Motor Trend MPG Analysis

Matt Dancho

July 8, 2016

Executive Summary

This analysis reviews data on a collection of vehicles to explore the relationship of various factors to fuel efficiency as measured in miles per gallon (MPG). Two questions are focused on:

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

This analysis concludes that transmission type is a confounding variable. While a manual tranmission appears to improve MPG (`model1` indicates a 7.24 MPG improvement in manual over automatic transmission) the true relationship is based on weight. Heavier vehicles tend to have automatic transmissions and lighter vehicles tend to have manual transmissions. In the optimal model, `model3`, transmission type is not included since weight is the driving factor of MPG.

Data Processing

The data and libraries are loaded first. The data comes from the `mtcars` data set, which is part of the `R` base package.

```r
# Load data
data("mtcars")
library(ggplot2)
library(knitr)
```

The first six rows of the data set are shown in **Table 1**. We will explore the relationship of transmission type to fuel efficiency. Transmission type is denoted `am`, where 0 = automatic and 1 = manual. Fuel efficiency is denoted `mpg` and is measured in miles per gallon.

```r
kable( head(mtcars), caption = "`mtcars` Dataset: First Six Rows" )      # See Appendix
```

Let's check out the data to see the format. Several of the features are factors incorrectly formatted as numbers.

```r
str(mtcars)     # See Appendix
# Factor features
mtcars.processed <- mtcars
mtcars.processed$am <- as.factor(mtcars.processed$am)
mtcars.processed$cyl <- as.factor(mtcars.processed$cyl)
levels(mtcars.processed$am) <- c("Automatic", "Manual")
```

Exploratory Analysis

Next, an exploratory analysis is performed on the data set. From **Figure 1**, it appears that manual transmissions yield a better fuel efficiency than automatic. However, further analysis is needed to support the significance of this trend.

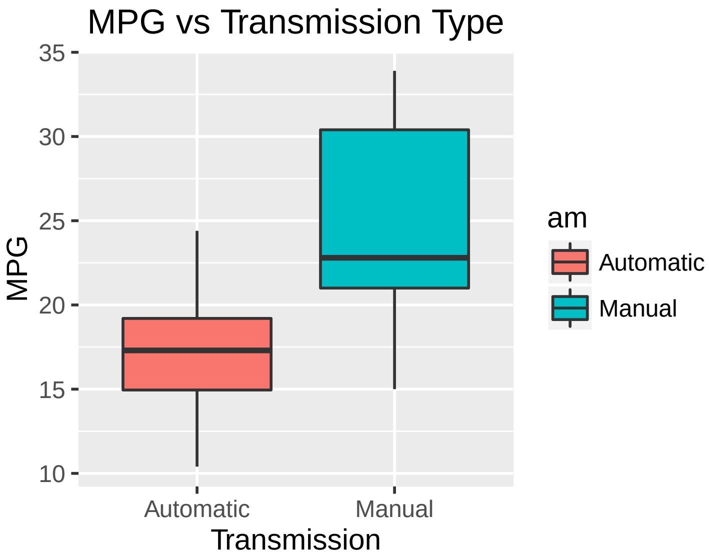


Figure 1: MPG vs Transmission Type

```
ggplot(data=mtcars.processed, aes(x=am, y=mpg)) +
        geom_boxplot(aes(fill=am)) +
        ggtitle("MPG vs Transmission Type") +
        xlab("Transmission") + ylab("MPG")
```

Regression Modeling

First, a model is developed only looking at transmission type. A summary shows that manual transmission results in a 7.245 mpg increase versus the baseline of automatic. The Pr(>|t|) is below 0.001 indicating 99% significance. However, the adjusted R-squared value is low at 0.3385 indicating that only 34% of the variance is explained by the transmission type. This could be indicative that the model needs further analysis. (See Appendix for summary and residual plots).

```
model1 <- lm(mpg ~ am, data=mtcars.processed)
summary(model1)          # See Appendix
```

A second model was developed adding in vehicle weight. The model indicates that a one unit increase in weight (1000 lbs) results in a 5.35 reduction in MPG. The Adjusted R-squared value spikes to 0.7358. Further, the new model indicates that transmission is no longer statistically significant. The coefficient of manual transmission type is very close to zero. This suggests that weight is more predictive, and transmission type could be a confounding variable. The residuals plot shows a bowed pattern, which indicates that the model fit could be improved. (See Appendix for summary and residual plots).

```
model2 <- lm(mpg ~ am + wt, data=mtcars.processed)
summary(model2)          # See Appendix
```

To better understand the trend, the fuel efficiency by weight was plotted in **Figure 2**. The scatter plot points were colored by transmission type. From the plot, weight and fuel efficiency appear closely related. It appears that lighter vehicles tend to have manual transmissions whereas heavier vehicles tend to have automatic transmissions. Further, it appears that the relationship between fuel efficiency and weight may be quadratic in nature.

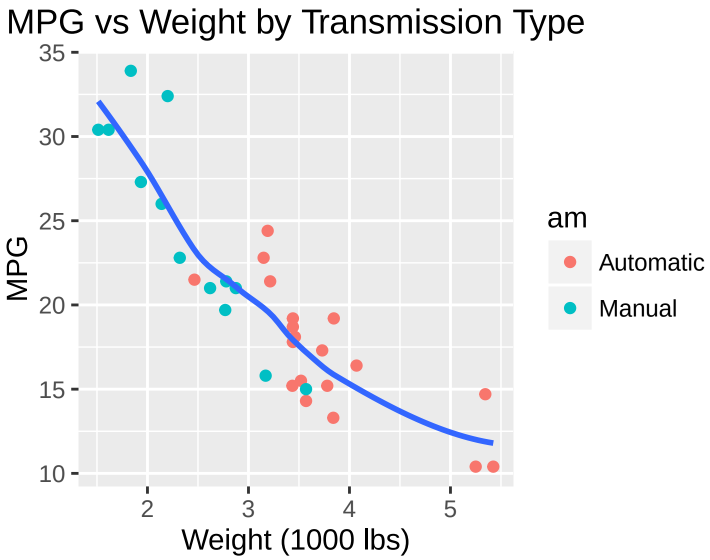


Figure 2: MPG vs Weight by Transmission Type

```r
ggplot(data=mtcars.processed, aes(x=wt, y=mpg)) +
        geom_point(aes(color=am)) +
        geom_smooth(se=FALSE) +
        ggtitle("MPG vs Weight by Transmission Type") +
        xlab("Weight (1000 lbs)") + ylab("MPG")
```

A final model was constructed using a second order quadratic of weight alone for the best combination of predictability and simplicity. The resultant adjusted R-squared is 0.8066 indicating over 80% of the variance of the modeled data is explained. The residuals plot looks much more uniform with no apparent pattern, indicating an optimal model has been achieved. (See Appendix for summary and residual plots).

```r
model3 <- lm(mpg ~ poly(wt, 2), data=mtcars.processed)
summary(model3)         # See Appendix
```

Conclusions

Transmission type is a confounding variable. While a manual tranmission appears to improve MPG (`model1` indicates a 7.24 MPG improvement in manual over automatic transmission) the true relationship is based on weight. As shown in **Figure 2**, heavier vehicles tend to have automatic transmissions and lighter vehicles tend to have manual transmissions. In the optimal model, `model3`, transmission type is not included since weight is the driving factor of MPG.

Appendix

Appendix - Data Processing

```r
kable( head(mtcars), caption = "`mtcars` Dataset: First Six Rows" )
```

Table 1: `mtcars` Dataset: First Six Rows

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

```r
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Appendix - Regression Modeling

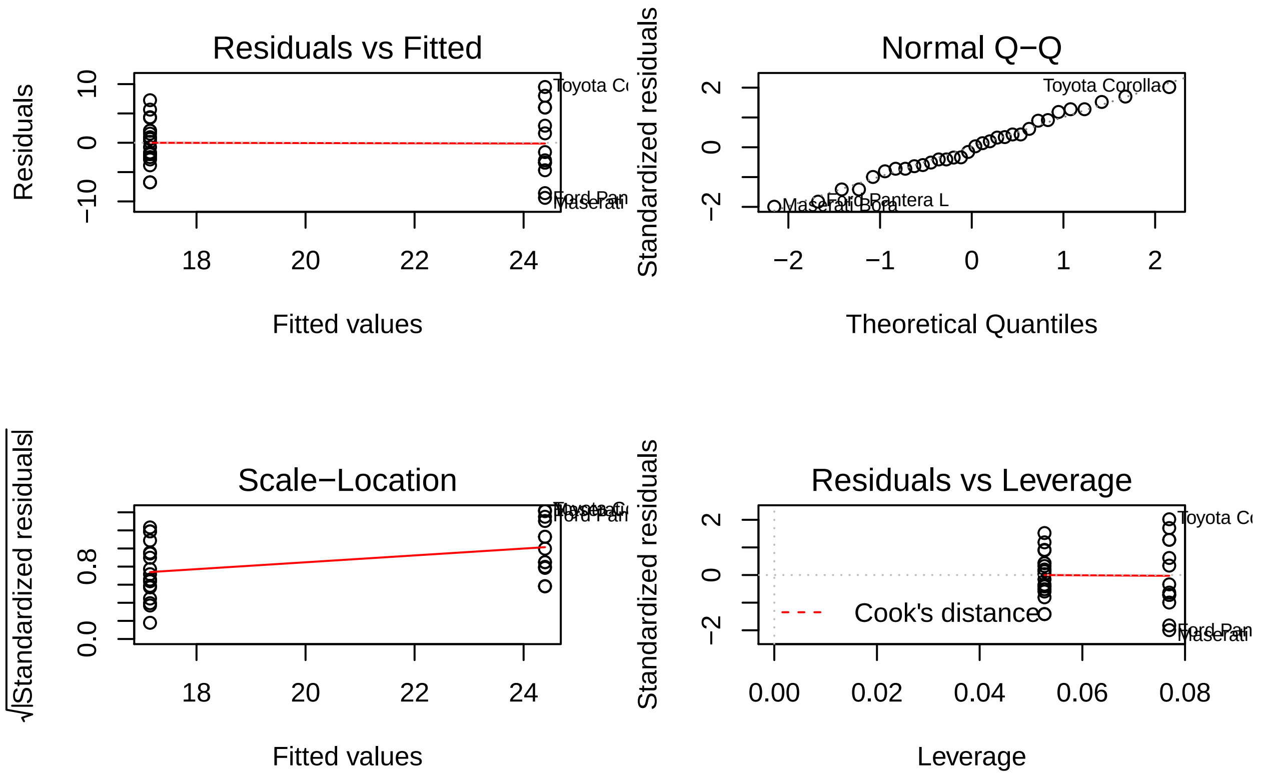
Formula: mpg ~ am

```r
summary(model1)
```

```
## 
## Call:
## lm(formula = mpg ~ am, data = mtcars.processed)
## 
## 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
```

```r
par(mfrow = c(2,2))
plot(model1)
```

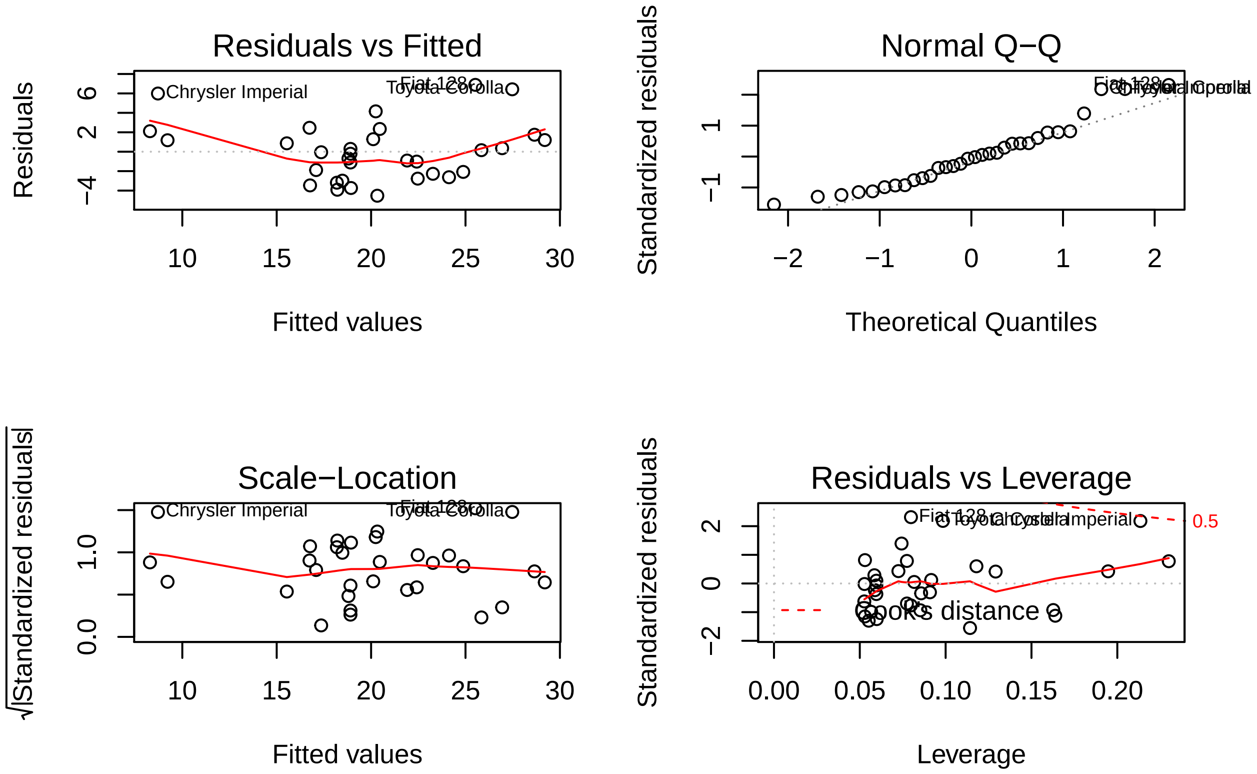


Formula: mpg ~ am + wt

```r
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## amManual    -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```r
par(mfrow = c(2,2))
plot(model2)
```

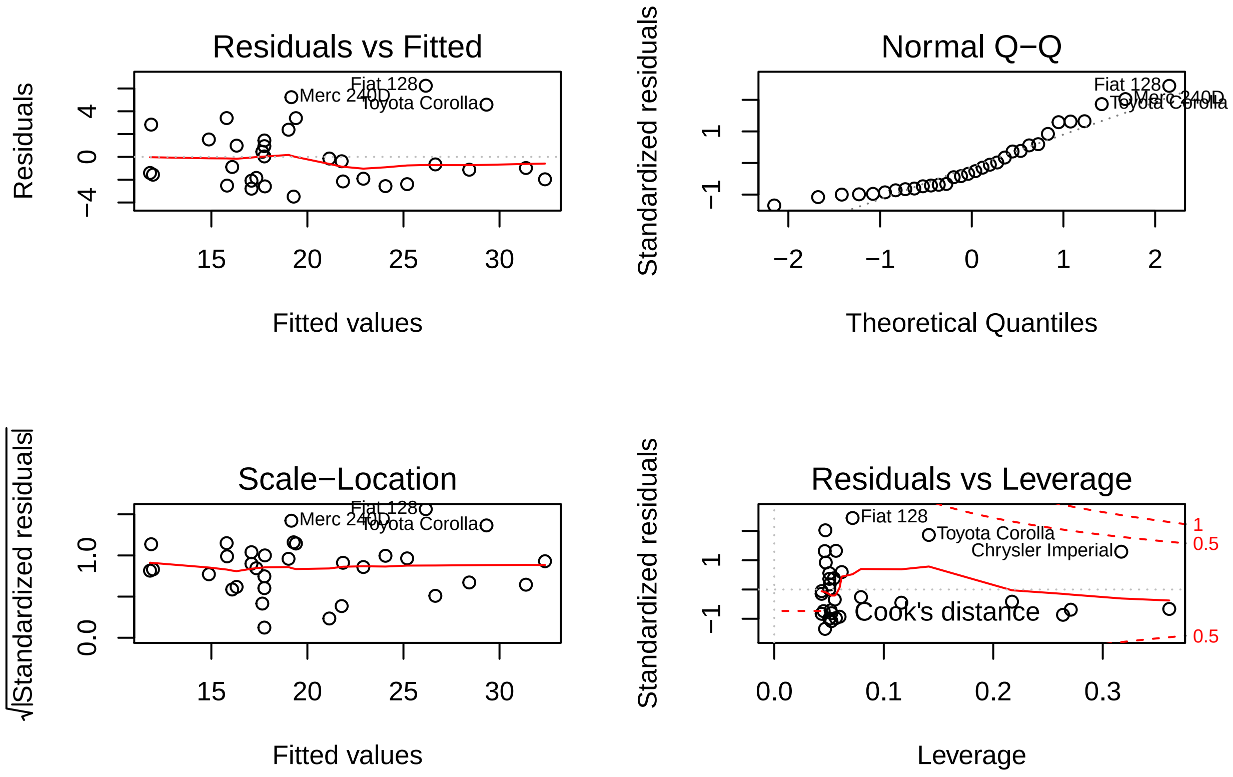


Formula: mpg ~ poly(wt, 2)

```r
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ poly(wt, 2), data = mtcars.processed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.483 -1.998 -0.773  1.462  6.238
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.0906     0.4686  42.877  < 2e-16 ***
## poly(wt, 2)1 -29.1157     2.6506 -10.985 7.52e-12 ***
## poly(wt, 2)2   8.6358     2.6506   3.258  0.00286 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.651 on 29 degrees of freedom
## Multiple R-squared:  0.8191, Adjusted R-squared:  0.8066
## F-statistic: 65.64 on 2 and 29 DF,  p-value: 1.715e-11
```

```
par(mfrow = c(2,2))
plot(model3)
```



Residuals vs Fitted
Residuals
Fitted values
−4 0 4
15 20 25 30
Fiat 128
Merc 240D
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1 1
−2 −1 0 1 2
Fiat 128
Merc 240D
Toyota Corolla
Scale−Location
√|Standardized residuals|
Fitted values
0.0 1.0
15 20 25 30
Fiat 128
Merc 240D
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
−1 1
0.0 0.1 0.2 0.3
Fiat 128
Toyota Corolla
Chrysler Imperial
Cook's distance
1
0.5
0.5