Regression Models Course Project

Axel Espinosa

7 de marzo de 2018

Executive Summary

In this project we will explore the relationship between a set of variables and Miles per Gallon (MPG). We are particularly interested in these two questions

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

For this study we will use `mtcars` R dataset

```
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
```

First we check the dataset, this dataset was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.

We have the next variables set

1. mpg: Miles/(US) gallon
2. cyl: Number of cylinders

3. disp: Displacement
4. hp: Gross horsepower
5. drat: Rear axle ratio
6. wt: weight (1000 lb)
7. qsec: 1/4 mile time
8. vs: V/S
9. am: Transmission (0 = automatic, 1 = manual)
10. gear: Number of forward gears
11. carb: Number of carburetors

Then we will do our exploratory analysis which consists in proof that there are differences on the means of cars that have manual or automatic transmission.

```
data <- mtcars

for(i in 1:11)
{
    if(i==9)
    {
        data$am <- factor(data$am)
    }else{
        data[,i] <- as.numeric(data[,i])
    }
}
```

We have done the last for cycle to convert the transmission in a factor so we can do more complex analysis on the data.

So for Example if we make a t.test we get the following results.

```
dataTest <- t.test(mpg ~ am, data = data)
dataTest

##
##  Welch Two Sample t-test
##
## data:  mpg by am
## 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 in group 0 mean in group 1
##        17.14737        24.39231
```

We see that the test rejects the null hypothesis, by its p-value and because its mean estimates have a difference of 7.245

```
abs(dataTest$estimate[1]-dataTest$estimate[2])

## mean in group 0
##        7.244939
```

We can watch this from a not t.test side calculating the mean of the mpg with the following criteria

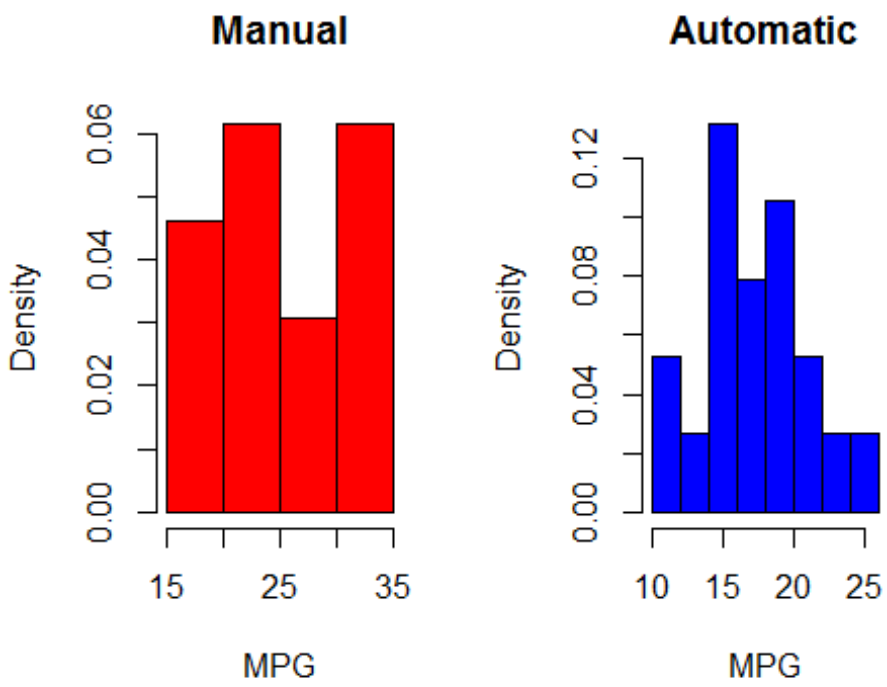
```
mean(as.numeric(data$mpg[data$am==1]))

## [1] 24.39231

mean(as.numeric(data$mpg[data$am==0]))

## [1] 17.14737

par(mfrow = c(1,2))
hist(as.numeric(data$mpg[data$am==1]), freq=FALSE, col="red",
main="Manual", xlab="MPG")
hist(as.numeric(data$mpg[data$am==0]), freq=FALSE, col="blue",
main="Automatic", xlab="MPG")
```



Regression

First we are fitting the linear model between the outcome MPG and the predictors (all of them)

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

##
## Call:
## lm(formula = mpg ~ ., data = data)
##
## 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
## am1          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

max(summary(fit)$coefficients[,4])

## [1] 0.9160874
```

We can see that there is no significance in the model, so the first thing we should consider is to ommit some variables, so we check which have a greater p-value, in this first case we see that the less significant predictor is 'cyl', so if we remove this variable from the model and we remodel we get the following result.

```
fit1 <- lm(mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb,
data = data)
summary(fit1)

##
## Call:
```

```
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + vs + am + gear +
##     carb, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4286 -1.5908 -0.0412  1.2120  4.5961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.96007   13.53030   0.810   0.4266
## disp         0.01283    0.01682   0.763   0.4538
## hp          -0.02191    0.02091  -1.048   0.3062
## drat         0.83520    1.53625   0.544   0.5921
## wt          -3.69251    1.83954  -2.007   0.0572 .
## qsec         0.84244    0.68678   1.227   0.2329
## vs           0.38975    1.94800   0.200   0.8433
## am1          2.57743    1.94035   1.328   0.1977
## gear         0.71155    1.36562   0.521   0.6075
## carb        -0.21958    0.78856  -0.278   0.7833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 22 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8153
## F-statistic: 16.21 on 9 and 22 DF,  p-value: 9.031e-08
```

```
max(summary(fit1)$coefficients[,4])
```

```
## [1] 0.8432585
```

Fortunately for us R has a function that delete those variables and we have no need to go manually until the result, this function is called step, that selects the model using the Akaike's info criterion that works as follows.

```
st_fit1 <- step(fit1)
```

```
## Start:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
```

```
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
```

```
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

if we want to know if we would get at the same model deleting the variables by its p-value we can achieve that using

```
summary(lm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb, data =
data))

##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear +
##     carb, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.356 -1.576 -0.149  1.218  4.604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.76828   11.89230   0.821   0.4199
## disp         0.01214    0.01612   0.753   0.4590
## hp          -0.02095    0.01993  -1.051   0.3040
## drat         0.87510    1.49113   0.587   0.5630
## wt          -3.71151    1.79834  -2.064   0.0505 .
## qsec         0.91083    0.58312   1.562   0.1319
## am1          2.52390    1.88128   1.342   0.1928
## gear         0.75984    1.31577   0.577   0.5692
## carb        -0.24796    0.75933  -0.327   0.7470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 23 degrees of freedom
## Multiple R-squared:  0.8687, Adjusted R-squared:  0.823
## F-statistic: 19.02 on 8 and 23 DF,  p-value: 2.008e-08
```

```r
summary(lm(mpg ~ disp + hp + drat + wt + qsec + am + gear, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1200 -1.7753 -0.1446  1.0903  4.7172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.19763   11.54220   0.797  0.43334
## disp         0.01552    0.01214   1.278  0.21342
## hp          -0.02471    0.01596  -1.548  0.13476
## drat         0.81023    1.45007   0.559  0.58151
## wt          -4.13065    1.23593  -3.342  0.00272 **
## qsec         1.00979    0.48883   2.066  0.04981 *
## am1          2.58980    1.83528   1.411  0.17104
## gear         0.60644    1.20596   0.503  0.61964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 24 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8296
## F-statistic: 22.56 on 7 and 24 DF,  p-value: 4.218e-09
```

```r
summary(lm(mpg ~ disp + hp + drat + wt + qsec + am, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2669 -1.6148 -0.2585  1.1220  4.5564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.71062   10.97539   0.976  0.33848
## disp         0.01310    0.01098   1.193  0.24405
## hp          -0.02180    0.01465  -1.488  0.14938
## drat         1.02065    1.36748   0.746  0.46240
## wt          -4.04454    1.20558  -3.355  0.00254 **
## qsec         0.99073    0.48002   2.064  0.04955 *
## am1          2.98469    1.63382   1.827  0.07969 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 25 degrees of freedom
```

```
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8347
## F-statistic: 27.09 on 6 and 25 DF,  p-value: 8.637e-10

#At this point we see that the model begins to win significance
summary(lm(mpg ~ disp + hp + wt + qsec + am, data = data))

##
## Call:
## lm(formula = mpg ~ disp + hp + wt + qsec + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## am1          3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10

summary(lm(mpg ~ hp + wt + qsec + am, data = data))

##
## Call:
## lm(formula = mpg ~ hp + wt + qsec + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4975 -1.5902 -0.1122  1.1795  4.5404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.44019    9.31887   1.871  0.07215 .
## hp          -0.01765    0.01415  -1.247  0.22309
## wt          -3.23810    0.88990  -3.639  0.00114 **
## qsec         0.81060    0.43887   1.847  0.07573 .
## am1          2.92550    1.39715   2.094  0.04579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.435 on 27 degrees of freedom
```

```
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8368
## F-statistic: 40.74 on 4 and 27 DF,  p-value: 4.589e-11
```

```
summary(fit1_model <- lm(mpg ~ wt + qsec + am + am*wt, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + am * wt, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am1           14.079      3.435   4.099 0.000341 ***
## wt:am1        -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

Thus we get the same model as if we could use the step function Now what we are trying to do is to assign a new variable to the model, in this case we are considering the interaction between weight and transmission

```
summary(fit_model <- lm(mpg ~ wt + qsec + am, data = data))
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am1           2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

As we can see this model is significant to because the extra variable which we are considering is a transformation between the weight and the transmission Comparing both models with an ANOVA test with a 0.05 as a type I error significance benchmark we get.

```
anova(fit_model,fit1_model)$'Pr(>F)'[2]

## [1] 0.001808576
```

Because 0.001 < 0.05 we can confirm that statistic significance exist in our model

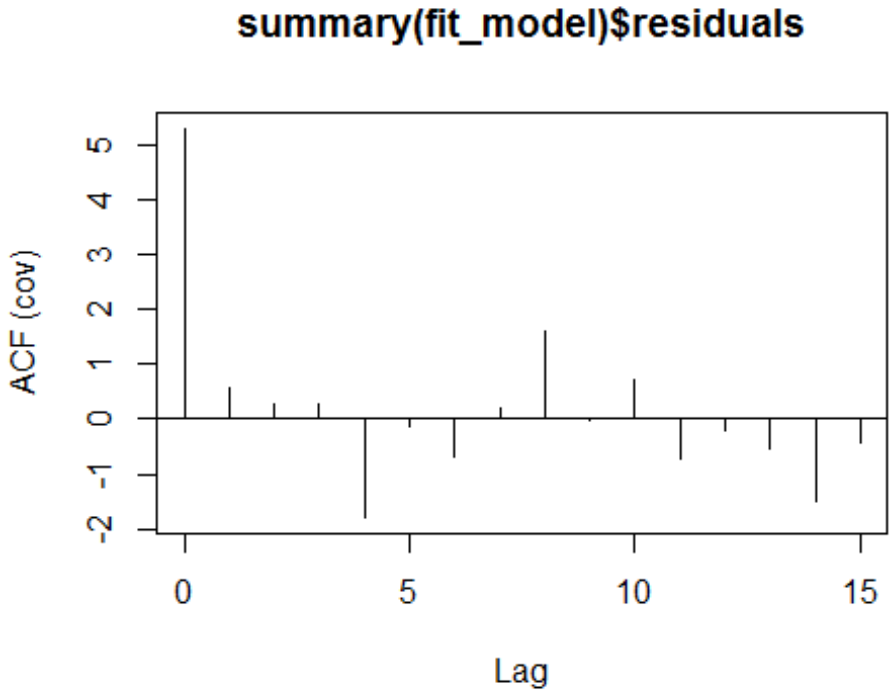
Residuals Analysis

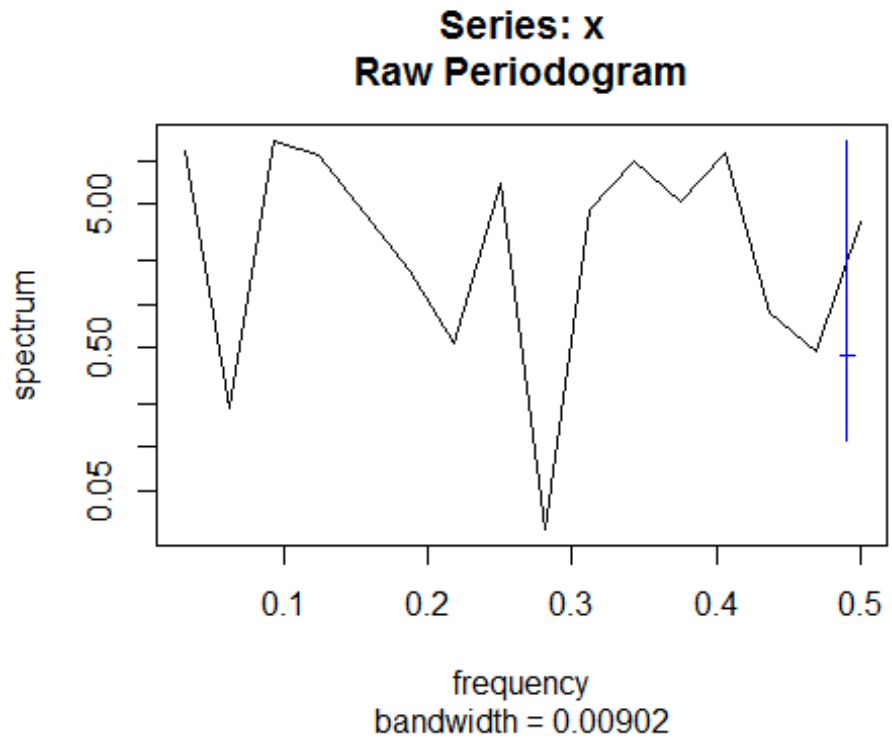
Once we diagnosed the significance of our model we have to proof that our residuals follow a normal distribution

```
library(normwhn.test)
summary(fit_model)$residuals

##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##          -1.4704610          -1.1582487          -3.4810670
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##           0.5425557           1.6904131          -2.7540920
##          Duster 360           Merc 240D            Merc 230
##          -0.7538960           2.7581469          -2.5535825
##            Merc 280           Merc 280C          Merc 450SE
##           0.6212790          -1.5142526           1.3919737
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##           0.7151853          -1.6793439          -0.6975657
## Lincoln Continental   Chrysler Imperial            Fiat 128
##           0.1800477           4.6609983           4.5946906
##         Honda Civic      Toyota Corolla       Toyota Corona
##           1.4681276           4.1380358          -2.9935771
##    Dodge Challenger         AMC Javelin          Camaro Z28
##          -1.0123837          -2.1724175          -0.1693090
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##           3.7398205          -0.8444279           1.3554045
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##           3.0545795          -2.1136475          -1.0061349
##       Maserati Bora          Volvo 142E
##          -1.4696346          -3.0672164

whitenoise.test(as.data.frame(summary(fit_model)$residuals))
```



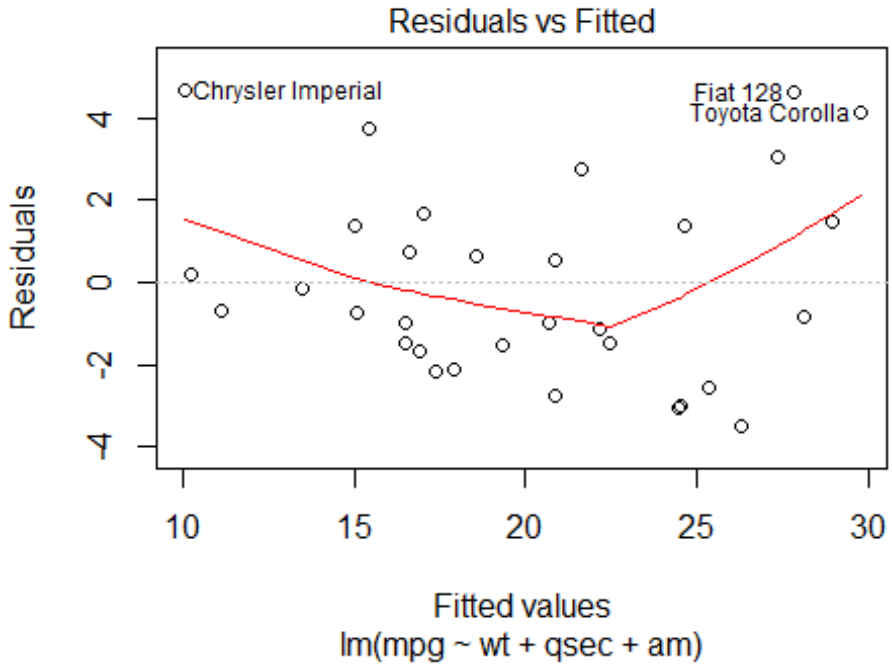


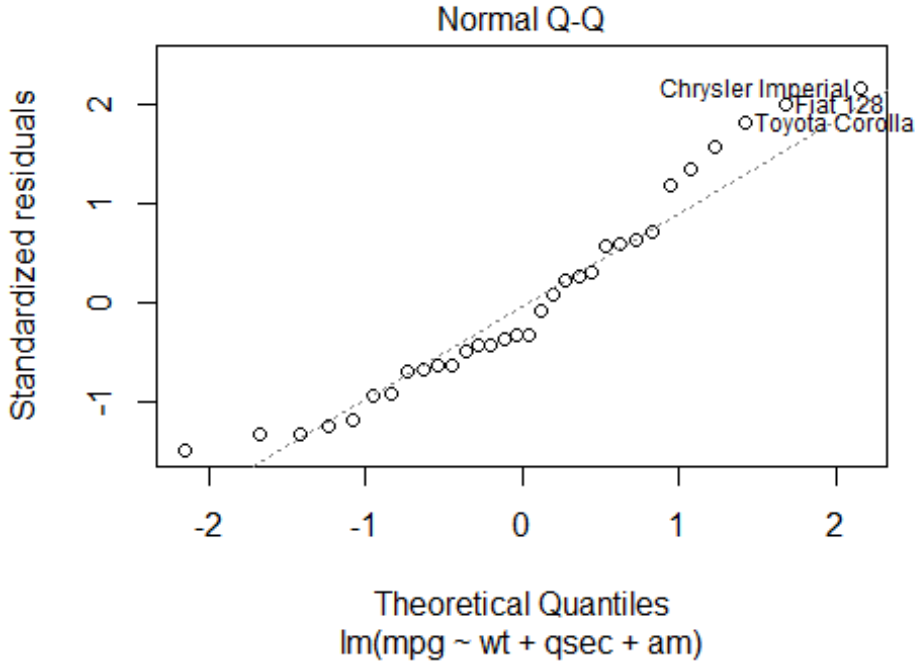

```
## [1] "no. of observations"
## [1] 1
## [1] "T"
## [1] 16
```

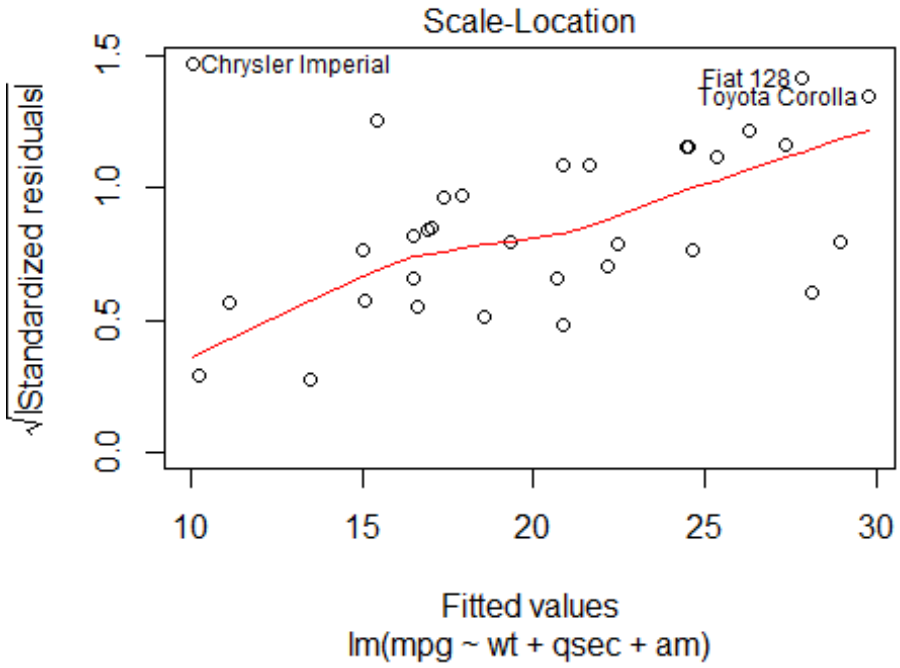
```
## [1] "CVM stat MN"
## [1] 0.7953912
## [1] "tMN"
## [1] -0.8184352
## [1] "test value"
## [1] 0.68238
```

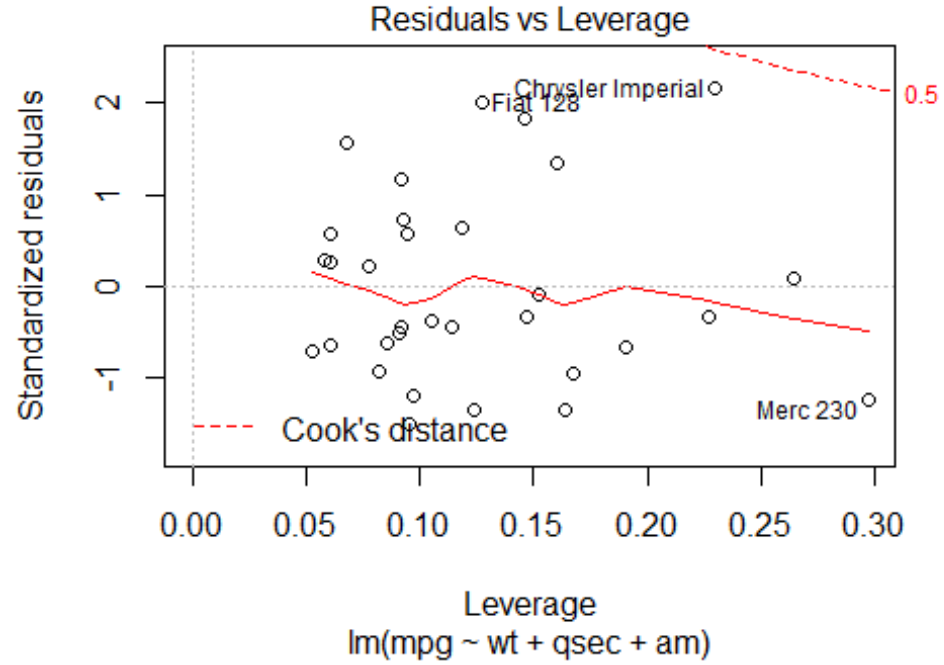
Using whitenoise.test we proof that the residuals follow a gaussian distribution by the plots we have the following conclutions the assumption of Independence is supported by the plot Residuals vs Fitted here we see that there may exist 3 outliers in the data, but we will ensure this later the Gaussian distribution shows us that the residuals closely follow the line, so we can conclude that the residuals follow a normal distribution (This is supported by the whitenoise.test) in the Scale Location we can confirm that the variance is constant (In time series this property is known as Homocedasticity) in the residuals vs leverage plot we see that every point is under the 0.5 threshold so because of this we can ensure that there is no outliers

```
plot(fit_model)
```


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals
4
2
0
-2
-4
10
15
20
25
30
Fitted values
lm(mpg ~ wt + qsec + am)


Normal Q-Q
Chrysler Imperial
Fiat 128
Toyota Corolla
Standardized residuals
2
1
0
-1
-2
-1
0
1
2
Theoretical Quantiles
lm(mpg ~ wt + qsec + am)





Conclusions

We have seen that a car with automatic transmission has more MPG than a manual one based on the mean differences for each option In a first linear model we saw that

our R^2 coefficient was near 0.83 and in our last model that value grew a little, we can assert that the model is optimal because this factor does not decrease from its extended form, also we can say that the weight, the 1/4 mile time and the transmission are more statistically significant when determining MPG.