Motor Trend Regression Analysis

May 5, 2016

Executive Summary

Using a collection of 32 data points for a variety of vehicles, we will attempt to answer the following questions:

- Is an automatic or manual transmission better for MPG?
- If the manual transmission has a significant impact on MPG, what is the quantifiable impact?

In order to answer these two questions, we will take the following steps:

- Import and pre-process the data
- Perform initial qualitative analysis through exploratory graphs
- Select an appropriate model to investigate the impact to MPG
- Analyze the model results for quality of fit
- Quantify significance and impact to MPG in our selected model

Importing Data

The data in the `mtcars` dataset has eleven fields. All the fields are represented as numeric data in the original data set. In order to simplify the model building process, we will create factor variables where appropriate.

```r
data(mtcars)

# Create copy of mtcars with factored variables
mtcarsF <- mtcars
mtcarsF$cyl <- factor(mtcarsF$cyl, levels=c(4, 6, 8))
mtcarsF$am <- factor(mtcarsF$am)
levels(mtcarsF$am) <- c("auto", "manual")
mtcarsF$gear <- factor(mtcarsF$gear, levels=c(3, 4, 5))
mtcarsF$carb <- factor(mtcarsF$carb, levels=c(1,2,3,4,5,6,7,8))
mtcarsF$vs <- factor(mtcarsF$vs)
levels(mtcarsF$vs) <- c("V", "S")
```

Exploratory Analysis

Let's take an initial look at the `mtcars` dataset.

```r
dim(mtcarsF)
```

```
## [1] 32 11
```

```r
head(mtcarsF)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs     am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  V manual    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  V manual    4    4
```

```
## Datsun 710         22.8   4  108  93 3.85 2.320 18.61  S manual    4    1
## Hornet 4 Drive     21.4   6  258 110 3.08 3.215 19.44  S   auto    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  V   auto    3    2
## Valiant            18.1   6  225 105 2.76 3.460 20.22  S   auto    3    1
```

As an initial exploratory tool, we can create a simple scatterplot of MPG vs transmission type (Figure 1 in Appendix):

```
library(lattice)
figure1 <- dotplot(mpg~ am, data=mtcars, main="MPG vs Transmission Type")
```

From this plot, it seems that there is a trend, but we need to control for confounding variables. We can explore potential confounding variables by looking at correlation:

```
corToMPG <- cor(mtcars$mpg, mtcars)
# Order correlation terms
(corToMPG <- corToMPG[,order(-abs(corToMPG[1,]))])
```

```
##        mpg         wt        cyl       disp         hp       drat
##  1.0000000 -0.8676594 -0.8521620 -0.8475514 -0.7761684  0.6811719
##         vs         am       carb       gear       qsec
##  0.6640389  0.5998324 -0.5509251  0.4802848  0.4186840
```

Note that there are many variables showing higher correlation to `mpg` than transmission type (`am`). This suggests that there may be strong confounding factors affecting the impact of transmission on MPG.

Model Selection

To start with, let's try a simple linear regression with `mpg` as the outcome and `am` as the regressor:

```
simpleModel <- lm(mpg ~ am, data=mtcarsF)
```

We can see from the summary (see appendix) that the p-value is pretty low. However, the R-squared is very poor. Based on this, the model is probably not the best fit. Let's try adding in all the variables (summary in appendix):

```
complicatedModel <- lm(mpg ~ .-1, data=mtcars)
```

Now that all the variables are included in the model, none of the p-values are very significant, and the R-squared value is only marginally improved.

The selection of regressor variables can be simplified using the step method (summary in appendix):

```
newModel <- step(complicatedModel, direction="both", trace=F)
```

It looks like the step method included engine displacement (`disp`), but the p-value is not very significant. Let's manually run one more iteration with that variable removed:

```
mtcars$am <- factor(mtcars$am)
levels(mtcars$am) <- c("auto", "manual")
finalModel <- lm(mpg ~ am + wt + qsec, data=mtcars)
```

Model Examination

The final model we selected includes transmission type (`am`) as well as weight (`wt`) and quarter mile time (`qsec`). All have p<0.05, and the R-squared value is also pretty good.

We can take a look at the quality of the fit using `autoplot()` (Figure 2 in appendix):

The normal Q-Q plot appears to be pretty well fitted, but the residual values appear to have quite wide distributions.

We can also construct a confidence interval on the impact of transmission type:

```
confint.lm(finalModel, level=0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## ammanual     0.04573031  5.825944
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
```

Conclusions

Based on the analysis above, we could conclude that a manual transmission is in fact better for MPG performance. However, it is tough to truly quantify the impact.

Based on the `finalModel` fit, we could conclude with 95% confidence that a manual transmission can squeeze out 0.05 - 5.82 MPG more than an automatic transmission. However, the residual plots suggest that there may be more going on that we have not accounted for in the model.

It is possible that the fit appears poor due to the small number of samples (32), or that the data collection did not include variables that confound the MPG regression.

Appendix

Model #1 Summary

```
summary(simpleModel)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcarsF)
##
## 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 ***
## ammanual       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
```

Model #2 Summary

```r
summary(complicatedModel)
```

```
##
## Call:
## lm(formula = mpg ~ . - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7721 -1.6249  0.1699  1.1068  4.4666
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## cyl   0.35083    0.76292   0.460   0.6501
## disp  0.01354    0.01762   0.768   0.4504
## hp   -0.02055    0.02144  -0.958   0.3483
## drat  1.24158    1.46277   0.849   0.4051
## wt   -3.82613    1.86238  -2.054   0.0520 .
## qsec  1.19140    0.45942   2.593   0.0166 *
## vs    0.18972    2.06825   0.092   0.9277
## am    2.83222    1.97513   1.434   0.1656
## gear  1.05426    1.34669   0.783   0.4421
## carb -0.26321    0.81236  -0.324   0.7490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 22 degrees of freedom
## Multiple R-squared:  0.9893, Adjusted R-squared:  0.9844
## F-statistic:   203 on 10 and 22 DF,  p-value: < 2.2e-16
```

Model #3 Summary

```r
summary(newModel)
```

```
##
## Call:
## lm(formula = mpg ~ disp + wt + qsec + am - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7169 -1.4638 -0.5382  1.7825  4.3566
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
```

```
## disp  0.012020   0.008891   1.352 0.187238
## wt   -4.612795   1.158173  -3.983 0.000440 ***
## qsec  1.705510   0.127486  13.378  1.1e-13 ***
## am    4.180854   1.013616   4.125 0.000301 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.462 on 28 degrees of freedom
## Multiple R-squared:  0.9879, Adjusted R-squared:  0.9862
## F-statistic: 572.1 on 4 and 28 DF,  p-value: < 2.2e-16
```

Figure 1

```
figure1
```

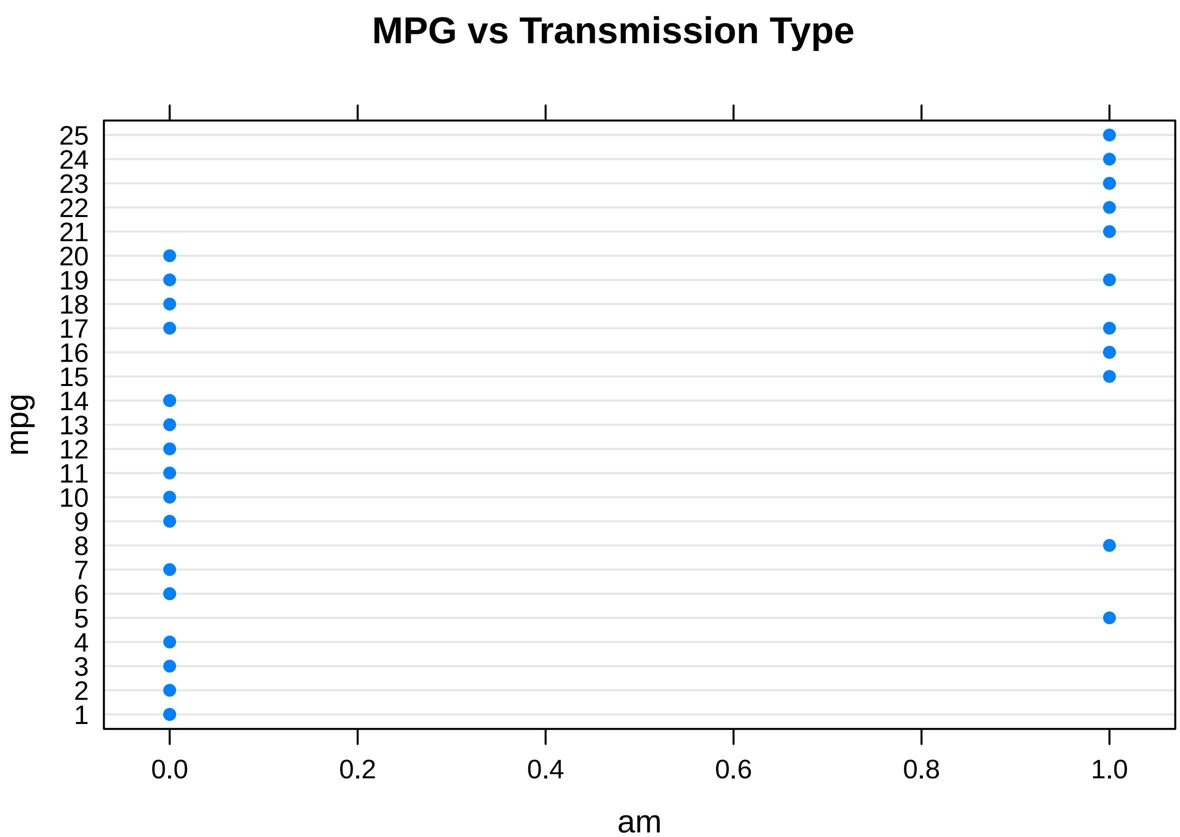

Figure 2

```
library(ggplot2)
library(eeptools)
```

```
## Warning: package 'eeptools' was built under R version 3.2.5
```

```
figure2 <- autoplot(finalModel)
```