Analysis of mtcars dataset in R

Hairizuan Bin Noorazman

Sunday, December 21, 2014

Executive Summary

The mtcars data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

The dataset contains 32 observations with 11 variables

1. mpg : Miles/(US) gallon
2. cyl : Number of cylinders
3. disp : Displacement (cu.in)
4. hp : Gross Horsepower
5. drat : Rear axle ratio
6. wt : Weight (lb/1000)
7. vs : V/S
8. am : Transmission (0 = automatic, 1 = manual)
9. gear : Number of forward gears
10. carb : Number of carburetors

A descriptive/exploratory data analysis is done to understand the structure and nature of the data before creating the appropiate linear models for the data.

Exploratory data analysis

Before starting to model the data, the data is first analyzed with descriptive tools to understand what the data is about. The code for this portion is available in the appendix.

From the quick descriptive analysis, it is possible to roughly say that it might be true that the cars with manual transmission has a higher mpg as compared to cars with automatic transmission. With this in mind, we can go ahead and attempt to model this.

```
## Modification of data
corrdata<-mtcars ## This is to allow working with the PlotCorr function in DescTools
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Modelling of the data

Before modelling the data, we can need to take note of the correleations between the variables that affect the mpg. Some of variables are highly correlated e.g. cyl, disp and hp.

However, the number of variables to select is still a lot, so instead of selecting the variables one at time by reasoning (Require expert knowledge to know what the variables mean), we would utilize a technique known as stepwise regression. Stepwise regression allows one to slowly test the variables and remove them one at a time in order to obtain the best fit for the y-variables and to select the best set of x variables.

```r
# backward elimination
model1 <- lm(mpg ~ ., data = mtcars)
final_model <- step(model1, direction = "backward")
```

```
## Start:  AIC=76.4
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  5   13.5989 134.00 69.828
## - gear  2    3.9729 124.38 73.442
## - am    1    1.1420 121.55 74.705
## - qsec  1    1.2413 121.64 74.732
## - drat  1    1.8208 122.22 74.884
## - cyl   2   10.9314 131.33 75.184
## - vs    1    3.6299 124.03 75.354
## <none>              120.40 76.403
## - disp  1    9.9672 130.37 76.948
## - wt    1   25.5541 145.96 80.562
## - hp    1   25.6715 146.07 80.588
##
## Step:  AIC=69.83
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  2    5.0215 139.02 67.005
## - disp  1    0.9934 135.00 68.064
## - drat  1    1.1854 135.19 68.110
## - vs    1    3.6763 137.68 68.694
## - cyl   2   12.5642 146.57 68.696
## - qsec  1    5.2634 139.26 69.061
## <none>              134.00 69.828
## - am    1   11.9255 145.93 70.556
## - wt    1   19.7963 153.80 72.237
## - hp    1   22.7935 156.79 72.855
##
## Step:  AIC=67
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1    0.9672 139.99 65.227
## - cyl   2   10.4247 149.45 65.319
## - disp  1    1.5483 140.57 65.359
## - vs    1    2.1829 141.21 65.503
## - qsec  1    3.6324 142.66 65.830
## <none>              139.02 67.005
## - am    1   16.5665 155.59 68.608
## - hp    1   18.1768 157.20 68.937
## - wt    1   31.1896 170.21 71.482
##
## Step:  AIC=65.23
## mpg ~ cyl + disp + hp + wt + qsec + vs + am
```

```
##
##        Df Sum of Sq    RSS    AIC
## - disp  1    1.2474 141.24 63.511
## - vs    1    2.3403 142.33 63.757
## - cyl   2   12.3267 152.32 63.927
## - qsec  1    3.1000 143.09 63.928
## <none>              139.99 65.227
## - hp    1   17.7382 157.73 67.044
## - am    1   19.4660 159.46 67.393
## - wt    1   30.7151 170.71 69.574
##
## Step:  AIC=63.51
## mpg ~ cyl + hp + wt + qsec + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - qsec  1     2.442 143.68 62.059
## - vs    1     2.744 143.98 62.126
## - cyl   2    18.580 159.82 63.466
## <none>              141.24 63.511
## - hp    1    18.184 159.42 65.386
## - am    1    18.885 160.12 65.527
## - wt    1    39.645 180.88 69.428
##
## Step:  AIC=62.06
## mpg ~ cyl + hp + wt + vs + am
##
##        Df Sum of Sq    RSS    AIC
## - vs    1     7.346 151.03 61.655
## <none>              143.68 62.059
## - cyl   2    25.284 168.96 63.246
## - am    1    16.443 160.12 63.527
## - hp    1    36.344 180.02 67.275
## - wt    1    41.088 184.77 68.108
##
## Step:  AIC=61.65
## mpg ~ cyl + hp + wt + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              151.03 61.655
## - am    1     9.752 160.78 61.657
## - cyl   2    29.265 180.29 63.323
## - hp    1    31.943 182.97 65.794
## - wt    1    46.173 197.20 68.191
```

```
summary(final_model)
```

```
##
## 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
```

From the stepwise regression technique, we somewhat get mpg being related to cyl and horsepower and weight and am (This is the variable that has to be in there in order to obtain an understanding of it.)

However, at the same time, to ensure that our results are somewhat consistent with our normal manual methods, we would also create other models, all of which is available in the appendix.

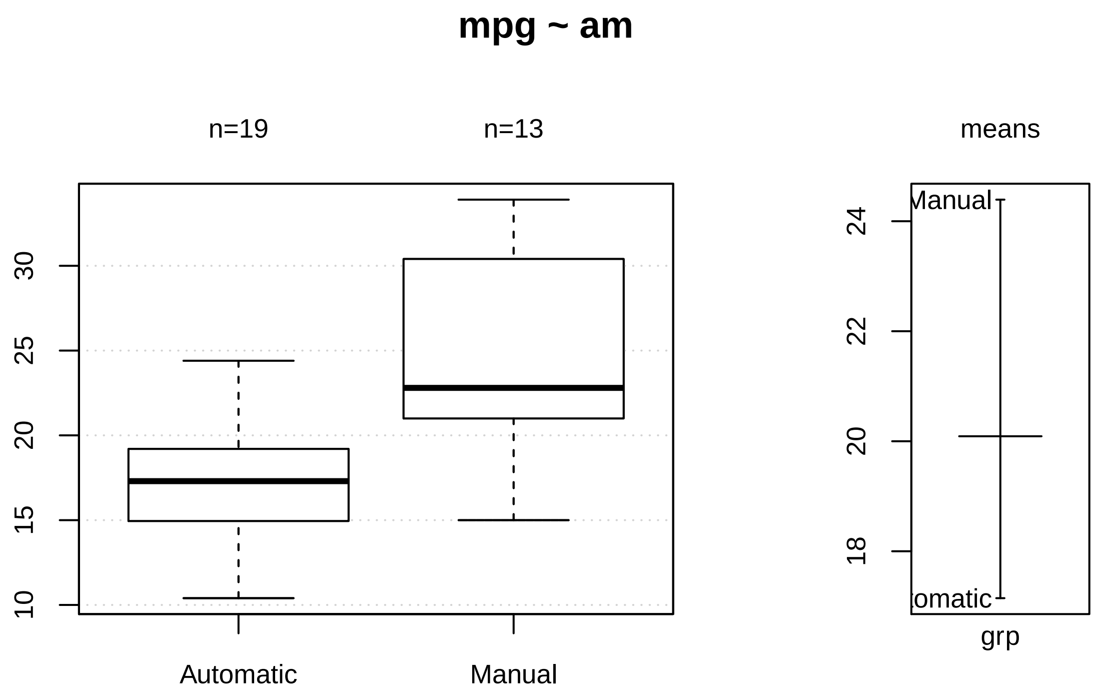
Assuming every other variable is constant, the mpg of a car that is manual can be said to be an additional 1.8 mpg more as compared to an auto car. However, further work on this model is still needed to be done to prove the relationship between mpg and the am variable. This is because of the probability vailue being more than 0.001 which proves that the variable is not that signifiacant. This will not be discussed in this paper.

Therefore, to conclude: 1. Manual transmission cars are better that auto transmission cars. 2. Manual transmission cars can be quantified to have at least 1.8 more mpg to them as compared to automatic transmission cars, keeping all other variables constant.

Appendix

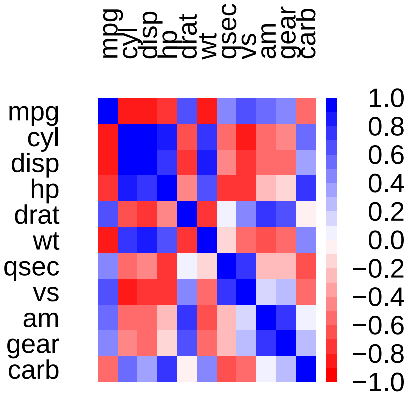
```r
## Understanding the data using summary
summary(mtcars)

## Utilize Desctools to understand the data. Use Desc to quickly create the plots to understand relatio
## Import package if unavailable
library(DescTools)
Desc(mpg ~ am, data=mtcars, plotit=TRUE)
```



```
PlotCorr(cor(corrdata))

## Results not shown in this paper due to lack of space
```



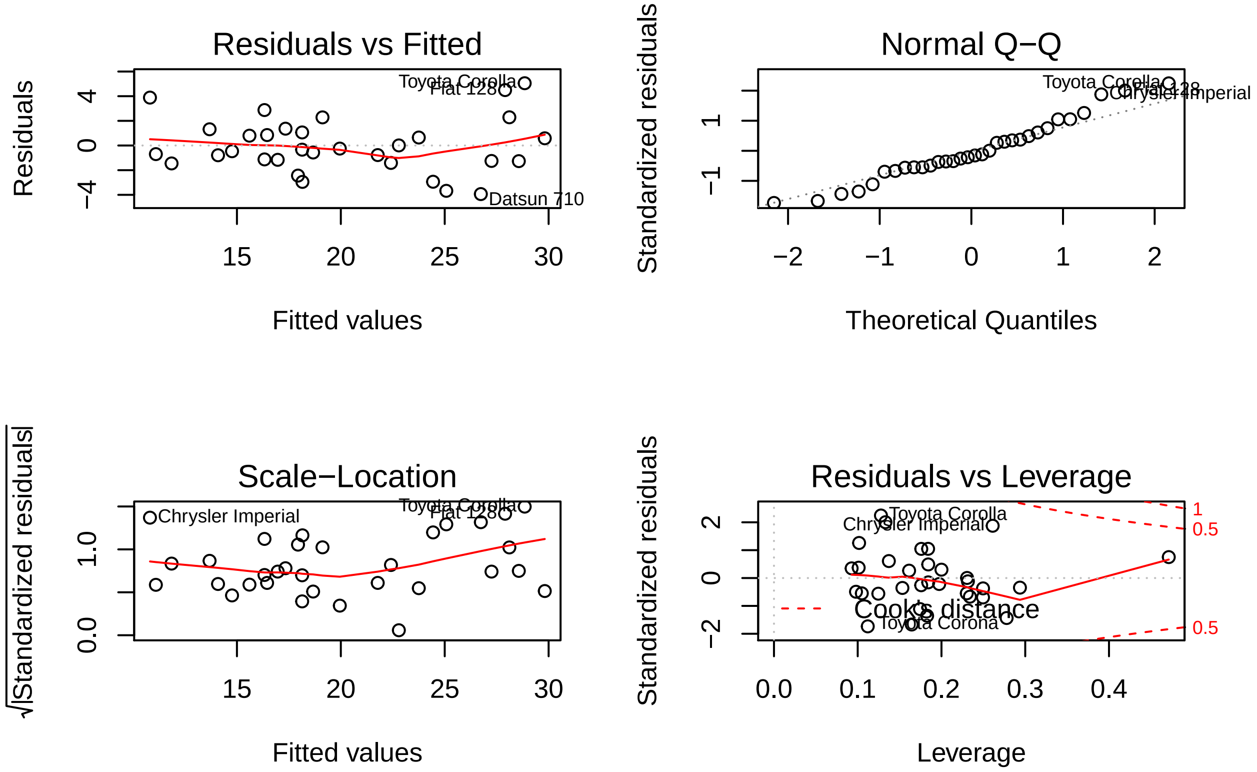
Additional models

```
fit1 <- lm(mpg ~ cyl, data = mtcars)
fit2 <- lm(mpg ~ wt, data = mtcars)
fit3 <- lm(mpg ~ cyl + wt, data = mtcars)
fit4 <- lm(mpg ~ disp, data = mtcars)
fit5 <- lm(mpg ~ disp + cyl, data = mtcars)
fit6 <- lm(mpg ~ disp + cyl + wt + am, data = mtcars)
fit7 <- lm(mpg ~ cyl + hp + wt + am, data = mtcars) ## This was given by the stepwise regression
```

As expected by the stepwise regression analysis, most of the variance is covered by the final linear fit. This is done by observing the Adjusted R-squared value and seeing whether it is able to hit close to 1 for that value.

Residual plots

```
par(mfrow = c(2, 2))
plot(fit7)
```


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