Linear Regression

What is Linear Regression ?

- A statistical measure that determines the strength of relationships between a dependent variable(Y) and a series of changing independent features (X)

- Relationship between two coefficient of an independent variable (X) and a dependent variable (Y)

- Relationship can be modelled as
  - Linear
  - Other functions like Polynomial, Quadratic etc.

Simple Linear Regression

Year	Profit
2001	20
2002	24
2003	33
2004	36
2005	55
2006	29
2007	47
2008	**?**

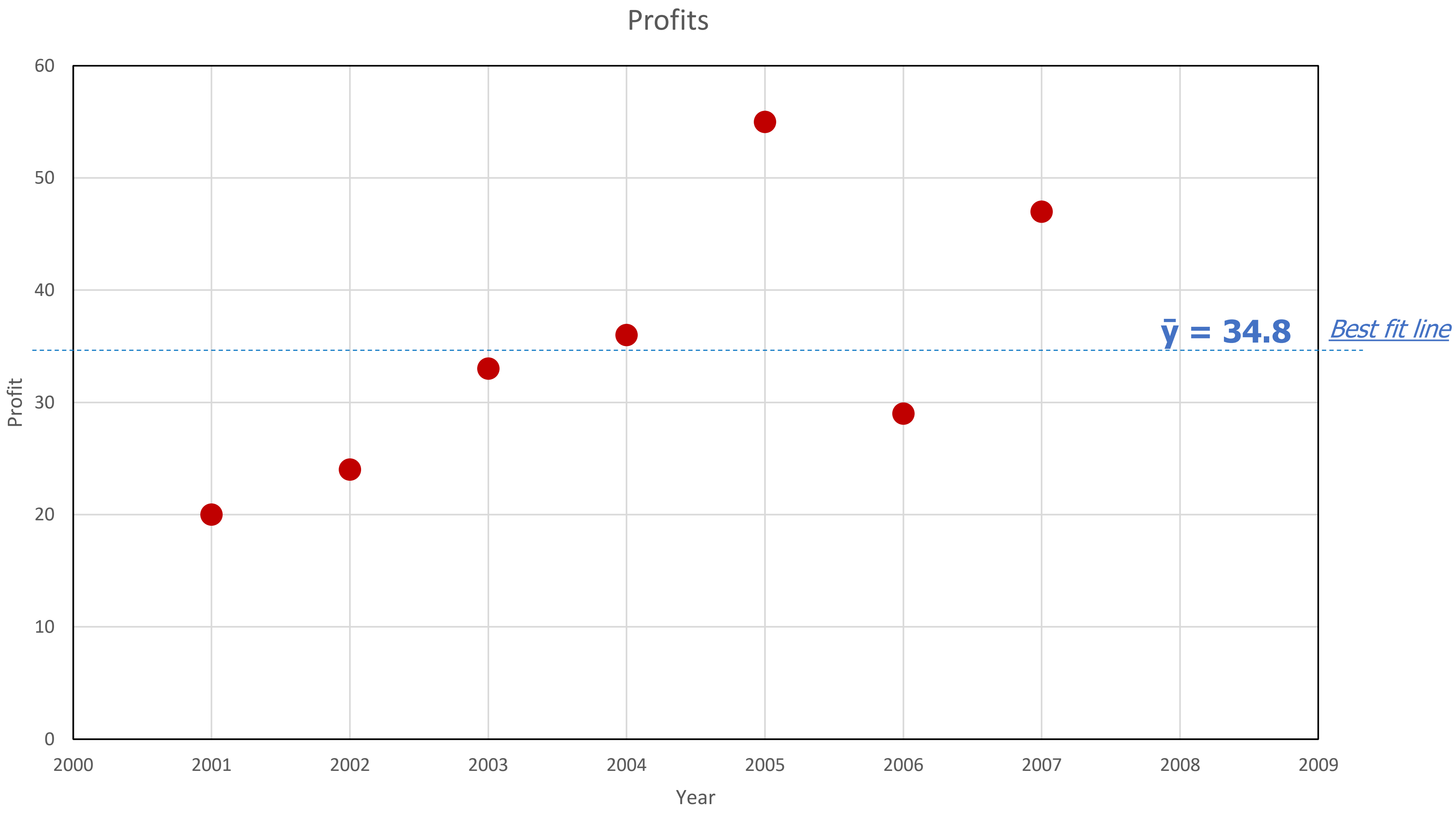
Table I

With the given data,
predict the Profit

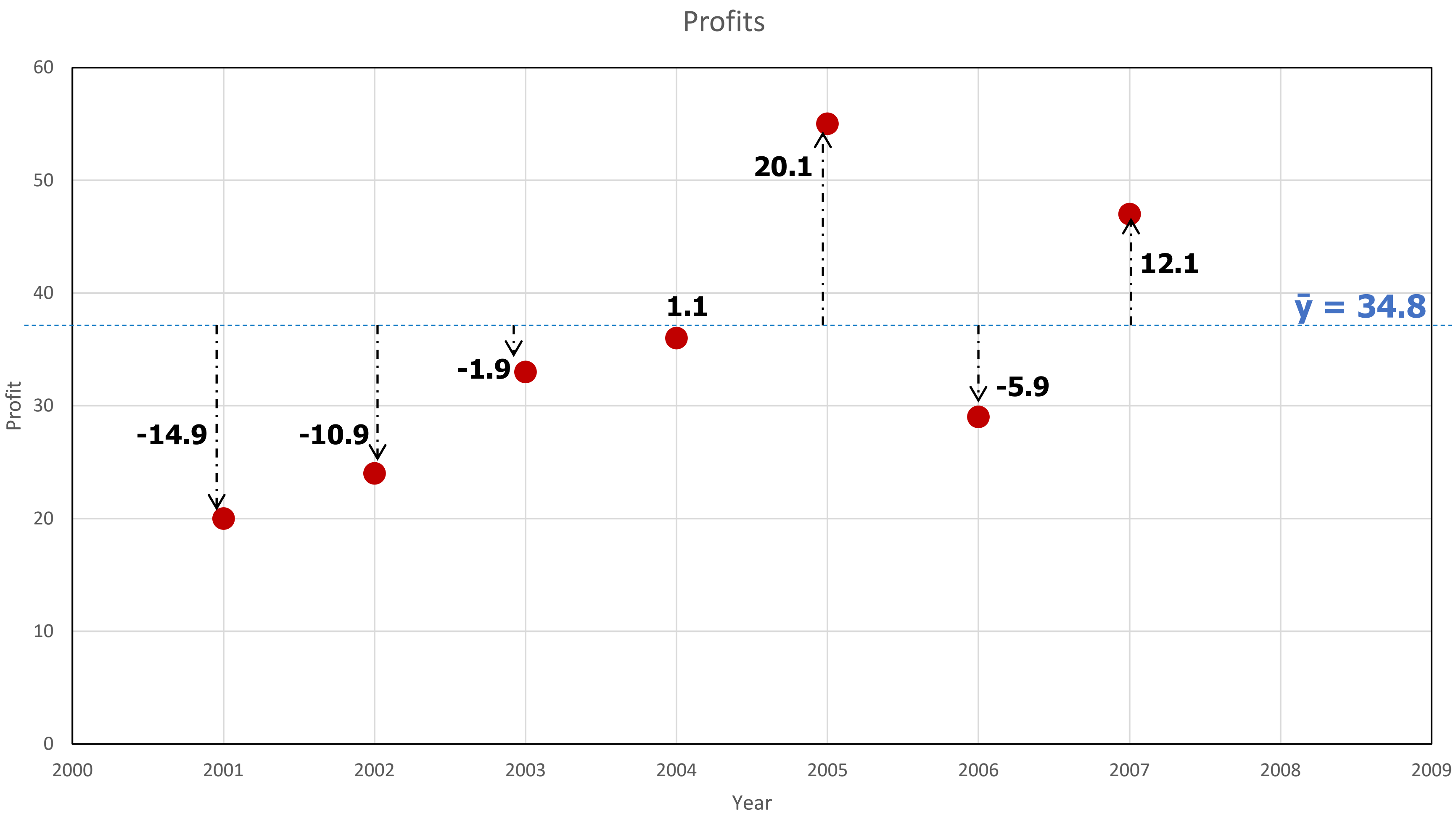
(20+24+33+36+55+29+47) / 7

34.8

Sample problem 1 : No Independent variables



Best Fit Line ?



Y - ȳ (e)

Year	Profit (y)	Residuals/Error
2001	20	20-34.8 = -14.9
2002	24	24-34.8 = -10.9
2003	33	33-34.8 = -1.9
2004	36	36-34.8 = 1.1
2005	55	55-34.8 = 20.1
2006	29	29-34.8 = -5.9
2007	47	47-34.8 = 12.1
2008	?	

$\Sigma e = 0$

- **With only 1 variable to predict, the predicted value (Profit) = mean (Profit)**
- **Variability in the Profit can be explained only by Profit**

Squaring the Errors (Method of Least Squares)

Year	Error	$(\text{Error})^2$
2001	-14.9	220.73
2002	-10.9	117.88
2003	-1.9	3.45
2004	1.1	1.31
2005	20.1	405.73
2006	-5.9	34.31
2007	12.1	147.45
SSE (Sum of Square of Errors)		930.86

Why square the errors ?

- Make them all positive
- Exaggerate the larger deviations

Goal of Simple Linear Regression

- To create a model that will minimise the Sum of Square of Errors (SSE)
- A new line will be introduced (Independent variables / x variables) that will minimise the size of the squares. This will then be the "Best Fit Line" ($\hat{\mathbf{Y}}$)
- A Linear Regression model is considered "GOOD" when the model reduces the SSE

What is done in Regression ?

Develop an estimating equation → *A mathematical formula relating X and Y* → $Y = a + bX + \varepsilon$

Independent variable

Error

Slope of the line

Y-intercept

Dependent variable

Apply correlation analysis → *How well the estimating equation describes the relationship (Strength)*

Choose coefficients **'a'** and **'b'** such that **Y** is close to the training examples of (x,y)

a = (intercept) → to move the line up and down the graph
b = (slope) → to change the steepness of the line
x = (explanatory/independent variable)
y = (predicted variable/dependent variable)

A few points in the interpretation of Linear Regression

- Relationships caused by regression is to be considered as "*relationships of association*"
- Relationships caused by regression is not always "*causal*" – Independent values (x) causes the dependent variable (Y) to change

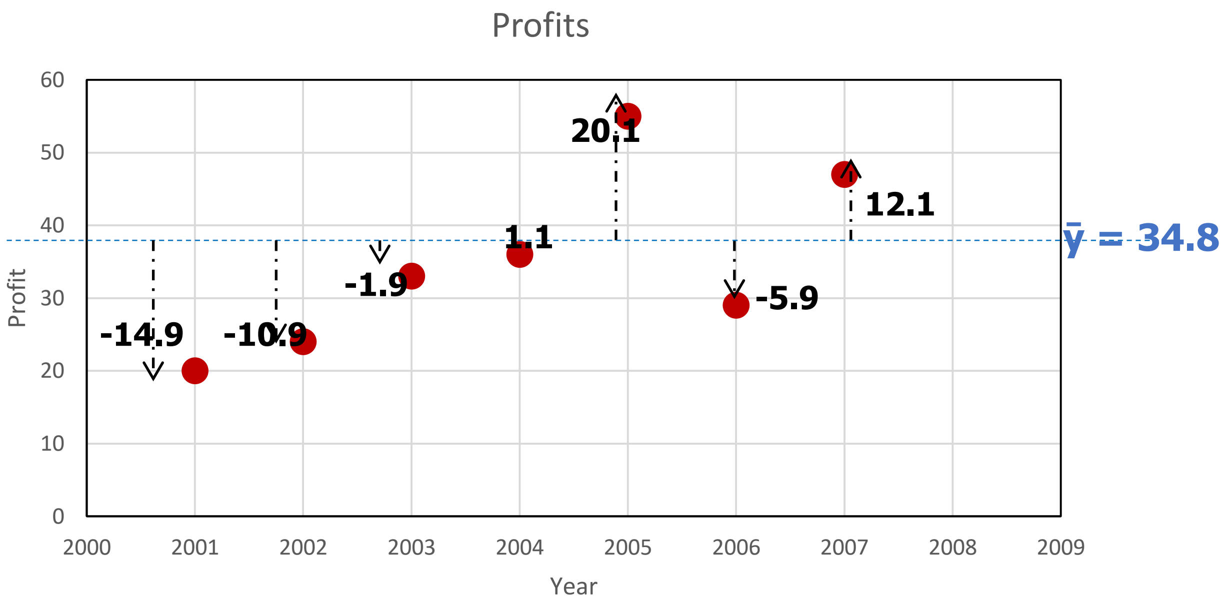
Sample problem 2 : With Independent variables

Year	Amt spent on R&D (x)	Profit (y)
2001	2	20
2002	3	24
2003	5	33
2004	9	36
2005	14	55
2006	11	29
2007	13	47
2008	19	**?**

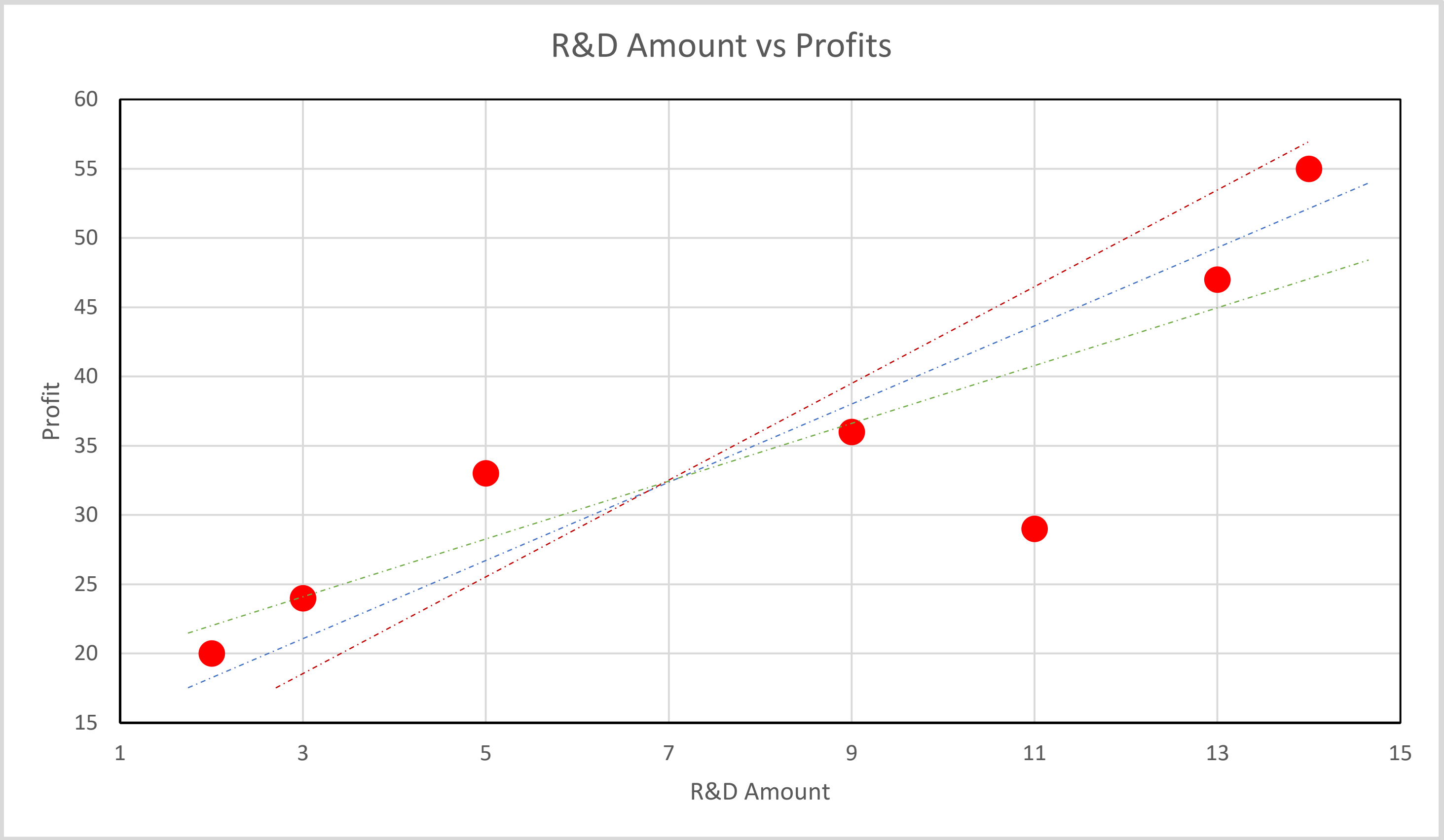
Table II

Predict the **Profit** given the *"Amount spent on R&D"*

Profit Y / Dependent variable
R&D Amount X / Independent variable



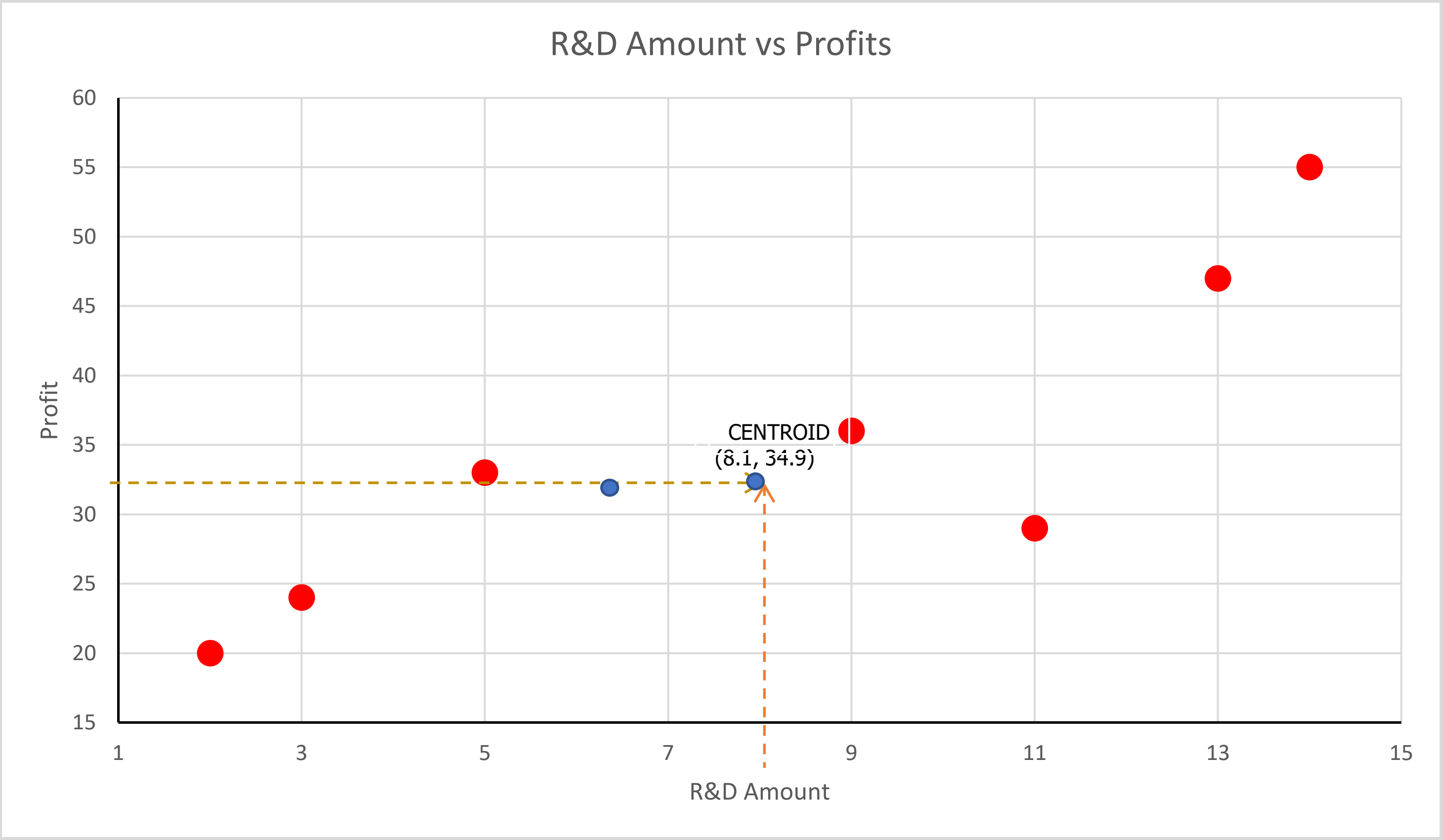
- The regression model with the new Independent variable will be compared with this model to see how good it is
- The error component should be < 930.86



amt_r_d	profit
2	20
3	24
5	33
9	36
14	55
11	29
13	47

Is there a linear pattern along the data points ?

Is there a Correlation between X and Y ?



amt_rd (X)	Profit (Y)
2	20
3	24
5	33
9	36
14	55
11	29
13	47

$\bar{x}$	$\bar{y}$
8.1	34.9

- The best fitting regression line MUST / WILL pass through this centroid
- From [regression calculations](),
  a = 16.6968
  b = 2.2302
- $\hat{Y} = 16.6968 + (2.2302 * X_1)$

Exercise

Calculate Profit for X_1 = 15, 16, 18

$\hat{Y}$ = 16.6968 + (2.2302 * 15) = **50.14**

$\hat{Y}$ = 16.6968 + (2.2302 * 16) = **52.38**

$\hat{Y}$ = 16.6968 + (2.2302 * 18) = **56.84**

Calculation of 'a' and 'b'

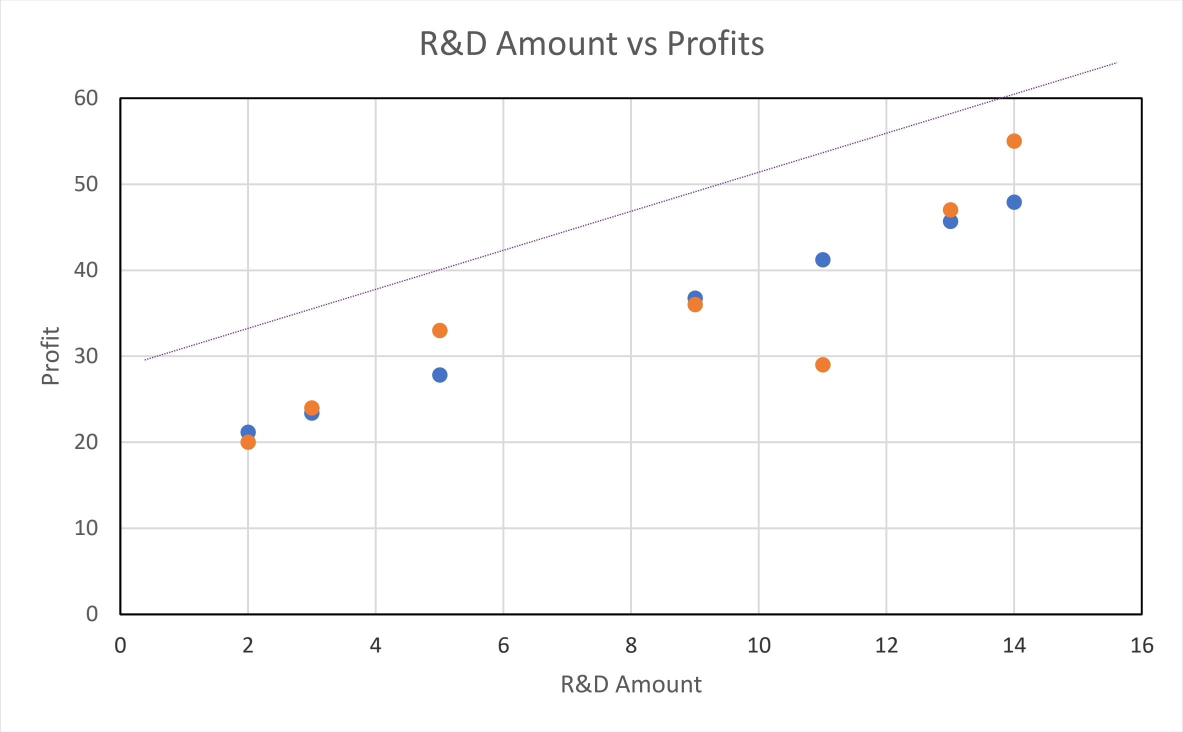
$$b = \frac{N\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{N\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

N: Number of observations
x_i: Independent feature
y_i: Dependent feature
$\bar{x}$: Mean of x
$\bar{y}$: Mean of y

Prediction using Regression

Year	R&D (X)	Profit (Y) (ACTUAL)	$\hat{Y}$ (PREDICTED) 16.6968 + (2.2302 * X)	Residual (e)	e^2
2001	2	20	21.15	-1.15	1.32
2002	3	24	23.38	0.62	0.38
2003	5	33	27.84	5.16	26.63
2004	9	36	36.76	-0.76	0.58
2005	14	55	47.91	7.09	50.27
2006	11	29	41.22	-12.22	149.33
2007	13	47	45.68	1.32	1.74
					230.25
Mean Square Error (MSE) (COST FUNCTION = SSE/n)					**32.892**



SSE without X	SSE with X	SSR
930.86	230.25	700.61

SSE : Sum of Squares of Errors
SSR : Sum of Squares due to Regression
SST : Total Sum of Squares

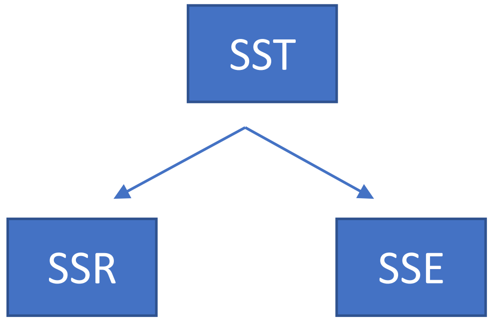
Comparing Residuals / Errors (e^2) - SSE

Without X

SSE	SSR	SST
930.86	-	**930.86**

With X

SSE	SSR	SST
230.25	700.61	**930.86**



SSE : $\Sigma(Y - \hat{Y})^2$: Unexplained deviation

SSR : $\Sigma(\hat{Y} - \bar{y})^2$: Explained deviation from mean

SST : $\Sigma(Y - \bar{y})^2$: Total Error (SSR + SSE)

It is the relation between SSR, SSE and SST that represents each value of the independent variable

Standard Error

The difference between the Actual (**Y**) and Predicted (**Ŷ**) Value of a regression

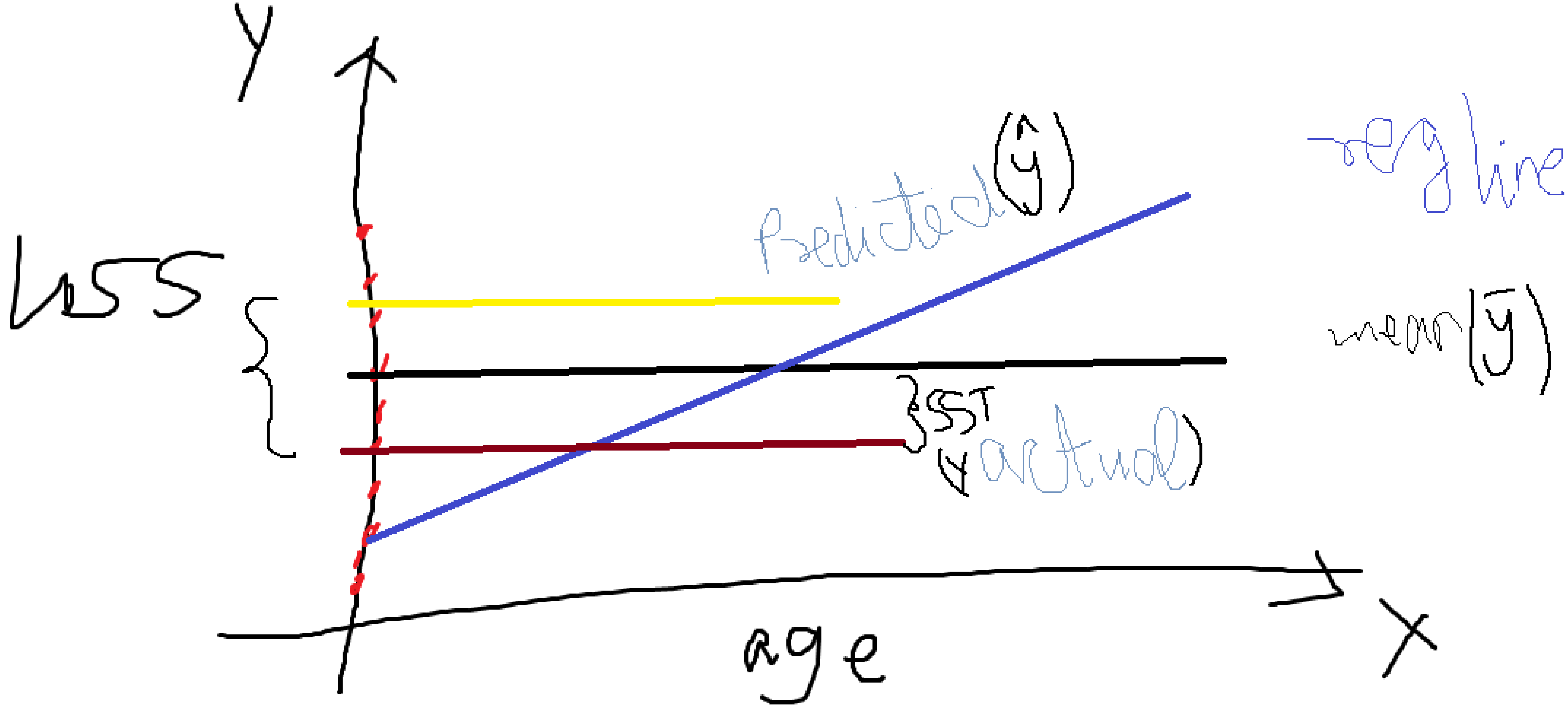
Formula: $$\sqrt{\frac{\Sigma(Y - \hat{Y})2}{(n - k - 1)}}$$

Where
Y = actual value
Ŷ = predicted value
N = sample size

	Y	Ŷ	(Y - Ŷ)	(Y - Ŷ)²
	3	3.8	-0.8	0.64
	5	4.3	0.7	0.49
	6	7.8	-1.8	3.24
	8	7.8	0.2	0.04
	6	5.2	0.8	0.64
Total				5.05
n				5
SE				1.297433

Individual Features

$$\frac{\sqrt{\frac{\Sigma(Y - \hat{Y})2}{(n - k - 1)}}}{\sqrt{(x - X)^2}}$$



Y
loss
Predicted(ŷ)
reg line
mean(ȳ)
SST
Y (actual)
age
X

How well does the regression equation fit data ?

Coefficient of Determination (R^2)
R^2 = SSR / SST
Proportion of total variation explained

SSE	SSR	SST	R^2	R^2
230.25	700.61	**930.86**	0.7526	**75.26 %**

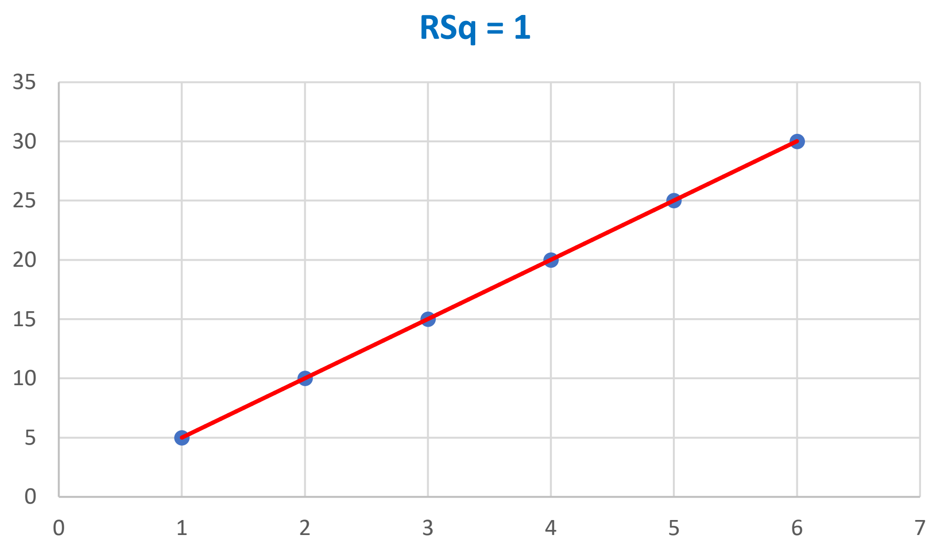
High SSE -→ Low R^2
Low SSE → High R^2

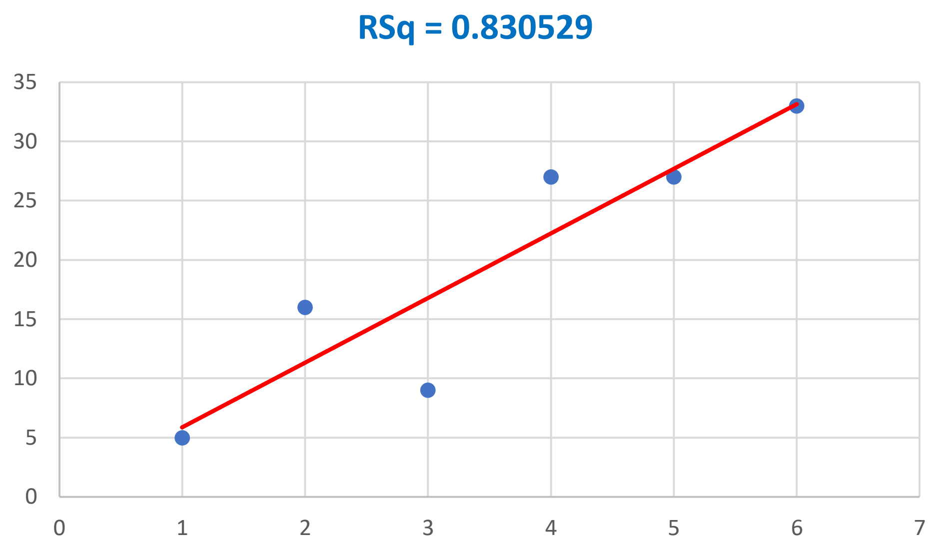
Interpretation of Coefficient of Determination (R^2)

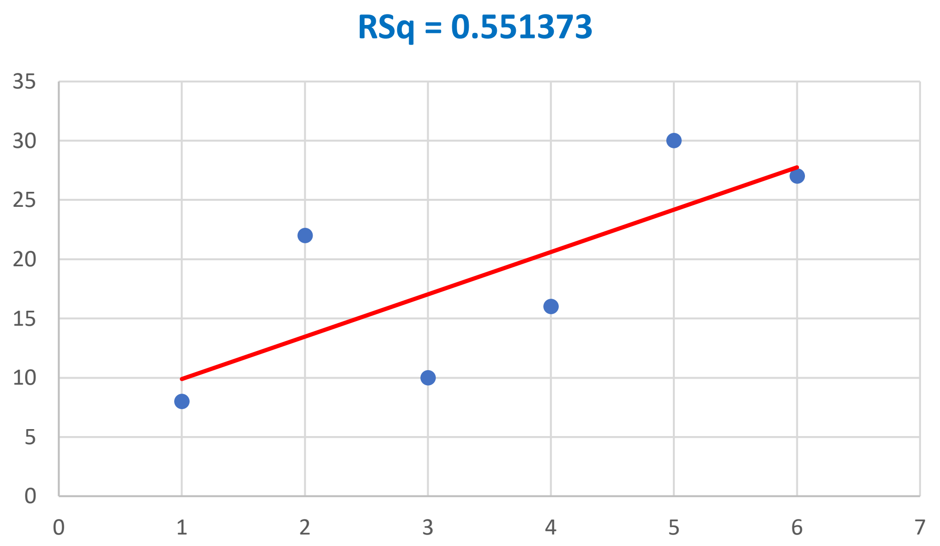
75.26% of the total sum of squares can be explained by the estimated Regression equation ($\hat{Y} = 16.6968 + (2.2302 * X_1)$) to predict the Profit. (Y). The remainder is the error.

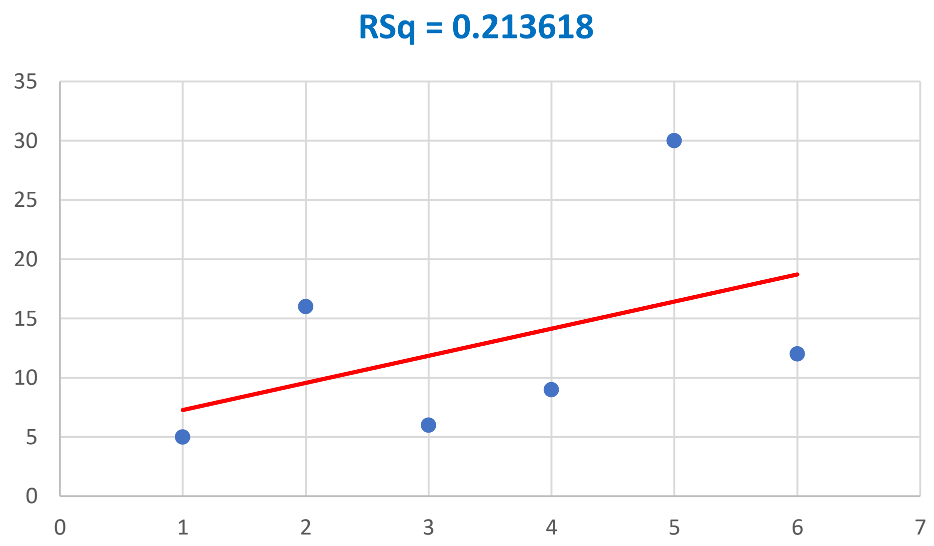
Proportion of variability in Y (Dependent variable) that is explained by the independent variables (X)

This model is a Good Fit


RSq = 1
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7

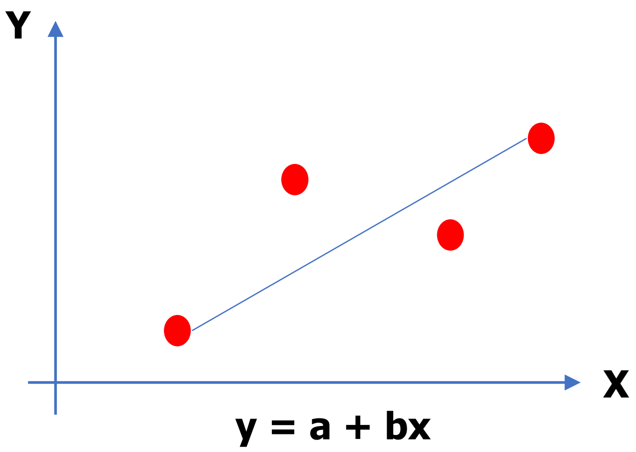

RSq = 0.830529
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7


RSq = 0.551373
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7

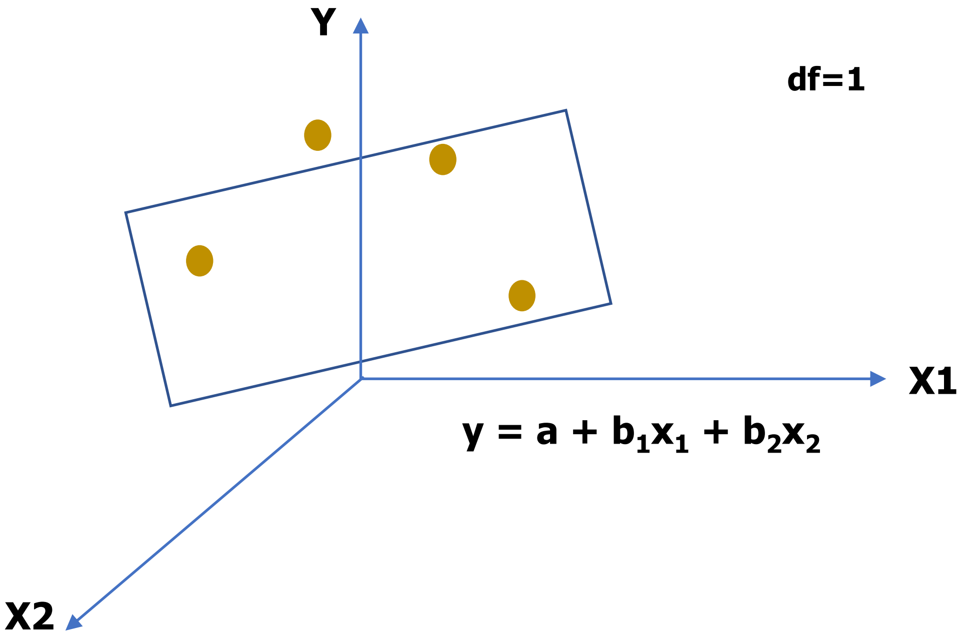

RSq = 0.213618
35
30
25
20
15
10
5
0
0 1 2 3 4 5 6 7

Degrees of Freedom

- The minimum number of observations required to estimate a regression equation $\mathbf{y = a + b_n x_n}$



R^2=1, no strength
R^2=0.8, df=1
R^2=0.78, df=2



Formula for DOF
DOF = n-k-1
where

n = number of observations
k = number of independent variables

- As **k (number of features)** increases, **DOF** decreases

More factors, more DOF is lost
Eg: to make a decision alone, there is no DOC
When you add FATHER, you lose full degree (1)
F+M, lose more freedom
....

Adjusted R^2

- Provides an unbiased estimate of the population R^2
- Modified version of R^2 adjusted for the number of Xs in the model
- Increases only if a newly added X is significant
- Compares the explanatory power of regression models having multiple Xs
- Can be negative, but usually positive
- Value is always lesser than R^2
- Formula

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N-1)}{n-k-1}$$

where
n = sample size
k = number of predictors

As **k (number of features)** increases, $R^2_{adjusted}$ decreases; holding everything else constant

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -7.802859  31.345516  -0.249   0.8035
cementcomp        0.119625   0.010020  11.939  < 2e-16 ***
slag              0.102261   0.012003   8.520  < 2e-16 ***
flyash            0.088446   0.014925   5.926 4.80e-09 ***
water            -0.190903   0.047096  -4.053 5.59e-05 ***
superplastisizer  0.156929   0.110440   1.421   0.1558
coraseaggr        0.009265   0.011063   0.837   0.4026
finraggr          0.021343   0.012717   1.678   0.0937 .
age               0.125699   0.006810  18.457  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 724 degrees of freedom
Multiple R-squared:  0.6234,    Adjusted R-squared:  0.6193
F-statistic: 149.8 on 8 and 724 DF,  p-value: < 2.2e-16
```

R^2 vs Adjusted R^2

R^2

- When new features (X) are added to a model, R^2 only increases or remains constant but never decreases.
- Difficult to judge the model accuracy

Adjusted R^2

- The Adjusted R-Square is the modified form of R-Square
- Adjusted for the number of predictors in the model using the model's degree of freedom
- The adjusted R-Square only increases if the new term improves the model accuracy.

Linear Regression assumptions

- Regression model is linear in it's coefficients (**y** has a linear relationship with **b**)

  $$\mathbf{y = a + b_1x_1 + b_2x_2^2}$$

  Equation is linear even with x raised to power 1 and 2

- Mean of residuals (of the linear model) is 0 (or near 0)

- Residuals have equal variance – This is known as **Homoscedasticity**
  - *Residuals not having equal variance is known as **Heteroscedasticity***
  - *Identify by plotting the residuals against the predicted Y*

- Residuals are normally distributed

- Residuals are independent of each other
  - If not independent, it is known as **Auto Correlation**

- Number of observations must be greater than number of X's

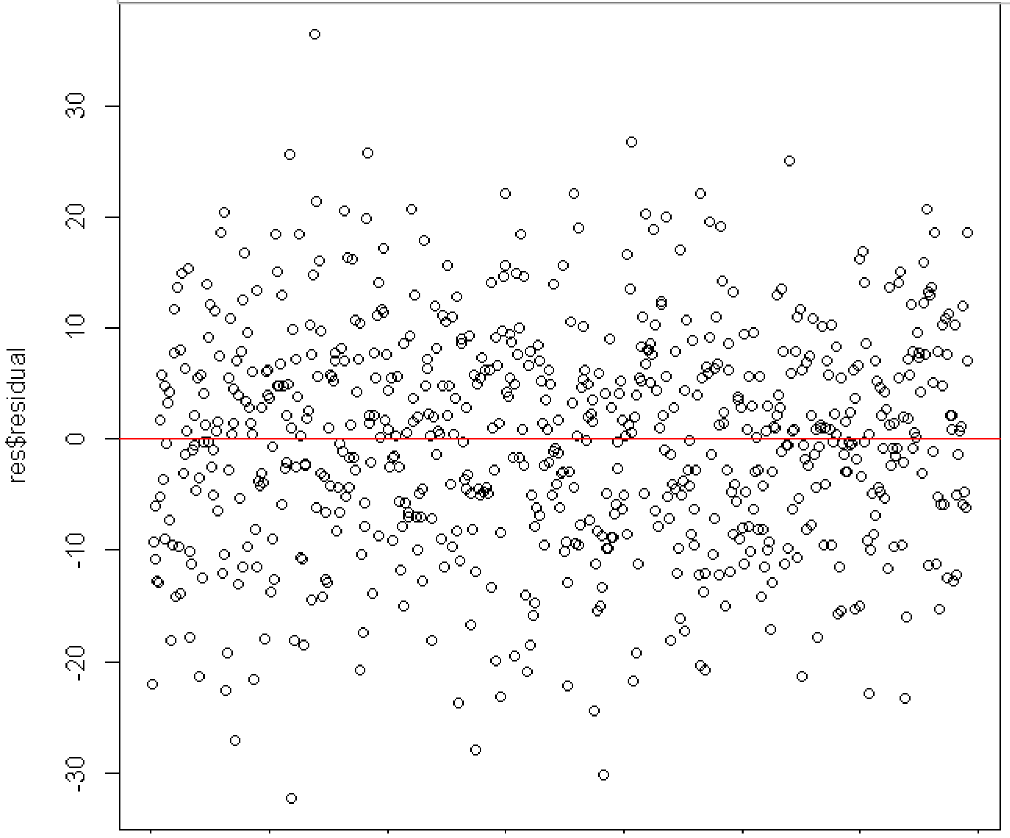
- Absence of outliers

*These assumptions are important. It is these assumptions that **differentiate** Linear Regression with other regression models like Logistic Regression etc.*

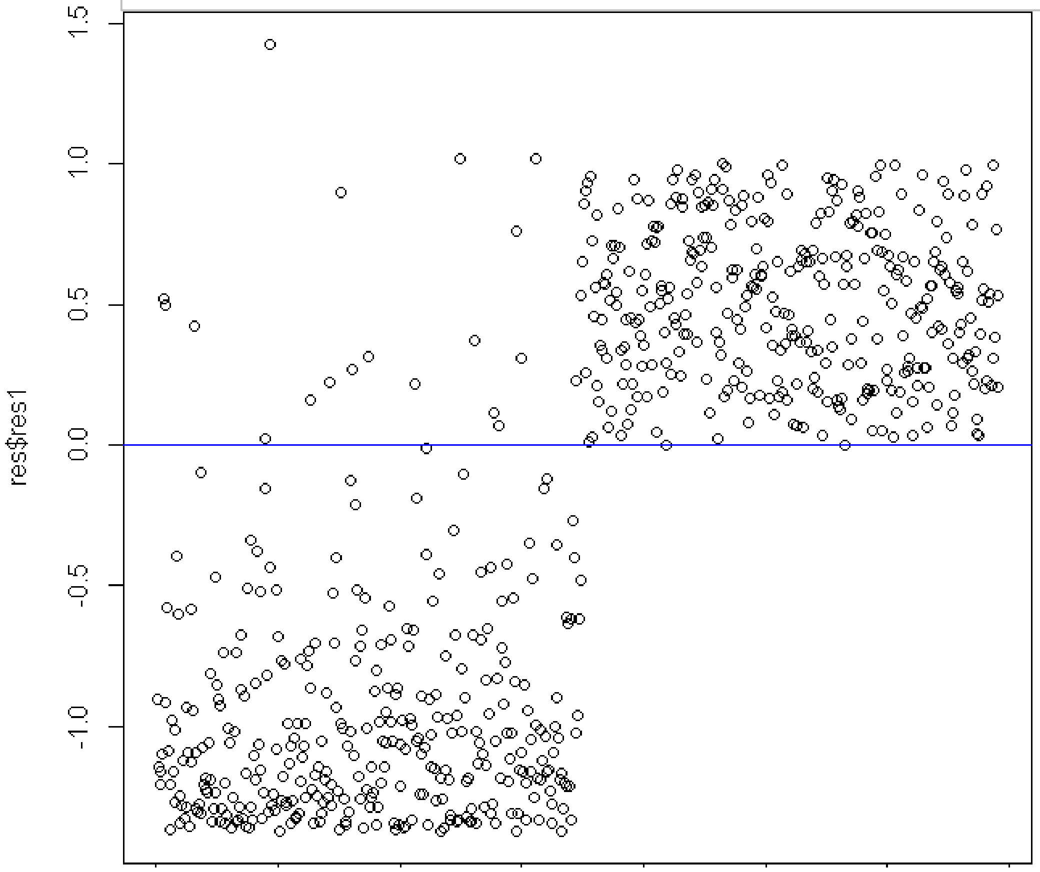
Heteroscedasticity

- A situation where the residuals / errors exhibit **unequal variance**
  - The errors are not constant
  - Can see patterns
  - Errors increases / decreases with every record predicted
  - Generally seen in cross-section data, not in Time Series

Constant Variance in errors / Homoscedasticity



Patterns in Variance in errors / Heteroscedasticity



Examples of Heteroscedasticity

1. **Age vs Salary**
   - Increase in Age causes an exponential increase in salary
   - Increase in Age causes a gradual increase in salary
   - Increase in Age causes a little increase in salary

2. **Earnings vs Expenditure**
   - More earnings causes more expenditure
   - More earnings causes controlled expenditure
   - More earnings causes little expenditure

Consequences of Heteroscedasticity

- **Coefficient estimates may show significance; where as in reality they may be insignificant**

Test for Heteroscedasticity

- Using the **Residuals plot** (plot the predicted Y against the residuals)
- **Park Test**
- **Glejser Test**
- **Goldfeld-Quandt Test**
- **Breusch-Pagan-Godfrey test**
- **NCV (Non-Constant Error Variance) Test**
- **Whites Test**

Hypothesis Testing

- $\mathbf{H_0}$**:** Homoscedasticity (Error Variances are equally distributed)
- $\mathbf{H_1}$**:** Heteroscedasticity (Error Variances are not equally distributed)

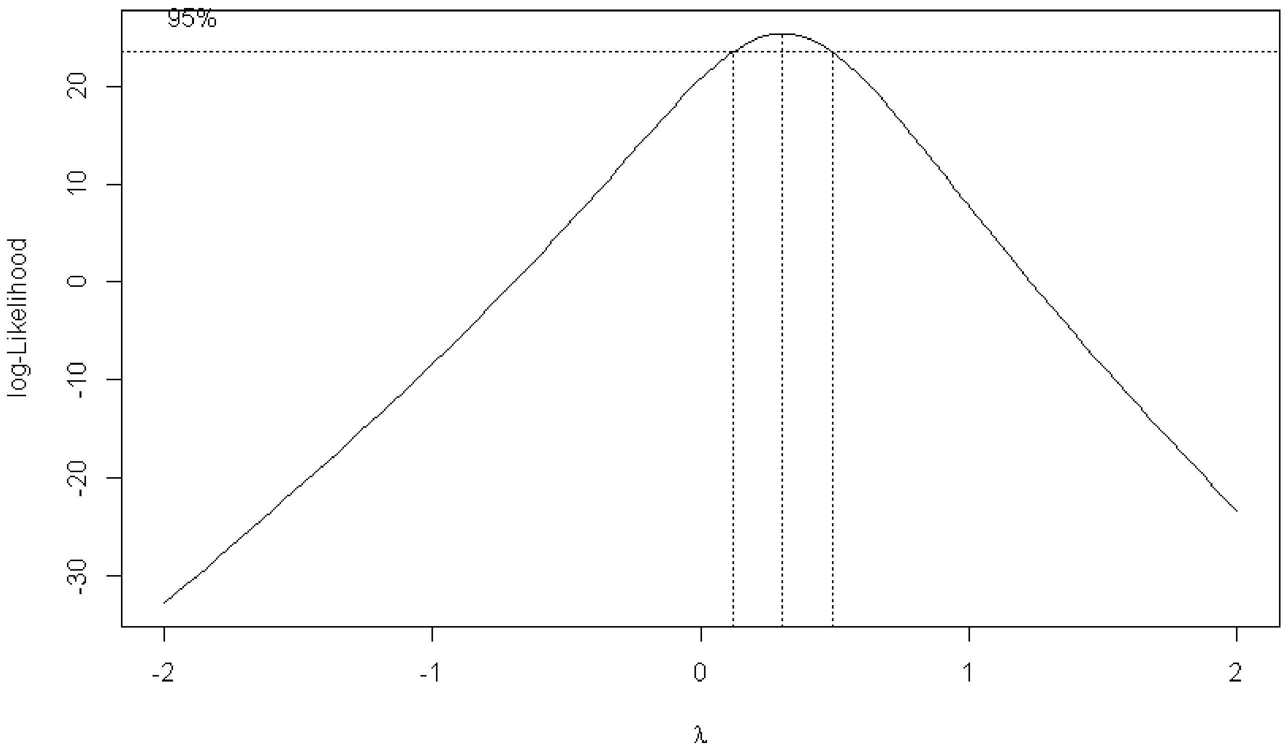
How to remove Heteroscedasticity

1. **Data transformation of the features and y-value**
   - Transform the dataset into **log** or other relevant transformation
   - Re-build model using these transformed values

2. **Box-Cox transformation**
   - Transformation of the y-variable by selecting an appropriate Gamma
   - Re-build model using the transformed y-value

BoxCox Transformation

- A technique to identify an appropriate exponent to transform the data
  - Improve the normality
  - This exponent is called the ***lambda***
  - Lambda value indicates to what power the data should be raised

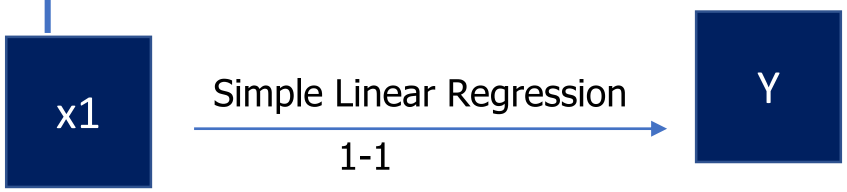


BoxCox Transformation formula

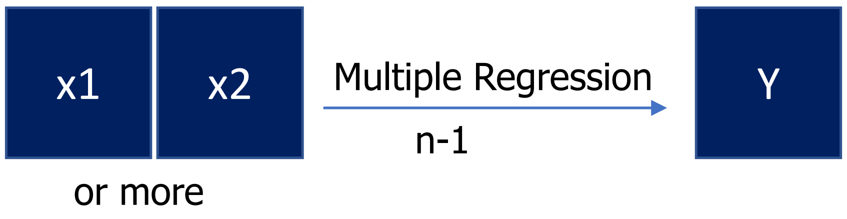
lambda	Y
> 0	((x^lambda)-1) / lambda
< 0	log(x)

Multiple Linear Regression

- It is an extension of the Simple Linear Regression
- Two or more Independent variables (x_1, x_2, …x_n) are used to <u>predict</u> or <u>explain the variance</u> in Y – the dependent variable



Year	Amt spent on R&D	Profit
2001	2	**20**
2002	3	**24**
2003	5	**33**



Predict **"Profit"** based on the input variables
"**R&D Amt, Employees, Advertisement Amt**"

Year	Amt spent on R&D	No_Emp	Adv	Profit
2001	2	10	6	**20**
2002	3	13	9	**24**
2003	5	20	13	**33**

A few points on Multiple Regression

- Adding new independent variables can help build a good model with better predictions, but this hypothesis need not be true always
- Eg: Adding Y-variables to improve R^2 from 60% to 80% (variation) may sound good, but it may be misleading
- Potential problems :
  - **Multicollinearity**
    - Correlation among the X-variables (X_n – X_n No relationship should exist)
    - Also referred to as "between-predictor correlation"
  - **Overfitting**
    - Incorrect predictions

  - **Solution**: Pick the best X-variables using *Variable selection techniques*

- Before implementing Multiple Regression, carry out a list of checks to ensure data is clean

- Estimated Multiple Regression Equation : $\hat{Y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_nx_n$

  Notice there is no error (ε) term. In MR, it is assumed to be 0

- Interpretation of the equation

  An estimated change in Y, corresponding to a 1-unit change in one x-variable, keeping other (x) variables constant

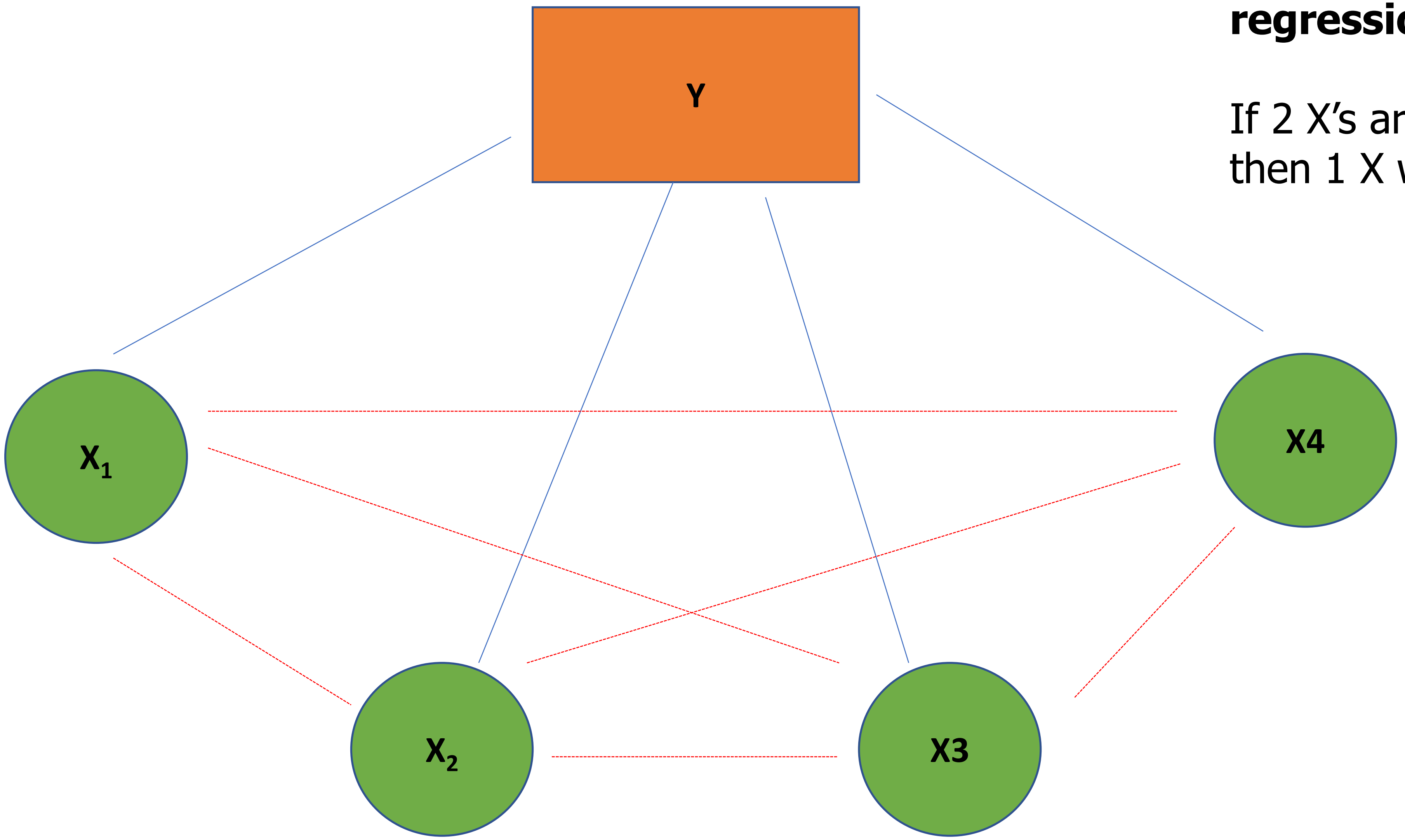
Identifying Multicollinearity

Variance Inflation Factor (VIF)

- Is a measure to identify the presence of multicollinearity in the independent variables
- Higher the value of VIF for a variable, greater the problem of multicollinearity
- As a general rule, **$VIF(X_n) > 5$** is considered as highly collinear and removed from the model
- Check other factors also before feature selection

```
> # variable inflation factor
> # to check Multicollinearity
> vif(lm1)
      cementcomp             slag           flyash            water 
          7.6158           7.1786           6.0867           6.6952 
superplastisizer       coraseaggr         finraggr              age 
          2.9123           5.0513           6.7309           1.1181 
> 
```

Multicollinearity



Elias property of linear regression

If 2 X's are multicollinear, then 1 X will be supressed

Predicting using the Linear regression formula

X_1 (lab_hrs)	X_2 (comp_hrs)	X_3 (reward)	$\hat{Y}$ (unpaid_tax)
60	65	25	76.535
62	75	30	91.512
70	90	45	119.995

ŷ = (intercept) + b1*lab_hrs + b2*comp_hrs + b3*reward

$$= -45.79 + (0.596)*x_1 + (1.176)*x_2 + (0.405)*x_3$$

Interpreting the Linear regression formula

The rate of change in **ŷ** for every 1 unit change in x_n, keeping other variables constant

X_1 (lab_hrs)	X_2 (comp_hrs)	X_3 (reward)	$\hat{Y}$ (unpaid_tax)
1	0	0	-45.194
0	1	0	-44.614
0	0	1	-45.385

Interpreting the model summary

Linear regression

```
Call:
lm(formula = unpaid_tax ~ ., data = tax)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29080 -0.11604 -0.09998  0.09102  0.44452
                           2         3        4
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.79635    4.87765  -9.389 8.29e-05 ***
lab_hrs       0.59697    0.08112   7.359 0.000323 ***
comp_hrs      1.17684    0.08407  13.998 8.29e-06 ***
reward        0.40511    0.04223   9.592 7.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
        1
Residual standard error: 0.2861 on 6 degrees of freedom
Multiple R-squared:  0.9834, 5  Adjusted R-squared:  0.9751 6
F-statistic: 118.5 on 3 and 6 DF,  p-value: 9.935e-06
```

$$\hat{y} = a + b_1x_1 + b_2x_2 + \ldots.. + b_nx_n$$

1) Residual standard error of regression
It is the estimated standard deviation of the "noise" in the dependent variable that is unexplainable by the independent variable(s)

2) Standard error of coefficient
It is the *estimated standard deviation of the error.* The higher the coefficient of determination, lower the standard error; and the more accurate predictions

3) t-value
Measure of the likelihood that the actual value of the parameter is not zero. Large t(|t|) == less likely parameter is 0

4) p-value
- P-values evaluate how well the sample data support the argument that the NULL hypothesis is true
- Sample provides enough evidence that the NULL hypothesis can be rejected for the entire population
- Probability of the likelihood that the actual value of the parameter is not zero. Small p == less likely parameter is 0
- P-value in the last line indicates if the model is good enough to be modelled

5) R^2 (COD – Coefficient of Determination)
Square of correlation between X and Y. Metric to evaluate the goodness of fit. Higher R^2, better model

6) Adjusted R^2
Unbiased estimate of the fraction of variable explained, taking into account the sample size and number of variables in the model, and it is always slightly smaller than unadjusted R-squared

ŷ = (intercept) + b1*lab_hrs + b2*comp_hrs + b3*reward

$$= -45.79 + (0.596)*X_1 + (1.176)*X_2 + (0.405)*X_3$$

Gradient Descent Optimization

- Optimization method used to find the values of the parameters (a, b_n) [coefficients] of a function $\hat{Y}$ that minimises the cost function
- Gradient descent is used when the parameters cannot be calculated analytically
- Searched using an optimization algorithm
- Regression uses Gradient Descent to minimise the Error terms

- Can also be used as a function that needs to be maximized:
  - MLE(Maximum Likelihood Estimate)

- By taking small / big steps, we get closer to the minimum – by adjusting the learning rate
  - Too small a value for learning rate → more number of iterations to arrive at the minimum value
    - The difference between Learning rate 0.1 and 0.01 is huge, though both are small numbers
  - Too big a value for learning rate → overshoot the minimum value
    - Need to go back and forth and keep readjusting the rates

- To decrease the cost function, take steps in the negative direction of the gradient

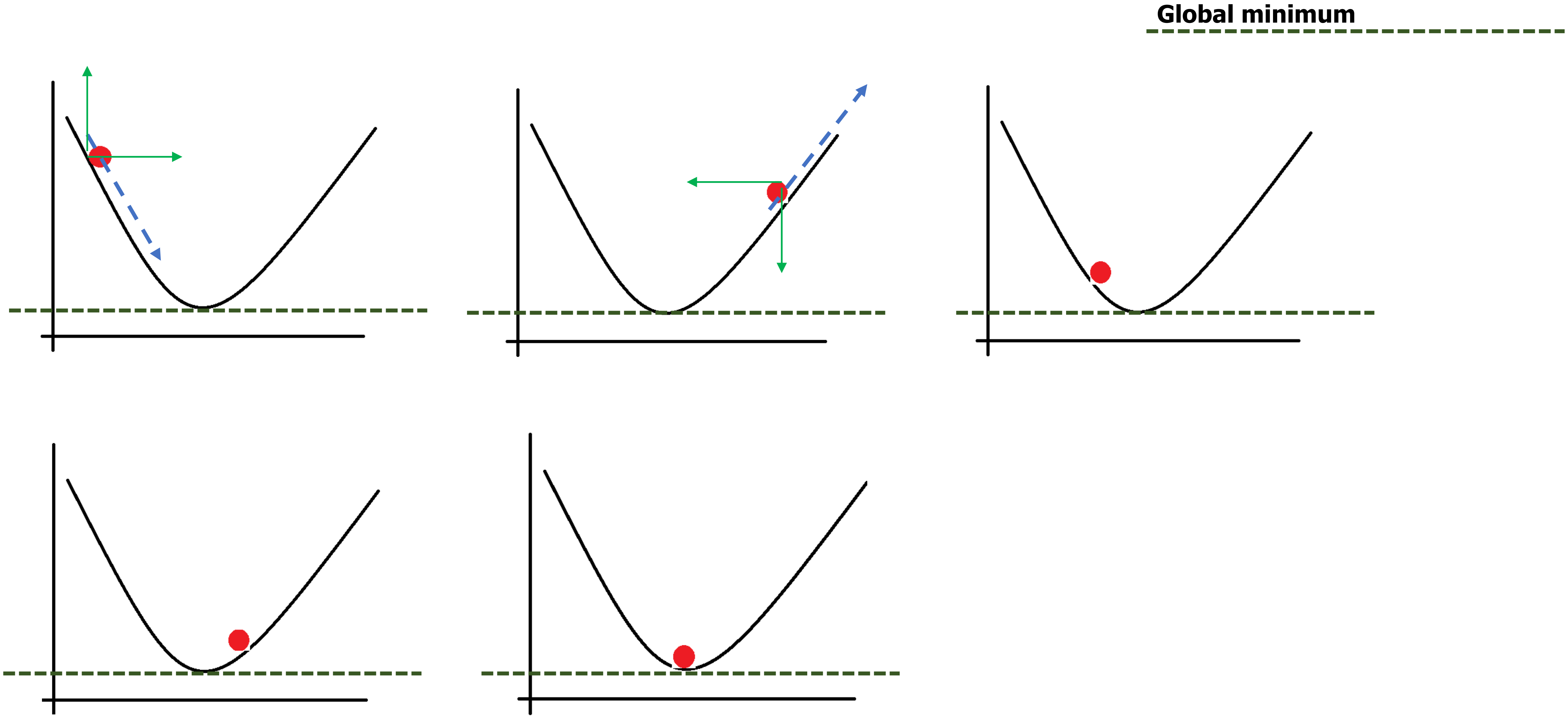
Cost function

$$\text{Cost} = \mathbf{1/m}\left(\sum_{i=1}^{m}(Y - \hat{Y})^2\right)$$

where
m = number of observations
Y = expected value
$\hat{Y}$ = predicted value

Also called **<u>Batch Gradient Descent</u>** as all the observations are taken as a single batch

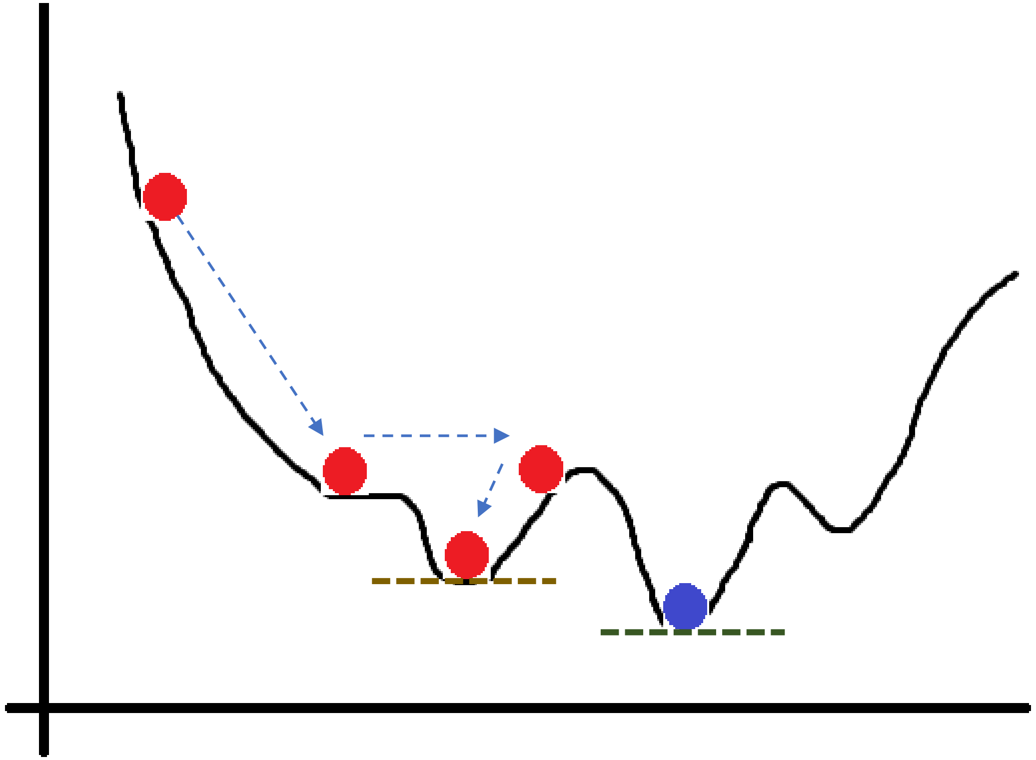
Gradient Descent – simple illustration



Stochastic Gradient Descent

Global minimum

Local minimum



- In this method, the weights are adjusted for every record / observation
- Finds the global minimum rather than the local minimum
- Local minimum will not be the best optimisation value
- Fluctuations are higher; so it is convenient to select the Global minimum
- Faster than batch process

Loss Function

- Loss is the difference between the Actual/Expected value (y) and Predicted value ($\bar{y}$)

- **Residual**
  - $l = (y - \bar{y})$ (also called residual $\hat{e}$)
  - $l(\hat{e}) = 0$ when the difference between Actual and Predicted values are 0

- **Sum of Square of Errors (Residuals)**
  - $\hat{e} = (y - \bar{y})^2$

- **Absolute / Laplace Loss**
  - $\hat{e} = |(y - \bar{y})|$