605_HW11.Rmd

Kumudini Bhave

April 18, 2017

LINEAR REGRESSION IN R

Problem Set 1 :

Using R's lm function, perform regression analysis and measure the significance of the independent variables for the following two data sets. In the first case, you are evaluating the statement that we hear that Maximum Heart Rate of a person is related to their age by the following equation:

$$MaxHR = 220 - Age$$

The sample we have

Age	MaxHR
18	202
23	186
25	187
35	180
65	156
54	169
34	174
56	172
72	153
19	199
23	193
42	174
18	198
39	183
37	178

Perform a linear regression analysis fitting the Max Heart Rate to Age using the lm function in R. What is the resulting equation? Is the effect of Age on Max HR significant? What is the significance level? Please also plot the fitted relationship between Max HR and Age.

Solution :

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

# library for displaying data in tabular format
library(DT)
```

```
## Warning: package 'DT' was built under R version 3.3.3
```

```r
# library for plotting
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```r
age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)

maxhr <- c(202, 186, 187, 180, 156, 169, 174, 172, 153, 199, 193, 174, 198, 183, 178)

agehr <- data.frame(cbind(age, maxhr))

head(agehr)
```

```
##   age maxhr
## 1  18   202
## 2  23   186
## 3  25   187
## 4  35   180
## 5  65   156
## 6  54   169
```

```r
# Taking the basic model of Max heart rate to Age
base.model <- lm(maxhr ~ age, data = agehr)

summary(base.model)
```

```
##
## Call:
## lm(formula = maxhr ~ age, data = agehr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9258 -2.5383  0.3879  3.1867  6.6242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
## age          -0.79773    0.06996  -11.40 3.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.578 on 13 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9021
## F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

From the summary we get the fitted line equation

$$\hat{y} = \beta_0 + \beta_1 x$$

as

$$\hat{maxhr} = 210.04846 + (-0.79773) \times age$$

The Hypothesis Test

H_0 : Age does not affect the max heart rate i.e. $\beta_1 = 0$

H_A : Age does affect the max heart rate i.e. $\beta_1 \neq 0$

From the summary we find the p value to be 3.85e-08 which is very much significant as it is much lower than 0.05 Hence age does seem to have a strong relationship with max heart rate for any individual.

Also we see that the coefficient for age is $\beta_1 \neq 0$ as β_1 = -0.79773. The negative symbolizes that as age increases we find a decrease in the max heart rate for any individual.

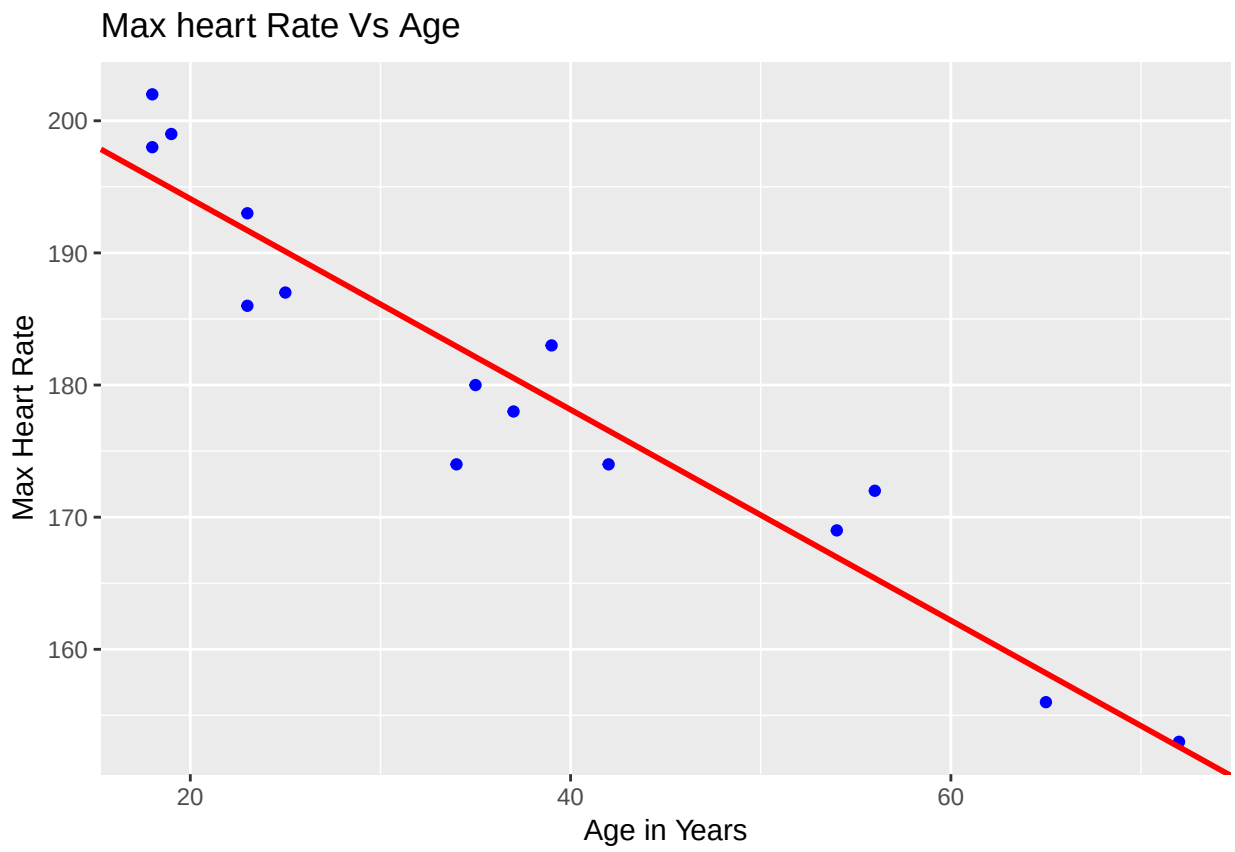
Hence we reject the null hypothesis H_0.

The fitted relationship between Maximum heart rate and Age can be plotted as below

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

plotagehr <- ggplot(data = agehr, aes(y = maxhr, x = age)) + geom_point(colour = "blue") + scale_x_conti
    geom_abline(intercept = base.model$coefficients[1], slope = base.model$coefficients[2], color = "re


plotagehr
```



Problem Set 2 :

Using the Auto data set from Assignment 5 (also attached here) perform a Linear Regression analysis using mpg as the dependent variable and the other 4 (displacement, horsepower, weight, acceleration) as independent variables. What is the final linear regression fit equation? Which of the 4 independent variables have a significant impact on mpg? What are their corresponding significance levels? What are the standard errors on each of the coefficients? Please perform this experiment in two ways. First take any random 40 data points from the entire auto data sample and perform the linear regression fit and measure the 95% confidence intervals. Then, take the entire data set (all 392 points) and perform linear regression and measure the 95% confidence intervals. Please report the resulting fit equation, their significance values and confidence intervals for each of the two runs.

Solution :

Extracting Raw Data From GitHub Data File And Reading In CSV Format

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

# Loading RCurl package to help scrape data from web (stored on GitHub).
library(RCurl)

# Loading plyr package to help map abbreviated values to explained.
library(plyr)

data.giturl <- "https://raw.githubusercontent.com/DataDriven-MSDA/DATA605/master/auto-mpg.data"
autompg.data <- read.table(data.giturl)
head(autompg.data)
```

```
##   V1 V2  V3    V4   V5   V6 V7 V8                        V9
## 1 18  8 307 130.0 3504 12.0 70  1 chevrolet chevelle malibu
## 2 15  8 350 165.0 3693 11.5 70  1         buick skylark 320
## 3 18  8 318 150.0 3436 11.0 70  1        plymouth satellite
## 4 16  8 304 150.0 3433 12.0 70  1             amc rebel sst
## 5 17  8 302 140.0 3449 10.5 70  1               ford torino
## 6 15  8 429 198.0 4341 10.0 70  1          ford galaxie 500
```

```
# Attempt extract the required data columns of displacement, horsepower, weight, acceleration, mpg from
autompg.datastudy <- subset(autompg.data, select = c(V3, V4, V5, V6, V1), V4 != "?")
autompg.datastudy <- na.omit(autompg.datastudy)


# Verifying the number of attributes
length(autompg.datastudy)
```

```
## [1] 5
```

```
# Verifying the number of observations selected
nrow(autompg.datastudy)
```

```
## [1] 392
```

```
# Renaming the attributes
colnames(autompg.datastudy) <- c("disp", "hp", "wt", "acc", "mpg")
```

```r
# Viewing the data
head(autompg.datastudy)
```

```
##   disp    hp   wt  acc mpg
## 1  307 130.0 3504 12.0  18
## 2  350 165.0 3693 11.5  15
## 3  318 150.0 3436 11.0  18
## 4  304 150.0 3433 12.0  16
## 5  302 140.0 3449 10.5  17
## 6  429 198.0 4341 10.0  15
```

```r
dim(autompg.datastudy)
```

```
## [1] 392   5
```

```r
str(autompg.datastudy)
```

```
## 'data.frame':    392 obs. of  5 variables:
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : Factor w/ 94 levels "?","100.0","102.0",..: 17 35 29 29 24 42 47 46 48 40 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
```

```r
# Converting the horsepower as numeric as required , since it is quantitative variable
autompg.datastudy$hp <- as.numeric(autompg.datastudy$hp)
str(autompg.datastudy)
```

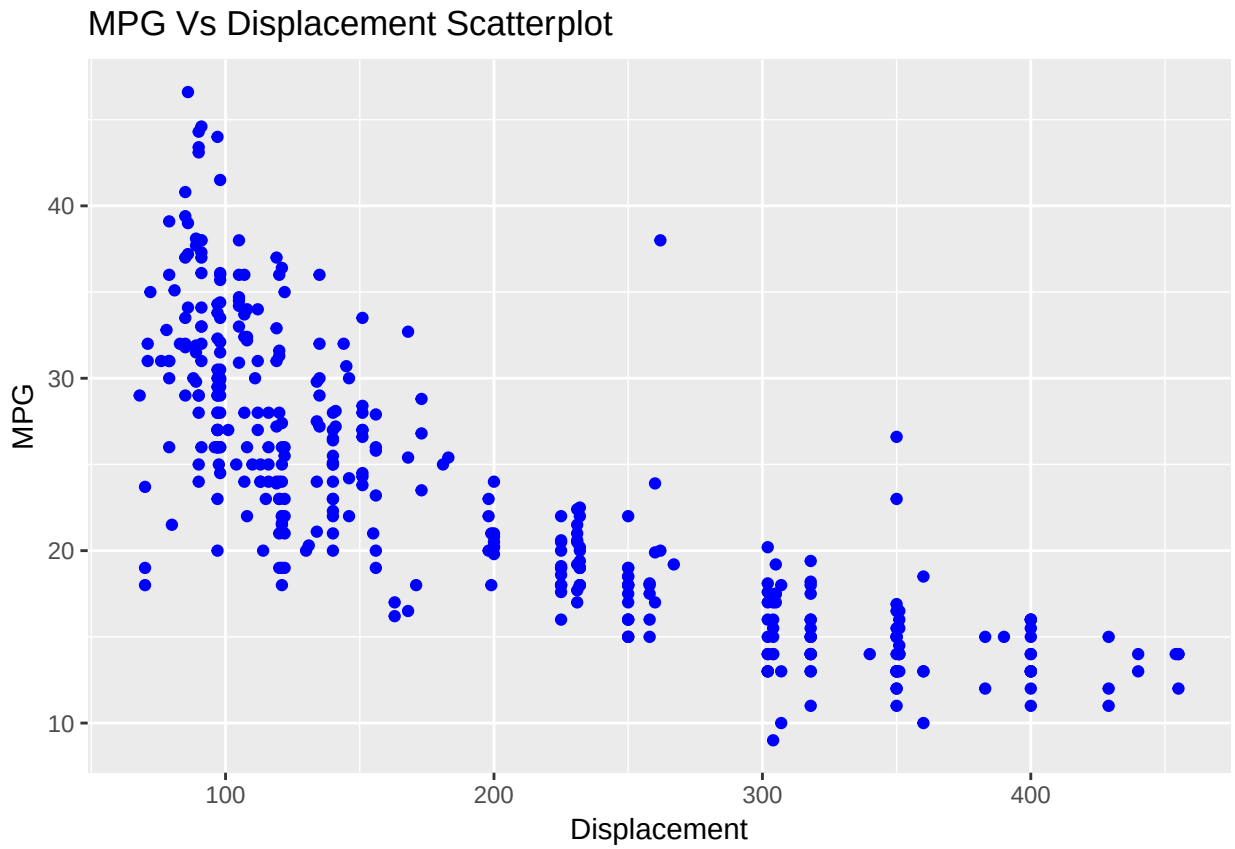
```
## 'data.frame':    392 obs. of  5 variables:
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : num  17 35 29 29 24 42 47 46 48 40 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
```

To understand the data and the relation between the various predictor variables (displacement, horsepower, weight , acceleration) on mpg, we plot the following scatterplots for mpg vs each of the predictor variables

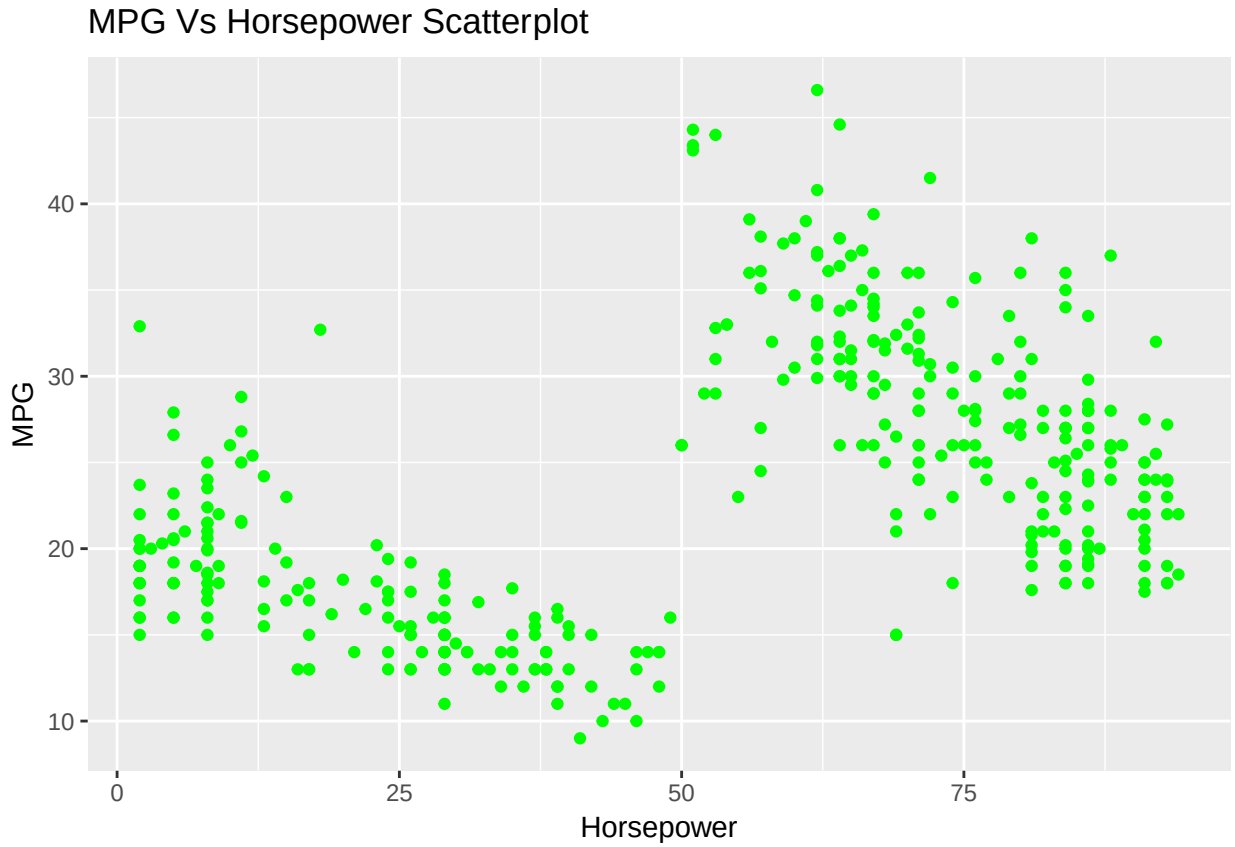
```r
knitr::opts_chunk$set(message = FALSE, echo = TRUE)


par(mfrow = c(1, 2))


ggplot(data = autompg.datastudy, aes(y = mpg, x = disp)) + geom_point(colour = "blue") + scale_x_continu
    ggtitle("MPG Vs Displacement Scatterplot")
```

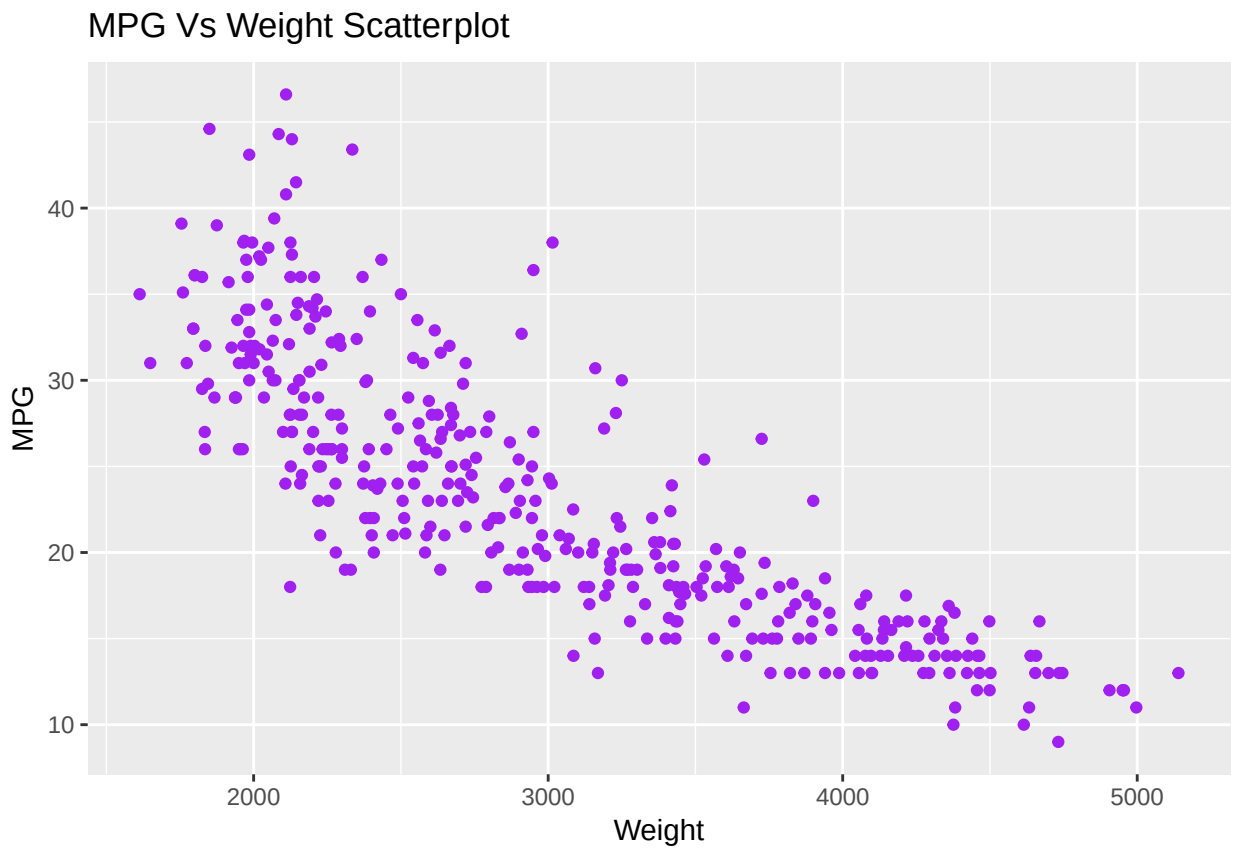



```
ggplot(data = autompg.datastudy, aes(y = mpg, x = hp)) + geom_point(colour = "green") + scale_x_continu
    ggtitle("MPG Vs Horsepower Scatterplot")
```

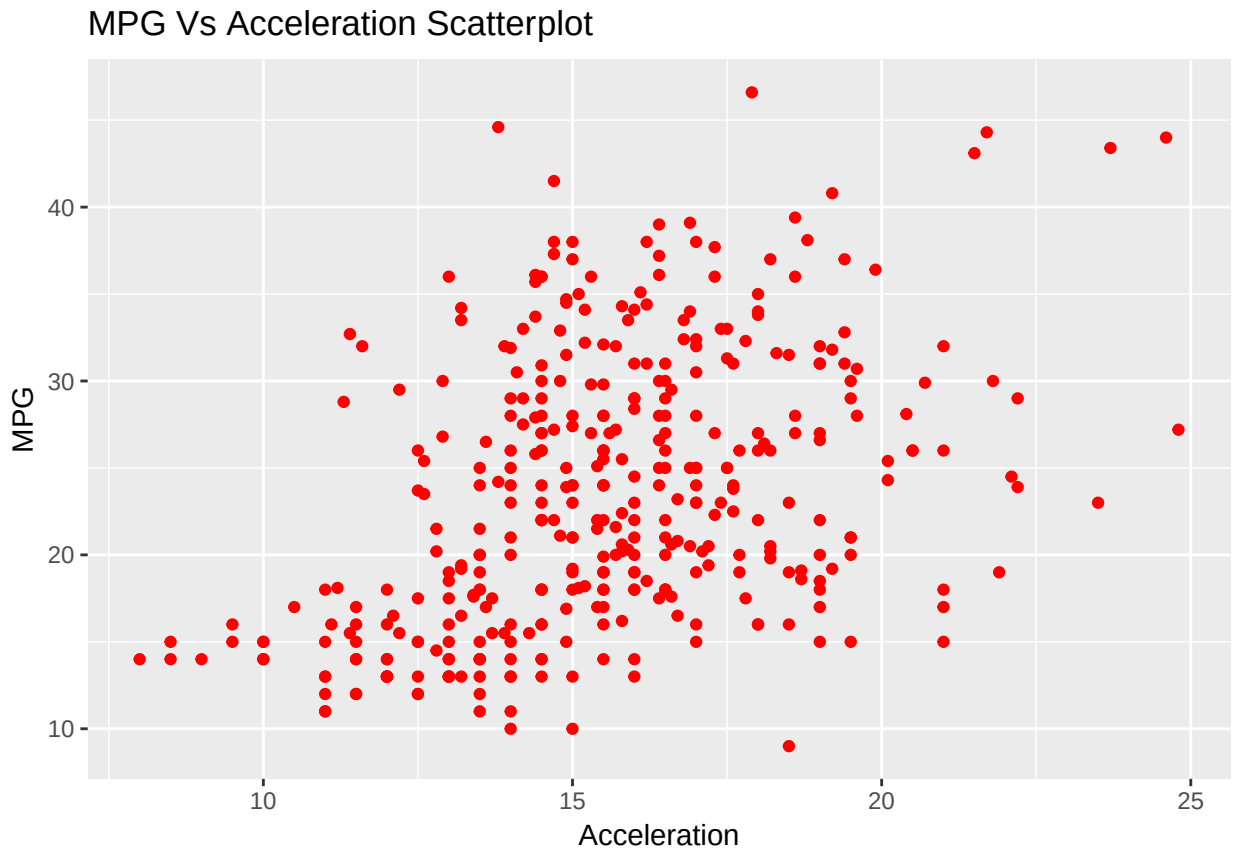



```
par(mfrow = c(1, 2))

ggplot(data = autompg.datastudy, aes(y = mpg, x = wt)) + geom_point(colour = "purple") + scale_x_contin
    ggtitle("MPG Vs Weight Scatterplot")
```




```
ggplot(data = autompg.datastudy, aes(y = mpg, x = acc)) + geom_point(colour = "red") + scale_x_continuou
    ggtitle("MPG Vs Acceleration Scatterplot")
```



Auto Data Sample Study

Lets take the sample data from the whole population of auto data.

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

# Function to randomly select samples from a given population dataset @param1 dataframe @param2 number
rsamples <- function(df, num) {
    set.seed(10)
    return(df[sample(nrow(df), num), ])

}

# Taking 40 random samples form autodata population
autodatasamples <- rsamples(autompg.datastudy, 40)

head(autodatasamples)
```

```
##     disp hp   wt  acc  mpg
## 201  250 74 3574 21.0 18.0
## 121  121  9 2868 15.5 19.0
## 169  140 79 2639 17.0 23.0
## 272  156  5 2745 16.7 23.2
## 35   225  5 3439 15.5 16.0
## 89   302 21 4042 14.5 14.0
```

```r
dim(autodatasamples)
```

```
## [1] 40  5
```

```r
str(autodatasamples)
```

```
## 'data.frame':    40 obs. of  5 variables:
##  $ disp: num  250 121 140 156 225 302 350 360 98 302 ...
##  $ hp  : num  74 9 79 5 5 21 39 37 79 16 ...
##  $ wt  : num  3574 2868 2639 2745 3439 ...
##  $ acc : num  21 15.5 17 16.7 15.5 14.5 12.5 13 15.9 12 ...
##  $ mpg : num  18 19 23 23.2 16 14 12 13 33.5 13 ...
```

```r
# Taking the model of mpg to all the predictors i.e. displacement, horsepower, weight, acceleration
auto.samplemodel <- lm(mpg ~ ., data = autodatasamples)

samplesummary <- summary(auto.samplemodel)
samplesummary
```

```
##
## Call:
## lm(formula = mpg ~ ., data = autodatasamples)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9914 -2.9370  0.0734  2.1250 12.4163
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 41.354484   7.358303   5.620 2.45e-06 ***
## disp        -0.018761   0.025384  -0.739   0.4648
## hp           0.024381   0.026272   0.928   0.3598
## wt          -0.005674   0.002603  -2.180   0.0361 *
## acc          0.129268   0.318987   0.405   0.6878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.448 on 35 degrees of freedom
## Multiple R-squared:  0.7624, Adjusted R-squared:  0.7352
## F-statistic: 28.07 on 4 and 35 DF,  p-value: 1.715e-10
```

From the summary we get the fitted line equation

$$\hat{mpg} = \beta_0 + \beta_1 \times disp + \beta_2 \times hp + \beta_3 \times wt + \beta_4 \times acc$$

as

$$\hat{mpg} = 41.354484 + (-0.018761) \times disp + (0.024381) \times hp + (-0.005674) \times wt + (0.129268) \times acc$$

The significance of the impact of the 4 independant variables on mpg can be gauged by the corresponding p values derived from the sample data, these are

Independant Variable	p-value	Significance of Impact
displacement	0.4648	not significant
horsepower	0.3598	not significant

Independant Variable	p-value	Significance of Impact
weight	0.0361	quite significant
accelaration	0.6878	not significant

The p-values for each of the independant variables for sample data are

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

samplesummary$coefficients[, 4]
```

```
##  (Intercept)         disp           hp           wt          acc
## 2.446714e-06 4.647674e-01 3.597543e-01 3.610395e-02 6.877635e-01
```

Form the p-values we derive that only weight seems to impact the mpg significantly. The displacement, horsepower, and acceleration do not have any significant impact on the mpg.

The standard error for sample data for the independant variables are

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

samplesummary$coefficients[, 2]
```

```
## (Intercept)        disp          hp          wt         acc
## 7.358302593 0.025383840 0.026272435 0.002603368 0.318986501
```

The confidence interval for sample data can be calculated as

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

confint(auto.samplemodel, level = 0.95)
```

```
##                    2.5 %        97.5 %
## (Intercept) 26.41633571 56.2926325697
## disp        -0.07029343  0.0327704383
## hp          -0.02895459  0.0777171669
## wt          -0.01095939 -0.0003891498
## acc         -0.51830915  0.7768448995
```

Auto Data Population Study

Now considering the entire populaiton data for the autodata

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

dim(autompg.datastudy)
```

```
## [1] 392   5
```

```
str(autompg.datastudy)
```

```
## 'data.frame':    392 obs. of  5 variables:
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : num  17 35 29 29 24 42 47 46 48 40 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
```

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)


# Taking the full model for the population

auto.popmodel <- lm(mpg ~ ., data = autompg.datastudy)

popsummary <- summary(auto.popmodel)
popsummary
```

```
##
## Call:
## lm(formula = mpg ~ ., data = autompg.datastudy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7797  -2.7837  -0.3145   2.3950  16.2489
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.4449493  2.0088844  20.133  < 2e-16 ***
## disp        -0.0102652  0.0065455  -1.568   0.1176
## hp           0.0064381  0.0085818   0.750   0.4536
## wt          -0.0061061  0.0007478  -8.165 4.58e-15 ***
## acc          0.1828906  0.0981158   1.864   0.0631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.281 on 387 degrees of freedom
## Multiple R-squared:  0.7022, Adjusted R-squared:  0.6991
## F-statistic: 228.1 on 4 and 387 DF,  p-value: < 2.2e-16
```

From the summary we get the fitted line equation as

$$\hat{mpg} = 40.4449493 + (-0.0102652) \times disp + 0.0064381 \times hp + (-0.0061061) \times wt + 0.1828906 \times acc$$

The significance of the impact of the 4 independant variables on mpg can be gauged by the corresponding p values derived from the population data, these are

Independant Variable	p-value	Significance of Impact
displacement	0.1176	not significant
horsepower	0.4536	not significant
weight	4.58e-15	quite significant
accelaration	0.0631	not significant

The p-values for population data for each of the independant variables are

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

popsummary$coefficients[, 4]
```

```
##  (Intercept)         disp           hp           wt          acc
```

```
## 3.428043e-62 1.176332e-01 4.535837e-01 4.583547e-15 6.307469e-02
```

As we observed with sample data, the population data for autodata also suggests that it is the weight that impacts the mpg significantly rather than the displacement, horsepower , acceleration.

The standard error for the independant variables for population data are

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

popsummary$coefficients[, 2]
```

```
##  (Intercept)         disp           hp           wt          acc
## 2.0088843769 0.0065455205 0.0085817837 0.0007478385 0.0981157949
```

The confidence interval for population data can be calculated as

```
knitr::opts_chunk$set(message = FALSE, echo = TRUE)

confint(auto.popmodel, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) 36.495256067 44.394642538
## disp        -0.023134437  0.002604026
## hp          -0.010434610  0.023310899
## wt          -0.007576471 -0.004635801
## acc         -0.010016073  0.375797362
```

We also observe that the standard error has significantly reduced from sample to population data since the 'n' of the data under study increases. This also results in a decrease in width of the confidence interval.
