

Coursera: Regression Models Course Project

Craig Rowley

December 19, 2014

Executive Summary

I work for Motor Trend (MT), a magazine about the automobile industry. Looking at a data set of a collection of cars, MT is interested in exploring the relationship between vehicle attributes and miles per gallon (MPG). Two specific areas (below) are summarized for you, followed by a detailed explanation of our conclusions.

Is an automatic or manual transmission better for MPG?

Our analysis shows that a manual transmission correlates with increased fuel efficiency. Transmission alone is not the best way to increase fuel efficiency. Here are some other possibilities discovered in the data.

- 1) Combining an Inline-engine and fewer Carburetors with a Manual transmission may also help increase MPG. On average, these cars may get up to 29 MPG, with typical manual transmissions. Toyota Corolla, Fiat, and Datsun do well here.
- 2) The number of cylinders in a car also does a good job of indicating fuel efficiency. A 4-cylinder car can commonly have up to 10 mpg increase over an 8-cylinder vehicle. Average MPG ranges between 15 and 26 mpg for highest and lowest cylinder counts, respectively.

Quantify the MPG difference between automatic and manual transmissions.

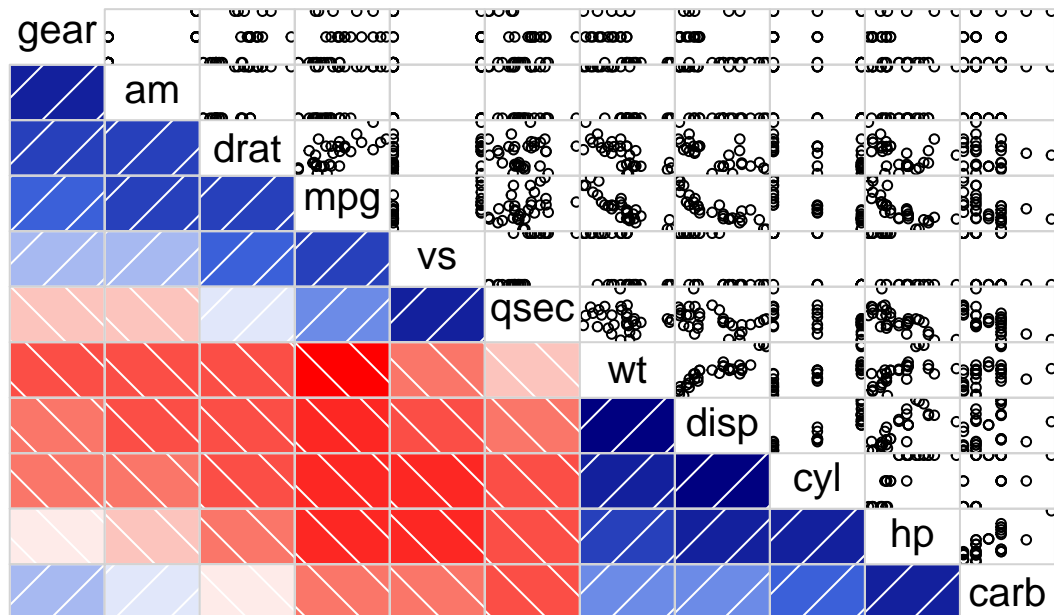
Using a model that considers only engine type (V or Inline) and the number of Carburetors, we can quantify the mpg increase attributable to a manual transmission. If we train the same model on two mutually exclusive data sets (automatic versus manual transmission), we see that a *manual transmission gets more than 7 mpg better fuel efficiency*. Switching to an Inline engine has a more positive effect on automatics (+4.34 mpg), but that is insufficient to overtake the manual transmission. In fact, our model suggests that a manual transmission could opt for more power by switching to a V-engine and dual Carbs and still likely outperform the fuel efficiency of the average automatic (if they weren't racing or hauling loads).

```
at <- mtcars[mtcars$am==0,c("mpg","vs","carb")]; lm_at <- lm(mpg ~ ., at)
mt <- mtcars[mtcars$am==1,c("mpg","vs","carb")]; lm_mt <- lm(mpg ~ ., mt)
coef(lm_at); coef(lm_mt)
```

```
## (Intercept)      vs      carb
##   19.461710    4.347200   -1.430825
```

```
## (Intercept)      vs      carb
##   26.738430    3.772314   -1.497521
```

Model Design



Thoughts on Correlation:

- MPG correlates with Manual transmission (am), Rear-axle ratio (drat), and an Inline engine (vs)
- MPG doesn't correlate with Vehicle weight (wt), Engine Displacement (disp), Cylinders (cyl), and Horsepower (hp)
- Weight may have collinearity with issues with Gear count, Transmission type, and Rear-axle ratio

```
ks<-lm(mpg~.,mtcars); sort(coef(ks),decreasing=TRUE); sqrt(vif(ks))
```

```
## (Intercept)      am      qsec      drat      gear      vs
## 12.30337416  2.52022689  0.82104075  0.78711097  0.65541302  0.31776281
##      disp      hp      cyl      carb      wt
##  0.01333524 -0.02148212 -0.11144048 -0.19941925 -3.71530393

##      cyl      disp      hp      drat      wt      qsec      vs      am
##  3.920948  4.649757  3.135608  1.837014  3.894212  2.743712  2.228424  2.156035
##      gear      carb
##  2.314617  2.812249
```

Thoughts on Kitchen Sink model Coefficients (above):

- Increased Weight, Carburetor count, Cylinder count, and Horsepower result in less MPG
- Increased Displacement, Straight engine, gears, Rear axle ratio, and Manual tranny result in higher MPG
- Quarter-mile time does “not” match intuition as faster acceleration should have negative impact on MPG

Thoughts on Colinearity (above) :

- Variance inflation factors all reasonably high (too many variables)
- Weight is correlated, but we find that it has Heteroskedascity issues with mpg (see Appendix)

Appendix: Exploratory Data Analysis (output grossly reduced to meet 5 page requirement)

Omitted Variables (due to 5-page limit) :: disp, drat, hp, qsec, gear

Critical Variables

mpg :: Miles/(US) gallon

Our Dependent Variable.

cyl :: Number of cylinders

Larger, more powerful vehicles typically have more cylinders. More cylinders typically means more fuel-per-second consumed in modern cars (and more horsepower), but have a negative impact on mpg.

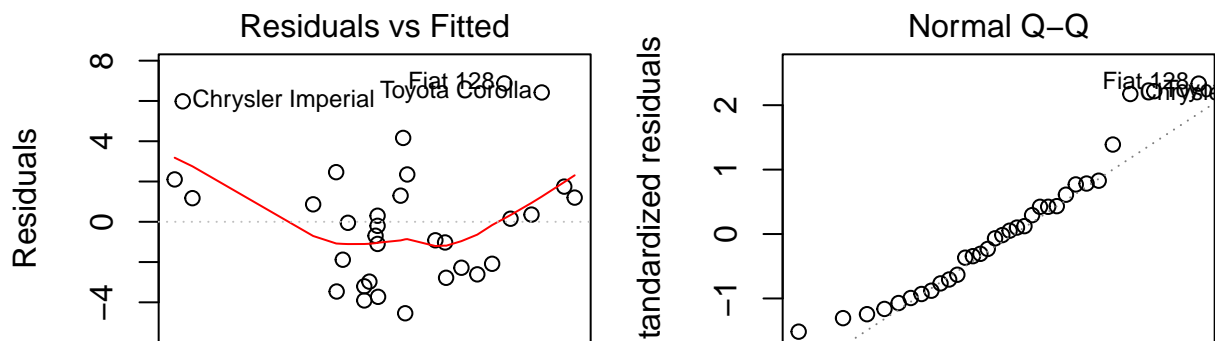
```
#...Model and output interesting EDA
fit <- lm(mpg ~ cyl, mtcars)
summary(fit)$r.squared
```

```
## [1] 0.72618
```

wt :: Weight (lb/1000)

Weight of vehicle expressed in units of 1k lbs. Note that Weight's residual plot shows non-random U-shape (i.e. Heteroskedasticity)

```
fit <- lm(mpg ~ wt, mtcars)
par(mfrow=c(1,2),mar=c(0,5,15,0))
plot(fit,which=c(1,2))
```



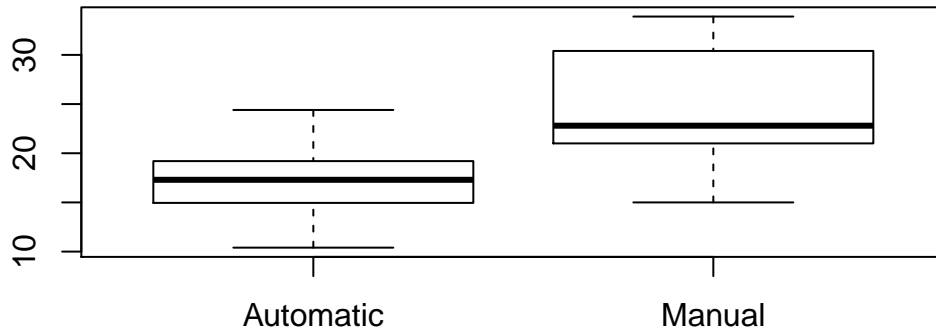
am :: Transmission (0 = automatic, 1 = manual)

Manual transmission should help mpg (based on Consumer Reports research)

```
fit <- lm(mpg ~ am, mtcars)
summary(fit)$r.squared
```

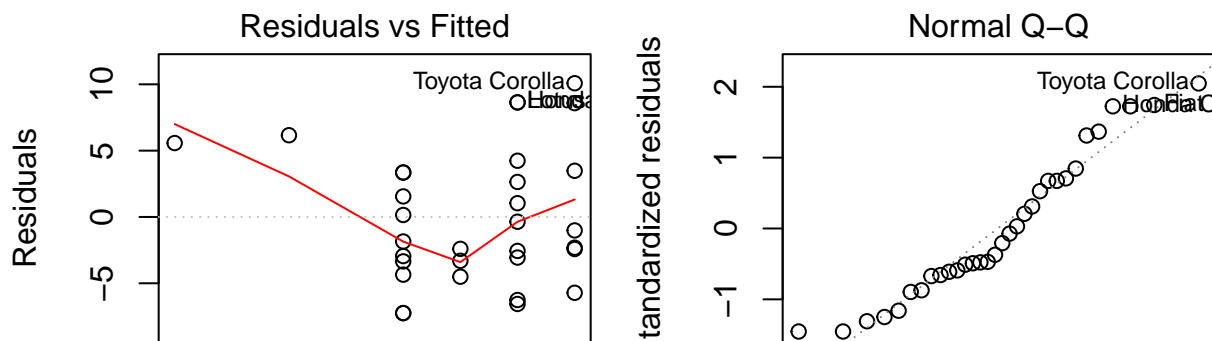
```
## [1] 0.3597989
```

```
par(mfrow=c(1,1),mar=c(8,5,8,5))  
boxplot(mtcars$mpg~mtcars$am,names=c("Automatic","Manual"))
```



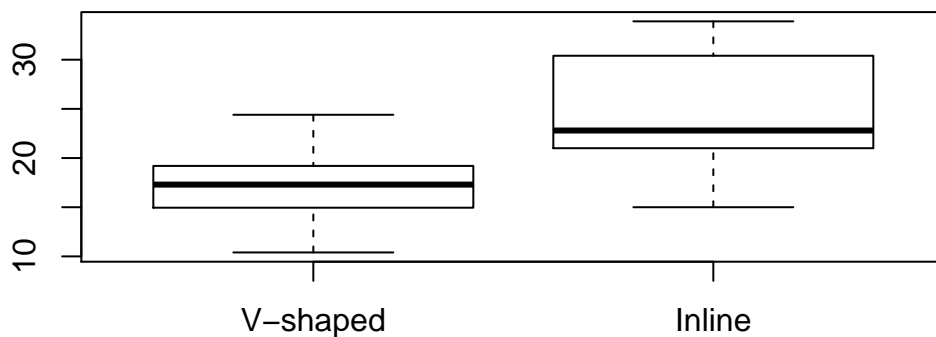
carb :: Number of carburetors

Number of carbs decreases fuel efficiency



vs :: V-shaped or Straight (Inline) engine shape

Inline engine increases fuel efficiency



Modeling

Remove Heteroskedastic, Colinear, and Uncorrelated variables

- Coefficients match intuition, but P-values for T-stats are much too large to give confidence
- Variance Inflation Factors are also reasonably large

```
mtcars_1 <- mtcars[,c("mpg", "drat", "vs", "am", "gear", "carb")]
model_1 <- lm(mpg ~ ., mtcars_1)
sqrt(vif(model_1))
```

```
##      drat      vs      am      gear      carb
## 1.685073 1.484628 1.887650 2.076755 1.495089
```

(Selected Model) Remove Gear and Drat due to correlation

- Coefficients match research and individual trends
- p-values and variable inflation factors are appropriately small
- No significant colinearity issues
- No Heteroskedasticity issues

```
mtcars_3 <- mtcars[,c("mpg", "vs", "am", "carb")]
model_3 <- lm(mpg ~ ., mtcars_3)
s <- summary(model_3)
s$coefficients
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 19.517399   1.6090815 12.129528 1.155904e-12
## vs          4.195736   1.3245867  3.167581 3.695735e-03
## am          6.797956   1.1014890  6.171606 1.154742e-06
## carb       -1.430783   0.4081085 -3.505890 1.552505e-03
```

```
s$r.squared
```

```
## [1] 0.7818462
```

```
sqrt(vif(model_3))
```

```
##      vs      am      carb
## 1.254946 1.033173 1.239088
```

Remove VS since it correlates with am and carb

Thoughts :: R^2 took a dive, so removing VS was not necessary

```
mtcars_5 <- mtcars[,c("mpg", "am", "carb")]
model_5 <- lm(mpg ~ ., mtcars_5)
summary(model_5)$r.squared
```

```
## [1] 0.7036726
```