Analysis of miles per gallon by vehicle transmission type

Summary

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”

Analysis

The data used for the analysis comes from the mtcars data set provided from the 1974 Motor Trend US magazine, and consists of 32 observations using 11 variables, see Appendix 1 for data structure.

We conduct a preliminary exploration of the relationship of MPG to transmission, taking into account weight and number of cylinders, which are also generally related to displacement. A plot (see Figure 1 in the first Appendix) reveals clear distinctions in the 4 and 6 cylinder cars, but that it is only 4 cylinder cars that markedly show a preference for manual transmission to deliver a better MPG performance. The 8 cylinder cars show a preference for automatic transmission for MPG performance, but an investigation of the two manual cars in the group reveal that they are high performance sports cars, whereas most of the 8 cylinder automatics are large family cars or mid-range performance sports cars.

In order to fairly evaluate the efficiency of manual vs. automatic transmission we would need to have identical models of cars with the alternative transmissions driven under controlled conditions to get the best mileage per gallon. We do not have such a set of data, so the best we can do is to identify what other factors might contribute to a better MPG performance, and try to account for them when modelling the effect of transmission on MPG.

We have already identified number of cylinders and weight as determining factors. Therefore rather than using a simple MPG to transmission relationship we should try to find a more complete set of variables in order to establish the most accurate and parsimonious model to quantify the relationship between MPG and transmission.

For this we can conduct a multivariate regression analysis.

To assist with visualizing the effects of all of the factors we add a column hp.wt to the data to show power to weight ratio (hp/wt), which together with quarter mile performance (qsec) can assist in understanding performance. In Appendix 2 we have printed lists of cars based on transmission type and sorted by MPG and performance.

As a comparative starting point we perform a simple evaluation by testing a two sided hypothesis that MPG from manual transmissions is better than from automatics, by supplying manual and auto transmission data as two samples to the R function **t.test** (default 95% confidence level):

diff in means	mean Manual	mean Auto	Test Statistic	P value	DF	low conf	high conf
7.245	24.392	17.147	3.767	0.00137	18.33	3.21	11.28

Our simple model’s prediction for MPG improvement between manual vs automatic transmission is the difference in the means, so we can predict an average improvement of **7.245 MPG** for manual transmission cars.

However this test ignores various other variables, so to account for sufficient of the variables involved in the MPG performance we need to perform some multivariate tests. We can use the step function of R to test all of the possible useful variations, and after evaluating the results suggest the best model. See Appendix 3 for results:

Model development

The suggested variables returned by **step** were cylinders, horsepower, weight, and transmission. After analysis of the results we re-introduced i/4 mile time (qsec) to test as an **alternative model**, as it seems help account for effects of high performance vehicles. The adjusted R squared score of the alternative model was only 0.01 better than the suggested step model, but as the estimate of transmission effect was **55% higher** and seems likely more accurate, it seems a better model. Comparisons of the residuals plots in Appendix 3 seem to indicate a better fit for our alternative model, the QQ plot is aligned along more points compared to the suggested model, the scale location regression line is flatter, and the residuals vs leverage regression line is flatter and centered closer to the zero intercept.

Some of the residuals exert significant leverage, the Maserati Bora has the highest performance specification of the manual transmission cars, while the Toyota Corona has the worst MPG of the 4 cylinder cars (though still better than any 6 or 8 cylinder cars). If we take out just the Maserati Bora from the data set then step produces what seems a more balanced **medium model** based just on weight, 1/4 mile time, and transmission and shows a **further trend** towards manual transmission being better than automatic transmission - column 4 of the comparative summary in Appendix 3.

Discarded variables

Many of the variables discarded by testing through step seemed to have significant effects, but were also variable in tandem with the key values of horsepower, weight and 1/4 mile time. For instance rear axle ratio (drat) was given a high score, but it seems to be confounded with the simultaneous low horsepower and weight of the Honda Civic, which also has the highest rear axle ratio of all of the cars, and achieves very good MPG. Therefore drat is noise, and is removed from from the model. Likewise for numbers of carburetors, they achieve a number of significant looking scores but are really linked to the performance of a vehicle better measured via horsepower.

Conclusion and MPG prediction

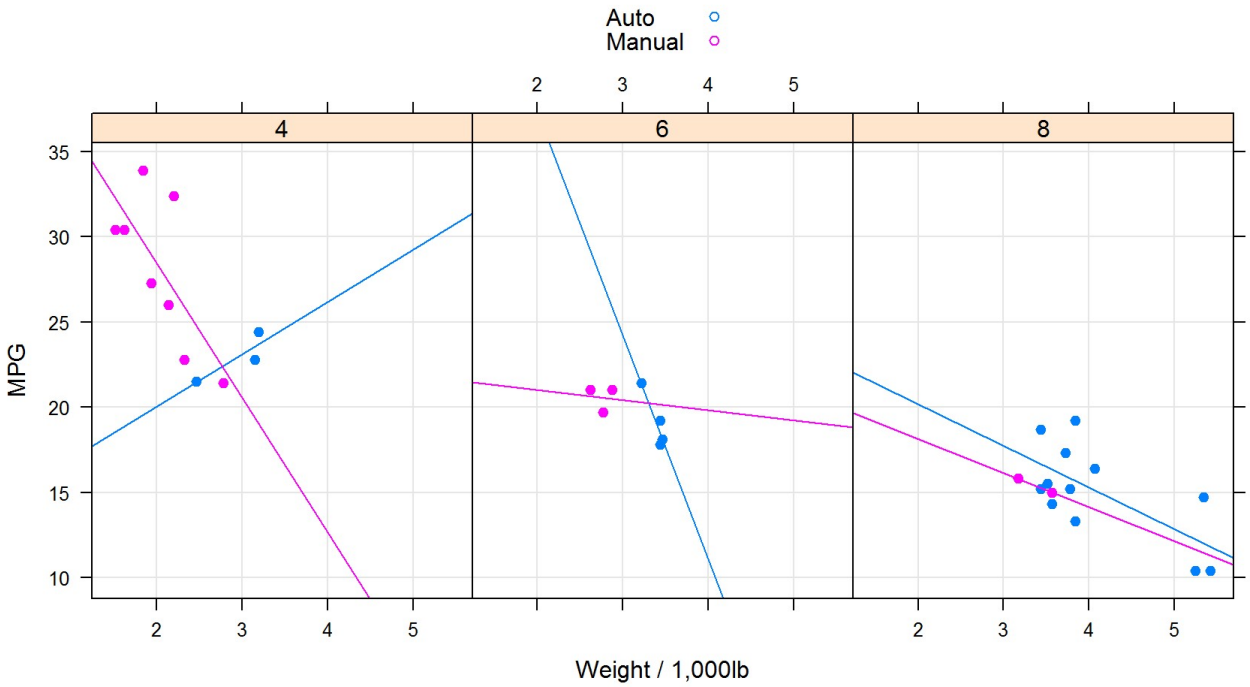
We have seen that a simple comparison of mean MPG between manual and automatic transmissions produces a rather high value for the predicted MPG effect, and have concluded that a multivariate regression analysis is required to take into account other variables beside transmission. We have taken the suggested variables which the step function determined as our first model, as an alternative we have added 1/4 mile time to try to better account for high performance vehicles, and finally we created a model that eliminates the extremely powerful Maserati Bora from the data. These 3 models suggest different values for MPG improvement from manual over automatic transmission, but applying **predict** to the models shows the range of possible values for the **alternative model** to be closest to the recorded figures (Appendix 4).

The range of possibilities available in the multivariate comparisons suggest an improvement in MPG performance for manual transmissions from between 1.8 to 3.2 MPG, with **2.8 MPG** as the most likely choice considering all of the vehicles, but **3.2 MPG** as most representative if the high performance Maserati Bora is excluded. Based on the adjusted R squared value the 3 models exhibit between **84 and 83 R squared** which we can use as a measure of certainty (Appendix 5).

Appendix 1 - Structure of mtcars table and MPG vs weight analysis

Field	Description	Field	Description
1. mpg	Miles/(US) gallon	7. qsec	1/4 mile time
2. cyl	Number of cylinders	8. vs	V/S (0 = V, 1 = Straight configuration)
3. disp	Displacement (cu.in.)	9. am	Transmission (0 = automatic, 1 = manual)
4. hp	Gross horsepower	10. gear	Number of forward gears
5. drat	Rear axle ratio	11. carb	Number of carburetors
6. wt	Weight (lb/1000)		.

Figure 1 - MPG vs Weight



Appendix 2 - data to assist with comparitive analysis

Figure 2 - Manual and Automatic cars

Manual Transmission

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	hp.wt
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1	35.42234
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1	30.00000
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2	32.19814
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2	74.68605
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1	34.10853

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	hp.wt
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2	42.52336
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1	40.08621
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2	39.20863
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4	38.26087
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4	41.98473
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6	63.17690
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4	83.28076
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8	93.83754

Automatic Transmission

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	hp.wt
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2	19.43574
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2	30.15873
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1	39.35091
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1	34.21462
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4	35.75581
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2	45.51365
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2	50.87209
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1	30.34682
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4	35.75581
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3	48.25737
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3	44.22604
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2	42.61364
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2	43.66812
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3	47.61905
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4	43.03087
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4	68.62745
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4	63.80208
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4	39.04762
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4	39.63864

Appendix 3 - modelling the effects of variables on MPG

Figure 3 - comparisons of model variables

Model Name	Variables	Description
model	lm(mpg ~ ., data=testCars)	Test all variables
stepmodel	cyl + hp + wt + am	Variables of the step model
altmodel	cyl + hp + wt + qsec + am	Variables of the alternative model
medmodel	wt + qsec + am	Variables of the medium model

	Dependent variable:			
	mpg			
	(1)	(2)	(3)	(4)
	model	stepmodel	altmodel	medmodel

(Intercept)	3.966 (37.306)	35.518 (2.032)***	24.409 (10.246)**	13.073 (6.170)**
6 cylinders	-2.057 (3.197)	-3.031 (1.407)**	-1.909 (1.730)	
8 cylinders	-0.316 (7.276)	-2.164 (2.284)	-0.227 (2.870)	
Displacement	0.036 (0.032)			
Gross horsepower	-0.154 (0.122)	-0.032 (0.014)**	-0.025 (0.015)	
Rear axle ratio	3.967 (4.604)			
Weight	-0.253 (6.454)	-2.497 (0.886)***	-2.963 (0.977)***	-3.821 (0.733)***
1/4 mile time	0.247 (0.965)		0.619 (0.560)	1.191 (0.296)***
V engine configuration	0.015 (3.970)			
Automatic transmission	0.213 (3.815)	-1.809 (1.396)	-2.833 (1.670)	-3.176 (1.471)**
4 forward gears	2.114 (4.101)			
5 forward gears	0.643 (4.607)			
2 carburetors	-2.884 (3.534)			
3 carburetors	0.816 (5.307)			
4 carburetors	-2.513 (6.731)			
6 carburetors	2.719 (6.929)			
8 carburetors	6.960 (8.506)			
Power to weight ratio	0.294 (0.407)			
Observations	32	32	32	31
R^2	0.897	0.866	0.872	0.848
Adjusted R^2	0.772	0.840	0.841	0.832
Residual Std. Error	2.879 (df = 14)	2.410 (df = 26)	2.400 (df = 25)	2.484 (df = 27)
F Statistic	7.166*** (df = 17; 14)	33.571*** (df = 5; 26)	28.420*** (df = 6; 25)	50.380*** (df = 3; 27)
Note:				*p<0.1;* ***p<0.05;*** p<0.01

Figure 4 - Residuals plots from the step model

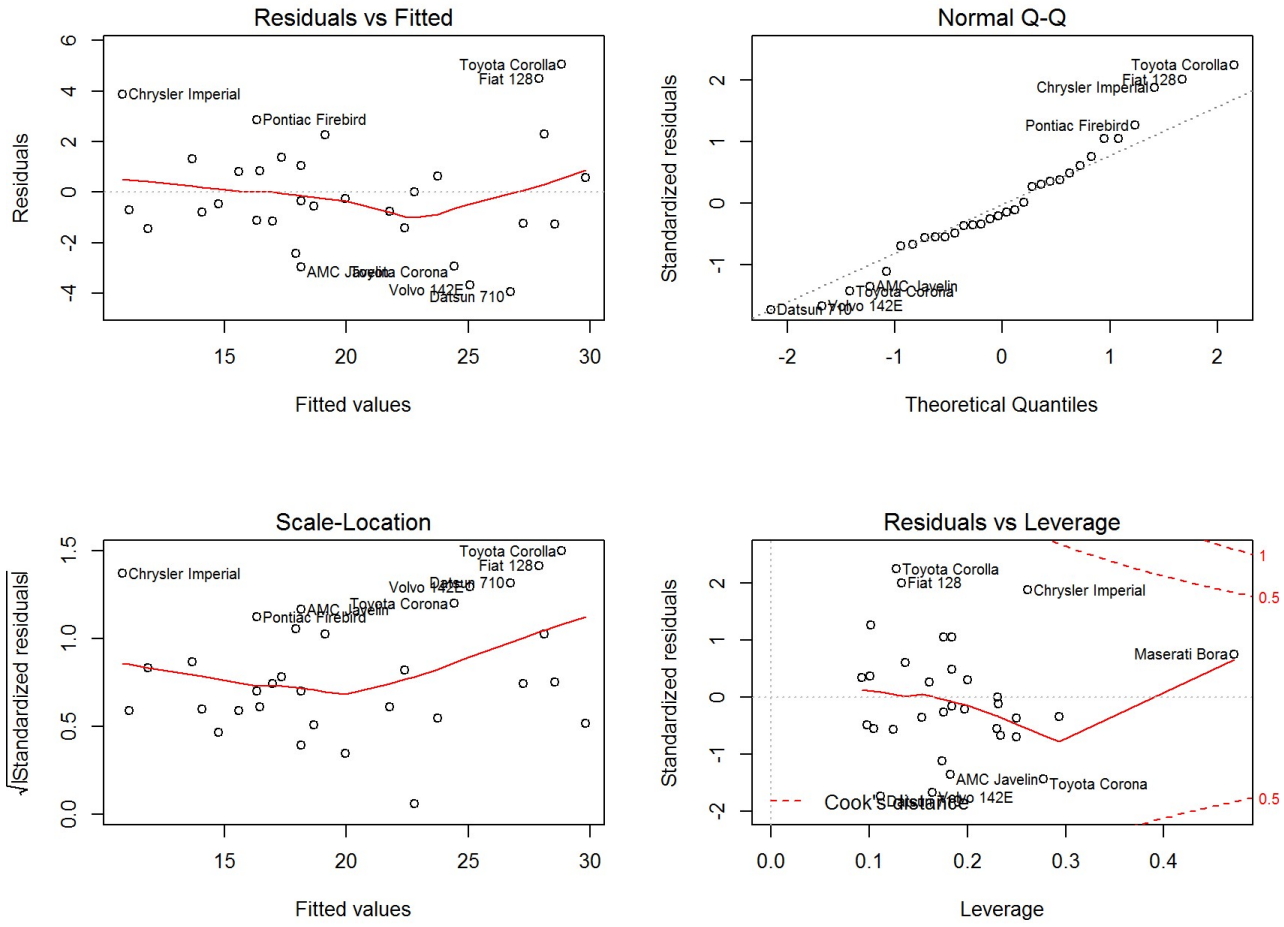
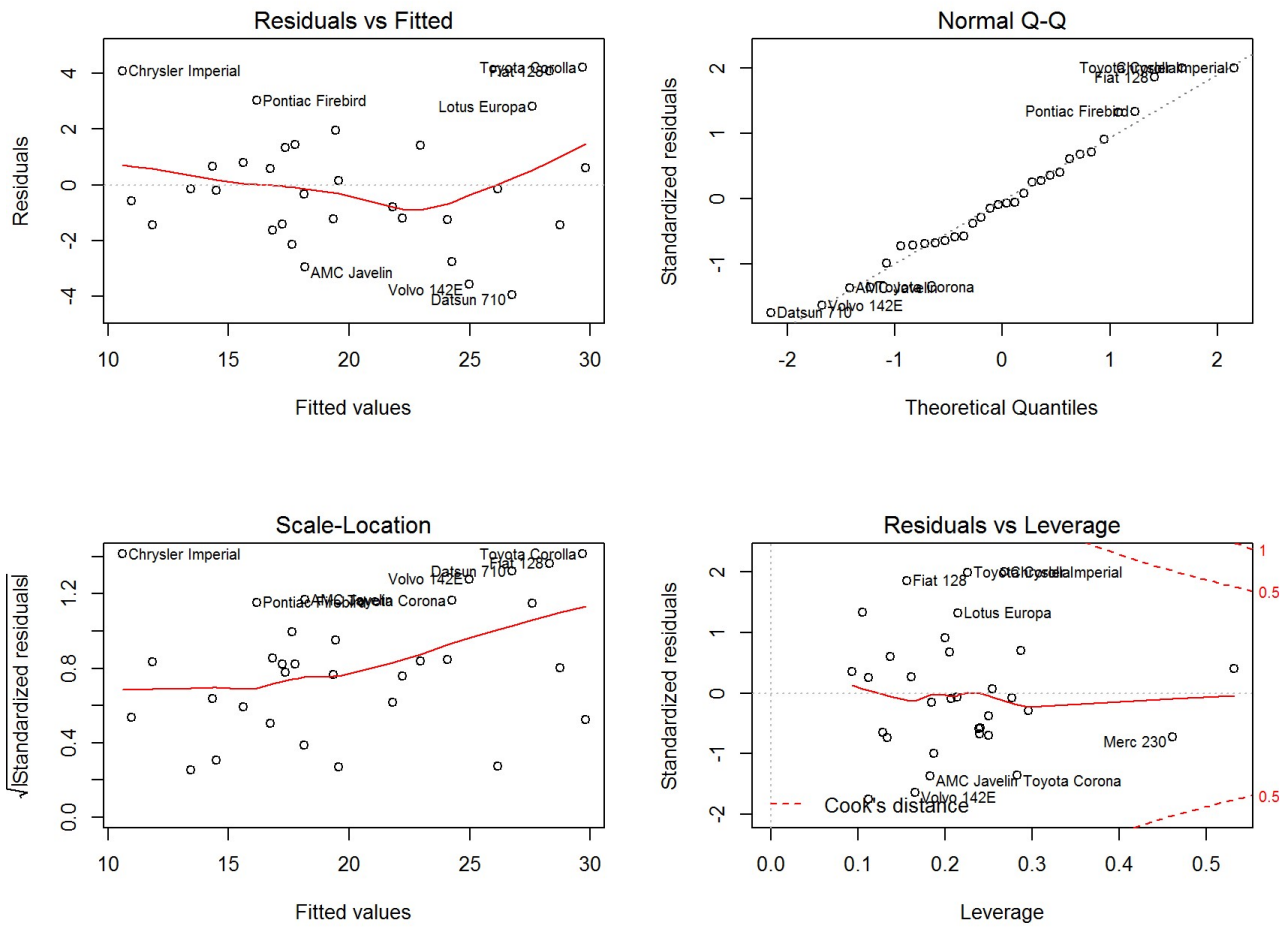


Figure 5 - Residuals plots from the alternative model, which includes qsec:



Appendix 4 - Toyota Corolla (actual MPG 33.9) - Predict and Statistics

Model	fit	lwr	upper	SE	DF	Residual Scale
stepmodel	28.84874	27.078182	30.61930	0.8613639	26	2.41012
altmodel	29.68133	27.330907	32.03176	1.1412400	25	2.399848
medmodel	29.75772	27.807767	31.70766	0.9503457	27	2.484144

Appendix 5 - summary of MPG values obtained

Model Name	Advantage manual transmission	Adjusted R squared
stepmodel	1.809 MPG	84.0
altmodel	2.833 MPG	84.1
medmodel	3.176 MPG	83.2

The regression table in appendix 3 was created using the stargazer library: http://CRAN.R-project.org/package=stargazer (http://CRAN.R-project.org/package=stargazer)

This document was created using R markdown and knitr to create an HTML file, which was then converted to PDF. For reasons of space most of the R code is not reproduced here, but the original R markdown file with the embedded R code is available:

https://github.com/Zohaggie/CourseraRegressionModels/blob/master/Regression_Models.Rmd (https://github.com/Zohaggie/CourseraRegressionModels/blob/master/Regression_Models.Rmd)