Analysis of Factors Contributing to Vehicle MPG

Austin Tidmore

September 26, 2015

Executive Summary

As an analyst for Motor Trend magazine I am providing this report as an analysis of factors contributing to vehicle fuel efficiency (i.e. "MPG"). The analysis herein has been compiled from a dataset of 32 vehicles from 1973-74.

My research found that while a Manual transmission (as opposed to Autmatic) results in significantly better fuel efficiency, there were other variables in the data which explain the relationship with MPG better than just transmission type.

The initial model shows an increase of 7.2 MPG for manual transmission cars but cannot explain much of the variability in MPG. The final model shows an increase of 2.9 MPG for manual transmission vehicles when weight and quarter-mile time are considered.

To ensure that the results are legitimate, the Appendix (Figure 4) contains a plot of the model diagnostics. The residual values look randomly distributed (a sign of a good model), and the diagnostics do not show anything out of the ordinary.

Exploring the Data

First we need to explore the key points of the dataset as well as start to understand the effect of Transmission type on MPG rating. The data will be loaded and slightly cleansed for analysis

```
data(mtcars)
mtcars$am <- as.factor(mtcars$am)
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
require(ggplot2)
```

A summary of the data has been provided in the Appedix, Figure 1. The "am" field refers to the transmission type where "0" represents Automatic and "1" represents Manual. MPG values are in Miles per U.S. gallon.

The Boxplot (Appendix, Fig. 2) indicates that the median MPG rating for manual transmission vehicles is higher. Also due to the non-overlap of the interquartile ranges I suspect that the higher median MPG for manuals would be statistically significant. This will be verified in the following section.

Since I suspect that more variables other than transmission type could help predict better MPG, I have analyzed a pairs plot for all variables in the dataset but have included a plot for only a few of the seemingly significant variables: weight, transmission and quarter-mile time. This is Figure 3.

Modeling the Relationships

To quantify the increased MPG as it relates to transmission type we can build a simple linear model with MPG as the outcome and "am" as the regressor. The model's p-values will indicate a significant difference among transmission-type values as they are able to explain MPG values.

```r
base.lm <- lm(mpg ~ am, data = mtcars)

summary(base.lm)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

A function to represent these results is: `MPG = 17.147 + 7.244*AM (for AM = 1 or 0)` This means that all things equal, a manual transmission car gets 7.244 MPG better than its automatic counterpart. I am not quite satisfied with the analysis, however, because the R-squared value indicates that only 0.3597989 (or 36%) of the variability in MPG has been explained by transmission type.

For increased robustness in the model I created a linear model with all available variables, then I utilized R's ability to perform stepwise regression to build the "best" model.

```r
base.lm.all <- lm(mpg ~ . ,data = mtcars)

best.lm <- step(base.lm.all, direction = "backward")
```

```r
summary(base.lm.all)$coef
```

```
##               Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 17.81984325 16.30602324  1.09283809 0.28745417
## cyl6        -1.66030673  2.26229662 -0.73390320 0.47152449
## cyl8         1.63743980  4.31573451  0.37941162 0.70838075
## disp         0.01391241  0.01740176  0.79948296 0.43340363
## hp          -0.04612835  0.02712018 -1.70088685 0.10446190
## drat         0.02635025  1.67648954  0.01571752 0.98761549
## wt          -3.80624757  1.84664309 -2.06117121 0.05252853
## qsec         0.64695710  0.72195025  0.89612421 0.38084614
## vs1          1.74738689  2.27267212  0.76886889 0.45095593
## am1          2.61726546  2.00474936  1.30553251 0.20653091
## gear         0.76402917  1.45668015  0.52450029 0.60569589
## carb         0.50935118  0.94244181  0.54045902 0.59484874
```

The model that includes all available variables to predict MPG is peculiar. None of the individual regressors are significant (according to their respective p-values) yet the entire model is significant. This is because our model has been overfit: numerous inter-related variables almost always guarantee a significant model.

```r
summary(best.lm)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am1          2.935837  1.4109045  2.080819 4.671551e-02
```

Performing step-wise regression on the model has shown that the most accurate way to predict MPG values is to consider vehicle weight, transmission type, and quarter-mile time. This model explains nearly 85% of the variance of MPG which is a great improvement from the initial model.

```r
anova(base.lm.all, base.lm, best.lm)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model 2: mpg ~ am
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
## 1     20 133.32
## 2     30 720.90 -10   -587.57  8.8143 2.249e-05 ***
## 3     28 169.29   2    551.61 41.3739 7.809e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of our ANOVA comparison show that R generated the best linear model: where MPG is an outcome of weight ("wt"), transmission type ("am") and quarter-mile time ("qsec").

Appendix

Figure 1: Dataset Summary

```
## [1] "MPG:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.42   19.20   20.09   22.80   33.90
```

```
## [1] "Transmission:"
```

```
##  0  1
## 19 13
```

```
## [1] "Vehicle Weight:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.513   2.581   3.325   3.217   3.610   5.424
```

```
## [1] "Quarter-mile Time:"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.50   16.89   17.71   17.85   18.90   22.90
```

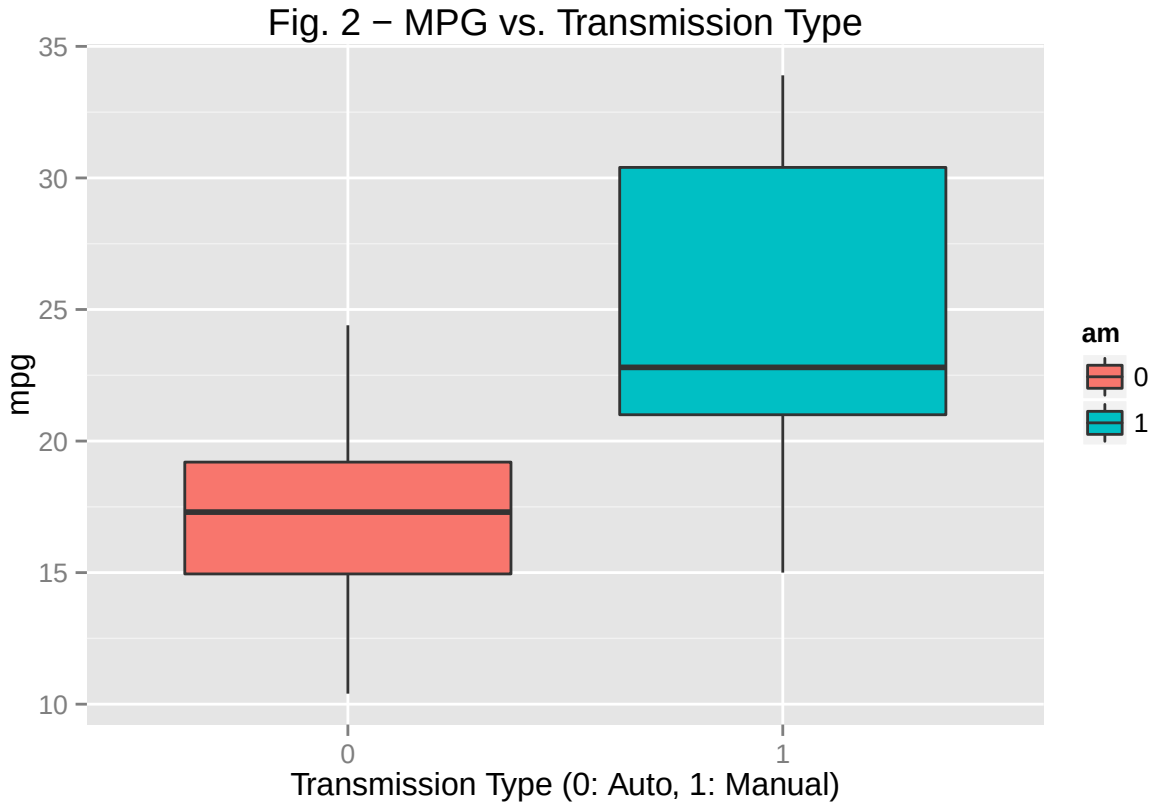


Fig. 3 - Pairs Plot

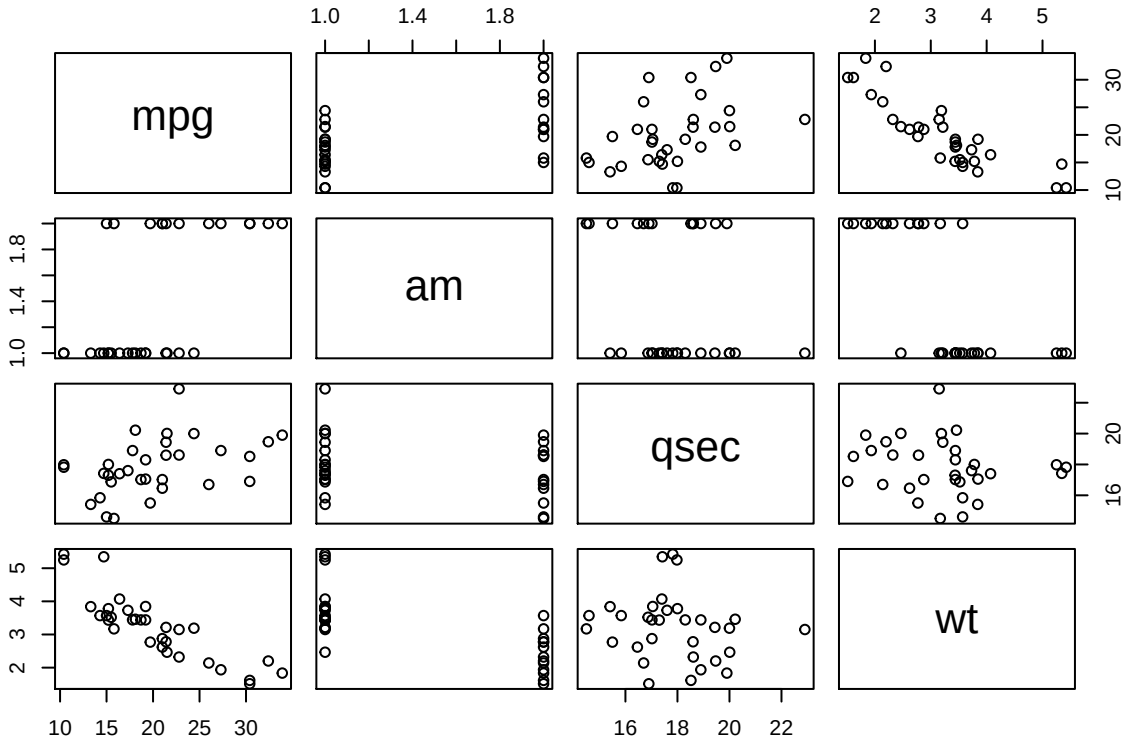
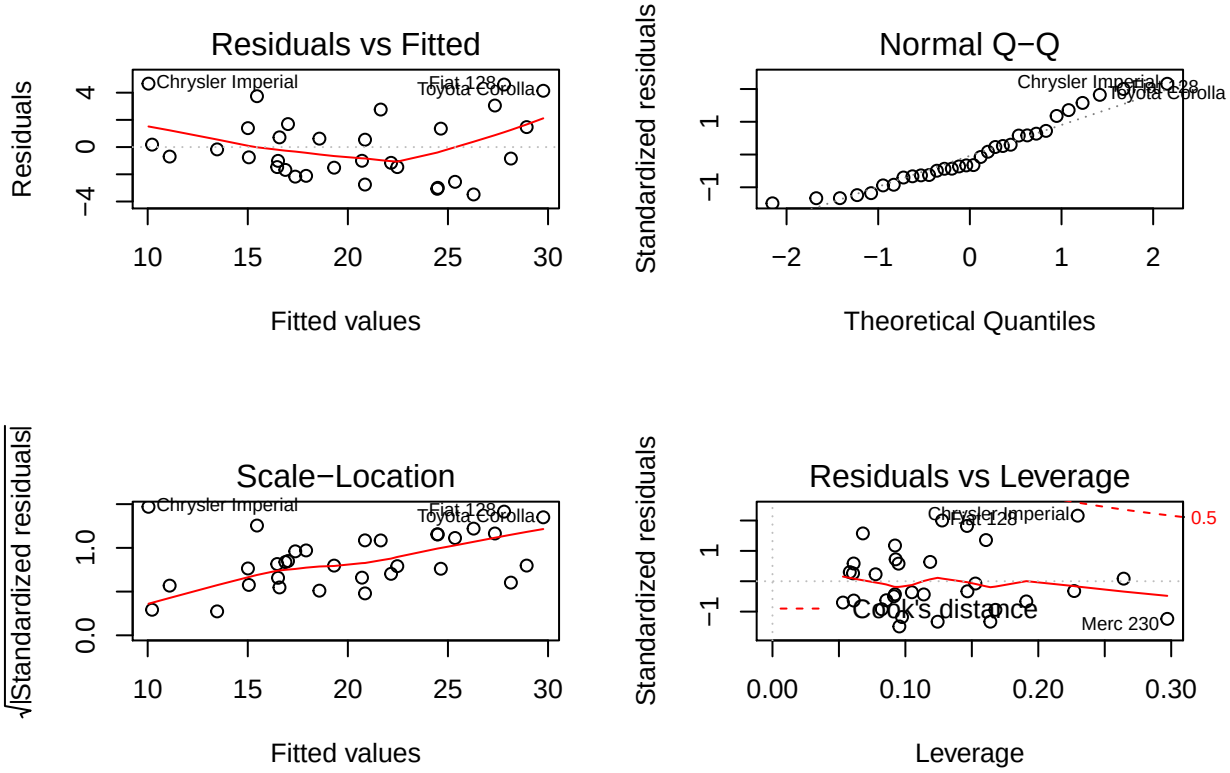


Fig. 4 - Best Model Diagnostics


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage