# Influence of Transmission Type on Vehicle Fuel Consumption - A foray into Regression Analysis

## Executive Summary

This analysis fits various linear regression models to the vehicle data to explore the relationship between variables and fuel effiiency (miles per gallon). Initial exploratory analysis results indicated that a manual transmission is more fuel efficient. Of interest is that the weight of the vehicle has a larger impact on fuel efficiency than the transmission type of the vehicle. The heavier the vehicle, the less fuel efficient it becomes. The linear regression analysis has borne out the exploratory results. A manual vehicle is more fuel efficient than an automatic one.

## Data Exploration

A basic box plot (Appendix - Fig 1a) of a direct correlation between Miles per Gallon (*mpg*) and Transmission type (*am*) appears to show that a Manual (*am - 1*) transmission provides better fuel efficiency. An overview of the QQ Plot indicates a relatively even distribution of residuals around the mean (Fig 1b).

A pair plot with correllation coefficients (Fig 2) for one-to-one comparisons of the various pairs of available variables indicate that transmission type has little correlation with any of the other variables, other than the 1/4 mile time (*qsec*). Therefore other variable interactions need to be explored in conjunction with the mpg ~ am baseline. For the purposes of this analysis & exploration, only correlations of an absolute value of 0.7 or higher will be considered potential candidates.

## Model Fitting, Testing and Selection

In the basic correlation done during exploratory analysis, transmission type appeared to not be highly correlated with vehicle fuel efficiency taken in isolation. There are other variables that are more significant. Given that the transmission type is a requirement for the analysis, the baseline model used for all model fitting is **mpg ~ am**. Covariance and confounding factors were explored through two modelling schemas, and then combined to achieve the final fit.

### Modelling schemas:

For a complete output of the various linear models used in the schemas below, please go to https://github.com/Chaendryn/Regression_Project - only the signifcant outputs are detailed below.

**1. Backward step wise removal of variables until only the most significant other than the baseline remained.**
The strategy followed started with a linear regression model (fitAll), to evaluate each variable against mpg based on the p-value. Thereafter iteratively removing the values in order of least significant (as judged from p-value in resulting coefficient summaries).

**2. Interactions between variables directly tied to engine efficiency**
The efficiency of a vehicle engine has an impact on the vehicle's fuel efficiency. While we do not have an engine efficiency variable in the dataset, the interaction of the following variables in addition to the baseline was assessed to see whether there were any that had a significant impact on fuel consumption (as judged from the p-value in the resulting coefficent summaries) - number of cylinders (*cyl*), displacement (*disp*), gross horsepower (*hp*), rear axle ratio (*drat*). The impact of weight on the initial modeling has indicated that an adjustment for the influence of weight on the other variables needs to be made. Baseline model for this schema - mpg ~ am:weight.

**3. Combination of best performers in schemas 1 & 2 - evaluation for final fit.**

Anova nested model evaluations were done and the best performers identified by the significance of the Pr(>F) output (see Table 2 below).

The best performer from modelling schema 1 (fitAllh) was the addition of a single variable **weight** (*wt*). The best performer from schema 2 was the addition of the variable **gross horsepower** (*hp*) adjusted for weight (fit1a).

In this schema, we are evaluating the following models:
- the impact of on mpg keeping transmission type and weight constant
- the impact on mpg keeping transmission type, weight and raw horsepower constant
- the impact on mpg keeping raw horsepower constant, while adjusting for the influence of weight.

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 37.3216 | 3.0546 | 12.22 | 0.0000 |
| wt | -5.3528 | 0.7882 | -6.79 | 0.0000 |
| am | -0.0236 | 1.5456 | -0.02 | 0.9879 |

(a) Table - Coef Schema 1

|  | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 28.2016 | 1.4489 | 19.46 | 0.0000 |
| hp | -0.0669 | 0.0084 | -7.93 | 0.0000 |
| am:wt | 1.7324 | 0.4575 | 3.79 | 0.0007 |

(b) Table - Coef Schema 2

Table 1: Coefficients of best performers

We can quantify the MPG difference between automatic and manual transmissions as follows:

The intercepts indicated by the tables above are the MPG for an automatic vehicle. In Schema 1, having the exact same vehicle in all aspects other than transmission type, fuel efficiency will decrease i.e. the car will get 0.02 miles less per gallon. This is not significant as there is an almost 100% chance that an observed value will be larger than this (p-value of 0.9879).

Of interest is the impact that the weight of the vehicle has on fuel efficiency. Across various models, this aspect has been influential on the outcome of the modelling. In schema 1 the vehicle weight decreases fuel efficiency by 5.3528 per ton of increase in weight (p-value of 1.8674 × 10-7)

In Schema 2, the fuel efficiency for a manual vehicle will be 1.7324 miles per gallon better (adjusted for the influence of weight) with a very small margin either way (p-value of 7.1213 × 10-4). The intercept for schema 2 (mpg for an automatic vehicle) has been adjusted for the influence of weight.

In evaluating the models for Schema 3 - the best model for Schema 2 (fuel efficiency keeping horse power constant and adjusting for weight) out performed both the other models tested.

|  | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 21 | 147.49 |  |  |  |  |
| 2 | 22 | 156.36 | -1 | -8.86 | 1.26 | 0.2739 |
| 3 | 23 | 158.59 | -1 | -2.23 | 0.32 | 0.5793 |
| 4 | 24 | 158.65 | -1 | -0.07 | 0.01 | 0.9223 |
| 5 | 25 | 159.52 | -1 | -0.87 | 0.12 | 0.7288 |
| 6 | 26 | 163.12 | -1 | -3.60 | 0.51 | 0.4820 |
| 7 | 27 | 179.91 | -1 | -16.79 | 2.39 | 0.1370 |
| 8 | 28 | 180.29 | -1 | -0.38 | 0.05 | 0.8175 |
| 9 | 29 | 278.32 | -1 | -98.03 | 13.96 | 0.0012 |

(a) Table - Anova Schema 1

|  | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 30 | 949.85 |  |  |  |  |
| 2 | 29 | 299.57 | 1 | 650.28 | 120.95 | 0.0000 |
| 3 | 27 | 208.59 | 2 | 90.98 | 8.46 | 0.0035 |
| 4 | 23 | 147.90 | 4 | 60.69 | 2.82 | 0.0628 |
| 5 | 15 | 80.65 | 8 | 67.25 | 1.56 | 0.2169 |

(b) Table - Anova Schema 2

|  | Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|---|
| 1 | 29 | 278.32 |  |  |  |  |
| 2 | 28 | 180.29 | 1 | 98.03 | 15.22 | 0.0005 |
| 3 | 29 | 299.57 | -1 | -119.28 | 18.52 | 0.0002 |

(c) Table - Anova Schema 3

Table 2: Anova output from best fit

In evaluating the plot (Fig 3a) the fitted values show a general linearity consistent with our model. The residual plot (Fig 3b) shows no patterned spread around the mean, therefore normality can be assumed.

# Conclusions

In conclusion - the initial exploratory analysis indicated a potential for manual vehicles to be more fuel efficient (mpg) than automatic transmission vehicles. Analysis of various linear models has shown that this is indeed the case, accounting for variability in other measures with weight being the most significant of those other variables.

# Appendix - Figures

Fig 1a

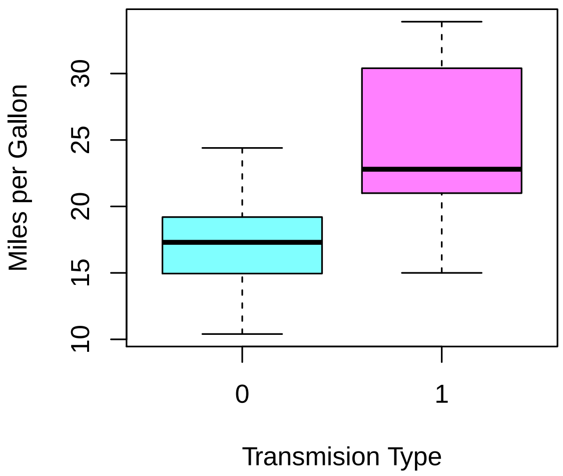


Fig 1b

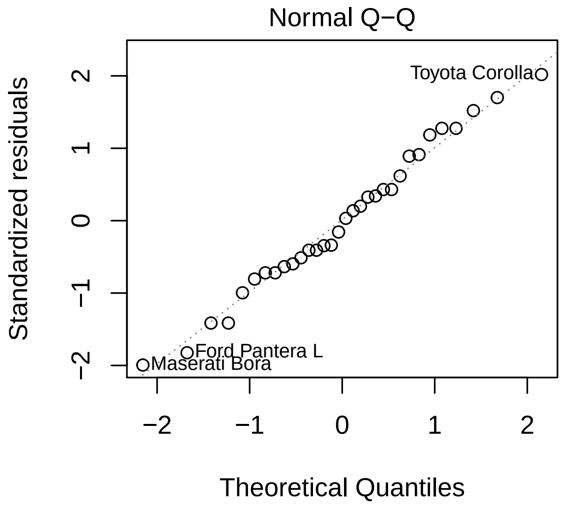
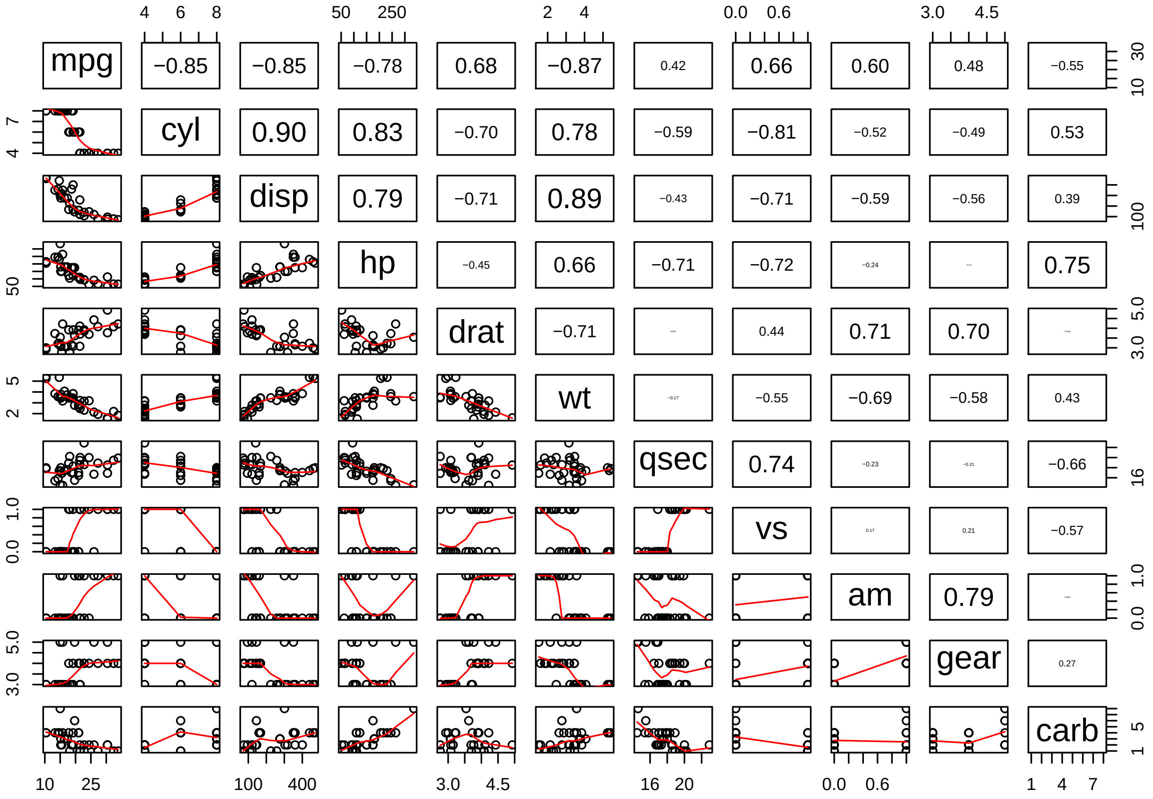
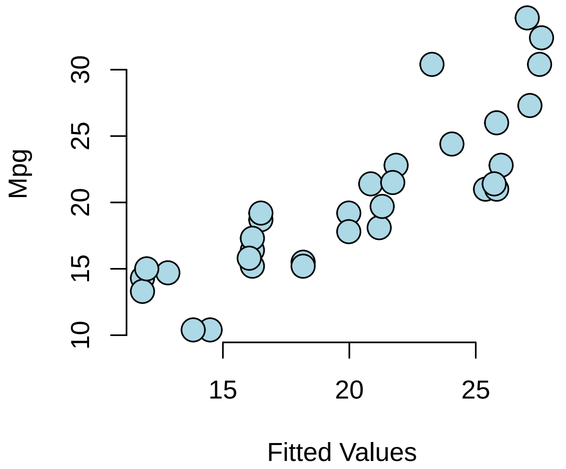


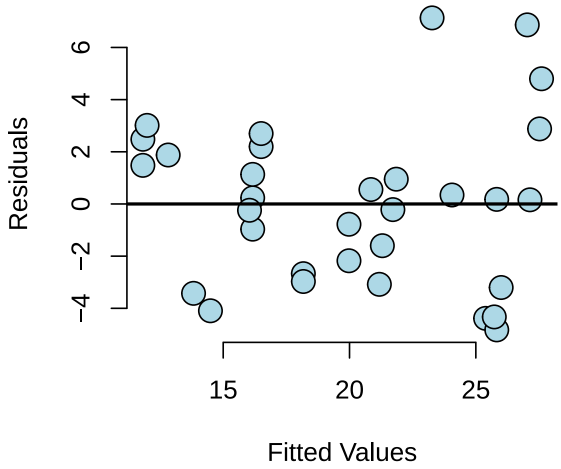
Fig. 2 − Correlation Pairs Plot



### Fig. 3a − Fitted Values

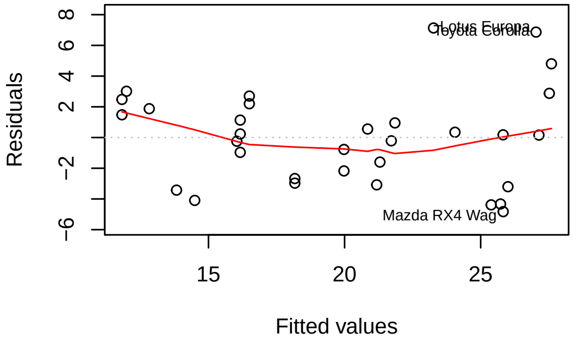


### Fig. 3b − Residuals



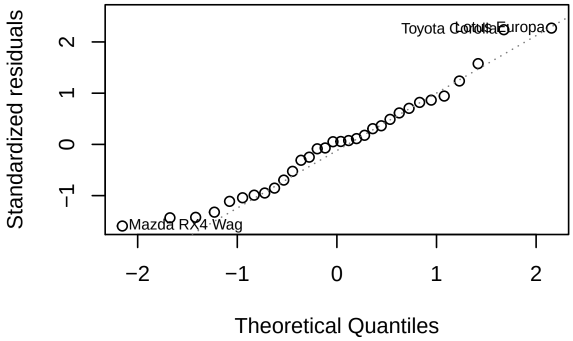
### Fig 4a

Residuals vs Fitted



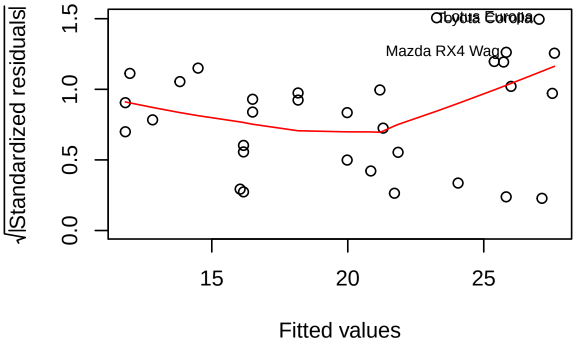
### Fig 4b

Normal Q−Q



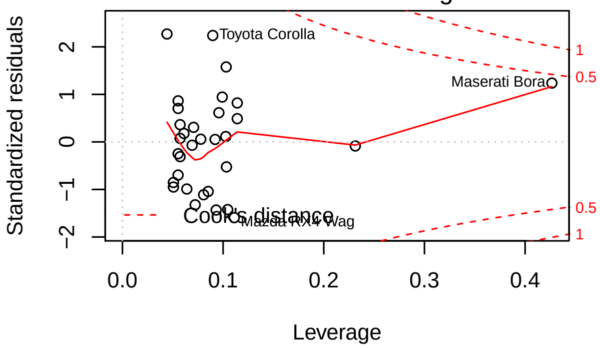
### Fig 4c

Scale−Location



### Fig 4d

Residuals vs Leverage


Standardized residuals
Toyota Corolla
Maserati Bora
Mazda RX4 Wag
Cook's distance
Leverage