Regression Model Course Project

Mohab Diab

September 15, 2018

Libraries & Backages:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
library(stats)
library(statsr)
```

Data Summary:

we are looking at a data set of car collection assigned with some specifications such as MPG, Number of cylinders, and so on…, and we need to answer some questions such as . Is an Automatic or manual transmission better of MPG consumbtion? . Quantify the difference between automatic and manual transmissions and to figure this out, I will follow these following steps 1. process and prepare data. 2. Explore data through visualization to gain sense of what I'm doing. 3. model selection, to figure which model is better from the other for the MPG. 4. Model exaamination to figure whether my model holds up against standarsd and conditions. 5. Jumbing to a conclusion depending on my answers to the questions.

Data Processing & Preparing:

```
data(mtcars)
mtcarsdata<- mtcars
remove(mtcars)
mtcarsdata$am<- as.factor(mtcarsdata$am)
levels(mtcarsdata$am)<-c("Automatic", "Manual")
mtcarsdata$cyl<- as.factor(mtcarsdata$cyl)
mtcarsdata$gear<- as.factor(mtcarsdata$gear)
mtcarsdata$vs<- as.factor(mtcarsdata$vs)
levels(mtcarsdata$vs)<-c("V", "S")
```

Exploratory Data Analysis

. Exploring datat dimensions and summary

```
head(mtcars)
```

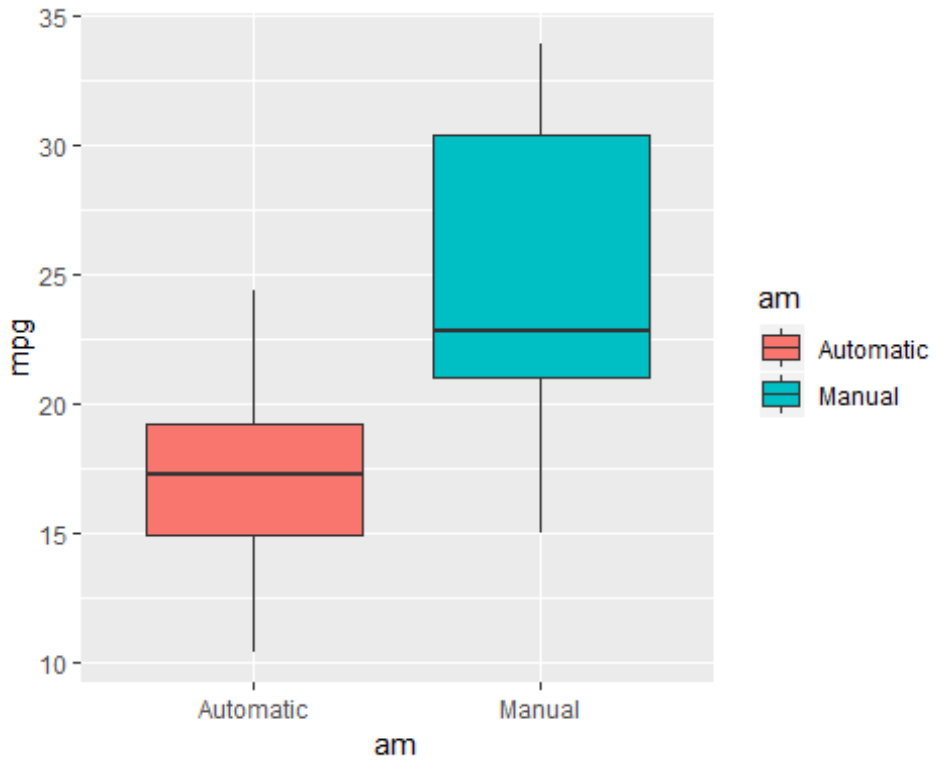
```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

```
dim(mtcars)
```

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

Visualising the relationship between the columns that I'm interested in:

```
ggplot(data=mtcarsdata, aes(am, mpg)) +
        geom_boxplot(aes(fill= am))
```

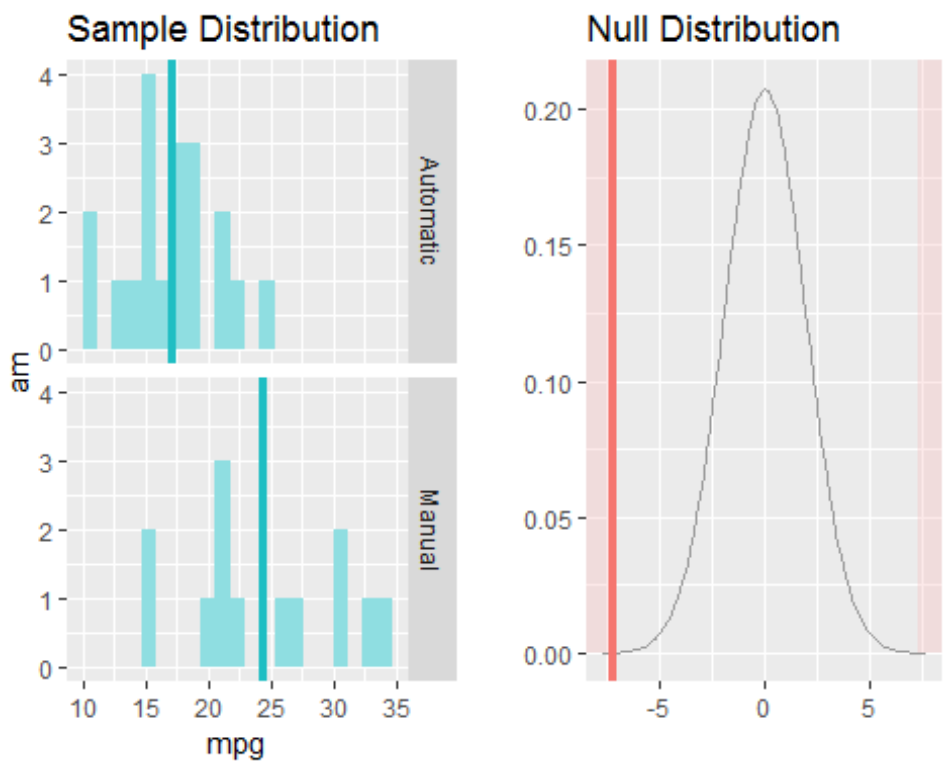


we can notice the difference between the two types and that Manual Transmission has a higher Consumption than the Automatic one but we dont really know whether it's a noticable difference or not so I'm calling a T test to compare between them using 95% Interval of Confidence

```
#Hypothesis Test
inference(mpg, am, data= mtcarsdata,statistic="mean", alt="twosided", type="ht", method="theoretical")

## Warning: Missing null value, set to 0

## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_Automatic = 19, y_bar_Automatic = 17.1474, s_Automatic = 3.834
## n_Manual = 13, y_bar_Manual = 24.3923, s_Manual = 6.1665
## H0: mu_Automatic =  mu_Manual
## HA: mu_Automatic != mu_Manual
## t = -3.7671, df = 12
## p_value = 0.0027
```



this T test shows that there is a difference between the two types of trasmissions but this crystal clear as data might be biased or the sample size might be not enough

Choosing The Best Model

To Choose the best regression model for this case I have to assign all the predictors that might Influence my model beside the transmission type so I'm the (Backward Methodology) for to achieve the best regression model depending on the P-Value

assigning all variables to the model

```
model1<- lm(mpg ~ am + wt + cyl + hp + drat + disp , data= mtcarsdata)
summary(model1)

##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp + drat + disp, data = mtcarsdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.8267 -1.4366 -0.4153  1.1649  5.0671
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.611986   6.274227   5.198 2.52e-05 ***
## amManual     1.681130   1.554386   1.082   0.2902
## wt          -2.726729   1.200207  -2.272   0.0323 *
## cyl6        -3.026760   1.576680  -1.920   0.0669 .
## cyl8        -2.541967   3.059145  -0.831   0.4142
## hp          -0.033038   0.014476  -2.282   0.0316 *
## drat         0.326616   1.471086   0.222   0.8262
## disp         0.004395   0.013090   0.336   0.7400
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.501 on 24 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8278
## F-statistic: 22.29 on 7 and 24 DF,  p-value: 4.768e-09
```

I can see that adjusted R squared is 82% which is sufficient as it indecates the percentage of variability that can be explaind by the predictors But i can Enhance my model by removing the mhighest P-Value which is related to drat

```
model2<- lm(mpg ~ am + wt + cyl + hp  + disp , data= mtcarsdata)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp + disp, data = mtcarsdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9374 -1.3347 -0.3903  1.1910  5.0757
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.864276   2.695416  12.564 2.67e-12 ***
## amManual     1.806099   1.421079   1.271   0.2155
## wt          -2.738695   1.175978  -2.329   0.0282 *
## cyl6        -3.136067   1.469090  -2.135   0.0428 *
## cyl8        -2.717781   2.898149  -0.938   0.3573
## hp          -0.032480   0.013983  -2.323   0.0286 *
## disp         0.004088   0.012767   0.320   0.7515
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.453 on 25 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8344
## F-statistic: 27.03 on 6 and 25 DF,  p-value: 8.861e-10
```

you can see that R squared has jumped from 82% to 83.4% which is perfection in this case but I'll have to try removig the second hight P- value which is related to the Piston Displacement

```
model3<-lm(formula = mpg ~ am + wt + cyl + hp , data = mtcarsdata)
summary(model3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + hp, data = mtcarsdata)
##
## 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 ***
## amManual     1.80921    1.39630   1.296  0.20646
## wt          -2.49683    0.88559  -2.819  0.00908 **
## 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 *
## ---
## 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
```

and jumping again to 84 which i think the best model to predict MPG.

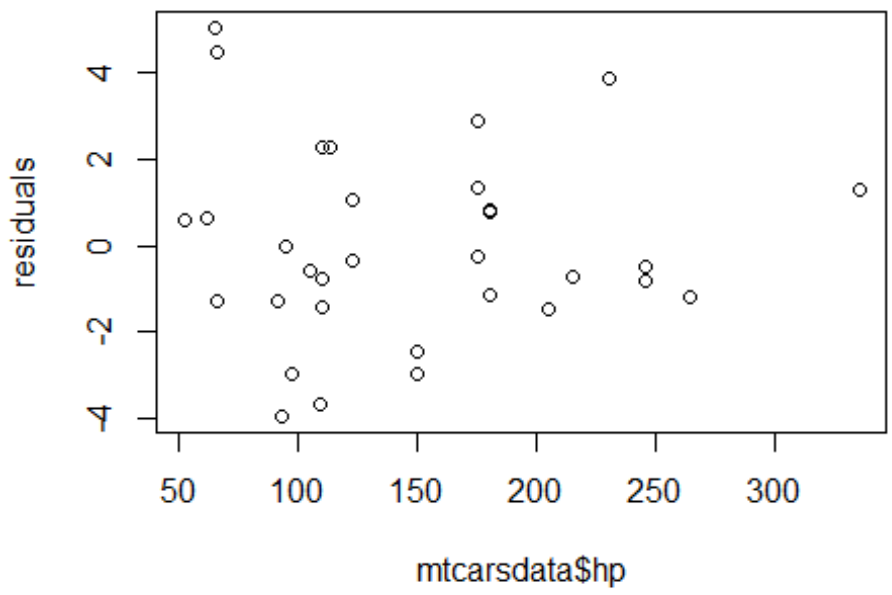
so my residuals are:

```
residuals<-residuals(model3)
```

Checking for Standars & Conditions

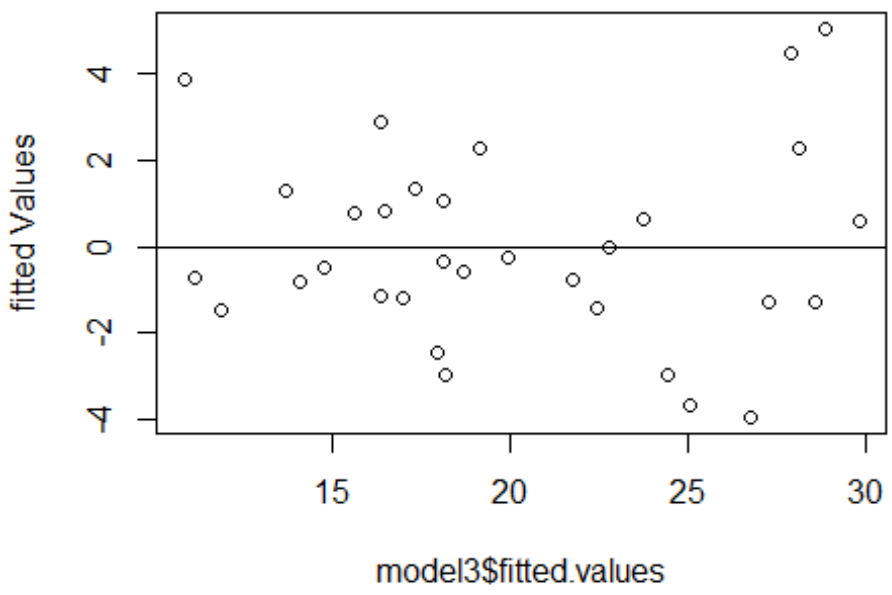
1. linearity

```
plot(residuals~mtcarsdata$hp)
```



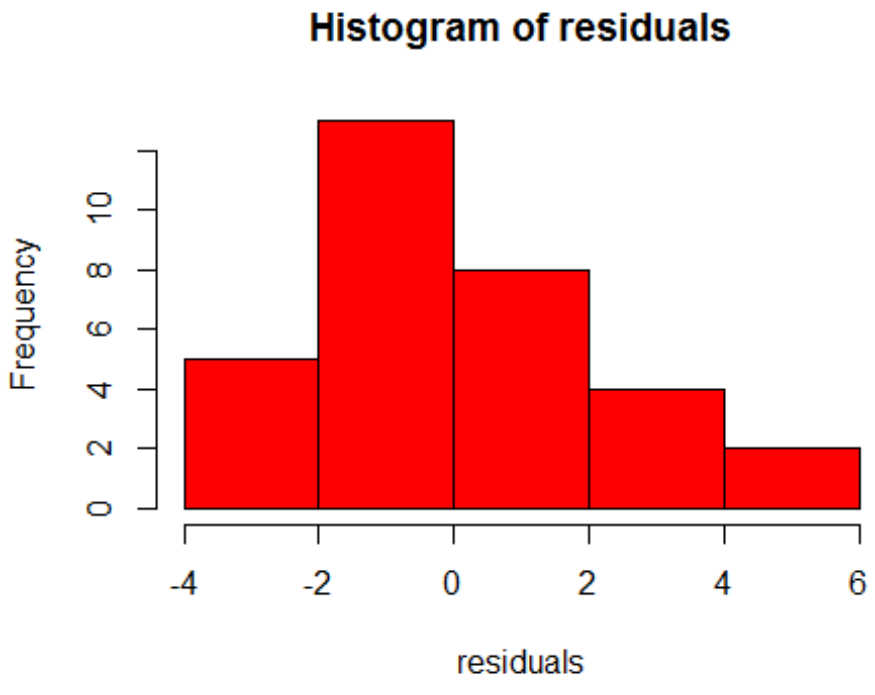
2. constants residuals. standard deviation

```
plot(residuals~ model3$fitted.values, ylab = "fitted Values")
abline(h=0)
```

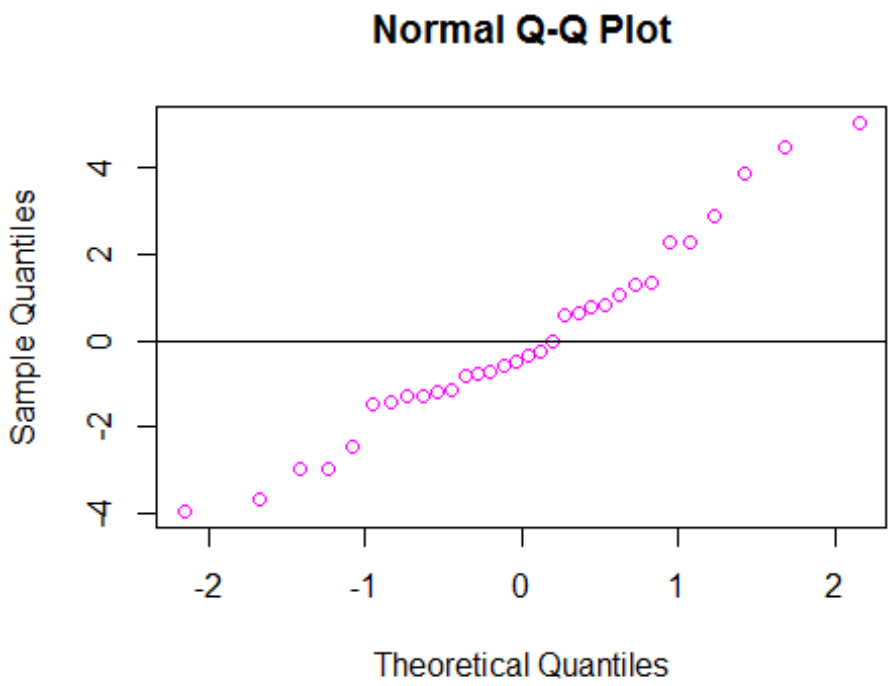


3. normally distributed around 0

```
mean(residuals) # near to ZERO
## [1] 8.326673e-17
hist(residuals ,col=2)
```




```
qqnorm(residuals, col= 22)
qqline(0)
```



Conclusion

1. for the first and second questions, I can tell there is adifference between the two types of transmission and this difference appears in the H.T i did above, and apparently Automatic is way better tha Manual Transmission for the consumbtion which appears in the regression model above too.

2. the models can explain 85 of the variablility in the prdicted values depending on the predictors which is great, as it indecates that, this model may be reliable to predict outsiders, in other word to predict the population movemnts.
3. I can tell the model need to be more enhanced, maybe by increasing the numbers in the sample to get more normality, and avoid skewness.