An investigation of the effect of automatic/manual transmission on MPG

Executive summary

This report is concerned with investigating relationships between automatic/manual transmisision and miles per gallon (MPG). *Motor Trend*, an automobile industry magazine is particularly interested in the following two questions: 1) Is an automatic or manual transmission better for MPG? 2) Quantify the MPG difference between automatic and manual transmission. We have a dataset of a collection of cars to assess the relationship between MPG and transmission; the dataset includes many variables (see below) which may be potential confounders; these could adversely influence the results of our study. The report includes a discussion on feature selection and a comparison of various mulitvariate regression models.

Data exploration & analysis

The data set we have is the `mtcars` data set that can be found in `R` while a brief description of the variables may be obtained by typing `?mtcars`. We begin by looking at correlations between the quantitative variables and the outcome, `mpg`. Prior to assessing correlations however, we add to our existing dataset the squares of potential predictors. The reason behind the transformation is that relationships in physics are rarely linear; a correlation matrix can show whether a variable in higher power is more correlated with the outcome compared to the same variable in first degree. We select the continuous variables and create a correlation plot (see Appendix, Fig. 1). In Fig. 2 (refer to Appendix), we have three panels wherein we explore the effect of transmission on MPG. Since we are interested in the effect of transmission on MPG, we illustrate through a violin plot the relationship of the MPG variable and transmission in panel (a). The black marker indicates the mean value in each distribution and we observe that cars with a manual transmission offer higher MPG than those with automatic transmission.

Next, we perform a statistical test to check whether the two means we observe from the two populations, MPG for automatic and MPG for manual transmission, as seen in Fig. 3 are reliably different from each other.

We assume that the data is sampled from normally distributed populations and that the two populations have *unequal* variance. The results of the *Welch Two Sample t-test* are shown below.

t-statistic	CI (lower)	CI (upper)	p-value
3.767	3.21	11.28	0.0014

Given the p-value, at a significance level of 0.05, we reject the null hypothesis and conclude that the difference in means is statistically significant such that manual transmission is associated with *higher* MPG than automatic transmission.

Multivariate regression

In model selection, we seek a relationship between a response (here, this is MPG) and predictors that is *parsimonious*. We are interested in assessing the effect of the transmission on gas mileage hence a first, simple model is

$$y_1 = \beta_0 + \beta_1 X_i + \varepsilon$$

where $X_0 = 0$ if transmission is automatic and $X_1 = 1$ if manual, ε (ε_i, where $i = 1, ..n$) are the unobserved errors assuming to be independent and identically distributed (i.i.d). The subsript "1" in the outcome represents the model we are considering (here, this is model **1**).

From the table below, we observe that only 35% of the variability has been accounted for by using the transmission factor as a regressor. The coefficient estimate $\hat{\beta_0}$ gives the `mpg` for automatic transmission while $\hat{\beta_1}$ gives the change in `mpg` for manual transmission *relative* to automatic. This model shows that manual transmission is better for `mpg` by approximately 7 units.

$\hat{\beta_0}$	$\hat{\beta_1}$	R^2	Adj. R^2
17.147	7.245	0.3598	0.3385

We build on the model by including variables that are highly correlated with outcome but disregarding predictors that are highly correlated with each other to avoid variance inflation. Based on the exploratory data analysis, the quantitative variables with the highest correlations with `mpg` are: `wt`, `disp`, `wt`^{2}, and `hp`. However, the variables `disp` and `wt` are highly correlated (`corr = 0.888`) thus `disp` is removed on account of a higher mean absolute correlation compared to `wt`.

Our second model builds on the first one by including the variables `hp` and `wt`: $y_2 = \beta_0 + \beta_1 X_i + \beta_2 \text{hp} + \beta_3 \text{wt} + \varepsilon$.

The model coefficients are shown below as well as the R^2 and adjusted R^2 values. We look at the adjusted R^2 as this quantity factors in the fact that we have two more variables in our model now compared to the first model. This value is now up to 0.8227 and an **analysis of variance** calculation on the two models suggests with a p-value of $3.744703e^{-09}$ that the inclusion of the two variables has improved on our first model.

$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$	R^2	Adj. R^2
34.003	2.084	-0.037	-2.879	0.8399	0.8227

Now, $\hat{\beta_0}$ gives the MPG for automatic transmission, holding `hp` and `wt` constant. The estimate for $\hat{\beta_1}$ represents the change of MPG for manual transmission compared to automatic, again, keeping all other variables constant. This model suggests that MPG is better for manual transmission by about 2 units. Beyond the adjusted R squared measure, we look at diagnostic plots to assess whether a linear fit to the aforementioned variables is a sensible one (see Appendix for figures). The upper left plot shows the residuals (the vertical distance from a point to the regression line) against the fitted values, $\hat{y_2}$. The smoothed blue line shows a distinct U-shape indicating that the linear model is not a good fit. The upper left plot shows that our normality assumption is valid, and that the variance of the residuals can be considered constant.

The third model we considered is the addition of wt^2, based on the correlations we observed before and the fact that diagnostics on model 2 indicate a nonlinear relationship; $y_3 = \beta_0 + \beta_1 X_i + \beta_2 \text{hp} + \beta_3 \text{wt} + \beta_4 \text{wt}^2 + \varepsilon$.

We observe some improvement in the adjusted R squared measure, the p-value associated with the addition of `wt`^{2} is 0.015. The results seem to suggest that the MPG is better for manual transmission by 0.3 units; however, the residual plot (Fig. 4, top) suggests that the nonlinear model is a better fit. Finally, we compare the performance of models 1-3 against an all-subsets regression model (model 4) obtained using the `leaps` package which ranks all the possible models from each subset. For instance, if we choose to have only one regressor in our model, the procedure in `leaps` finds the best single-regressor model. We used the **exhaustive** method to obtain the following model:

$$y_4 = \beta_0 + \beta_1 X_i + \beta_2 \text{wt} + \beta_3 \text{qsec}$$

.

$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$	R^2	Adj. R^2
9.618	2.936	-3.917	1.226	0.8497	0.8336

This shows that `mpg` is better for manual transmission by 3 units.

Appendix

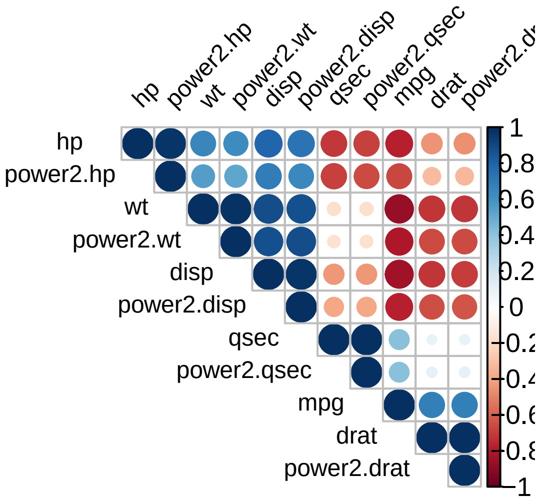

Figure 1: Graphical display of correlation matrix: MPG has a strong negative correlation with the variables associated with gross horsepower, displacement, and weight. Further, MPG is positively correlated with rear axle ratio (`drat`) and has low positive correlation with `qsec` which measures the 1/4 mile time.

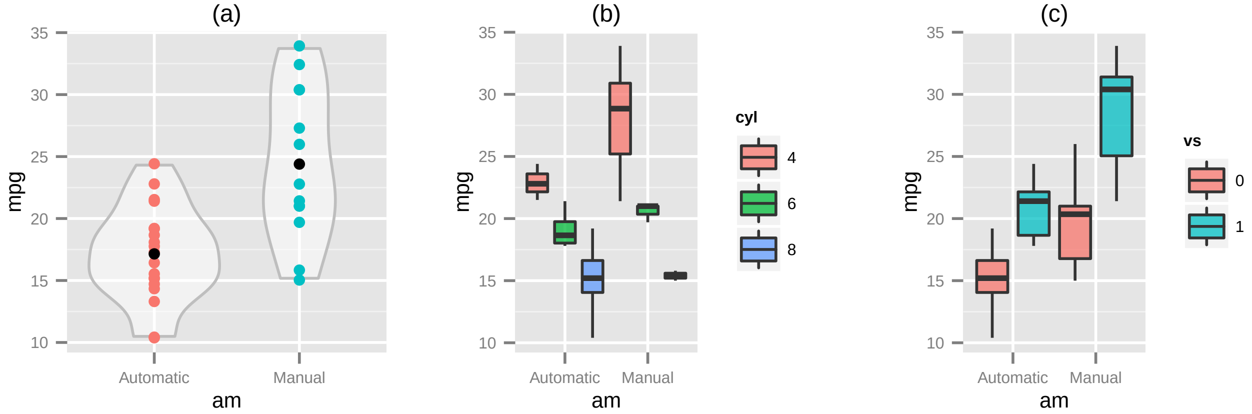

Figure 2: (a) Violin plot showing MPG against automatic and manual transmission. Here, the black marker indicates the *mean* value. Transmission vs MPG grouped by (b) number of cyclinders and (c) type of combustion engine

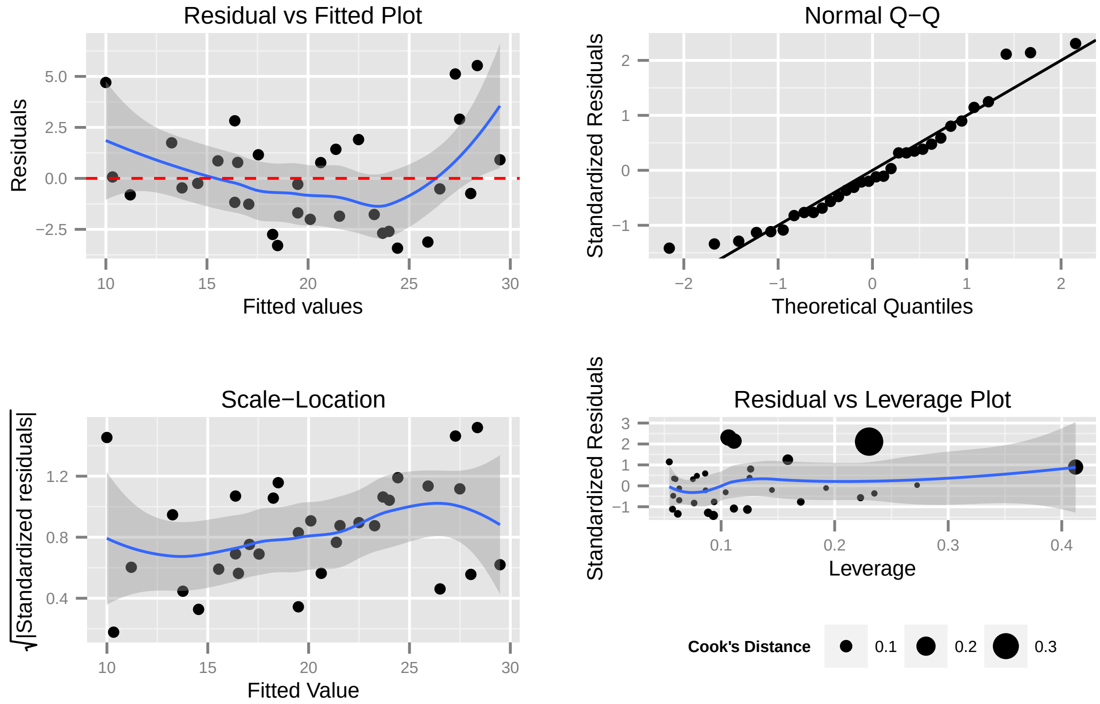

Figure 3: Diagnostic plots for model 2

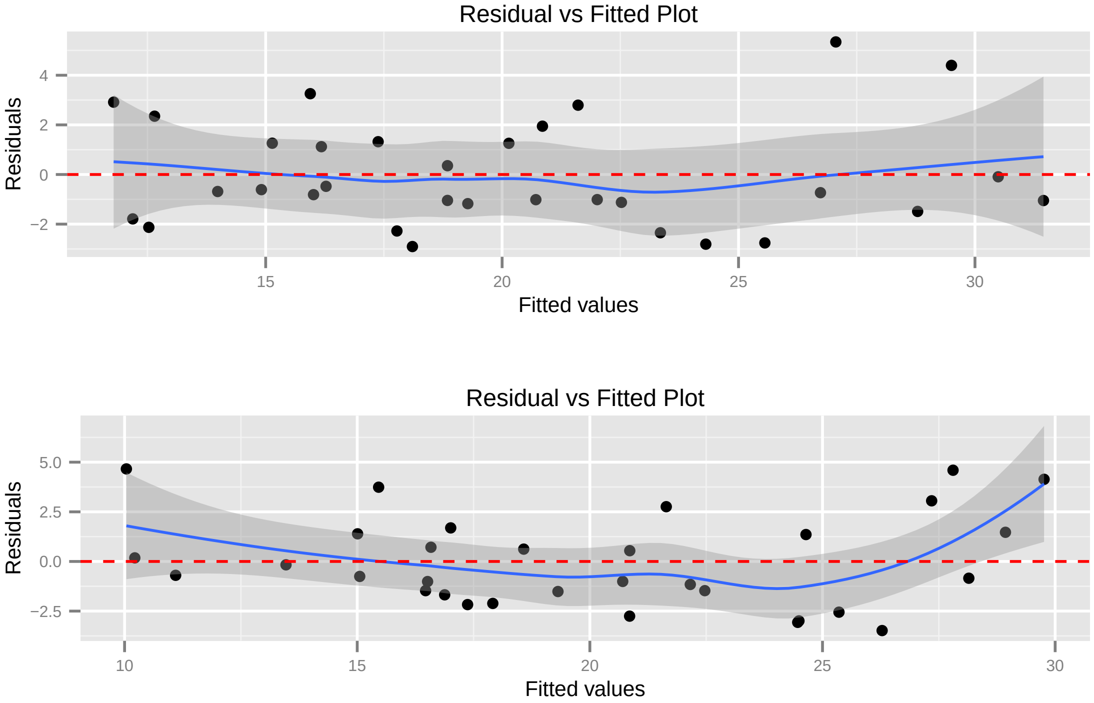

Figure 4: Residual plots for model 3 (top) and model 4 (bottom)