Miles per gallon analysis

Gissur Jónasson

Saturday, May 23, 2015

Exectuvive summary

This report tries to analyzi and explain the miles per gallon usages of cars based on their designs e.g. horsepower, weight, transmission etc. By analyzing the data it can clearly can be seen that mpg are mostly predicted by weight and horsepower of the cars but if we dig litle deeper it seems that automatic vs manual transmission does play a part although a vague one. By doing hypothesis testing both t test to compare mean between auto vs manual transmission and anova test to compare different predict models for mpg we end by saying that the manual transmission plays a part so by shifting gears yourself you can 2.083710 miles per gallon.

Note that: the statment is vague as the dataset is not big enough and to get more comfortable with the result I would conduct research on more cars as this dataset does have to much of a heavy automatic transmission cars relative to the manual cars that are lighter. Github project directory can be found here

Data processing and exploratory analysis

The data set in this project is a built in data set so we only need to do load it

The help file for the data set says (see apendix)

It can be seen from summary of the data and from the help file that few of the variables are more like factor variables e.g. cyl, vs, am, gear, carb so I convert theses field to factors and have that dataset as well for back up when deciding on model.

First off we start by analyzing the mpg variables with the respect to automatic vs manual transmission (see apendix) There are cleary differences in the mpg values on automatic vs manual. Is this difference significant lets make a t-test

```
autoData <- mtcars[mtcars$am == 0,]
manualData <- mtcars[mtcars$am == 1,]
t_test <- t.test(autoData$mpg, manualData$mpg)
t_test[[3]]
```

```
## [1] 0.001373638
```

It can be seen by the p value that it is significan now lets look at the mpg based on other variables (see apendix)

it looks from these pictures that the data is little skewed as with automatic transmission tend to be heavier and with more horse power in the data set there are though few data points that we can base on.

Model building

lets now build or models - first off lets check to see what variables will be significant in the linear regression and correlation as well as variance inflation for the whole data set.

```
rcorr(as.matrix(mtcars))[[1]][1:4,1:6]
```

```
##             mpg        cyl       disp         hp       drat         wt
## mpg   1.0000000 -0.8521619 -0.8475513 -0.7761683  0.6811719 -0.8676594
## cyl  -0.8521619  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958
## disp -0.8475513  0.9020329  1.0000000  0.7909486 -0.7102139  0.8879799
## hp   -0.7761683  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479
```

```r
fitall <- lm(mpg ~ ., data = mtcars)
#summary(fitall)
fitall_clean <- lm(mpg ~ ., data = mtcars_clean)
#summary(fitall_clean)
vif(fitall)[c(1,2,3,5,7,8,9)]
```

```
##       cyl      disp        hp        wt        vs        am      gear
## 15.373833 21.620241  9.832037 15.164887  4.965873  4.648487  5.357452
```

It can be seen by looking at beta coefficents for all variables, the correlation matrix as well as the variance inflation that **wt** and **hp** are the best variables for describing mpg which makes sense intuitively as cars get heavier and with more hoursepower they should have lower mpgs.

if we analyze the numbers above further it can be seen that **hp** is highly corralted with cyl and disp and we dont want collinearity so we dont consider adding these parameters to the model even though they could be good as seen above the variance inflation numbers also support that we will leave these parameters out.

but for the purpose of this project lets first chek how model based on **am** only comes out

```r
fit1 <- lm(mpg ~ am, data = mtcars_clean)
#summary(fit1)
```

This model does not do that good of a job predciting **mpg** as it only has adjusted r-squared as 33,85% so this model does not fit the data that well but based on that we would conclude that manual transmission on average had 7,245 mpg then the automatic ones.

Lets make a model based on **hp** and **wt** as we had found out they are pretty strong predictors. Model summary can be seen in the apendix.

```r
fit2 <- update(fit1,mpg ~ wt + hp , data = mtcars)
#summary(fit2)
fit3 <- update(fit1,mpg ~  wt + hp + am, data = mtcars)
print(anova(fit1,fit2,fit3))
```

```
# Analysis of Variance Table
#
# Model 1: mpg ~ am
# Model 2: mpg ~ wt + hp
# Model 3: mpg ~ wt + hp + am
#   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
# 1     30 720.90
# 2     29 195.05  1    525.85 81.6666 8.567e-10 ***
# 3     28 180.29  1     14.76  2.2918    0.1413
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We conclude that the mpg is mainly based on **wt** and **hp** but the model gets litle bit better by adding **am** as well so we use that as our final model and for the purpose of this project even thoug the anova test implies that the model with **am** is not significantly different from the one without it by the p value = 0.1413

Before we report the details of our model, it is important to check the residuals for any signs of non-normality and examine the residuals vs. fitted values plot to spot for any signs of heteroskedasticity. See appendix (residual analysis)

Our residuals are normally distributed and homoskedastic. the **HP**, **wt**, **am** model explains *82,3%* of the mpg usages

mpg = 34.002875 - 0.037479 * hp - 2.878575 * wt + 2.083710 * am

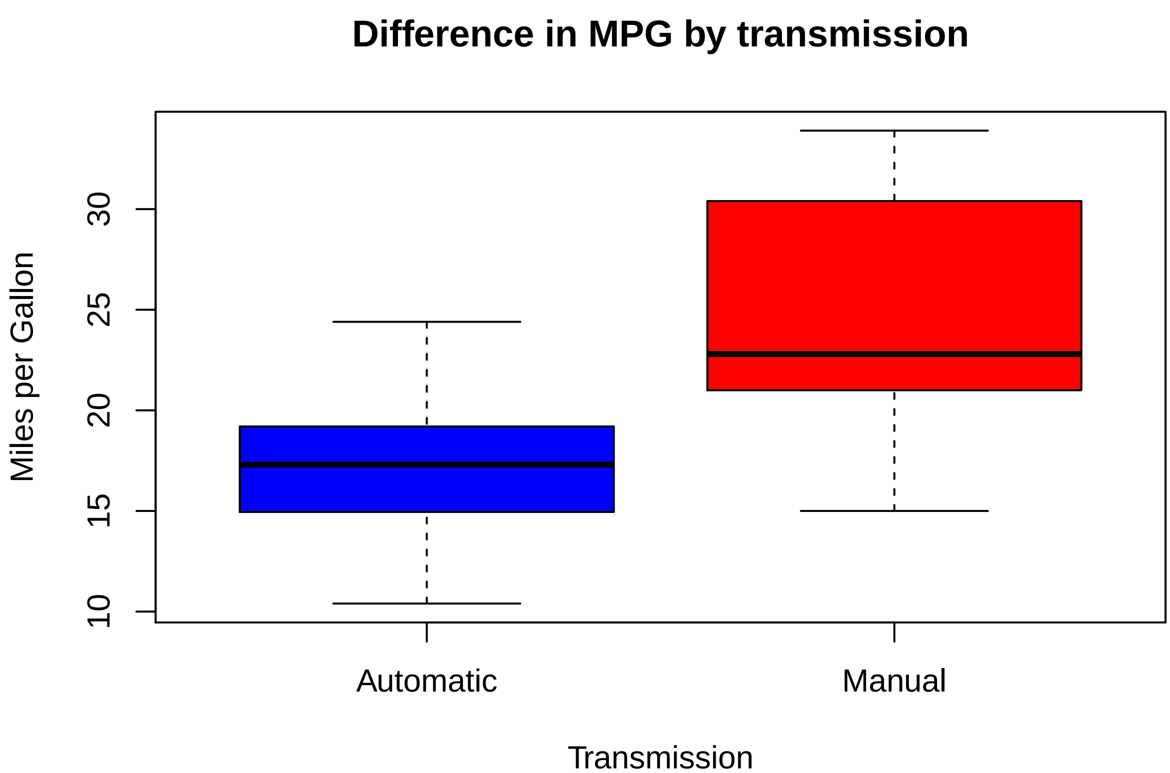
so one might say that manual transmission cars on average have 2.083710 more miles then the automatic ones

Appendix

data explain:

Category	Explanation
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

auto vs manual:

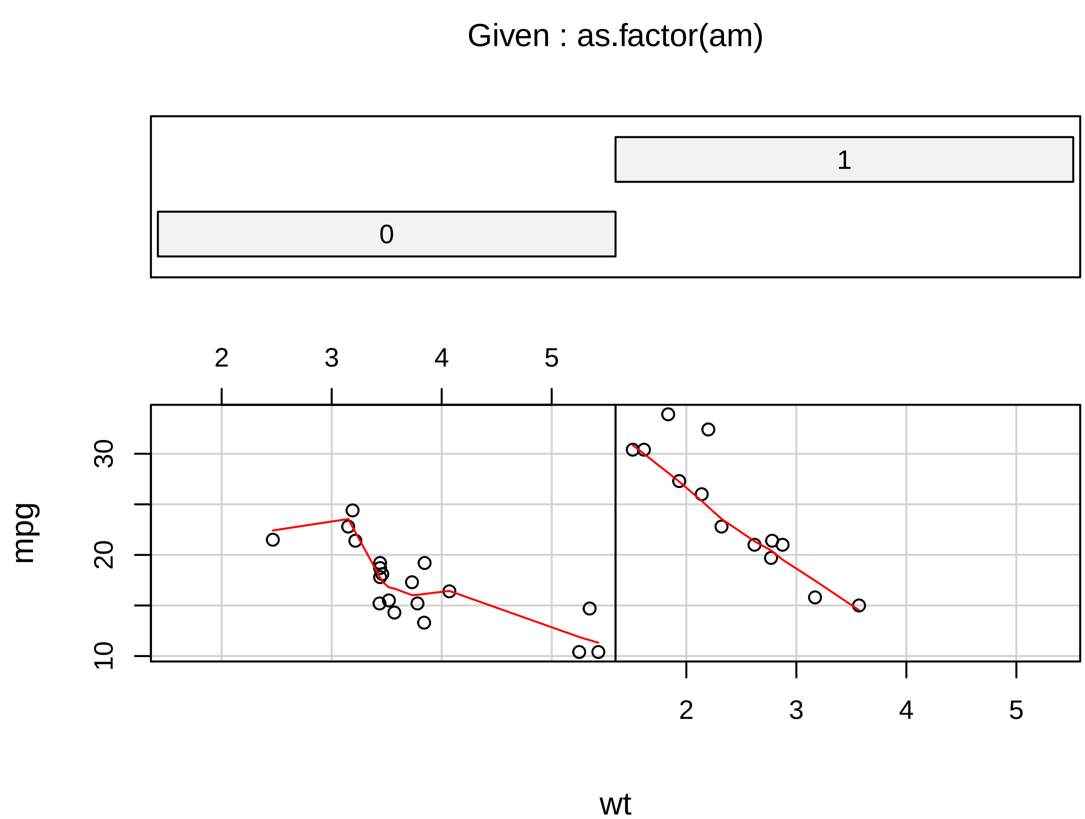


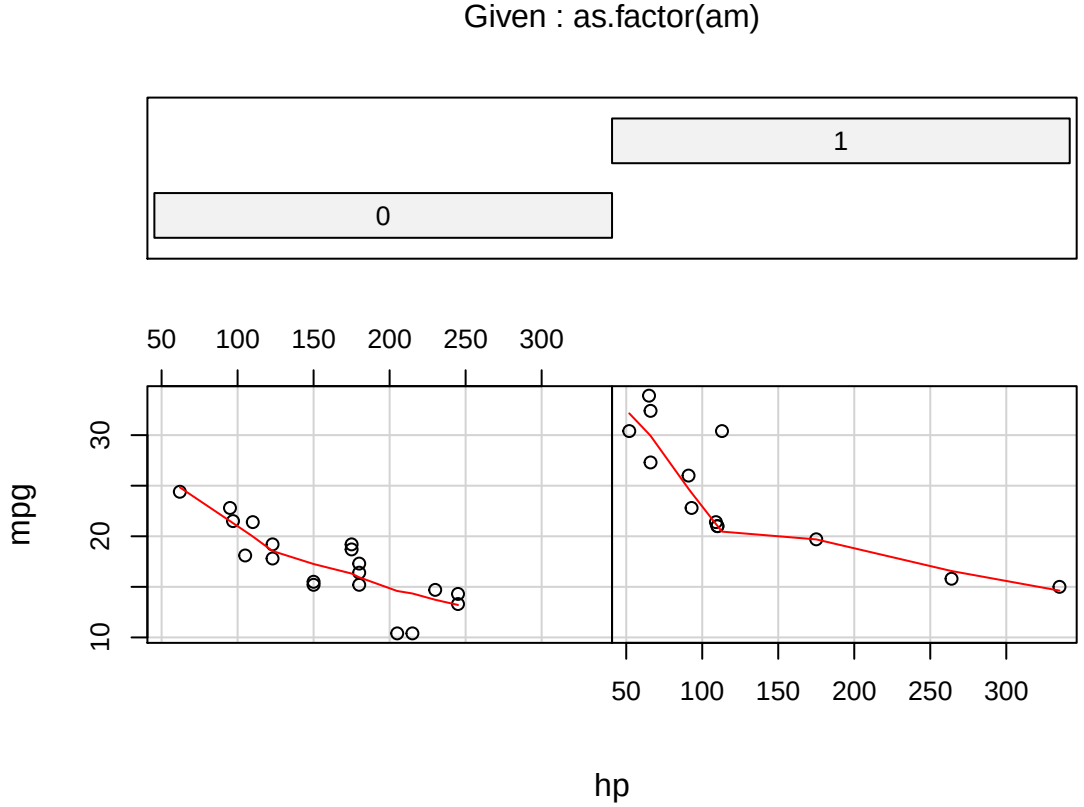
model fitting table

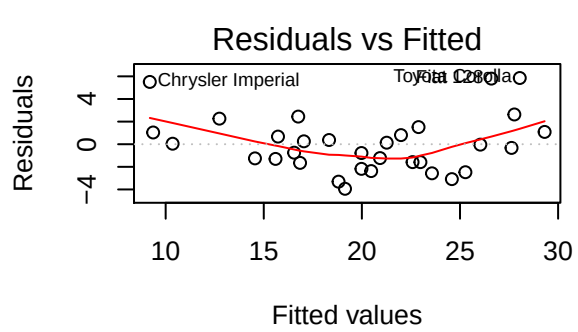
```
##
## Models fitting results
## ==========================================================================================
##                                                  Dependent variable:
##                       -------------------------------------------------------------------
```

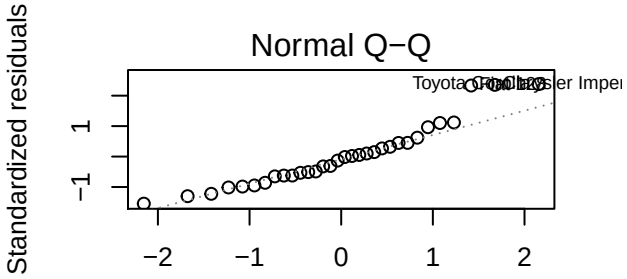
```
##                                                 mpg
##                              (1)                (2)                (3)
## -----------------------------------------------------------------------------------------
## am1                       7.245***
##                           (1.764)
##
## wt                                           -3.878***          -2.879***
##                                               (0.633)            (0.905)
##
## hp                                           -0.032***          -0.037***
##                                               (0.009)            (0.010)
##
## am                                                                2.084
##                                                                  (1.376)
##
## Constant                 17.147***           37.227***          34.003***
##                           (1.125)             (1.599)            (2.643)
##
## -----------------------------------------------------------------------------------------
## Observations                32                  32                 32
## R2                         0.360               0.827              0.840
## Adjusted R2                0.338               0.815              0.823
## Residual Std. Error    4.902 (df = 30)     2.593 (df = 29)    2.538 (df = 28)
## F Statistic         16.860*** (df = 1; 30) 69.211*** (df = 2; 29) 48.960*** (df = 3; 28)
## =========================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

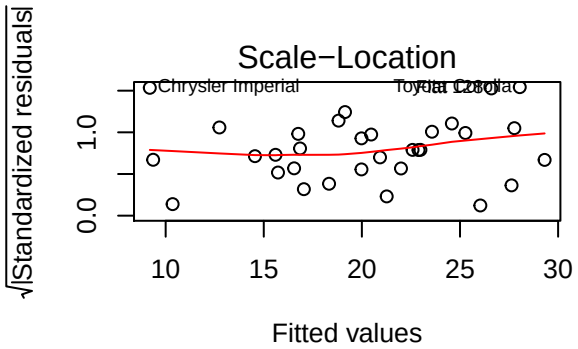
auto vs manual 2:

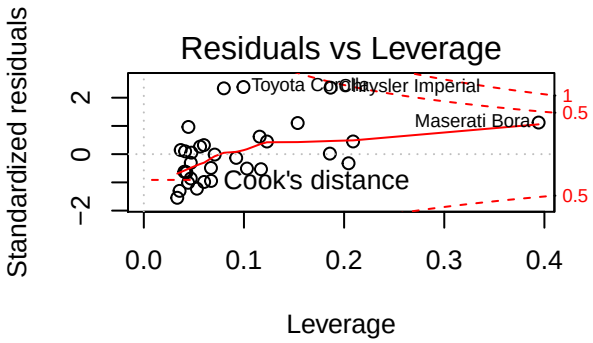












residual analysis: