Effect of transmission type on efficiency of classic cars

Vineet W. Singh

15 April 2018

Summary

The mtcars data set provides technical data and efficiency for a set of classic cars sold in the US, prior to the fuel crises of the 1970's.

Using this data set, this analysis tries to address the following questions:
1) Does the transmission type contribute to the efficiency of classic cars? and
2) What is the average difference in mpg of cars with different transmission types?

The efficiency of a car, is usually measured in the average number of miles per gallon (***mpg***) it runs and the ***mpg*** depends upon a number of actors. Linear models were made and all factors in the data set were analysed in a systemic way, to find out which variables/factors have the most significant effects on the efficiency (***mpg***) of these classic cars.

It is observed that besides the transmission type(***am***), ***mpg*** mainly depends upon the weight (***wt***) and the horse power (***hp***) of the car and this relationship is approximated by the following linear equation/model:
mpg = 34.003+2.084(am)-2.879(wt)-0.038(hp)

Analysis

Exploratory data analysis involved making box plots (Appendix) in which the efficiency (in ***mpg***) was grouped by the transmission type (***am***). From the box plots itself, it can be seen that manual transmission cars (MT/***am*** = 1) are more efficient and give higher ***mpg*** than automatic transmission (AT/***am*** = 0) cars. However, there is considerable variation in the ***mpg*** within each transmission type (***am***) and effects of other factors should also be analysed. To begin with, the simplest of models is made, and this calculates as to how ***mpg*** varies by ***am*** (AT/MT). Computation 1 (Appendix) provides the first linear model: ***mpg=17.147+7.245(am)***. The coefficients of this simple model show that a MT car runs 24.392 ***mpg*** compared to 17.145 ***mpg*** for an AT car.

Any model that calculates car ***mpg***, should also take into consideration, that ***mpg*** of any car is related to it's weight. *Should weight be included in the model?*
In computation 2 (Appendix), a new model is made in which weight is included, residuals are tested for normalility and ANOVA done for the previous model and the new model.
The null hypothesis (NH) proposed for ANOVA is that omitting ***wt*** will not increase the bias in the model. Alternative hypothesis (AH) is that omitting ***wt*** will increase the bias. From ANOVA, we find that the F score is 46.115 (P value <.0001), which is significant. We therefore, reject the NH, accept the AH and include ***wt*** to improve the model. The model changes to: ***mpg=37.322-0.024(am)-5.353(wt)***

Next we need to test, whether to include other engine variables (like horse power ***hp***) that might effect ***mpg***. In computation 3 (Appendix), ***hp*** is added to the model and evaluated. Including ***hp*** as a regressor in the model, increases variance of both the ***wt*** and ***am*** coefficients, thereby increasing the overall variance of the model. *Should **hp** be included?*
From the shapiro test, residuals of the updated model are normal so ANOVA can be performed. ANOVA tests the NH that omitting the ***hp*** will not increase bias in the model and gives us a F score of 15.224 (P value < .001), which is significant. The NH is rejected and ***hp*** is included in the model which now transforms to: ***mpg = 34.003+2.084(am)-2.879(wt)-0.038(hp)***

***hp** normally depends upon **disp** and **cyl**. Should these be included?*
We include ***disp*** and ***cyl*** in the old model and evaluate it - computation 4 (Appendix). As suspected, since

hp depends upon ***disp*** and ***cyl***, it is strongly correlated with both and including them increases variances of the coefficients of ***am,wt,hp***. From the shapiro test, residuals are normally distributed, ANOVA can be done. NH is the same as before i.e. omitting ***disp*** and ***cyl*** will not increase the bias of the model. ANOVA of the two models gives F score of 1.369 (P values .27) which is not significant. Also, there is no significant reduction in the RSS (residual sum of squares): 180 vs 163, between the old and new models.
Therefore, we can accept the NH and omit ***disp*** and ***cyl*** from our model.

Similarly, we include all variables (in addition to ***am , wt and hp***) in the final testing model. Based on results of computation 5 (Appendix), it is inferred that including all the other variables in the testing model, increases the variances of the coefficients that matter but does not significantly reduce the RSS: 180 vs 147, between the last model and the testing model. There is no gain in including any regressors other than ***am , wt and hp*** in the model.

Conclusion

Based on the analysis, the linear model specified by ***mpg = 34.003+2.084(am)-2.879(wt)-0.038(hp)*** with a standard error of 2.538 ***mpg*** is the best fit to the data provided.
Based on this model, we can conclude that a MT car will run 2.084 miles more than an AT car with the same ***wt/hp***. Each ton increase in ***wt*** will decrease the car ***mpg*** by 2.879 ***mpg*** and one ***hp*** increase in engine power will decrease the car ***mpg*** by 0.038 mpg.
Finally, the error between the ***mpg*** for any data point in the data set and the ***mpg*** calculated using the model should not exceed +/- 2.538 ***mpg***.

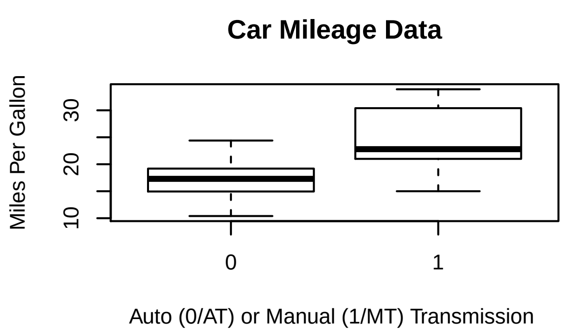
Appendix

EDA-Box Plot

```
data(mtcars)
require('car')
```

```
## Loading required package: car
```

```
par(cex=0.7)
boxplot(mpg~factor(am),data=mtcars, main="Car Mileage Data",
    xlab="Auto (0/AT) or Manual (1/MT) Transmission ", ylab="Miles Per Gallon")
```



Computation 1

```
mdl1<-lm(mpg~factor(am),mtcars)
mdl1$coefficients
```

```
## (Intercept) factor(am)1
##   17.147368    7.244939
```

```r
shapiro.test(mdl1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdl1$residuals
## W = 0.98208, p-value = 0.8573
```

Computation 2

Shapiro tests the null hypothesis that the data provided is normal and if so ANOVA may be done to compare different models.

```r
mdl2<-lm(mpg~factor(am)+wt,mtcars)
mdl2$coefficients
```

```
## (Intercept) factor(am)1          wt
## 37.32155131 -0.02361522 -5.35281145
```

```r
shapiro.test(mdl2$residuals)
```

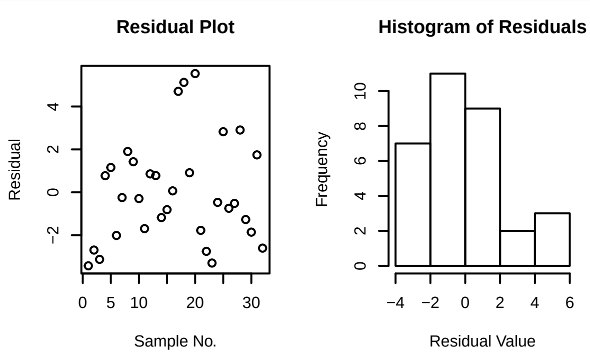
```
##
##  Shapiro-Wilk normality test
##
## data:  mdl2$residuals
## W = 0.94478, p-value = 0.1024
```

```r
anova(mdl1,mdl2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 46.115 1.867e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Computation 3

```r
mdl9<-lm(mpg~factor(am)+wt+hp,mtcars)
par(mfrow=c(1,2), cex=0.5)
plot(mdl9$residuals,main="Residual Plot",xlab="Sample No.", ylab="Residual" )
hist(mdl9$residuals, main="Histogram of Residuals",ylab="Frequency",
     xlab="Residual Value")
```



```r
shapiro.test(mdl9$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdl9$residuals
## W = 0.9453, p-value = 0.1059
```

Plot shows even dispersion of residuals and Histogram shows residuals are approximately normal.

```r
vif(mdl2)
```

```
## factor(am)         wt
##   1.921413   1.921413
```

```r
vif(mdl9)
```

```
## factor(am)         wt         hp
##   2.271082   3.774838   2.088124
```

```r
anova(mdl2,mdl9)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt
## Model 2: mpg ~ factor(am) + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 278.32
## 2     28 180.29  1    98.029 15.224 0.0005464 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(mdl9)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + wt + hp, 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|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## factor(am)1  2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## 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:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Computation 4

```r
mdl11<-lm(mpg~factor(am)+wt+hp+disp+cyl,mtcars)
vif(mdl11)
```

```
## factor(am)         wt         hp       disp        cyl
##   2.553064   6.079452   4.501859  10.401420   7.209456
```

```r
shapiro.test(mdl11$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdl11$residuals
## W = 0.94786, p-value = 0.1253
```

```r
anova(mdl9,mdl11)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt + hp
## Model 2: mpg ~ factor(am) + wt + hp + disp + cyl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     26 163.12  2    17.171 1.3685 0.2722
```

Computation 5

```r
mdl12<-lm(mpg~factor(am)+wt+hp+disp+cyl+drat+qsec+factor(vs)+carb+gear,mtcars)
vif(mdl12)
```

```
## factor(am)         wt         hp       disp        cyl       drat
##   4.648487  15.164887   9.832037  21.620241  15.373833   3.374620
##       qsec factor(vs)       carb       gear
##   7.527958   4.965873   7.908747   5.357452
```

```r
shapiro.test(mdl12$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mdl12$residuals
## W = 0.95694, p-value = 0.2261
```

```r
anova(mdl9,mdl12)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + wt + hp
## Model 2: mpg ~ factor(am) + wt + hp + disp + cyl + drat + qsec + factor(vs) +
##     carb + gear
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     21 147.49  7    32.797 0.6671 0.6973
```