The effect of transmission type on motor fuel efficiency

th

Saturday, November 22, 2014

Executive Summary

Do cars with manual transmission or cars with automatic transmission perform more fuel efficiently ? In this document we build a multivariate regression model on the Motor Trend Car Road Tests (1973-74) data, which shows that on average a car with a manual transmission will get you 2.08 miles further per gallon.

But does it really? Certainly marginal association has the manual transmission beating the automatic transmission. But light cars tend to have a manual transmission, while heavy cars tend to have an automatic one. Even if we consider the weight class where we have samples for both sets, manual cars drive 0.7 mile further on the same gallon!

Data cleaning and processing

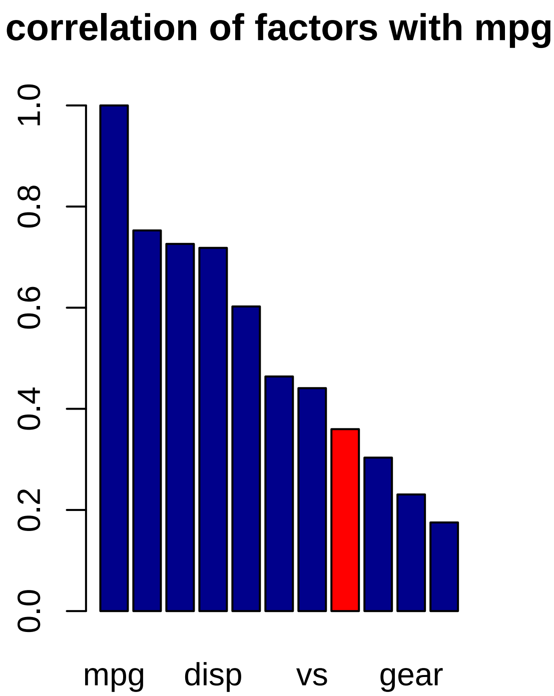
The factor *am* which is of primary interest a numeric value, while it really represent 2 categories of cars - those with either manual or automatic transmission. We transform it into a factor to make it more clear, and label the values A and M.

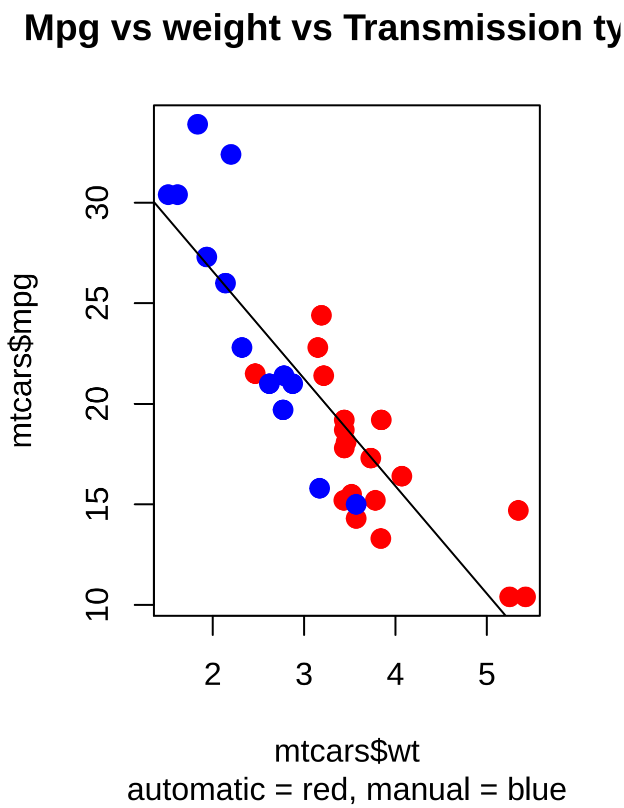
```r
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("A", "M")
```

Exploratory Data Analysis

We are using the 1974 Motor Trend Car Road Tests data set (mtcars) for our analysis, and are mostly interested in how transmission type affects the fuel efficiency (Miles per Gallon).

Lets explore the model to see which factors contribute to mpg. We notice that several factors are correlated to mpg. We will need to create a model that shows which factors are confounding between am and mpg, and will do so in detail later in this document.





Lets visualize the data to see Below we plot a regression line (mpg against weight) and indicate the transmission types in different colours.
It seems that heavier cars tend to have automatic transmission, while lighter cars tend to have manual transmission. The data set has a range between 2.5lb and 3.5lb where both manual and automatic transmission samples are available. In this range, the automatic cars tend to have a higher mpg.

==> diferent slopes for blue and red ==> weight affects automatic cars less

Creating the model

We consider the correlation values to determine which factors we need to include in our model. Highly correlated factors are good candidates to be added. We will perform nested model testing to determine which factors should be added.

```
##    names        r2        cor         pval sig
## 1    mpg 1.0000000  1.0000000 0.000000e+00 ***
## 6     wt 0.7528328 -0.8676594 1.293959e-10 ***
## 2    cyl 0.7261800 -0.8521620 6.112687e-10 ***
## 3   disp 0.7183433 -0.8475514 9.380327e-10 ***
## 4     hp 0.6024373 -0.7761684 1.787835e-07 ***
## 5   drat 0.4639952  0.6811719 1.776240e-05 ***
## 8     vs 0.4409477  0.6640389 3.415937e-05 ***
## 9     am 0.3597989  0.5998324 2.850207e-04 ***
## 11  carb 0.3035184 -0.5509251 1.084446e-03 ***
## 10  gear 0.2306734  0.4802848 5.400948e-03 ***
```

```
## 7   qsec 0.1752963  0.4186840 1.708199e-02  **
```

The factors wt, cyl, disp, hp, drat, vs and am are all correlated with mpg. The p values indicate that indeed these are all in the 99% significance interval. Still, in the model creation we will need to take care not to include too many variables, in order to avoid the standard error to increase.

```
  m1 <- lm(mpg~ wt -1, data=mtcars)
  m2 <- lm(mpg~ wt + am -1, data=mtcars)
  best <- lm(mpg~ wt + am + hp -1, data=mtcars)

  anova(m1, m2, best)
```

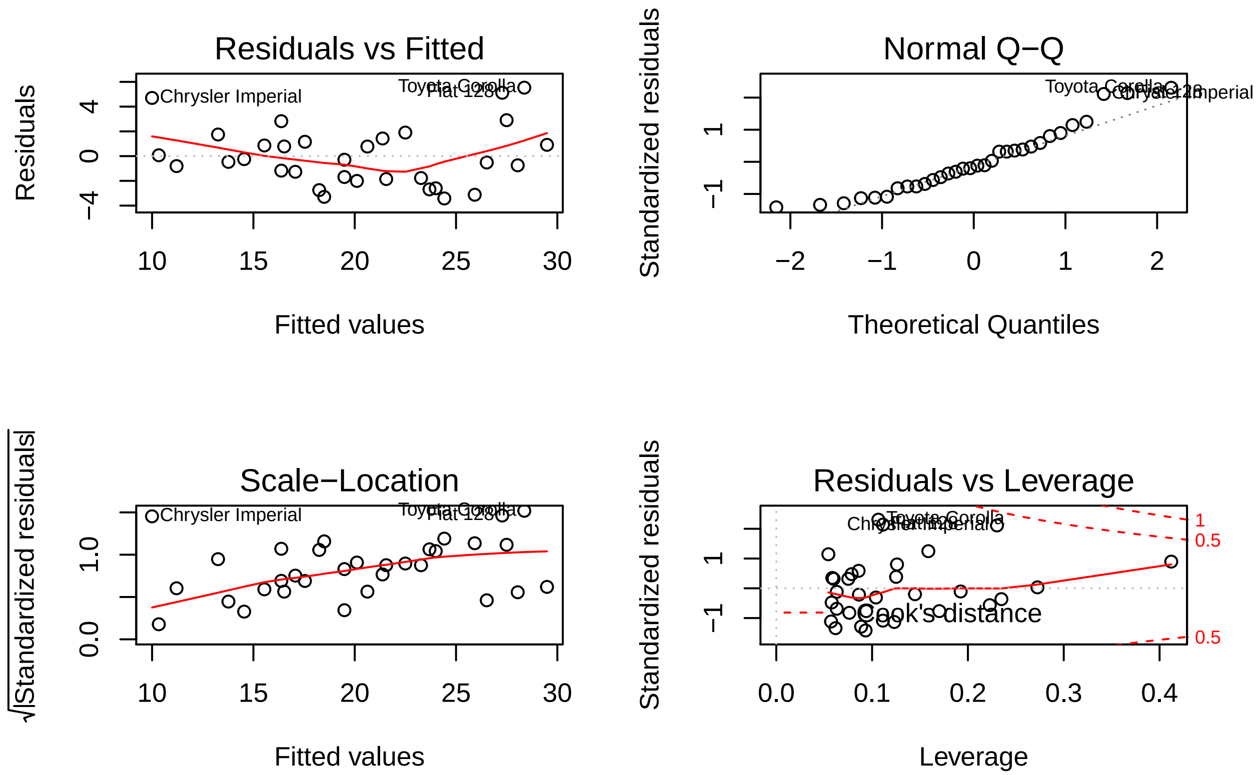
```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt - 1
## Model 2: mpg ~ wt + am - 1
## Model 3: mpg ~ wt + am + hp - 1
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     31 3936.6
## 2     29  278.3  2    3658.3 284.075 < 2.2e-16 ***
## 3     28  180.3  1      98.0  15.224 0.0005464 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.0.0005464, we reject the null hypothesis and conclude that adding hp is creating a significantlty different multivar model than the model without hp.

Residual analysis

Lets review the residuals of our model, to check for signs of temporal patterns, heteroskedasticity and non-normality.

```
par(mfrow = c(2, 2))
fit<-(lm(mpg~ wt + am + hp -1, data=mtcars)); plot(fit)
```



The plots of Fitted values vs Residuals or vs Standardized residuals show no residual pattern, so we can conclude there is no residual heteroskedasticity. The Normal Q-Q plot shows a normal distribution of the residuals.

Interpreting the model

```
best <- lm(mpg~ wt + am + hp -1, data=mtcars)
summary(best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + hp - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## wt  -2.878575   0.904971  -3.181 0.003574 **
## amA 34.002875   2.642659  12.867 2.82e-13 ***
## amM 36.086585   1.736338  20.783  < 2e-16 ***
## hp  -0.037479   0.009605  -3.902 0.000546 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.9872, Adjusted R-squared:  0.9853
## F-statistic: 538.2 on 4 and 28 DF,  p-value: < 2.2e-16
```

Our model explains 98% of the variance. The most significant influence on mpg is wt. For every half ton increase, mpg reduces with 2.88.

Keeping all other factors constant, then the model predicts that changing from manual to automatic transmission will reduce the mpg with 2.08

If we only take the weight range into account for which we have samples of both automatic and manual cars, the difference reduces to a drop by 0.709

```r
min <- min(mtcars$wt[mtcars$am == "A"])
max <- max(mtcars$wt[mtcars$am == "M"])
mtcars_mix <- mtcars [mtcars$wt <= max & mtcars$wt >=min,]

best <- lm(mpg~ wt + am + hp -1, data=mtcars_mix)
summary(best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + am + hp - 1, data = mtcars_mix)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8849 -1.0114 -0.3166  1.3606  3.1205
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## wt  -2.508884   2.055383  -1.221  0.24390
## amA 31.100540   5.975025   5.205  0.00017 ***
## amM 31.809632   5.033472   6.320 2.66e-05 ***
## hp  -0.029318   0.009737  -3.011  0.01003 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.808 on 13 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9911
## F-statistic: 476.6 on 4 and 13 DF,  p-value: 5.923e-14
```