Regression Models Course Project - May 2015

The purpose of this document is to answer the following question: "Is an automatic or manual transmission better for MPG?" Based on regression analysis performed on the mtcars dataset, the answer to this question is that a manual transmission gives a better MPG rating. Specifically, with 95% confidence and with all other variables held constant, a manual transmission automobile is expected to have 2.94 more MPGs than an automatic.

Model Selection And Analysis

The model is based on the mtcars dataset, which was extracted from the 1974 issue of *Motor Trend* magazine. A total of 32 observations of 11 variables were collected in this dataset: rate of fuel consumption and 10 other variables about the automobiles, such as engine displacement, automobile weight and others.

The regression model predicts mpg using transmission type, weight and quarter-mile time as regressors. Models' predictive power were assessed using adjusted R squared (R^2_{adj}) values and predictive R squared (R^2_{pred}) values (obtained from the PRESS statistic), and adding variables were confirmed significant by examining likelihood ratios from ANOVA and by comparing coefficient p-values before and after variable addition.

1. The dataset only has 32 observations, so it doesn't have the statistical power to merit a model with more than 3 coefficient terms. See this whitepaper by Minitab statisticians for more information.

2. Theory about automobile fuel efficiency dictates that weight of the automobile and the engine performance are primary factors. Naturally, weight is included in the model. A base model comparing just MPG and transmission results in highly significant coefficients but a low predictive power, as evidenced by the adjusted R^2 value of 0.3384589. Furthermore, ANOVA analysis of a model including the addition of weight versus a model that just compared MPG values to transmission results in a highly significant likelihood ratio value of 1.867415×10^{-7}.

3. Weight cannot be the only included variable besides transmission in the model. The model including wt and transmission type only as predictors has high predictive power (adjusted $R^2 = 0.7357889$), but its coefficient showing the difference in MPG from automatic to manual is -0.0236152, with a very low significance value of 0.9879146. Given the non signficance of this coefficient and the fact that there is a clear difference in MPG ratings (as can be seen in Figure 1 in the appendix), the model must be further adjusted.

4. Higher performing engines, i.e., those with higher accelerations typically consume fuel at a much higher rate than those with less performance. Factors that affect engine performance include the number of cylinders, horsepower and engine size (displacement), and all of these are variables in the mtcars dataset. Given the size constraints of the model, however, all three cannot be included. But, another variable in the dataset, qsec, or quarter-mile time, assesses engine performance as well. It shows engine performance by timing how long an automobile takes to travel one-fourth of a mile starting from rest. Higher performing engines will have smaller qsec values, and vice-versa. The qsec variable is a succinct way to assess engine performance, and so it is added to the model as well. ANOVA shows that the addition of the qsec term is highly significant: the likelihood ratio is 2.1617371×10^{-4}.

The following table summarizes all coefficients of the model and appropriate companion statistics. Note that the model uses automatic transmission as the basis for comparison.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	9.617781	6.9595930	1.381946	0.1779152
Transmission-Manual	2.935837	1.4109045	2.080819	0.0467155
Weight	-3.916504	0.7112016	-5.506882	0.0000070
Quarter-mile time	1.225886	0.2886696	4.246676	0.0002162

This table shows that coefficients of the three regressors are significant; the intercept's high p-value puts it outside the realm of standard accepted significance. I attribute this observation to the relatively low number of observations in the dataset. However, this does not affect what is more important: the significant difference in MPG values between automatic and manual transmissions. The coefficient of manual transmissioned automobiles in this model is, as mentioned above, 2.9358372 and meets the 95% confidence scientific standard. Furthermore, it has high R^2_{adj} and predictive R^2 values of 0.8335561 and 0.7945881, respectively. Finally, the residuals of this particular model do not appear to show any cause for concern; they show no correlation with regressors and appear to be normally distributed (see Figures 3-6)

Caveats

There are, as with any model, certain caveats that must be addressed.

- This dataset is considerably dated. Its 1974 acquisition date puts it at just over 40 years old, and so, given the many changes in automobile design, engineering and manufacturing technologies in those 4 decades, this model should not be used to make predictions for more modern automobiles.
- The size of this dataset does not allow for a lot of statistical power. As I mentioned earlier, given only the 32 observations does not give leeway to make models with large numbers of terms. In fact, in no way does this set meet the recommended minimum of 40 observations needed to make a 1-3 term model whose R^2_{pred} value is within 0.2 units of the population R^2_{pred} value 90% of the time.

Thus, while the model predicts just shy of 3 mpg difference between automatic and manual transmission automobiles on average, in no way should this figure be applied to predict current automotive settings. Many more observations are required to make a more accurate model, and these observations need to be made with more current information.

Appendix

Figure 1

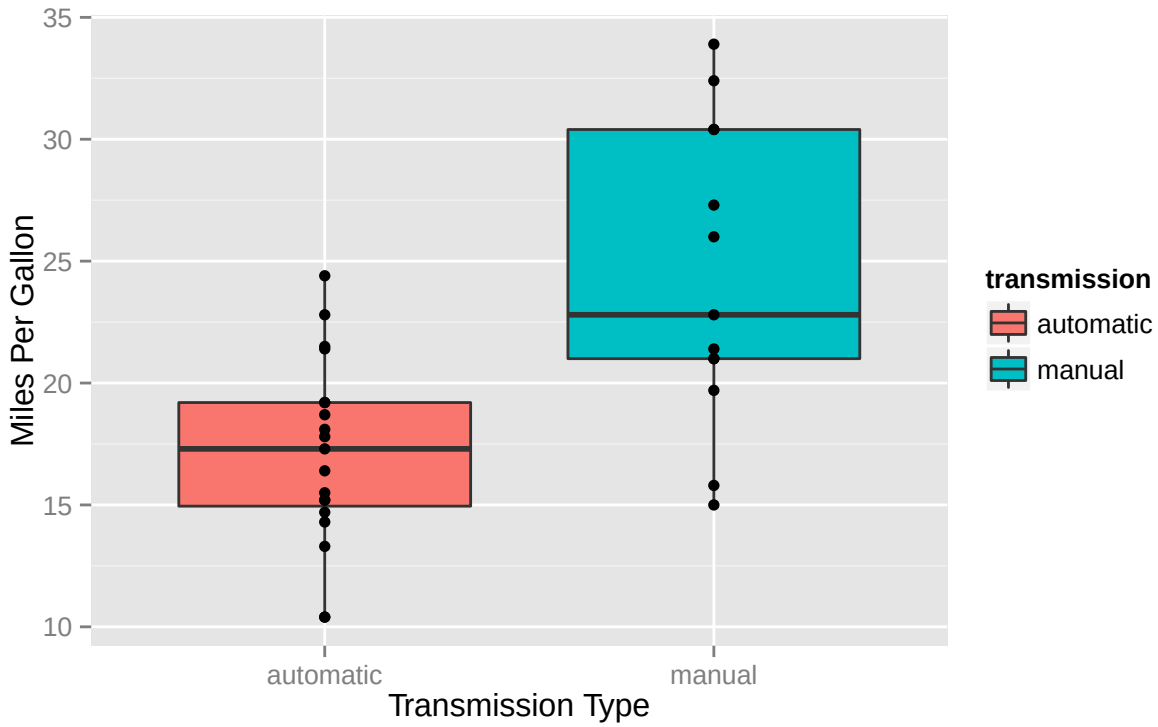

Figure 2

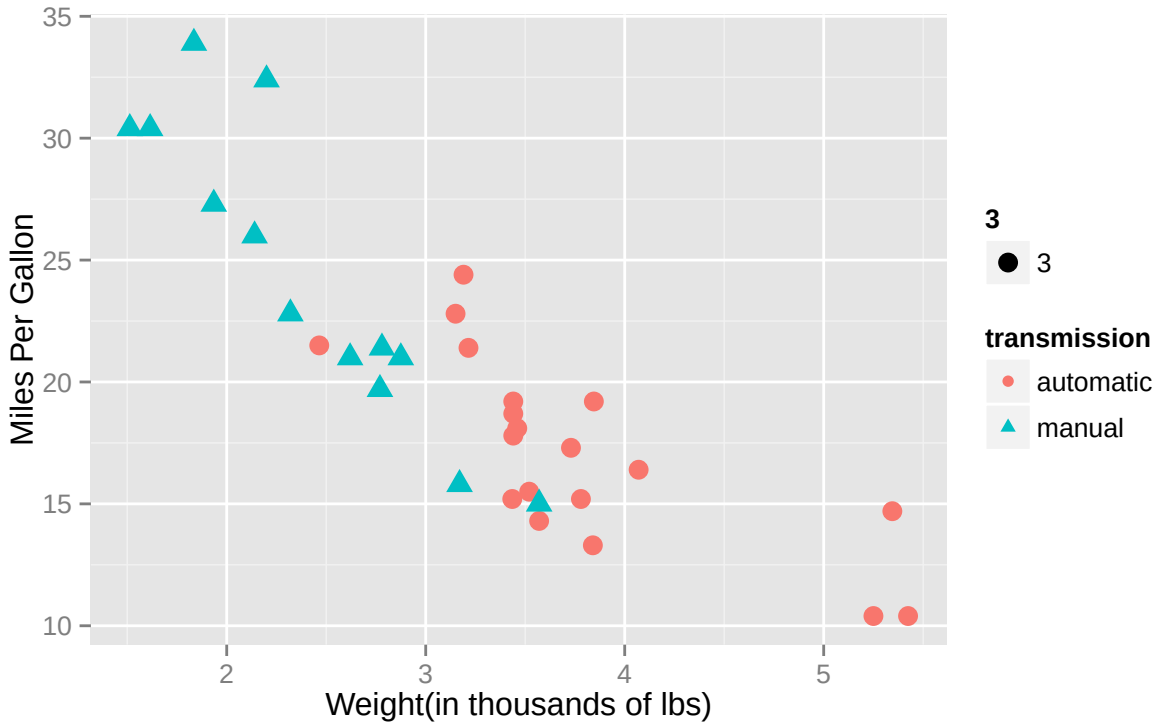

Figure 3

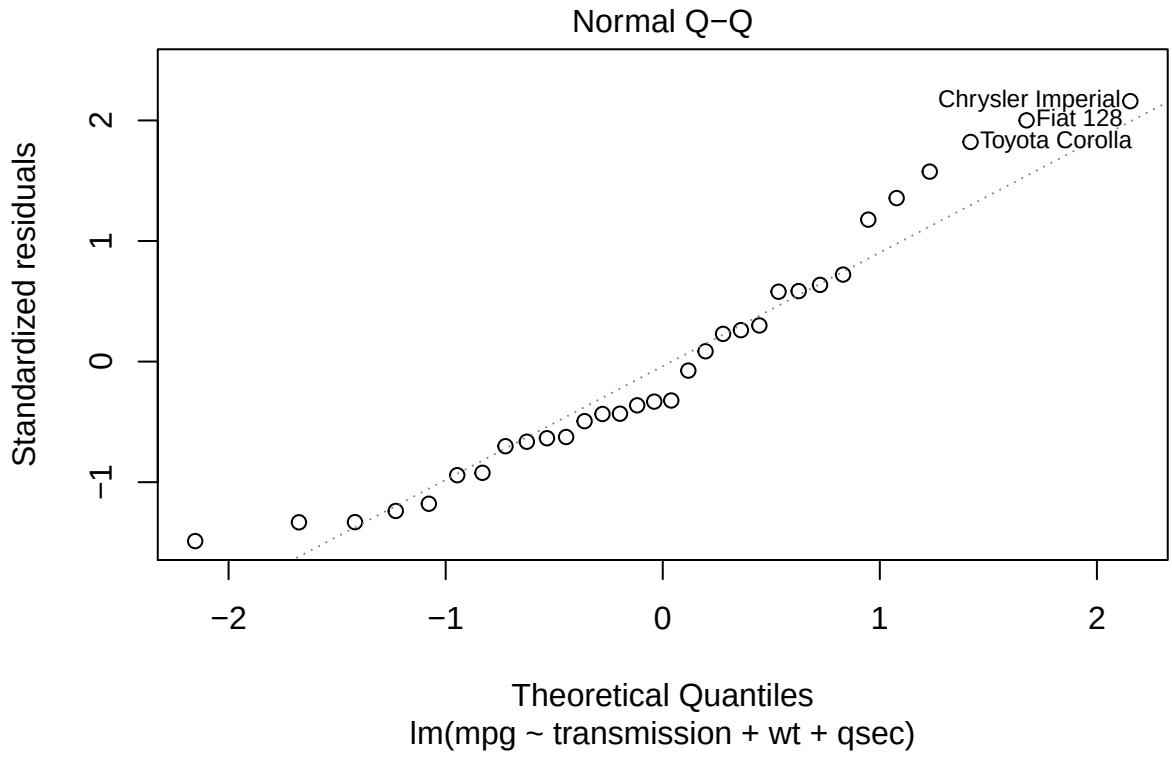

Figure 4

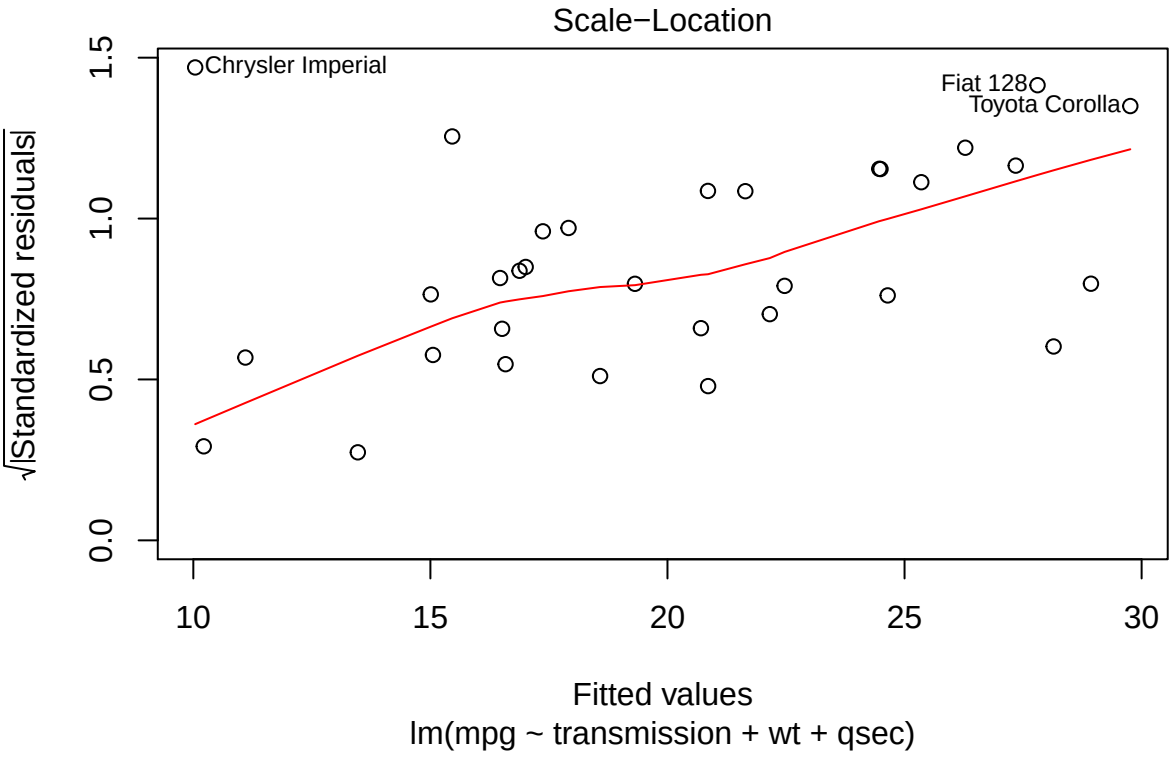

Figure 5

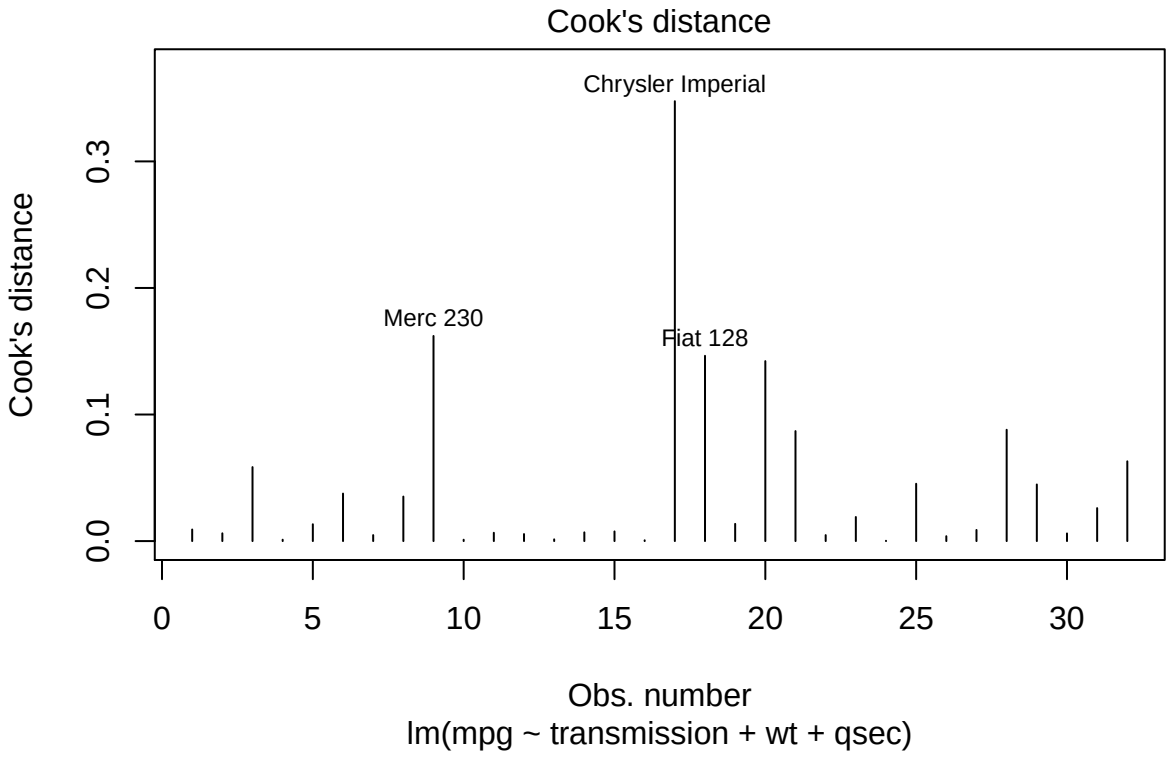

Figure 6