Gas Mileage Regression Analysis

Coursera: Regression Models

06/21/2014

Executive Summary

This report will analyse the data provided in the Motor Trend dataset (`mtcars`), found here http://bit.ly/1svZI4M, and provide some insight into the following two research questions:

1. Is an automatic or manual transmission better for MPG?
2. How different is the MPG between automatic and manual transmissions?

Linear regression models and other techniques will be applied to the data to explore and quantify any differences between automatic and manual trasmission vehcles.

Exploratory Analysis

The `mtcars` dataset contains 32 rows (instances) and 11 columns (attributes). Each instance represents a vehicle type and the attributes describe various properties of the vehicles. This analysis will focus specifically on `mpg` (miles per gallon) and `am` (transmission type) which are continuous and binary categorical attributes, respectively. `am` takes on the following values: `0 = automatic, 1 = manual`. `mpg` is the response variable and `am` is the explanatory variable. Of the 32 vehicles, 19 of them have automatic transmissions and 13 have manual transmissions. The boxplot in the appendix shows the spread of the `mpg` variable with respect to each transmission type. It is clear from the boxplot that some difference exists and that it may be significant. Upon first glance we can see that the medians of both groups are quite far from each other and the IQR's of each transmission type do not overlap. This means the middle 50% of each group is complete separated from each other. Additionally there do not seem to be any outliers at this point in the analysis. We can also see that the variance of the `automatic` group is smaller than the variance of the `manual` group. In the next section we will walk through a regression analysis and formally quantify the differences between the two groups.

Regression Analysis

As stated earlier, `mpg` is the response variable and `am` is the explanatory varible. We will begin by looking at the most basic model first and then consider interactions and confounding variables. `am = 'automatic'` is the reference level so the linear model we use is:

$$Y_i = \beta_0 + I(X_i = manual)\beta_1 + \beta_2 \cdot X_i + I(X_i = manual) \cdot \beta_3 \cdot X_i$$

where $I()$ is an indicator function which equals 1 if the condition is met.

Analysis 1 in the appendix shows the summary of the model based on the equation above. **The expected `mpg` of an automatic transmission vehicle is** 17.7 **and the expected `mpg` for a manual transmission vehicle is** 24.4**.** Figure 2 shows the Q-Q plot of the residuals and they are near-normally distributed as they should be. At first glance it seems like driving a manual transmission vehicle gives you an extra 7 miles per gallon, this is great news, however, the finer details of the model suggest a deeper investigation is needed. The adjusted R^2 value is a low .3385 meaning the model does not do a good job explaining the variance seen in `mpg`. The rest of this report will consider adjusted R^2 values over standard R^2 as a metric for model selection since the former penalizes complex models and the latter increases monotonically

Confounding Variables and Model Selection

There are 9 variables in the `mtcars` dataset, each of which may have an impact on `mpg`. I know from experience that the weight of a car and the horsepower of an engine both decrease gas mileage as they increase. The `mtcars` dataset abbreviates these two variables as `wt` and `hp`, respectively. Figure 3 in the appendix plots the distributions of weight and horsepower against transmission type. In both cases we can see that vehicles in this dataset with a automatic transmission also tend to have larger weights and higher horse power. This raises the obvious question: Do these two variables have a larger impact on gas mileage than transmission type? In an effort to keep this report short, I am going to assume the other variables have no effect.

Analysis 2 and 3 in the appendix aims to answer the new question. We fit three models. The base model is the model we have used up to this point and the two successive models each add one additional variable, weight followed by horse power. Two methods are used to evaluate if these two variables should be added, the first (Analysis 2) is based on `anova` and the second (Analysis) is based on adjusted R^2 values. The p-values (Analysis 2) of the successive models in the `anova` suggest that the new variables are significant at the 95% significance level. Reinforcing this result is Analysis 3 which shows the adjusted R^2 value increasing significantly for each new variable, more than doubling when `wt` is included.

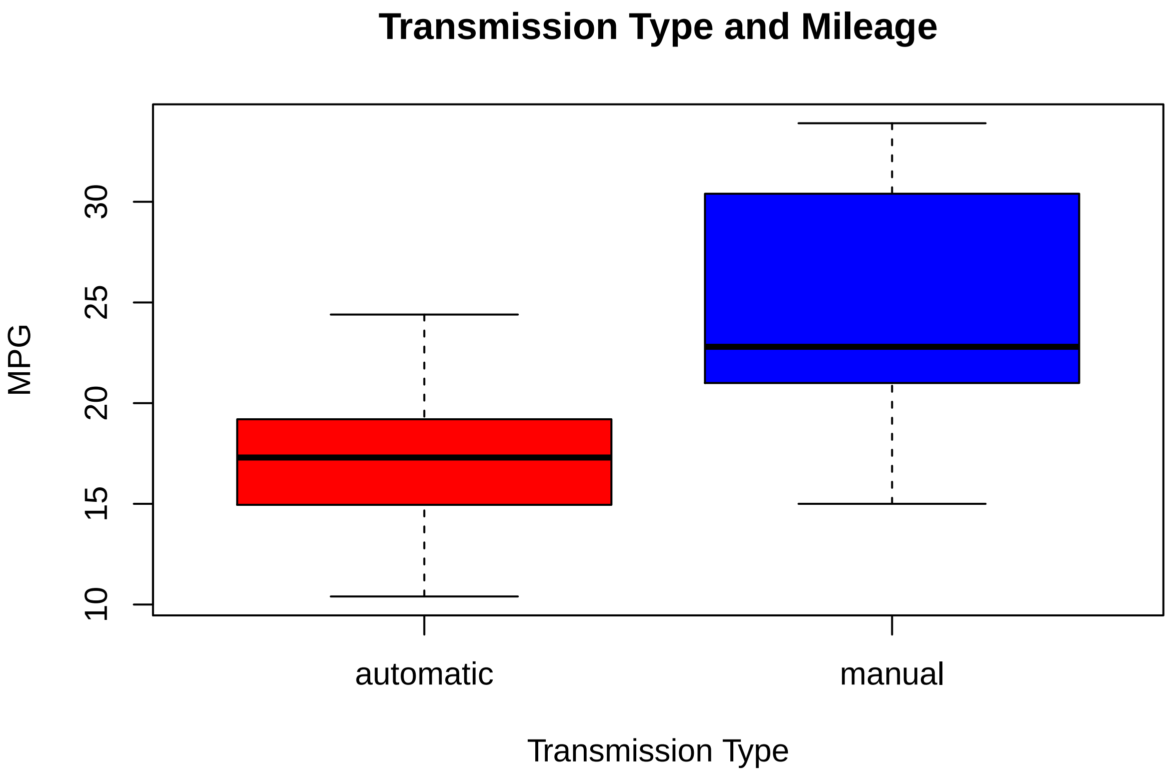
Conclusion

Analysis 4 shows the summary of the final model which includes `am`, `wt`, and `hp` as regressors of `mpg`. The adjusted R^2 is much better than the original model, .823 versus .338 and we get additional insights into the true effect of transmission type on gas mileage. Interpreting the properties of the final model with respect to the orignal research questions, we arrive at the following conclusion: **Holding weight and horsepower constant, switching to a manual transmission increases gas mileage by 2 mpg when compared to an automatic transmission. The p-value is only significant at the 85% confidence level so this may not be the strongest result.** These results may be different if we did not make the assumption that all other variables had no effect on mpg but recall that this assumption was made to avoid a long analysis. Only a couple examples were needed to demonstrate the role of confounding variables in model selection.

Appendix

```r
# setup the dataset
data(mtcars)
mtcars$am[mtcars$am == '0'] = 'automatic'
mtcars$am[mtcars$am == '1'] = 'manual'
groups = split(mtcars, mtcars$am)
```

```r
# ----- Figure 1
boxplot(mpg ~ am, data=mtcars, col=c('red', 'blue'), ylab='MPG', xlab='Transmission Type', main='Transmi
```



```r
summary(groups$automatic$mpg) # summary of mpg for automatic transmission vehicles
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.4    15.0    17.3    17.1    19.2    24.4
```

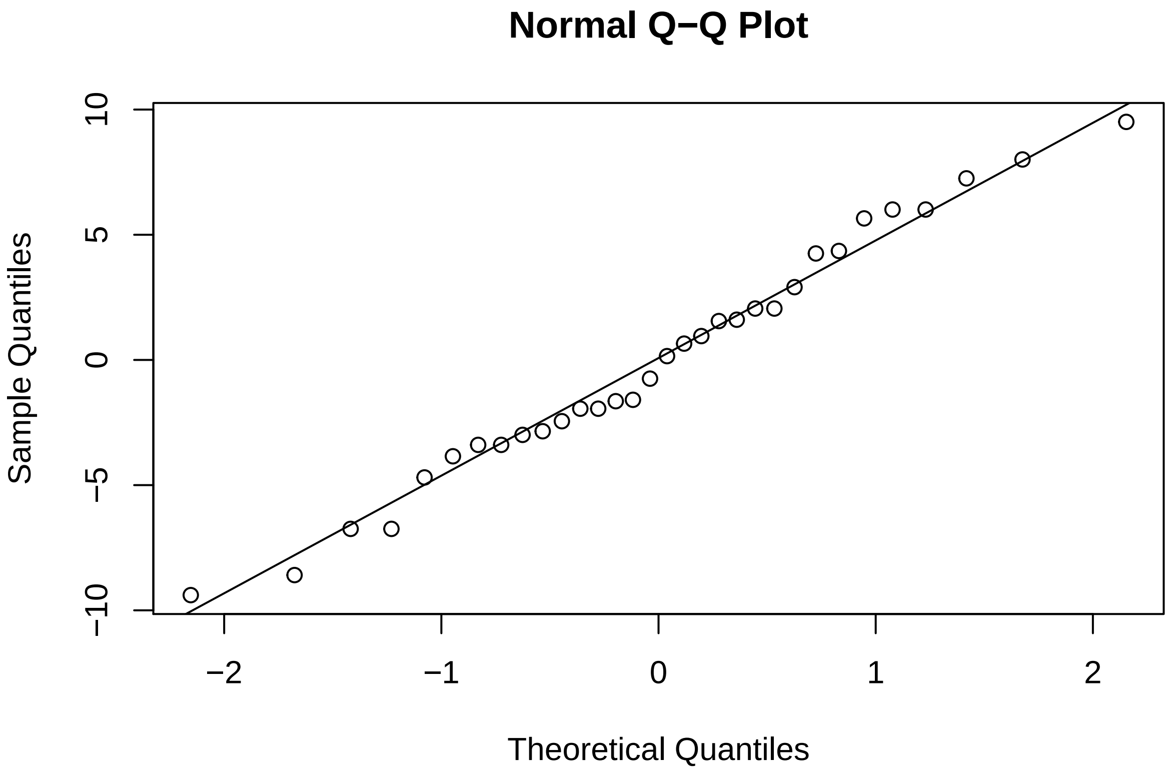
```r
summary(groups$manual$mpg) # summary of mpg for manual transmission vehicles
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0    21.0    22.8    24.4    30.4    33.9
```

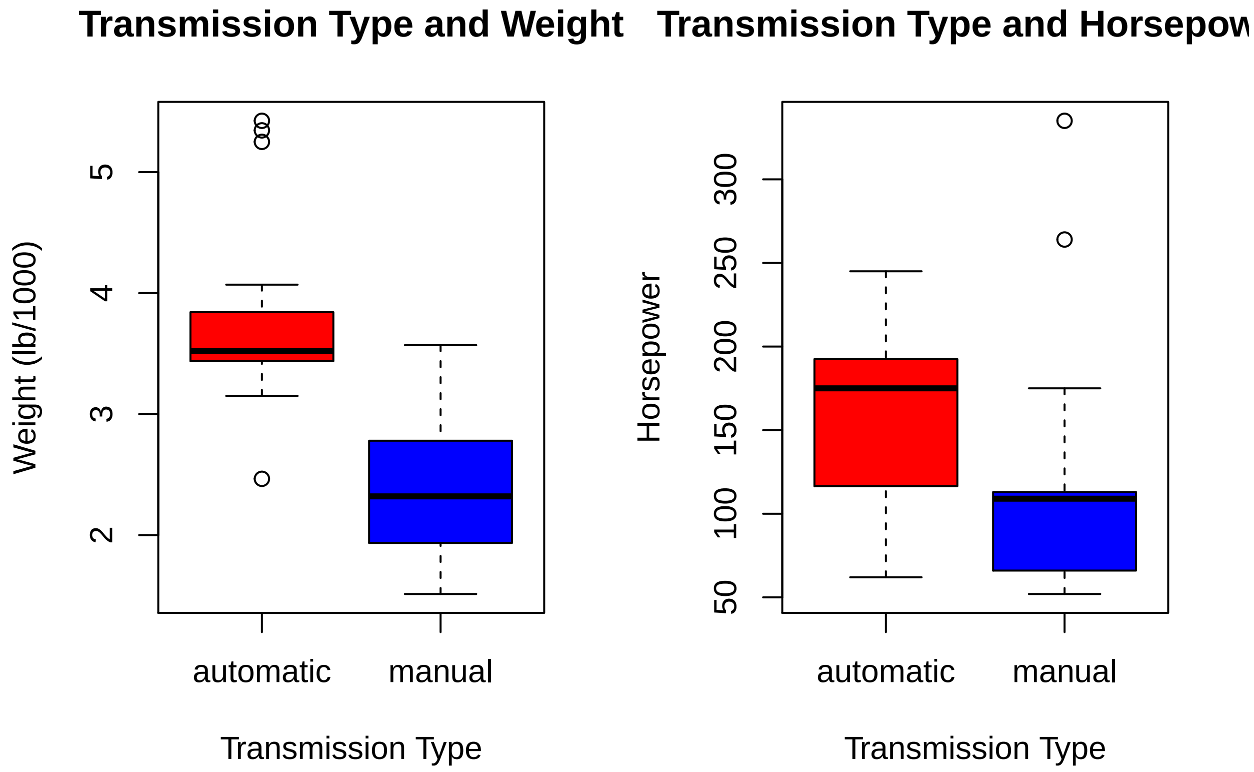
```r
# basic linear regression with mpg as the response and am as the regressor
fit = lm(mpg ~ as.factor(am), data=mtcars)

# ----- Analysis 1
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(am), data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              17.15       1.12   15.25  1.1e-15 ***
## as.factor(am)manual       7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

```r
# ----- Figure 2
qqnorm(fit$residuals); qqline(fit$residuals)
```



```
# ----- Figure 3
par(mfrow=c(1, 2))
boxplot(wt ~ am, data=mtcars, col=c('red', 'blue'), ylab='Weight (lb/1000)', xlab='Transmission Type', m
boxplot(hp ~ am, data=mtcars, col=c('red', 'blue'), ylab='Horsepower', xlab='Transmission Type', main=''
```



```
fit1 = lm(mpg ~ as.factor(am), data=mtcars)
fit2 = lm(mpg ~ as.factor(am) + wt, data=mtcars)
fit3 = lm(mpg ~ as.factor(am) + wt + hp, data=mtcars)

# ----- Analysis 2
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ as.factor(am)
## Model 2: mpg ~ as.factor(am) + wt
## Model 3: mpg ~ as.factor(am) + wt + hp
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     29 278  1       443 68.7 5.1e-09 ***
## 3     28 180  1        98 15.2 0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ----- Analysis 3
# Adjusted R^2 values for each successive model from above
print(c(summary(fit1)$adj.r.squared, summary(fit2)$adj.r.squared, summary(fit3)$adj.r.squared))
```

```
## [1] 0.3385 0.7358 0.8227
```

```
# ----- Final Analysis 4
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(am) + wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.422 -1.792 -0.379  1.225  5.532
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         34.00288    2.64266   12.87  2.8e-13 ***
## as.factor(am)manual  2.08371    1.37642    1.51  0.14127
## wt                  -2.87858    0.90497   -3.18  0.00357 **
## hp                  -0.03748    0.00961   -3.90  0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.54 on 28 degrees of freedom
## Multiple R-squared:  0.84,   Adjusted R-squared:  0.823
## F-statistic:   49 on 3 and 28 DF,  p-value: 2.91e-11
```