What factors affect fuel effiency of a car?

DomR

Saturday, September 13, 2014

Executive Summary

This report analyis data from 32 1973-1974 model cars and tries to explore if there are other factors other than transmission type that can be used to predict fuel efficiency of a car.

- This report specially tries to answer the following questions:
  - Is an automatic or manual transmission better for MPG?
  - Quantify the MPG difference between automatic and manual transmissions

Data Preparation

```
require(MASS)
require(plyr)
require(ggplot2)
require(lattice)
require(knitr)
options("scipen"=100, "digits"=4)
opts_chunk$set(fig.width=7, fig.height=5, tidy=FALSE, size='small',width=100 )
```

Load the data

```
data(mtcars)
#Convert factor variables into factors
mtcars$cyl <- as.factor(mtcars$cyl) #4, 6 or 8
mtcars$vs <- as.factor(mtcars$vs) #0- V engine or 1= straight engine.
mtcars$am <- as.factor(mtcars$am) #0 - automatic and 1 - gear
mtcars$gear <- as.factor(mtcars$gear) #3,4,5
mtcars$carb <- as.factor(mtcars$carb) #1,2,3,4
```

Exploratory Data Analysis

Base model using am as the only predictor variable keeping all other factors constant

```
baseFit <- lm(mpg~am, data=mtcars)
summary(baseFit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)    17.15       1.12   15.25 0.0000000000000011 ***
## am1             7.24       1.76    4.11            0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

```
autoSummary <- summary(subset(mtcars, am==0)$mpg)
manualSummary <- summary(subset(mtcars, am==1)$mpg)
```

- The mean, median, minimum and maximum mpg (see also Figure 1 in Appendix) for
  - automatic cars is 17.1,17.3,10.4 and 24.4 respectively.
  - manual cars is 24.4,22.8,15 and 33.9 respectively.
- The manual tranmission (am1) cars provide additional **7.245** miles per gallon of gas compared to automatic cars. However, adjusted R-squared value is only 0.3385. *Is there a better model than the baseFit?*

Identify the best model to predict miles per gallon for a car.

```
#Get a lm fit of mpg against all other factors
lmFit <- lm(mpg~., data=mtcars)
#Perform a step wise model selection of mpg versus other factors
steplmFit <- stepAIC(lmFit, direction="both")
steplmFit$anova
```

```
attr(terms(steplmFit),"term.label")
```

```
## [1] "cyl" "hp"  "wt"  "am"
```

- As shown, the significant predictor variables to predict mpg outcome is cyl, hp, wt, am and the best Fit model is

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

```
##
## 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 0.00000000000077 ***
## 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: 0.000000000151
```

- Adjusted R-square for best fit model is 0.8401 which is much better than base model. How much transmission type has effect on mpg? Let's build multiple models with cyl, hp and wt as predictor variables.

How much does transmission type really impacts mpg?

```
bestFitWithCylHpWt <- lm(mpg~cyl + hp + wt, data=mtcars)
#summary(bestFitWitCylHpWt)
bestFitWithCylWt <- lm(mpg~cyl + wt, data=mtcars)
#summary(bestFitWithCylWt)
bestFitWithWt <- lm(mpg~wt, data=mtcars)
#summary(bestFitWithWt)
bestFitWithCyl <- lm(mpg~cyl, data=mtcars)
#summary(bestFitWithCyl)

rSquaredValues <- rbind(
c("Model : mpg~cyl + hp + wt + am",summary(bestFitWithCylHpWtAm)$adj.r.squared),
c("Model : mpg~cyl + hp + wt ",summary(bestFitWithCylHpWt)$adj.r.squared),
c("Model : mpg~cyl + wt",summary(bestFitWithCylWt)$adj.r.squared),
c("Model : mpg~wt",summary(bestFitWithWt)$adj.r.squared),
c("Model : mpg~cyl ",summary(bestFitWithCyl)$adj.r.squared)
)
colnames(rSquaredValues ) <-c("Model", "Adj. R-Squared")
```

Build few models model without using am as predictor variable

- Adjusted R-square for the new models

```
rSquaredValues
```

```
##      Model                              Adj. R-Squared
```

```
## [1,] "Model : mpg~cyl + hp + wt + am" "0.840087540272603"
## [2,] "Model : mpg~cyl + hp + wt "     "0.836066778752893"
## [3,] "Model : mpg~cyl + wt"           "0.820014581578736"
## [4,] "Model : mpg~wt"                 "0.744593886780206"
## [5,] "Model : mpg~cyl "               "0.71400902925487"
```

Let's further analyze these models using anova.

```
modelCompare <- anova(bestFitWithCylHpWtAm, bestFitWithCylHpWt,bestFitWithCylWt,bestFitWithWt,bestFitWi
modelCompare
```

Compare models using pValues

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ cyl + hp + wt
## Model 3: mpg ~ cyl + wt
## Model 4: mpg ~ wt
## Model 5: mpg ~ cyl
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     26 151
## 2     27 161 -1      -9.8 1.68 0.2065
## 3     28 183 -1     -22.3 3.84 0.0610 .
## 4     30 278 -2     -95.3 8.20 0.0017 **
## 5     29 301  1     -22.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- As seen above from pValues, model differences between *bestFitWithCylHpWtAm* and *bestFitWitCylHpWt* is not significant (pValue of $0.206 > 0.05$ (95% confidence level), Null hypothesis of models been equal cannot be rejected). Model differences between *bestFitWithCylHpWt* and *bestFitWithCylWt* is also not significant(pValue of $0.06 > 0.05$). However, model differences bestFitWithCylWt and bestFitWithWt are significant (pValue of $0.001733 > 0.05$ and hence Null hypothesis of models been same can be rejected). Also, rValue of model *bestFitWithCylWt* (.82) is much higher than that of *bestFitWithWt* (0.74). Since we prefer a model with least number of predictor variables, we can conclude model *bestFitWithCylWt* (with cylinder and weigth) is the best model for predicting fuel efficiency rather than tramission type.

Quantity mpg differences between automatic and manual cars

Let's add *am* to the best model *bestFitWithCylWt*

```
bestFitWithCylWtAm <- lm(mpg~cyl + wt+am, data=mtcars)
summary(bestFitWithCylWtAm)
```

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

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.490 -1.312 -0.504  1.416  5.776
##
## Coefficients:
##             Estimate Std. Error t value        Pr(>|t|)
## (Intercept)   33.754      2.813   12.00 0.0000000000025 ***
## cyl6          -4.257      1.411   -3.02          0.0055 **
## cyl8          -6.079      1.684   -3.61          0.0012 **
## wt            -3.150      0.908   -3.47          0.0018 **
## am1            0.150      1.300    0.12          0.9089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 27 degrees of freedom
## Multiple R-squared:  0.838,  Adjusted R-squared:  0.813
## F-statistic: 34.8 on 4 and 27 DF,  p-value: 0.000000000273
```

As seen from above, the manual cars get just **0.150** additional miles per gallon when number of cylinders and weight of a car are taken into consideration.

Residual Analysis

```
bestFitWithCylWt$residuals
```

Residual analysis for model *bestFitWithCylWt* using wt and cyl as predictor variables

```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##             -0.3365              0.4809             -3.7538
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##              1.9708             1.8074             -0.5438
##          Duster 360           Merc 240D            Merc 230
##             -2.1759             0.6351             -1.0931
##            Merc 280           Merc 280C          Merc 450SE
##              0.4921            -0.9079              1.5269
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##              1.3370            -0.6027             -0.6905
## Lincoln Continental   Chrysler Imperial            Fiat 128
##             -0.1327             3.9141              5.4616
##         Honda Civic      Toyota Corolla       Toyota Corona
##              1.5863             5.7915             -4.5890
##    Dodge Challenger         AMC Javelin          Camaro Z28
##             -1.1362            -1.7087             -2.3104
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##              3.6056            -0.4879             -1.1308
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##              1.2593            -1.9581             -1.1557
##       Maserati Bora          Volvo 142E
##             -1.4759            -3.6792
```

Toyoto Corolla (5.79), Fiat 128 (5.46), Chrysler Imperial(3.91), Toyoto Corona (-4.58) (see also Figure 2 in the Appendix) are the outliers in the dataset with either very high/low efficieny which will have an effect of using model *bestFitWithCylWt* for prediction.

Conclusion

Using *baseFit* with only transmission type as the predictor variable shows that manual tranmission type cars achieve higher efficiency of **7.2** miles per gallon than automatic cars. However, as shown above, transmission type is a not a good predictor for fuel efficiency, but rather weight and number of cylinders of a car. Taking number of cylinders and weight into consideration, manual cars provide just **0.15** additional miles per gallon than automatic cars.

Appendix

Figure 1 showing a boxplot comparing mpg for automatic and manual transmission cars

```
with(mtcars,{
     boxplot(mpg ~ am,
             ylab = "miles per gallon",
             xlab = "0 - auto, 1 - manual",
             main = "Fig. 1 - mpg for auto and manual tramission cars")
    abline(baseFit, col="red")
     })
```

Fig. 1 − mpg for auto and manual tramission cars

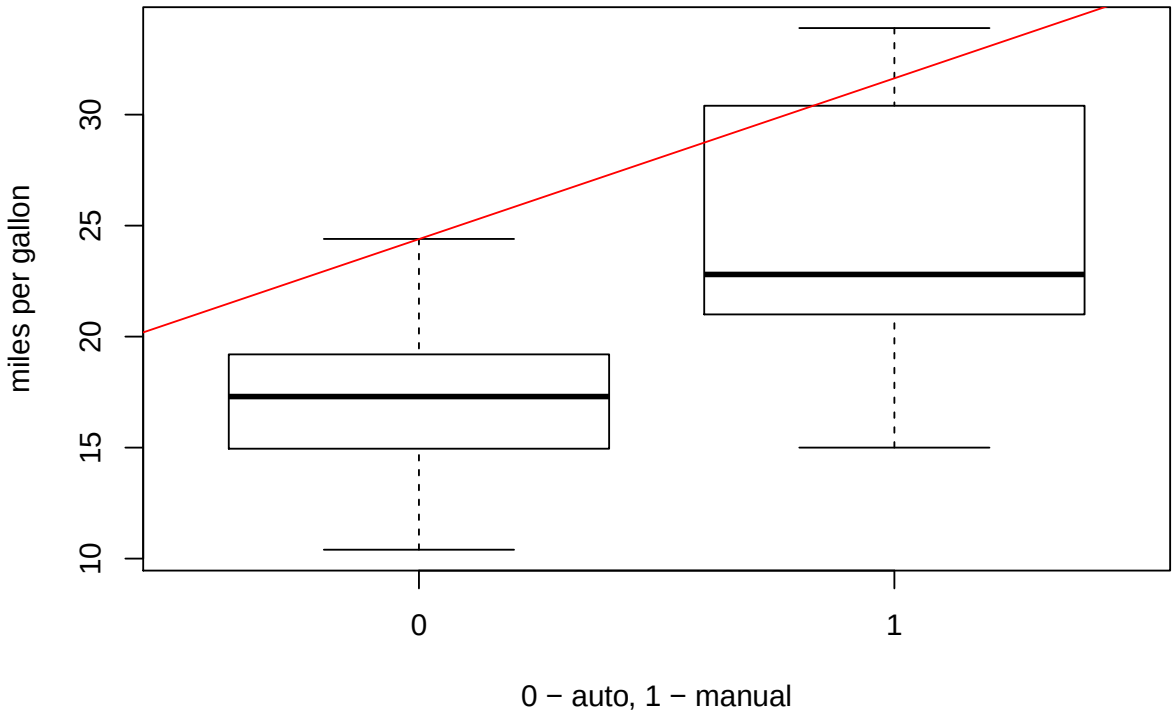
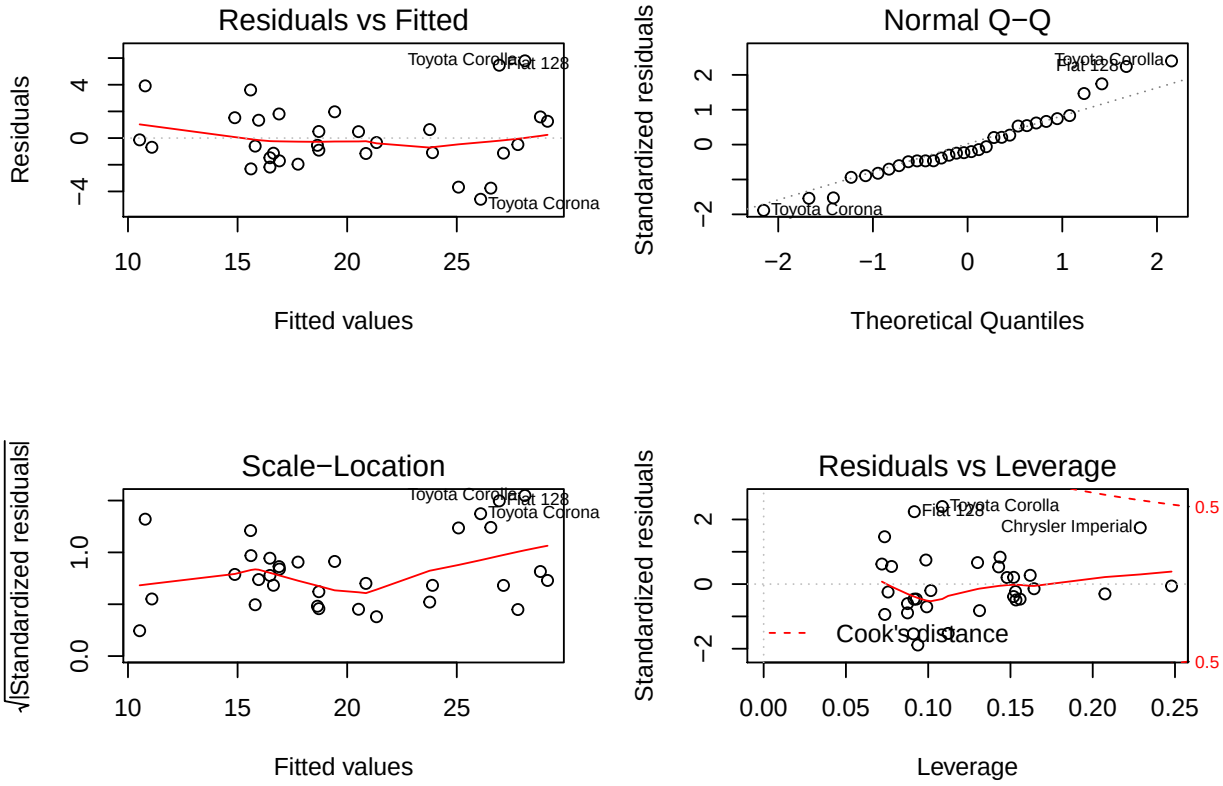


Figure 2 showing diagnostic plots for best model (mpg ~ cyl + wt).

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


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Fiat 128
Toyota Corona
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Chrysler Imperial
Cook's distance