Statistical Evidence on Manual versus Automatic Cars for better Fuel Consumption

Mohammed K. Barakat

July 31, 2015

Executive Summary

By analyzing a dataset of a collection of cars (*mtcars*), this study explores the relationship between miles per gallon (MPG) feature and a set of other car features. We are particularly interested in finding out if an automatic or a manual transmission is better for MPG. The study proves preference of transmission type and quantifies the difference.

The study uses the *mtcars* dataset and employs several statistical techniques to reach to a robust conclusion. In summary, the study concluded that using manual-transmission cars is better than automatic for MPG. Besides, MPG has a statistically significant relationship with car weight and quarter mile time (acceleration).

About the *mtcars* dataset

mtcars dataset was extracted from the 1974 *Motor Trend* US magazine. It comprises fuel consumption and 10 other aspects of automobile features for 32 automobiles. It can be downloaded from R datasets library.

```
library(datasets)
data("mtcars")
```

The dataset consists of 32 observations for different automobiles. *See Appendix: Fig.1 Data description* to know the variables with their descriptions. Besides, the *Appendix (Fig.2 A snapshot of data observations)* shows the first six rows of the data.

Exploratory Data Analysis (EDA)

Boxplot and Pair-Panel plot are two EDA tools used to explore data properties and find possible patterns or correlations. A boxplot is used to see the variation of MPG across both types of transmission. On average, using the manual type yields higher mpg compared to automatic. *(See Appendix: Fig.3 Boxplot)*. The Pair Panel Plot gives information about the correlation between pairs of variables. E.g. the resulted plot shows a negative correlation between mpg and weight. *(See Appendix: Fig.4 Pair Panel Plot)*

Inference about MPG in relation to automatic and manual cars using Hypothesis Testing

A statistical evidence is still required to verify difference between means of manual and automatic cars. We assumed the null hypothesis of no difference in means. Whereas the alternative hypothesis assumes difference in means.

mean.Auto	mean.Manual	t.statistic	p.value	LCL	UCL
17.147	24.392	-3.767	0.001	-11.28	-3.21

Using t.test the p-value is **0.001** (<0.05 α error rate), which provides a statistically significant difference in means where the manual mpg mean (**24.39**) is higher than that of the automatic (**17.15**).

Exploring effect of other variables on MPG using Multivariable Linear Regression

MPG may be affected by other regressors (variables). Hence, modelling the MPG vs transmission type should be tested by adjusting for other variables in the model. We will fit multiple models and select the best one. Our **Model Selection strategy** goes in the following steps:

1. Create the initial regression model including all regressors

```
f0<-lm(mpg~factor(am)+factor(cyl)+disp+hp+drat+wt+qsec+factor(vs)+gear+carb,data=pmtc
ars)
```

2. Perform preliminary screening to select the potential significant regressors using the Stepwise Regression method

Stepwise Regression reduces the number of input variables to those significant ones using a specific algorithm. We will use the *Backward* approach which starts with all variables, tests the effect on the model by deleting each variable, then deletes the variable that improves the model the most. This process is repeated until no further improvement is possible. **The stepwise method will be used only for initial screening of variables.**

After running the Stepwise method the three variables (Transmission, wt, and qsec) seem to have significant effect on MPG. (p-values are **4.67e-02**, **6.95e-06**, and **2.16e-04**, respectively (<0.05 error rate). *(See Appendix: Fig.5 Stepwise Regression results)*

3. Test the preliminary model against other models using the Nested Model Testing method

So far we have a preliminary model of MPG versus transmission type, wt, and qsec. This model needs to be tested against other models by adjusting for other variables. 10 models are created by adjusting for a new additional variable in each model. Based on the Nested modelling we can confirm that the model of including *wt* and *qsec* remains significant by comparing its p-value to those of other models. This model (fit3- third fit from the top in the Appendix) gets a p-value of **0.00063431**. *(See Appendix: Fig.6 (Nested Model Testing) for the entire nested fits results)*

4. Confirm validity of the selected model by checking specific parameters

4.1 Low Variance Inflation Factors (VIF)

One way to measure multicollinearity is through the Variance Inflation Factor (VIF). The lower the VIF, the better the model is. VIF of each of the three regressors are all below 5, which confirms absence of multicollinearity.

```
library(car);VIF.value<-round(vif(fit3),3);VIF.value
```

```
## factor(am)         wt       qsec
##      2.541      2.483      1.364
```

4.2 Low Standard Error, significant p-value, and high R-squared of the model

The last step to confirm model validity is by testing if the model has the lowest variation around the fitted line (residual standard error), most significant model (lowest p-value), and the highest ratio of explained variation (Adjusted R-squared) compared to other models. The results show that "fit3" is the best fit with optimum values of p-value = **1.2104e-11**, Residual Standard Error = **2.459**, and R-Squared = **0.834**. *See Appendix: Fig.7 (Fits parameters) for the entire table of fits parameters*.

Interpreting the final model

```
kable(summary(fit3)$coefficients,align = 'c')
```

The coefficients table *(Appendix: Fig.8 Final model coefficients)* of the selected model (fit3) shows that the three regressors (Transmission type, wt, and qsec) are all significant in affecting the output (mpg) where p-values are all < 0.05. Besides, the table shows that on average, automatic cars have **9.618** mpg fuel consumption. Whereas, manual cars are **2.936** mpg higher than that of automatic cars. Besides, MPG decreases by **3.917** for an increase of 1000 lb in weight (wt). Whereas, MPG increases by **1.226** for an increase of one unit acceleration (qsec).

Model statistics and confidence intervals

```
ci<-confint(fit3,level=0.95);kable(ci,align = 'c')
```

Based on the CI results we can say that 95% of the time MPG of manual cars will be **0.046** higher than that of automatic cars at minimum and **5.826** higher than that of automatic cars at maximum. *See Appendix: Fig.9 (Model statistics and CI) for confidence intervals of each of the significant variables.*

Model residual plots and diagnostics

Model diagnostics using Variance Inflation Factor (VIF)

As explained earlier the VIF of each of the three regressors is below 5. (Transmission = 2.541, wt = 2.483, and qsec = 1.364).

Residuals, leverage, and normality plots

Both *Residual vs Fitted* and *Residual vs Leverage* plots *(See Appendix: Fig.10 Residuals, leverage, and normality plots)* show no specific patterns, and residuals are symmetrical around zero and, hence, randomly distributed.

The points of the model Q-Q Plot lie pretty close to the dashed line which implies good normality of residuals. The Cook's distance plot shows how individual observations can influence the estimated regression coefficients of the model.

Conclusion and answers to questions raised by the study (with 0.05 error rate of uncertainty)

- Our Hypothesis Testing showed that manual transmission is better for MPG than automatic where the MPG mean is (**24.39**) for manual and (**17.15**) for automatic type.

- When the model is adjusted for other variables weight (wt) and acceleration (qsec) proved significant in affecting the MPG vs Transmission relationship. The final model showed that, on average, automatic cars have **9.618** mpg, whereas manual cars are **2.936** mpg higher than that of automatic cars with a confidence interval for MPG difference of (**0.046**, **5.826**) using 95% confidence level (0.05 α error rate). Hence, manual transmission cars are still better than automatic for MPG.

APPENDIX

Fig.1 Data description

Var	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time (quarter mile time (acceleration))
vs	V/S (V-engine/Straight engine) (0/1)
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Fig.2 A snapshot of data observations

For better readability the 0/1 levels for factor variables are converted into texts.

```
kable(head(pmtcars),align = 'c')
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	V	Manual	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	V	Manual	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	S	Manual	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	S	Auto	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	V	Auto	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	S	Auto	3	1

Fig.3 Boxplot

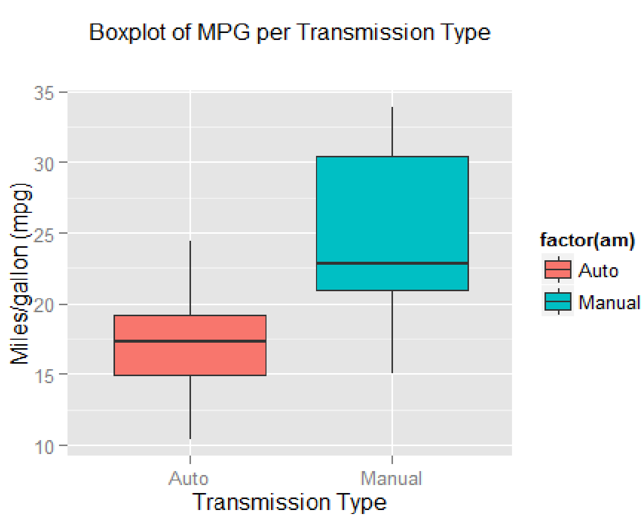


Fig.4 Pair Panel Plot

```
pairs(mtcars,main = "Pair Panel - Mtcars variables", panel=panel.smooth,upper.panel =
NULL)
```

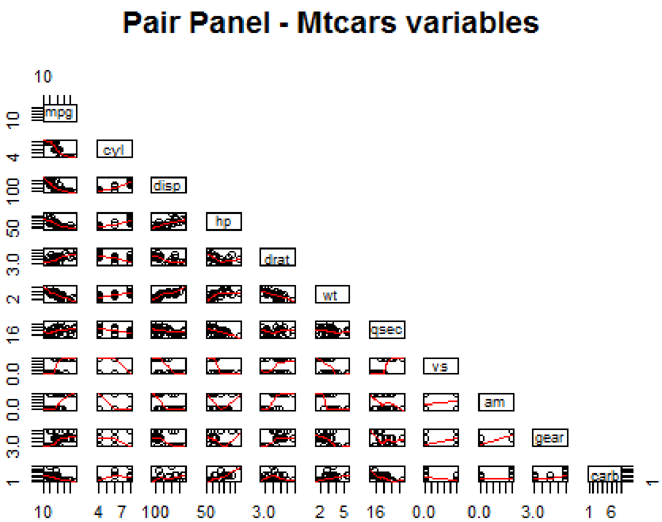


Fig.5 Stepwise Regression results

```
library(MASS);step <- stepAIC(f0, direction="backward", trace=FALSE)
kable(summary(step)$coeff,align = 'c')
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
factor(am)Manual	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

Fig.6 Nested Model Testing

```
fit1 <- lm(mpg ~ factor(am), data = pmtcars)
fit2 <- lm(mpg ~ factor(am)+wt, data = pmtcars)
fit3 <- lm(mpg ~ factor(am)+wt+qsec, data = pmtcars)
fit4 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl), data = pmtcars)
fit5 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp, data = pmtcars)
fit6 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp+hp, data = pmtcars)
fit7 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp+hp+drat, data = pmtcars)
fit8 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp+hp+drat+factor(vs), data = pmtca
rs)
fit9 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp+hp+drat+factor(vs)+gear, data =
pmtcars)
fit10 <- lm(mpg ~ factor(am)+wt+qsec+factor(cyl)+disp+hp+drat+factor(vs)+gear+carb, d
ata = pmtcars)

nested<-anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10)
kable(nested,align = 'c')
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
30	720.8966	NA	NA	NA	NA
29	278.3197	1	442.576902	66.3914559	0.0000001
28	169.2859	1	109.033768	16.3562774	0.0006343
26	159.4244	2	9.861565	0.7396722	0.4898800
25	157.7339	1	1.690499	0.2535936	0.6200576
24	142.3306	1	15.403276	2.3106626	0.1441415
23	141.2059	1	1.124688	0.1687157	0.6856232
22	139.0230	1	2.182858	0.3274530	0.5735394
21	135.2706	1	3.752430	0.5629063	0.4618276
20	133.3235	1	1.947162	0.2920960	0.5948487

Fig.7 Fits parameters

Fit	pv	sdErr	adjRsq
fit3	1.2104e-11	2.459	0.834
fit4	3.0067e-10	2.476	0.831
fit2	1.5788e-09	3.098	0.736
fit5	1.5837e-09	2.512	0.826
fit6	2.5657e-09	2.435	0.837
fit7	1.2059e-08	2.478	0.831
fit8	4.8142e-08	2.514	0.826
fit9	1.5991e-07	2.538	0.823
fit10	5.7224e-07	2.582	0.816
fit1	0.00028502	4.902	0.338

Fig.8 Final model coefficients

```
kable(summary(fit3)$coefficients,align = 'c')
```

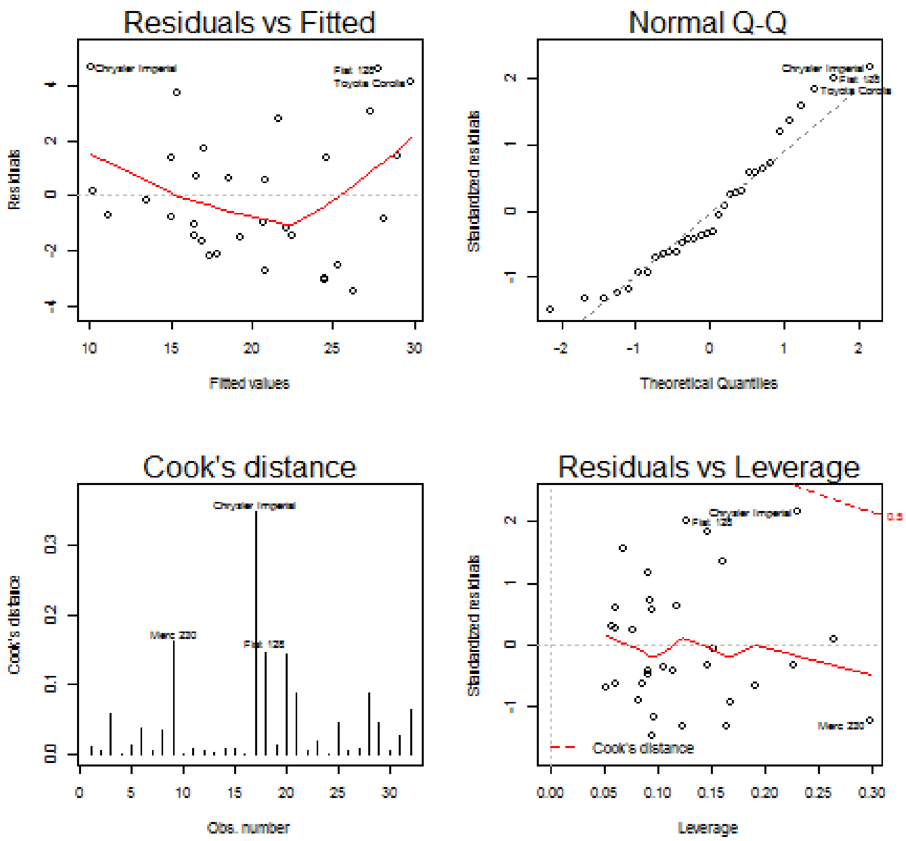
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.617781	6.9595930	1.381946	0.1779152
factor(am)Manual	2.935837	1.4109045	2.080819	0.0467155
wt	-3.916504	0.7112016	-5.506882	0.0000070
qsec	1.225886	0.2886696	4.246676	0.0002162

Fig.9 Model statistics and CI

	2.5 %	97.5 %
(Intercept)	-4.6382995	23.873860
factor(am)Manual	0.0457303	5.825944
wt	-5.3733342	-2.459673
qsec	0.6345732	1.817199

Fig.10 Residuals, leverage, and normality plots

```
par(mfrow = c(2, 2),cex=.5);plot(fit3,which=c(1,2,4,5))
```



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Cook's distance
Obs. number
Merc 230
Residuals vs Leverage
Leverage
0.5