Motor Trend Data Analysis

Calvin Seto

October 21, 2015

Executive Summary

Most drivers know that manual transmissions get better gas mileage than automatic transmissions. Some say that automatic transmissions do as well or better with regard to miles per gallon. There are many theories about the effects of different car specifications on gas mileage.

- In general, cars with more horsepower or displacement mean the car is more powerful, but that doesn't mean they are more fuel efficient.
- A rear axle ratio of 3.42 is associated with everday use cars and better gas mileage and higher speeds.
  - A rear axle ratio of 4.10 is for cars that tow heavy loads with more torque at lower speeds.
- A heavier car is harder to move and less efficient.
- Lower quarter mile times mean faster cars.
- Both V and straight shape engines are high performing.
- More gears and carburetors can mean higher performance.

We'll perform a data analysis to answer these questions:

"Is an automatic or manual transmission better for MPG"

"Quantify the MPG difference between automatic and manual transmissions"

Regression models can be very useful statistical tools that help in this situation. We'll use them to predict a car's miles per gallon with the other variables in the data set, for example weight. The model we choose will be a parsimonious, easily described mean relationship between miles per gallon and the variables we choose, so we can quantify the average difference between variables.

The basis of regression models is to try to fit a line through our data by finding the middle or mean via least squares, that is, the point that minimizes the sum of the squared distances between the observed data and itself.

Methods

Exploratory data analysis

The data set includes 10 aspects of 32 cars (1973-74 models) like small cars Fiat 128, Honda Civic and Toyota Corolla; sedans like the Volvo 142E and Lincoln Continental; sports cars like the Pontiac Firebird, Dodge Challenger, and Camaro Z28; and exotic cars like the Ferrari Dino and Maserati Bora. See tables 1, 2, and 3 in the Appendix for the counts of cars with each specification in the data set.

The data set contains continuous and categorical variables, specifically the dependent variable, miles per gallon (mpg) and the continuous independent variables weight (wt), displacement (disp), gross horsepower (hp), rear axle ratio (drat), and quarter mile time (qsec). There are categorical variables, transmission (am 0=automatic, 1=manual) and V-shape or straight engine (vs 0=V, 1=straight). The remaining variables, number of cylinders (cyl), number of forward gears (gear), and number of carburetors (carb), could be interpreted as either continuous or categorical. There were no missing values in the data set.

Model Selection

Regression model analysis compares these statistics to help us choose the best fit model.

1. residuals - the difference between the observed and fitted data - we want these to be normally distributed around 0
2. significance stars - our statistical software, R marks significant coefficients of our model with one to three asterisks, more is better
3. estimated coefficient - the value of the slope in the regression model (spot check to make sure it seems reasonable)
4. standard error of the coefficient estimate - measures variability in the coefficient estimate (want it to be at least an order of magnitude (power of ten) less than the coefficient estimate)
5. t-value of the coefficient estimate - score used to calculate probability coefficient estimate is significant
6. variable p-value - probability variable is not significant (we want this number to be as small as possible)
7. significance legend - explanation of 2.
8. residual std error/degrees of freedom - standard deviation of the residuals (want this number to be proportional to the quantiles of the residuals for a normal distribution, the 1st and 3rd quantiles should be 1.5 +/- the std error); degrees of freedom is the difference between the number of observations and the number of variables in the model
9. R-squared-metric to evaluate the goodness of fit of the model. Higher is better with 1 being the best. (it's the variability in your prediction explained by the model)
10. F-statistic and p-value-F-test comparing our model to a model that has fewer parameters. (a low p-value means our model performs better)
11. Predicted R-squared-metric to evaluate how well the model predicts the future
12. Variance Inflation Factor-assesses how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be 1. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

We started with a model using transmission only to predict mpg, and the coefficient had a low p-value (0.000285), but it's R-squared is also low(0.3598). Let's to try improve it by adding variables.

We used a correlation table to see how strongly each variable is related to miles per gallon. We added variables with high correlation to mpg to the model and examined the model's statistics. To reduce the effects of multicollinearity, we also looked at the correlation amongst each variable. We omitted variables that were related to each other to reduce the variance of the estimated coefficients, if the predictors were correlated.

Using the top three most correlated variables with mpg, (i.e. wt, cyl, and disp) we created three more models combining each with transmission. These models had significant p-values for the coefficients for wt, cyl, and disp, high R-squared, and F-test values, but all showed that transmission was NOT significant in predicting miles per gallon.

To control the confounding effects of transmission, we'll design models by restricting and matching the other variables. Additionally, we'll try an automatic model selection method and the nested modeling/ANOVA technique. We did not use these models because the transmission coefficient was not significant, the variance inflation factors were too high, indicating the variables were correlated, or other variable's coefficients were not significant.

I arrived at the final model by "Reducing the model" or the practice of including all candidate predictors, and then systematically removing the term with the highest p-value one-by-one until only significant predictors are left.

The Appendix shows the final model residuals, coefficients, p-values, R-squared values (multiple, adjusted, and predicted), F-test values, and variance inflation factors.

The R-squared is 0.8497, The Adjusted R-squared is 0.8336 and the Predicted R-squared is 0.7946.

The F-statistic is 52.75 and the p-value is 1.21e-11.

The variance inflation factor for weight is 2.482952, quarter mile time is 1.364339, and transmission is 2.541437.

Figure 1 in the Appendix shows plots of our model's residuals vs fitted values and Normal Q-Q standardized residuals.

Results

The final parsimonious regression model that easily describes the mean relationship between weight, quarter mile time, transmission and miles per gallon for automatics was

$$mpg = 9.6178 - 3.9165wt + 1.2259qsec$$

and for manuals was

$$mpg = 12.5536 - 3.9165wt + 1.2259qsec$$

The intercept of our model is not significant with a p-value of 0.177915, but we are not interested in the cases where the predictors weight or quarter mile time is zero. There are statistically significant relationships between weight, quarter mile time, transmission and miles per gallon (p-values of 6.95e-06, 0.000216, and 0.046716).

A change of 1,000 pounds in the weight of the car corresponds to a change of -3.9165 in miles per gallon, holding quarter mile time constant. A change of 1 sec in the quarter mile time of the car corresponds to a change of 1.2259 miles per gallon, holding the weight of the car constant. A car with a manual transmission is expected to have a change of 2.9358 miles per gallon.

The 95% confidence intervals of the coefficients of our regression model:

weight -5.37333423 -2.459673

quarter mile time 0.63457320 1.817199

manual transmission 0.04573031 5.825944

With 95% confidence, we estimate a 1,000 pound change in weight results in a -5.3733 to -2.4597 increase in miles per gallon.

With 95% confidence, we estimate a 1 second change in quarter mile time results in a 0.6346 to 1.8172 increase in miles per gallon.

With 95% confidence, we estimate a manual transmission results in a 0.0457 to 5.8260 increase in miles per gallon.

Conclusions

Our analysis arrived at a model that adjusts the confounding transmission variable and suggests a linear model relating weight, quarter mile time, and transmission to miles per gallon.

On average, a manual transmission is better than an automatic transmission.

For a car with weight of 2.168 and quarter mile time of 16.8, the average miles per gallon, if using an automatic transmission is 21.72168. For a manual transmission with the same weight and quarter mile time, the average miles per gallon is 24.65752. It's estimated that this car with a manual transmission will get 2.93584 more miles per gallon than an automatic.

This is based on a limited and outdated sample of cars that don't represent many of the newer vehicles today which use the technology that makes automatic transmission cars equally or more fuel efficient than manual transmissions.

Appendix

Table 1		
Transmission	automatics	manuals
	19	13
Engine	V engine	straight
	18	14

Table 2			
Cylinders	4 cyl	6 cyl	8 cyl
	11	7	14
Gears	3 gear	4 gear	5 gear
	15	12	5

Table 3						
Carburetors	1 carb	2 carb	3 carb	4 carb	6 carb	8 carb
	7	10	3	10	1	1

Final Model Regression Analysis

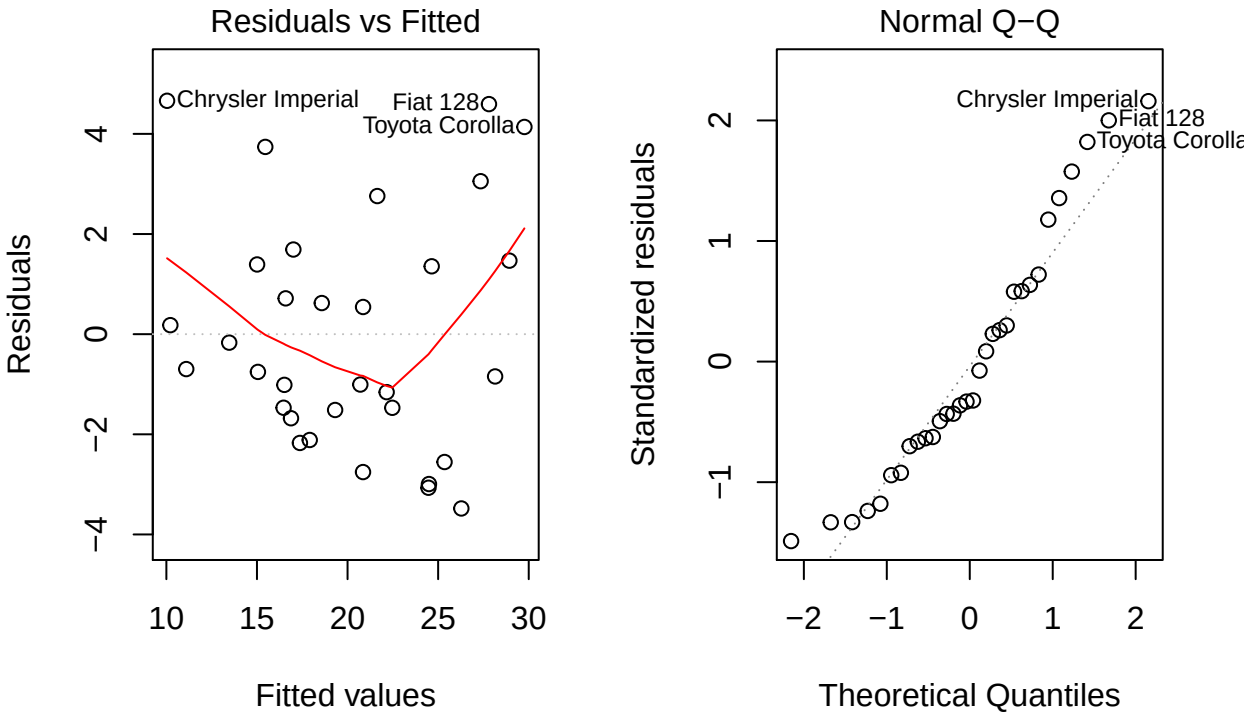
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## factor(am)1   2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
```

```
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```
## [1] 0.7945881
```

```
##          wt        qsec factor(am)
##    2.482952    1.364339    2.541437
```

Figure 1 - Residuals Plots



The mtcarsAnalysis.Rmd file can be found at my github repo here