Is Fuel Consumption for Automatic Transmissions Higher?

Student 972147 / McReyar

Sunday, August 24, 2014

Executive Summary[1]

Using data from the 1974 Motor Trend US magazine the relation between fuel consumption and automatic transmission is examined. At first glance automatic transmission seem to use more fuel, however the correlation with several other variables is very high.
Therefore the based regression model is evaluated based on the Akaike information criterion (AIC) by using forward and backward selection. Finally interaction terms are included and as those models don't use transmission type, there is not enough evidence to conclude that the transmission type is the cause for higher gasoline consumption.

Exploratory Data Analysis

Following variables are available in the dataset:

Variable	Description
mpg	Gasoline Milage (MPG)
cyl	Number of Cylinders
disp	Engine Size (Cubic Inches)
hp	Horespower
drat	Final Drive Ratio
wt	Weight (in 1000 lbs)
qsec	Quarter Mile Time (Seconds)
vs	Engine Shape (V, Straight)
am	Transmission Type (Automatic, Manual)
gear	Number of Transmission Speeds
carb	Number of Carburetor Barrles

Table 1: Data Description

As can be seen in following boxplot, it looks like automatic transmission types have lower gasoline mileage than manual ones:

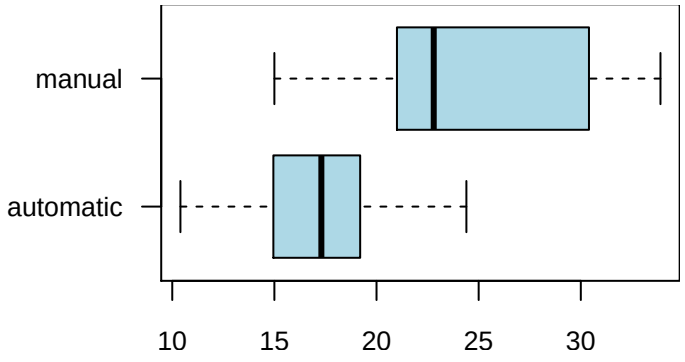

Figure 1: Gasoline Mileage by Transmission Type

However they are also highly correlated with other variables (especially number of transmission speeds (0.79), final drive ratio (0.71), weight and (-0.69)), which can be seen in Figure 3. As single variables weight (-0.87), engine size (-0.85), number of cylinders (-0.85) and horespower (-0.78) seem to have the highest impact on fuel consumption, but again all of these variables are highly correlated with each other.
This is reflected in the variance inflation factors - the increase in standard deviation for the respective regressor can be seen in following table:

am	cyl	disp	hp	drat
2.16	3.92	4.65	3.14	1.84

wt	qsec	vs	gear	carb
3.89	2.74	2.23	2.31	2.81

Table 2: Increase in Standard Deviation

Regression Models

To better estimate the impact of the variables on gasoline mileage, the best model for each number of predictors is chosen based on Akaike information criterion (AIC). First a backward selection approach is taken:

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			21.00	147.49	70.90
- cyl	1.00	0.08	22.00	147.57	68.92
- vs	1.00	0.27	23.00	147.84	66.97
- carb	1.00	0.69	24.00	148.53	65.12
- gear	1.00	1.56	25.00	150.09	63.46
- drat	1.00	3.34	26.00	153.44	62.16
- disp	1.00	6.63	27.00	160.07	61.52
- hp	1.00	9.22	28.00	169.29	61.31

Table 3: Backward Selection

This results in model based on weight, quarter mile time and transmission type that has R^2 0.8497 which means that 84.97% of the variance of the gasoline mileage can be explained.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.62	6.96	1.38	0.18
ammanual	2.94	1.41	2.08	0.05
wt	-3.92	0.71	-5.51	0.00
qsec	1.23	0.29	4.25	0.00

Table 4: Best Model Based on Backward Selection

Based on this model, we are 95% confident that cars with an automatic transmission drive get 0.05 to 5.83 miles less out of a gallon with everything else beeing equal.
By forward selection, following predictors are chosen:

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			31.00	1126.05	115.94
+ wt	-1.00	847.73	30.00	278.32	73.22
+ cyl	-1.00	87.15	29.00	191.17	63.20
+ hp	-1.00	14.55	28.00	176.62	62.66

Table 5: Forward Selection

[1]This report is written as course project for Regression Models (regmods-004) taught by Prof. Brian Caffo, PHD on Coursera. For reproducibility the RMD-file is available on GitHub.

For this model, transgression type isn't considered at all, but it still explains 84.31% of the variability (R^2=0.8431).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.75	1.79	21.69	0.00
wt	-3.17	0.74	-4.28	0.00
cyl	-0.94	0.55	-1.71	0.10
hp	-0.02	0.01	-1.52	0.14

Table 6: Best Model Based on Forward Selection

Although number of cylinders and horse power are included in the model, they don't seem to be significant.
As weight is used in both models it can be assumed that it has an impact on gasoline mileage. What is more it would be logical that the influence of horse power is bigger if the car weights more. Therefore adding an interaction term $wt * hp$ is examined.

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			20.00	98.04	59.83
- vs	1.00	0.03	21.00	98.07	57.84
- disp	1.00	0.12	22.00	98.19	55.88
- am	1.00	0.65	23.00	98.84	54.09
- drat	1.00	1.38	24.00	100.22	52.53
- carb	1.00	4.81	25.00	105.03	52.03

Table 7: Backward Selection with Interaction Term

With the interaction term transgression type isn't used in the best model evaluated by backward selection, which explains 90.67% of the variability (R^2=0.9067)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.01	12.08	1.91	0.07
cyl	0.98	0.76	1.29	0.21
hp	-0.13	0.03	-4.21	0.00
wt	-9.29	1.68	-5.54	0.00
qsec	0.94	0.46	2.04	0.05
gear	1.84	0.94	1.95	0.06
hp:wt	0.03	0.01	4.01	0.00

Table 8: Best Model Based on Backward Selection with Interaction Term

Number of cylinders, number of transmission speeds and quarter mile time have a p-value over 0.05, but the interaction between weight and horse power seems to be a good predictor. With using forward selection following variables are chosen as predictors:

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
			31.00	1126.05	115.94
+ wt	-1.00	847.73	30.00	278.32	73.22
+ cyl	-1.00	87.15	29.00	191.17	63.20
+ hp	-1.00	14.55	28.00	176.62	62.66
+ hp:wt	-1.00	49.29	27.00	127.33	54.19

Table 9: Forward Selection with Interaction Term

Although this model uses only 4 predictors, it can still explain 88.69% of the variability (R^2=0.8869).

The interaction term again has a very low p-value and is therefore highly significant.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.49	3.66	13.51	0.00
wt	-7.63	1.52	-5.01	0.00
cyl	-0.37	0.51	-0.72	0.48
hp	-0.11	0.03	-3.64	0.00
wt:hp	0.03	0.01	3.23	0.00

Table 10: Best Model Based on Forward Selection with Interaction Term

To choose the best model, the model without interaction term that includes transmission type, as well as both models with interaction term are compared with ANOVA:

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
28	169.29				
27	127.33	1	41.96	9.99	0.0041
25	105.03	2	22.30	2.65	0.0901

Table 11: Model Comparison with Anova

The model with 6 predictors has a p-value that is larger than 0.05 and therefore isn't used. Hower there is enough evidence to take the model that uses weight, number of cylinders, horsepower and the interaction between weight and horsepower over the model which uses transmission type.

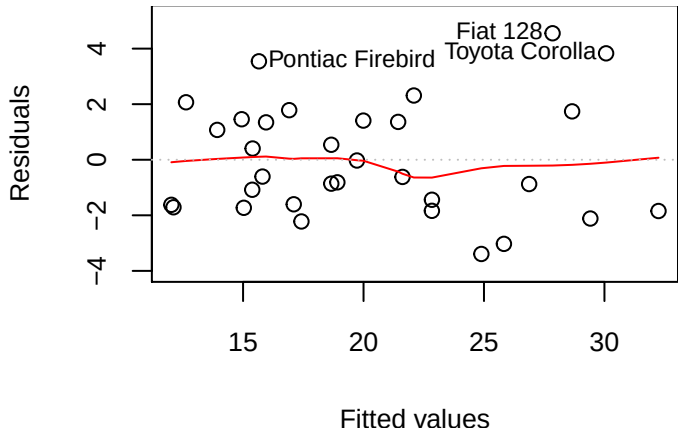

Figure 2: Residuals vs Fitted

The variance of the residuals is approximately constant, and as can be seen in Figure 4 the residuals are roughly normal distributed and there are no outliers that have high influence.

Conclusion

Based on this analysis gasoline mileage is affected by weight (4.51 to 10.75 less mpg per 1000lbs)[2], horsepower (0.05 to 0.17 less mpg per hp) and the product of weight and horse power (0.01 to 0.04 more mpg per $wt * hp$). This means that heavier cars actually are more economic when they have more horse power which could have to do with the fact that the motor doesn't run at its limit.
As far as transgression type is concerned, there is not enough evidence to conclude that it has an impact on gasoline mileage. This doesn't mean that there definitely is no impact, but based on this data no relation could be found.

[2]basd on a 95% confidence interval with everything else beeing equal (this applies to all numbers in this paragraph)

Appendix

Correlation of Variables

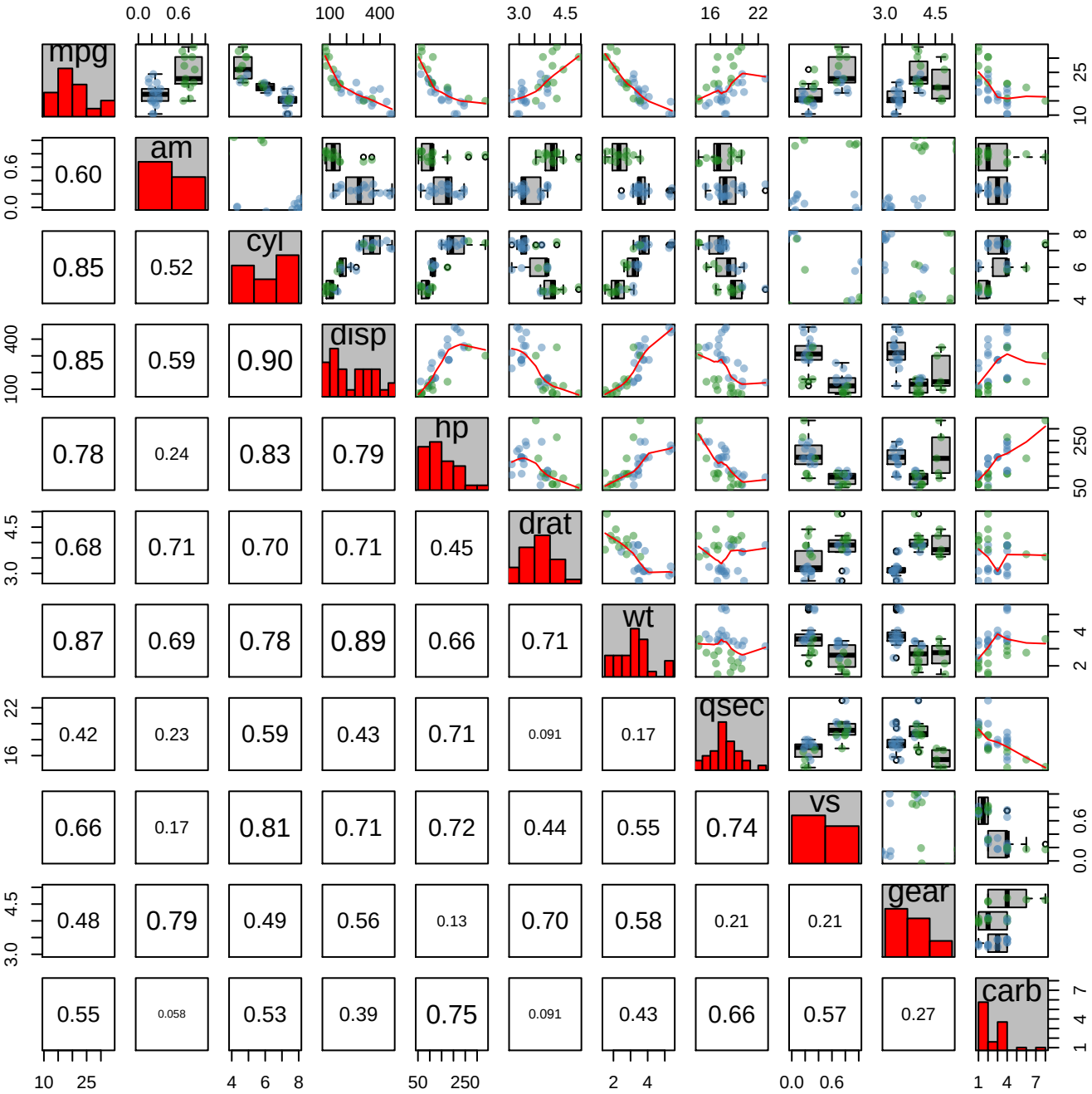

Figure 3: Correlation between Variables

Diagnostics for Linear Regression

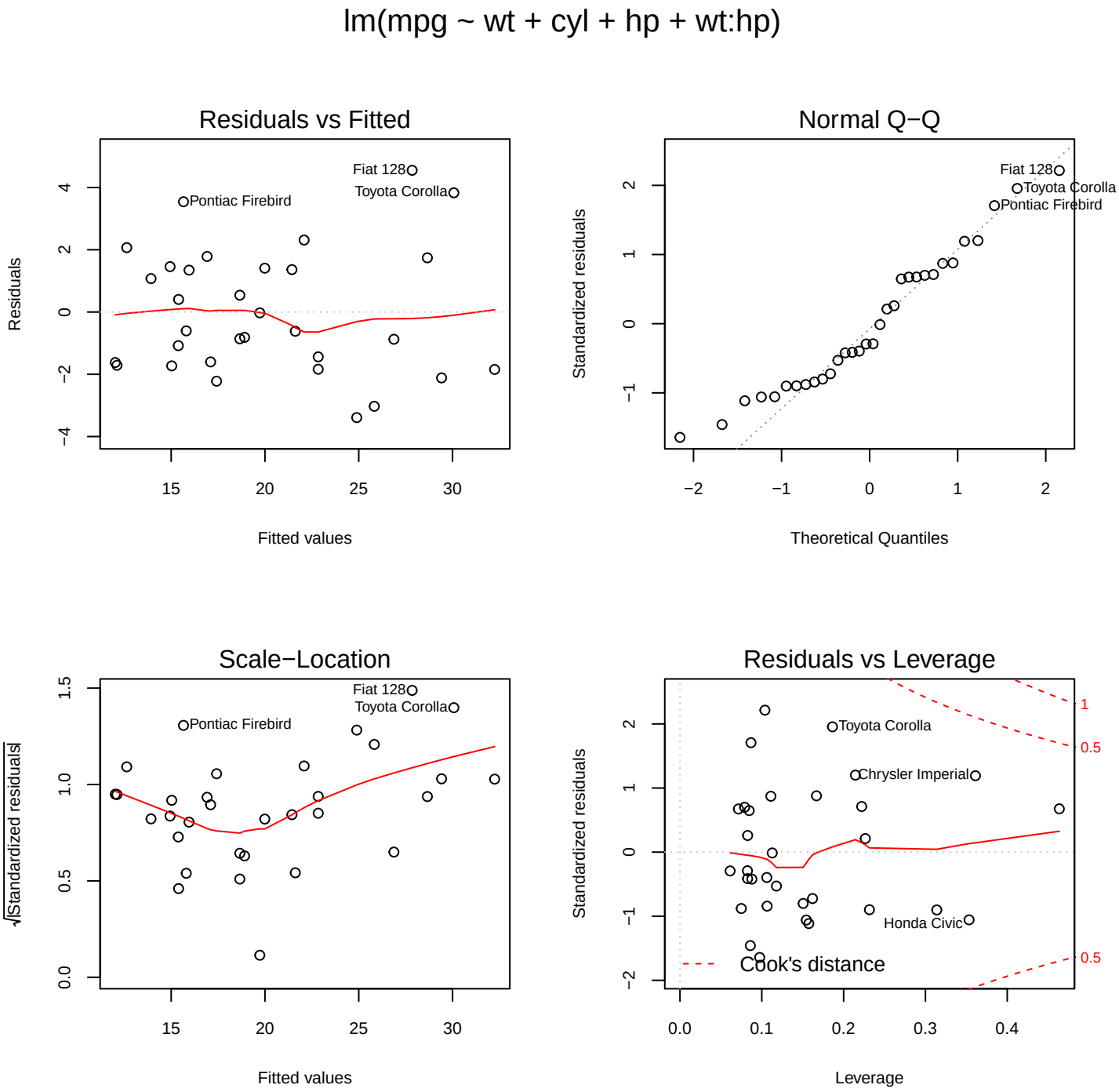

Figure 4: Diagnostics for Linear Regression