Car Fuel Economy and Transmission Type

Michal Siwek

Thursday, October 22, 2015

Executive Summary

The report aims at ascertaining the effect of the difference between automatic and manual transmission on the car fuel consumption. The analysis is based on a series of tests performed on 32 cars. The conclusion is that a separate experiment should be conducted to establish the character of the relation between *mpg* and car weight. If this relation is quadratic then transmission type is irrelevant for *mpg*. If this relation is linear then transmission type has a significant impact on *mpg* and manual transmission should be used for lighter cars while automatic for heavier.

Data Exploration

There are **10 possible regressors** for our dependent variable *mpg* in the data. Two of these are nominal (*am* and *vs*) and the rest is numerical, either continuous or discrete (e.g. denoting number of cylinders, gears etc.).

First let's see if there's any **straightforward relation between *am* (transition type) and *mpg*** (see **Plot 1a in the Appendix**). It looks promising - the mean *mpg* is much lower for the automatic than for the manual transmission sample. Before jumping to conclusions we should check for confounders as most probably the study wasn't properly randomized and the two groups may significantly differ with respect ot other features.

Following pairs of variables are strongly correlated (you may find correlation values in List 1 in the Appendix):

1. *cyl* and *disp* - number of cylinders and their total volume
2. *disp* and *wt* - total engine volume and car weight
3. *cyl* and *hp* - number of cylinders and gross horsepower
4. *cyl* and *vs* - number of cylinders and vee or straight engine

The first pair measures the same thing so we shouldn't include both *cyl* and *disp* in one regression. **Following variables are strongly correlated with our dependent variable** *mpg*: *wt* (weight), then *cyl*, *disp* and *hp* (you may find correlation values in List 2 in the Appendix).

Regression and Models Selection

Following **bottom-up selection strategy** will be applied:

1. Fit the minimal model including *am* (transmission type) as the only regressor.
2. One by one add more regressors as potential confounders for *am* effect on *mpg*.
3. Use anova method to select the best model.

As for the order of potential confounders **I will follow the order of correlation strength**. It might be wrong - removing the linear effect of one of them may yield different relations between the residuals of mpg and the other variables. Still I will follow this path. Starting with *wt* (weight) is reasonable from the domain knowladge viewpoint as weight seems to have biggest impact on the fuel consuption.

Let's fit **the simplest models**, i.e. without *wt* (model 1) and with it (model 2):

```r
attach(mtcars)
fit1 <- lm(mpg ~ am)
fit2 <- lm(mpg ~ am + wt)
```

Model 1 shows *am* significant, one should expect additional 7.2 mpg by switching from automatic to manual transmission. In model 2 *wt* is significant while *am* is not. It looks like *wt* is the true driver of *mpg* and *am* is irrelevant (see **plot 1b in the Appendix**). However the residual plot (see **plot 2 in the Appendix**) shows nonlinearity in the data (*mpg* is underestimated for extreme values of *wt*). Let's try **two ways of accouting for this nonlinearity**: adding interactions (model 2a) or wt^2 (model 2b):

```r
fit2a <- lm(mpg ~ am * wt)
fit2b <- lm(mpg ~ am + wt + I(wt^2))
```

Both models give big increase in R^2, however in model 2b *am* becomes insignificant. Let's have a look on the data and the estimated regression lines to understand this situation (see **plot 3 in the Appendix**). We see that the problem is that **manual transmission was tested for lighter cars while the automatic transmission for heavier**. We have very little overlap there so at this point **deciding whether transmission type has any significant impact on *mpg* is up to the model assumptions**. Either the relationship between *mpg* and weight is quadratic and transmission type is irrelevant or the relation between *mpg* and weight is linear and then transmission type significantly affects parameters of this relation and one should use automatic type for heavier cars and manual for lighter cars to save fuel.

Let's fit **models with more regressors** and run *anova* to select the best one.

```r
fit3a <- lm(mpg ~ am * wt + cyl)
fit3b <- lm(mpg ~ am + wt + I(wt^2) + cyl)
fit4a <- lm(mpg ~ am * wt + cyl + hp)
fit4b <- lm(mpg ~ am + wt + I(wt^2) + cyl + hp)
anova(fit1, fit2, fit2a, fit3a, fit4a)
anova(fit1, fit2, fit2b, fit3b, fit4b)
```

In both lines of models anova suggest stopping at *hp* as the last regressor.

Results

As p-value for *am* in model 3b is very high (0.3819692) I only print **the coefficients for model 3a**:

```r
summary(fit3a)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 34.282998  2.7964507 12.259468 1.518742e-12
## am          11.938516  3.8453256  3.104683 4.438319e-03
## wt          -2.368930  0.8243992 -2.873523 7.815636e-03
## cyl         -1.181366  0.3802985 -3.106417 4.419268e-03
## am:wt       -4.197434  1.3115498 -3.200362 3.496375e-03
```

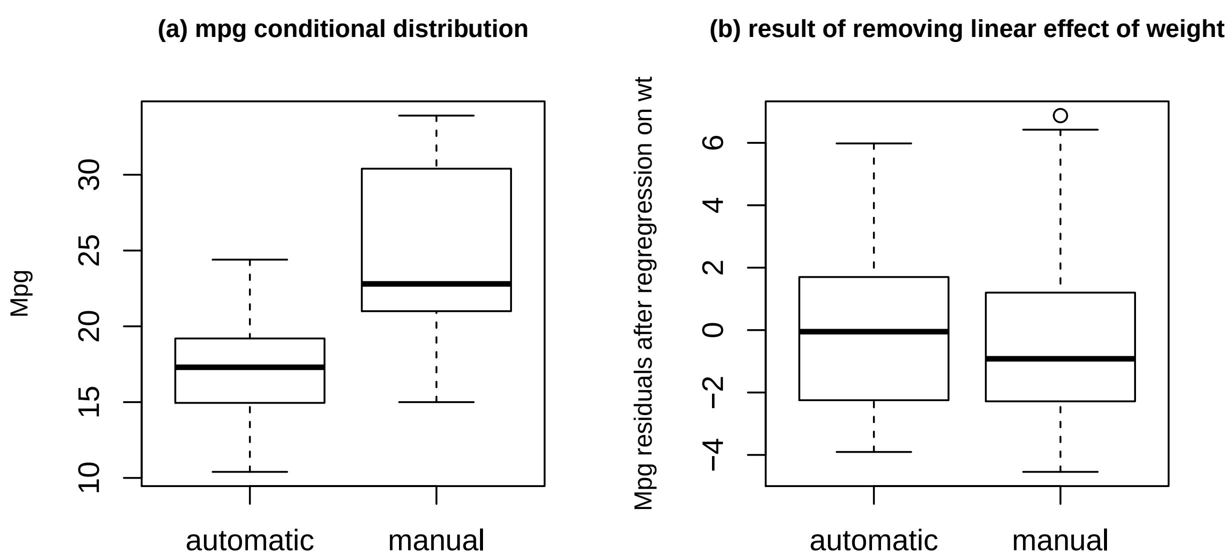
The conclusion is that either there's a quadratic relation between *mpg* and weight (and transmission is irrelevant for *mpg*, as shown in model 3b) or this relation is linear and then the transmission type has following impact on the mpg:

- the expected mpg is 11 miles higher for cars with manual transmission comparing with cars with automatic transmission, if the cars weights were 0
- every additional 1000 lb in weight decreases the mpg by 2.4 mile for cars with automatic transmission and by 6.6 mile for cars with manual transimission

Appendix

Plot 1: Conditional distribution of *mpg* on *am* (transmission type)

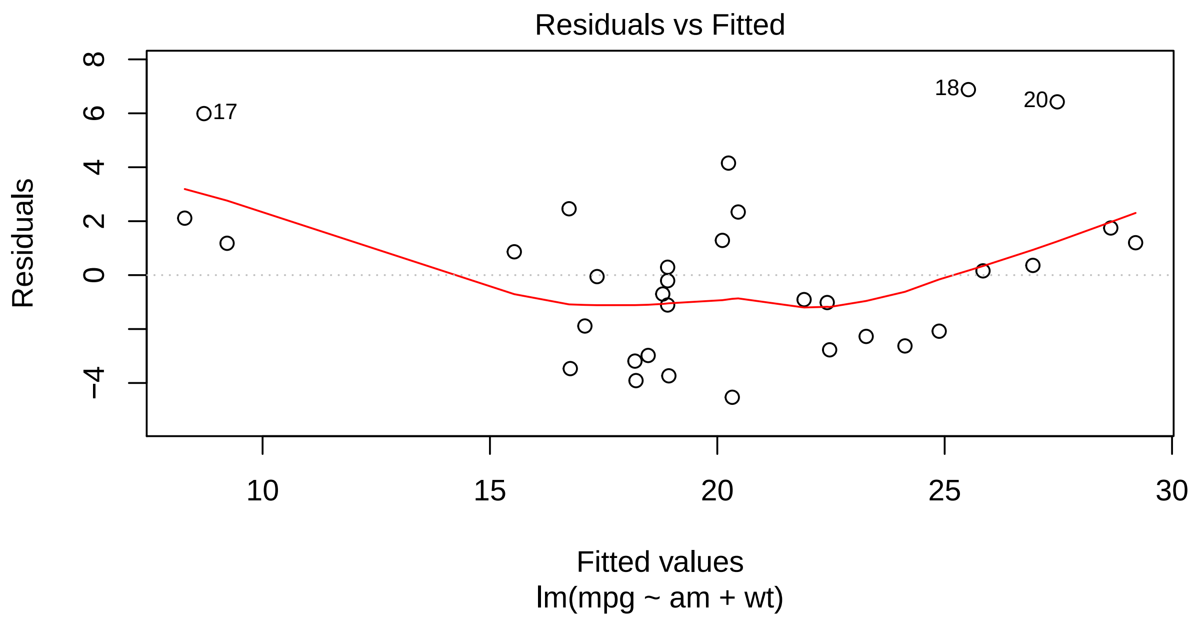
```r
am2 <- factor(am, levels = c(0, 1), labels = c("automatic", "manual"))
par(mfrow = c(1, 2))
boxplot(mpg ~ am2, ylab = "Mpg", cex.lab = .8)
title(main = "(a) mpg conditional distribution", cex.main = .8)
boxplot(lm(mpg ~ wt)$residuals ~ am2,
        ylab = "Mpg residuals after regregression on wt", cex.lab = .8)
title(main = "(b) result of removing linear effect of weight",  cex.main = .8)
```



It looks like accouning for linear effect of weight makes transmission type irrelevant.

Plot 2: residuals from model 2

```r
plot(fit2, which = 1)
```

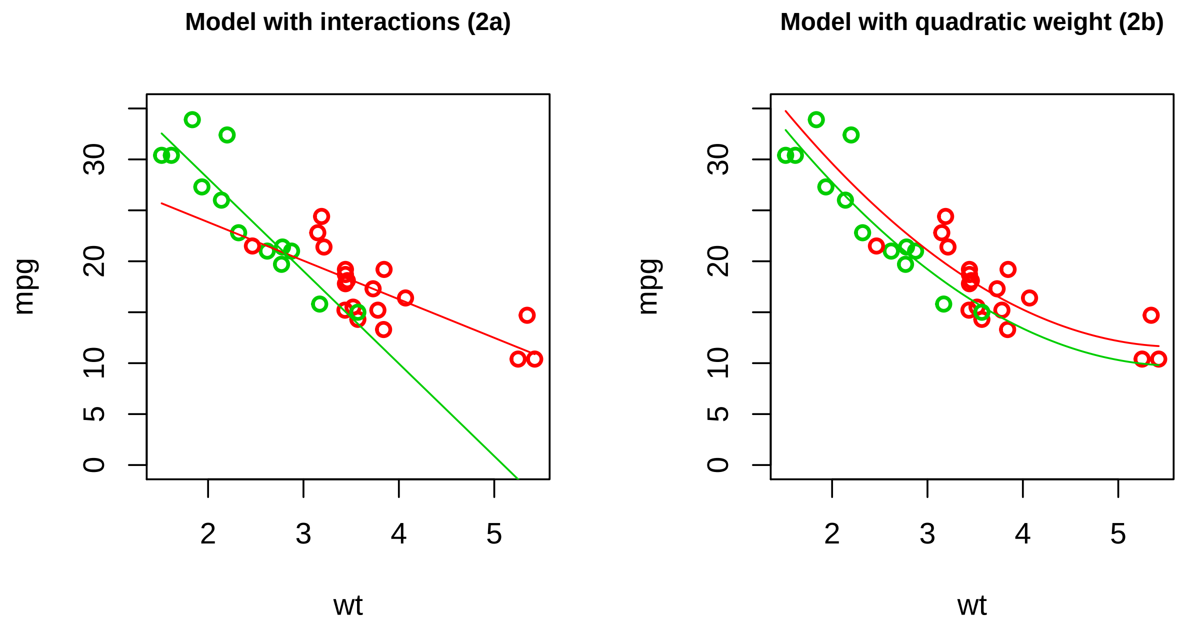


Residuals suggest nonlinearity in the data as mpg is underestimated for extreme values of weight.

Plot 3: The estimated regression lines for models 2a and 2b

```
par(mfrow = c(1, 2))

plot(wt, mpg, xlim = range(wt), ylim = c(0, 35), col = am + 2, lwd=2)
curve(fit2a$coefficient[1] + fit2a$coefficient[3] * x,
      xlim = range(wt), ylim=c(0,35), add = T, col = 2)
curve(fit2a$coefficient[1] + fit2a$coefficient[2] +
          (fit2a$coefficient[3] +fit2a$coefficient[4]) * x,
      xlim = range(wt), ylim=c(0,35), add = T, col = 3)
title(main = "Model with interactions (2a)", cex.main = .8)
plot(wt, mpg, xlim = range(wt), ylim = c(0, 35), col = am + 2, lwd = 2)
curve(fit2b$coefficient[1] + fit2b$coefficient[3] * x + fit2b$coefficient[4] * x^2,
      xlim = range(wt), ylim=c(0,35), add = T, col = 2)
curve(fit2b$coefficient[1] + fit2b$coefficient[2] +
          fit2b$coefficient[3] * x + fit2b$coefficient[4] * x^2,
      xlim = range(wt), ylim=c(0,35), add = T, col = 3)
title(main = "Model with quadratic weight (2b)", cex.main = .8)
```



Red color shows data points and the estimated regression lines for cars with automatic transmission, **green** shows the same features for cars with manual transmission.

List 1: Correlation between variables

```
library(tidyr); x <- as.data.frame(abs(cor(mtcars))); x$var1 <- rownames(x)
x <- gather(x, var2, cor, -var1); colnames(x)[3] <- "abs(cor)"
x$var2 <- as.character(x$var2); x <- x[x$var1 < x$var2,]
x <- x[order(x["abs(cor)"], decreasing = T),]; head(x, 4)
```

```
##    var1 var2  abs(cor)
## 24  cyl disp 0.9020329
## 58 disp   wt 0.8879799
## 56  mpg   wt 0.8676594
## 2   cyl  mpg 0.8521620
```

List 2: Correlation between *mpg* and other variables

```
head(x[x$var1 == "mpg" | x$var2 == "mpg",],4)
```

```
##    var1 var2  abs(cor)
## 56  mpg   wt 0.8676594
## 2   cyl  mpg 0.8521620
## 3  disp  mpg 0.8475514
## 4    hp  mpg 0.7761684
```