Miles per gallon analyzis

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.

Problem Statement

To determine whether mpg is better for automatic or manual transmission in the mtcars dataset.

Methodology

First of we start by looking ad what variable seems to have the best estimate of the mpg variable

steps taken are 1. process the data 2. explore the data set with the auto vs manual transaction in mind 3. Build a model 4. find out new model and compare 5. analyze residuals

Data processing

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

```r
mtcars <- mtcars
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

The help file for the data set says

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

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.

```
mtcars_clean <- mtcars
mtcars_clean$am<- as.factor(mtcars_clean$am)
mtcars_clean$vs<- as.factor(mtcars_clean$vs)
mtcars_clean$cyl<- as.factor(mtcars_clean$cyl)
mtcars_clean$gear<- as.factor(mtcars_clean$gear)
mtcars_clean$carb<- as.factor(mtcars_clean$carb)
```

Exploratory data analysis

First off we start by analyzing the mpg variables with the respect to automatic vs manual transmission

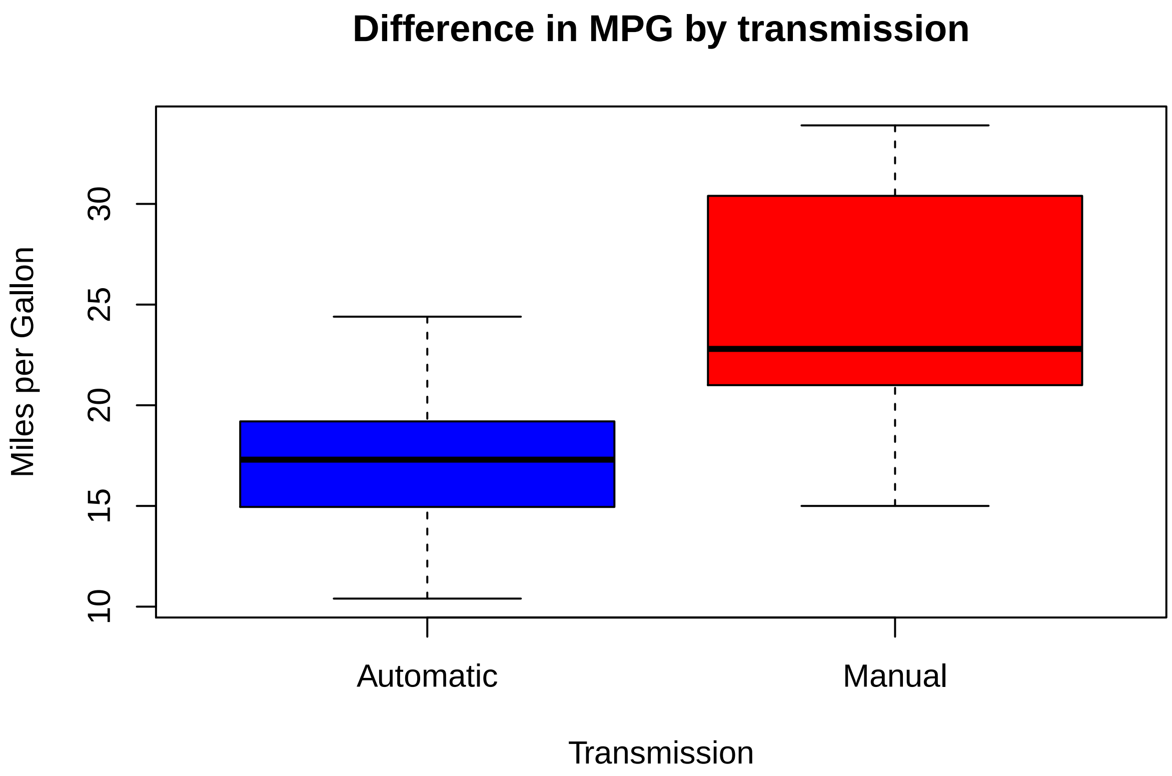
```
attach(mtcars)
am_mean <-aggregate(mtcars, by=list(am),
                    FUN=mean, na.rm=TRUE)
am_mean
```

```
##   Group.1      mpg      cyl     disp       hp     drat       wt     qsec
## 1       0 17.14737 6.947368 290.3789 160.2632 3.286316 3.768895 18.18316
## 2       1 24.39231 5.076923 143.5308 126.8462 4.050000 2.411000 17.36000
##          vs am     gear     carb
## 1 0.3684211  0 3.210526 2.736842
## 2 0.5384615  1 4.384615 2.923077
```

```
am_std <-aggregate(mtcars, by=list(am),
                    FUN=sd, na.rm=TRUE)
am_std
```

```
##   Group.1      mpg      cyl      disp       hp      drat        wt
## 1       0 3.833966 1.544657 110.17165 53.90820 0.3923039 0.7774001
## 2       1 6.166504 1.552500  87.20399 84.06232 0.3640513 0.6169816
##       qsec        vs am      gear     carb
## 1 1.751308 0.4955946  0 0.4188539 1.147079
## 2 1.792359 0.5188745  0 0.5063697 2.177978
```

```
mtcars_clean$am<-factor(mtcars$am,levels=c(0,1), labels = c("Automatic","Manual"))
boxplot(mpg~am, data = mtcars_clean,
                col = c("blue", "red"),
                main = "Difference in MPG by transmission",
                xlab = "Transmission",
                ylab = "Miles per Gallon"
                )
```



There are cleary differenc in the mpg values on automatic vs manual. Is this difference significant lets make a t-test

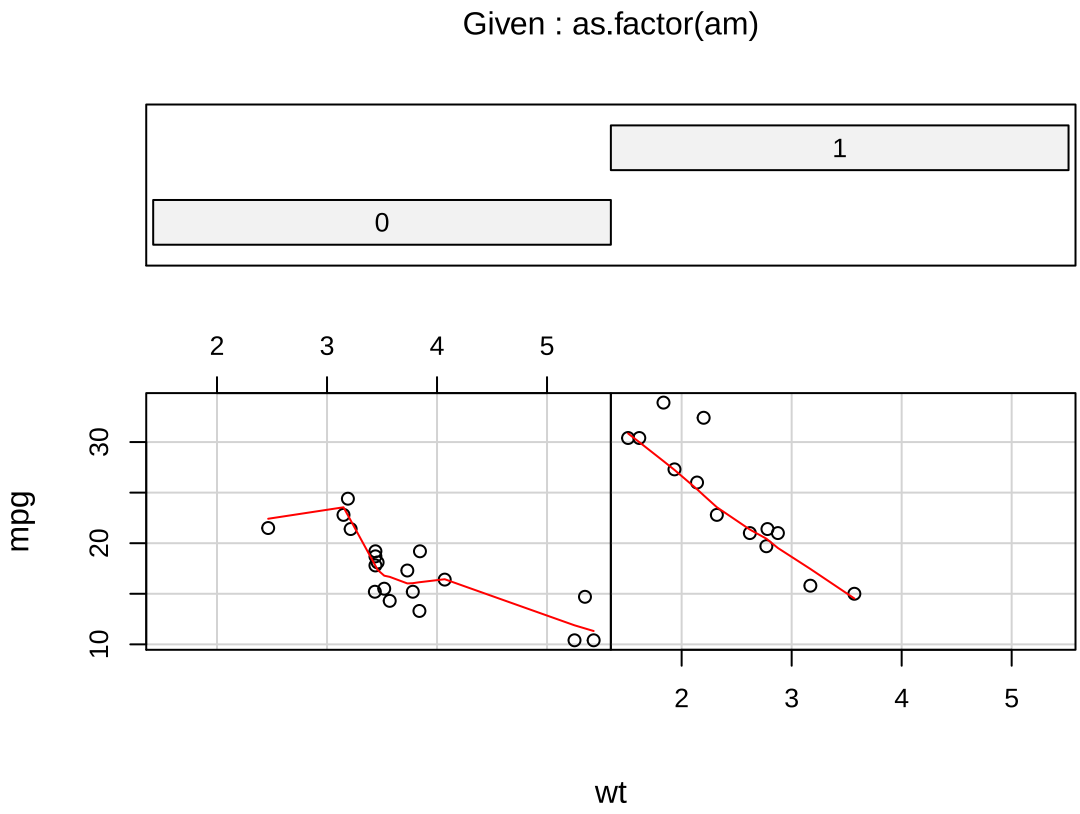
```
autoData <- mtcars[mtcars$am == 0,]
manualData <- mtcars[mtcars$am == 1,]
t.test(autoData$mpg, manualData$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autoData$mpg and manualData$mpg
## 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 of x mean of y
##  17.14737  24.39231
```

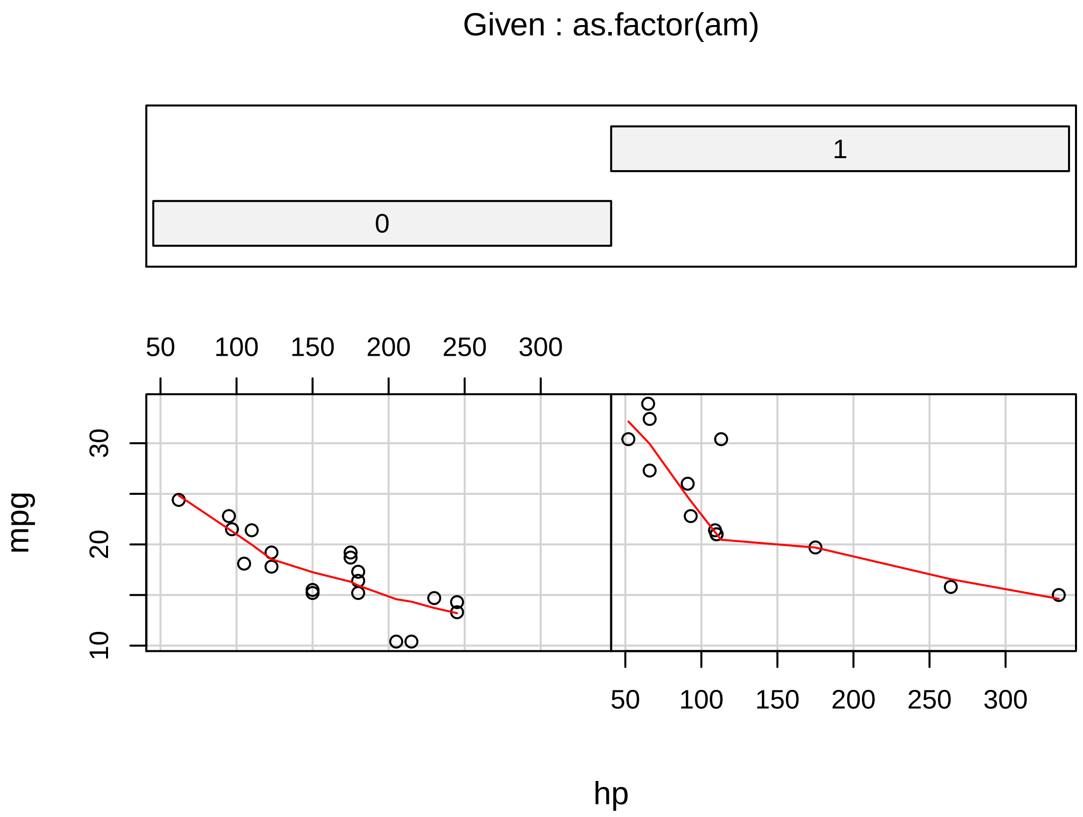
yes it is

Now let look at the mpg based on other variables

```
coplot(mpg ~ wt | as.factor(am), data = mtcars,
     panel = panel.smooth, rows = 1)
```



```
coplot(mpg ~ hp | as.factor(am), data = mtcars,
      panel = panel.smooth, rows = 1)
```



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.

```
#install.packages("Hmisc")
library("Hmisc")
library("car")
correlation<-rcorr(as.matrix(mtcars))
correlation
```

```
##        mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
## mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
## cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
## disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
## hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
## drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
## wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
## qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
## vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
## am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
## gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
## carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
##
## n= 32
##
##
## P
##      mpg    cyl    disp   hp     drat   wt     qsec   vs     am     gear
## mpg         0.0000 0.0000 0.0000 0.0000 0.0000 0.0171 0.0000 0.0003 0.0054
## cyl  0.0000        0.0000 0.0000 0.0000 0.0000 0.0004 0.0000 0.0022 0.0042
## disp 0.0000 0.0000        0.0000 0.0000 0.0000 0.0131 0.0000 0.0004 0.0010
## hp   0.0000 0.0000 0.0000        0.0100 0.0000 0.0000 0.0000 0.1798 0.4930
## drat 0.0000 0.0000 0.0000 0.0100        0.0000 0.6196 0.0117 0.0000 0.0000
## wt   0.0000 0.0000 0.0000 0.0000 0.0000        0.3389 0.0010 0.0000 0.0005
## qsec 0.0171 0.0004 0.0131 0.0000 0.6196 0.3389        0.0000 0.2057 0.2425
## vs   0.0000 0.0000 0.0000 0.0000 0.0117 0.0010 0.0000        0.3570 0.2579
## am   0.0003 0.0022 0.0004 0.1798 0.0000 0.0000 0.2057 0.3570        0.0000
## gear 0.0054 0.0042 0.0010 0.4930 0.0000 0.0005 0.2425 0.2579 0.0000
## carb 0.0011 0.0019 0.0253 0.0000 0.6212 0.0146 0.0000 0.0007 0.7545 0.1290
##      carb
## mpg  0.0011
## cyl  0.0019
## disp 0.0253
## hp   0.0000
## drat 0.6212
## wt   0.0146
## qsec 0.0000
## vs   0.0007
## am   0.7545
## gear 0.1290
## carb
```

```
fitall <- lm(mpg ~ ., data = mtcars)
summary(fitall)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```r
fitall_clean <- lm(mpg ~ ., data = mtcars_clean)
summary(fitall_clean)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

```r
vif(fitall)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

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

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

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

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.

```
fit2 <- update(fit1,mpg ~ wt + hp , data = mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

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

```
##
## Call:
## lm(formula = mpg ~ wt + hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## am           2.083710   1.376420   1.514 0.141268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

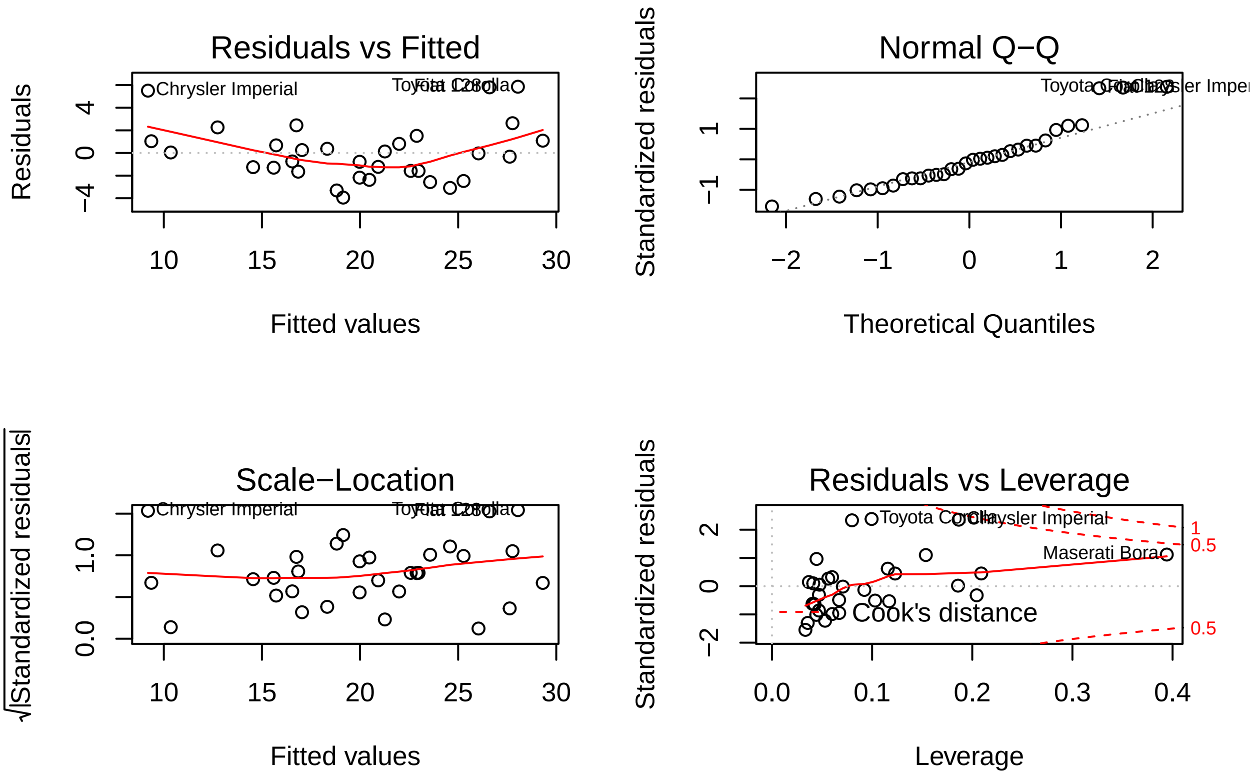
```r
anova(fit1,fit2,fit3)
```

```
## 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.

```
par(mfrow = c(2,2))
plot(fit2)
```



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