Relationship between mles per gallon and transmission type

Gaurav Bansal

Saturday, July 26, 2014

Executive Summary

Through stepwise model selection through minimizing AIC, the output model turned out to be mpg ~ as.factor(cyl) + hp + wt + as.factor(am). However, performing ANOVA anlysis to compare the above model with and without the am variable showed that such two models are likely to be similiar. A Shapiro-Wilk test of normality over the residues the above model failed to reject the null hypothesis, validating the anova analysis.Therefore, basing on all the studies, the conclusions are: 1. with the given data, the manual and automatic transmission types do not significantly impact the MPG; 2. with the given data, holding all other variables constant, vehicle of manual transmission have 1.81 increase in MPG compared to vehicle of automatic transmission.

```
#loading the data
data("mtcars")
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.4   Min.   :4.00   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.4   1st Qu.:4.00   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.2   Median :6.00   Median :196.3   Median :123.0  
##  Mean   :20.1   Mean   :6.19   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.8   3rd Qu.:8.00   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.9   Max.   :8.00   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.76   Min.   :1.51   Min.   :14.5   Min.   :0.000  
##  1st Qu.:3.08   1st Qu.:2.58   1st Qu.:16.9   1st Qu.:0.000  
##  Median :3.69   Median :3.33   Median :17.7   Median :0.000  
##  Mean   :3.60   Mean   :3.22   Mean   :17.8   Mean   :0.438  
##  3rd Qu.:3.92   3rd Qu.:3.61   3rd Qu.:18.9   3rd Qu.:1.000  
##  Max.   :4.93   Max.   :5.42   Max.   :22.9   Max.   :1.000  
##        am              gear            carb      
##  Min.   :0.000   Min.   :3.00   Min.   :1.00  
##  1st Qu.:0.000   1st Qu.:3.00   1st Qu.:2.00  
##  Median :0.000   Median :4.00   Median :2.00  
##  Mean   :0.406   Mean   :3.69   Mean   :2.81  
##  3rd Qu.:1.000   3rd Qu.:4.00   3rd Qu.:4.00  
##  Max.   :1.000   Max.   :5.00   Max.   :8.00
```

```
#changing the values of factor varables to factors
mtcars$cyl<- as.factor(mtcars$cyl)
mtcars$vs<- as.factor(mtcars$vs)
mtcars$am<- as.factor(mtcars$am)
mtcars$gear<- as.factor(mtcars$gear)
mtcars$carb<- as.factor(mtcars$carb)
```

Fit the model taking all factors and then find AIC, the one having the lowest AIC will be the best fit model

```r
fit <- lm(mpg~.,mtcars)
library("MASS")
x <- stepAIC(fit)
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq RSS  AIC
## - carb  5     13.60 134 69.8
## - gear  2      3.97 124 73.4
## - am    1      1.14 122 74.7
## - qsec  1      1.24 122 74.7
## - drat  1      1.82 122 74.9
## - cyl   2     10.93 131 75.2
## - vs    1      3.63 124 75.4
## <none>              120 76.4
## - disp  1      9.97 130 76.9
## - wt    1     25.55 146 80.6
## - hp    1     25.67 146 80.6
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq RSS  AIC
## - gear  2      5.02 139 67.0
## - disp  1      0.99 135 68.1
## - drat  1      1.19 135 68.1
## - vs    1      3.68 138 68.7
## - cyl   2     12.56 147 68.7
## - qsec  1      5.26 139 69.1
## <none>              134 69.8
## - am    1     11.93 146 70.6
## - wt    1     19.80 154 72.2
## - hp    1     22.79 157 72.9
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq RSS  AIC
## - drat  1      0.97 140 65.2
## - cyl   2     10.42 149 65.3
## - disp  1      1.55 141 65.4
## - vs    1      2.18 141 65.5
## - qsec  1      3.63 143 65.8
## <none>              139 67.0
## - am    1     16.57 156 68.6
## - hp    1     18.18 157 68.9
## - wt    1     31.19 170 71.5
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
##
##        Df Sum of Sq RSS  AIC
```

```
## - disp  1      1.25 141 63.5
## - vs    1      2.34 142 63.8
## - cyl   2     12.33 152 63.9
## - qsec  1      3.10 143 63.9
## <none>              140 65.2
## - hp    1     17.74 158 67.0
## - am    1     19.47 160 67.4
## - wt    1     30.72 171 69.6
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq RSS  AIC
## - qsec  1       2.4 144 62.1
## - vs    1       2.7 144 62.1
## - cyl   2      18.6 160 63.5
## <none>              141 63.5
## - hp    1      18.2 159 65.4
## - am    1      18.9 160 65.5
## - wt    1      39.6 181 69.4
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq RSS  AIC
## - vs    1       7.3 151 61.7
## <none>              144 62.1
## - cyl   2      25.3 169 63.2
## - am    1      16.4 160 63.5
## - hp    1      36.3 180 67.3
## - wt    1      41.1 185 68.1
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq RSS  AIC
## <none>              151 61.7
## - am    1       9.8 161 61.7
## - cyl   2      29.3 180 63.3
## - hp    1      31.9 183 65.8
## - wt    1      46.2 197 68.2
```

The model with cyl, hp, wt, am is the best fit model, so let us find the summary of this model, wether the values are significant or not

```
fit1 <- lm(mpg~cyl+hp+wt+am,mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## am1           1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

After fitting the variable the p-value for AM is found to be 0.20646 which is not significant and we fail to reject the null hpothesis

```
fit1 <- lm(mpg~cyl+hp+wt+am,mtcars)
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## am1           1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

```
#Also, let us test the anova of two models, one with am and one without am
fit1 <- lm(mpg~cyl+hp+wt+am,mtcars)
fit2 <- lm(mpg~cyl+hp+wt,mtcars)
anova(fit1,fit2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     26 151
## 2     27 161 -1     -9.75 1.68   0.21
```

P-value is 0.21 so we fail to reject the null hypothesis, these both the models are similar

```
#test for normality of residues
shapiro.test(fit1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit1$residuals
## W = 0.9681, p-value = 0.4479
```

P-value = 0.45, fail to reject the null. So the residues are likely to follow a normal distribution. So our anova test is likely to be valid.

Apendix

```
par(mfrow = c(2, 2))
plot(fit1)
```