Regression models:Project-Cars dataset analysis

RP

7 June 2017

Executive summary

In this project, we analyse mtcars dataset to find if 1."Automatic or manual transmission is better for mpg" 2."Quantify MPG difference between automatic and manual transmissions". First of all,We load dataset and do exploratory analysis.Then we use hypothesis testing and linear regression to analyse the data.We do both single and multivariate linear regression modelling but find multivariable regression analysis to fit the model better.

Exploratory Data analysis

```
data(mtcars)
dim(mtcars)  ##dimensions of dataset
```

```
## [1] 32 11
```

Structure ,head and Summary of dataset is available in Table 1 ,2 & 3 respectively.

Data Processing

1.Is an automatic or manual transmission better for mpg

We plot mpg vs transmission for the dataset as shown in Plot 2 in Appendix.It is clear from the plot that for transmission type "manual" mpg is more than auto type. We find means for auto and manual transmission type groups (Table 3-Appendix) and see mean for manual transmission type(24.4) is more than auto type(17.1) We conduct t-test for above two groups (Table 4 & 5-Appendix) and see that there is a significant difference in two groups as p-value is .06.

Checking correlations of different variables in mtcars dataset

```
p <- cor(mtcars)
p [1,]
```

```
##        mpg        cyl       disp         hp       drat         wt 
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594 
##       qsec         vs         am       gear       carb 
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

```
fit0<-lm(mpg~.,data=mtcars)
#print(paste(("VIF values"),sqrt(vif(fit))))
fit1<-lm(mpg~am,data=mtcars)
```

In the above linear regression model(Table5-Appendix), manual transmission cars get 7.245 mpg more than automatic transmission cars.R-squared value shows 36% of predicted variables are explained using "am" variable.So, we will explore multivariable regression also.In addition to am variable(default), we see that variables wt, cyl,disp,hp are highly corelated with mpg.We look at corelation among variables(as shown in appendix-Table 6), we see cyl and disp are highly correlated with each other, so we leave disp and we include wt and hp and cyl variables in our model also.

Multivariable linear regression analysis

```
fit2<-lm(mpg~am+wt+hp,data=mtcars)##multivariable regression
fit3<-lm(mpg~am+wt+hp+cyl,data=mtcars)
anova(fit1,fit2,fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt + hp + cyl
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 42.9310 4.112e-09 ***
## 3     27 170.00  1     10.29  1.6348    0.2119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results with p-value of 4.1e-09 show the model is significantly different than linear fit1 model.Also, it r-squared value (as shown in Table 7 in appendix) is 0.84,which explains 84% of variance.We will cross check with residuals for any signs of non-normality and examine residuals vs fitted values plot to check hetroskedasicity.

On checking the plots(Plot 4-Appendix), we find plots are normally distributed and not hetroskedastic.We report estimates of this model.

Conclusion

```
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, 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 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## 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
```

The above model explains 84% of variance .Thus difference between automatic and manual transmissions is 2.08mpg

Appendix

Table 1-Dataset structure

```r
str(mtcars)  ##structure of mtcars dataset
```

```
## '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 ...
```

Table 2 -Dataset glimpse

```r
head(mtcars)  ## a glimpse of the mtcars dataset
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```r
##Histogram of miles per gallon of dataset
```

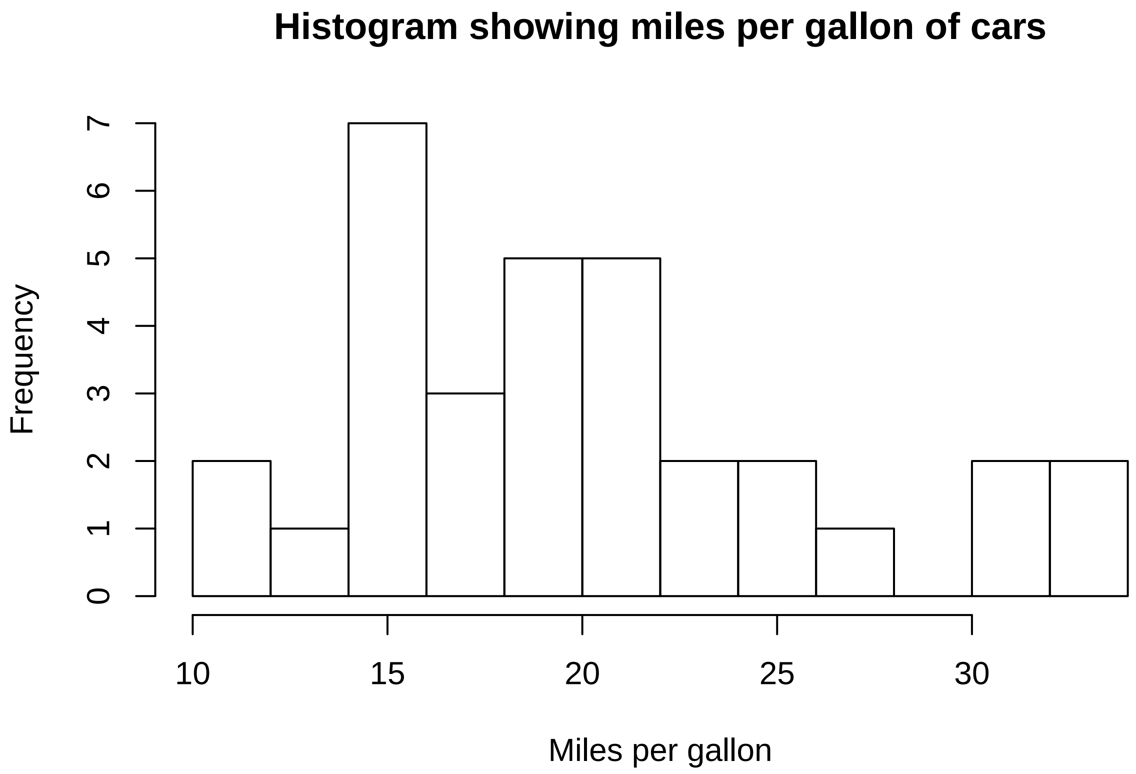
Table 3 -Dataset summary

```r
summary(mtcars)  ##summary of dataset
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Plot 1 -histogram -mpg

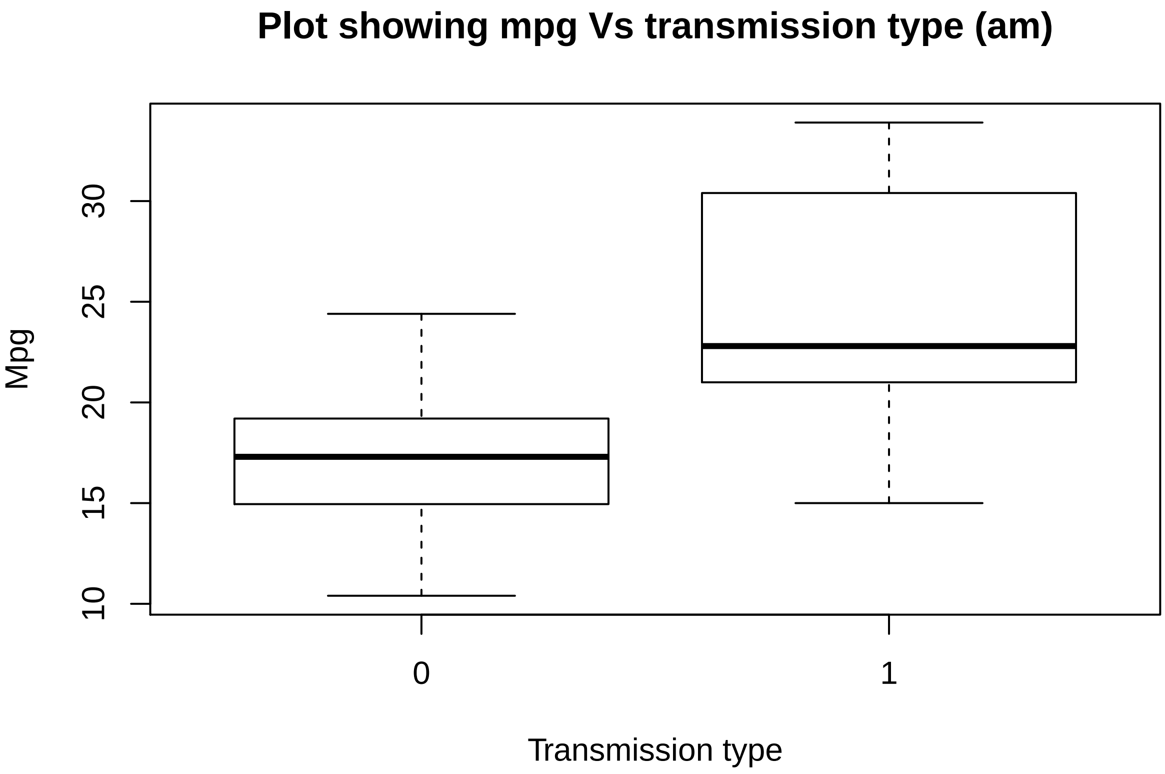
```
hist(mtcars$mpg,breaks=10,xlab="Miles per gallon",main="Histogram showing miles per gallon of cars")
```



Distribution seems to be normal.

Plot 2 -Plotting mpg Vs transmission type (am)

```
plot(mpg~as.factor(am),data=mtcars,xlab="Transmission type",ylab="Mpg",main="Plot showing mpg Vs transmi
```



It appears transmission type “1”(Manual) gives better mpg

Table 3 -Mean of manual vs auto

```
aggregate(mpg~am,data=mtcars,mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

Table 4-t-test for manual vs auto

```
#modelling with single variable (am)
data_auto<-mtcars[mtcars$am==0,]
data_manual<-mtcars[mtcars$am==1,]
t.test(data_auto,data_manual)
```

```
##
##  Welch Two Sample t-test
##
## data:  data_auto and data_manual
## t = 1.8772, df = 348.4, p-value = 0.06132
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7538182 32.3497623
## sample estimates:
## mean of x mean of y 
##  46.02645  30.22848
```

Table 5- T-test coefficients summary

```
f1<-lm(mpg~as.factor(am),data=mtcars)
##summary of t-test coefficients
print("Table 5")
```

```
## [1] "Table 5"
```

```
summary(f1) ## a look at the coefficients
```

```
## 
## Call:
## lm(formula = mpg ~ as.factor(am), data = mtcars)
## 
## 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 ***
## as.factor(am)1    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
```

Table 6-Correlation among variables

```
print("Corelation of variables with each other")
```

```
## [1] "Corelation of variables with each other"
```

```
p[2,]
```

```
##        mpg        cyl       disp         hp       drat         wt 
## -0.8521620  1.0000000  0.9020329  0.8324475 -0.6999381  0.7824958 
##       qsec         vs         am       gear       carb 
## -0.5912421 -0.8108118 -0.5226070 -0.4926866  0.5269883
```

```
p[4,]
```

```
##        mpg        cyl       disp         hp       drat         wt 
## -0.7761684  0.8324475  0.7909486  1.0000000 -0.4487591  0.6587479 
##       qsec         vs         am       gear       carb 
## -0.7082234 -0.7230967 -0.2432043 -0.1257043  0.7498125
```

```
p[6,]
```

```
##        mpg        cyl       disp         hp       drat         wt 
## -0.8676594  0.7824958  0.8879799  0.6587479 -0.7124406  1.0000000 
##       qsec         vs         am       gear       carb 
## -0.1747159 -0.5549157 -0.6924953 -0.5832870  0.4276059
```

Plot 3 -Residual plot

```
par(mfrow=c(2,2))
plot(fit1)
```

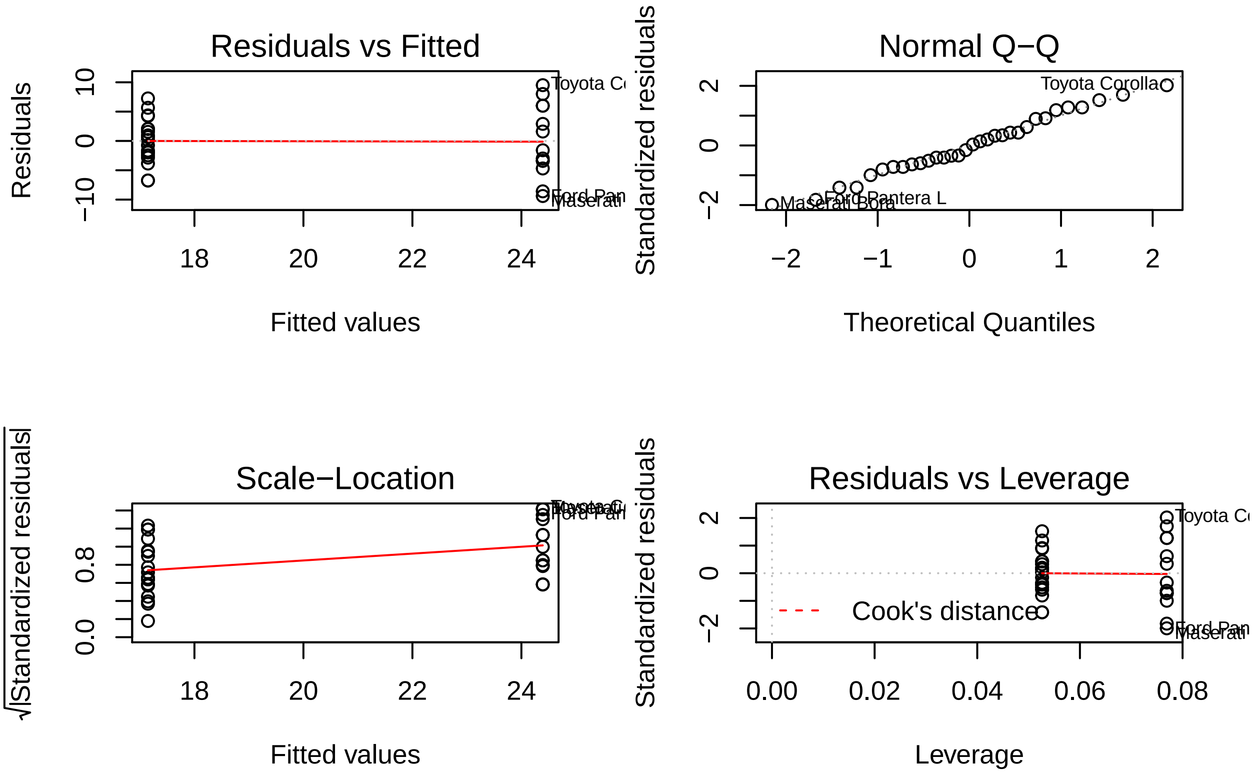


Table 7-Multivariable regression summary

```
summary(fit2)
```

```
## 
## Call:
## lm(formula = mpg ~ am + wt + hp, 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 ***
## am           2.083710   1.376420   1.514 0.141268    
## wt          -2.878575   0.904971  -3.181 0.003574 ** 
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## 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
```