Analysis of Automatic vs. Manual Transmission Affect on MPG

mail@darrenbishop.com

28 February 2016

Executive Summary

This analysis aimed to answer the question of which transmission, automatic or manual, is better for mpg performance of cars, quantifying and validating any difference.

The analysis suggests the choice of automatic vs. manual transmission has ***no significant*** affect on mpg performance.

Despite a ***0.15 mpg difference in favour of manual-transmission***, the ***-2.518 to 2.818 confidence interval is centred at 0-zero***, strongly suggesting there is no effect.

Two significant covariates were identified, both having confidence interval wholly below 0-zero:

Weight A continuous variable; a one unit change representing 1000 lbs
    Per unit increase, MPG performance is reduced by 3.15 mpg

No. of Cylinders A discrete numeric variable; values of 4, 6 and 8 cylinders
    A 6 cylinder engine reduces MPG performace by 4.23 mpg, compared to a 4 cylinder engine
    An 8 cylinder engine reduces MPG performace by 6.08 mpg, compared to a 4 cylinder engine
    An 8 cylinder engine reduces MPG performace by 1.82 mpg, compared to a 6 cylinder engine

If buying from a similar set of cars i.e. having the same no. of cylinders, pick the lighter one to get better MPG performance!

Exploratory Data Analysis

The documentation for the `mtcars` dataset (see Aappendix) reveals that the `vs` and `am` variables are binary and `cyl`, `gear` and `carb` variables are discrete; they can all thus be represented as factor variables.

`vs` and `am` are re-labelled and also renamed to `engine` and `trans`, respectively; Table 2. of the Appedix shows a subset of the `mtcars` data with engine (`engine`) and transmission (`trans`) presented with self-describing levels.

Figure 1. of the Appendix shows a preliminary analysis, capturing the distribution of cars' `mpg` performance, grouped by transmission.

It appears that more than 75% of automatic cars get less than 20 mpg, while more than 75% of manual cars perform better than 20 mpg, suggesting manual cars are better for fuel efficiency.

Model Design

Considering how cars are constructed and what they are used for, the available variables can be placed into two groups, *performance metrics* and *mechanical properties*, which in statistics terms translates to *response* and *regressor* variables, respectively.

Performance Metrics `mpg`, `hp` and `qsec`
    Intuition: these are derived, that is, they can not be changed directly, but indirectly through changes to the *mechanical properties*

Mechanical Properties `cyl`, `disp`, `drat`, `wt`, `engine`, `trans`, `gear` and `carb`
    Intuition: these are structural aspects to be built into a car, analogous to the no. of bedrooms built into a house

Any correlation between *performance metrics* would likely be correlated and confounded by one of the *mechanical properties*.

Thus for the sake of parsimony and simplified design, this analysis disregards all *performance metrics* when considering regressor selection.

ANOVA Testing

Table 3. of the Appendix shows the results of running the *mechanical properties* through a step-wise-inclusion Analysis of Variance test; 0.05 Type-I Error rate is used for determining significance:

Model 1. contains the regressor of interest, transmission (`trans`)

Model 2. introduces Weight (`wt`) with a significant P-value of 0.0000012

We reject the Null-Hypothesis that Weight does not explain any variance in the model

Model 4. introduces No. of Cylinders (`cyl`) with a significant P-value of 0.0468712

We reject the Null-Hypothesis that No. of Cylinders does not explain any variance in the model

All interim and subsequent inclusions produced large P-values, suggesting including them would result in an over-fitted model.

With so few regressor variables as candidates for inclusion, there is the risk of producing a biased model. Variance Inflation Factor analysis and Regression Diagnostics are used to address this.

Variance Inflation Factor Analysis

Tables 4. and 5. of the Appendix shows the result of VIF anaylsis, where all 8 *mechanical properties* and just the 3 candidate variables (`trans`, `wt` and `cyl`) are included, respectively.

Table 4. shows large standard error inflation across the board, suggesting the presence of strongly correlated regressors; Table 5. shows much diminished standard error inflation in comparison for `trans`, `wt` and `cyl`.

Regression Diagnostics

To support the selection of `trans`, `wt` and `cyl` as regressor variables, Figure 2. of the Appendix gives a: Residual plot showing the abscence of pattern, which would otherwise indicate a poor model fit; and a Q-Q plot showing the standardized residuals are approximately Normal with mean zero, meaning there is no remaining variance still to be accounted for (by some excluded confounder).

Data Interpretation

With these *ANOVA Test* findings, validated by the *Variance Inflation Factor Analysis* and *Regression Diagnostics*, the final model is:

$$\mathtt{Model} : mpg \sim \mathtt{trans} + wt + \mathtt{cyl}$$

Table 1: Coefficients and 95% Confidence Intervals

	Coefficients	2.5 %	97.5 %
(Intercept)	33.7535920	27.980801	39.526382
transManual	0.1501031	-2.517734	2.817941
wt	-3.1495978	-5.012761	-1.286434
cyl6	-4.2573185	-7.152943	-1.361695
cyl8	-6.0791189	-9.533813	-2.624425

Knowing that both `trans` and `cyl` are factors of 2 and 3 levels each, respectively, 4 parallel regression lines are expected i.e. 4 different intercepts, all with the same slope determined by the coefficient of `wt`; note that the first represents automatic with 4 cylinders.

Table 1. shows a model intercept of 33.7535920, which as mentioned, represents cars with automatic transmission and 4 cylinder engines.

According to this model, the difference in switching from an automatic to a manual transission car, also with a 4 cylinder engine, is merely an extra 0.15 mpg

While keeping weight and transmission fixed, going for a bigger engine i.e. 6 cylinders or 8 cylinders, reduces mpg performance by 4.23 and 6.08 each, respectively.

It is intuitive that, while all other variables are kept the same, a heavier vehicle needs more energy/power to move, thus burning more fuel; this model indicates that per 1000 lbs increase in weight, the mpg performance is reduced by 3.15.

Conclusions

This analysis shows the choice of automatic vs. manual is somewhat moot, where the 95% confidence interval contains 0-zero i.e. likely no effect on mpg.

This finding contradicts the preliminary analysis, which suggested that manual cars exhibited far superior mpg performance.

The deeper analysis shows there are stronger covariates than transmission e.g. weight and # of cylinders, that better explain the variance in mpg.

Appendix

`mtcars` Documentation Excerpt

```
A data frame with 32 observations on 11 variables.

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

V/S (`vs`) is taken to mean *V-Shape* or *Straight* engine; to be consistent with Transmission (`am`), 'V-Shape' is assumed to be 0 and 'Straight' as 1.

Post-processed `mtcars` Sample

Table 2: `mtcars` sample, with renamed and relabelled engine and trans.

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

Exploratory Analysis Plots

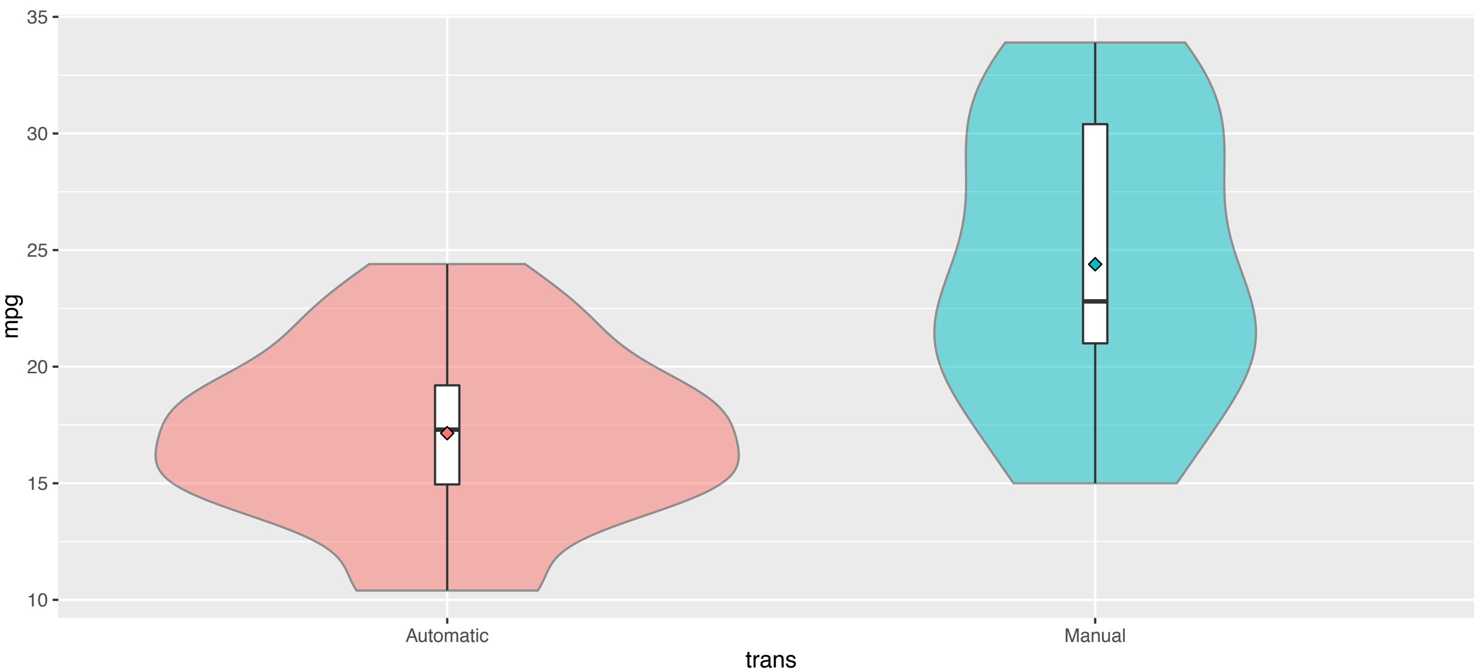

Figure 1: A Violin plot enriched with mean, median and inter-quartile range for transmission

Analysis of Variance Test Table

Table 3: ANOVA Test of the 8 *Mechanical Properties*

Model	Formula	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	mpg ~ trans	30	720.8966	NA	NA	NA	NA	
2	mpg ~ trans + wt	29	278.3197	1	442.5769020	50.6604204	0.0000012	***
3	mpg ~ trans + wt + disp	28	246.5563	1	31.7634352	3.6358630	0.0726380	.
4	mpg ~ trans + wt + disp + cyl	26	182.8693	2	63.6869319	3.6450239	0.0468712	*
5	mpg ~ trans + wt + disp + cyl + carb	21	163.8593	5	19.0100237	0.4352038	0.8181389	
6	mpg ~ trans + wt + disp + cyl + carb + engine	20	161.5583	1	2.3010305	0.2633919	0.6140409	
7	mpg ~ trans + wt + disp + cyl + carb + engine + gear	19	161.5045	1	0.0537513	0.0061528	0.9383439	
8	mpg ~ trans + wt + disp + cyl + carb + engine + gear + drat	18	157.2507	1	4.2538708	0.4869275	0.4942172	

Variance Inflation Factor Analysis Tables

Table 4: Variance and Standard Deviation Inflation Factors

	trans	wt	cyl	disp	drat	engine	gear	carb
σ^2	5.867926	14.520038	57.210000	35.912628	6.367638	6.060509	6.286963	78.242257
σ	2.422380	3.810517	7.563729	5.992715	2.523418	2.461810	2.507382	8.845465

Table 5: Variance and Standard Deviation Inflation Factors

	trans	wt	cyl
σ^2	1.925620	3.611208	2.585745
σ	1.387667	1.900318	1.608025

Regression Diagnostic Plots

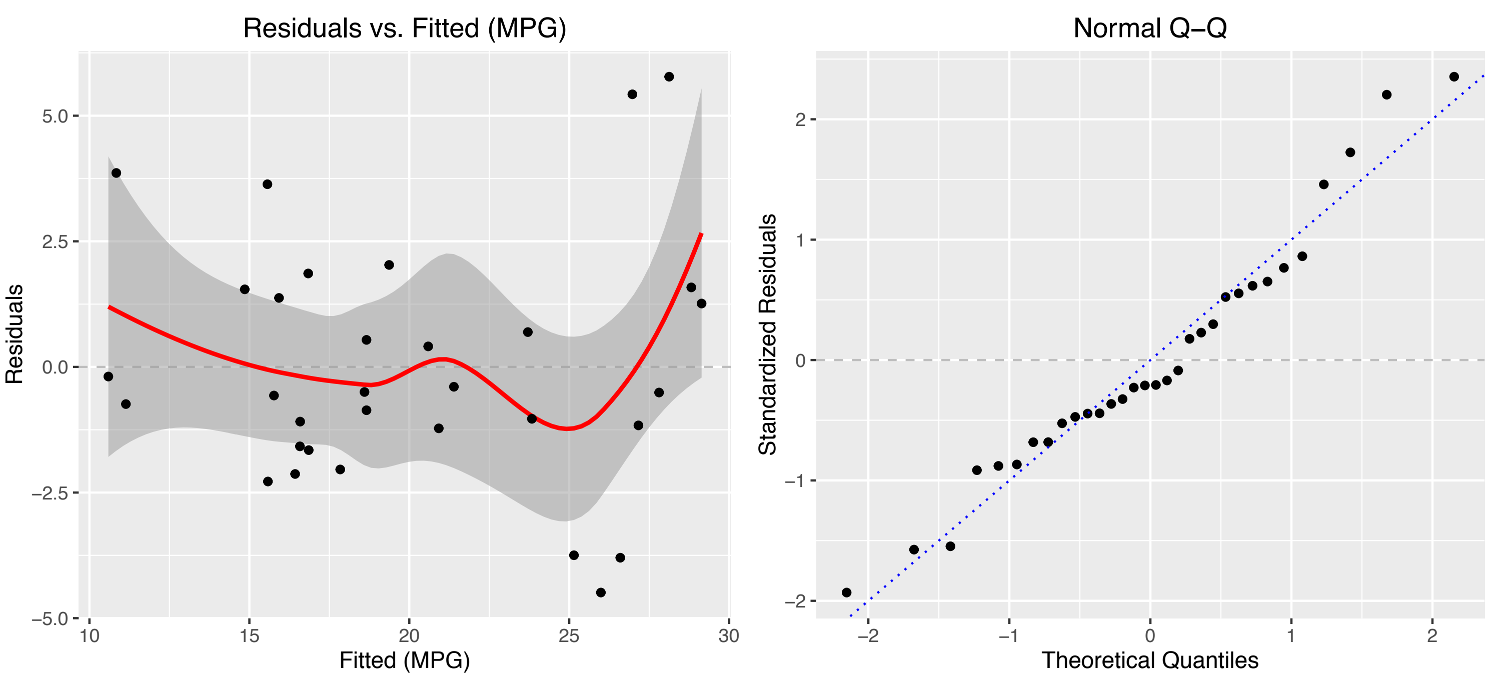

Figure 2: Diagnostic plots showing the absence of pattern in and normality of the residuals