MotorTrend - TransmissionType effect over MPG

andhdo: AndresHurtado

18 de octubre de 2015

Executive Summary

Motor Trend, a magazine about automobile industry is interested in explore the relationshp between a set of 11 variables extracted over a collection of 32 car measures with its oil consumption (expressed in miles per gallon). For the analysis is used the 1974's MotorTrend dataset to answer the following questions:

- "Is a transmission type (automatic or manual) better for MPG"
- "How different is the MPG betweeen both"

The conclusion is that Manual Transmission achieves a higher value of MPG compared to Automatic Transmission (near 1.8 MPG), based on a linear regression model; however this difference is statistically insignificant

Exploratory Data Analysis

First we load the dataset (mtcar), look for information about the dataset (?mtcars) and do some preprocessing tasks. Specially we need to set cylinders (cyl) and type of transmission - automatic/manual (am) variables as categorical (factor variable).

```
library(ggplot2)
data(mtcars)

mtcars$cyl  <- as.factor(mtcars$cyl)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
mtcars$vs   <- as.factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

Then we explore relationships between variables (see appendix), where we can see a strong correlation between mpg and am, starting first from a visualization between all variables of the dataset (appendix 1) and second, if we go to a more detailed diagram between the variables of interest for this study (mpg and am shown in appendix 2). Now lets see the application of the model.

Model Analysis: Using Regression

We build linear regressions over the variables of the dataset to explore it as a predictors of mpg (using AIC algorithm, that is calling lm repeatedly and finding the better selection for mpg prediction)

```
initialmodel <- lm(mpg ~ ., data = mtcars)
summary(initialmodel) # results hidden
```

The initial model tell us that the model can explain near 78% of the variance of the mpg variable but none of the coeficients are significant at 0.05-significant-level.

```
bestmodel <- step(initialmodel, direction = "both")
```

see the results below

```
summary(bestmodel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## 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:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

It tell us that cyl,hp,wt cofounders and am as independent variable are candidates to fit the best model (act as predictors) and more than 84% of variability is explained combining those values as predictors (0.84 r-squared value). Now let's consider only the am variable and compare it with the previous model using anova:

```
basemodel <- lm(mpg ~ am, data = mtcars)
anova(basemodel, bestmodel)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is highly significant and we reject the hypothesis that cofounder variables (cyl,hp,wt) don't contribute to the accuracy of the model.

Model Analysis: Residuals

Residuals plots help us to find some potential problems with the model (see appendix 3). it tell us: - residuals-vs-fitted: are distributed random, so, it verifies independence condition - normal-q-q: (points near of the line): residuals normally distributed - scale-location: constant variance - increased leverage of outliers in the top right of the plots

To find points with more leverage, let's do some computations:

```
leverage <- hatvalues(bestmodel)
tail(sort(leverage),3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##           0.2777872           0.2936819           0.4713671
```

To find points that influences the model, let's do some computations:

```
influential <- dfbetas(bestmodel)
tail(sort(influential[,6]),3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

Model Analysis: Statistical Inference

We perform a t-test of the manual and automatic subsets. Now test the following: H(0): come from same distribution

```r
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

So we reject H0 (that is: we reject that mpg distributions for manual and automatic are the same)

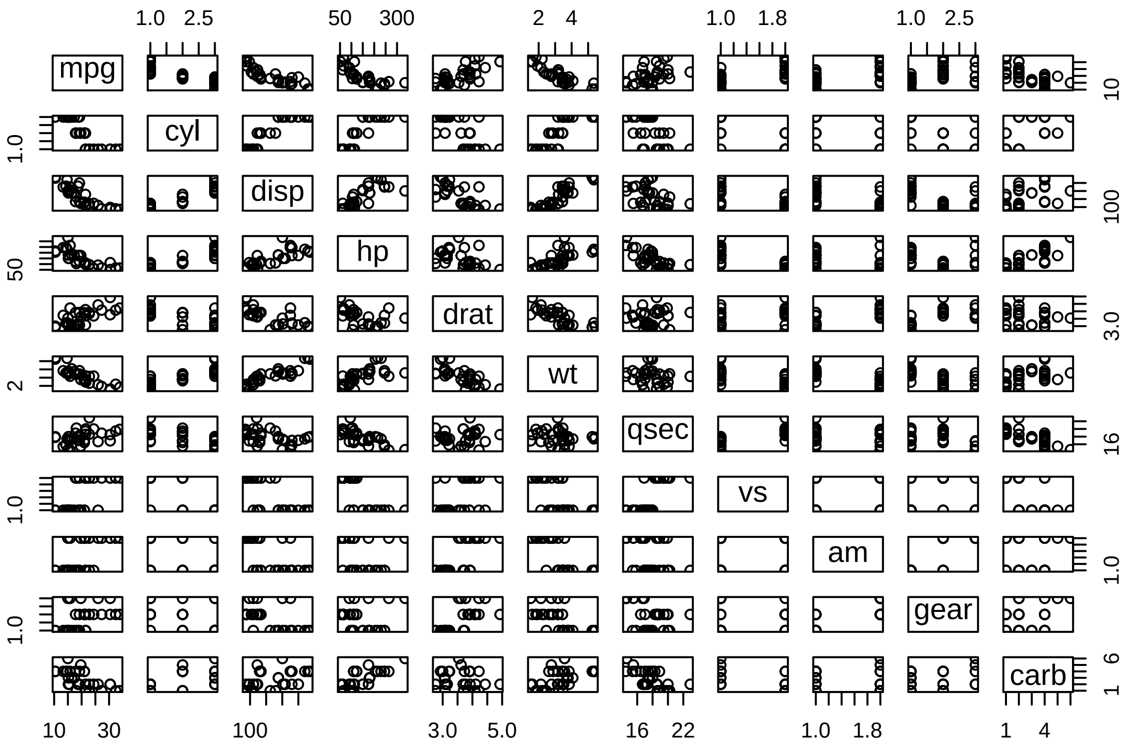
Conclusions

- mean for mpg of manual transmission is about 7 more than automatic transmission cars (24.392 - 17.147). best model assumption tell us that:
- increasing cylinders to 6 or 8, then mpg will decrease in a factor of -3.03 or -2.16 respectively (summary(bestmodel)$coef)
- increments in hp, causes only a decrease in 0.32 mpg (unit: 10hp)
- increments in weight, causes a decrease of -2.49 mpg (unit:1000lb)

Appendix

1. Relationship between variables of the dataset

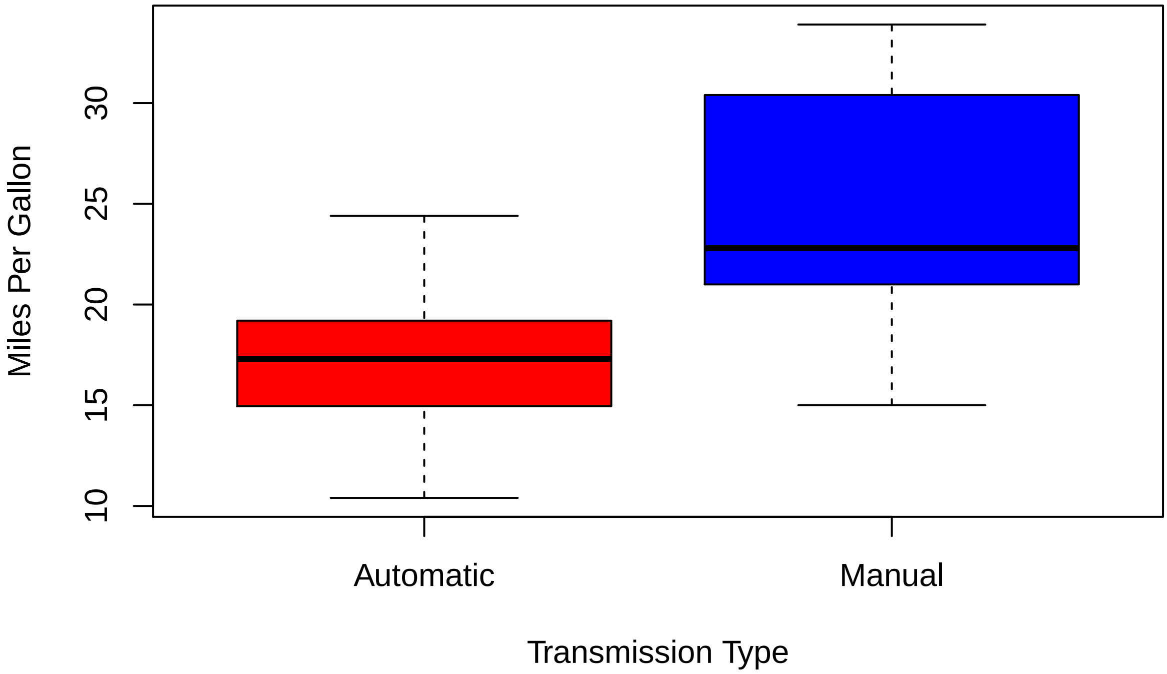
```r
pairs(mpg ~ ., data = mtcars)
```



We can see higher correlations between variables like wt,disp,cyl,hp In the quadrant mpg vs am, we can see two lines, one for automatic and one for manual. It shows variations over mpg influence. Let's zoom it with a box plot:

2. Visualization of automatic vs manual transmission

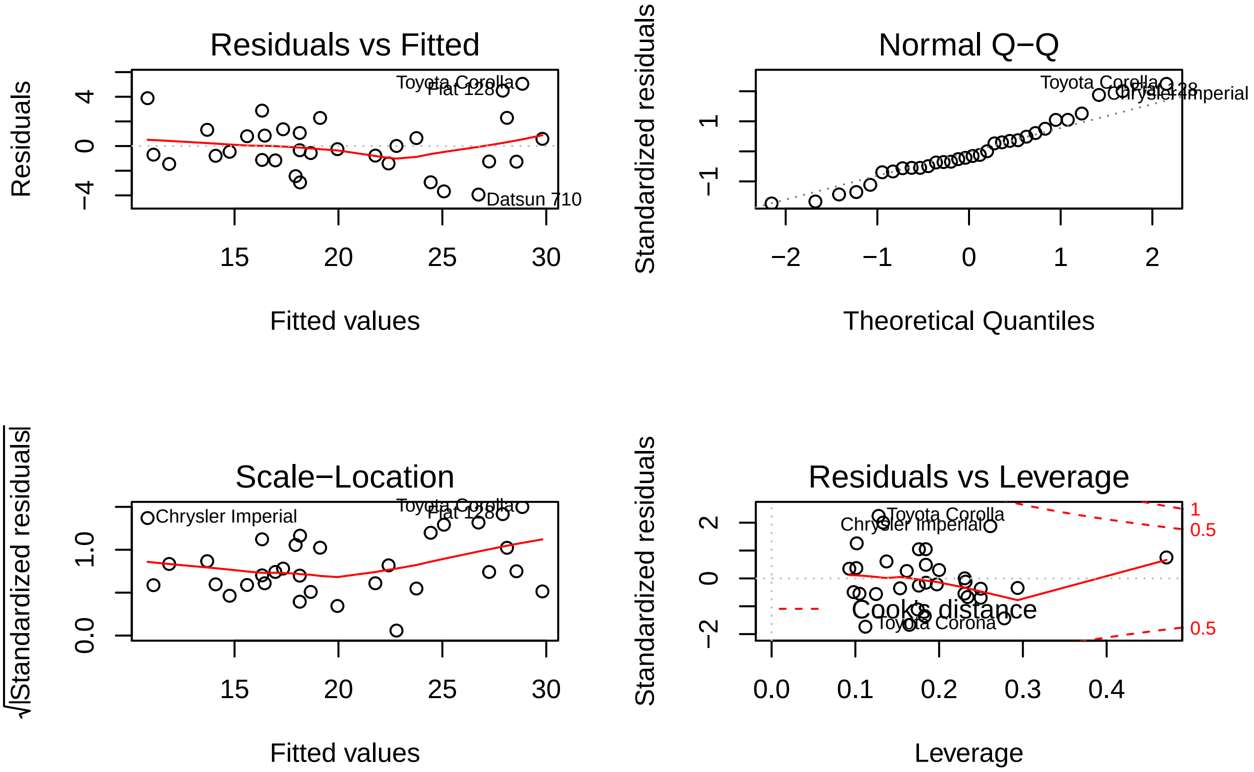
```
boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmission Typ
```



It shows that manual transmission yields higher values of mpg

3. Visualization of residuals

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Datsun 710
4
0
−4
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Fiat 128
Chrysler Imperial
1
−1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
1.0
0.0
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
2
0
−2
0.0
0.1
0.2
0.3
0.4