MPGTransmissionStudy.Rmd

Jim Callahan

October 7-19, 2015

Executive Summary

This project is intended to answer the following two questions:

1. "Is an automatic or manual transmission better for MPG?"

2. "Quantify the MPG difference between automatic and manual transmissions?"

using statistical regression analysis in **R** on the **"Motor Trend"**, **"mtcars"** data set included with the **R** system. This study does not show transmission type (automatic vs. manual) to be significant once one accounts for weight and number of engine cylinders. This result, however, may represent a flaw in the study design; this regression study, in effect uses group averages and not paired data. Paired data would be closer to the consumer experience of evaluating one model of car with (standard) manual or (optional) automatic transmission. There may be a more distinct effect when one examines one model of car at a time with manual or automatic transmission rather than pooling several models of cars together in one data set.

Data Vintage

The source of the **"mtcars"** data set (as described in the documentation **help(mtcars)**) is Henderson and Velleman (1981), **Building multiple regression models interactively.** Biometrics, 37, 391–411. http://www.mortality.org/INdb/2008/02/12/8/document.pdf

The **help(mtcars)** documentation states:

> "The data was extracted from the **1974 Motor Trend** US magazine,
> and comprises fuel consumption and 10 aspects of automobile design and performance
> for 32 automobiles (**1973–74 models**)."

So, it should be noted the **"mtcars"** data set is vintage **mid-1970s** and is therefore unlikely to be representative of the contemporary state of the automotive art.

Exploratory Data Analysis

According to the **help(mtcars)** documentation, **"mtcars"** is

> "A **data frame** with 32 observations on 11 variables.
>
> [, 1] **mpg** Miles/(US) gallon
> [, 2] **cyl** Number of cylinders
> [, 3] **disp** Displacement (cu.in.)
> [, 4] **hp** Gross horsepower
> [, 5] **drat** Rear axle ratio
> [, 6] **wt** Weight (lb/1000)
> [, 7] **qsec** 1/4 mile time
> [, 8] **vs** V/S
> [, 9] **am** Transmission (0 = automatic, 1 = manual)
> [,10] **gear** Number of forward gears
> [,11] **carb** Number of carburetors"

The documentation is confirmed using the **str()** (structure) function in **R**:

```
data(mtcars)
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 ...
```

Preliminary Analysis

On the surface the minimum requirements of this project are trivially simple:
1. Convert the zero-one transmission indicator variable, **"am"** to an **R "factor"**.
2. Run a regression with mpg = f(am) or in **R** notation **lm(mpg ~ am))**
I have supressed the intercept ("0 +"), so the coefficients can be read off directly without having to calculate the automatic transmision as a base plus an offset.

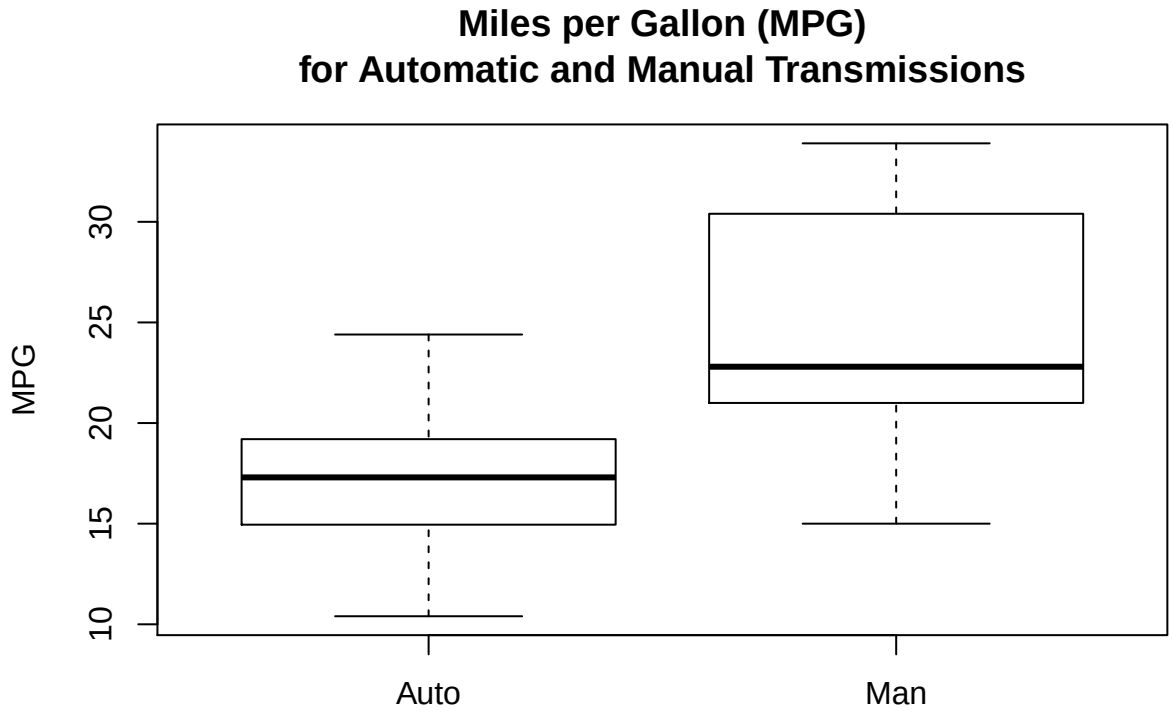
```
# MPG Model zero "000" -- our "quick and dirty" literal regression
mtcars$am <- factor(mtcars$am,levels=c(0,1), labels=c("Auto","Man"))
MPGmod000 <- lm(mpg ~ 0 + as.factor(am), data=mtcars)
MPGmod000
```

```
##
## Call:
## lm(formula = mpg ~ 0 + as.factor(am), data = mtcars)
##
## Coefficients:
## as.factor(am)Auto   as.factor(am)Man
##             17.15              24.39
```

So, the "quick and dirty" interpretation our base model zero, would be that
the average 1975 vintage car with automatic transmission gets 17 miles per gallon while the average 1975 vintage car with manual transmisson gets an additional 7 miles per gallon for a total of 24 miles per gallon.

We can picture this with a box plot.

```
plot(as.factor(mtcars$am), mtcars$mpg,
     main = "Miles per Gallon (MPG)\nfor Automatic and Manual Transmissions",
          ylab = "MPG")
abline(mtcars$mpg ~ as.factor(mtcars$am))
```



Clearly, as indicated by the dark horizontal line, the mean mpg of the manual transmission cars is higher than the mean mpg of the automatic transmission cars. But, the "whiskers" of the "box and whiskers" plot (the interquartile range) shows that the two ranges overlap; in other words, some cars with manual transmissions have mpgs as low or lower than some cars with automatic transmissions. If manual transmission cars always had higher mpg, there would be no overlap of the interquartile ranges.

Of course to accept this analysis at face value, one would have to invoke the economist's assumption of "***ceteris paribus***" (all other things being equal).

Of course we know all other things are **NOT EQUAL**. There are **confounding variables**. For instance, the cars vary in weight, number of cylinders in their engines and the size of their engines measured in cubic inch displacement.

One, **low tech** way of seeing what is going on is simply to **sort the data set by mpg** and look at the data.

```
mtcars[order(-mtcars$mpg), ]
```

```
##                      mpg cyl  disp  hp drat    wt  qsec vs   am gear carb
## Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  Man    4    1
## Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  Man    4    1
## Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  Man    4    2
## Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  Man    5    2
## Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  Man    4    1
## Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  Man    5    2
## Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1 Auto    4    2
## Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  Man    4    1
## Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1 Auto    4    2
```

```
## Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1 Auto    3    1
## Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1 Auto    3    1
## Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  Man    4    2
## Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  Man    4    4
## Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  Man    4    4
## Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  Man    5    6
## Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1 Auto    4    4
## Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0 Auto    3    2
## Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0 Auto    3    2
## Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1 Auto    3    1
## Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1 Auto    4    4
## Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0 Auto    3    3
## Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0 Auto    3    3
## Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  Man    5    4
## Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0 Auto    3    2
## Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0 Auto    3    3
## AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0 Auto    3    2
## Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  Man    5    8
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0 Auto    3    4
## Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0 Auto    3    4
## Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0 Auto    3    4
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0 Auto    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0 Auto    3    4
```

The **top 5 high mileage cars** tend to have **smaller engines (as measured by cylinders (cyl) displacemnet (disp) and horsepower (hp))** and **weigh less than 2,200 pounds.** The **high milage cars** also tend to be **slower** (as measured by their quarter mile times (**qsec**)), have **manual transmissions (am = 1 or "Man")** with more gears (**gear**) and fewer carburetors (**carb**).

```
head(mtcars[order(-mtcars$mpg), ], 5)
```

```
##                 mpg cyl disp  hp drat    wt  qsec vs  am gear carb
## Toyota Corolla 33.9   4 71.1  65 4.22 1.835 19.90  1 Man    4    1
## Fiat 128       32.4   4 78.7  66 4.08 2.200 19.47  1 Man    4    1
## Honda Civic    30.4   4 75.7  52 4.93 1.615 18.52  1 Man    4    2
## Lotus Europa   30.4   4 95.1 113 3.77 1.513 16.90  1 Man    5    2
## Fiat X1-9      27.3   4 79.0  66 4.08 1.935 18.90  1 Man    4    1
```

While the **bottom 5 low mileage cars** tend to have **bigger engines (as measured by cylinders (cyl) displacemnet (disp) and horsepower (hp))** and **weigh more than 3,500 pounds.** The **low milage cars** also tend to be **faster** (as measured by their quarter mile times (**qsec**)), have **automatic transmissions (am = 0 or "Auto")** with fewer gears (**gear**) and more carburetors (**carb**).

```
tail(mtcars[order(-mtcars$mpg), ], 5)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs   am gear carb
## Chrysler Imperial   14.7   8  440 230 3.23 5.345 17.42  0 Auto    3    4
## Duster 360          14.3   8  360 245 3.21 3.570 15.84  0 Auto    3    4
## Camaro Z28          13.3   8  350 245 3.73 3.840 15.41  0 Auto    3    4
## Cadillac Fleetwood  10.4   8  472 205 2.93 5.250 17.98  0 Auto    3    4
## Lincoln Continental 10.4   8  460 215 3.00 5.424 17.82  0 Auto    3    4
```

```r
attach(mtcars)
# wt = Weight (lb/1000)
summary( wt )
```

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

```r
# cyl = Number of Cylinders
# Count of cars by Number of Cylinders
summary( as.factor(cyl) )
```

```
##  4  6  8
## 11  7 14
```

```r
# mpg by Number of Cylinders
tapply(mpg, as.factor(cyl), mean )
```

```
##        4        6        8
## 26.66364 19.74286 15.10000
```

```r
# disp = Displacement (cu.in.)
summary(disp )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.1   120.8   196.3   230.7   326.0   472.0
```

```r
detach(mtcars)
```

Historical Context

Moreover, keep in mind the historical context, this is the late 1970s/early 1980s when there was still a huge difference between American, European and Japanese auto technology.

> "In the mid-1980s, Toyota took over the Fremont [,California] plant, one of GM's worst, a factory known for sex, drugs and defective vehicles. And as part of an historic joint venture [NUMMI], Toyota turned the plant into one of GM's best, practically overnight.
> Along the way — remarkably — Toyota even shared its production secrets. . . . In 1985, after NUMMI opened, Car and Driver magazine ran the following
> headline: **'Hell Freezes Over.'**"
> **"The End Of The Line For GM-Toyota Joint Venture"** by Frank Langfitt,
> National Public Radio (NPR), MARCH 26, 2010 3:00 PM ET
> http://www.npr.org/templates/story/story.php?storyId=125229157

GM's Saturn was not introduced until the 1991 model year, ten years after the 1981 vintage of the "mtcars" data set.
https://en.wikipedia.org/wiki/Saturn_Corporation

Electric vehicle hybrids, such as Toyota's Prius NHW11, were not introduced to the US market until the 2001 model year.
https://en.wikipedia.org/wiki/Toyota_Prius

A Web search for ggplot2 facet examples found a QuickR blog post, "Graphics with ggplot2" by Robert I. Kabacoff, PhD. http://www.statmethods.net/advgraphs/ggplot2.html

```r
# Separate regressions of mpg on weight for each number of cylinders
# create factors with value labels
library(ggplot2)
data(mtcars)
mtcars$gear <- factor(mtcars$gear,levels=c(3,4,5),
                      labels=c("3gears","4gears","5gears"))
mtcars$am <- factor(mtcars$am,levels=c(0,1),
                    labels=c("Automatic","Manual"))
mtcars$cyl <- factor(mtcars$cyl,levels=c(4,6,8),
                     labels=c("4 cylinder","6 cylinder","8 cylinder"))

# All -- use everything
MPGModAll <- lm(mpg  ~ . , data = mtcars);
summary(MPGModAll)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2015 -1.2319  0.1033  1.1953  4.3085
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   15.09262   17.13627   0.881   0.3895
## cyl6 cylinder -1.19940    2.38736  -0.502   0.6212
## cyl8 cylinder  3.05492    4.82987   0.633   0.5346
## disp           0.01257    0.01774   0.708   0.4873
## hp            -0.05712    0.03175  -1.799   0.0879 .
## drat           0.73577    1.98461   0.371   0.7149
## wt            -3.54512    1.90895  -1.857   0.0789 .
## qsec           0.76801    0.75222   1.021   0.3201
## vs             2.48849    2.54015   0.980   0.3396
## amManual       3.34736    2.28948   1.462   0.1601
## gear4gears    -0.99922    2.94658  -0.339   0.7382
## gear5gears     1.06455    3.02730   0.352   0.7290
## carb           0.78703    1.03599   0.760   0.4568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.616 on 19 degrees of freedom
## Multiple R-squared:  0.8845, Adjusted R-squared:  0.8116
## F-statistic: 12.13 on 12 and 19 DF,  p-value: 1.764e-06
```

```r
# Weignt is significant
MPGmod001 <- lm(mpg ~ as.factor(am)+wt, data=mtcars)
summary(MPGmod001)
```

```
##
## Call:
## lm(formula = mpg ~ as.factor(am) + wt, data = mtcars)
##
```

```
## 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 ***
## as.factor(am)Manual -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
```

```
# Cylinders helps
MPGmod002 <- lm(mpg ~ as.factor(am)+wt+as.factor(cyl), data=mtcars)
summary(MPGmod002)
```

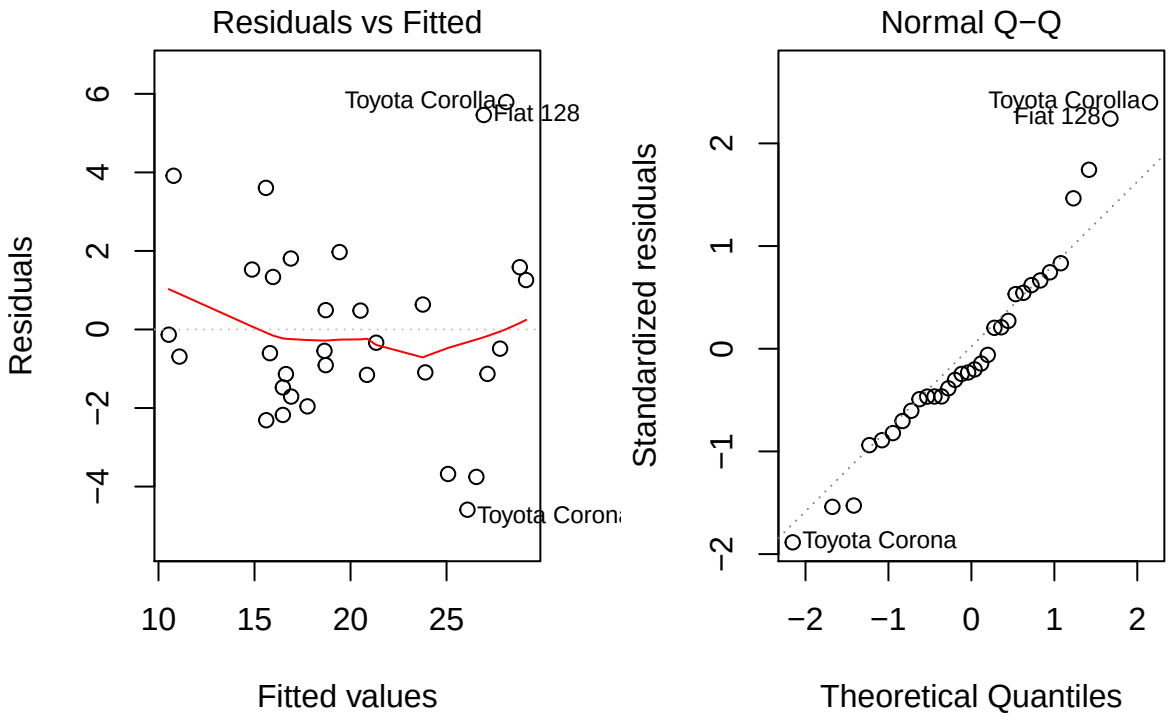
```
##
## Call:
## lm(formula = mpg ~ as.factor(am) + wt + as.factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4898 -1.3116 -0.5039  1.4162  5.7758
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                33.7536     2.8135  11.997  2.5e-12 ***
## as.factor(am)Manual         0.1501     1.3002   0.115  0.90895
## wt                         -3.1496     0.9080  -3.469  0.00177 **
## as.factor(cyl)6 cylinder   -4.2573     1.4112  -3.017  0.00551 **
## as.factor(cyl)8 cylinder   -6.0791     1.6837  -3.611  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.603 on 27 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8134
## F-statistic: 34.79 on 4 and 27 DF,  p-value: 2.73e-10
```

```
# drop Transmission (am) *** BEST MODEL ***
par(mfrow = c(1,2))
MPGmod003 <- lm(mpg ~ wt+as.factor(cyl), data=mtcars)
summary(MPGmod003)
```

```
##
## Call:
## lm(formula = mpg ~ wt + as.factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                33.9908     1.8878  18.006  < 2e-16 ***
## wt                         -3.2056     0.7539  -4.252 0.000213 ***
## as.factor(cyl)6 cylinder   -4.2556     1.3861  -3.070 0.004718 **
## as.factor(cyl)8 cylinder   -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

```
plot(MPGmod003, which = 1)
plot(MPGmod003, which = 2)
```




```
MPGmod005 <- lm(mpg ~ log(wt), data=mtcars)
summary(MPGmod005)
```

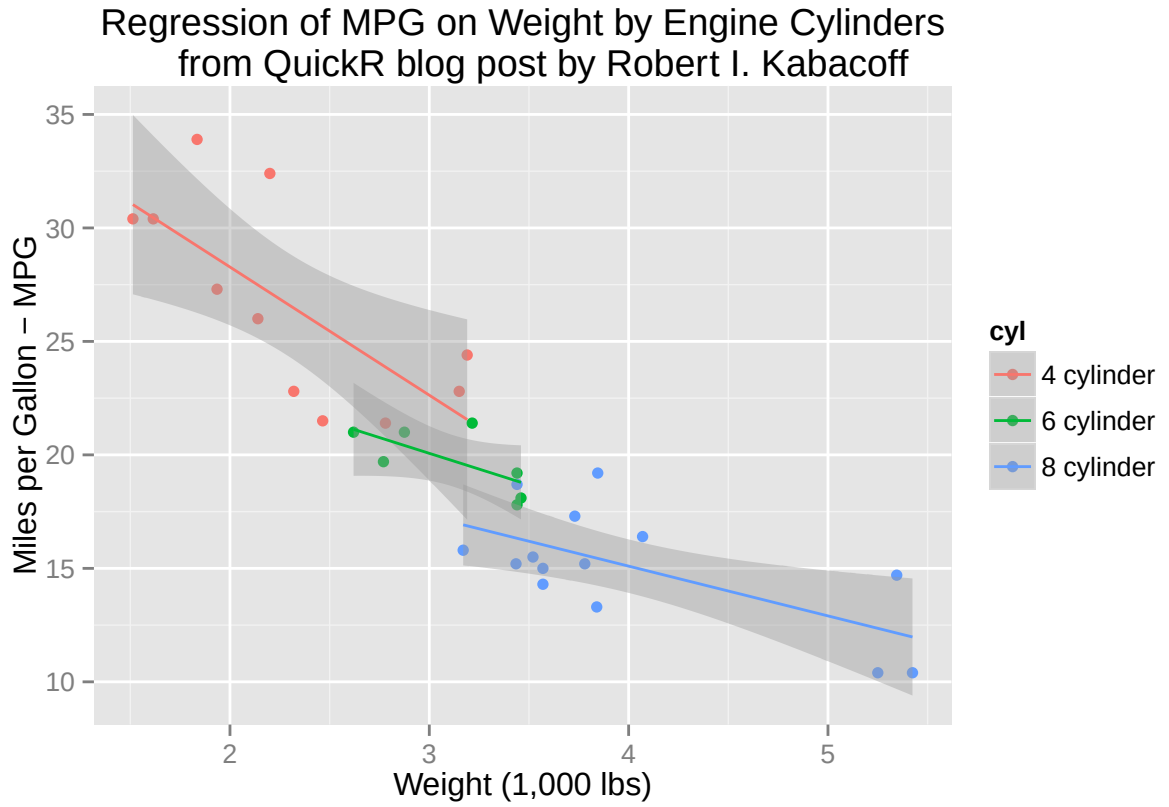
```
##
## Call:
## lm(formula = mpg ~ log(wt), data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7440 -2.0954 -0.3672  1.0709  6.6150
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.257      1.758   22.32  < 2e-16 ***
## log(wt)      -17.086      1.510  -11.31 2.39e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.669 on 30 degrees of freedom
## Multiple R-squared:  0.8101, Adjusted R-squared:  0.8038
## F-statistic:   128 on 1 and 30 DF,  p-value: 2.391e-12
```

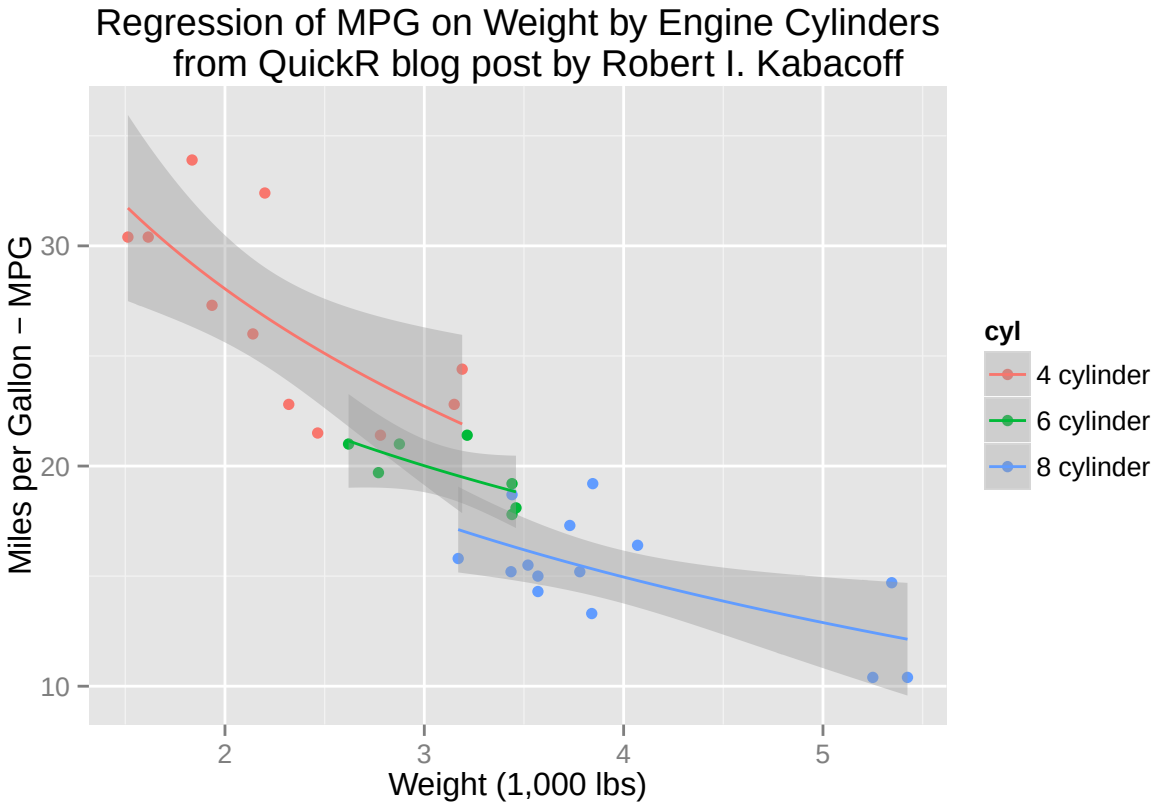
```
MPGmod005 <- lm(mpg ~ log(wt)+as.factor(cyl), data=mtcars)
summary(MPGmod005)
```

```
##
## Call:
## lm(formula = mpg ~ log(wt) + as.factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9830 -1.3486 -0.6479  1.6017  5.6220
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 35.755      1.941  18.418  < 2e-16 ***
## log(wt)                    -11.386      2.260  -5.039  2.5e-05 ***
## as.factor(cyl)6 cylinder    -3.133      1.373  -2.283  0.03025 *
## as.factor(cyl)8 cylinder    -5.045      1.609  -3.135  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.375 on 28 degrees of freedom
## Multiple R-squared:  0.8597, Adjusted R-squared:  0.8446
## F-statistic: 57.18 on 3 and 28 DF,  p-value: 4.633e-12
```

```
qplot(wt, mpg, data=mtcars, geom=c("point", "smooth"),
      method="lm", formula=y~x, color=cyl,
      main="Regression of MPG on Weight by Engine Cylinders
    from QuickR blog post by Robert I. Kabacoff",
      xlab="Weight (1,000 lbs)",
      ylab="Miles per Gallon - MPG")
```




```
qplot(wt, mpg, data=mtcars, geom=c("point", "smooth"),
      method="lm", formula=y~log(x), color=cyl,
      main="Regression of MPG on Weight by Engine Cylinders
    from QuickR blog post by Robert I. Kabacoff",
      xlab="Weight (1,000 lbs)",
      ylab="Miles per Gallon - MPG")
```


Regression of MPG on Weight by Engine Cylinders
from QuickR blog post by Robert I. Kabacoff
Miles per Gallon – MPG
30
20
10
2
3
4
5
Weight (1,000 lbs)
cyl
4 cylinder
6 cylinder
8 cylinder