The Relationship of Transmission to MPG in Cars

Stephanie Lum

February 28, 2016

Executive Summary

In this report, the relationship between the type of automobile transmission and its effect on miles per U.S. gallons (MPG) are investigated. Ultimately, the question of which transmission type (i.e., automatic or manual) results in a better, or larger, MPG will be answered.

The analysis and inverstigation found the following:

- Manual transmission results in a better MPG in automobiles than automatic transmission.
- There may be other factors that will result in better MPG than just transmission type alone.

Data

The source of the data used in this analysis is from the 1974 *Motor Trend* US magazin, which compared fuel consumption and other automobile design variables for 32 automobile models from 1973 to 1974. This data can be found in `mtcars`, a dataset built-in to R, as part of the `datasets` package.

The initial two variables of interest for this investigation and their definitions are as follow:

1. `mpg`: Miles/(US) gallon
2. `am`: Transmission (0 = automatic, 1 = manual)

From the boxplot showing the relationship between the two variables (found in Appendix 1), it can initially be inferred that manual transmission has a better MPG than automatic transmission.

Analysis

Regression Models

A linear regression with the factorized variable `am` as the predictor and the variable `mpg` as the outcome is conducted.

```
fit <- lm(mpg ~ factor(am), data = mtcars)
```

From the summary of the regression (found in Appendix 2), it can be seen that an automobile with manual transmission has an average of 7.25 more MPG than an automobile with automatic transmission. Despite the agreement of this model with the results shown in box and whisker plot in the initial analysis, the adjusted R^2 value of the model was 0.3385, meaning that only 34% of the variance of the residuals were explained. Further models of the data need to be explored.

A multivariable regression of the outcome `mpg` with all the other variables used in `mtcars` as predictors (henceforth known as the full model) is conducted.

```
fit_fullmodel <- lm(mpg ~ ., data = mtcars)
```

From the summary of the regression (found in Appendix 3), the adjusted R^2 value for the full model showed that about 80% of the residual variance can be explained using all the variables as predictor.

Although using the full model provided a significantly better fitted model (based on the adjusted R^2 value) than using only the `am` variable, there is still a risk of using unnecessary variables and overfitting the model. This would result in an increase of the variance estimate than that of a correctly fitted model.

A step-wise search algorithm can be applied to the full model to determine which variables should be included in the linear model. This method works by finding the best model with the least Akaike Information Criterion (AIC). The summary of the results can be found in Appendix 4.

```
fit_fullmodel <- lm(mpg ~ ., data = mtcars)
fit_stepmodel <- step(fit_fullmodel, k = log(nrow(mtcars)))
```

The step-wise search algorithm revealed that the best model (henceforth known as step model) includes `qsec` and `wt` as confounders to `am` as predictors for the outcome `mpg`. The definitions for the variables are as follow:

- `qsec`: 1/4 mile time
- `wt`: Weight (1000 lbs)

The adjusted R^2 value for the step model showed that 83% of the variance can be explained using this model. This is an additional improvement over the full model.

Thus, it is concluded that the step model, `lm(mpg ~ am + wt + qsec, mtcars)`, is the best model for the investigation into the relationship between transmission type and MPG.

Residuals and Diagnostics

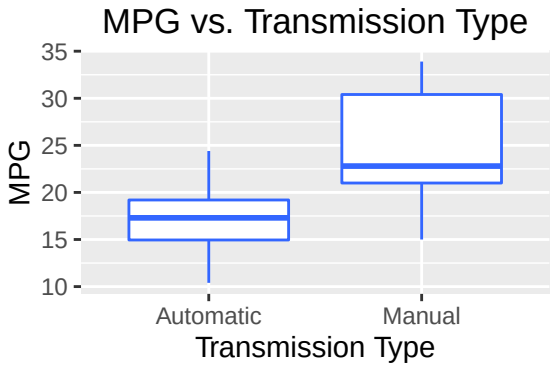
The following diagnostic plots (see Appendix 5) are made and analyses are performed on their behavior:

- The **Residual vs. Fitted plot** showed the characteristics of a well-fitted graph, meaning that no residual stood out (i.e., no outliers). Additionally there should have been no discernable pattern to the points on the plot. Ideally, the regression line across the points would have been horizontal, and the regression line was very close to that on this plot.
- The **Scale-Location plot** showed homoscedasticity, which is the assumption that the variance of the residuals will not change with the change as a function of X. This plot showed a relatively flat line, which reinforced that idea.
- The **Normal Q-Q plot** showed the normality of the errors of the model. It plotted the theoretical quantiles for the standard error vs. the actual quantiles of the standardized residuals. Because the points lied on a linear line, this suggests that the data was normally distributed.
- The **Residuals vs. Leverage plot** showed the amount of leverage that each individual points had with relation to the dotted grey line across the zero axis. Because the regression line stayed close to the dotted line, it suggested that the chosen fit was a good fit with no outliers that may have had high leverage against the other regression points.

Although the diagnostic plots show a good fit, indicating that the chosen model is a decent one, the generalized pairs plot (found in Appendix 6) show that there is a large positive correlation between `wt`, the weight of the automobile and the MPG, `mpg`, and also a large negative correlation between `wt` and `am`. This suggests that the weight of the vehicle may be a greater predictor over just transmission type alone of whether the automobile gets a higher MPG. Futher investigation will be required.

Appendix

Appendix 1. Plot of Relationship Between Transmission Type and MPG



Appendix 2. Initial Regression Summary

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17.147      1.125  15.247 1.13e-15 ***
## factor(am)Manual    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Appendix 3. Full Model Regression Summary

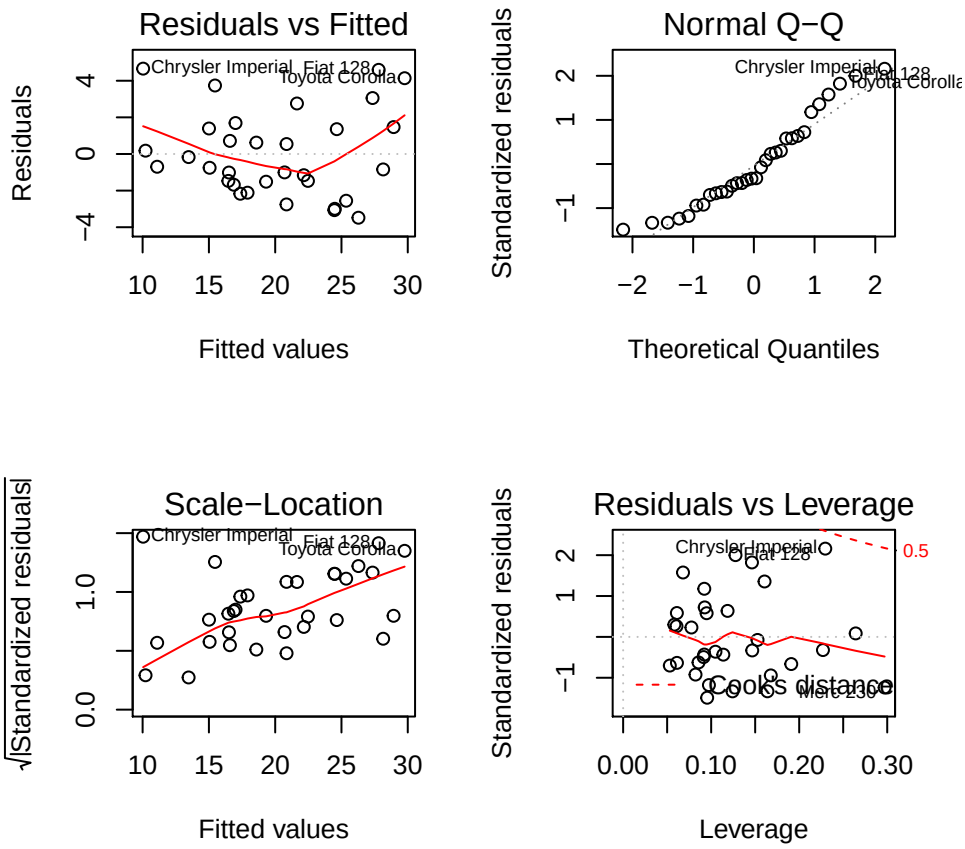
```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
```

```
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Appendix 4. Step Model Regression Summary

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + 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 ***
## amManual      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
```

Appendix 5. Diagnostic Plots



Appendix 6. Pairs Plot for Predictors in Step Model