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 factors. 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* also 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

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

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 efficiency of any car is proportional to it's weight. Should weight be included in the model?
In computation 2 (Appendix), a new model is made in which weight (*wt*) is included.
The model residuals are tested for normalility by using the shapiro test. The NH for the shapiro test is that the residuals follow a normal distribution. Shapiro results in a p-value of .10 and the NH is accepted. ANOVA is possible between 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*. *hp* is added to the model and evaluated as below:

```r
mdl9<-lm(mpg~factor(am)+wt+hp,mtcars)
vif(mdl2)
```

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

```r
vif(mdl9)
```

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

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

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

```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
```

Including *hp* increases variance of both the *wt* and *am* coefficients, so do we retain *hp* as a regressor in our model?
The residuals of the testing model are plotted in the Appendix and they show an even dispersion.The histogram of the residuals shows an approximately normal distribution.
From the shapiro test, residuals of the updated model are normally distributed (p-value: .11) so ANOVA can be performed between this model and the previous one. ANOVA tests the NH that omitting the *hp* will not increase bias in the model.
ANOVA results in 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)***
From the summary of the model (computation 3, Appendix), we can observe that the residual standard error is 2.538 *mpg*.

In the final test, computation 4 (Appendix), we include all variables (in addition to *am , wt and hp*) in the testing model.
The shapiro test on residuals gives a P-value of .23 which implies that the residuals are normally distributed and we can perform ANOVA on the testing model and the previous model.
The NH for the ANOVA is: Omitting all regressors except *am , wt and hp* will not increase bias in the model. Subsequent ANOVA between the previous model and the final testing model gives an F score of 0.667 (P-value: .70) which is not significant, so we cannot reject the NH.
It is inferred that including 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 models. Therefore there is no gain in including any regressors other than *am , wt and hp* in the model.

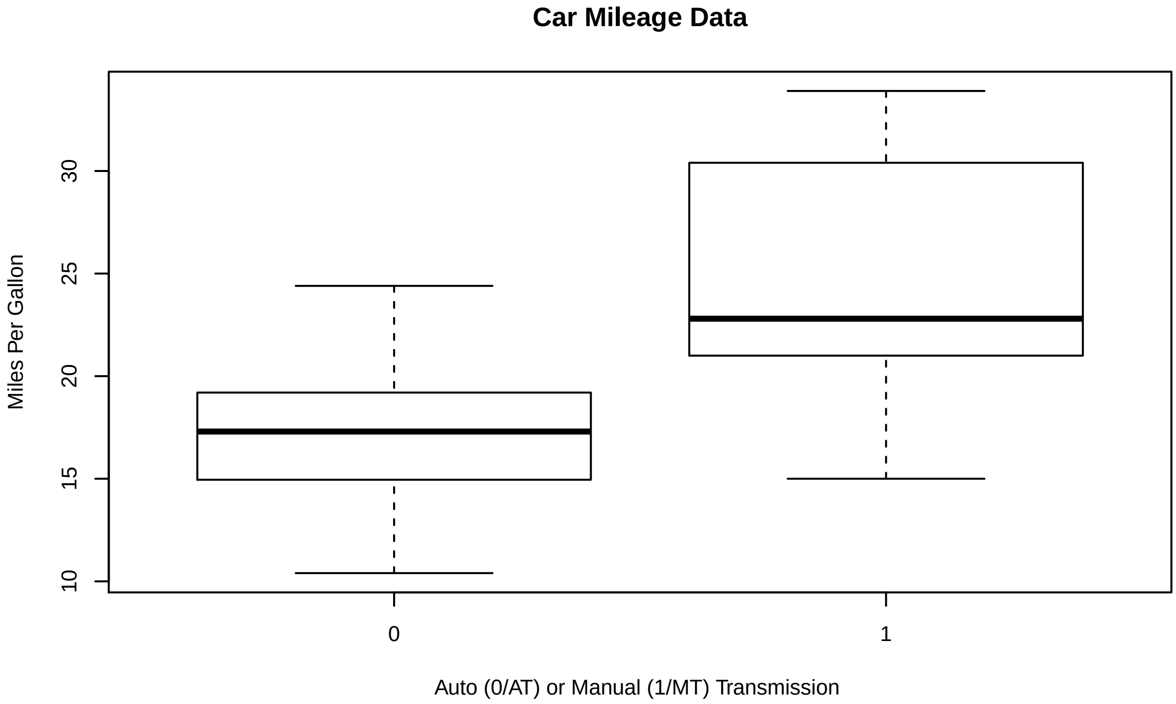
Conclusion

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.
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.
Approximately 66% of cars (data points) of this sample set will have an *mpg* that is within an error margin of +/- 2.538 *mpg* of the *mpg* predicted by the equation/model.

Appendix

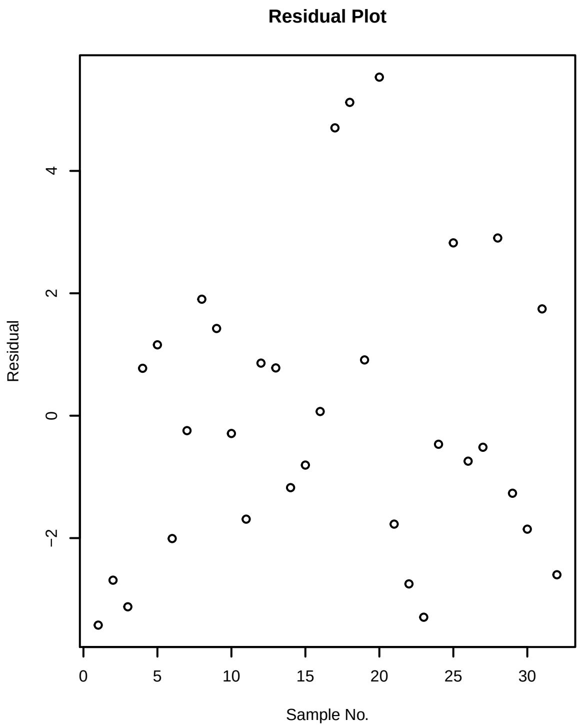
EDA-Box Plot

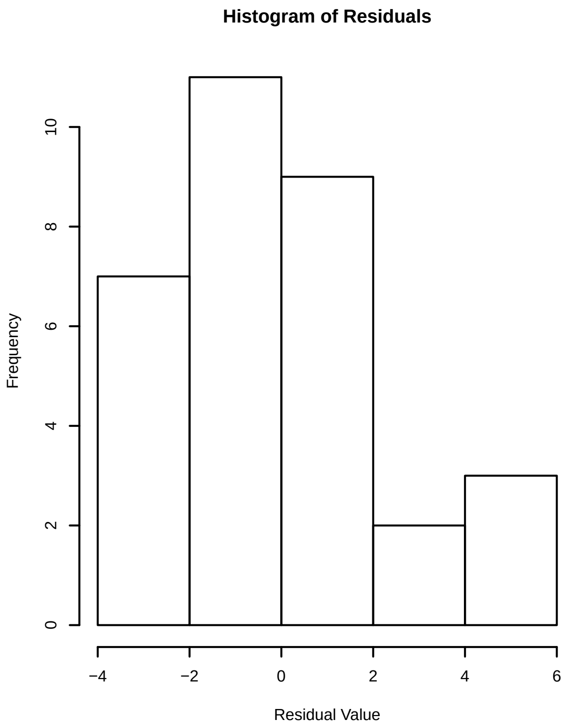
```
par(cex=0.7)
boxplot(mpg~factor(am),data=mtcars, main="Car Mileage Data",
    xlab="Auto (0/AT) or Manual (1/MT) Transmission ", ylab="Miles Per Gallon")
```



Residual Plot and Residual Histogram for the final model

```
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")
```





Computation 1

```
mdl1<-lm(mpg~factor(am),mtcars)
mdl1$coefficients
shapiro.test(mdl1$residuals)
```

Computation 2

```
mdl2<-lm(mpg~factor(am)+wt,mtcars)
mdl2$coefficients
shapiro.test(mdl2$residuals)
anova(mdl1,mdl2)
```

Computation 3

```
summary(mdl9)
```

Computation 4

```
mdl12<-lm(mpg~factor(am)+wt+hp+disp+cyl+drat+qsec+factor(vs)+carb+gear,mtcars)
vif(mdl12)
shapiro.test(mdl12$residuals)
anova(mdl9,mdl12)
```