Regression

- **Regression** is a supervised learning technique which helps in finding the **correlation between variables** and enables us to **predict** the continuous output variable based on the one or more predictor variables.
- It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effectt relationship between variables**.
- **Dependent Variable:** The main factor in Regression analysis that we want to predict or understand is called the dependent variable. It is also called the **target variable**.
- **Independent Variable:** The factors which affect the dependent variables, or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.

Underfitting and Overfitting:

Overfitting

- If our algorithm works well with the training dataset but not well with test dataset, then such a problem is called **Overfitting**.
- The model is too complex and memorizes training data, but it fails on unseen data.
- Training error is low, but test error is high.
- Example: Suppose you're training a model to classify images of cats vs. dogs.If the dataset has background patterns (e.g., all cat images have wooden floors, and all dog images have grass), the model may learn to classify based on the background instead of the animal itself.
- The model performs well on training data but fails on new images where backgrounds are different.

underfitting

- if our algorithm does not perform well even with a training dataset, then such a problem is called **underfitting**.
- The model is too simple and fails to capture patterns in the data.
- Both training and test errors are high.
- Example: Imagine predicting house prices using only the number of rooms as the feature, ignoring location, square footage, and amenities. The model is too simple to make accurate predictions.
- Signs of Underfitting: Poor performance on both training and test data.
- High bias (strong assumptions, missing key patterns).

Why do we use Regression Analysis?

- **Regression analysis** helps in the prediction of a continuous variable.

- There are various scenarios in the real world where we need some future predictions such as weather conditions, sales predictions, marketing trends, etc., for such cases we need some technology that can make predictions more accurately.

- So, for such a case we need Regression analysis which is a statistical method used in machine learning and data science.

Some other reasons for using Regression analysis:

- ✓ Regression estimates the relationship between the target and the independent variable.
- ✓ It is used to find the trends in data.
- ✓ It helps to predict real/continuous values.

- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

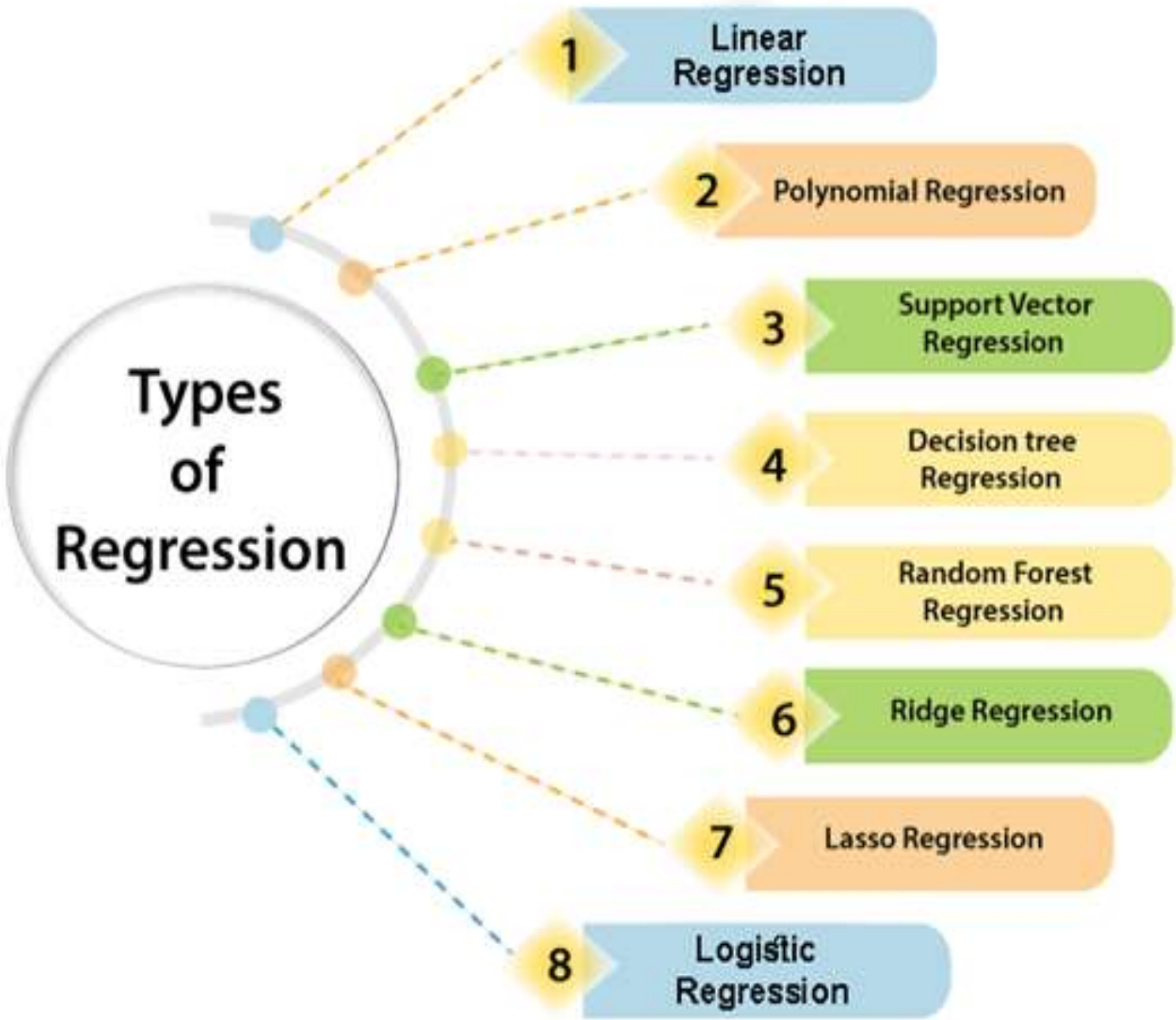
Types of Regression

- There are various types of regressions that are used in data science and machine learning.

- Each type has its importance in different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables.

- Some important types of regression which are given below:

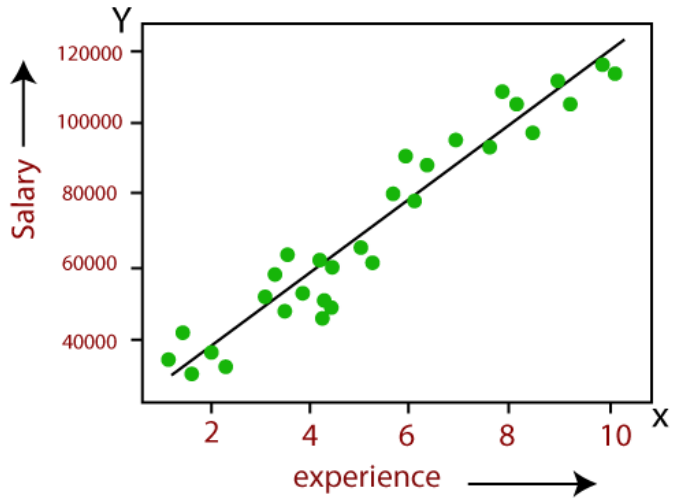
  - **Linear Regression**
  - **Logistic Regression**
  - **Polynomial Regression**
  - **Support Vector Regression**
  - **Decision Tree Regression**
  - **Random Forest Regression**
  - **Ridge Regression**
  - **Lasso Regression:**


Types of Regression
1 Linear Regression
2 Polynomial Regression
3 Support Vector Regression
4 Decision tree Regression
5 Random Forest Regression
6 Ridge Regression
7 Lasso Regression
8 Logistic Regression

Linear Regression

- Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors).
- It is commonly used for prediction, trend analysis, and understanding relationships between variables.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.

- The relationship between variables in the linear regression model can be explained using the given image.



- Here we are predicting the salary of an employee based on **the year of experience**.

- The mathematical equation for Linear regression:

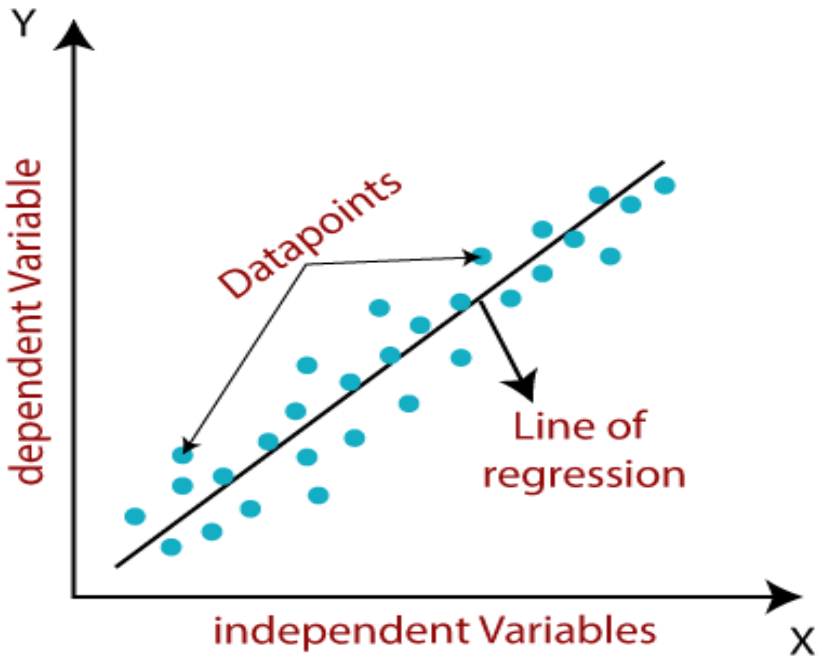
$$Y = aX + b$$

- **Here,**

  Y= dependent variables(target variables),
  X= Independent variables (predictor variables),

  a and b are the linear coefficients

- The linear regression model provides a sloped straight line representing the relationship between the variables.



How Linear Regression Works

- The model finds the best-fit line that minimizes the **sum of squared errors (SSE)** using **Ordinary Least Squares**
- The coefficients are estimated to minimize the error between predicted and actual value
- **Variables:**
  - **Dependent Variable (Y):** The thing you want to predict (for example, test scores).
  - **Independent Variable (X):** The factor you think influences the dependent variable (for example, hours of study).

- **The basic idea is to find a straight line that best fits your data points. The line is represented by:**

$$Y=mX+c$$

- **Goal of Linear Regression:**

  The aim is to find the best values for "m" and **" c"** so that the line is as close as possible to all the data points. The "best" line minimizes the differences between the actual data points and the predictions.

Types of Linear Regression

- Linear regression can be further divided into two types of the algorithm:
- **Simple Linear Regression:**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Types of Linear Regression

Simple Linear Regression – One independent variable is used to predict a dependent variable.

- $Y=b0+b1X+\epsilon$

Where:

- Y is the dependent variable
- X is the independent variable
- b0 is the intercept
- b1 is the slope (coefficient)
- ϵepsilonϵ is the error term

Multiple Linear Regression

Multiple independent variables are used to predict the dependent variable.

$$Y=b0+b1X1+b2X2+\cdots+bnXn+\epsilon$$

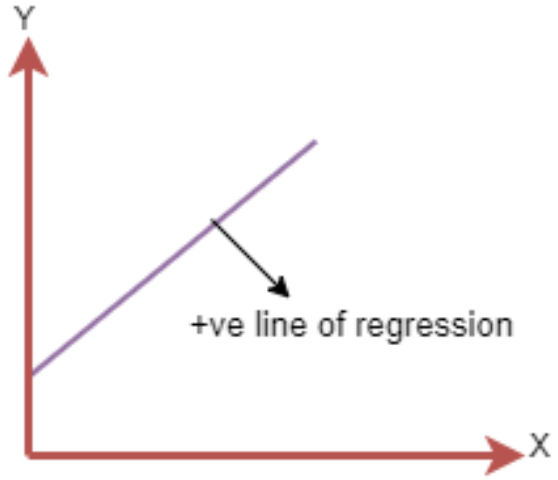
Regression line

- A linear line showing the relationship between the dependent and independent variables is called a **regression line**.

- A regression line can show two types of relationship:
  - ✓ Positive Linear Relationship
  - ✓ Negative Linear Relationship

Positive Linear Relationship:

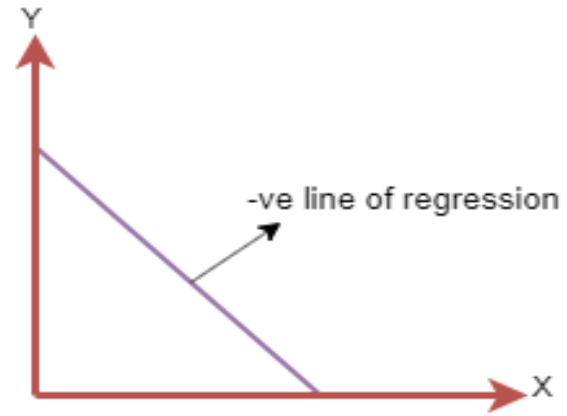
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Finding the Best Fit Line

- When working with linear regression**, our main goal is to find the best fit line which means the error between predicted values and actual values should be minimized**. The best fit line will have the least error.

- The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for $\mathbf{a_0}$ and $\mathbf{a_1}$ to find the best fit line, so to calculate this we use **cost function.**

Cost Function

- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.

- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as the **Hypothesis function**.

- For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error that occurred between the predicted values and actual values.

- It can be written as:

$$MSE = 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1 x_i + a_0))^2$$

Where,

N=Total number of observation

Yi = Actual value

(a1xi+a0)= Predicted value

- Measuring Model Performance:
  - MSE quantifies the average squared difference between actual and predicted values.
  - A lower MSE indicates better model accuracy.
  - MSE provides a standardized way to compare different regression models.

- **Residuals:** The distance between the actual value and predicted values is called **residual**.

- If the observed points are far from the regression line, then the residual will be high, and so the cost function will be high.

- If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Sum of Squared Errors (SSE) in Linear Regression

- When we fit a **linear regression** model, our goal is to find the best-fitting line that predicts values as accurately as possible. However, there will always be some differences (errors) between the predicted values and the actual values. One way to measure these errors is by using the **Sum of Squared Errors (SSE)**.

- An **error** (or residual) is the difference between the actual value (Yi) and the predicted value from the regression line.
- Yi = Actual value
- Y^i = Predicted value from the regression equation

Sum of Squared Errors (SSE) Formula

- To measure how well the regression line fits the data, we **square** each error (to make all values positive) and then sum them up:
- SSE=∑(Yi−Y^i)2
- where:
- Yi = Actual observed value
- Y^i = Predicted value from the regression line
- The **squared** term ensures that larger errors contribute more to the total error.

Hours Studied (X)	Test Score (Y)
1	50
2	60
3	65
4	70

Suppose our **regression equation** is:
$\hat{Y}=45+6X$

Now, we calculate SSE:

X	Y (Actual)	$\hat{Y}$ (Predicted)	Error $(Y - \hat{Y})$	Squared Error
1	50	$45 + 6(1) = 51$	$50 - 51 = -1$	$(-1)^2 = 1$
2	60	$45 + 6(2) = 57$	$60 - 57 = 3$	$3^2 = 9$
3	65	$45 + 6(3) = 63$	$65 - 63 = 2$	$2^2 = 4$
4	70	$45 + 6(4) = 69$	$70 - 69 = 1$	$1^2 = 1$

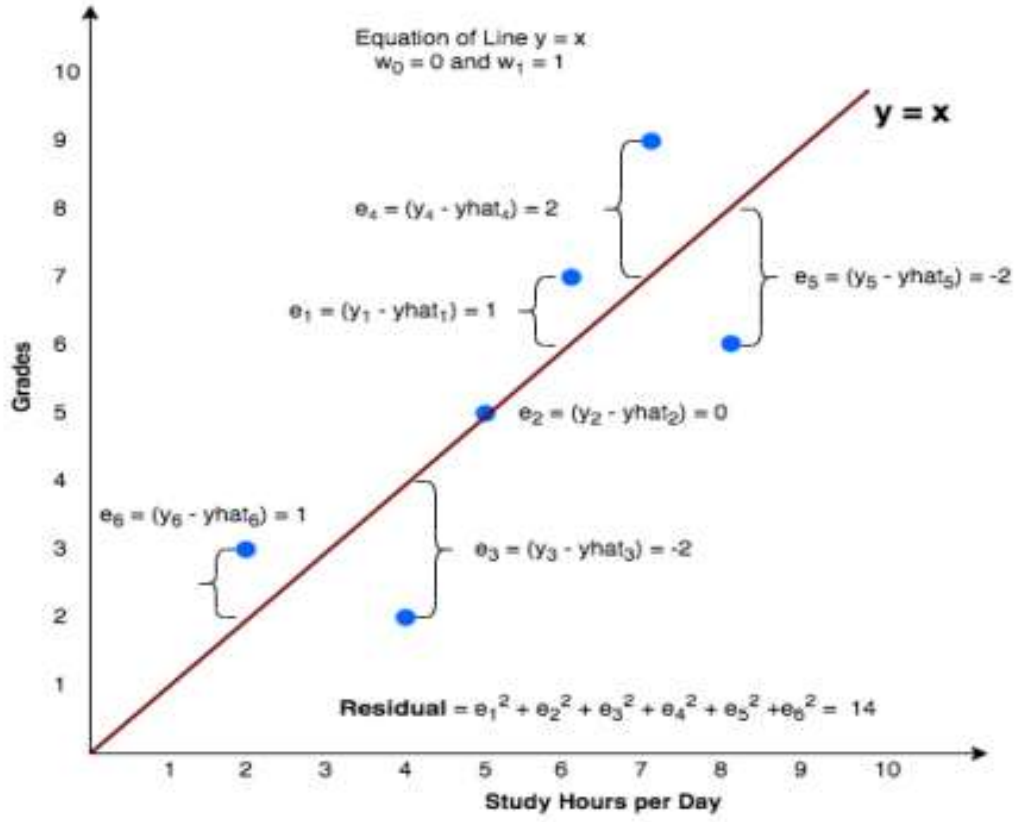
$$SSE = 1 + 9 + 4 + 1 = 15$$

How SSE is Used

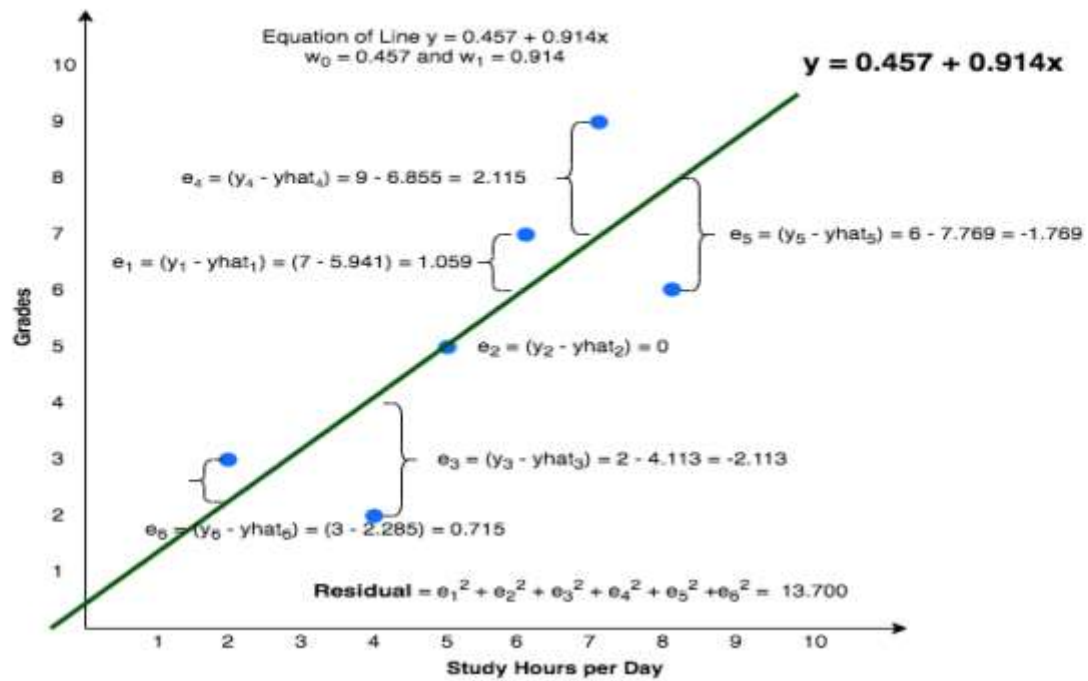
- The smaller the SSE, the better the regression model fits the data.
- However, **SSE alone is not enough** to compare models with different numbers of variables. Other metrics like **Mean Squared Error (MSE)** or **Root Mean Squared Error (RMSE)** are also used.

Best fit line

- Best fit line tries to explain the variance in given data. (minimize the total residual/error)



❖ Best fit line tries to explain the variance in given data. (minimize the total residual/error)



Linear Regression- Methods to Get Best

- Least Square
- Gradient Descent

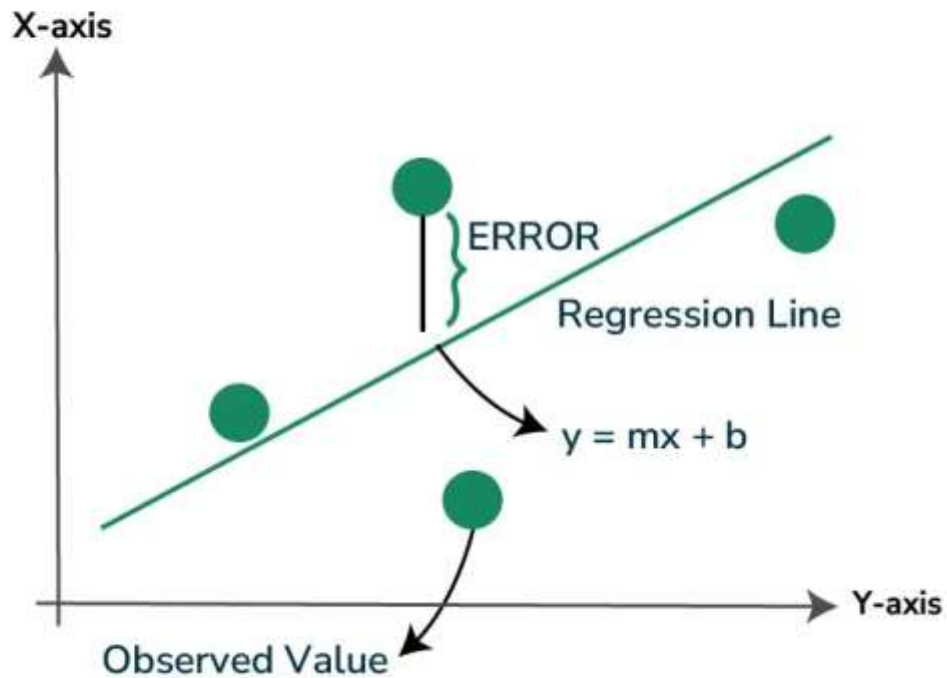
Least Square Method

- In statistics, when we have data in the form of data points that can be represented on a cartesian plane by taking one of the variables as the independent variable represented as the x-coordinate and the other one as the dependent variable represented as the y-coordinate, it is called **scatter data**.

- This data might not be useful in making interpretations or predicting the values of the dependent variable for the independent variable where it is initially unknown.

- So, we try to get an equation of a line that fits best to the given data points with the help of the **Least Square Method**

Least Square Method



- The least-squares method can be defined as a statistical method that is used to find the equation of the line of best fit related to the given data.

- This method aims at reducing the sum of squares of deviations as much as possible. The line obtained from such a method is called a **regression line**.

Steps:

- **Step 1:** Denote the independent variable values as $\mathbf{x_i}$ and the dependent ones as $\mathbf{y_i}$.

- **Step 2:** Calculate the average values of $\mathbf{x_i}$ and $\mathbf{y_i}$ as **X** and **Y**.

- **Step 3:** Presume the equation of the line of best fit as $\mathbf{y = mx + c}$, where **m** is the **slope** of the line and **c** represents the **intercept** of the line on the Y-axis.

- **Step 4:** The slope **m** can be calculated from the following formula:
- The Least Square Regression line formula is
-

ax + b

$$\text{Slope } a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$\text{Intercept } b = \frac{(\sum y) - a(\sum x)}{n}$$

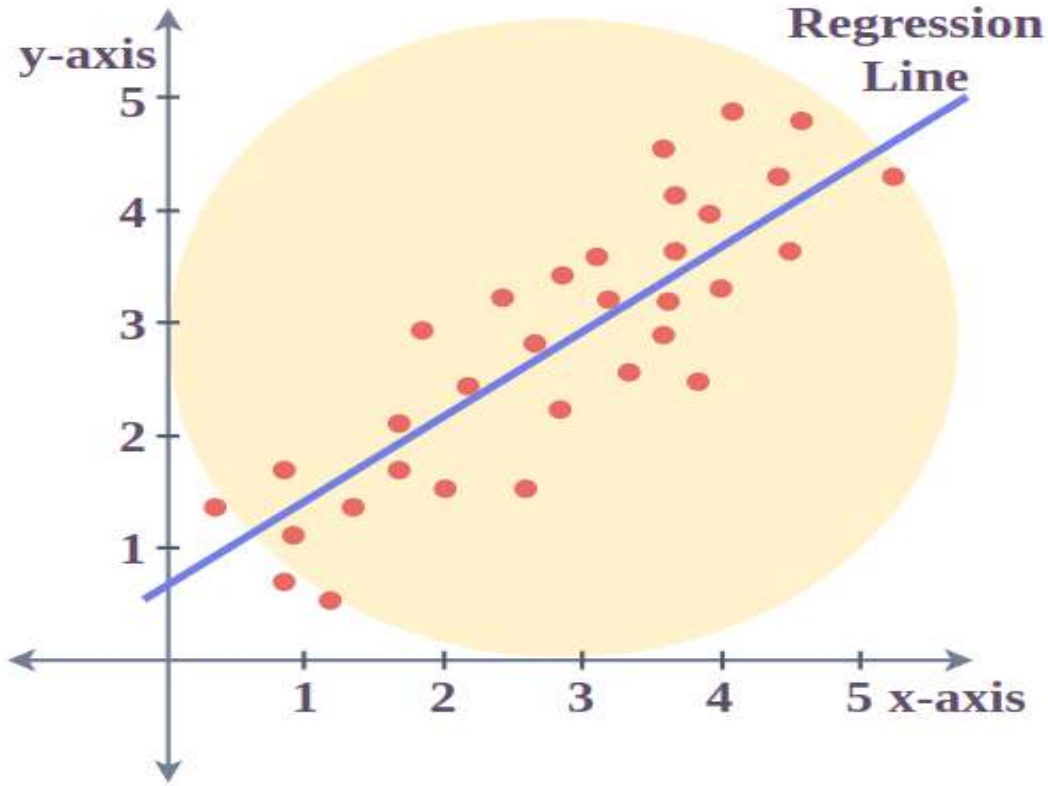
Where:

- n is the number of data points,
- $\sum xy$ is the sum of the product of each pair of x and y values,
- $\sum x$ is the sum of all x values,
- $\sum y$ is the sum of all y values,
- $\sum x^2$ is the sum of the squares of x values.

These formulas are used to calculate the parameters of the line that best fits the data according to the criterion of the least squares, minimizing the sum of the squared differences between the observed values and the values predicted by the linear model.

Least Square Method Graph

- Let us have a look at how the data points and the line of best fit obtained from the least squares method look when plotted on a graph.



- The red points in the above plot represent the data points for the sample data available. **Independent variables are plotted as x-coordinates** and **dependent ones are plotted as y-coordinates**. **The equation of the line of best fit obtained from the least squares method is plotted as the red line in the graph.**

- We can conclude from the above graph that how the least squares method helps us to find a line that best fits the given data points and hence can be used to make further predictions about the value of the dependent variable where it is not known initially.

- The least squares method assumes that the data is evenly distributed and doesn’t contain any outliers for deriving a line of best fit.

- However, this method doesn’t provide accurate results for unevenly distributed data or data containing outliers.

Problem - 1

- Sam found how many **hours of sunshine** vs how many **ice creams** were sold at the shop from Monday to Friday:

"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Find the number of ice cream to be made by Sam if the hours of sunshine are given as 8 hours.

Solution

- Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$$y = mx + b$$

Step 1: For each (x,y) calculate x^2 and xy:

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

Step 2: Sum x, y, x^2 and xy (gives us Σx, Σy, Σx^2 and Σxy):

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**Σx^2: 168**	**Σxy: 263**

Also **N** (number of data values) = **5**

Step 3: Calculate Slope **m**:

$$\mathbf{m} = \frac{N\,\Sigma(xy) - \Sigma x\,\Sigma y}{N\,\Sigma(x^2) - (\Sigma x)^2}$$

$$= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 26^2}$$

$$= \frac{1315 - 1066}{840 - 676}$$

$$= \frac{249}{164} = \mathbf{1.5183...}$$

Step 4: Calculate Intercept **b**:

$$\mathbf{b} = \frac{\Sigma y - m\,\Sigma x}{N}$$

$$= \frac{41 - 1.5183 \times 26}{5}$$

$$= \mathbf{0.3049\ldots}$$

Step 5: Assemble the equation of a line:

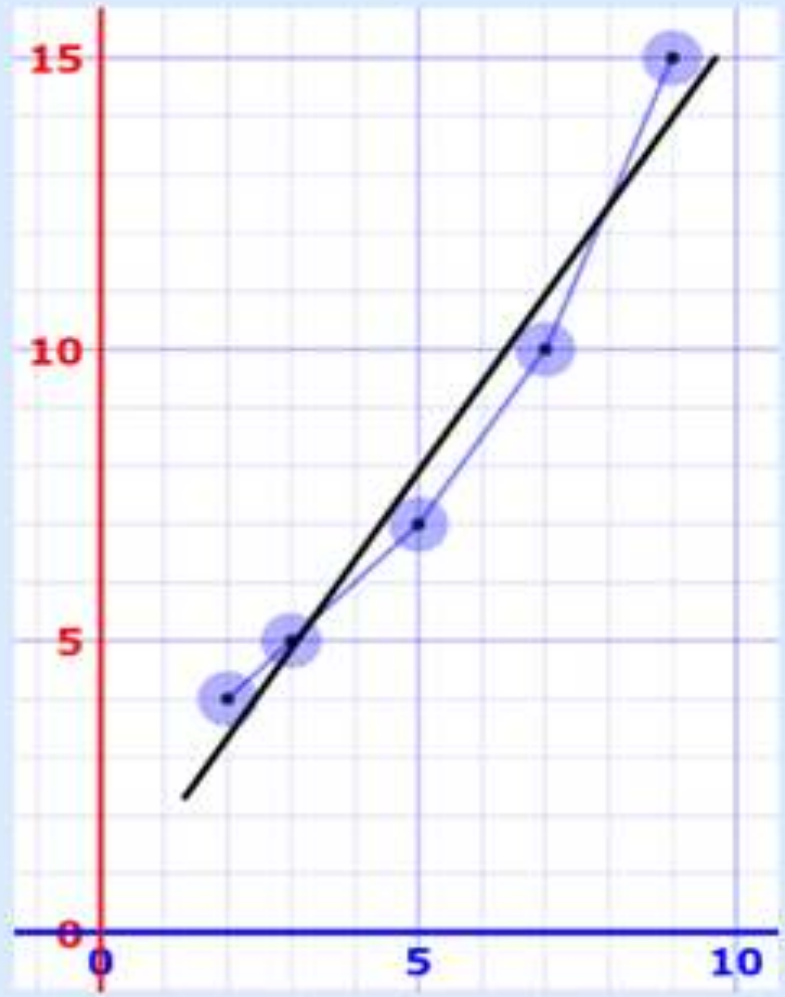
$$y = mx + b$$

$$y = 1.518x + 0.305$$

Let's see how well it works:

x	y	y = 1.518x + 0.305	error
2	4	3.34	−0.66
3	5	4.86	−0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	−1.03

Here are the (x,y) points and the line **y = 1.518x + 0.305** on a graph:



Nice fit!

Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1.518 \times 8 + 0.305 = 12.45 \text{ Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case. Yum.

Problem - 2

- Let's consider the data from the hours of studying along with the score obtained. Use the least squares regression line formulas to find the slope and constant for our model. Also, predict the score for studying 5 hours.

Hours (x)	Score (y)
1	11
3	16
4	15
6	20
8	18

Solution

- To start, we need to calculate the following sums:

  Σx, Σy, Σx^2, and the **Σxy.**

	Hours (x)	Score (y)	x^2	xy
	1	11	1	11
	3	16	9	48
	4	15	16	60
	6	20	36	120
	8	18	64	144
Sums	**22**	**80**	**126**	**383**
	Σx	**Σy**	**Σx^2**	**Σxy**

Next, we'll plug those sums into the slope formula.

$$m = \frac{5 * 383 - 22 * 80}{5 * 126 - 22^2} = \frac{155}{146} = 1.0616$$

Now that we have the slope (m), we can find the y-intercept (b) for the line.

$$b = \frac{80 - 1.0616 * 22}{5} = \frac{56.6448}{5} = 11.329$$

- Let’s plug the slope and intercept values in the least squares regression line equation:

$$y = 11.329 + 1.0616x$$

- Now, We can use this equation to make predictions.

- If we want to predict the score for **studying 5 hours**, we simply plug **x = 5** into the equation:

$$y = 11.329 + 1.0616 * 5 = 16.637$$

- Therefore, the model predicts that people studying for 5 hours will have an average test score of **16.637.**

Problem - 3

- Find the LSR line, y = ax + b for the following data. Also, estimate the value of y when x = 10

X	Y
0	2
1	3
2	5
3	4
4	6

Solution

- We get,

$$y = 0.9X + 2.2$$

Therefore, for x = 10, Y = 11.2

Problem -4

- **Find the line of best fit for the following data points using the least squares method:**

(x,y) = (1,3), (2,4), (4,8), (6,10), (8,15).

Important points to remember

- The least-squares method is used to predict the behavior of the dependent variable with respect to the independent variable.
- The sum of the squares of errors is called variance.
- The main aim of the least-squares method is to minimize the sum of the squared errors.

Multiple Regression

- Multiple regression is an extension of simple linear regression. It’s used to model the relationship between one dependent variable and two or more independent variables.
- The primary purpose is to understand how the dependent variable changes as the independent variables change.

Mathematical Equation

The mathematical representation of multiple regression is:

$$Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \ldots + C_nX_n + e$$

where,

- **Y:** Dependent Variable (target variable)
- $X_1, X_2, X_3, \ldots, X_n$: Independent Variable (input variable)
- C_0: Intercept (value of Y when X=0)
- $C_1, C_2, C_3, C_4, C_5, \ldots, C_n$: Slope of line
- ***e*:** Error term

Assumptions of Multiple Regression

- **Linearity:** A linear relationship exists between the dependent and independent variables.
- **Independence:** Observations are independent of each other.
- **No multicollinearity:** Independent variables aren't too highly correlated with each other.
- **Homoscedasticity:** Constant variance of the errors.
- **No Autocorrelation:** The residuals (errors) are independent.
- **Normality:** The dependent variable is normally distributed for any fixed value of the independent variables.

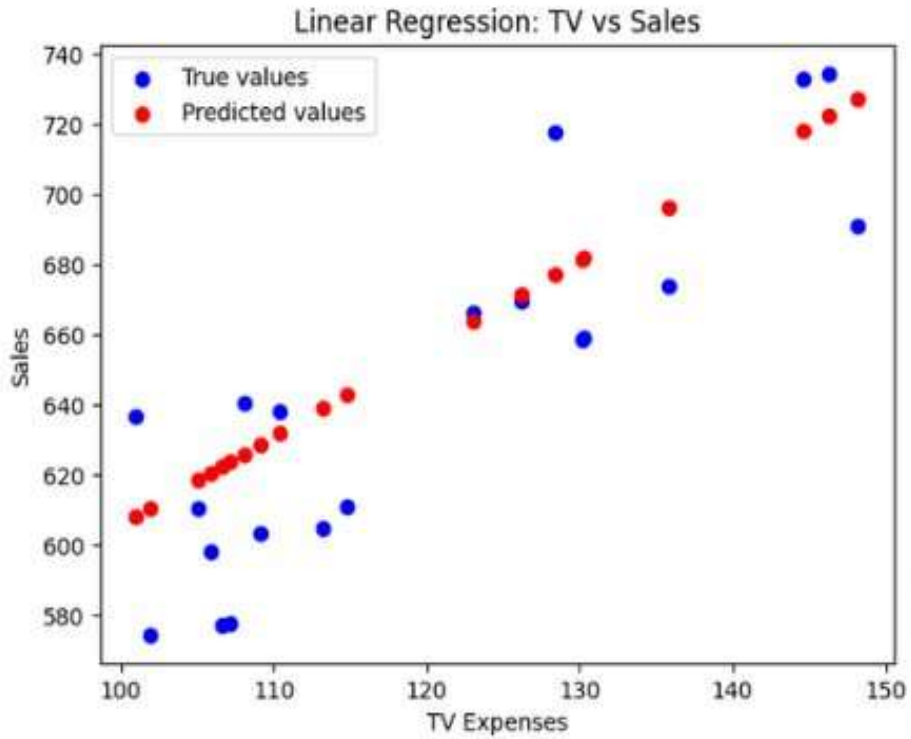
Limitations of Multiple Regression

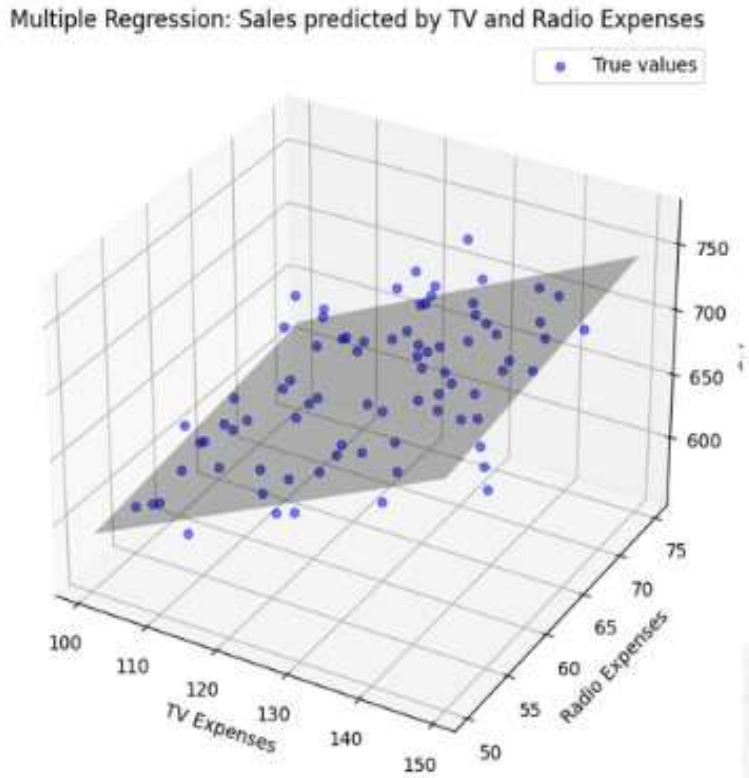
- **Overfitting:** Including too many independent variables can lead to a model that fits the training data too closely.
- **Omitted Variable Bias:** Leaving out a significant independent variable can bias the coefficients of other variables.
- **Endogeneity** occurs when an independent variable is correlated with the error term, leading to biased coefficient estimates.

Linear Regression vs Multiple Regression

Parameter	Linear (Simple) Regression	Multiple Regression
Definition	Models the relationship between one dependent and one independent variable.	Models the relationship between one dependent and two or more independent variables.
Equation	$Y = C_0 + C_1X + e$	$Y = C_0 + C_1X_1 + C_2X_2 + C_3X_3 + \ldots + C_nX_n + e$
Complexity	Simpler dealing with one relationship.	More complex due to multiple relationships.
Use Cases	Suitable when there is one clear predictor.	Suitable when multiple factors affect the outcome.
Assumptions	Linearity, Independence, Homoscedasticity, Normality	Same as linear regression, with the added concern of multicollinearity.
Visualization	Typically visualized with a 2D scatter plot and a line of best fit.	Requires 3D or multi-dimensional space, often represented using partial regression plots.
Risk of Overfitting	Lower, as it deals with only one predictor.	Higher, especially if too many predictors are used without adequate data.
Multicollinearity Concern	Not applicable, as there's only one predictor.	A primary concern; having correlated predictors can affect the model's accuracy and interpretation.
Applications	Basic research, simple predictions, understanding a singular relationship.	Complex research, multifactorial predictions, studying interrelated systems.

LINEAR REGRESSION VS MULTIPLE REGRESSION





Example of Linear and Multiple Regression

- Problem Statement: Suppose we have data for a retail company. The company wants to understand how their advertising expenses in various channels (e.g., TV, Radio) impact sales.

1. **Linear Regression**: Predict sales using only TV advertising expenses.

2. **Multiple Regression**: Predict sales using both TV and Radio advertising expenses.

Multiple Linear Regression Using Least Square Method

The **Multiple Linear Regression Model** with n independent variables is written as follows:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots\ldots\ldots + b_nX_n + u$$

Where,

Y = *The variable needs to be predicted (dependent variable)*

X = *The variable used to predict Y (independent variable)*

a = *The intercept*

b = *The slope*

u = *The regression residual*

MLR using LSM for Two Independent Variables

- Regression of two independent variables can be predicted by using the below formulas such as

- **Intercepts (a)**
- **Regression Coefficients (b1, b2)**

$$Intercepts\ a = \bar{Y} - b_1(\bar{X}_1) - b_2(\bar{X}_2)$$

Regression Coefficients (b1, b2)

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Where,

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

$$\overline{Y} = \frac{\sum Y}{N}$$

$$\overline{X}_1 = \frac{\sum X_1}{N}$$

$$\overline{X}_2 = \frac{\sum X_2}{N}$$

Problem - 1

Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Use the following steps to fit a multiple linear regression model to this dataset. `using LSM`

Solution

Step 1: Calculate X_1^2, X_2^2, X_1y, X_2y and X_1X_2.

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

X_1^2	X_2^2	X_1y	X_2y	X_1X_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
38767	2823	101895	25364	9859

Sum

Step 2: Calculate Regression Sums.

Next, make the following regression sum calculations:

- $\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38{,}767 - (555)^2 / 8 =$ **263.875**
- $\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2{,}823 - (145)^2 / 8 =$ **194.875**
- $\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101{,}895 - (555^*1{,}452) / 8 =$ **1,162.5**
- $\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25{,}364 - (145^*1{,}452) / 8 =$ **-953.5**
- $\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9{,}859 - (555^*145) / 8 =$ **-200.375**

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

X_1^2	X_2^2	X_1y	X_2y	X_1X_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
38767	2823	101895	25364	9859

Sum

Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 3: Calculate a , b_1, and b_2.

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$a = \overline{Y} - b_1(\overline{X}_1) - b_2(\overline{X}_2)$$

Thus, $b_1 = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{3.148}$

$b_2 = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{-1.656}$

$a = 181.5 - 3.148(69.375) - (-1.656)(18.125) = \mathbf{-6.867}$

Step 5: Place a b_1, and b_2 in the estimated linear regression equation.

The estimated linear regression equation is: $\hat{y} = a + b_1 * x_1 + b_2 * x_2$

In our example, it is $\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$

Problem - 2

- Predict the value of Y for subject 6 from the given dataset that contains values for X1, X2, and Y by using LSM on the Multiple Regression Model.

Subject	Y	X1	X2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Solution

Step 1 :- First, calculate all the values required in the above formulae.

Subject	Y	X_1	X_2	X_1X_2	X_1X_1	X_2X_2	X_1Y	X_2Y
1	-3.7	3	8					
2	3.5	4	5					
3	2.5	5	7					
4	11.5	6	3					
5	5.7	2	1					

Subject	Y	X_1	X_2	X_1X_2	X_1X_1	X_2X_2	X_1Y	X_2Y
1	-3.7	3	8	24	9	64	-11.1	-29.6
2	3.5	4	5	20	16	25	14	17.5
3	2.5	5	7	35	25	49	12.5	17.5
4	11.5	6	3	18	36	9	69	34.5
5	5.7	2	1	2	4	1	11.4	5.7
SUM	**19.5**	**20**	**24**	**99**	**90**	**148**	**95.8**	**45.6**

- **Step 2:** - Then put these values into the below mentioned formulae to get the exact predictable values to calculate Regression Coefficients b1 and b2

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 90 - \frac{20 \times 20}{5} = 10$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 148 - \frac{24 \times 24}{5} = 32.8$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N} = 95.8 - \frac{20 \times 19.5}{5} = 17.8$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N} = 45.6 - \frac{24 \times 19.5}{5} = -48$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 99 - \frac{20 \times 24}{5} = 3$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_1 = \frac{(32.8 \times 17.8) - (3 \times (-48)}{(10 \times 32.8) - (3)^2} = 2.2816$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_1 = \frac{(10 \times (-48)) - (3 \times 17.8)}{(10 \times 32.8) - (3)^2} = -1.672$$

- **Step 3 : -** Calculate the value of ***Intercept a***

$$a = \bar{Y} - b_1(\bar{X}_1) - b_2(\bar{X}_2)$$

$$= \frac{19.5}{5} - \frac{2.2816 \times 20}{5} - \frac{(-1.672 \times 24)}{5} = 2.796$$

•Step 4 : -

The final Regression Equation or Model looks as follows:

$$Y = 2.796 + 2.28x_1 - 1.67x_2$$

Therefore, for given **x1= 3** and **x2 = 2**, the value of **Y = ?** calculated as follows:

$$Y = 2.796 + (2.28 \times 3) - (1.67 \times 2)$$

$$Y = 6.296$$

Problem - 3

x1 Product 1 Sales	x2 Product 2 Sales	Y Weekly Sales
1	4	1
2	5	6
3	8	8
4	2	12

Solve using MLR by applying LSM

Solution

$$a_0 = -1.69$$
$$a_1 = 3.48$$
$$a_2 = -0.05$$

- $y = a_0 + a_1x_1 + a_2x_2$
- Hence, the constructed model is:
- $y = -1.69 + 3.48x_1 - 0.05x_2$

2. Gradient Descent -

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

- In models like polynomial regression cost function becomes highly complex and non-linear. Analytical solutions are not available. **That’s where gradient descent plays an important role, it** is an optimization algorithm used to minimize the cost function by iteratively updating the model parameters.
- It works well even for:Large datasets.
- Complex, high-dimensional problems.

How Does Gradient Descent Work in Linear Regression?

- **Initialize Parameters**: Start with random initial values for the slope (m*m*) and intercept (b*b*).
- **Calculate the Cost Function**: Compute the error using a cost function such as Mean Squared Error (MSE)

$$J(m,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i-(mx_i+b))^2$$

- **Compute the Gradient**: Find the gradient of the cost function with respect to m and b. These gradients indicate how the cost changes when the parameters are adjusted.

For m (slope): $\frac{\partial J}{\partial m} = -\frac{2}{n}\sum_{i=1}^{n} x_i(y_i-(mx_i+b))$

For b (intercept): $\frac{\partial J}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n}(y_i-(mx_i+b))$

- **Update Parameters**: Adjust m*m* and b*b* in the direction that reduces the cost:

$$\text{For } m \text{ (slope): } m = m - \alpha \cdot \frac{\partial J}{\partial m}$$

$$\text{For } b \text{ (intercept): } b = b - \alpha \cdot \frac{\partial J}{\partial b}$$

- **Repeat**: Iterate until the cost function converges i.e further updates make little or no difference.

Regularization

- Regularization is a technique used to **prevent overfitting** by adding a penalty term to the loss function. It helps improve a model's **generalization** to unseen data by controlling the complexity of the model.
- In **linear regression**, the best-fit line is found by minimizing the **Sum of Squared Errors (SSE)**. However, when a model is too complex (i.e., too many features or high-degree polynomials), it can fit the **training data perfectly** but fail on new data (overfitting).
- Regularization **adds a penalty term** to prevent extreme coefficients, reducing overfitting and making the model more stable.

Types of Regularization in Regression

1. Ridge Regression (L2 Regularization)

- Adds **L2 penalty**: sum of squared weights to the loss function.
- Encourages **smaller** but nonzero coefficients.

$$\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum w_j^2$$

- **Effect:**
- Reduces large coefficients but does **not** force them to zero.
- Suitable when all features are important.

- **Lasso Regression (L1 Regularization)**
- Adds **L1 penalty**: sum of absolute values of weights.
- Encourages **sparse models** by forcing some coefficients to be exactly zero.

$$\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum |w_j|$$

- **Effect:**
- Eliminates unimportant features by setting their coefficients to **zero**.
- Useful for **feature selection**.

How Regularization Works in Linear Regression

- Without regularization: **Large coefficients** → Overfitting
- With regularization: **Penalizes large coefficients** → Better generalization