Applied Regression Course Project

Bohdan Horak

Friday, August 22, 2014

Executive Summary

The purpose of this paper is to examine the effects of a variety of car related features on the miles per gallon performance of a vehicle. In order to convincingly illustrate the effects of factors on the car's gaes mileage a variety of statistical tools will be employed, using the native "mtcars" data provided with R. This paper will examine all factors provided in the data set, however the focus will lie on the type of transmission the car has equiped. There are two types of transmissions Manual and Automatic, this paper will exame which is better or wether the differences between the two statistically insignificant given the provided data. In our analysis we will use a significance level of 5.

Exploratory Analysis

In order to perform the exploratory analysis the variables were into two camps, continuous and discrete. The table below summarizes which variables are continuous and which are discrete.

	Continous	Discrete
1	mpg	cyl
2	disp	vs
3	hp	am
4	drat	gear
5	wt	carb
6	qsec	

For continous variables I used basic scatter plots with a fitted line to get a sense of the relationship. You can see the plots in figure 1. We can tell that as the weight(wt), the displacement(disp) and horsepower(hp) increase the mpg tends to decrease. The opposite is true for the ear axle ratio(drat) and quarter mile time(qsec).

For discrete variables, a series of bar plots (figure 2) were used. The bar plots show the mean mpg for each variable seperated by the factors in that variable. I also, included an error range which corresponds to a one deviation shift in the standard error of the mean. Please note, the carb variable has very few observation inclduded therefore any comparative anylysis between the factors is sure to be unreliable, so it will be ignored from now on.

Analysis

For the regression analysis the discrete variables were coded as factor variables. The regression system in R automaticaly converts these factor variables into dummy's. For example, the regression will treat the automatic transmission as a 0, in other words the base case, and the manual transmission is treated as a 1. So from looking at our boxplots we can expect the slope coefficient to be positive.

Regression 1

To get things started I did a multiple regression across all variables except carb. The summary of coefficients is provided in figure 4. The R-squared ended up being 0.881, which is quite high. However, the p-values of the T-tests on wether or not the coefficients calculated by the regression are significant or not are all extremely

high and not close to a reasonable significance level of 5%. This likely means that there are confounding variables which are not adding any explanatory power to our analysis. Note: according to the significance level used in order for variable coefficient to be statistically significant the p value of the test has to be less than 5%.

In order to weed out these confounding variables a correlation matrix was created. Variables that have really high or really low correlations have a high likelihood of being the culprits. Each time a potential confounding variable was excluded a new regression was run to see whether results improved.

Regression 2

Looking at the correletation matrix it can be seen that disp and wt have a high a high correlation. Displacement was removed and a new regression was created. The summary of which can be seen figure 5. As can be seen removing the displacemendid not cause the R^2 to decrease by much, it is now 0.8801, however our P values have decreased somewaht, however they are still unsatifactory.

Regressions 4-Final

In the interest of brevity I repeated the process until the p values of the model seemed reasonable. I ended up removing the all but "am"" and "wt"" variables. The results of which can be seen in Figure 7. We can see that both of the P values resulting from the T-test are effectively 0 meaning that they are statistically significant. At this point we can be reasonably certain that low horsepower vehicles with manual transmissions tend to be more gas efficient. The final R^2 is 0.782. This regression model stipulates that a one unit increase horsepower (everythin else held constant) will result in a 0.0589 decrease in mpg. Similarly going from automatic to manual you can expect to see a 5.2777 increase in mpg. The intercept in this case doesn't mean much as a car with an automatic transmission and no horspower is not really a car. If I had more time and space I could'va transformed the hp variable in order to get a more meaningful intercept coefficient.

Residual Analysis

From the residual plots figures 9 and 10 it can be seen that there is no obvious pattern to the variables, and it really does look quite random. Thus, we can be reasonably sure that our model is at least partially correct. Diagnostically speaking, the final model seems good.

Conclusion

To answer the initial question, the transmission type does have a statistically significant effect on the mpg performance of a vehicle. Both the exploratory and the regression analysis confirm that the manual transmission is more gas efficient than the automatic. In conclusion, the final model ended up consisting of one continous variable(weight) and one discrete variable(am). Both the intercepts and the coeficients were statistically significant.

For the RMD file that generated this report please refer to my Github repo.

Appendix

Figure 1:

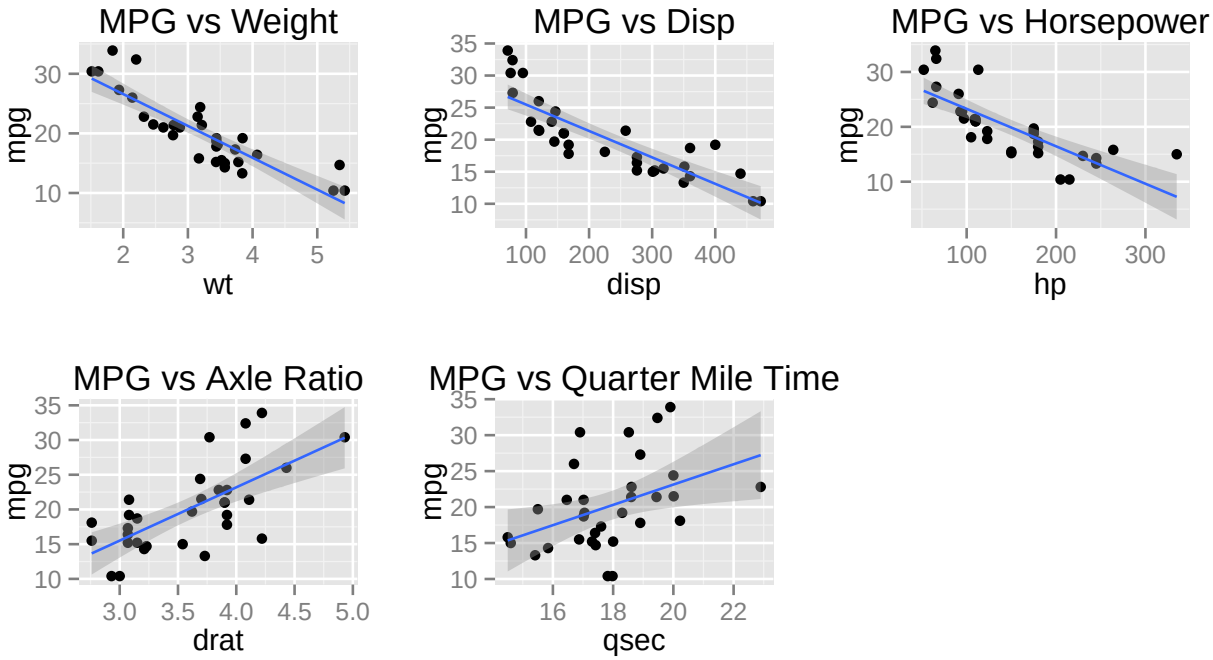

Figure 2:

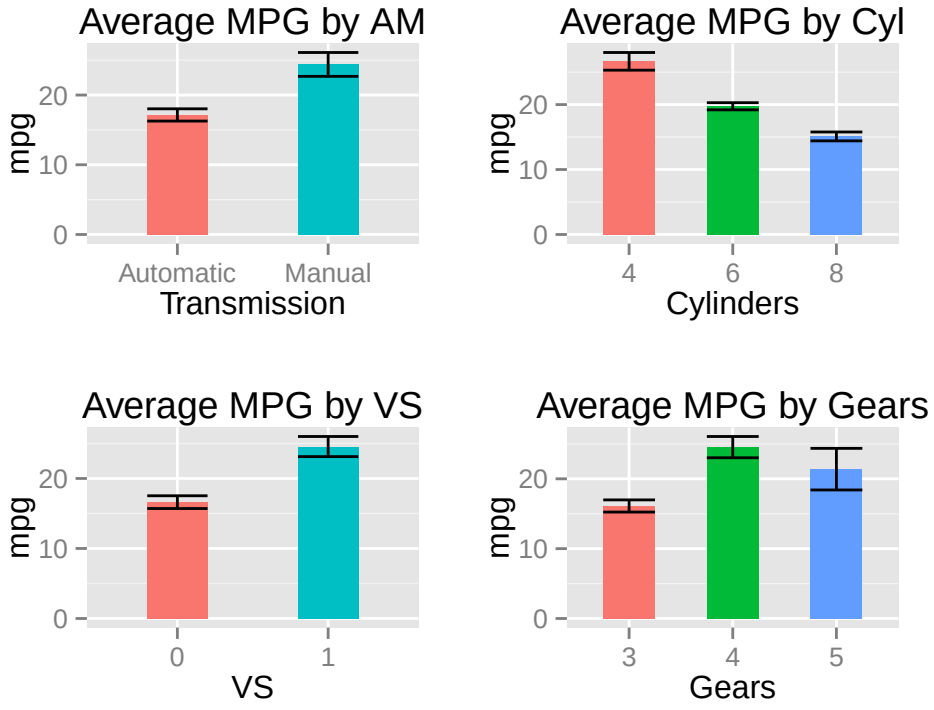

Figure 3: Regression #1

R-squared of the first model is 0.881.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.0444	16.9088	0.95	0.3540
cyl6	-0.7524	2.2891	-0.33	0.7458
cyl8	2.4930	4.7222	0.53	0.6033
disp	0.0059	0.0152	0.39	0.7043
hp	-0.0395	0.0214	-1.84	0.0800
drat	0.8245	1.9601	0.42	0.6785
wt	-2.8537	1.6602	-1.72	0.1011
qsec	0.6439	0.7265	0.89	0.3860
vs1	1.6984	2.2928	0.74	0.4675
amManual	2.9365	2.2010	1.33	0.1972
gear4	0.0998	2.5397	0.04	0.9690
gear5	1.9580	2.7598	0.71	0.4862

Figure 4: Covariance Matrix

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00	-0.85	-0.85	-0.78	0.68	-0.87	0.42	0.66	0.60	0.48	-0.55
cyl	-0.85	1.00	0.90	0.83	-0.70	0.78	-0.59	-0.81	-0.52	-0.49	0.53
disp	-0.85	0.90	1.00	0.79	-0.71	0.89	-0.43	-0.71	-0.59	-0.56	0.39
hp	-0.78	0.83	0.79	1.00	-0.45	0.66	-0.71	-0.72	-0.24	-0.13	0.75
drat	0.68	-0.70	-0.71	-0.45	1.00	-0.71	0.09	0.44	0.71	0.70	-0.09
wt	-0.87	0.78	0.89	0.66	-0.71	1.00	-0.17	-0.55	-0.69	-0.58	0.43
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00	0.74	-0.23	-0.21	-0.66
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00	0.17	0.21	-0.57
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00	0.79	0.06
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	0.27
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

Figure 5: Regression #2

R-squared of the second model is 0.8801.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.1513	16.5601	0.98	0.3405
cyl6	-0.5993	2.2082	-0.27	0.7887
cyl8	3.2183	4.2415	0.76	0.4564
hp	-0.0402	0.0209	-1.93	0.0676
drat	0.9352	1.8992	0.49	0.6275
wt	-2.4059	1.1604	-2.07	0.0506
qsec	0.5993	0.7025	0.85	0.4032
vs1	1.8766	2.1996	0.85	0.4032
amManual	3.0809	2.1244	1.45	0.1618
gear4	-0.2893	2.2822	-0.13	0.9003
gear5	1.7683	2.6599	0.66	0.5134

Figure 6: Regression Final

R-squared of the final model is 0.782.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	26.5849	1.4251	18.65	0.0000
hp	-0.0589	0.0079	-7.50	0.0000
amManual	5.2771	1.0795	4.89	0.0000

Figure 7: Residuals

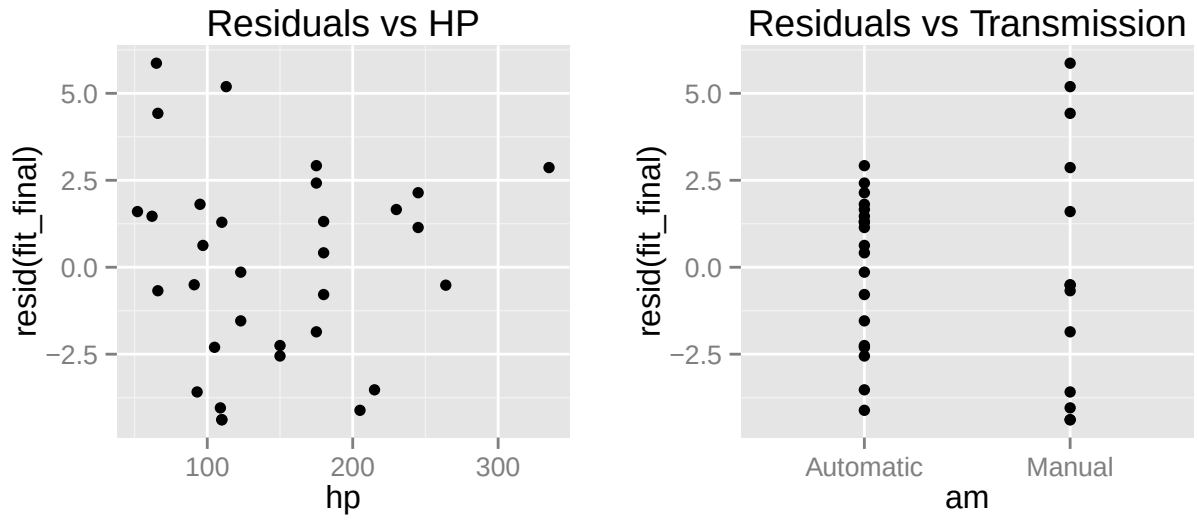

Figure 8: Residuals vs Fitted