

Simple Relationship Analysis using RStudio

Course Taught at SUAD

Dr. Tanujit Chakraborty
Ms. Madhurima Panja (TA)

@ Sorbonne
tanujitisi@gmail.com

This presentation includes...

- Introduction to Relationship Analysis
- Regression Analysis
 - Simple Linear regression
- Practical Implementation using R
 - Engineering Data
 - Advertisement Data

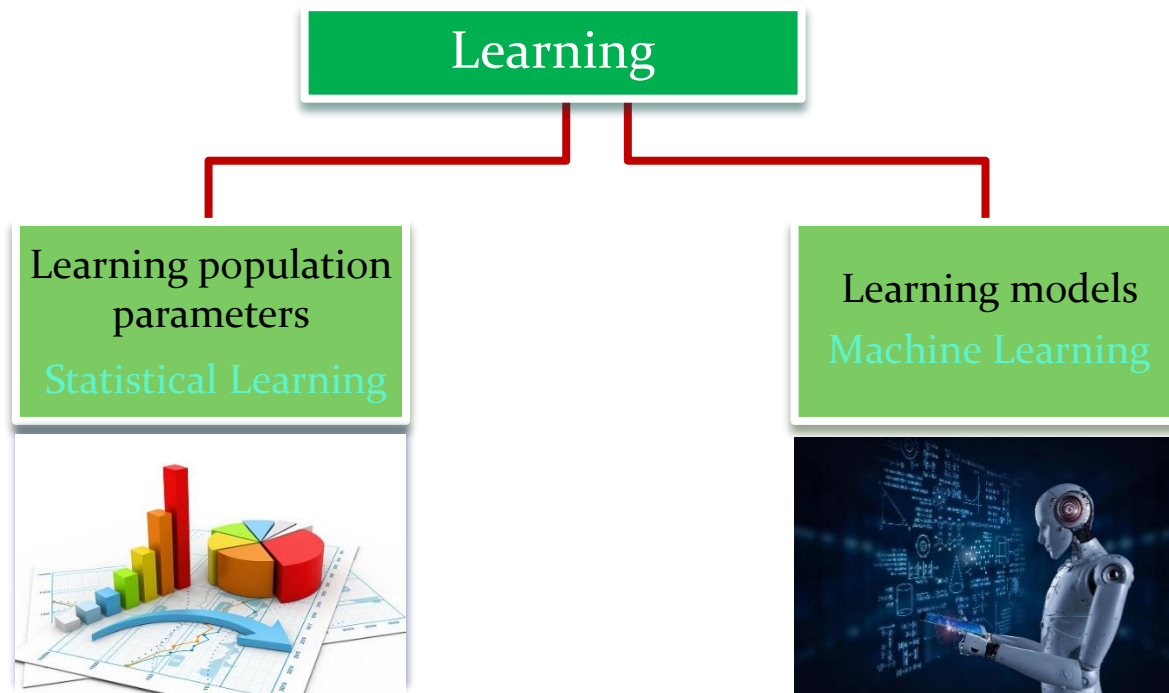
Regression Analysis



Learning Strategies







There are two types of learning concepts:

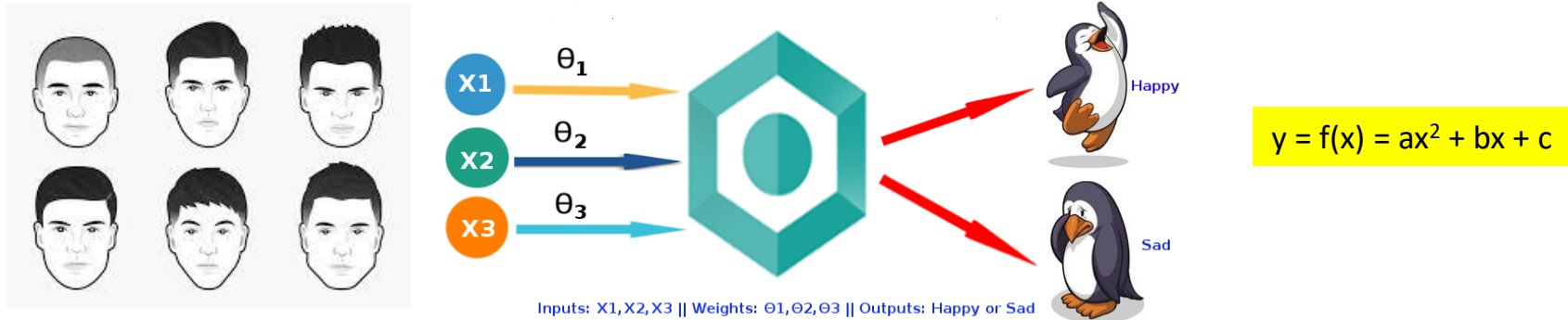


Statistical Learning

Usually assumes certain properties of the population from which we draw samples:

-  Observation come from a normal population.
-  Sample size is small.
-  Population parameters like mean, variance, etc. are hold good.
-  Requires measurement equivalent to interval scaled data.

Machine Learning



- Does not under any assumption
- Works well with high volume high dimensional data

Important Point

This learning strategy needs a very large sample data



Relationship Analysis

Relationship Analysis

- **Example: Wage Data**

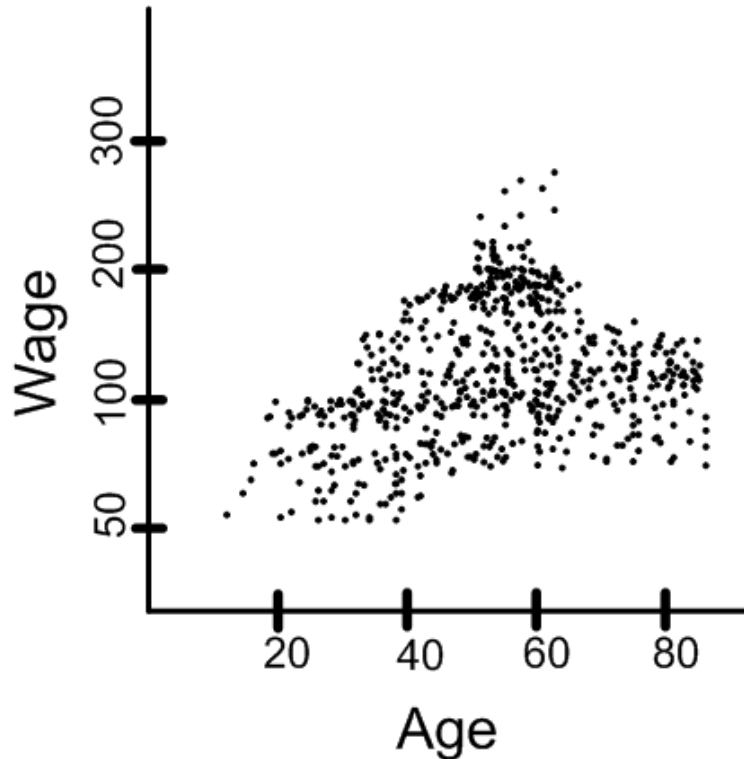
A large data regarding the wages for a group of employees from the eastern region of India is given.

In particular, we wish to understand the following relationships:

- *Employee's age and wage:* How wages vary with ages?
- *Calendar year and wage:* How wages vary with time?
- *Employee's age and education:* Whether wages are anyway related with employees' education levels?

Relationship Analysis

- Example: Wage Data
 - Case I. Wage versus Age
 - From the data set, we have a graphical representations, which is as follows:

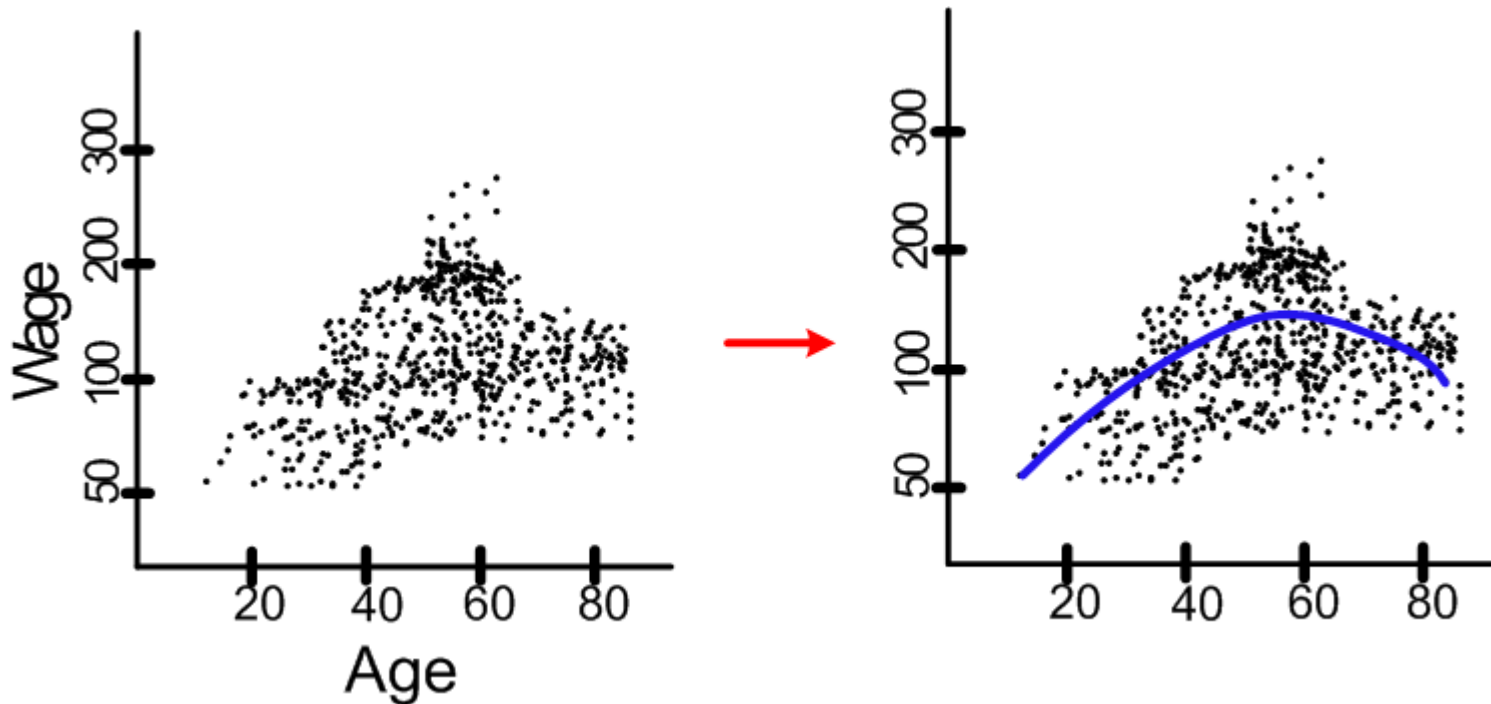


?

How wages vary with ages?

Relationship Analysis

- Example: Wage Data
 - *Employee's age and wage: How wages vary with ages?*



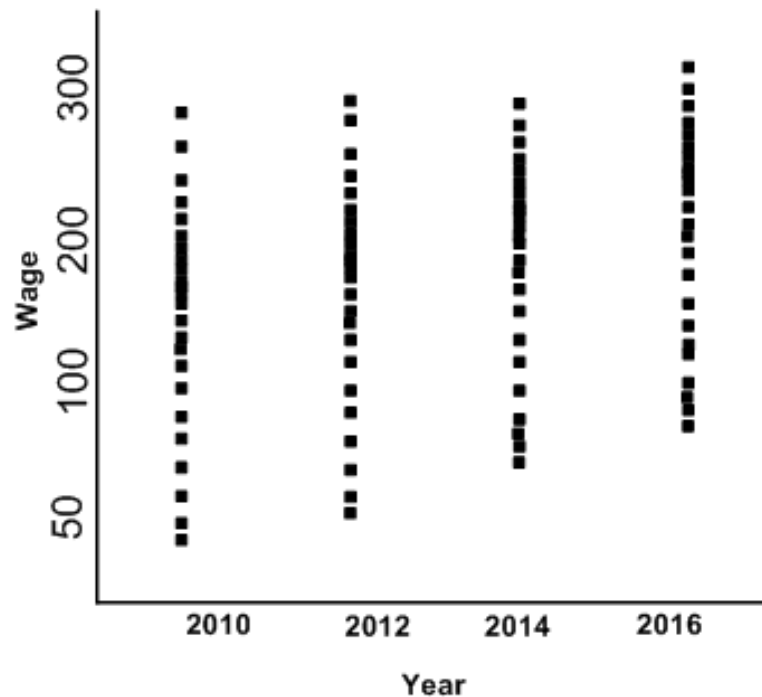
Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

Relationship Analysis

- Example: Wage Data

- Case II. Wage versus Year

- From the data set, we have a graphical representations, which is as follows:



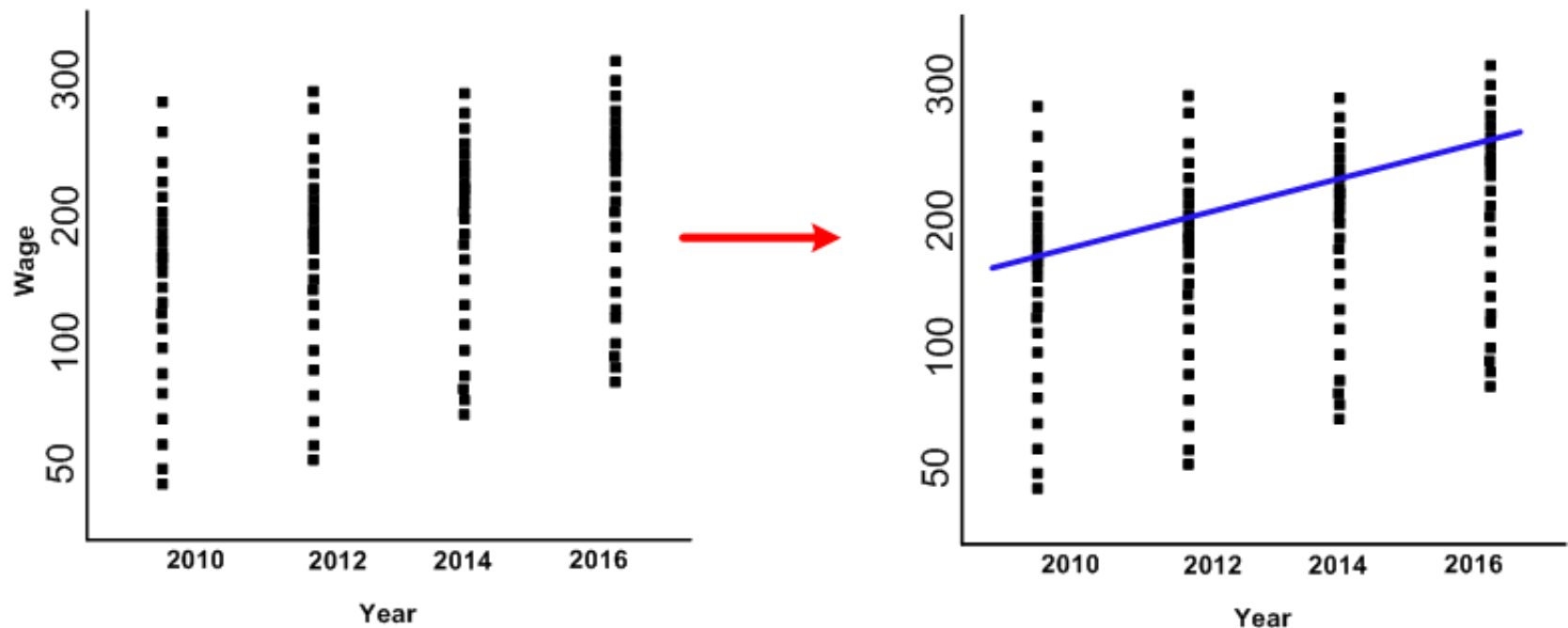
?

How wages vary with time?

Relationship Analysis



- Example: Wage Data
 - *Wage and calendar year: How wages vary with years?*

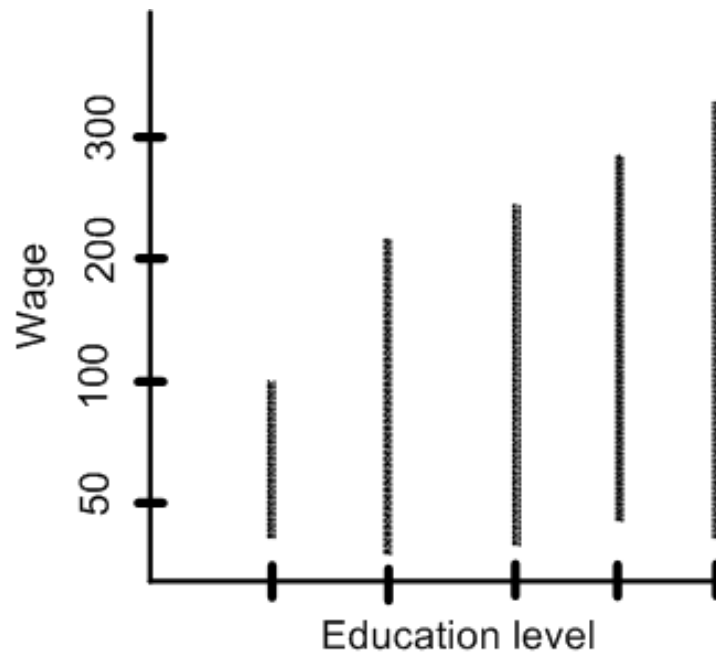


Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

Relationship Analysis



- Example: Wage Data
 - Case III. Wage versus Education
 - From the data set, we have a graphical representations, which is as follows:

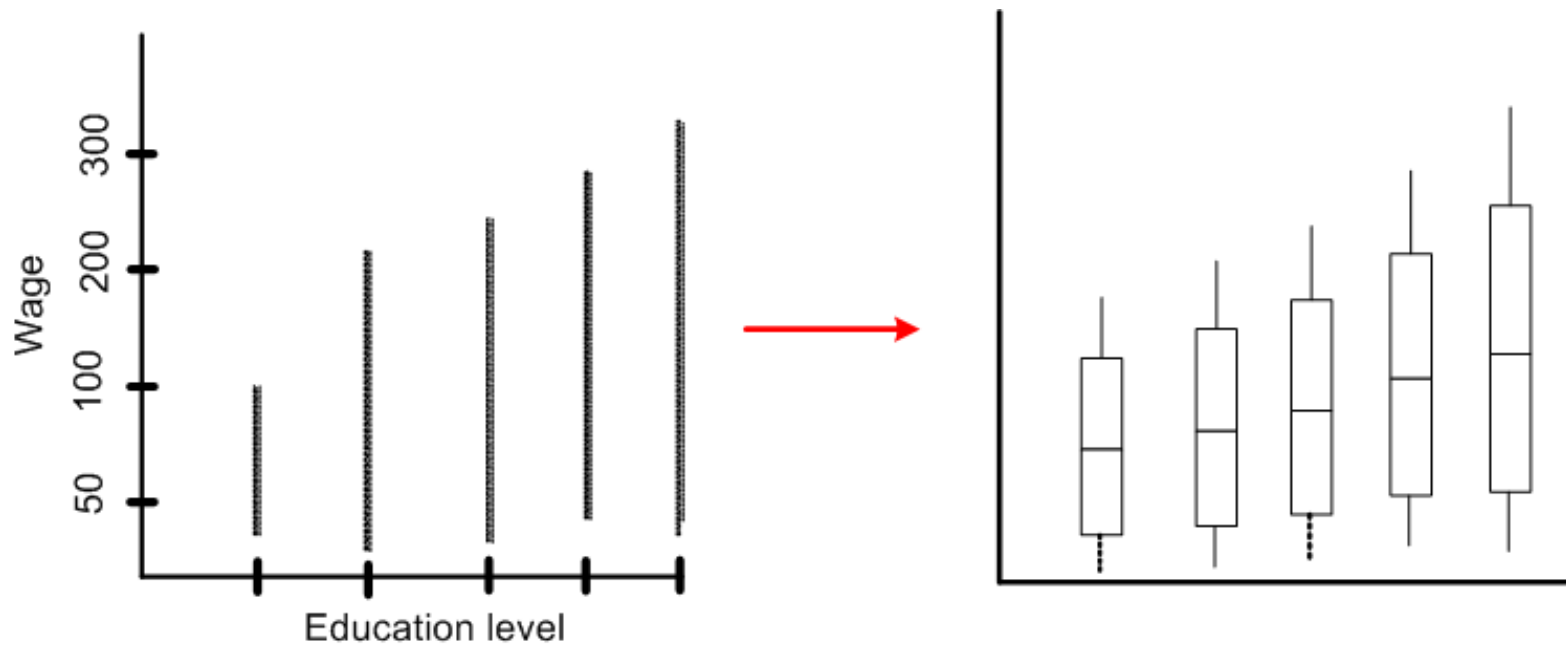


?

Whether wages are related with education?

Relationship Analysis

- Example: Wage Data
 - *Wage and education level: Whether wages vary with employees' education levels?*



Interpretation: On the average, wage increases with the level of education.

Relationship Analysis



What more information can we get?

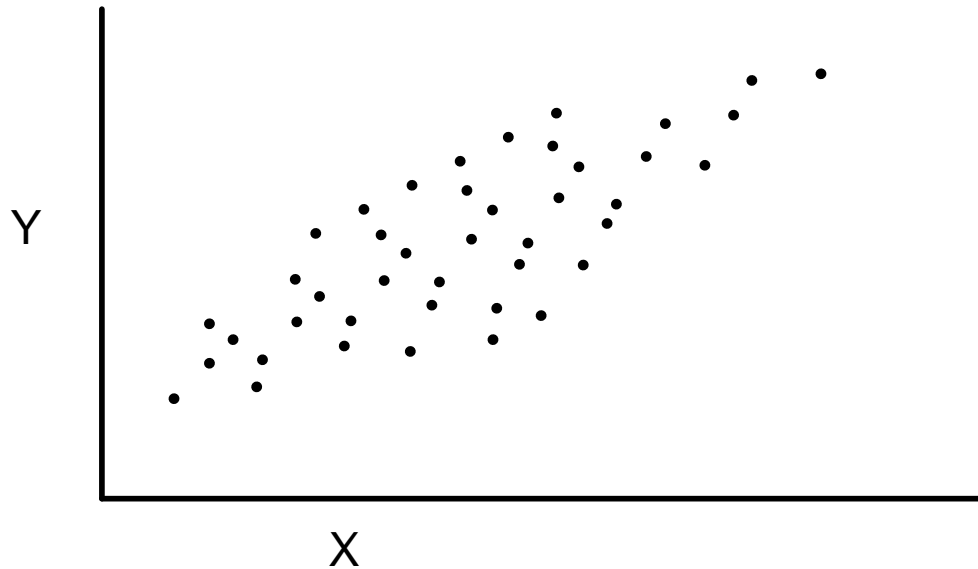
Whether wage has
any association
with both year and
education level?

Given an employee's
wage can we predict
his age?



... and what's more?

A curious Question!

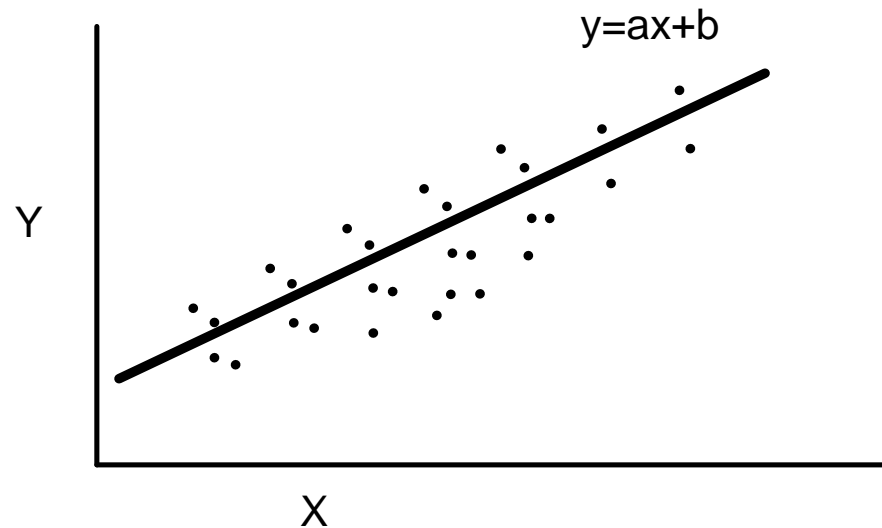


Suppose there are countably infinite points in the XY plane. We need a huge memory to store all such points.

Is there any way out to store this information with a least amount of memory?

Say, with two values only.

Yahoo!



Just decide the values of **a** and **b**
(as if storing one point's data only!)

Note: Here, the trick was to find a relationship among all the points.

Measures of Relationship



Univariate Population

<i>Temperature</i>	20	30	21	18	23	45	52
--------------------	----	----	----	----	----	----	----



Bivariate Population

<i>Temperature</i>	20	30	21	18	23	45	52
<i>Pressure</i>	1	1.5	1.05	0.96	1.2	2.5	2.8

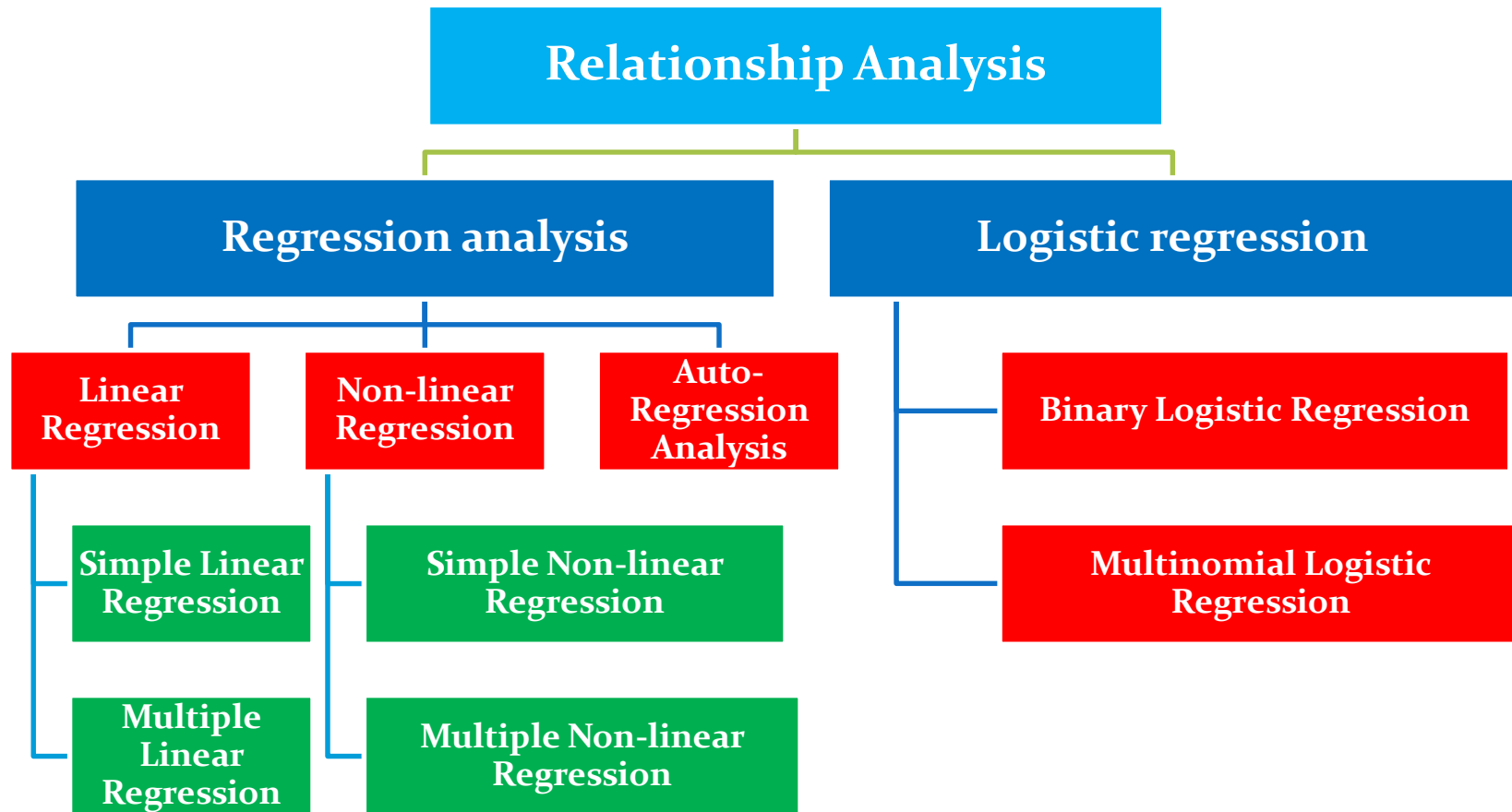


Multivariate Population

<i>Temperature</i>	20	30	21	18	23	45	52
<i>Pressure</i>	1	1.5	1.05	0.96	1.2	2.5	2.8
<i>Volume</i>	20	30	21	18	23	45	52



Measures of Relationship



Regression Analysis

Definition

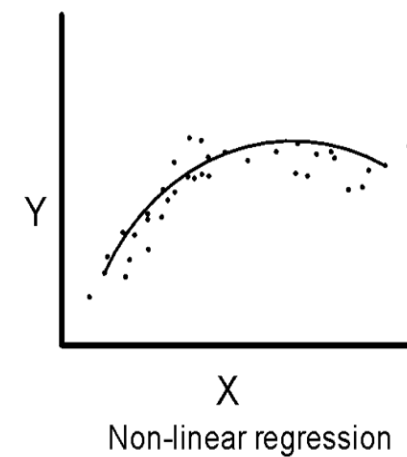
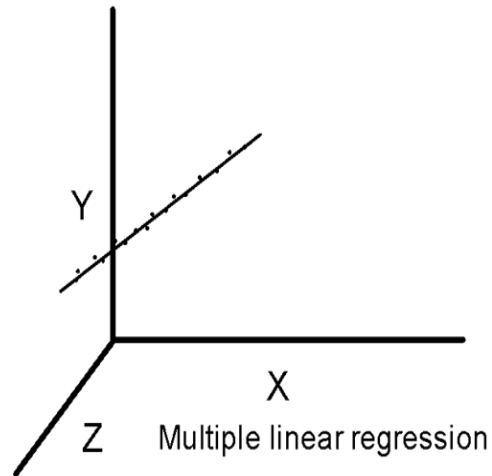
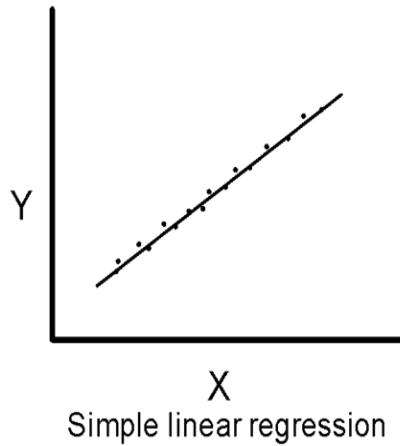
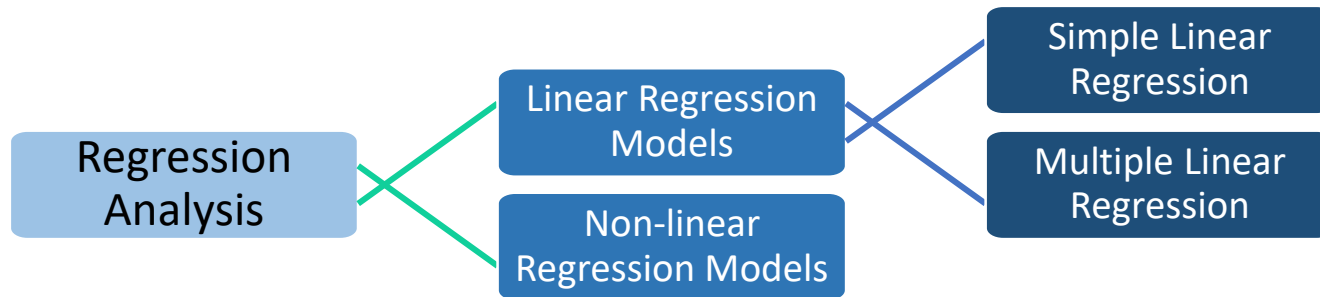
The regression analysis is a statistical method to deal with the formulation of mathematical model depicting relationship amongst variables, which can be used for the purpose of prediction of the values of **dependent variable**, given the values of **independent variable(s)**.

Example

How Exam Score is related to Hours of Study?

<i>Hours Study</i>	<i>Exam Score</i>
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100

Regression Analysis





Simple Linear Regression

Simple Linear Regression Model

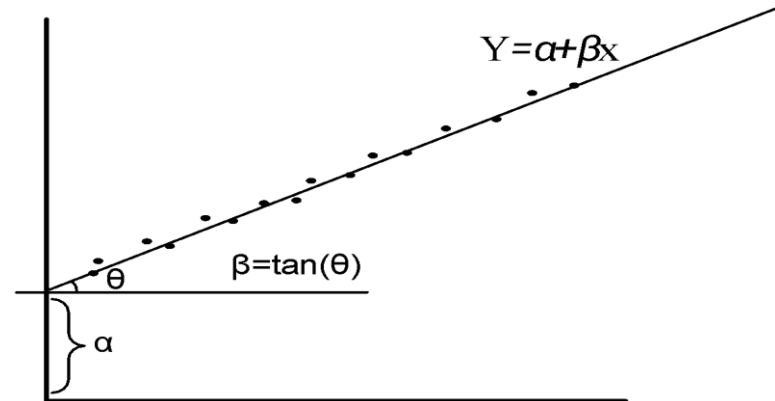
In simple linear regression, we have only two variables:

Dependent variable:

Also called *Response*, usually denoted as Y

Independent variable:

Also called *Regressor*, usually denoted as x



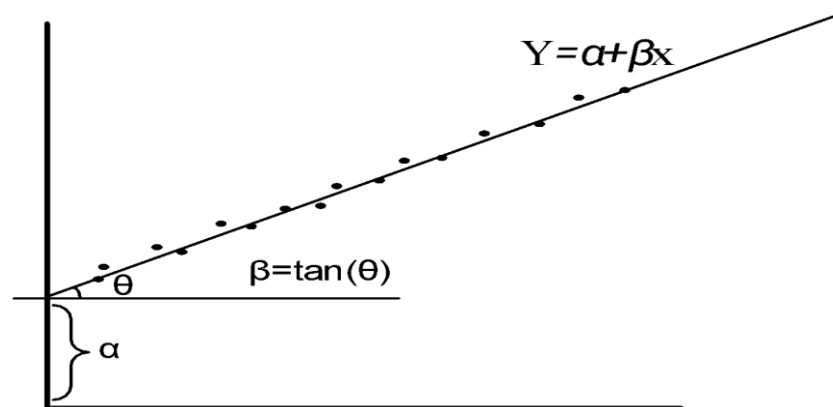
Linear regression

A reasonable form of a relationship between the Response Y and the Regressor x is the linear relationship, that is in the form $Y = \alpha + \beta x$

Simple Linear Regression Model



In simple linear regression, we have only two variables:



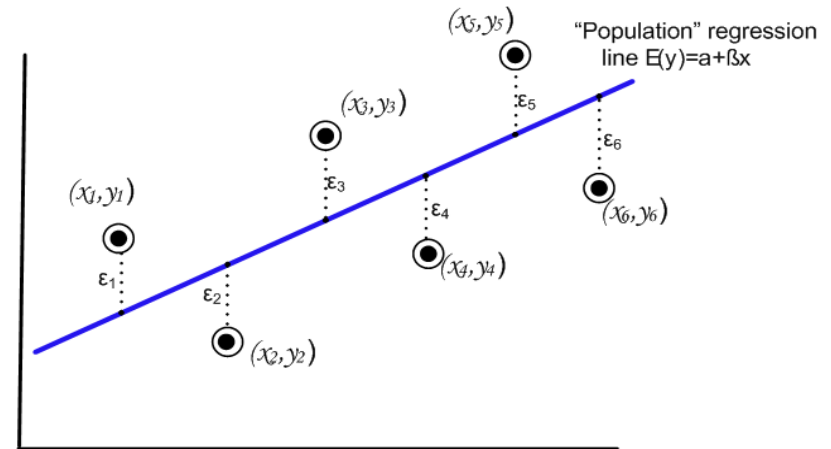
Note

- There are infinite number of lines (and hence α_s and β_s)
- The concept of regression analysis deal with finding the best relationship between Y and x (and hence best fitted values of α and β) quantifying the strength of that relationship.

Regression Analysis

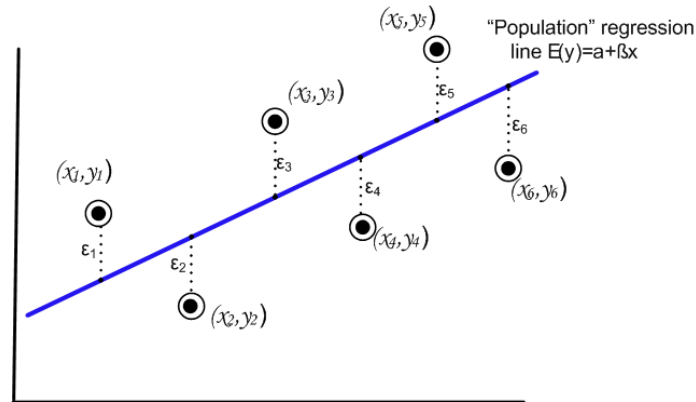
Given: The set $[(x_i, y_i), i = 1, 2, 3, \dots, n]$ of data involving n pairs of (x, y) values

Objective: To find “true” or population regression line, such that $Y = \alpha + \beta x + \epsilon$





Here, ϵ is a random variable with $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. The quantity σ^2 is often called the **error variance**.

Regression Analysis



Note

- 
 $E(\epsilon) = 0$ implies that at a specific x , the y values are distributed around the “true” regression line $Y = \alpha + \beta x$ (i.e., the positive and negative errors around the true line is reasonable).
- 
The values of the regression coefficients α and β to be estimated from data

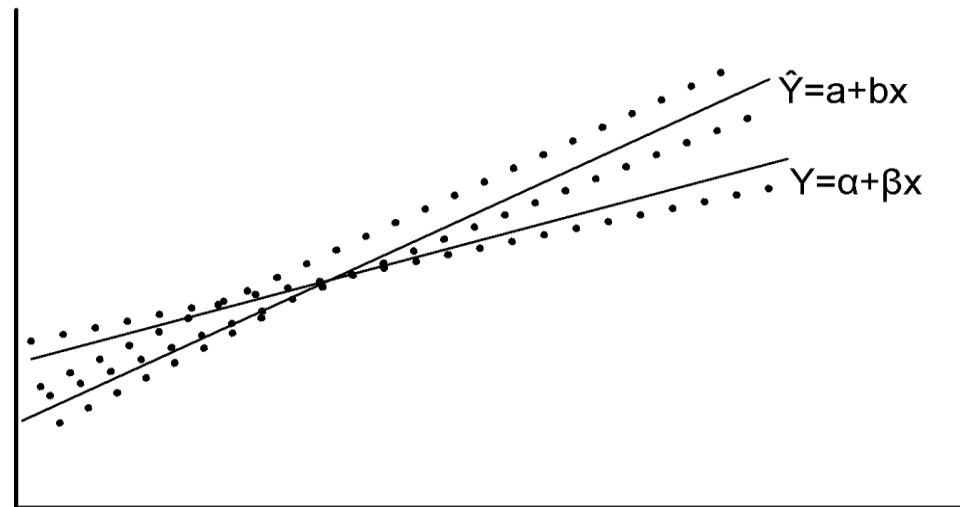
True versus Fitted Regression Line

📍 The task in regression analysis is to estimate the regression coefficients α and β .

📍 Suppose, we denote the estimates a for α and b for β . Then the fitted regression line is

$$\hat{Y} = a + bx$$

where, \hat{Y} is the predicted or fitted value



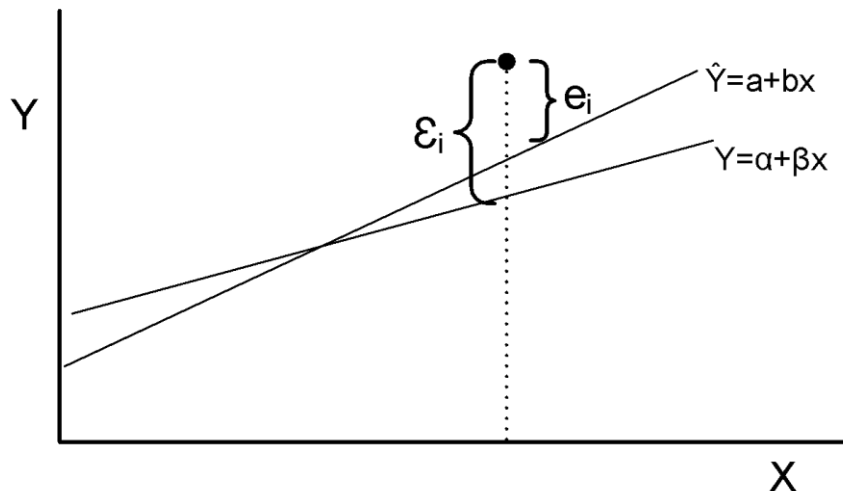
Least Square Method to estimate α and β



Concept of Residuals

This method uses the concept of residual. A residual is essentially an error in the fit of the model $\hat{Y} = a + bx$. Thus, i^{th} residual is

$$e_i = Y_i - \hat{Y}_i, i = 1, 2, 3, \dots, n$$



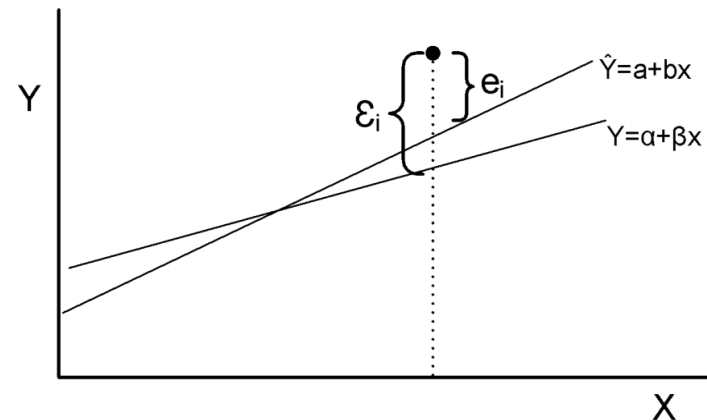
Least Square Method to estimate α and β

Sum of Squares Error (SSE)

The residual sum of squares is often called **the sum of squares of the errors** about the fitted line and is denoted as

SSE

$$\begin{aligned} SSE &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - a - bx_i)^2 \end{aligned}$$



We need to **minimize the value of SSE** and hence to determine the parameters of a and b .

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 1: Differentiation

Differentiating SSE with respect to a and b , we have

$$\frac{\partial(SSE)}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i)$$

$$\frac{\partial(SSE)}{\partial b} = -2 \sum_{i=1}^n (y_i - a - bx_i) \cdot x_i$$

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$, and $\frac{\partial(SSE)}{\partial b} = 0$

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$, and $\frac{\partial(SSE)}{\partial b} = 0$

Thus we get,

$$na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Least Square Method to estimate α and β

Minimizing the Sum of Squares Error (SSE)

Step 2: Equating the partial derivatives to zero

For minimum value of SSE, $\frac{\partial(SSE)}{\partial a} = 0$,
and $\frac{\partial(SSE)}{\partial b} = 0$

Thus we get,

$$na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Step 3: Solving for a and b

These two equations on the left can be solved to determine the values of a and b , and it can be calculated that

$$b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

Galton Board



- [Sir Francis Galton](#), Charles Darwin's half-cousin, invented the 'Galton Board' in 1874 to demonstrate that the normal distribution is a natural phenomenon.
- It specifically shows that the binomial distribution approximates a normal distribution with a large enough sample size.



**R^2 : Measure of Quality of
the Fitting**

R²: Measure of Fit Quality

Coefficient of Determination

A quantity R^2 , is called **coefficient of determination** is used to measure the proportion of variability of the fitted model.

Total corrected sum of squares:

We have $SSE = \sum_{i=1}^n (y_i - \hat{y})^2$. It signifies the **variability due to error**.

Now, the **total corrected sum of squares** is defined as $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

R^2 :

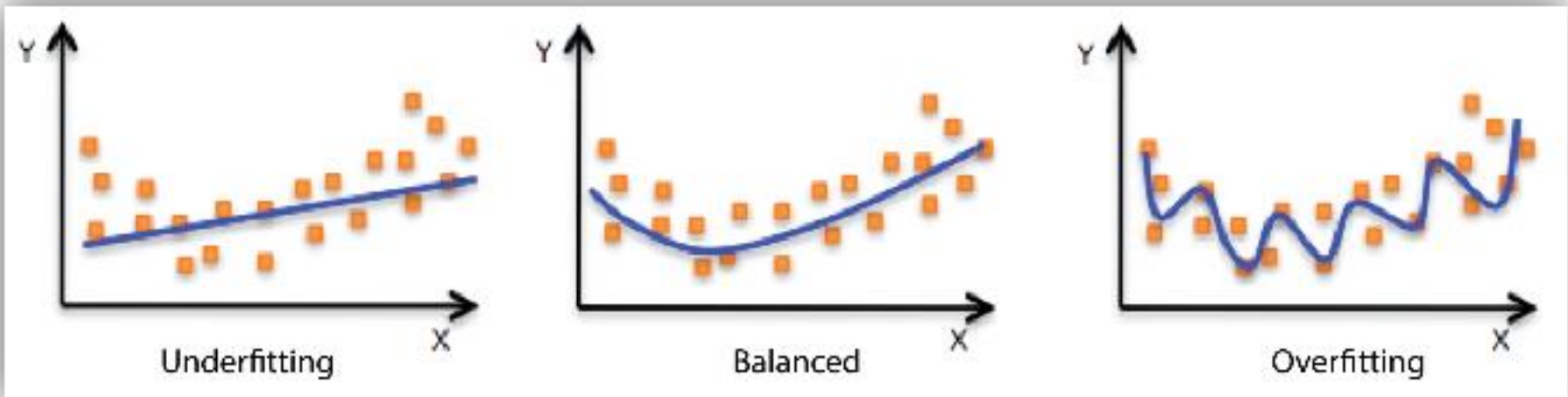
SST represents the variation in the response values. The R^2 is: $R^2 = 1 - \frac{SSE}{SST}$

R²: Measure of Quality Fit

Coefficient of Determination

Note

- ☺ If fit is perfect, all residuals are zero and thus $R^2 = 1.0$ (very good fit)
- ☺ If SSE is only slightly smaller than SST, then $R^2 \approx 0$ (very poor fit)





Simple Linear Regression in RStudio

Regression

- Correlation helps
 - To check whether two variables are related
- If related
 - Identify the type & degree of relationship
- Regression helps
 - To identify the exact form of the relationship
 - To model output in terms of input or process variables



Practical Problem

Exercise1: The data from the pulp drying process is given in the file DC_Simple_Reg.csv. The file contains data on the dry content achieved at different dryer temperature. Develop a prediction model for dry content in terms of dryer temperature.

Path to Solution

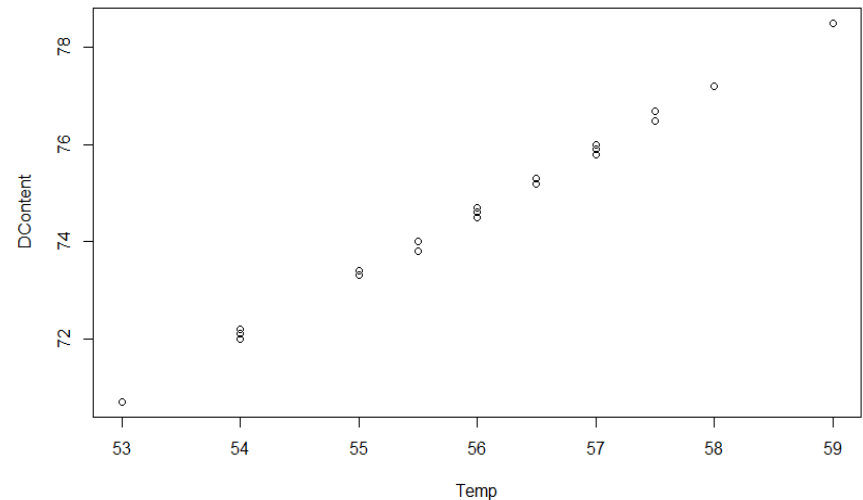
1. Reading the data and variables

```
> mydata = read.csv('DC_Simple_Reg.csv',header = T,sep = ",")  
> Temp = mydata$Dryer.Temperature  
> DContent = mydata$Dry.Content
```

2. Checking Correlation

Scatter Plot

```
> plot(Temp, DContent)
```



Path to Solution

2. Computing Correlation

```
> cor(Temp, DContent)
```

Attribute	Dry Content
Temperature	0.9992

Remark:

Correlation between y & x need to be high (preferably 0.8 to 1 to -0.8 to -1.0)

Path to Solution

3. Performing Regression

```
> model = lm(DContent ~ Temp)
> summary(model)
```

Statistic	Value	Criteria
Residual standard error	0.07059	
R-squared	0.9984	> 0.6
Adjusted R-squared	0.9984	> 0.6

Model	df	F	p value
Regression	1	2449 7	0.000
Residual	40		
Total	41		

Criteria:
P value < 0.05

Path to Solution

3. Performing Regression

Attribute	Coefficient	Std. Error	t Statistic	p value
Intercept	2.183813	0.463589	4.711	0.00
Temperature	1.293432	0.008264	156.518	0.00

Interpretation

The p value for independent variable need to be < significance level α (generally $\alpha = 0.05$)

Model: $\text{Dry Content} = 2.183813 + 1.293432 \times \text{Temperature}$

Path to Solution

4. Regression Anova

```
> anova(model)
```

ANOVA					
Source	SS	df	MS	F	p value
Temp	122.057	1	122.057	24497	0.000
Residual	0.199	40	0.005		
Total	122.256	41			

Criteria: P value < 0.05

Path to Solution

5. Residual Analysis

```
> pred = fitted(model)
> Res = residuals(model)
> write.csv(pred,"D:/SUAD/DataSets/Pred.csv")
> write.csv(Res,"D:/SUAD/DataSets/Res.csv")
```

SL No.	Fitted	Residuals	SL No.	Fitted	Residuals
1	73.32259	-0.02259	22	74.61602	-0.01602
2	74.61602	-0.01602	23	75.26274	-0.06274
3	73.96931	0.030693	24	73.96931	0.030693
4	78.49632	0.00368	25	75.90946	-0.00946
5	74.61602	-0.01602	26	75.26274	0.03726
6	73.96931	0.030693	27	73.96931	0.030693
7	75.26274	-0.06274	28	78.49632	0.00368
8	77.20289	-0.00289	29	76.55617	-0.05617
9	75.90946	-0.00946	30	74.61602	-0.11602
10	74.61602	-0.01602	31	75.90946	0.090544
11	73.32259	-0.02259	32	76.55617	-0.05617
12	75.90946	-0.00946	33	76.55617	0.143828
13	75.90946	0.090544	34	75.90946	0.090544
14	74.61602	-0.01602	35	75.90946	-0.10946
15	74.61602	0.083977	36	73.96931	-0.16931
16	74.61602	-0.11602	37	73.32259	-0.02259
17	70.73573	-0.03573	38	74.61602	-0.01602
18	72.02916	-0.02916	39	73.32259	0.077409
19	72.02916	0.070841	40	75.90946	0.090544
20	72.02916	0.170841	41	73.96931	0.030693
21	70.73573	-0.03573	42	75.26274	-0.06274

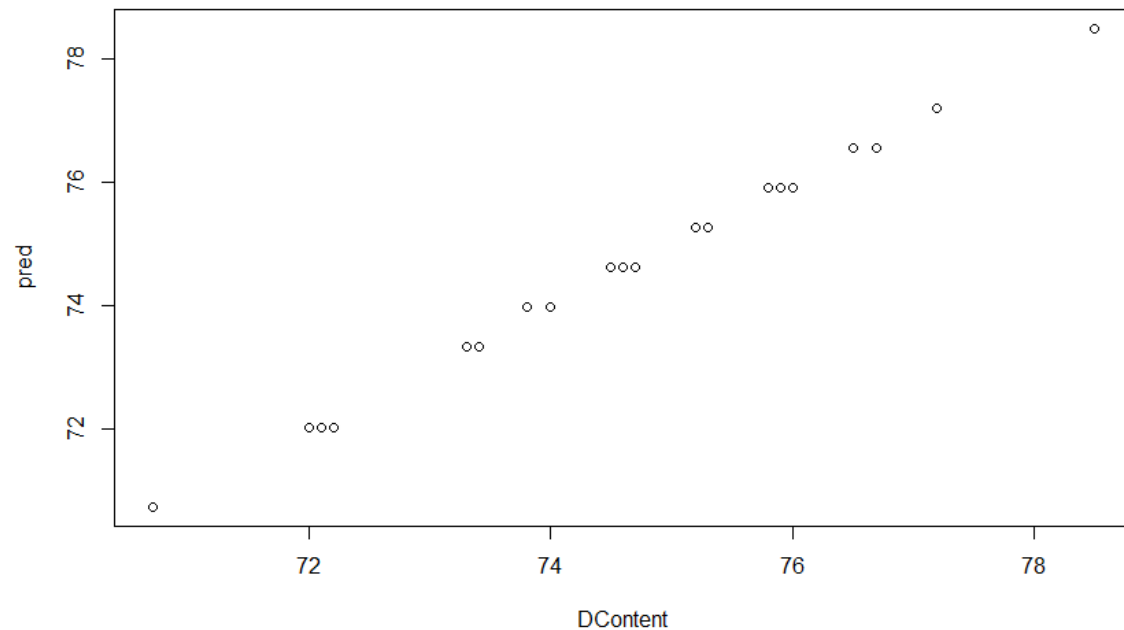
Path to Solution



5. Residual Analysis

Scatter Plot: Actual Vs Predicted (fit)

```
> plot(DContent, pred)
```



Path to Solution

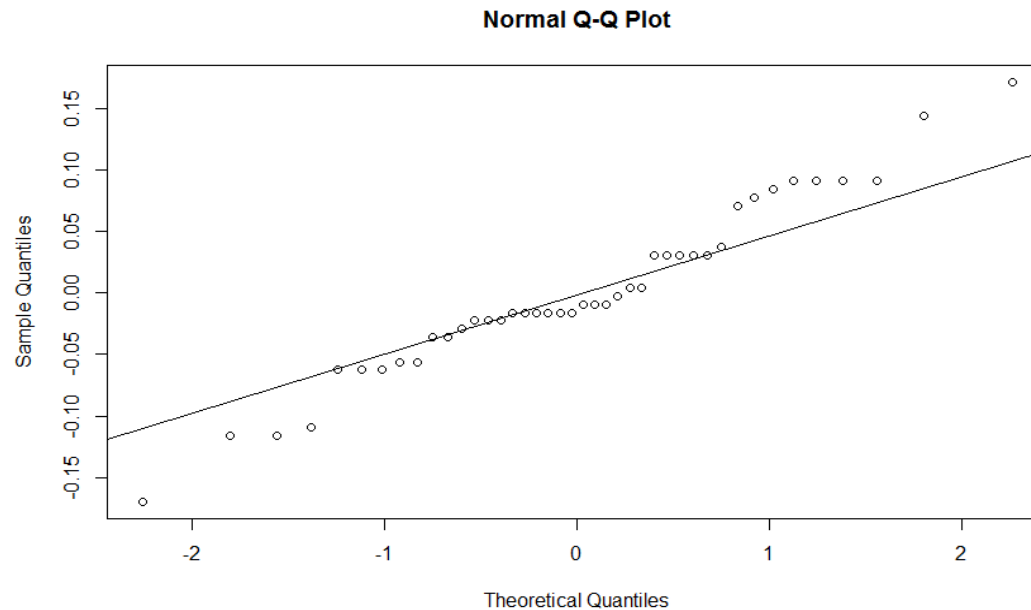


5. Residual Analysis

Normality Check on residuals

```
> qqnorm(Res)
```

```
> qqline(Res)
```



Residuals should be normally distributed or bell shaped

Path to Solution

5: Residual Analysis

Normality Check on residuals

```
> shapiro.test(Res)
```

Shapiro-Wilk normality Test:	
W	p value
0.9693	0.3132

Residuals should be normally distributed or bell shaped



Path to Solution

5: Residual Analysis

```
> plot(pred, Res)
```

```
> plot(Temp, Res)
```

Residuals should be independent and stable

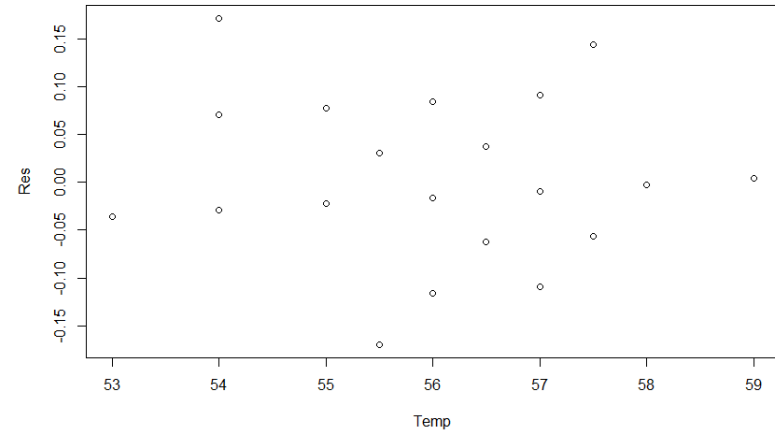
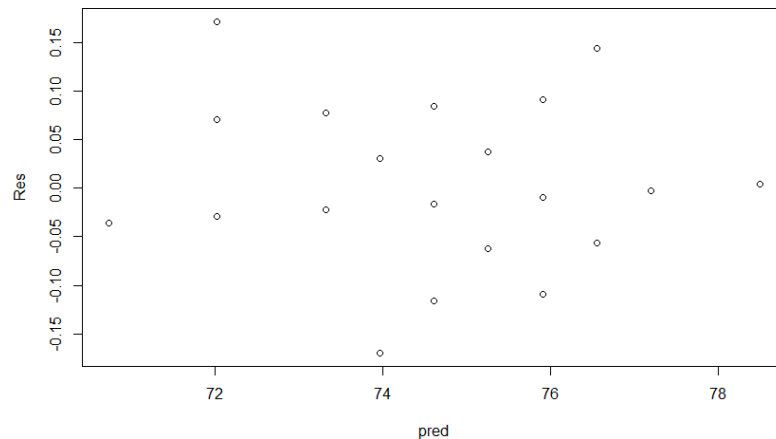
Plot the residuals against fitted value. The points in the graph should be scattered randomly and should not show any trend or pattern. The residuals should not depend in anyway on the fitted value.

If there is a pattern then a transformation such as $\log y$ or \sqrt{y} to be used

Similarly the residuals shall not depend on x . This can be checked by plotting residuals vs x . A pattern in this plot is an indication that the residuals are not independent of x . Instead of x , develop the model with a function of x as predictor (Eg: x^2 , $1/x$, \sqrt{x} , $\log(x)$, etc.)

Path to Solution

Residual Analysis



There is no trend or pattern on residuals vs fitted value ,residuals vs observation order or residuals vs x plot. Hence the assumptions of independence and stability of residuals are satisfied.

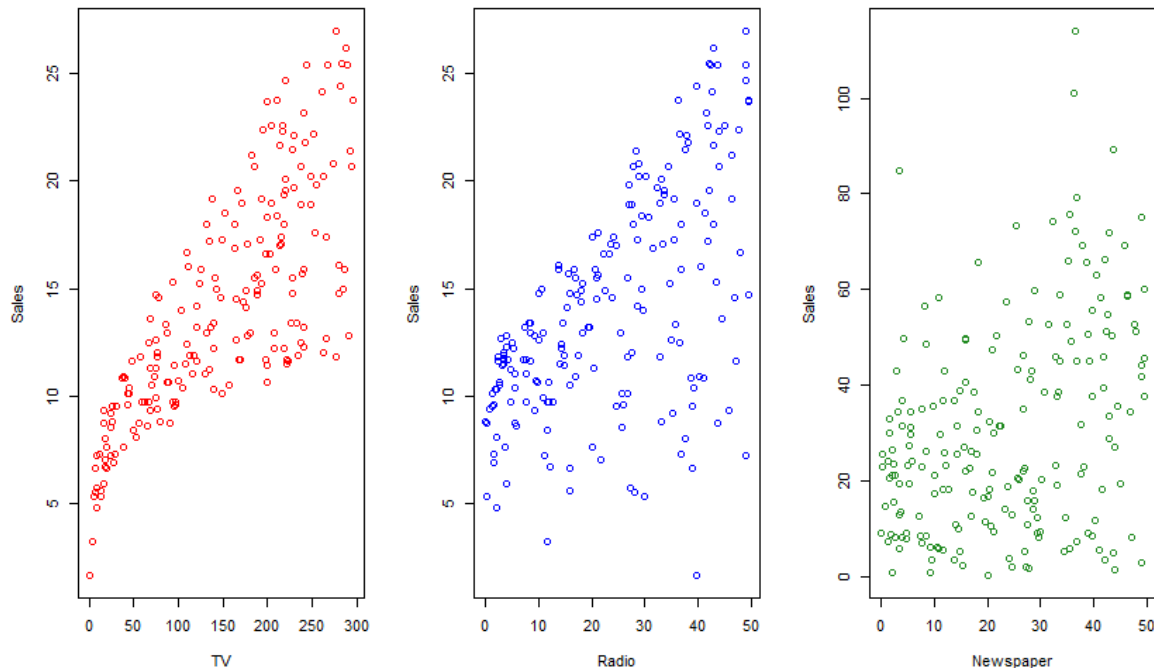
Another Motivating Example

- Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product.
- The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.
- It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media.
- Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales.
- In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.
- In this example, the variables TV, radio and newspaper are the feature vectors, also going by the names of predictors, covariates, independent variables, or just variables.
- The output variable, sales is the label, or the dependent variable, also called the response variable.

Let's begin with plotting the data

```
dat <- read.csv("Advertising.csv", header = TRUE)
```

```
par(mfrow=c(1,3)) # this creates three plots in a single row
plot(dat$TV, dat$sales, xlab = "TV", ylab = "Sales", col = "red")
plot(dat$radio, dat$sales, xlab="Radio", ylab="Sales", col="blue")
plot(dat$radio, dat$newspaper, xlab="Newspaper", ylab="Sales", col="forestgreen")
```



Fitting Linear Regression Model

```
dat <- as_tibble(dat) %>% select(-X) # X represents serial number hence can be removed
# Split the data into training and testing set
dat.split <- resample_partition(dat, c(test = 0.3, train = 0.7))
train <- as_tibble(dat.split$train)
test <- as_tibble(dat.split$test)
# Time for SLR model fitting
mod1 <- lm(sales ~ TV, data = train)
# Summary
```

```
summary(mod1)
## Call:
## lm(formula = sales ~ TV, data = train)
##
## Residuals:
##   Min     1Q   Median     3Q    Max
## -7.9082 -1.9914 -0.0741  2.0400  7.5511
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.152113   0.537792   13.30  <2e-16 ***
## TV           0.045378   0.003099   14.64  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.211 on 139 degrees of freedom
## Multiple R-squared:  0.6067, Adjusted R-squared:  0.6039
## F-statistic:  214.4 on 1 and 139 DF, p-value: < 2.2e-16
```

Fitting Linear Regression Model

Tidy summary (available in 'broom' package)

```
library(broom)
```

```
tidy(mod1)
```

A tibble: 2 × 5

##	term	estimate	std.error	statistic	p.value
##	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
##	1 (Intercept)	7.15	0.538	13.3	1.50e-26
##	2 TV	0.0454	0.00310	14.6	5.86e-30

We can also compute interval estimates of the coefficients.

```
confint(mod1)
```

##	2.5 %	97.5 %
## (Intercept)	6.08880236	8.21542290
## TV	0.03925107	0.05150479

Interpretation of coefficients

Checking Model accuracy

- We now explore goodness-of-fit of the model on the data.
- Quantitative assessment of the same may be achieved through the following measures:
 - Mean Squared Error – estimate of error variance
 - R^2 - proportion of variability explained by the predictor variable
 - F-statistic

```
MSE <- mse(mod1, train)
```

```
MSE
```

```
## [1] 10.16292
```

```
rsquare(mod1, train)
```

```
## [1] 0.6067217
```

In case of simple linear regression, R^2 is exactly equal to the square of the correlation between the response variable y and predictor variable X .

```
r = cor(train$sales, train$TV)
```

```
r^2
```

```
## [1] 0.6067217
```

Model Accuracy

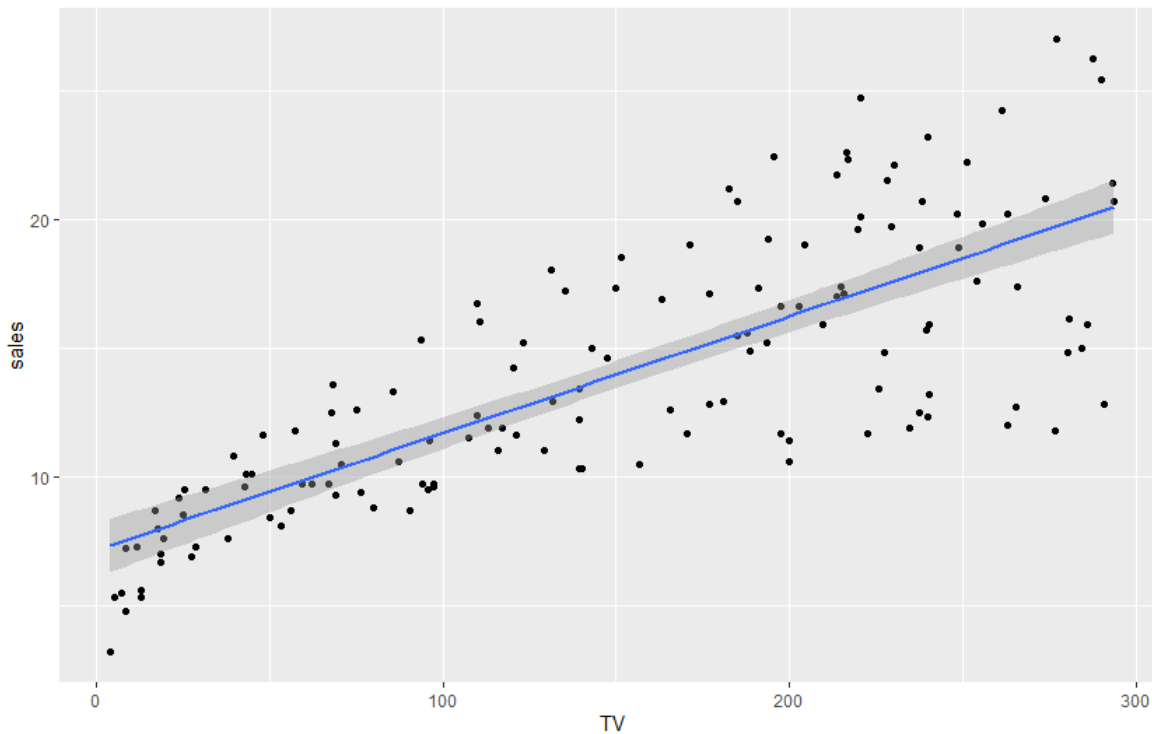
Checking Model accuracy

- The F-statistic provides an overall measure of model fit by assessing whether at least one of the predictors in the model is significant, that is, has non-zero coefficient.
- In the above example, the value of the F-statistic has an extremely small p-value indicating that there is strong evidence that at least one of the predictors (in this case the variable TV itself) in the model is significant.

Assessment of Model Fit

We can visually assess the model fit by including the fitted linear regression line on the scatter plot.

```
ggplot(train, aes(x = TV, y = sales)) + geom_point() + geom_smooth(method = "lm")  
## `geom_smooth()` using formula 'y ~ x'
```



Predictions

We can use the fitted model to make our predictions for new values of the predictor variable(s).

Note that we have a test data where we can assess the accuracy of predictions as well.

This is achieved via the *add_predictions* command.

```
test %>% add_predictions(mod1)
```

```
## A tibble: 59 × 5
##   TV radio newspaper sales pred
##   <dbl> <dbl>   <dbl> <dbl> <dbl>
## 1 44.5 39.3    45.1 10.4 9.17
## 2 17.2 45.9    69.3  9.3 7.93
## 3 120. 19.6    11.6 13.2 12.6
## 4 66.1  5.8    24.2  8.6 10.2
## 5 281. 39.6    55.8 24.4 19.9
## 6 218. 27.7    53.4 18  17.1
## 7 228. 16.9    26.2 15.5 17.5
## 8 267. 43.8     5  25.4 19.3
## 9 74.7 49.4    45.7 14.7 10.5
## 10 207.  8.4    26.4 12.9 16.5
## # ... with 49 more rows
```



Predictions

One important aspect to assess the accuracy of predictions is to look at the out-of-sample mean squared error and compare it with the in-sample mean squared error. The former is nothing but the MSE computed using the predictions on the test data.

```
pred.MSE <- mse(mod1, test)
c(Prediction_MSE = pred.MSE, Train_MSE = MSE)
```

```
## Prediction_MSE    Train_MSE
##      11.59707      10.16292
```

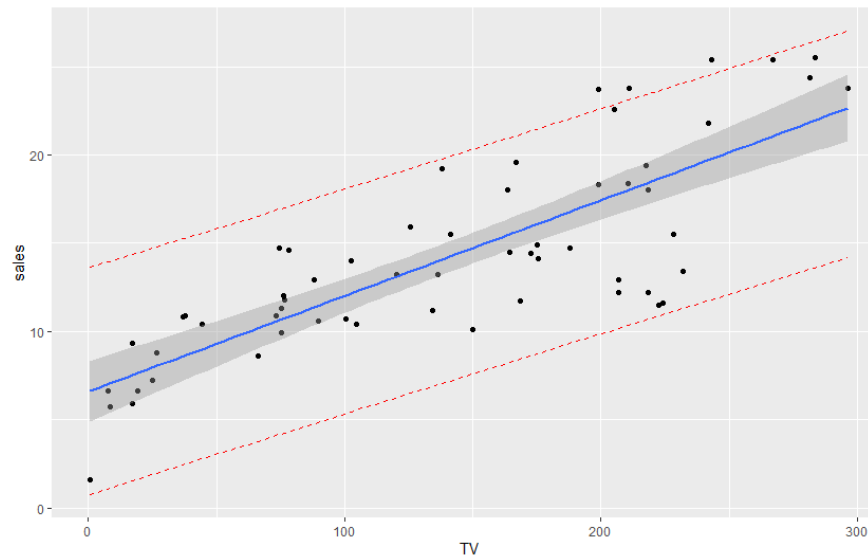
Confidence Interval Predictions

```
test.predict <- predict(mod1, test, interval = "prediction")
```

```
new_df <- cbind(test, test.predict)
```

```
ggplot(new_df, aes(x = TV, y = sales))+  
  geom_point() +  
  geom_line(aes(y=lwr), color = "red", linetype = "dashed")+  
  geom_line(aes(y=upr), color = "red", linetype = "dashed")+  
  geom_smooth(method=lm, se=TRUE)
```

`geom_smooth()` using formula 'y ~ x'



Homoscedasticity Checking

Violation of the equal variance assumption of the error terms is identified, if we observe a funnel or double-bow pattern in the residual plot.

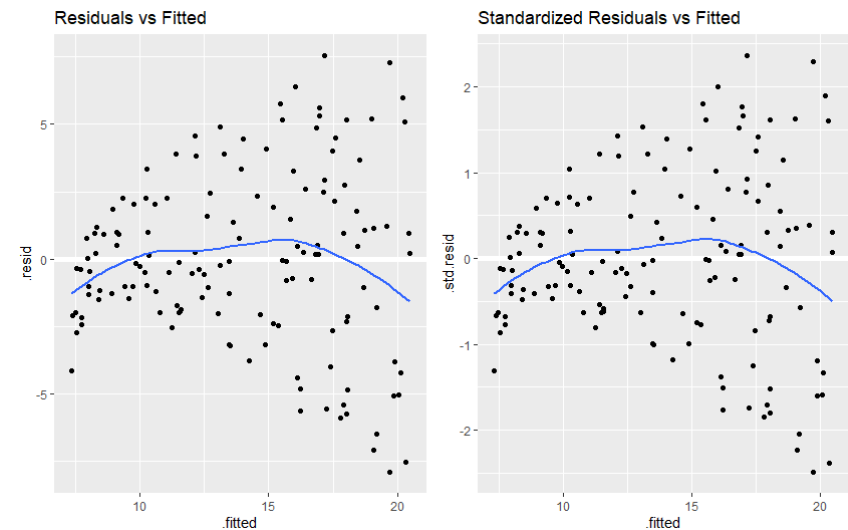
add model diagnostics to our training data
`mod1_results <- augment(mod1, train)`

```
p1 <- ggplot(mod1_results, aes(.fitted, .resid)) +  
  geom_ref_line(h = 0) +  
  geom_point() +  
  geom_smooth(se = FALSE) +  
  ggtitle("Residuals vs Fitted")
```

```
p2 <- ggplot(mod1_results, aes(.fitted, .std.resid)) +  
  geom_ref_line(h = 0) +  
  geom_point() +  
  geom_smooth(se = FALSE) +  
  ggtitle("Standardized Residuals vs Fitted")
```

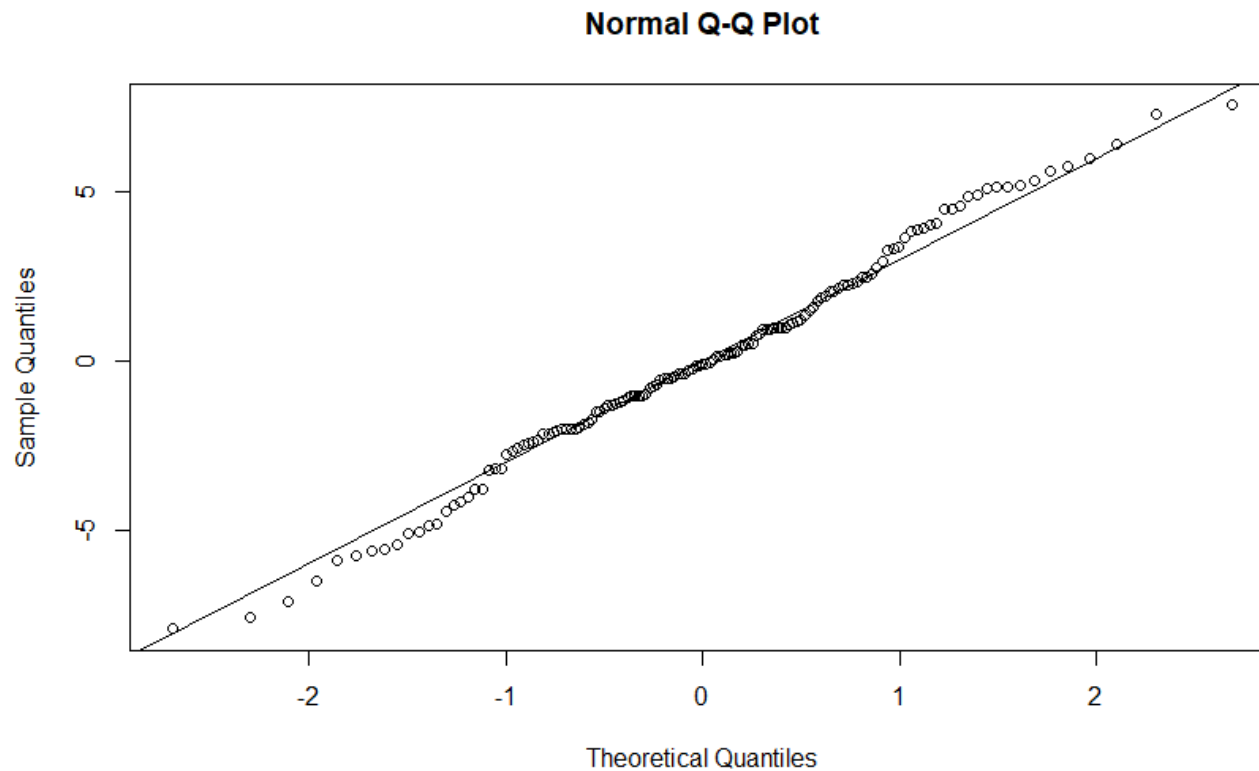
```
gridExtra::grid.arrange(p1, p2, nrow = 1)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'  
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Normality Checking

```
# Normality check of errors  
qqnorm(mod1_results$resid)  
qqline(mod1_results$resid)
```

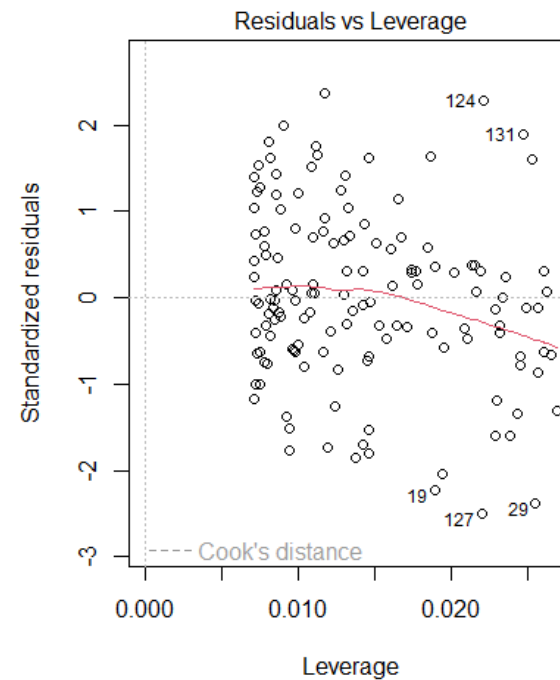
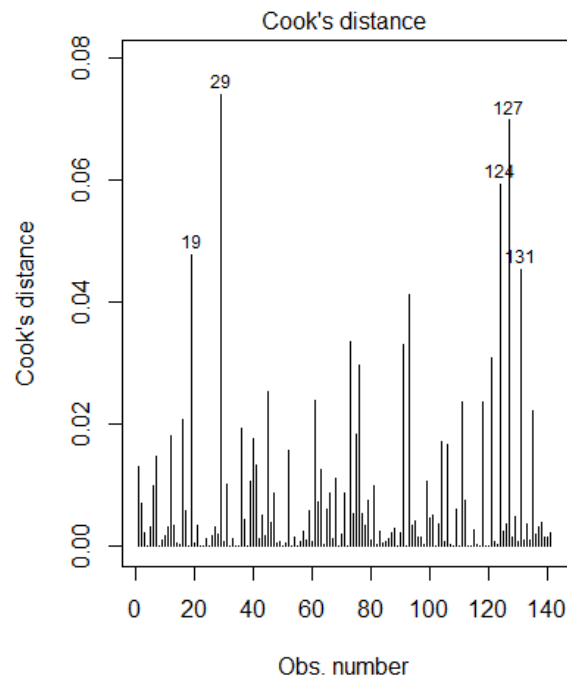


Outlier Analysis

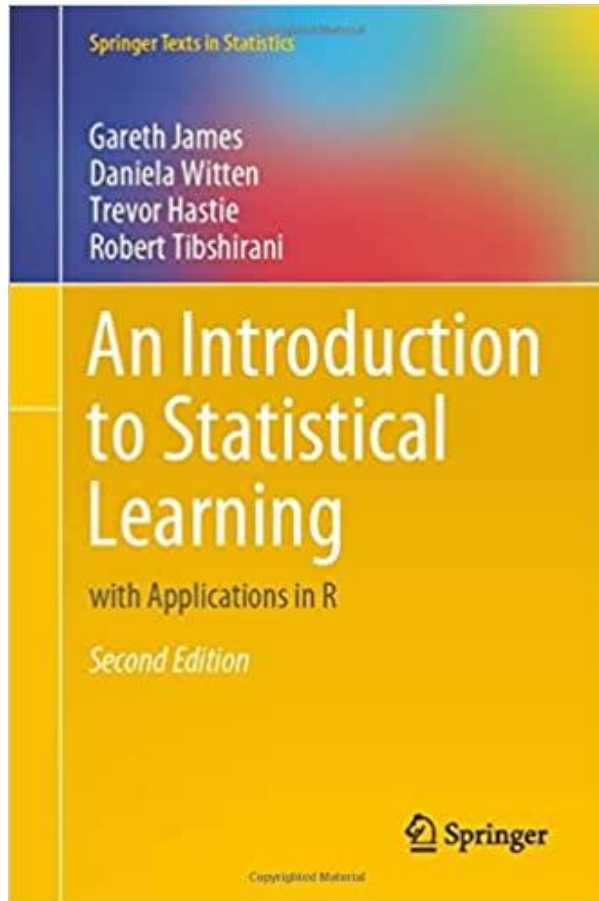
- A point is an influential point if its deletion, singly or in combination with others, causes substantial changes in the fitted model. We are interested in detecting those points.
- Observations with large standardized residuals are outliers in response variable because they lie far from the fitted equation in the Y direction. Since the standardized residuals are approximately normally distributed with mean zero and sd 1, points with standardized residuals larger than 2 or 3 sds away from the mean (zero) are outliers.
- Outliers can also occur in the predictor variables (X-space), that may affect the regression results. Observations that are outliers in the X space are known as high-leverage points. In general, high-leverage points should be flagged and examined to see if they are influential as well.
- An influence measure, known as Cook's distance is widely used to measure the influence of the observations. A practical operational rule is to classify points with Cook's distance greater than 1 as being influential.

Leverage Points

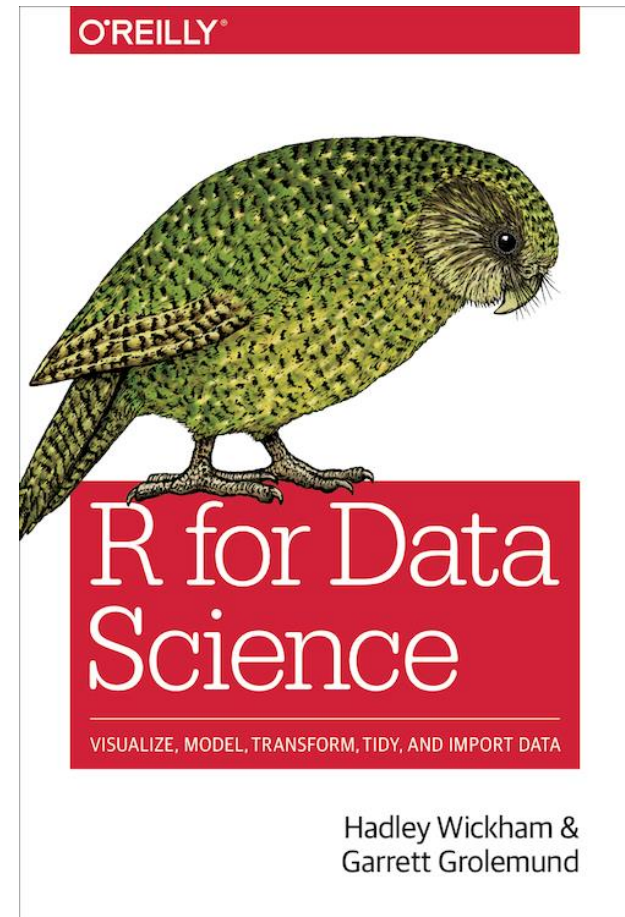
```
par(mfrow=c(1, 2))
plot(mod1, which = 4, id.n = 5)
plot(mod1, which = 5, id.n = 5)
```



References.....



<https://www.statlearning.com/>



<https://r4ds.had.co.nz/>