Mtcars Data Regression analysis

Executive Summary

In this report, we present an analysis of the relationship between a set of variables and **mpg**. The data we have used was extracted from 1974 Motor Trend US magazine. This report is primarily focussed on two questions:

- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

Exploratory Data Analysis

The graphs for all the exploratory work is present in the **Appendix** section. For getting a preliminary knowledge regarding the dataset, we load the dataset into R.

```r
library(datasets); data(mtcars)
```

We check for any possible correlation between transmission type **am** and other variables and dependent variable **mpg** as shown in Fig.1. We also take a look at the correlation between transmission type and **mpg**. The value 0.5998324 indicates a positive correlation between **mpg** and **am**.

```r
cor(mtcars$am, mtcars$mpg)
```

```
## [1] 0.5998324
```

Is an automatic or manual transmission better for MPG?

Since we have checked for correlation between variables and is difficult to verify which is better we perform a box plot analysis as shown in Fig.2 and conclude that manual is better for mpg than automatic.

Regression Model

Single Variable

Since we have verified the positive correlation between **am** and **mpg** we consider the first model with only these two variables, $Y_i = \beta_0 + X_i\beta_1 + \epsilon_i$.

Coefficient Interpretation: β_1 is the group mean for transmission, β_0 is the intercept and ϵ is the residual.

```r
fitSV <- lm(mtcars$mpg ~ mtcars$am); summary(fitSV)$coef; summary(fitSV)$adj
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## mtcars$am    7.244939   1.764422  4.106127 2.850207e-04
```

```
## [1] 0.3384589
```

The above value correspond to the R^2, which means our model only explains 33.8% of the variance.

Multi-Variable

To find a better fit we now use the variance inflation factor and correlation to determine the the variables for the model along with **am**. After checking the correlations with **mpg** we consider **cyl**, **disp**, **hp**, **wt** along with **am** as the variables for the model, $Y_i = \beta_0 + X_{i_1}\beta_1 + X_{i_2}\beta_2 + X_{i_3}\beta_3 + X_{i_4}\beta_4 + X_{i_5}\beta_5 + \epsilon_i$.

Coefficient Interpretation: β_1, β_2, β_3, β_4, β_5 are the group mean for transmission, number of cylinders, displacement, horsepower, weight respectively. β_0 is the intercept and ϵ is the residual.

```
fitMV <- lm(mpg ~ am+cyl+disp+hp+wt, data = mtcars); summary(fitMV)$coef; summary(fitMV)$adj
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 38.20279869 3.66909647 10.412045 9.084987e-11
## am           1.55649163 1.44053603  1.080495 2.898430e-01
## cyl         -1.10637984 0.67635506 -1.635797 1.139322e-01
## disp         0.01225708 0.01170645  1.047036 3.047194e-01
## hp          -0.02796002 0.01392172 -2.008374 5.509659e-02
## wt          -3.30262301 1.13364263 -2.913284 7.256888e-03
```

```
## [1] 0.8272816
```

The above value correspond to the R^2, which means our model explains 82.7% the variance. We are thus more likely to accept this as our model. The residual plots are presented in Fig.3 and the **residual summary** is given below:

```
summary(fitMV$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -3.5950 -1.5860 -0.7157  0.0000  1.2820  5.5720
```

The analysis of variance between the models is shown below:

```
anova(fitMV, fitSV)
```

```
## Warning in anova.lmlist(object, ...): models with response '"mtcars$mpg"'
## removed because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15 64.5778 1.626e-08 ***
## cyl        1 449.53  449.53 71.6522 6.037e-09 ***
## disp       1  19.28   19.28  3.0732  0.091376 .
## hp         1  35.71   35.71  5.6925  0.024609 *
## wt         1  53.25   53.25  8.4872  0.007257 **
## Residuals 26 163.12    6.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe that in the multi-variable model, the residuals are normally distributed and homoskedastic. Thus we conclude our report with the statement that it is better for **mpg** to have manual transmission and cars on an average have 1.55 **mpgs** more in case of **manual transmission** than **automatic transmission**. The entire summary of the accepted model i.e **multi-variable regression** is shown in the appendix.

Appendix

```r
require(graphics)
pairs(mtcars, main = "Correlation", panel=panel.smooth); title(sub = "Fig.1")
```

Correlation

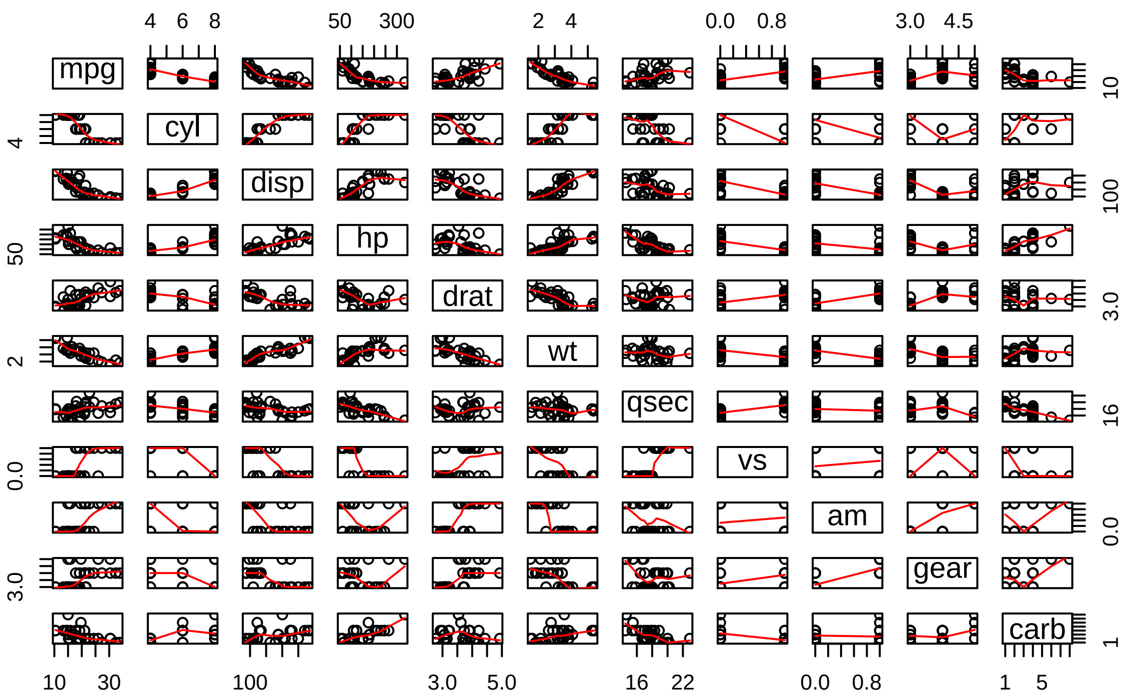


Fig.1

```r
with(mtcars, {plot(as.factor(am), mpg, main="MPG by transmissions", xlab="am", ylab="mpg")}); title(sub
```

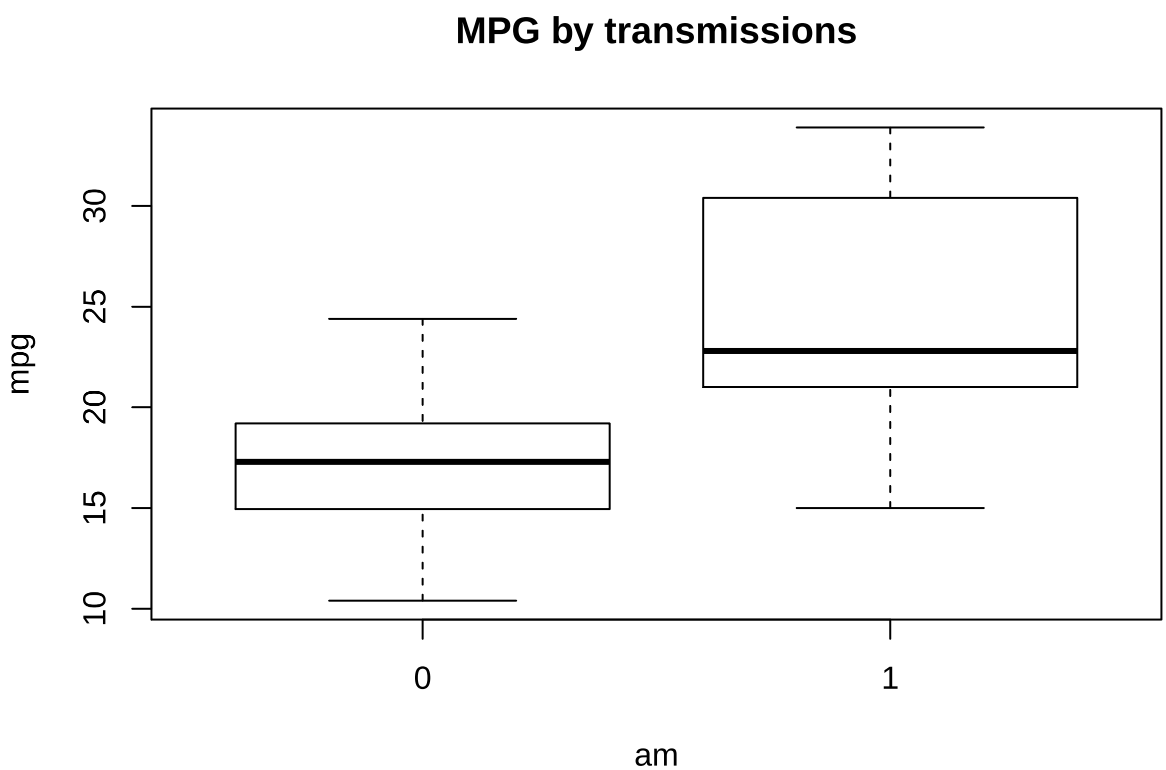
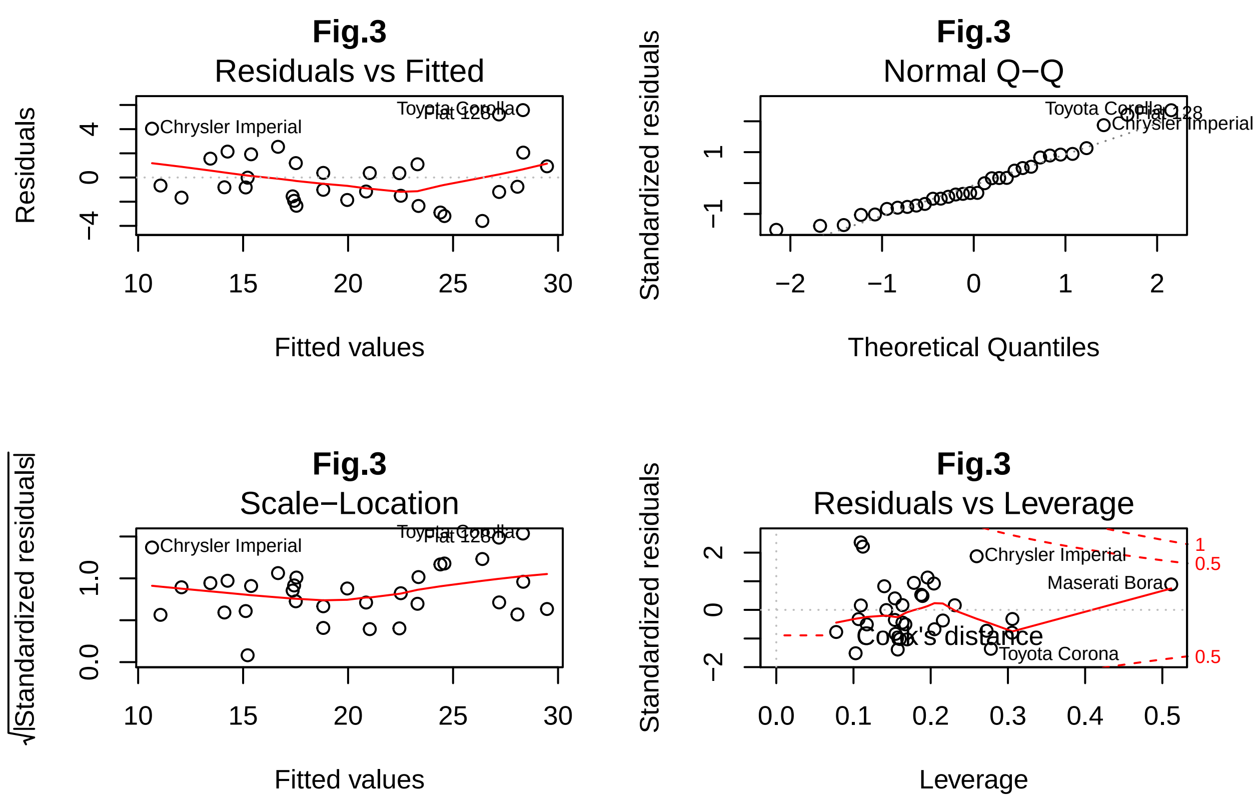


Fig.2

```r
par(mfrow = c(2, 2)); plot(fitMV, main = "Fig.3")
```



```r
summary(fitMV)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + disp + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am           1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```