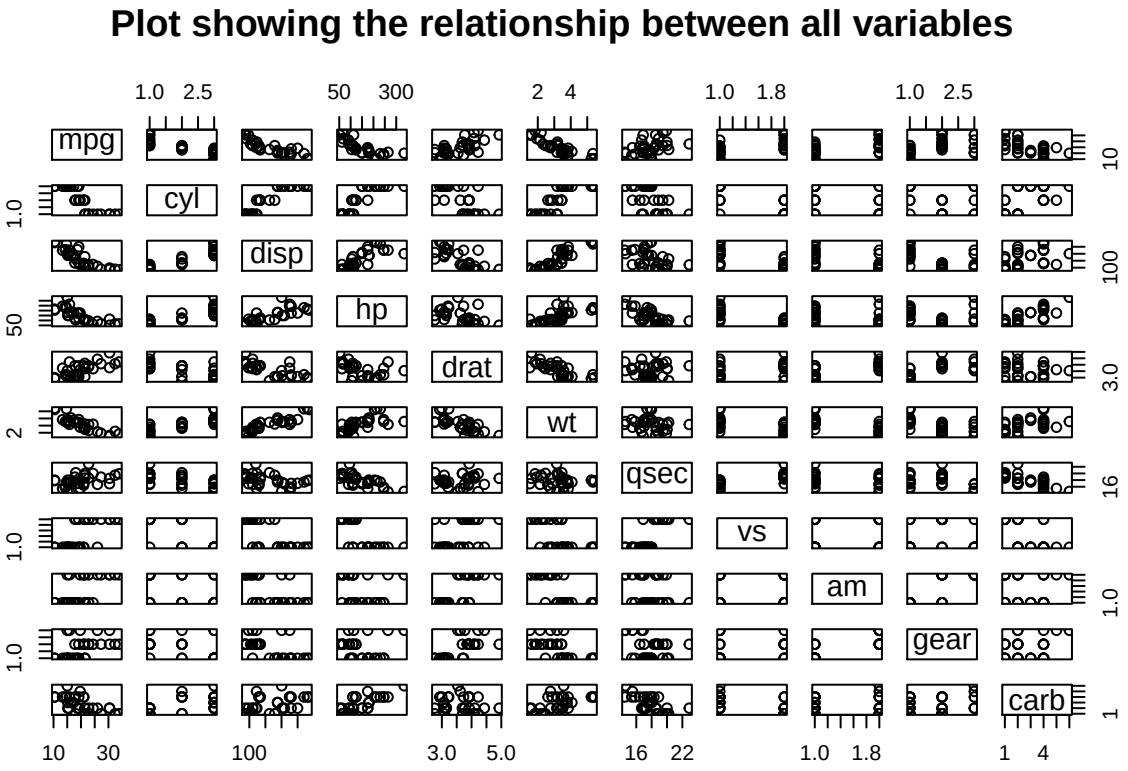
Results

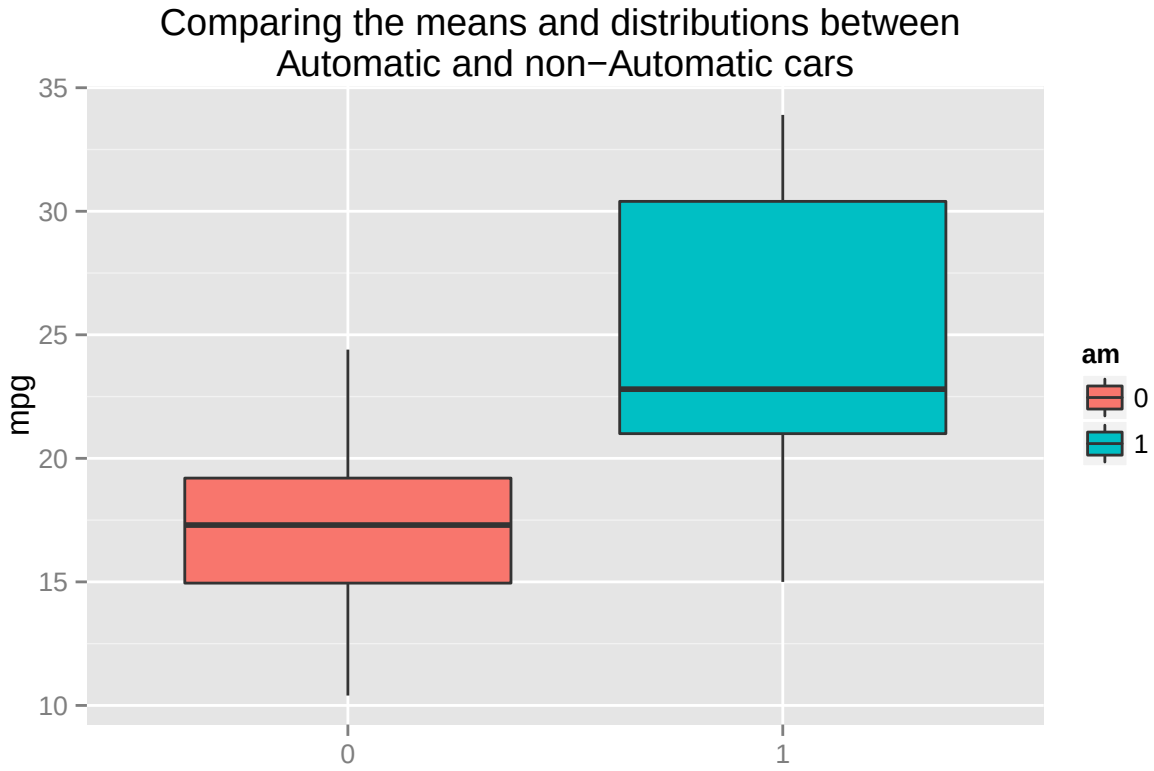
	mod0step$coefficients
(Intercept)	33.71
cyl6	-3.03
cyl8	-2.16
hp	-0.03
wt	-2.50
am1	1.81

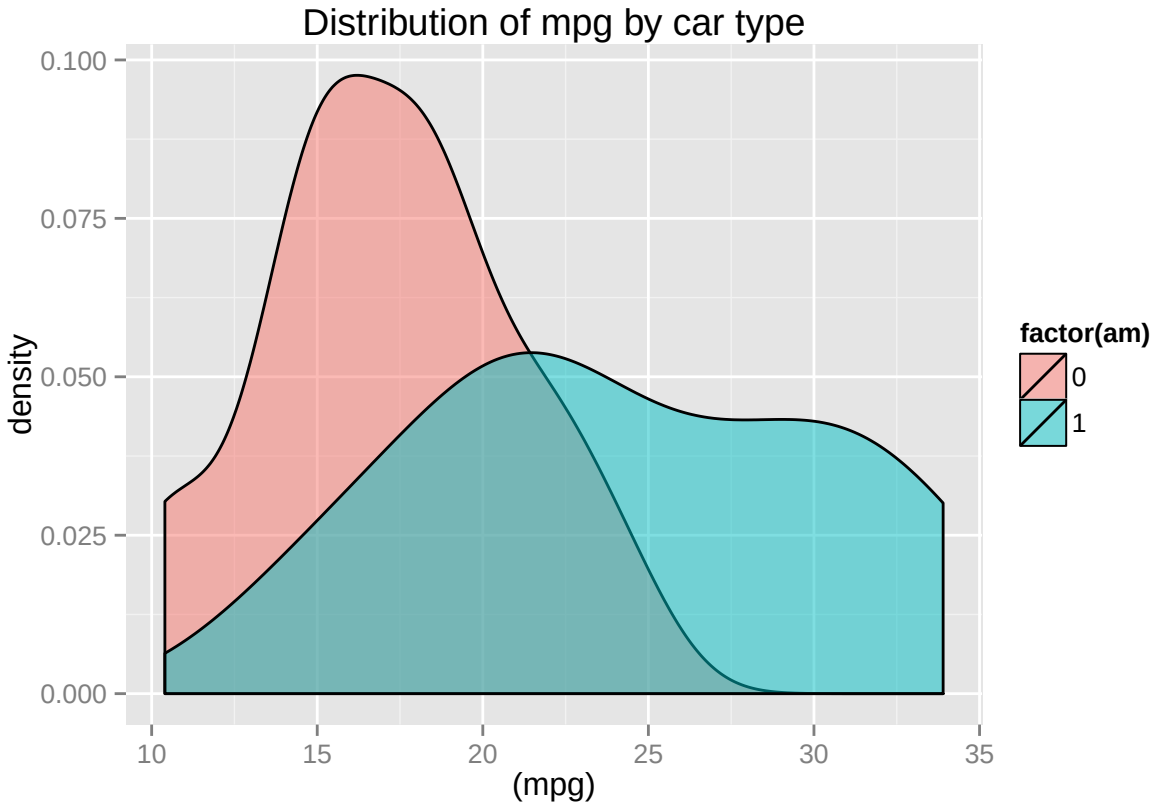
Conclusions

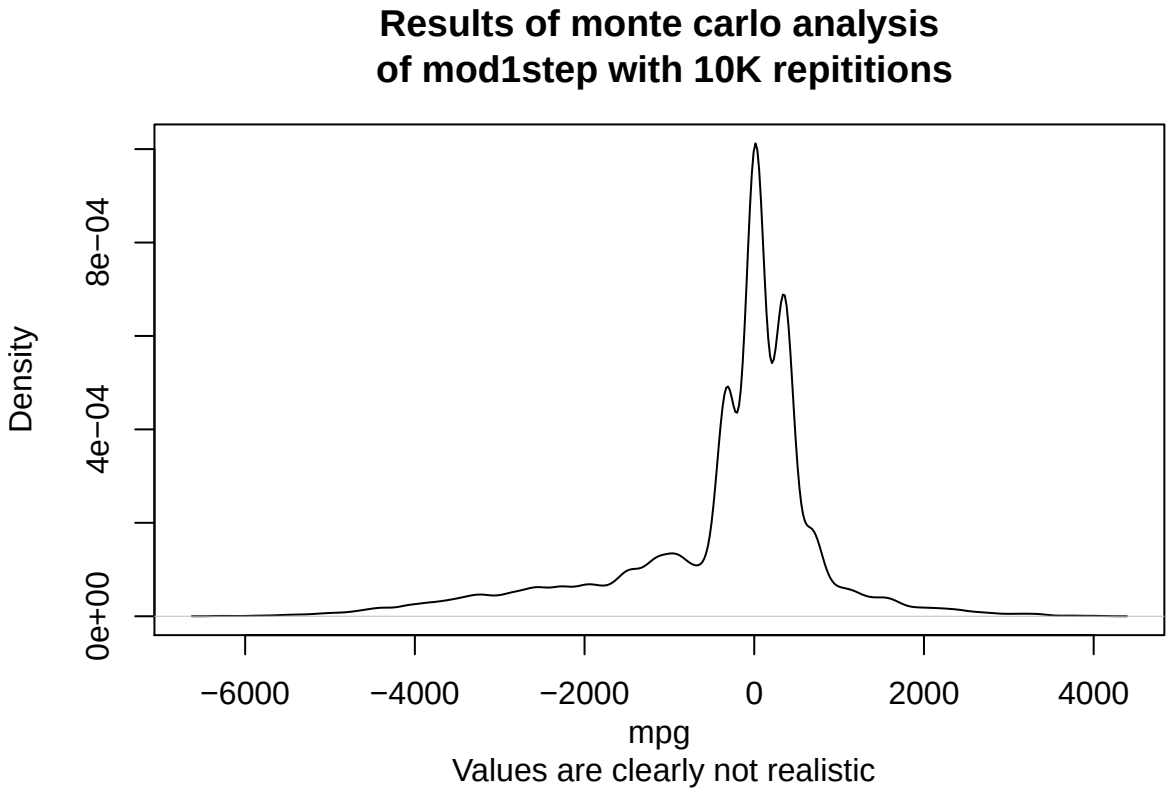
In conclusion it appears that in 1974 manual cars were 1.8092 mpg more fuel efficient than automatic cars when controlling for cylinders, horsepower and car weight. Futher work could explore how interaction plays a part in the fuel efficiency of a car. Although not shown in this report controlling for rear axel ratio (drat) appeared to reverse the sign on the fuel efficiency of the car, this is definately worth investigating further

Appendix


Plot showing the relationship between all variables
mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb


Comparing the means and distributions between
Automatic and non−Automatic cars
mpg
35
30
25
20
15
10
0
1
am
0
1


Distribution of mpg by car type
density
0.100
0.075
0.050
0.025
0.000
10
15
20
25
30
35
(mpg)
factor(am)
0
1


Results of monte carlo analysis
of mod1step with 10K repititions
Density
0e+00
4e−04
8e−04
−6000
−4000
−2000
0
2000
4000
mpg
Values are clearly not realistic