Regression-Models-Project

Abhijeet

Synopsis

The following analysis is done on the “mtcars” data in the R dataset package and addresses the following questions:

1. “Is an automatic or manual transmission better for MPG”
2. “Quantify the MPG difference between automatic and manual transmissions”

- Description: The 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).

Data Processing

```r
#libraries
library(datasets);library(ggplot2);require(stats); require(graphics)
data(mtcars)
```

Exploratory Data Analysis

```r
?mtcars
```

```
## starting httpd help server ... done
```

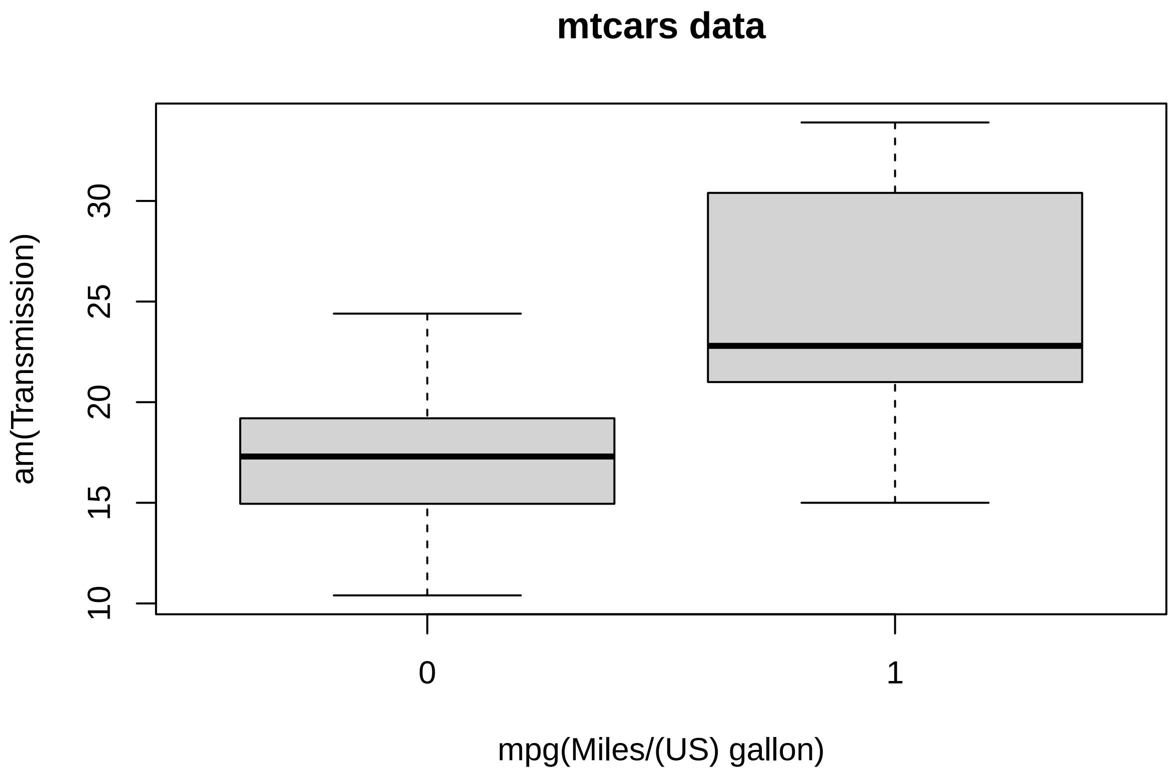
```r
summary(mtcars)
```

```
##       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  
```

The data frame contains 32 observations on 11 (numeric) 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 (1000 lbs)
7. qsec 1/4 mile time
8. vs Engine (0 = V-shaped, 1 = straight)
9. am Transmission (0 = automatic, 1 = manual)
10. gear Number of forward gears
11. carb Number of carburetors

```
boxplot(mtcars$mpg~mtcars$am,main="mtcars data",xlab="mpg(Miles/(US) gallon)",ylab="am(Transmission)")
```



Checking Normality

```
shapiro.test(mtcars$mpg)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mtcars$mpg
## W = 0.94756, p-value = 0.1229
```

- The P-value of the Shapiro-Wilk normality test must be greater than 0.05 for assuming normality. Hence, the mpg(Miles/(US) gallon) data is normally distributed.

```
t.test(mtcars$mpg~mtcars$am, paired = FALSE, var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

- From the above test it is ensured that null Hypothesis(p-value < Significance level) can be rejected at 5% significance level. The mpg(miles per gallon) is higher for manual transmission and there is a significant difference between the auto and manual transmission.

Regression model

logistic regression

First, lets check the dependency of miles per gallon(mpg) and transmission(am), particularly looking to answer the following question:

"Is an automatic or manual transmission better for MPG?"

In this test the outcome is binary or categorical variable 0 and 1.Thus use logistic regression.

```
fit<-glm(am~mpg,data=mtcars, family ="binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = am ~ mpg, family = "binomial", data = mtcars)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6035     2.3514  -2.808  0.00498 **
## mpg           0.3070     0.1148   2.673  0.00751 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.230  on 31  degrees of freedom
## Residual deviance: 29.675  on 30  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 5
```

```
exp(fit$coefficients)
```

```
## (Intercept)         mpg
## 0.001355579 1.359379288
```

AIC which similar to R^2 tells the significance of test but opposite of R^2 the lesser the value of AIC better fit to the data. In this test the AIC is 33.675.There is a 36% probability of a transmission to be manual for every additional mile per gallon. To improve AIC, trying multivariate regression.

multivariate regression

The best fit model suggested that the transmission, weight and 1/4 mile time/ performance measure of acceleration are the best fir variables.

```
fit_1<-lm(formula = mpg ~ wt + qsec + am, data = mtcars)
summary(fit_1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

```
summary(fit_1)$r.squared
```

```
## [1] 0.8496636
```

Conclusion

In logistic regression it can observed that there is a 36% probability of a transmission to be manual for every additional mile per gallon. While it is difficult to interpret in multivariate regression because at significance level alpha=0.05, miles per gallon and transmission are significantly influenced by other factors(weight and acceleration).

Appendix

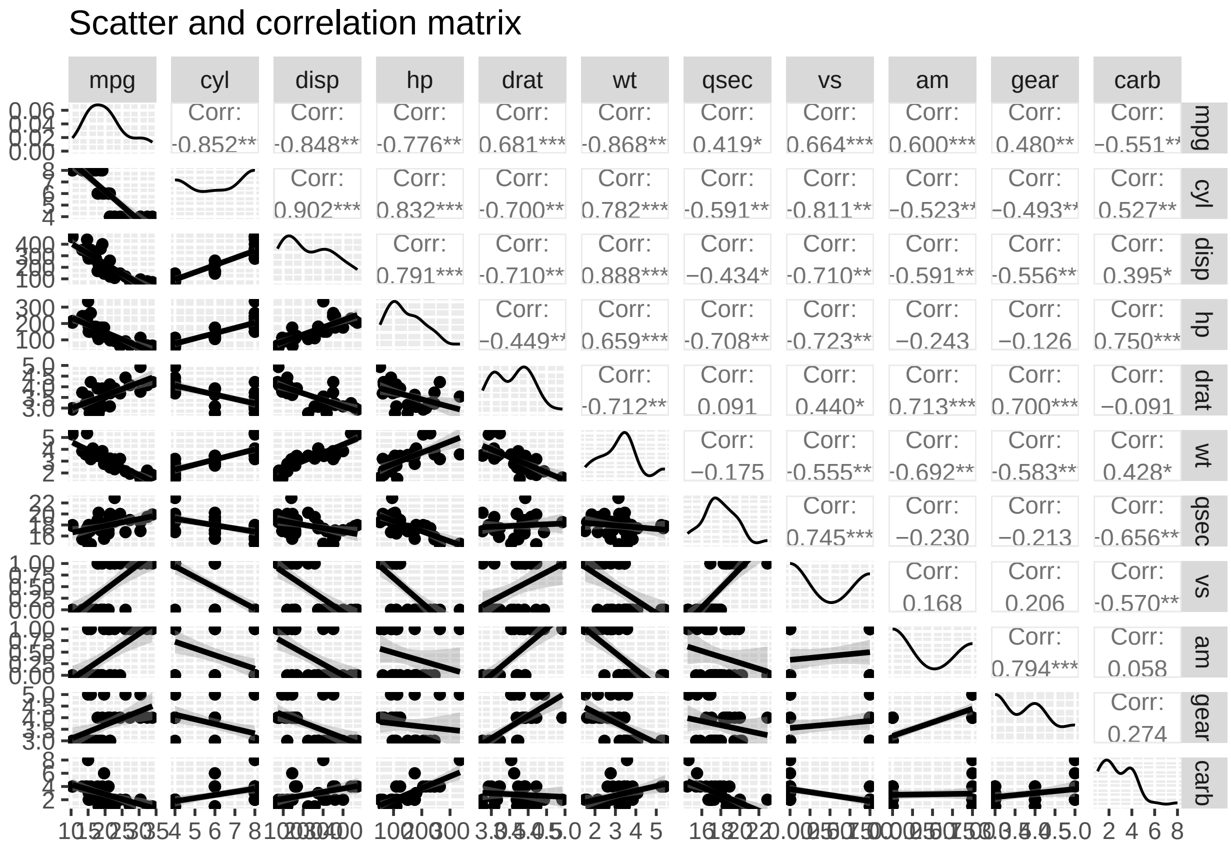
Appendix A -Data Visualization

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

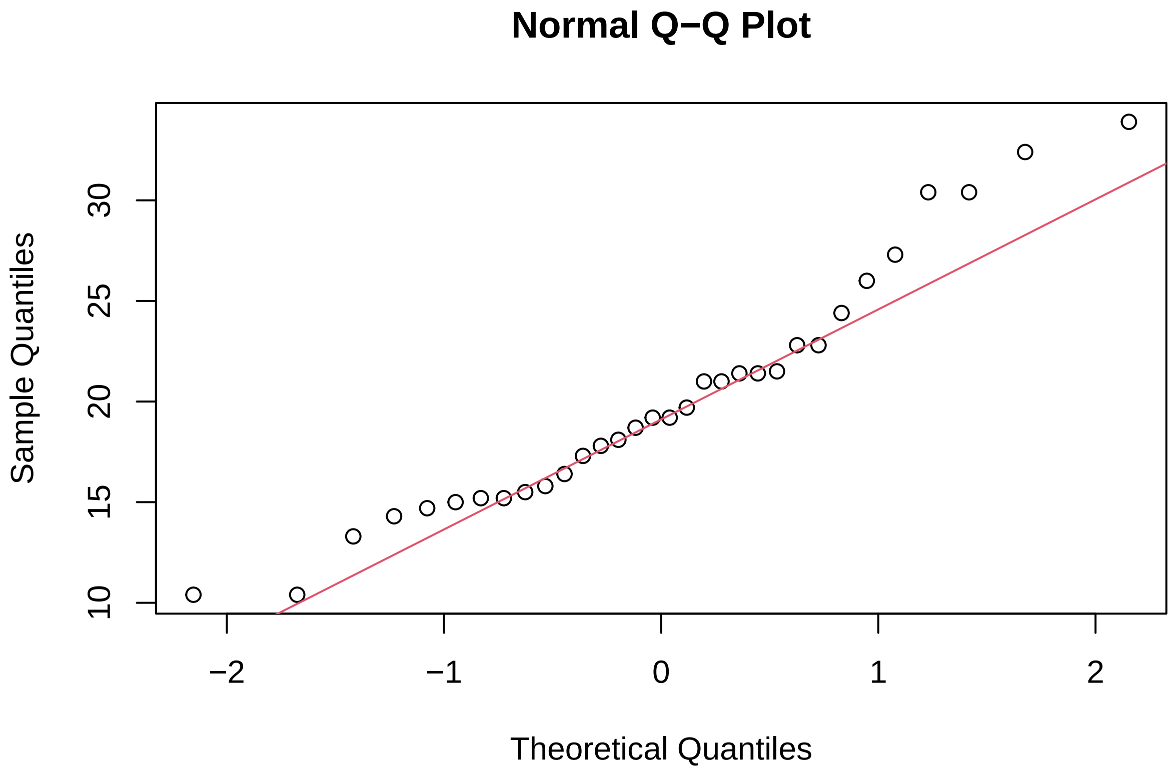
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
options(repr.plot.width = 30, repr.plot.height = 30)
ggpairs(mtcars,  title = "Scatter and correlation matrix",upper = list(continuous = wrap("cor",size = 3
    lower = list(continuous = wrap("smooth")))
```



Normality visualization

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg,col=2)
```



Appendix B - Model Selection

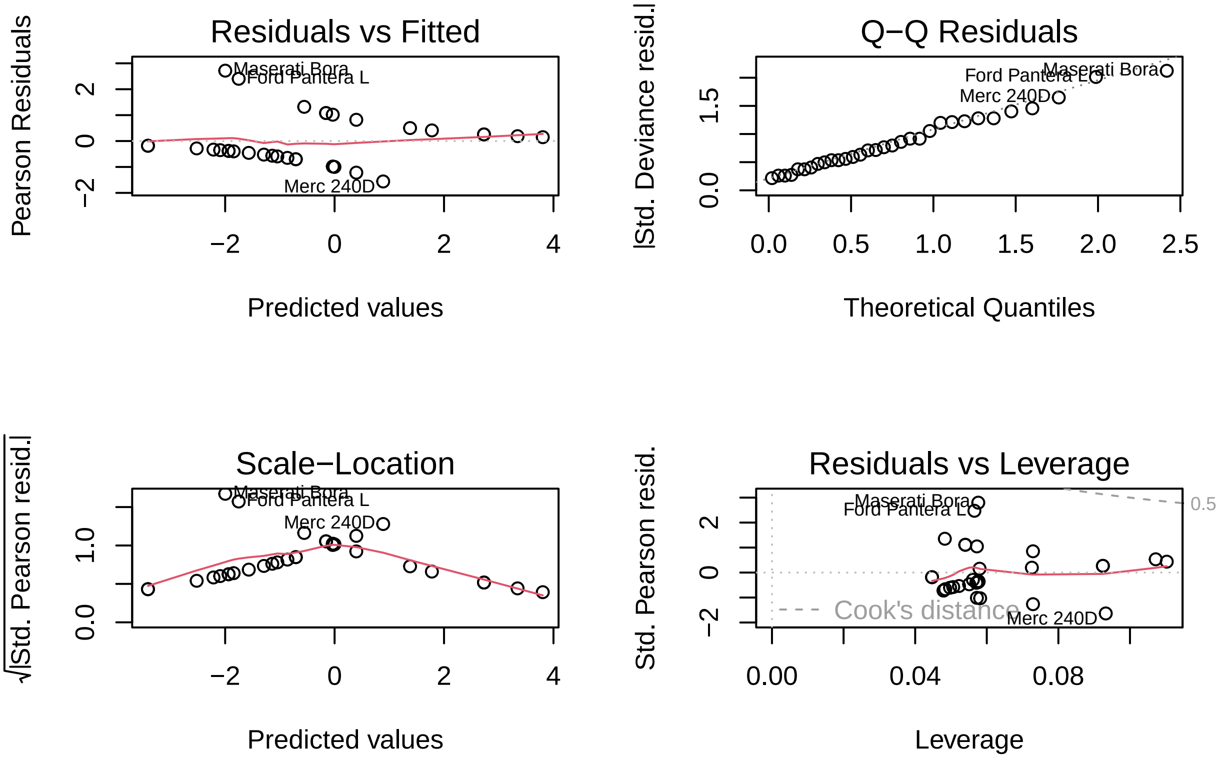
```
step(lm(mpg~.,data=mtcars),direction = "both",trace=0)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Coefficients:
## (Intercept)           wt         qsec           am
##       9.618       -3.917        1.226        2.936
```

Appendix C logistic regression Plots

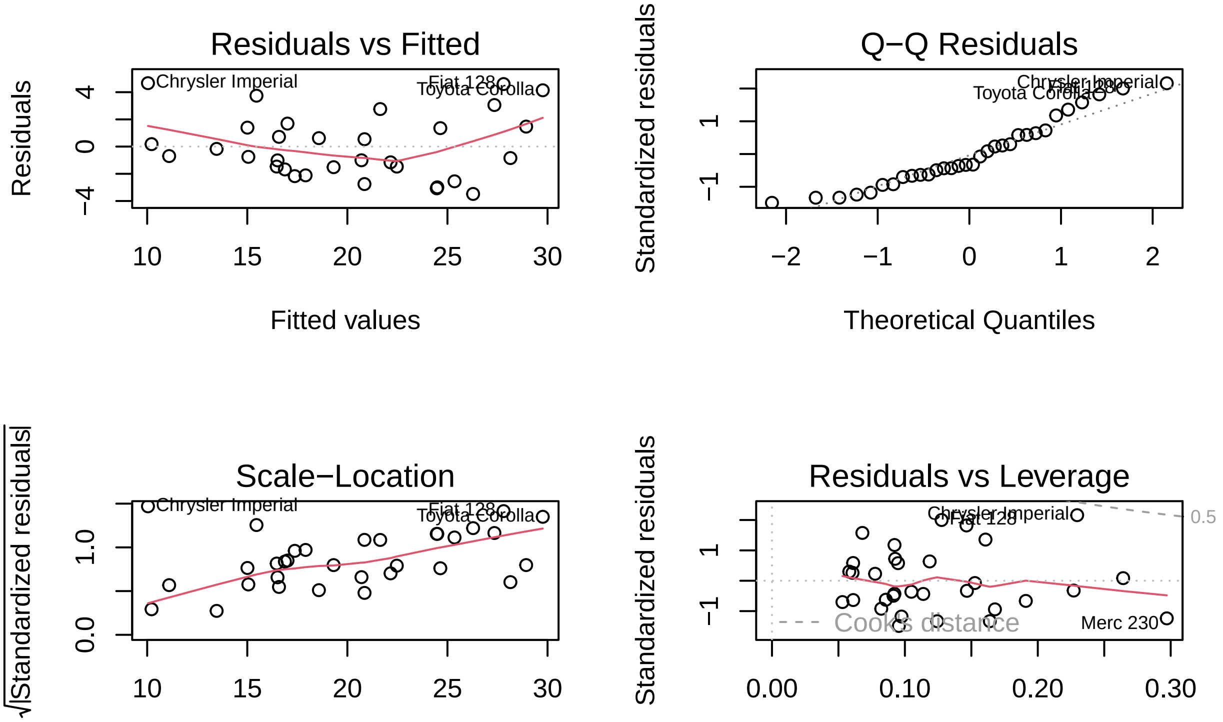
Plot of residuals for my multivariate regression.

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



Appendix D Multivariate Regression

```
par(mfrow = c(2, 2))
plot(fit_1)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
4
0
−4
10
15
20
25
30
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corolla
1
−1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
1.0
0.0
10
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Fiat 128
Merc 230
Cook's distance
0.5
1
−1
0.00
0.10
0.20
0.30