# Data Science in Statistical Methods using R

Md Sayeef Alam

21/09/2020

# Day 1

## Session 1: Application of Regression and Multiple Regression in Data Science

### Dr. R. K. Jana, IIM Raipur

Simple addition in R

```
1+1
```

```
## [1] 2
```

Some packages to be installed

```
install.packages("matlib", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("corpcor", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("GPArotation", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("psych", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("FactoMineR", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tseries", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("corrplot", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tseries", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("ggpubr", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("tidyverse", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("Hmisc", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("dplyr", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("ggplot2", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("lattice", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("grid", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("DMwR", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("stats", dependencies = T,repos = "http://cran.us.r-project.org")
install.packages("nortest", dependencies = T,repos = "http://cran.us.r-project.org")
```

Adding the libraries corresponding to packages.

```
library(dplyr)
library(tseries)
library(matlib)
library(corpcor)
library(GPArotation)
library(psych)
library(FactoMineR)
library(corrplot)
library(ggpubr)
library(lattice)
library(grid)
library(nortest)
library(stats)
library(DMwR)
library(ggplot2)
```

Reading xls and xlsx files

```
install.packages("gdata", dep = T,repos = "http://cran.us.r-project.org")
library(gdata)
xls.data = read.xls("file.xls")
```

You need to specify the sheetIndex (sheet number)

```
install.packages("xlsx", dep = T,repos = "http://cran.us.r-project.org")
library(xlsx)
xlsx.data = read.xlsx("file.xlsx", sheetIndex = 1)
```

## Linear Regression

Simple Linear Regression

1 dependent (y)

1 independent (x)

Assumptions

1. Relationships between the above two must be linear
2. Residuals should be normally distributed
3. Residuals should be homoscedastic
4. Residuals should be independent

Homoscedasticity means same variance, error term (i.e. distance of the points from the fitted line) should be same across all values of the independent variables.

Heteroscedasticity is when the error varies with the values of the independent variables.

Several measures are there to check for homoscedasticity

```
library(datasets)
data(cars)
```

Lets check the variables inside the dataset

```
names(cars)
```

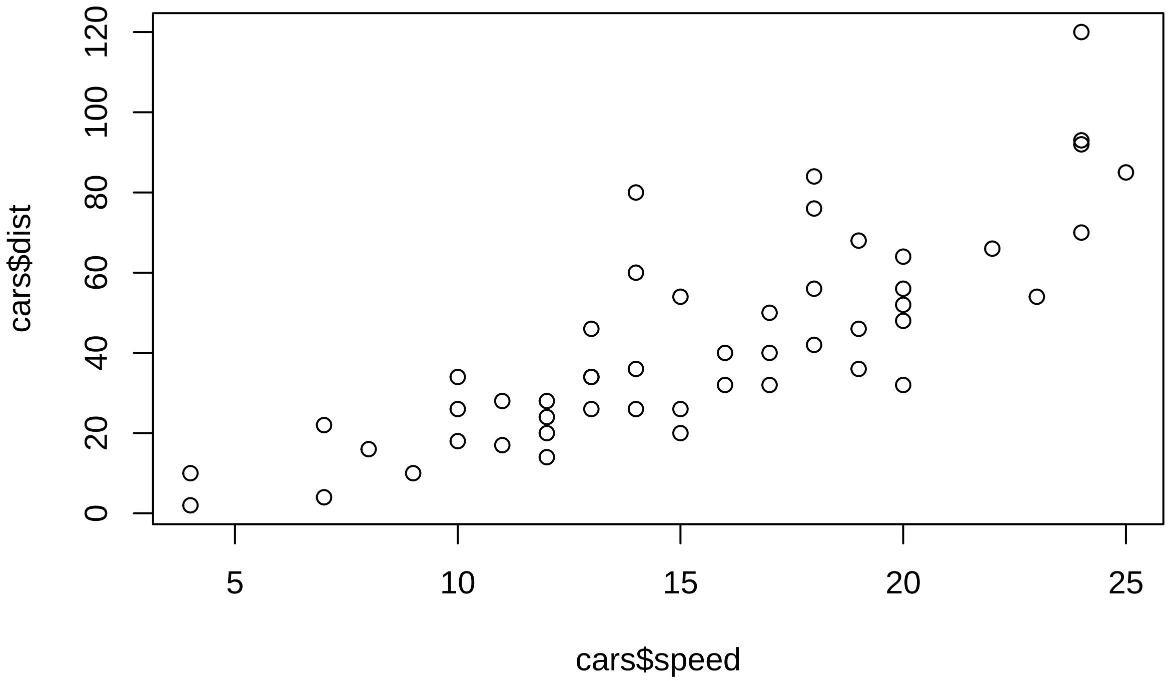
```
## [1] "speed" "dist"
```

```
head(cars)
```

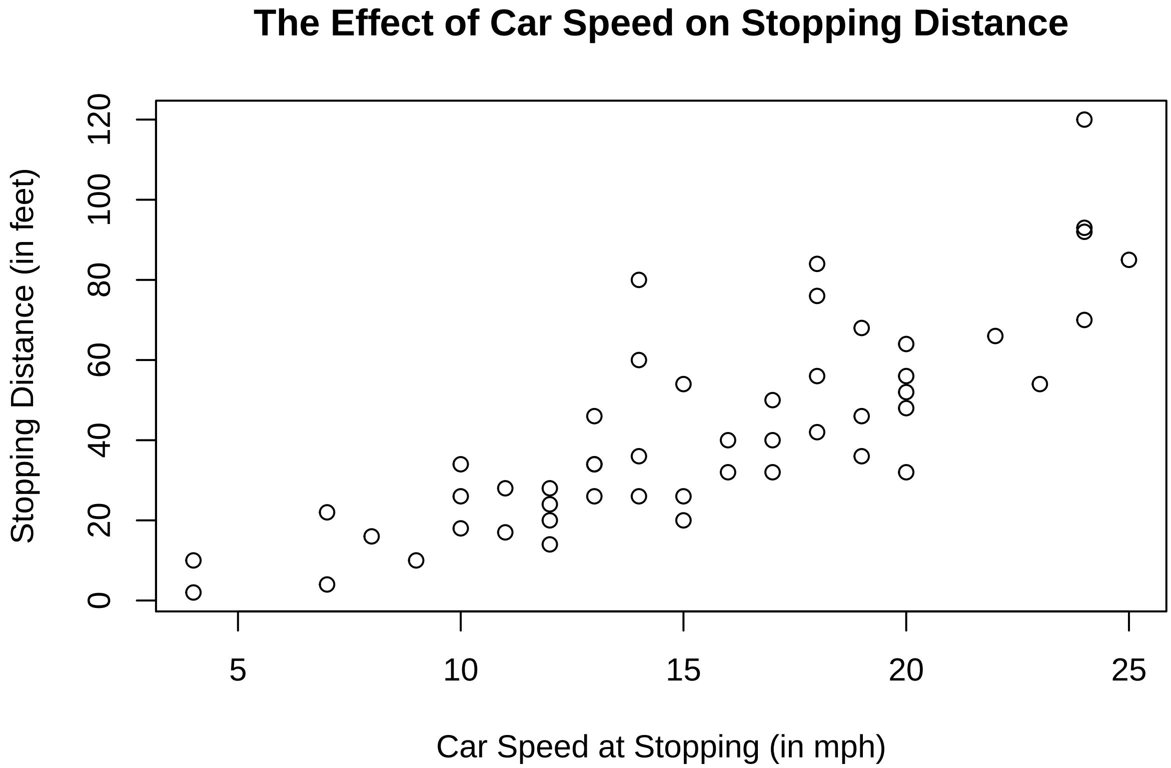
```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Lets plot some parameters specifically speed vs distance

```
plot(cars$speed, cars$dist)
```

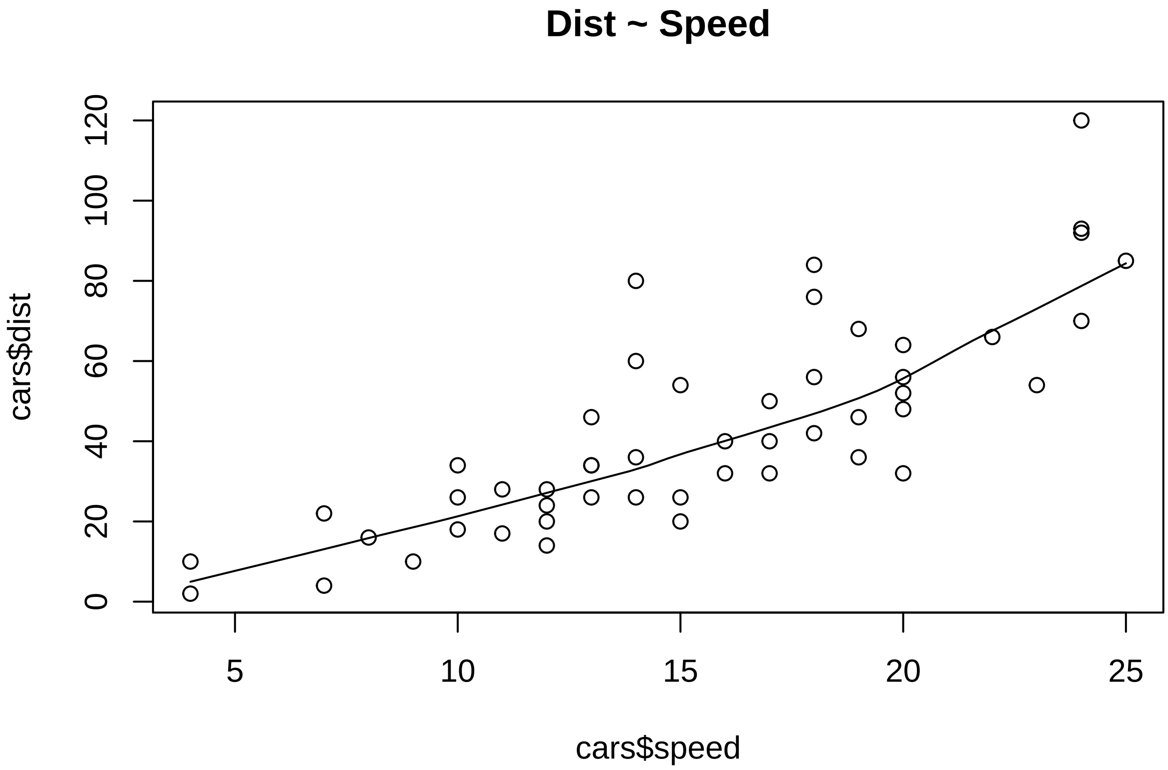


```
plot(cars$speed, cars$dist, xlab = "Car Speed at Stopping (in mph)",
     ylab = "Stopping Distance (in feet)", main = "The Effect of Car Speed on Stopping Distance")
```



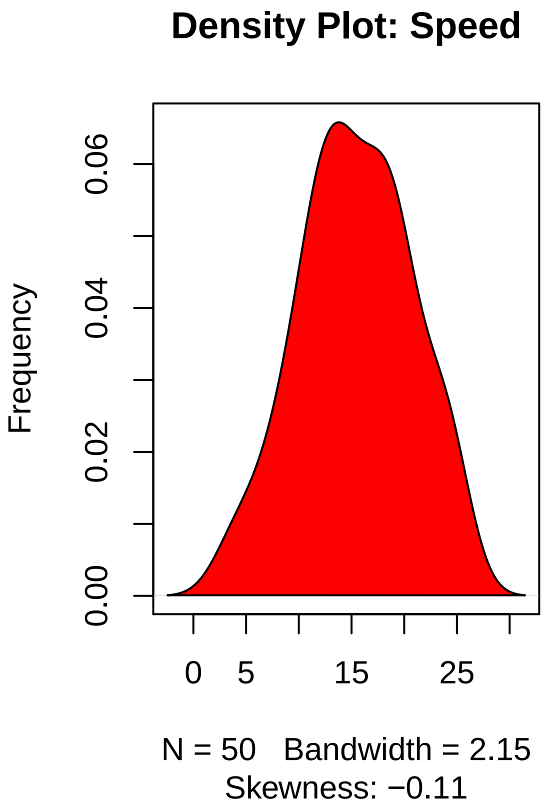
Fitting a smooth line

```
scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~ Speed")
```



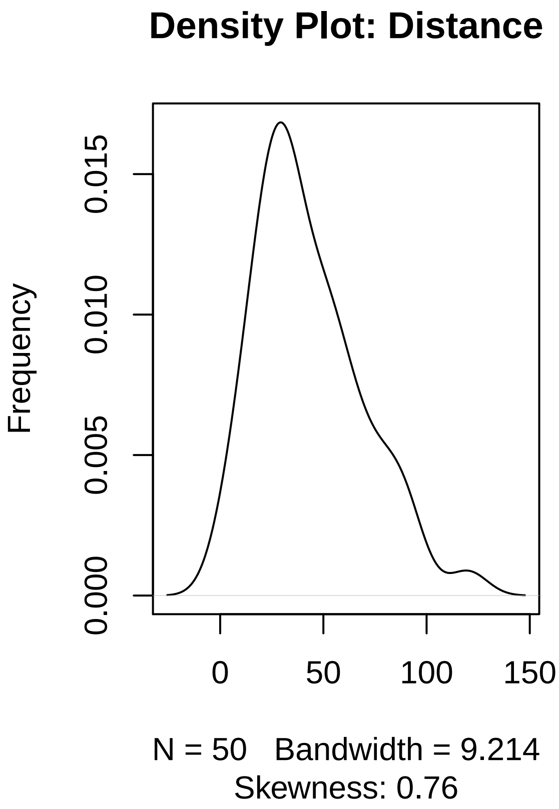
Density plots for speed and distance

```
library(e1071)
par(mfrow=c(1, 2))

plot(density(cars$speed), main="Density Plot: Speed", ylab="Frequency", sub=paste("Skewness:", round(e10
polygon(density(cars$speed), col="red")

plot(density(cars$dist), main="Density Plot: Distance", ylab="Frequency", sub=paste("Skewness:", round(e
```



Density Plot: Speed

N = 50 Bandwidth = 2.15
Skewness: −0.11



Density Plot: Distance

N = 50 Bandwidth = 9.214
Skewness: 0.76

Linear regression model fitting

```
carmod <- lm(dist ~ speed, data = cars)
summary(carmod)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

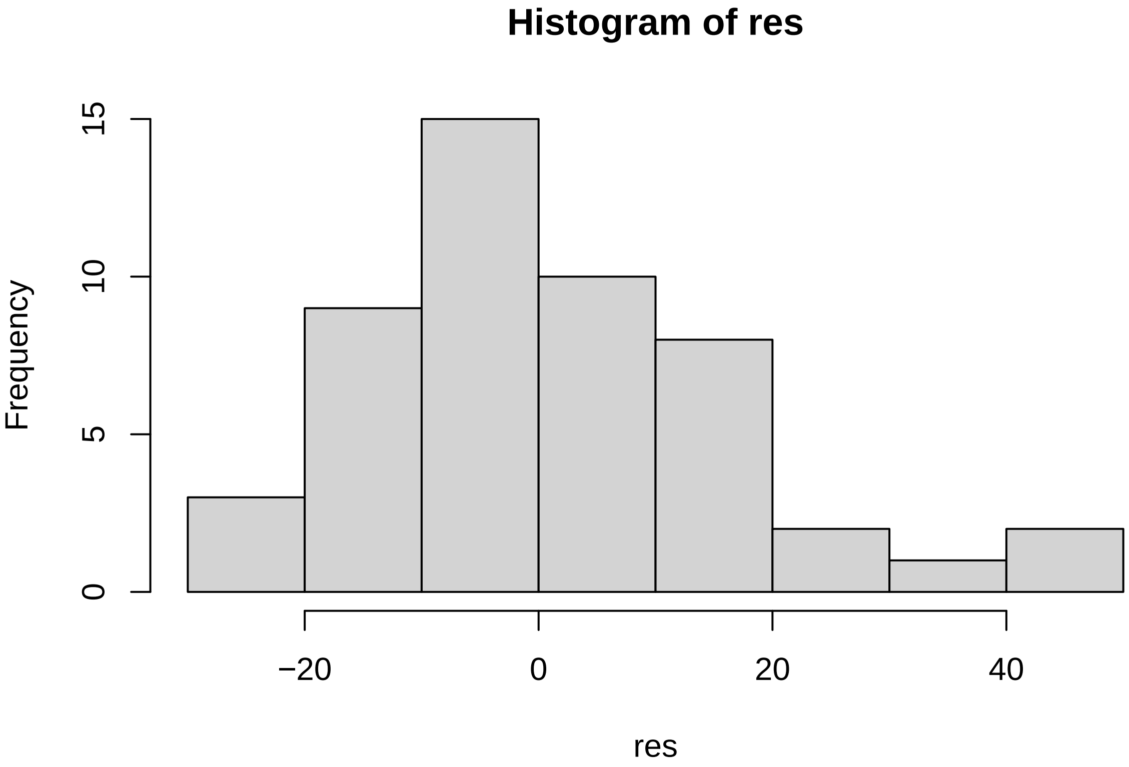
95% CI

```
confint(carmod, level = 0.95)
```

```
##                  2.5 %    97.5 %
## (Intercept) -31.167850 -3.990340
## speed         3.096964  4.767853
```

Normality of residuals check

```
res = carmod$residuals
hist(res)
```



**Interpretation**

The coefficients in linear regression model states that with a unit change in x how much change is expected in y.

## Session 2: Data Science & Sample Survey

### Prof. G. N. Singh, IIT (ISM) Dhanbad

Word Statistics

In a literal sense

Plural sense some sort of data numerical figures in our day to day arising, runs and all figures are called statistics

In singular collection of methods and principles in a book,

Procedure to collection, analyse and interpret the data is called statistics

Statistics never claims 100% accuracy

Statistics is the science of decision making. As no decision is free from error.

Hope that PPTs will be provided soon.