Regression Models Course Project

Fritz Lin

Summary

Today's consumers are increasingly becoming more conscious of their impact on the environment. One source of pollution is the car. The automobile industry has reacted to this trend by producing more and more fuel efficient cars. In this paper we will look at the 1974 Motor Trend Car data that comprises fuel consumption of 32 automobiles. We are interested in exploring the relationship between a set of variables and fuel efficiency, specifically we would like to address the following two questions:

- Is an automatic or manual transmission better for gas mileage?
- What is the difference in gas mileage between automatic and manual transmissions?

We come to the conclusion that fuel efficiency is not solely dependent on the transmission type (automatic or manual) alone, but also on a combination with weight and acceleration. We do believe, in general, that manual transmission is better for gas mileage.

Naive Model and Exploratory Data Analysis

We first investigate the naive model that considers only the difference in the transmission type (automatic versus manual). The independent group t-test reveals that the difference in the means of the gas mileage based on the transmission type is significant at the 5% level (p-value of 0.0014) with a mean value of 17.15 for automatic transmissions and 24.39 for manual transmissions. This is also evident in the boxplot in figure 1.

However, if we perform a regression on the gas mileage with only the transmission type as regressor, the naive model fares rather poorly with an adjusted R^2 value of only 0.3385.

It is evident that fuel efficiency as measured by gas mileage is not heavily dependent on the transmission type. We will now look at the different variables in our dataset to see which variables are the main contributors. Figure 3 shows a correlogram and a scatterplot of all the variables. However, we are mostly interested in the first column and first row.

We identify variables such as the number of cylinders (cyl), the displacement (disp), the gross horsepower (hp) and the weight (wt) to be highly (positively or negatively) correlated with gas mileage (mpg), and rear axle ratio (drat), the engine type (vs), the transmission type (am) and the number of carburetors to be also correlated with it, but to a lesser degree. Finally, the number of forward gears (gear) and acceleration measured as quarter mile time (qsec) are least correlated. We also notice that variables are also highly correlated among each other, especially displacement, horsepower, weight and engine type.

Regression and Model Selection

We would like to find a subset of variables that is most relevant to our dependent variable. We make use of the step-wise search algorithm using BIC as our information criterion to perform the model selection. As can be seen in figure 2, the algorithm identifies four parameters with weight (wt), acceleration (qsec) and transmission type (am) as our best model. Note that the algorithm includes therefore the most and least correlated variable with respect to the dependent variable.

Now looking at figure 4 which plots the regression lines on gas mileage with weight (wt) and acceleration (qsec) as regressors, it appears that there is an interaction between the variables, as indicated by the different slopes. We therefore specify additional models that include the interaction terms.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
am	1	405.15	405.15	89.91	0.0000
wt	1	442.58	442.58	98.22	0.0000
qsec	1	109.03	109.03	24.20	0.0001
am:wt	1	52.01	52.01	11.54	0.0024
am:qsec	1	0.80	0.80	0.18	0.6768
wt:qsec	1	7.50	7.50	1.66	0.2094
am:wt:qsec	1	0.83	0.83	0.18	0.6710
Residuals	24	108.14	4.51		

Table 1: ANOVA with Transmission Type, Weight and Acceleration as regressors

The results of our ANOVA analysis (table 1) suggest a model with transmission type (am), weight (wt), acceleration (qsec) and an additional interaction term between transmission type and weight as our regressors. However, it turns out that the intercept is not significant (p-value of 0.1109). The best fit we are able to achieve is given by

$$MPG_i = 13.97 - 3.18 I_i^{auto} WT_i - 6.1 I_i^{man} WT_i + 0.83 I_i^{auto} QS_i + 1.45 I_i^{man} QS_i$$

where I_i is an indicator variable. A summary of our regression model in table 2 suggests that all coefficients are significant at the 5% level. In addition, our model produces an adjusted R^2 of 0.879.

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	13.97	5.78	2.42	0.02	2.12	25.82
amAutomatic:wt	-3.18	0.64	-4.99	0.00	-4.48	-1.87
amManual:wt	-6.10	0.97	-6.30	0.00	-8.09	-4.11
amAutomatic:qsec	0.83	0.26	3.20	0.00	0.30	1.37
amManual:qsec	1.45	0.27	5.37	0.00	0.89	2.00

Table 2: Summary of Suggested Regression Model with Confidence Intervals

Residual Plot and Diagnostics

To evaluate our fit given by our specified model, we look at a series of diagnostic plots given in figure 5. We do not detect any outliers in the residual plots, except for point 3 in the upper left plot. In addition, if we examine some of the regression diagnostics provided by `influence.measure()` in R, we note that point 9 and 17 are also influential. These points represent the following cars: Datsun 710, Merc 230, Chrysler Imperial. However, our dataset is already small, any exclusion of the data would be not very meaningful.

Besides this, we are confident that our specified model provides a good fit, however we have one caveat: our analysis of residuals is sensitive to its assumption that the model residuals are approximately normal. We can however inspect this by looking at the QQ-plot in figure 5, which appears to be approximately normal, except in the tails. A Shapiro-Wilk test for normality will provide more assurance. The test returns a p-value of 0.4521, so we fail to reject the null indicating residuals are approximately normal.

Conclusion

Our selected model suggests that fuel efficiency is not solely dependent on the transmission type alone, but also on a combination with weight and acceleration. Cars with automatic transmissions appears to be more heavy, which of course will have a bigger negative impact on fuel efficiency (negative coefficients in the regression). In addition to that, if we fix the variable quarter mile time, our model suggests that cars with manual transmission tend to be more fuel efficient as compared to its counterpart (see right side of figure 4). On average the difference in fuel economy between cars with automatic and manual transmission is around 7.24 (= 24.39 - 17.15) miles per US gallon.

Appendix

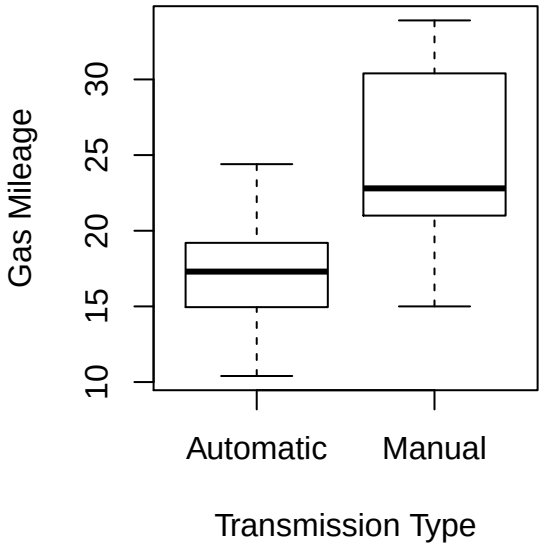

Figure 1: Boxplot of Transmission Type vs Gas Mileage

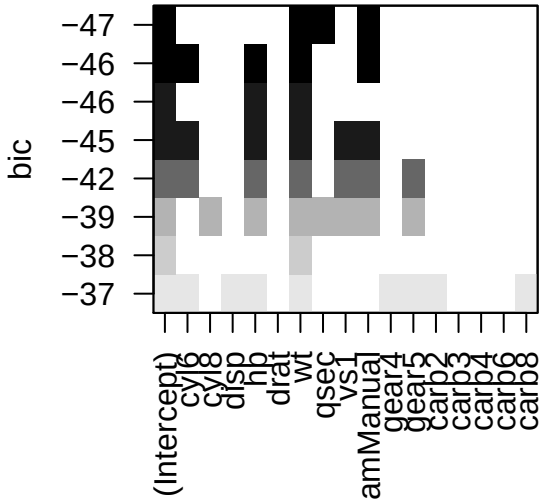

Figure 2: Model Selection by Step-Wise Search

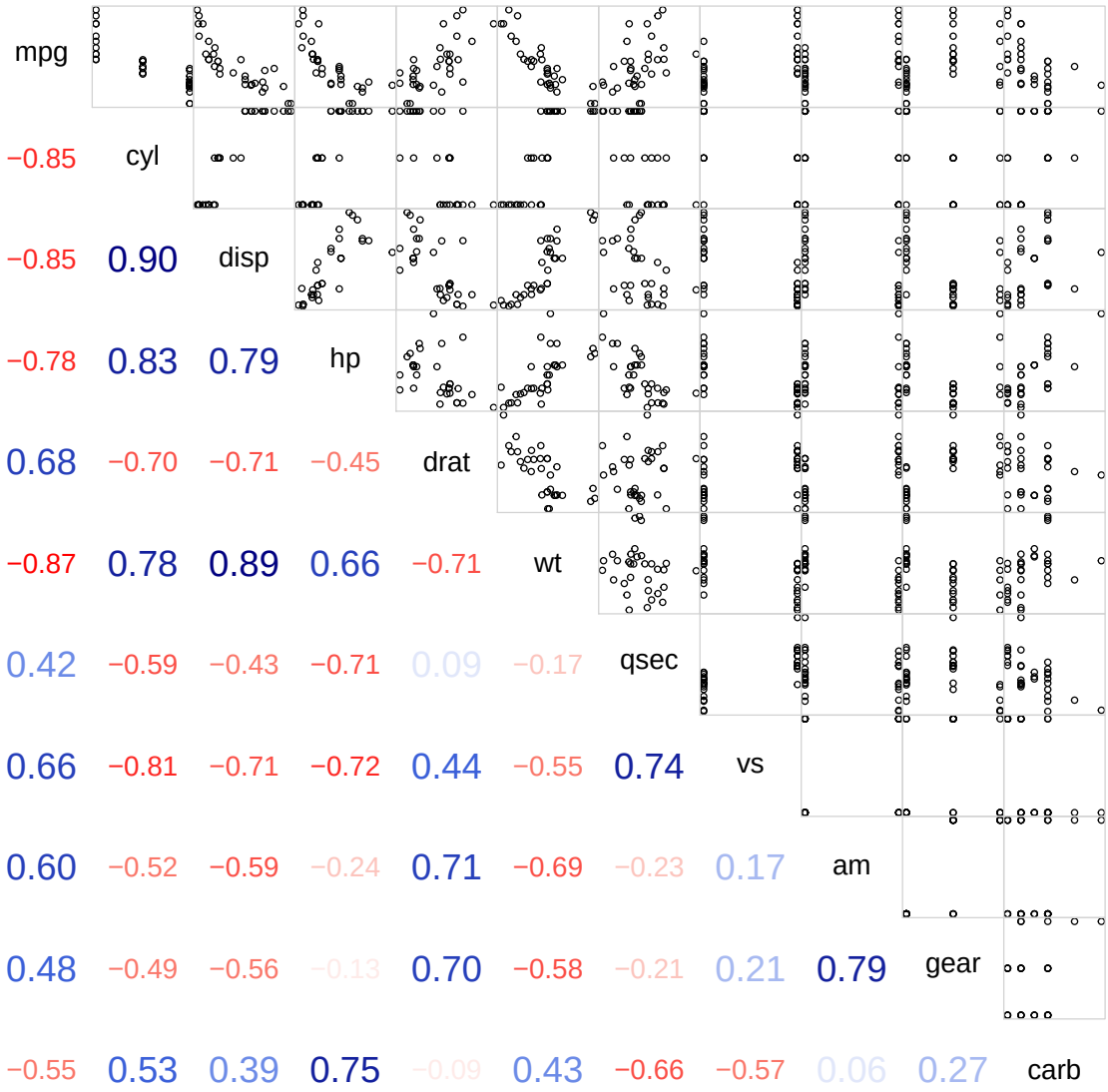

Figure 3: Correlogram & Scatterplot of all Variables

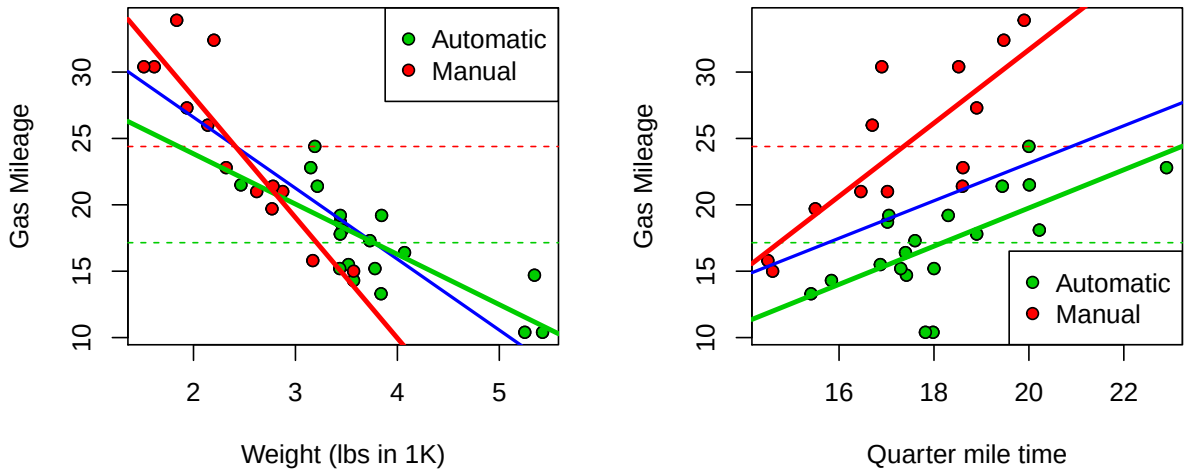

Figure 4: Regression on Gas Mileage

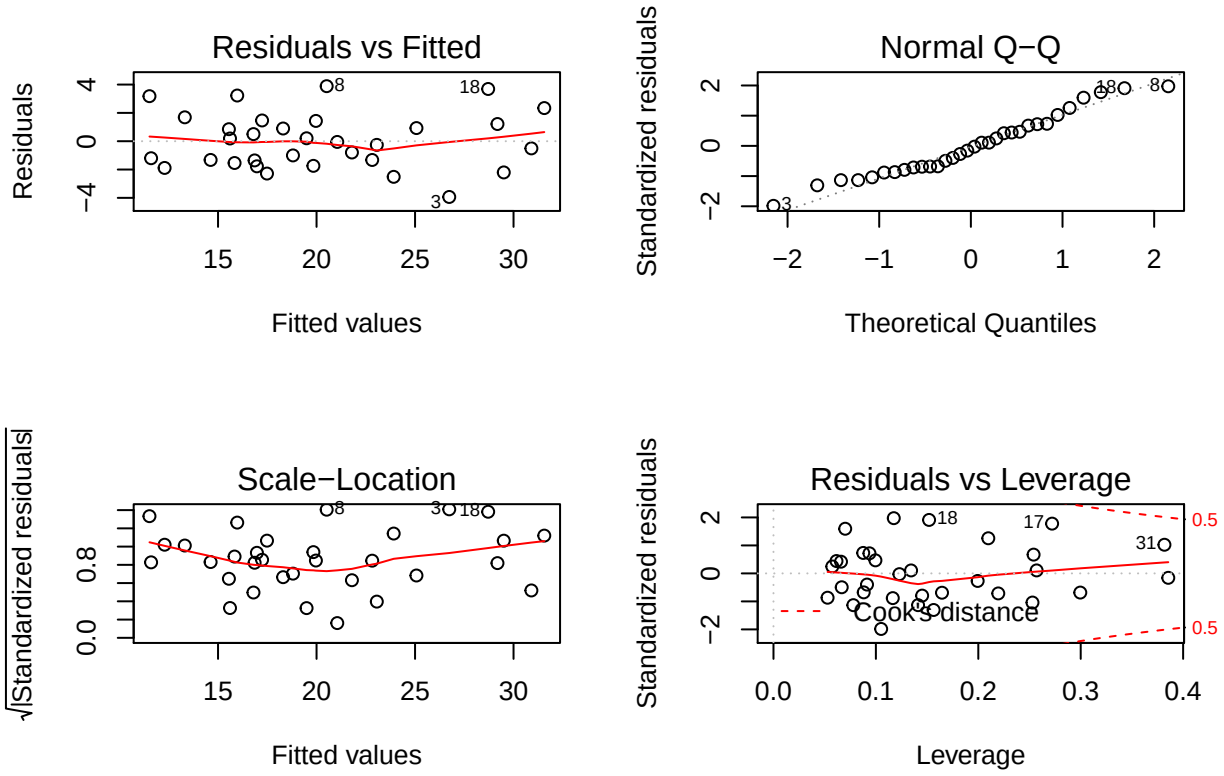

Figure 5: Residual Plots and QQ-Plot