Transmission Versus MPG Regression

Brian Stroh

October 11, 2018

Overview

This report explores fitting mpg from the mtcars dataset to a linear model.

Executive Summary

Upon testing various models against each other, it has been decided that the model regressing am (transmission type), wt (car weight) and qsec (quarter mile time) along with an intercept creates the most appropriate model for mpg. A nice balance is struck between the regressors p-Value coefficients and the variance inflation factors, and the anova test has confirmed that adding each of these three regressors results in a significant improvement to the model's fit. Upon fitting this model, we can conclude that the manual transmissions contribute an average increase of 2.9 mpg compared to vehicles with autmatic transmissions.

Setup

```
library(dplyr)
library(ggplot2)
library(grid)
library(gridExtra)
library(car)
data(mtcars)
```

Exploratory Analysis

```
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 ...
```

```
summary(filter(mtcars, am == 1)$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

```
summary(filter(mtcars, am == 0)$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
t.test(filter(mtcars, am == 1)$mpg, filter(mtcars, am == 0)$mpg, paired = FALSE,
       alternative = "greater")$p.val
```

```
## [1] 0.0006868192
```

If the transmission type were the sole predictor, this would allow us to conclude that manual transmissions have a greater average mpg.

Model Selection

We can not use logistic regression because that is concerned with binary **outcomes**, not **inputs**. Now we'll build various models and examine models of interest

```
fit1<-lm(mpg~am-1,mtcars)
fit2<-update(fit1,mpg~am+wt-1)
fit3<-update(fit2,mpg~am+wt+qsec-1)
fit4<-update(fit3,mpg~am+wt+qsec+hp-1)

fit1Int<-lm(mpg~am,mtcars)
fit2Int<-update(fit1Int,mpg~am+wt)
fit3Int<-update(fit2Int,mpg~am+wt+qsec)
fit4Int<-update(fit3Int,mpg~am+wt+qsec+hp)

summary(fit1Int)$coef
```

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

```
summary(fit2Int)$coef
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) 37.32155131  3.0546385 12.21799285 5.843477e-13
## am          -0.02361522  1.5456453 -0.01527855 9.879146e-01
## wt          -5.35281145  0.7882438 -6.79080719 1.867415e-07
```

```
summary(fit3Int)$coef
```

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

```
summary(fit4Int)$coef
```

```
##                Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 17.44019110  9.3188688  1.871492 0.072149342
## am           2.92550394  1.3971471  2.093913 0.045790788
## wt          -3.23809682  0.8898986 -3.638726 0.001141407
## qsec         0.81060254  0.4388703  1.847021 0.075731202
## hp          -0.01764654  0.0141506 -1.247052 0.223087932
```

```
summary(fit3)$coef
```

```
##       Estimate Std. Error   t value     Pr(>|t|)
## am    4.299519  1.0241147  4.198279 2.329423e-04
## wt   -3.185455  0.4827586 -6.598442 3.128844e-07
## qsec  1.599823  0.1021276 15.664944 1.091522e-15
```

At first glance, we can make the following observations about these models:

- The model with 4 regressors and an intercept has components that we would reject as being significant at even the 10% level. This will let us focus on just the models with 3 regressors or less.
- The transmission indicator regressor 'am' encounters Simpson's paradox when regressed along with weight.
  - Notice that the sign for the estimate switches from positive to negative and back to positive when going from a 1 regressor model to a 3 regressor model.
- The intercept becomes less significant as we add more regressors to the model.
  - In the next section, we'll explain why we want to keep the intercept.
- Intuitively, with the model without the intercept, the regressors have more impact on the predicted mpg than the model with the intercept.
- Focusing on the model fit3Int, we can see that a transmission type of 1 will result in a greater estimated mpg than a transmission type of 0.
  - The car weight has a great negative influence on the predicted mpg.
  - The quarter mile time appears to have a small impact, but this variable has larger values than wt and am.
  - One would think that quarter mile time would be a result of several other car components, but it turns out that qsec is an excellent consolidated metric for a car's power in this dataset.

Variance Inflation Factors

```
vif(fit3Int)
```

```
##       am       wt     qsec 
## 2.541437 2.482952 1.364339
```

```
vif(fit3)
```

```
## Warning in vif.default(fit3): No intercept: vifs may not be sensible.
```

```
##        am        wt      qsec 
##  2.186566 13.488703 17.217505
```

With the model that has no intercept component, the variance inflation factors are much greater than the model with the intercept. For this reason we will elect to keep the intercept even though our p-value of 3 regressor model for the intercept tells us that the intercept is insignificant.

Analysis of Variance

```
anova(fit1Int,fit2Int,fit3Int,fit4Int)
```

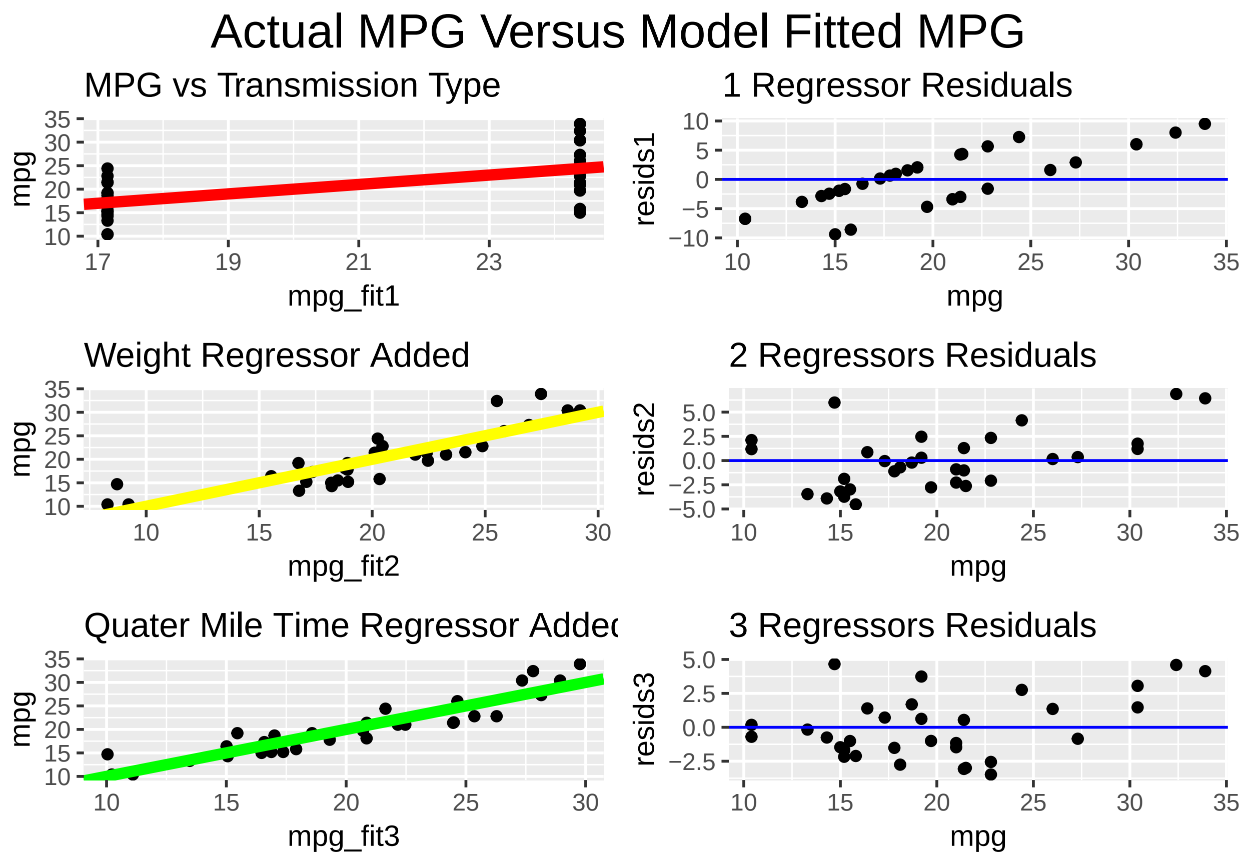
```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + qsec
## Model 4: mpg ~ am + wt + qsec + hp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     29 278.32  1    442.58 74.6538  2.96e-09 ***
## 3     28 169.29  1    109.03 18.3918 0.0002055 ***
## 4     27 160.07  1      9.22  1.5551 0.2230879    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This F test shows us that *the model with 3 regressors is significant* at the .1% level. The 4th regressor, hp, was tested in this report because it was the variable that had the next most significant F Test result of the remaining mtcars variables. This anova test proves that we should only be regressing 3 variables.

```
mtc2<-cbind(mtcars,mpg_fit1=fit1Int$coef[1]+mtcars$am*fit1Int$coef[2])
mtc2<-cbind(mtc2,mpg_fit2=fit2Int$coef[1]+mtcars$am*fit2Int$coef[2]+mtcars$wt*fit2Int$coef[3])
mtc2<-cbind(mtc2,mpg_fit3=fit3Int$coef[1]+mtcars$am*fit3Int$coef[2]+mtcars$wt*fit3Int$coef[3]
            +mtcars$qsec*fit3Int$coef[4])

mtc2<-cbind(mtc2,resids1=resid(fit1Int))
mtc2<-cbind(mtc2,resids2=resid(fit2Int))
mtc2<-cbind(mtc2,resids3=resid(fit3Int))

g1<-ggplot(mtc2,aes(x=mpg_fit1,y=mpg))+geom_point()+
      geom_abline(slope=1,intercept = 0,lwd=2,col="red") +
      ggtitle("MPG vs Transmission Type")
g2<-ggplot(mtc2,aes(x=mpg_fit2,y=mpg))+geom_point()+
      geom_abline(slope=1,intercept = 0,lwd=2,col="yellow") +
      ggtitle("Weight Regressor Added")
g3<-ggplot(mtc2,aes(x=mpg_fit3,y=mpg))+geom_point()+
      geom_abline(slope=1,intercept = 0,lwd=2,col="green") +
      ggtitle("Quater Mile Time Regressor Added")
g4<-ggplot(mtc2,aes(x=mpg, y=resids1))+geom_point()+geom_abline(intercept=0, slope = 0, col = "blue")+
      ggtitle("1 Regressor Residuals")
g5<-ggplot(mtc2,aes(x=mpg, y=resids2))+geom_point()+geom_abline(intercept=0, slope = 0, col = "blue")+
      ggtitle("2 Regressors Residuals")
g6<-ggplot(mtc2,aes(x=mpg, y=resids3))+geom_point()+geom_abline(intercept=0, slope = 0, col = "blue")+
      ggtitle("3 Regressors Residuals")
grid.arrange(g1,g4,g2,g5,g3,g6,ncol=2, nrow=3, top = grid.text("Actual MPG Versus Model Fitted MPG",
                                                            gp=gpar(fontsize=18,font=7)))
```



Actual MPG Versus Model Fitted MPG

MPG vs Transmission Type

1 Regressor Residuals

Weight Regressor Added

2 Regressors Residuals

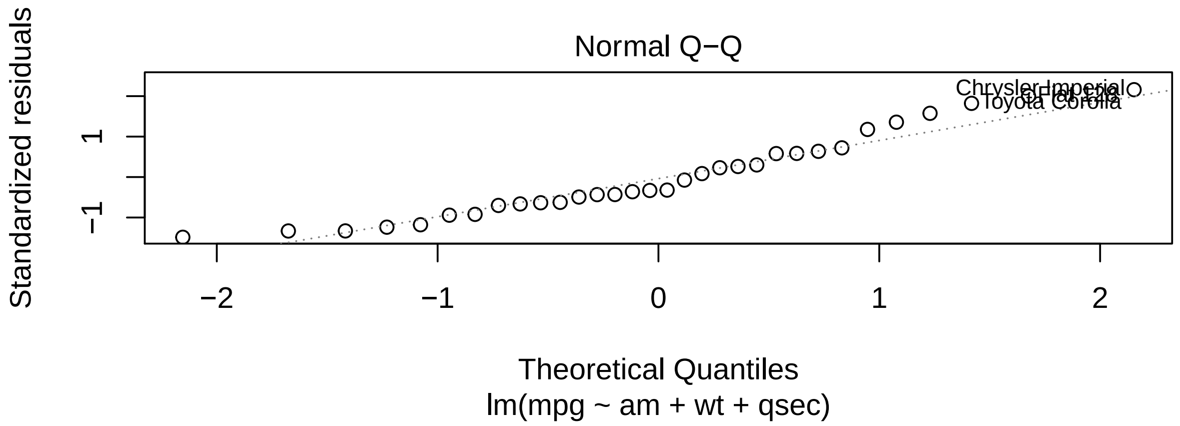
Quater Mile Time Regressor Addec

3 Regressors Residuals

Notice that as we add each of the selected regressors, our models' residuals are noticeably reduced and are more normally distributed.

Testing Resdiuals for Normality

```r
plot(fit3Int, which=2)
```



There is some cause for concern in the potential non-normality of this model's residiuals, but it is difficult to tell with such a small dataset.