Model Fitting for mtcars Dataset

Executive summary

The present report is focused on analysis of the mtcars dataset with the aim to provide answers to two questions:

1. Is an automatic or manual transmission better for MPG
2. Quantify the MPG difference between automatic and manual transmissions

By "better" in question 1 it is understood "higner" in terms of numerical value of MPG (miles per gallon).

The main part of the report contains only text and tables, whereas the supporting figures are available in the Appendix.The original Rmd file used to generate this report is available in my github repository [1]. The approach to variable transformations in this report was inspired by the "Building Multiple Regression Models Interactively" paper by Henderson and Velleman [2].

Dataset Exploration

The **Motor Trend Car Dataset** (called `mtcars` from now on) consists of **32** observations (corresponding to car models) of **11** variables. The observed variables are:

mpg, *cyl*, *disp*, *hp*, *drat*, *wt*, *qsec*, *vs*, *am*, *gear* and *carb*

Useful summary of the data for our purposes is presented in Figure 1. From this figure we see (upper panel) that the variables in the `mtcars` dataset can be divided into two groups: continuous and discrete. For better interpretability of the results, we'll center and scale the continuous variables by substracting their respective means and dividing by standard deviations.

The lower panel of Figure 1 depicts the correlation ellipses and smooth lines for the pairwise variable relations. For the purpose of the present report we are primarily interested in relation of `mpg` with the remaining variables. Ellipse plots indicate that `mpg` is particularly strongly inversly correlated with `wt` (r = -0.8677), `cyl` (r = -0.8522), `disp` (r = -0.8476) and `hp` (r = -0.7762). There is also quite strong positive correlation between each pair of those 4 variables, what indicates that they are not mutually independent. These observations would need to be taken into account to avoid effects of multicollinearity when fitting the models.

Looking at the smooth lines on Figure 1 it can be seen that there is mostly non-linear relationship between `mpg` and the other variables, especially those strongly inversly correlated. As noted in [2], this suggests that for fitting a linear model with `mpg` as the dependent variable it would be better to fit to its inverse, `gpm` or gallon per mile. We introduce therefore a variable equivalent to `mpg` which is gallons per 100 miles, centered and normalized and we call it `gpm`.

Figure 2 shows the reationships of `gpm` with all remaining variables but `am` (we control for `am`). The relation of `gpm` with its higly correlated variables `cyl`, `disp`, `wt` looks quite linear with the exception of `hp` which still retained some non-linear dependency.

Model Fitting

First the trivial simple linear model with `am` as predictor is fitted. It is equivalent to calculating and testing significance of the difference of the group means. Subsequent models are created using backward elimination procedure where in some models we take into account underlying physics governing the fuel consupmtion (see [2]), that is preserving `wt` as the predictor will be preferred. Lastly a few models are build by trying to incrementally improve the original simple linear model by adding new predictors. For all models Akaike Infomation Criterion and R^2 are used as the criteria to select the best model.

Trivial Model

	Estimate	Std. Error	t value	Pr(>|t|)	CI 2.5%	CI 97.5%
(Intercept)	0.4395416	0.1963006	2.239125	0.0327152	0.0386423	0.8404408
I(factor(am))1	-1.0819484	0.3079817	-3.513028	0.0014266	-1.7109310	-0.4529659

This model (**fit0**) is fit using `am` variable as the sole predictor. As the `am` variable assumes only two values (0 - automatic transmission, 1 - manual transmission), obtained coefficients are interpreted as follows: for the automatic transmission the average gallons consumed per 100 miles driven are 0.4395 SDs above the average consumption for all the cars in the dataset. The average consumption for the cars with manual transmission is 1.0819 SDs of overall gallons per 100 miles lower of the average consumption for the cars with automatic transmission. The p-value for the differece in means is statistically significant, so we can infer that for this particular

dataset there is a difference in gpm (and mpg) between automatic and manual transmission. For this model no diagnostic plots are presented as this case is trivial and equivalent to t-test of the sample means. Uncertainty of the model coefficients is quantified by the provided Confidence Intervals.

Backward elimination, p-value criterium, preserving `wt`

In this model we start with all independent variables as predictors and eliminate them one by one starting with the least significant variables as indicated by the variable's coefficient p-value, but preserving `wt` variable independently of the p-value it's coefficient might have during the process. At the end we arrive at the following model:

fit1: $gpm = \beta_0 + \beta_1 disp + \beta_2 wt + \beta_3 carb$

Backward elimination, VIF criterium, preserving `wt`

In this model we start with all independent variables as predictors and eliminate them one by one starting with the variable having highest VIF, but preserving `wt` variable independently of its VIF. At the end we arrive at the following model:

fit2: $gpm = \beta_0 + \beta_1 wt + \beta_2 qsec$

Backward elimination, VIF criterium, not preserving `wt`

This process is similar to the previous one with the difference that the variable `wt` is allowed to be eliminated. This resulted in the following model:

fit3: $gpm = \beta_0 + \beta_1 drat + \beta_2 carb$

Please note, that this model is uniterpretable.

Forward selection models with `am` as obligatory predictor

Following 3 models were selected for analysis by adding variables and performing ANOVA analysis of significance of added predictors, so that all are significant:

fit43: $gpm = \beta_0 + \beta_1 am + \beta_2 wt + \beta_3 disp$

fit53: $gpm = \beta_0 + \beta_1 am + \beta_2 wt + \beta_3 qsec$

fit62: $gpm = \beta_0 + \beta_1 am + \beta_2 hp$

The following table presents the AIC and R^2 values for all the models:

	AIC	Adj. R-Sq
fit0	84.77	0.2679
fit1	38.29	0.838
fit2	37.8	0.8361
fit3	66.12	0.6027
fit43	41.48	0.8211
fit53	39.77	0.8304
fit62	57.57	0.6959

From the above table it can be observed that the best model from those considered is **fit2** having the lowest AIC and one of the highest R^2 (84%). Model **fit1** has slightly higher R^2, but probably due to havig one regressor more. This result (**fit2** as best model) is consistent with findings in [2], where `qsec` can be treated as equivalent to car's overpower.

From models including `am` as a predictor the best results presents **fit53** which is basically a modification of **fit2** by including `am`. We can conclude that in this case the inclusion of `am` doesn't spoil the best model too much, so `am` basically offers no explanatory value.

Diagnostic printouts for **fit2** model are presented in Figure 3. It can be seen that there are some outlying values (principally rows 15 and 17), removing of which could potentially improve the model, but other than that the model has acceptable diagnostics.

Results and Conclusions

With reference to question 1, we can conclude that ***within the analysed dataset*** the cars with the manual transmission offer better mpg/gpm than those with automatic transmission. But we found no evidence in the data that this observation holds in general (compare models **fit2** and **fit53**).

As far as question 2 is concerned, its equivalent version of gallons per 100 miles was quantified. To this end, cars with automatic transmission have, on average, higher gas consumption of 1.777 gallons per 100 miles. See also discussion of **fit0** model.

Appendix

Figure 1. Motor Trend Car Data

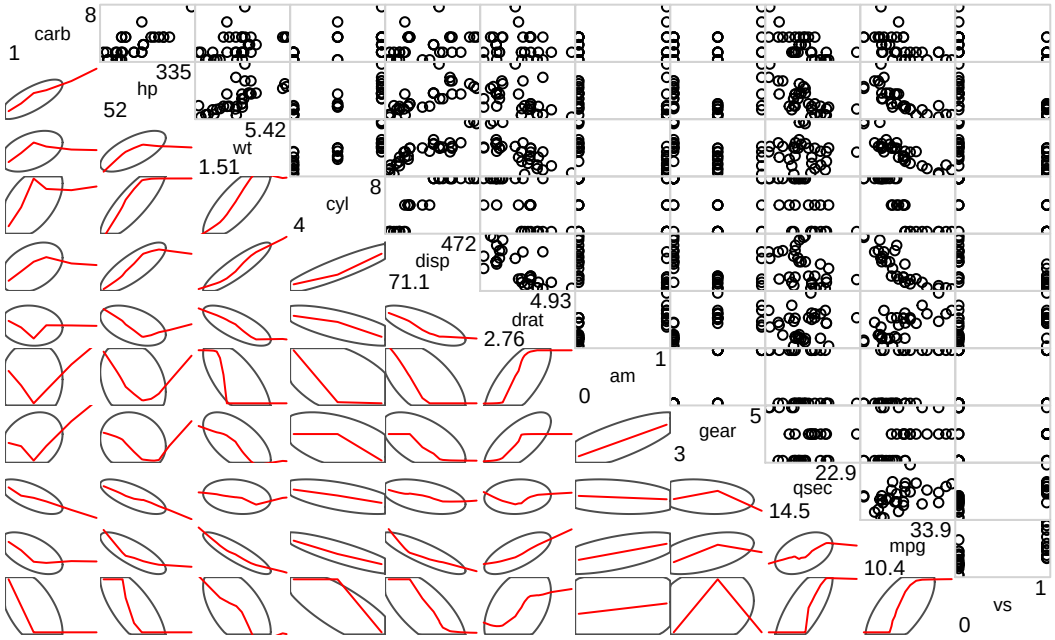


Figure 2. GPM vs other variables controlling for AM

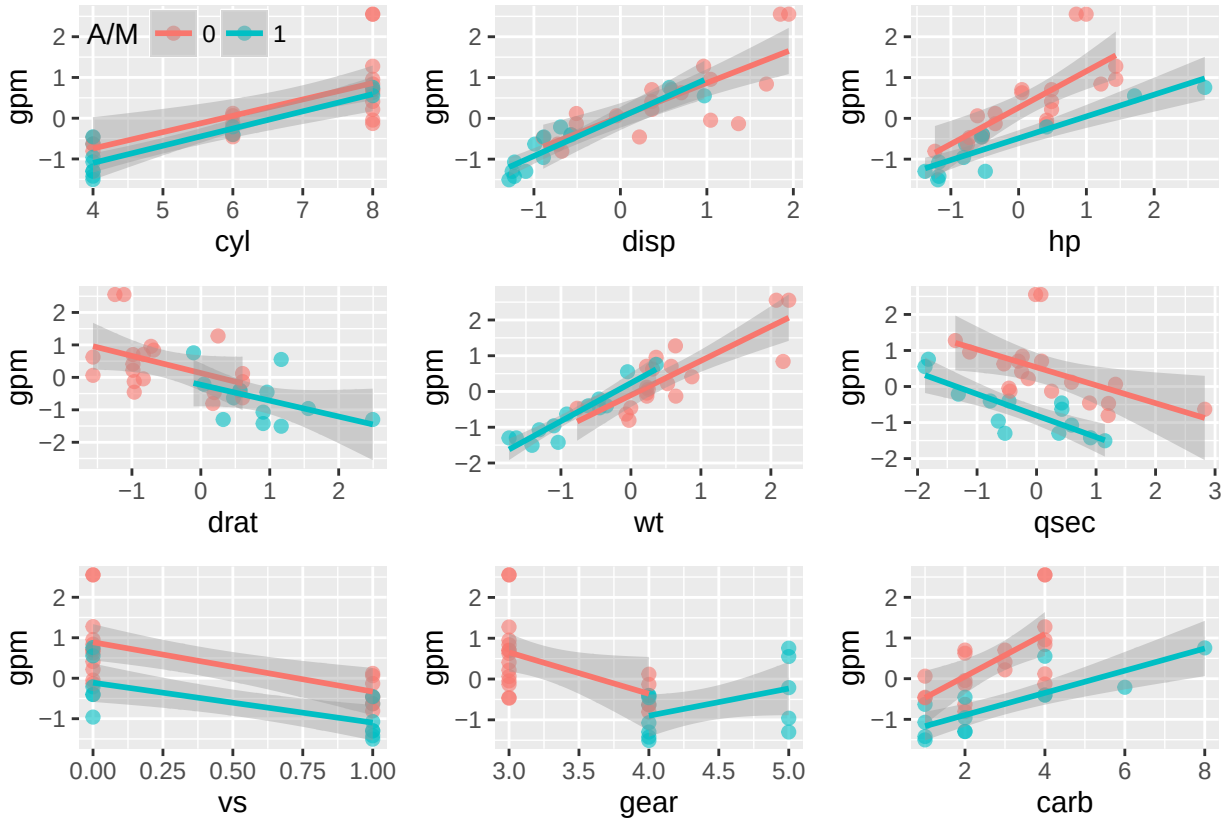
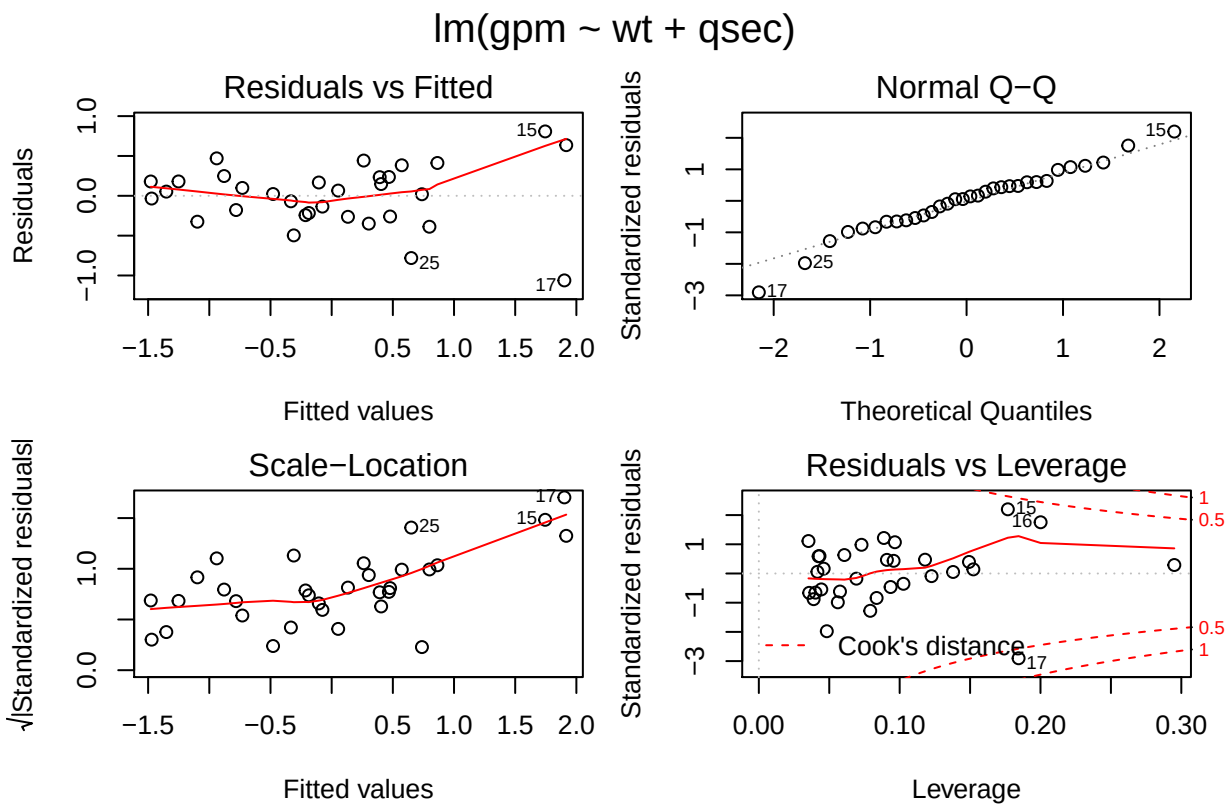


Figure 3. Diagnostic plots for model fit2



References

1. https://github.com/marioem/Regression-Models-Project.git
2. Harold V. Henderson, Paul F. Velleman (1981). *Building Multiple Regression Models Interactively.* Biometrics, Vol. 37, No. 2. (Jun., 1981), pp. 391-411.