Motortrend Study

Harland Hendricks

August 7, 2018

Executive Summary

Motor Trend magazine is interested in exploring the relationship between a set of variables from a data set of a collection of cars and miles per gallon (MPG).

Specifically, the magazine would like to focus on the following two questions:

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

When testing the predictors **wt**, **cyl**, and **am** using the nested model method, I found that **am** is not significant and should not be used to model affects on MPG. A simple plot of MPG per type of transmission is all that is required to determine that a manual transmission can result in higher gas mileage for the data set used. Additionally, simply subtracting the mean of automatic transmission MPG from manual transmission MPG quantifies the difference in MPG.

An appropriate model for MPG for the given data is one that includes the predictors **wt** and **cyl**. These predictors appear to influence MPG more than **am**. This model suggests that for every 1000 lbs increase in the automobile weight, MPG decreases by 3.2 MPG. Additionally, MPG will decrease by 4.3 when increasing cylinders from 4 to 6, and will decrease by 6.1 when increasing cylynders from 4 to 8.

mtcars Dataset

The instructions for this assignment are located here.

Built with R version 3.5.0 with the following system:

```
##           sysname           release           version          nodename 
##         "Windows"          "10 x64"     "build 17134" "DESKTOP-TPCQ5AJ" 
##           machine             login              user    effective_user 
##          "x86-64"           "harla"           "harla"           "harla"
```

Load the required libraries

```
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Explore mtcars dataset

```r
head(mtcars, 3)
```

```
##                 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
```

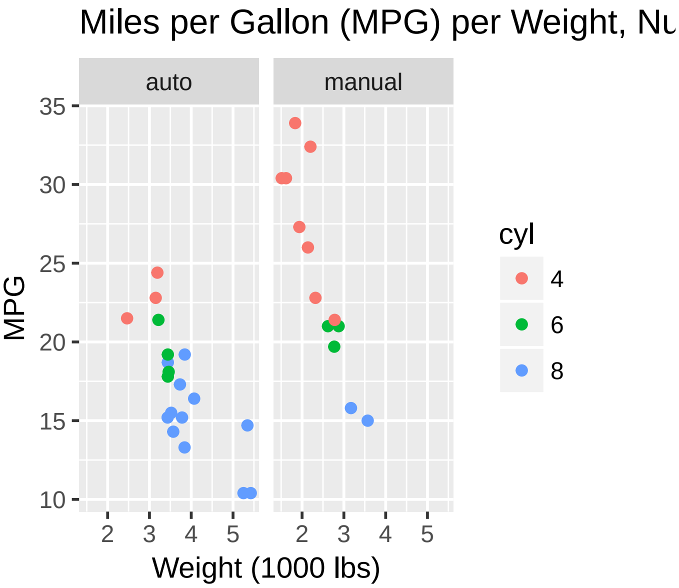
```r
str(mtcars)
```

```
## '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 ...
```

Create *data* object with **mpg**, **cyl**, **wt**, and **am** variables. Reclassify **cyl** and **am** from numeric to factor and add column **trans** to indicate transmission type.

```r
data <- mtcars[, c(1, 2, 6, 9)]
data$cyl <- as.factor(data$cyl)
data$am <- as.factor(data$am)
data <- mutate(data, trans = ifelse(am == 1, "manual", "auto"))
```

Plot **trans** vs **mpg**



The initial plot of MPG given a weight, number of cylinders, and specific transmission type indicates that as weight increases - mpg decreases. Additionally, as the number of cylinders increases - mpg decreases. Finally, a manual transmission provides greater miles per gallon.

Model Selection and Fit

I chose to use the predictors **cyl**, **wt**, and **am** due to the belief that these variables would have the most affect on **mpg**. *Motor Trend* is also specifically interested in **am**. I will build three models and then nest the models to determine which model contains necessary predictors.

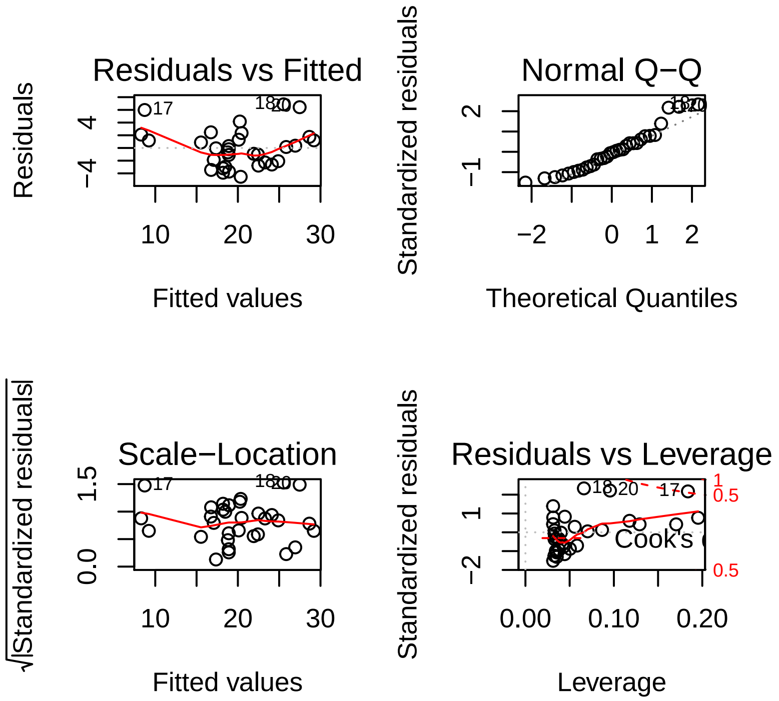
Model 1

```
m1 <- lm(mpg ~ wt, data = data)
summary(m1)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

We can then plot the residuals for diagnostics:

```
par(mfrow = c(2, 2))
plot(m1)
```



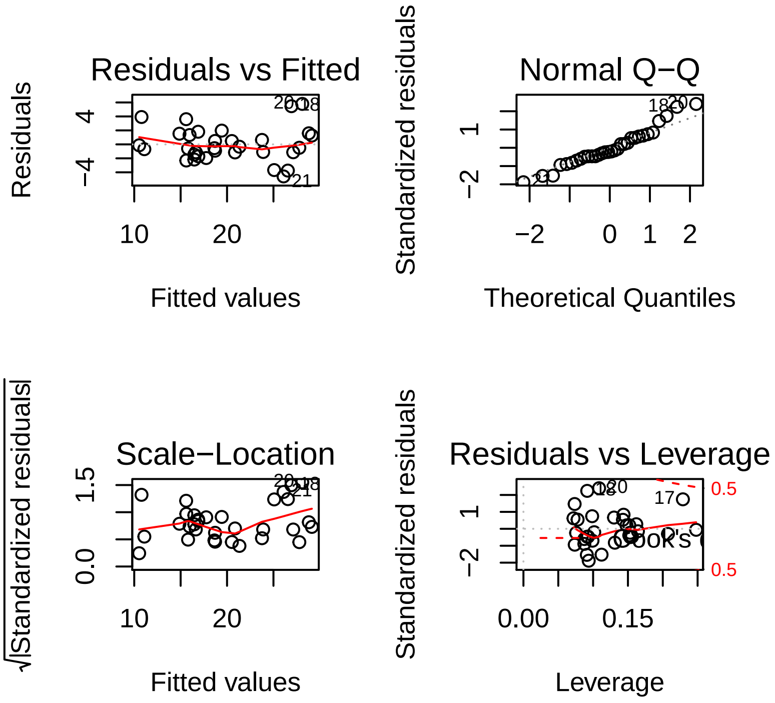
The top left and top right residual plots indicate that there may be a problem modeling with just **wt**

Model 2

```
## lm(formula = mpg ~ wt + cyl, data = data)
```

```
## (Intercept)          wt        cyl6        cyl8 
##   33.990794   -3.205613   -4.255582   -6.070860
```

We can then plot the residuals for diagnostics:

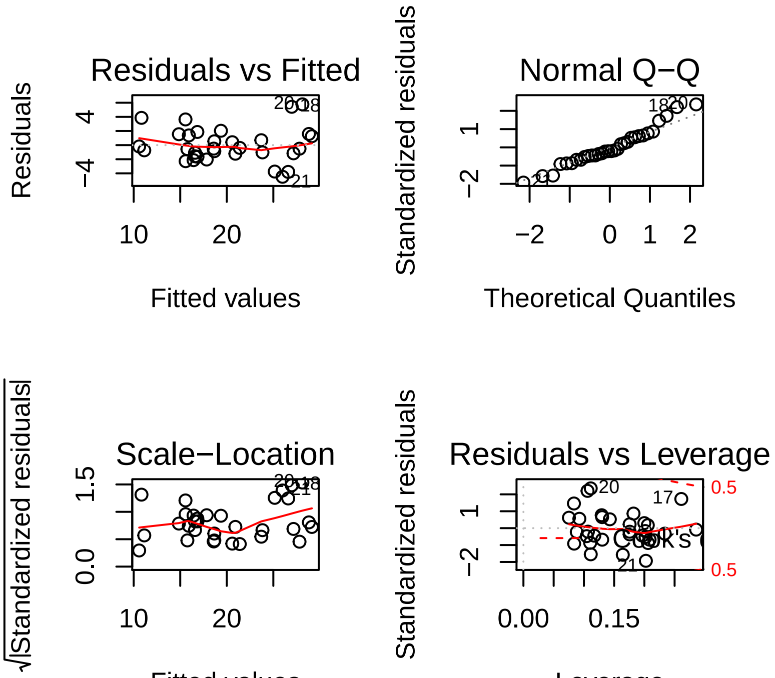


This model looks better, but the top right residual plot still indicates that the residuals may not be normally distributed and there may be a problem modeling with **wt** and **cyl**.

Model 3

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + am, data = data)
##
## 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 ***
## wt           -3.1496     0.9080  -3.469  0.00177 **
## cyl6         -4.2573     1.4112  -3.017  0.00551 **
## cyl8         -6.0791     1.6837  -3.611  0.00123 **
## am1           0.1501     1.3002   0.115  0.90895
## ---
## 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
```

We can then plot the residuals for diagnostics:



The residual plots for model 3 don't change much from 2, but the p values for **am** are not significant. There may be a problem modeling with **wt**, **cyl**, and **am**.

Nested Model Testing

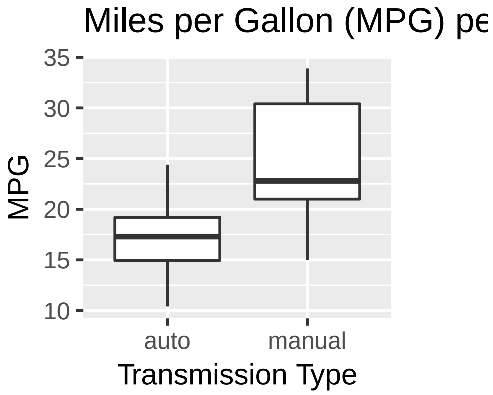
```r
anova(m1, m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + am
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     28 183.06  2    95.263 7.0288 0.003488 **
## 3     27 182.97  1     0.090 0.0133 0.908947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the nested model test, I determine that the model with predictors **wt** and **cyl** is significant with acceptable residual plots.

The predictor **am** is not significant and should not be used to model **mpg** in *Motor Trend's* study. The linear model with **am** as a predictor just predicts the mean of **mpg** when grouped by manual or automatic transmission and is not a very useful model

The mean MPG of each **am** show that a manual transmission is better for **mpg** and that the difference between manual and automatic translates to a difference in MPG:


Miles per Gallon (MPG) pe
35
30
25
20
15
10
MPG
auto
manual
Transmission Type