knowledgehut
upGrad

Introduction to Machine Learning and Linear Regression

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the end of this session, you will:

- Understand machine learning and its types
- Learn the difference between Supervised and Unsupervised ML
- Explore the difference between Regression and Classification
- Analyze Simple Linear Regression and its interpretation

Why Learn Machine Learning?

ML skills are in high demand across industries to make data-driven decisions, automate processes, improve customer experiences.

Multiple career opportunities, including data scientist, machine learning engineer, AI researcher, data analyst etc.

ML can help you develop cutting-edge applications and solutions across various fields, from healthcare to finance to entertainment.

ML powers recommendation systems that personalize user experiences, whether in online shopping, content streaming, or social media.

Pop Quiz

Q. One-way machine learning is used for which of the following in everyday life?

a. Baking bread

b. Tying shoelaces

c. Sending emails

d. Watering the plants

Pop Quiz

Q. One-way machine learning is used for which of the following in everyday life?

a. Baking bread

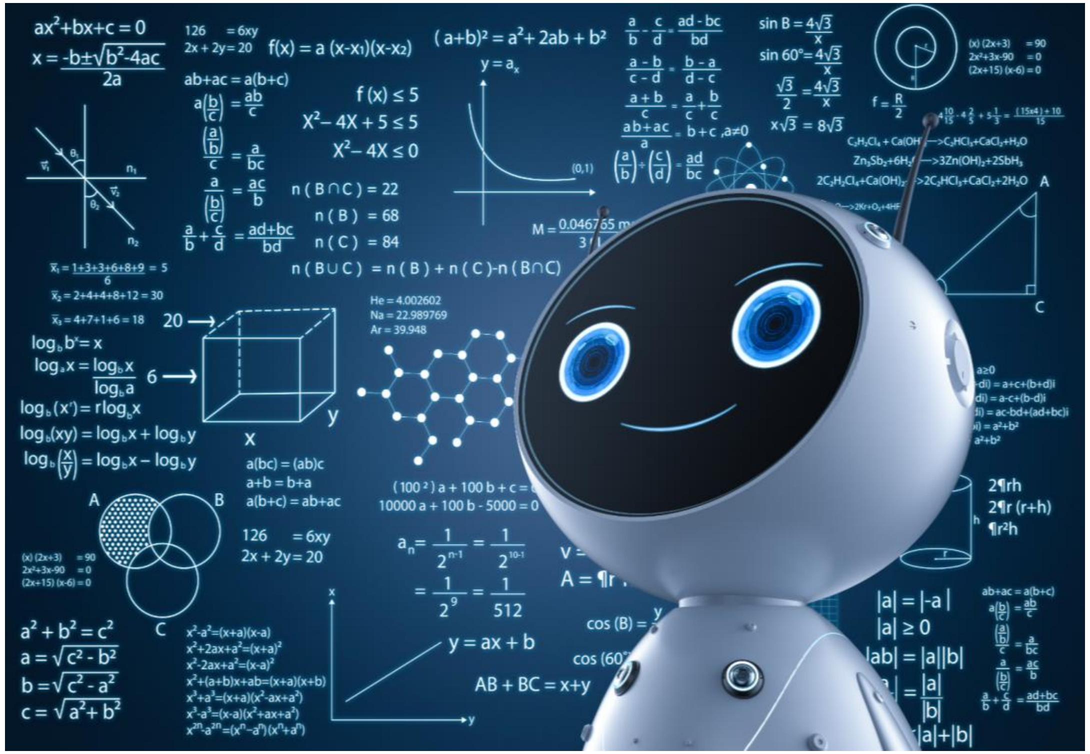
b. Tying shoelaces

[✓ **c. Sending emails**]

d. Watering the plants

Introduction to Machine Learning

What is Machine Learning?



1. Machine Learning is about teaching computers to learn from examples :

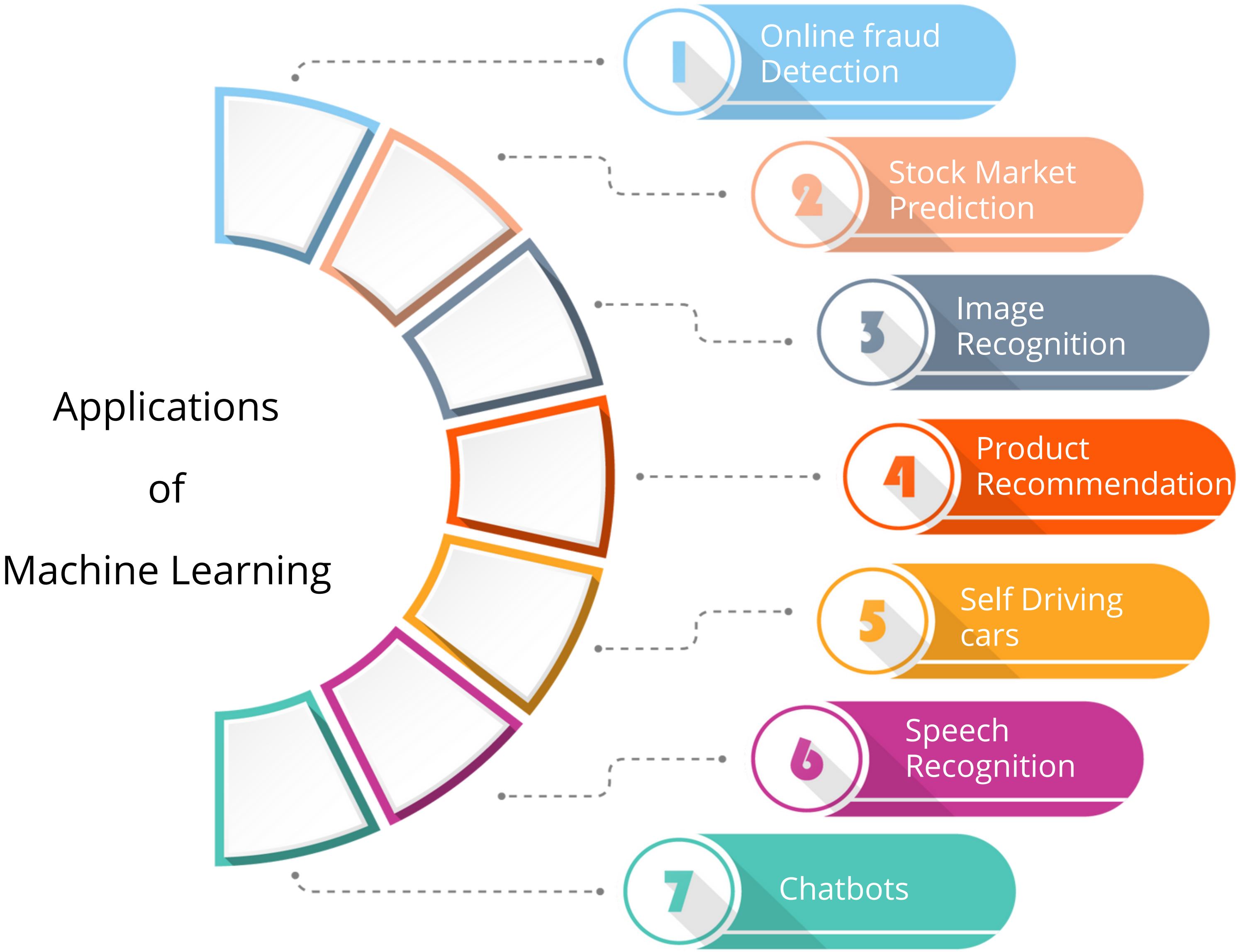
    a. So they can make predictions or decisions about new things they haven't seen before.

    b. To learn patterns and make decisions without being explicitly programmed for every single detail.

    c. To be capable of handling tasks that might be too complex or time-consuming for humans.

2. Example: Think like teaching a kid to recognize different fruits, but the "learner" here is a computer.

Applications of Machine Learning



Impact of Machine Learning

Machine Learning has a significant impact across various industries and domains, revolutionizing the way we approach tasks, make decisions, and solve problems.

01 **Social Impact**

Machine learning can be applied to address societal challenges, such as predicting disease outbreaks, optimizing resource distribution, and aiding in disaster response.

02 **Image and Video Analysis**

Machine learning has enabled the development of facial recognition, object detection, and image classification systems, with applications ranging from security to entertainment.

03 **Environmental Monitoring**

Machine learning aids in analyzing environmental data, such as satellite imagery, to monitor deforestation, climate change, and natural disaster prediction.

04 **Research and Discovery**

Machine learning helps scientists process and analyze large datasets in fields like astronomy, genomics, and materials science, leading to new discoveries and insights.

05 **Healthcare Advancement**

It has led to advancements in medical image analysis, disease diagnosis, drug discovery, and personalized treatment plans. It has also facilitated development of wearable health devices and remote patient monitoring.

Poll Time

Q. Which of the following tasks is commonly accomplished using machine learning?

a. Washing dishes

b. Reading a book

c. Identifying handwritten digits

d. Making a sandwich

Poll Time

Q. Which of the following tasks is commonly accomplished using machine learning?

a. Washing dishes

b. Reading a book

[✓ c. **Identifying handwritten digits**]

d. Making a sandwich

Introduction to Supervised and Unsupervised Learning

These are two fundamental approaches in the field of machine learning that involve different types of learning and data handling.

A Tale of Supervised Guidance and Unsupervised Discovery

Supervised ML

- It's like training a computer with examples to recognize patterns.
- It learns to make predictions or classifications on its own when given new, similar inputs.

Unsupervised ML

- It's about discovering interesting things in a big collection of things, without any knowledge.
- It's like finding hidden patterns in a puzzle without knowing what the complete picture looks like.

Supervised vs. Unsupervised ML

MACHINE LEARNING	
Supervised	**Unsupervised**
- It's like teaching a computer with a clear "supervision" or guidance.	- Giving computer a lot of data and letting it find patterns on its own without guidance.
- Provide computer with input data along with the correct answers which then learns to make predictions.	- There are no labeled answers; the computer explores the data and tries to group similar things together.
- Think of it as teaching a child to learn different animals and guess the new ones.	- Think of collecting various balls and expecting a kid to group similar color balls together.
- Examples: Email spam classification and fraud detection.	- Examples: Customer segmentation and product recommendations.

Use Cases of Supervised vs. Unsupervised ML

SUPERVISED ML

- Customer churn prediction
- Language translation
- Surveillance systems
- Medical image analysis
- Voice assistants
- Fraud detection

VS

UNSUPERVISED ML

- Customer segmentation
- Anomaly detection
- Geospatial Data Analysis
- E-commerce recommendations
- Climate change analysis
- Social Media pattern analysis

Pop Quiz

Q. What is the main difference between supervised and unsupervised machine learning?

a. Supervised learning uses labeled data, while unsupervised learning does not
b. Supervised learning doesn't use data, while unsupervised learning uses labeled data
c. Supervised learning requires human intervention, while unsupervised learning is fully automated
d. Supervised learning works only with images, while unsupervised learning works with text data

Pop Quiz

Q. What is the main difference between supervised and unsupervised machine learning?

a. **Supervised learning uses labeled data, while unsupervised learning does not**

b. Supervised learning doesn't use data, while unsupervised learning uses labeled data

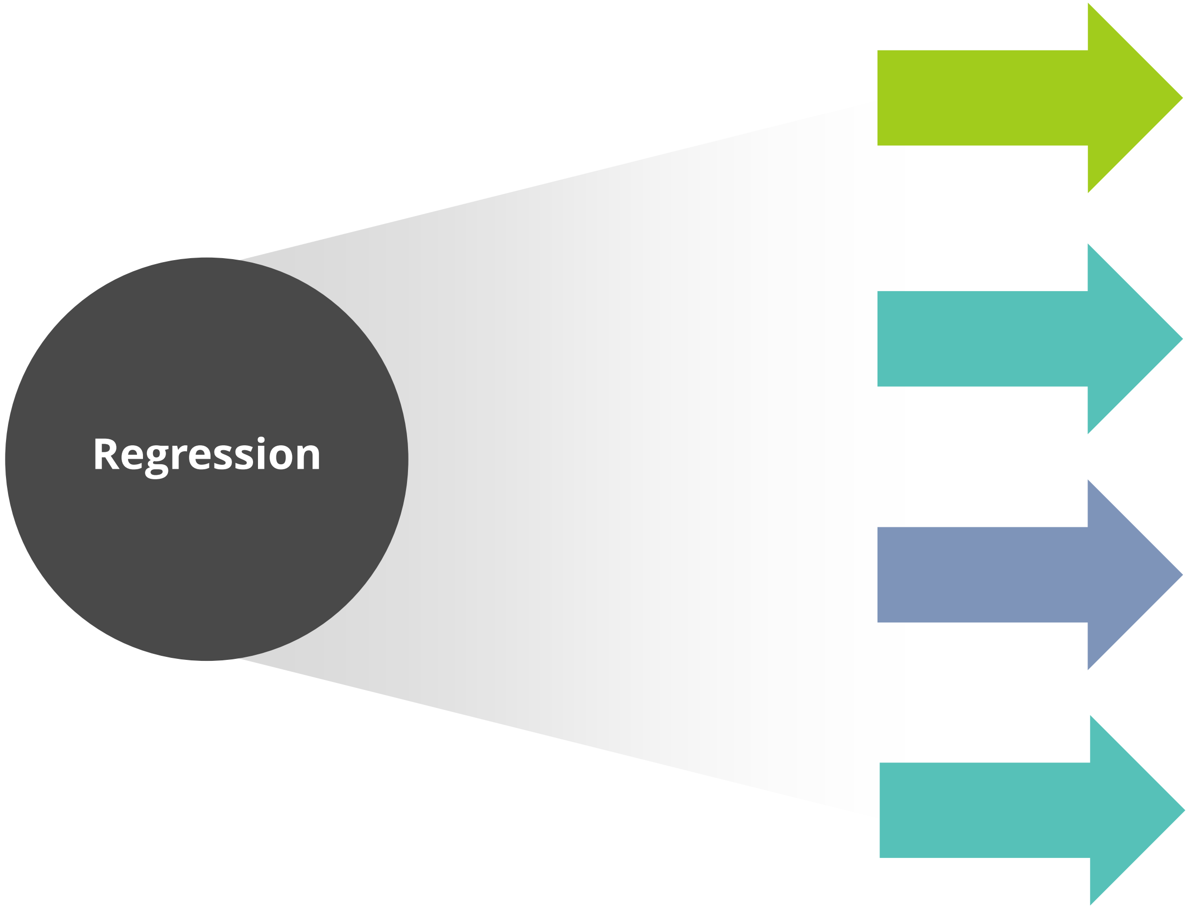
c. Supervised learning requires human intervention, while unsupervised learning is fully automated

d. Supervised learning works only with images, while unsupervised learning works with text data

Q&A

Introduction to Regression

Regression is a supervised machine learning technique used to predict or estimate a continuous numerical outcome based on input features.



Dependent Variable

Continuous numerical variable which you are trying to predict or explain.

Independent Variables

One or more variables that you use to predict or explain the behavior of the dependent variable.

Objective

Minimize difference between actual values and predicted values of dependent variable.

Applications

Prediction of continuous numerical variables like sales prediction, stock market forecasting, population growth projection, property value, etc.

Regression vs Classification

Regression

VS

Classification

Regression

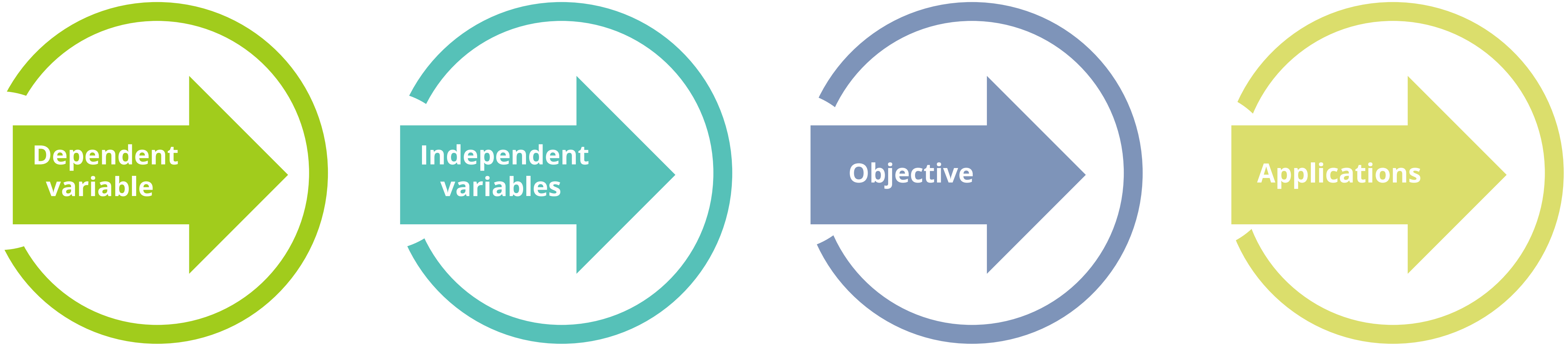
- ML technique to predict continuous numeric values
- Prediction of a quantity
- Goal is to establish a relationship between input variables and the continuous output variable
- E.g. : predicting stock prices, predicting a person’s weight,
  Predicting temperature

Classification

- ML technique to predict categorical discreet values
- Assigning labels to data points
- Goal is to categorize input data into specific classes based on patterns
- E.g. : predicting patient has a specific disease or not, email is spam or not, fraud detection

Introduction to Classification

Classification is a supervised machine learning technique used to categorize data into predefined classes or labels. Here, outcome is in the form of categories or classes.



Dependent variable

Categorical variable which you are trying to predict or explain.

Independent variables

One or more variables that you use to predict or explain the behavior of the dependent variable.

Objective

Accurately predict the class labels or categories of new or unseen data points.

Applications

Prediction of categorical variables like email spam detection, sentiment analysis, fraud detection, etc.

Poll Time

Q. What is the key difference between classification and regression?

a. Classification predicts continuous numeric values, while regression assigns data points to categories

b. Classification predicts continuous numeric values, while regression predicts labels for data points

c. Classification assigns data points to categories, while regression predicts continuous numeric values

d. Classification assigns data points to categories, while regression assigns data points to classes

Poll Time

Q. What is the key difference between classification and regression?

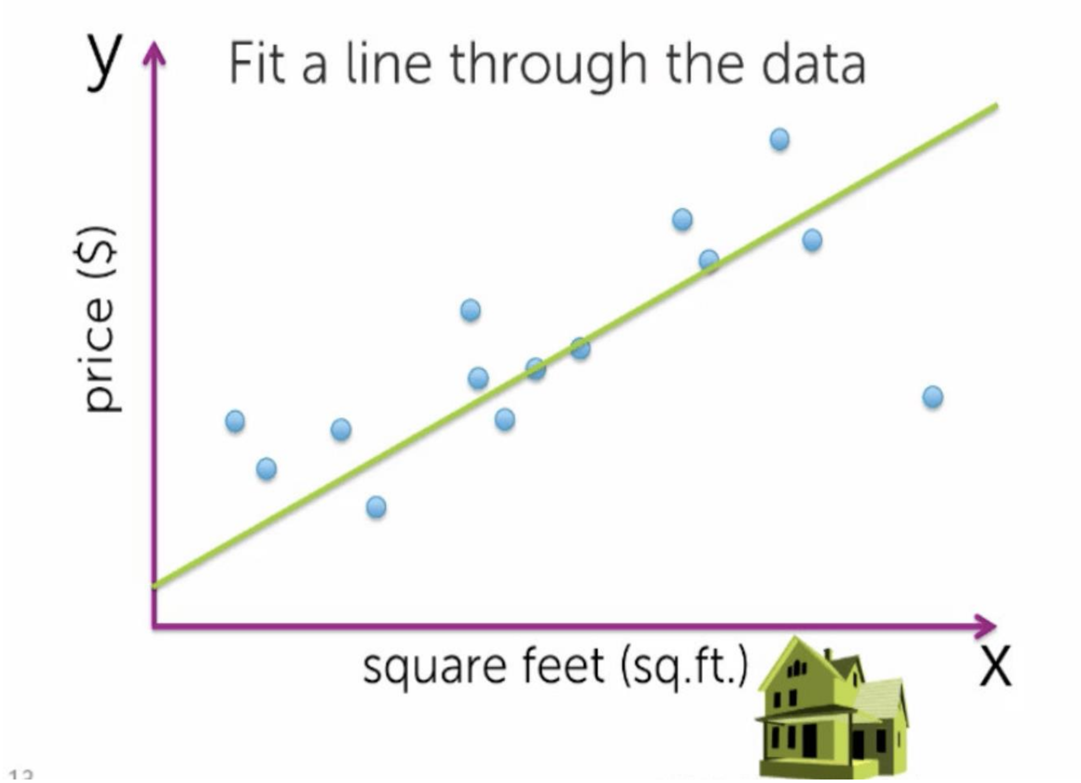
a. Classification predicts continuous numeric values, while regression assigns data points to categories

b. Classification predicts continuous numeric values, while regression predicts labels for data points

c. Classification assigns data points to categories, while regression predicts continuous numeric values

d. Classification assigns data points to categories, while regression assigns data points to classes

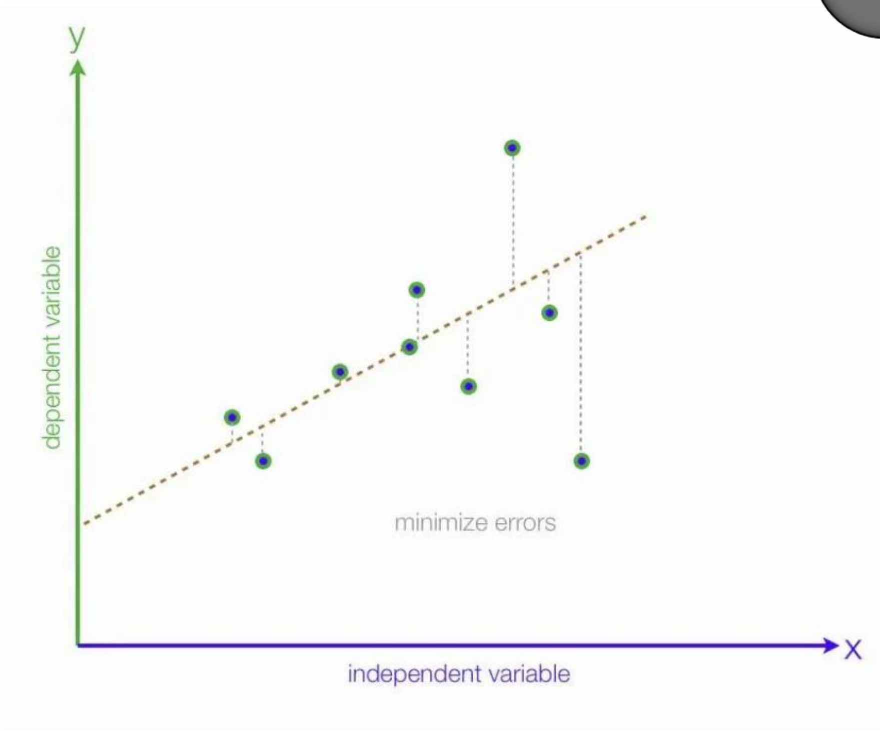
What is Linear Regression?



- A method to find relationship between two variables by drawing a straight line that represents how changes in one variable can predict changes in other.
- Example: Size of a house might be a big factor in deciding its price.
- Regression is like drawing a straight line through a set of values of house sizes and their corresponding prices.
- This line helps you estimate the price of a house based on its size.
- So, in simple terms, it's like finding a rule that helps you guess one thing based on the other.

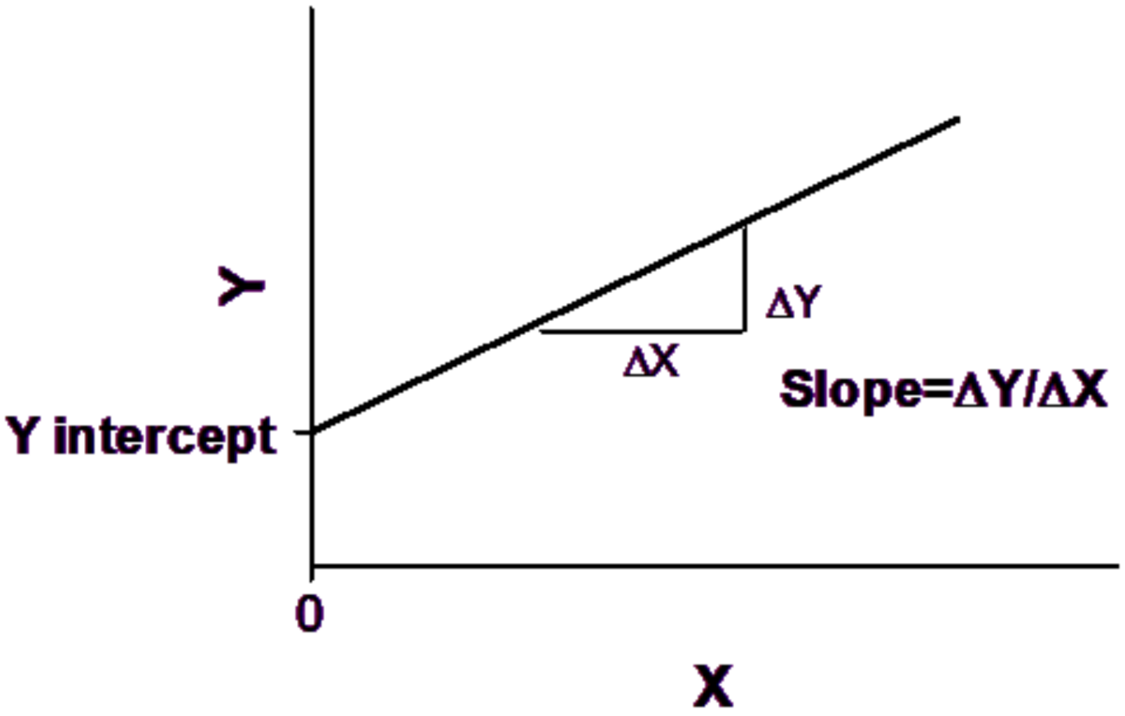
Linear Regression

Simple Linear Regression



- Method to understand and predict relationship between two variables.
- Goal is to find a straight-line equation that best represents the relation between two variables.
- We check how changes in independent variable might impact changes in dependent variable.
- Example: How change in number of hours (independent variable) will impact exam score (dependent variable).
- Simple linear regression would aim to find a line that best fits the set of values of number of hours & corresponding exam scores.
- "best fit" means finding straight line that minimizes errors between actual and predicted values.

Linear Regression : Slope and Intercept



Slope

- The slope in linear regression represents how much the dependent variable changes for a given change in the independent variable

Intercept

- The intercept is where the line crosses the y-axis. It's the value of the dependent variable when the independent variable is zero.

Pop Quiz

Q. In linear regression, what does the slope of the line represent?

a. The starting point of the line on the y-axis

b. The change in the dependent variable for a given change in the independent variable

c. The point where the line crosses the x-axis

d. The total number of data points in the dataset

Pop Quiz

Q. In linear regression, what does the slope of the line represent?

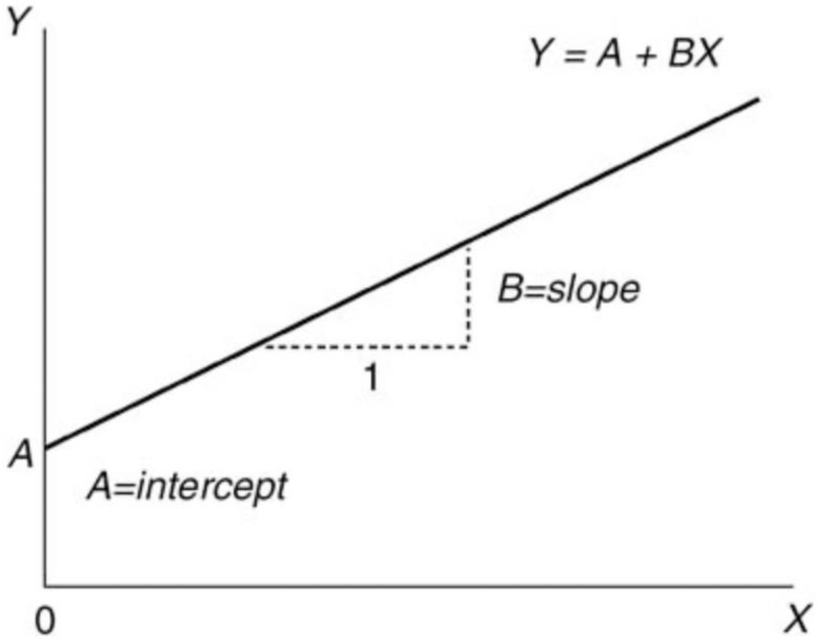
a. The starting point of the line on the y-axis

🟩 **b. The change in the dependent variable for a given change in the independent variable**

c. The point where the line crosses the x-axis

d. The total number of data points in the dataset

Interpretation of Slope



- Simple linear regression has a straight-line equation and is given as : Y = A + BX
  where X is independent variable, Y is dependent variable
  A is the intercept value and B is the slope value.

- The slope B represents how much dependent variable changes for a one-unit change in the independent variable.

- Slope tells you the direction (positive or negative) and steepness of the line.

- If value of slope is positive, it means if X increases, Y also increases. If its negative, it means if X increases, Y decreases.

- Large value of slope means steeper slope, indicating larger change in Y for given change in X and vice versa.

Calculation of Slope

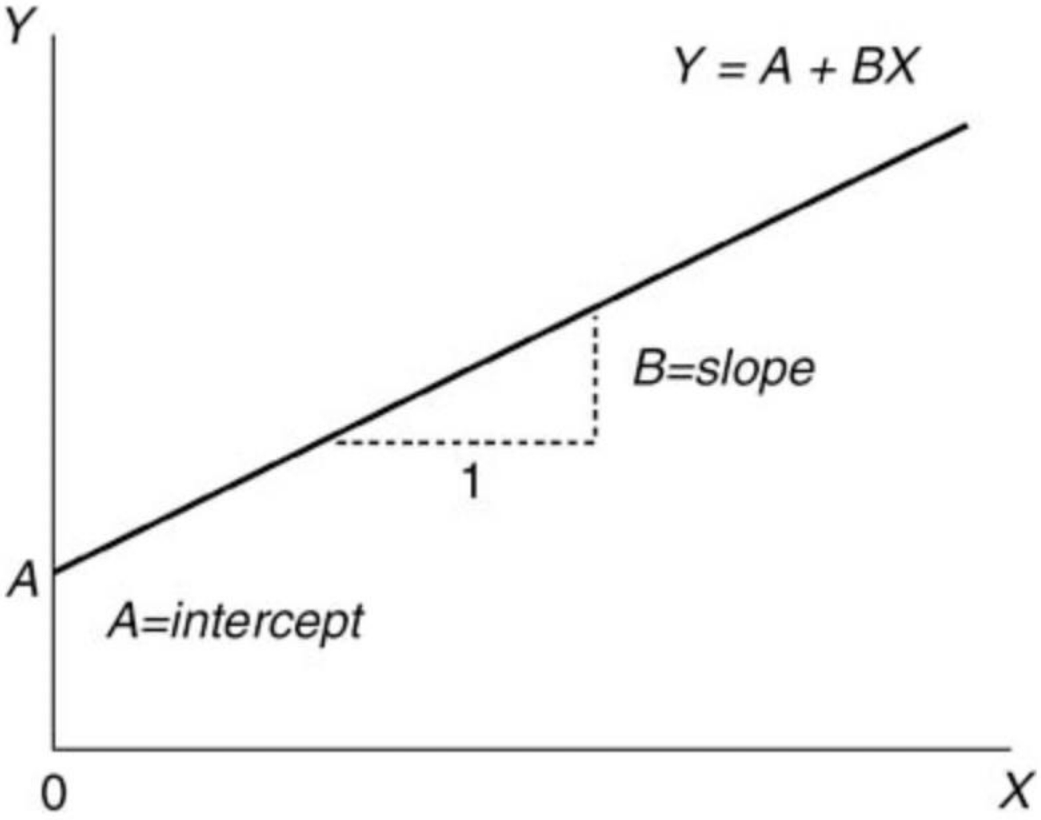
- Formula for Slope (B) in Simple Linear Regression is:

$$B = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$$

Where X represent independent variable and Y represent dependent variable.
X̄ and Ȳ represent the mean of all independent variable values and dependent variable values, respectively.

- Let's try to understand the numerator: $\sum(X-\bar{X})(Y-\bar{Y})$
  - Summation of this multiplication captures relationship between how X varies from its mean and how Y varies from its mean for each data point.
- If we look at the denominator: $\sum(X-\bar{X})^2$
  - Summation of this squaring the difference for each independent variable data point ensures that both positive and negative deviations contribute equally.

Interpretation of Intercept



- The intercept (A) is the value of the dependent variable (Y), when the independent variable (X) is zero.
- It's the point where the regression line crosses the y-axis.
- In the hours studied vs. exam score example, if intercept is 30, it means that if a student doesn't study at all (0 hours), their predicted exam score would be around 30.
- Interpretations of slope and intercept depend on the context of your data and variables you're working with.
- It's important to interpret them in a way that makes sense within the real-world scenario you're analyzing.

Poll Time

Q. In a simple linear regression equation (y = 2x + 10), what does the slope represent?

a. The starting point of the line on the y-axis

b. The change in the dependent variable (y) for a one-unit change in the independent variable (x)

c. The value of the dependent variable (y) when the independent variable (x) is zero

d. The overall spread of data points

Poll Time

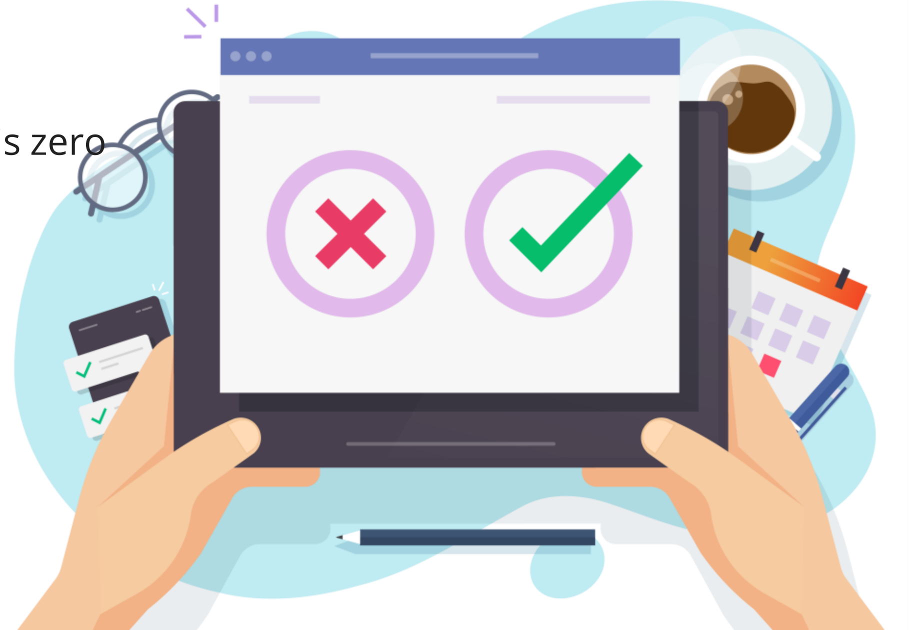
Q. In a simple linear regression equation (y = 2x + 10), what does the slope represent?

a. The starting point of the line on the y-axis

b. The change in the dependent variable (y) for a one-unit change in the independent variable (x)

c. The value of the dependent variable (y) when the independent variable (x) is zero

d. The overall spread of data points



Q&A
?
?
?
?
?
?
?
?

Activity 1

Pre-requisites:

Knowledge of difference between supervised and unsupervised machine learning.

Scenario:

You are a marketing executive at a finance firm, and you are trying to understand the differences between supervised and unsupervised learning as you are reading about it for the first time. You want to create 5 real world examples for both types of Machine Learning.

Data:

Consider any real-world use case in day-to-day life. Think of data related to finance, e-commerce, social media, stock market etc.

Expected Outcome:

A compiled a list of five real-world examples for both supervised and unsupervised learning, enriching your understanding of how these techniques are utilized.

Steps:

1) Research and understand the fundamental concepts and applications of each type of machine learning.
2) Identify and create a list of 5 real-world examples for each type of machine learning.
3) Enhance these examples with a brief description of each machine learning application.
4) Reflect on how machine learning techniques could enhance decision-making and strategy.

Activity 2

Pre-requisites:
Familiarity with interpreting linear regression slope and intercept mathematical relationships.

Scenario:

You are a manager at a local coffee shop. You're interested in understanding the relationship between the number of cups of coffee sold (Y) and the temperature in degrees Celsius (X). By applying simple linear regression, you'll calculate the slope and intercept to gain insights into how temperature affects coffee sales.

Data:

Consider the table shown on right side having temperature and corresponding coffee sales.

Temperature (X)	Coffee Sales (Y)
20	100
25	120
30	150
15	80

Expected Outcome:

Engage in hands-on calculations to determine the slope and intercept values. Understand how these values influence the interpretation of relationships between variables in a real-world business context.

Steps:

1) Use the formula of slope to calculate mean of X values and mean of Y values.
2) Calculate the numerator and denominator for the slope formula.
3) Calculate the slope (B) = Numerator / Denominator.
4) Use the calculated value of slope, mean values of X and Y to calculate intercept.
5) Interpret how slope and intercept values provide insights into how temperature influences coffee sales.

Summary

- ✓ Machine learning is a technique that empowers computers to learn patterns and make decisions from data.
- ✓ Supervised learning is effective for prediction tasks and unsupervised learning is useful for pattern discovery.
- ✓ Regression helps predict numerical outcomes while classification helps predict categorical outcomes.
- ✓ Linear regression is a predictive modeling technique. Simple linear regression is its basic form involving a single predictor variable and a continuous target variable.

Next Session:

Multiple Linear Regression

THANK YOU!

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut

upGrad

knowledgehut
upGrad

Multiple Linear Regression and OLS Assumptions

Pre-requisites

Hope you have gone through the self-learning content for this session on the PRISM portal.

By the End of this Session, You Will:

- Understand Multiple Linear Regression
- Evaluate how to find Beta coefficients
- Learn about the significance of gradient descent
- Apply the OLS Method and its assumptions

Recap

Poll Time

Q. In a simple linear regression, if the intercept (A) is 10 and the slope (B) is 3, what does it mean if the independent variable (X) is zero?

a. The dependent variable (Y) is 3

b. The dependent variable (Y) is 10

c. The dependent variable (Y) is 0

d. The dependent variable (Y) is undefined

Poll Time

Q. In a simple linear regression, if the intercept (A) is 10 and the slope (B) is 3, what does it mean if the independent variable (X) is zero?

a. The dependent variable (Y) is 3

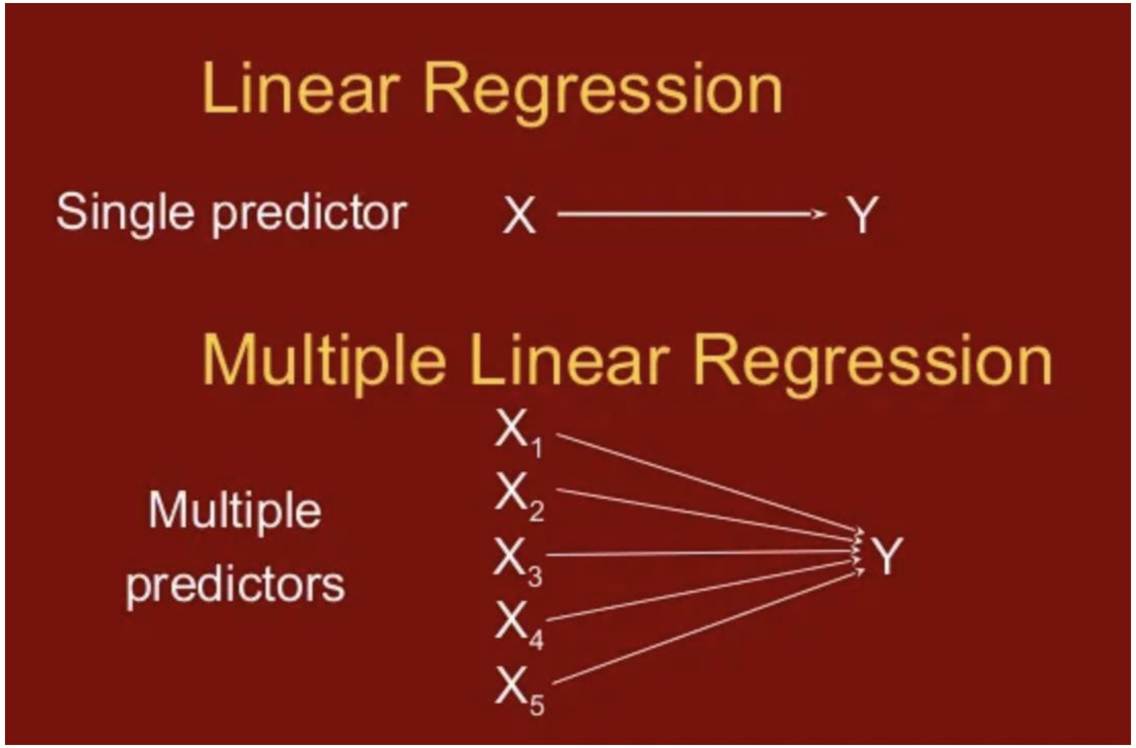
b. The dependent variable (Y) is 10

c. The dependent variable (Y) is 0

d. The dependent variable (Y) is undefined

Multiple Linear Regression

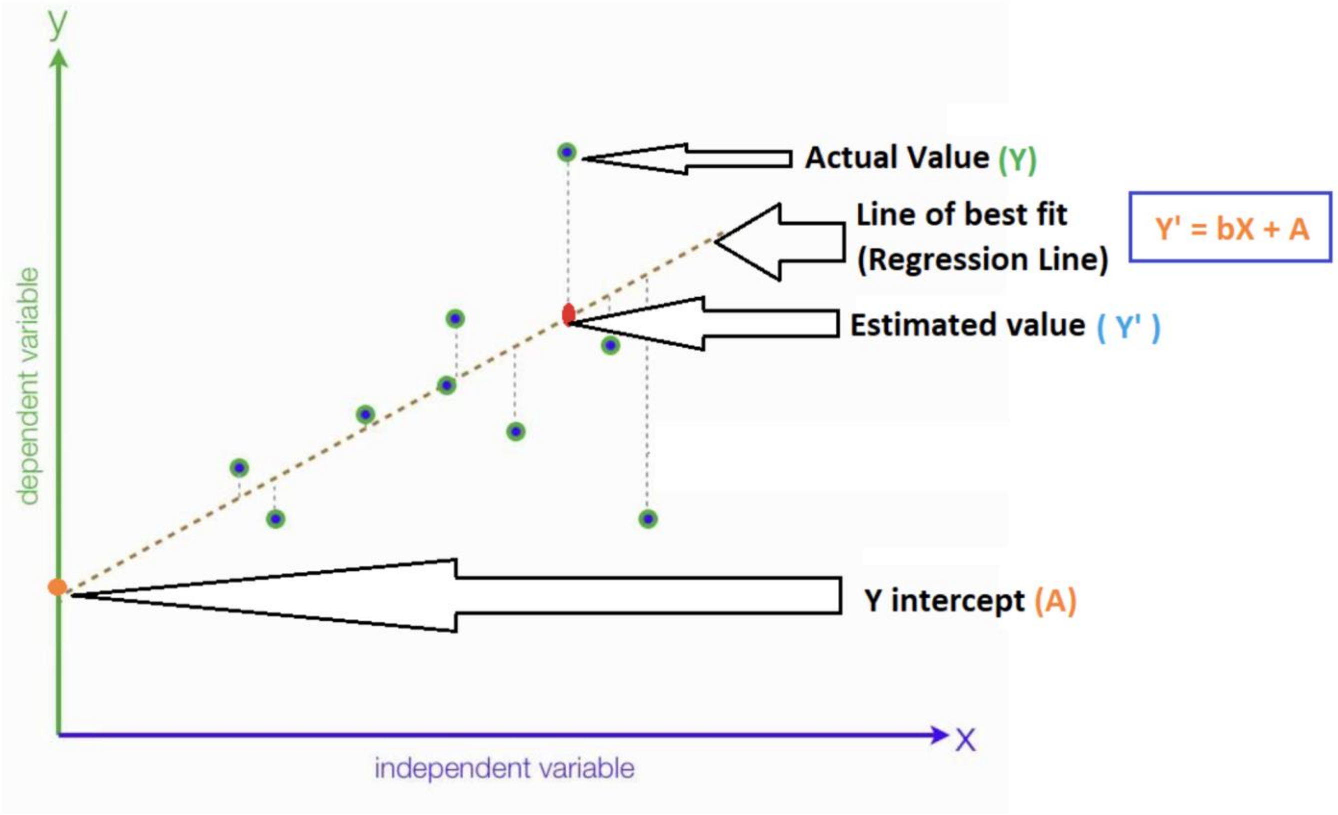
Multiple Linear Regression



- More than one factor impacts the variable to be predicted.
- E.g. : using the size, and location of the house to predict its price
- Its equation is similar to simple linear regression but has different slopes for different variables.

Y = A1 + B1X1+B2X2
Here, Y is dependent on X1 and X2.

Line of Best Fit for Linear Regression



- It's a straight line that best represents the relationship between the independent variable and the dependent variable.
- The line of best fit minimizes the sum of squared differences between observed data points and corresponding points predicted by the line.
- In multiple linear regression, this line becomes a multi-dimensional hyperplane that best fits the data.

Pop Quiz

Q. What does the coefficient for an independent variable represent in a multiple linear regression?

a. The y-intercept of the regression line

b. The change in the dependent variable for a unit change in that independent variable keeps other variables constant.

c. The total sum of squares of the data

d. The R-squared value of the model

Pop Quiz

Q. What does the coefficient for an independent variable represent in a multiple linear regression?

a. The y-intercept of the regression line

b. The change in the dependent variable for a unit change in that independent variable keeps other variables constant

c. The total sum of squares of the data

d. The R-squared value of the model

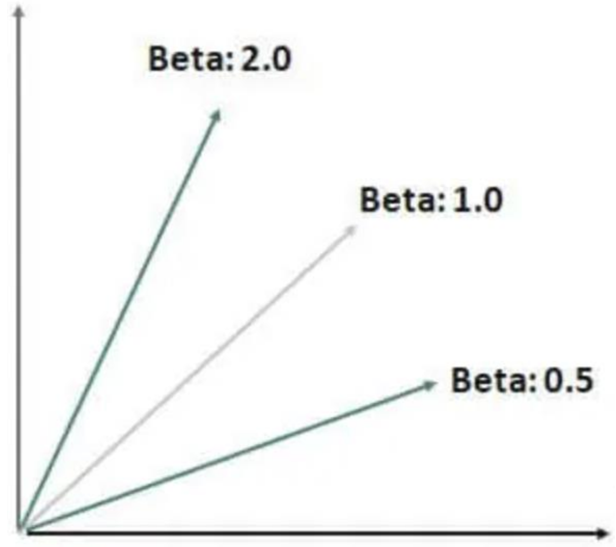
How to Find the Beta Coefficients?

- Beta coefficients are also known as regression coefficients or slope coefficients. They represent slope of each independent variable.

- In this equation :
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$
  $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are beta coefficients
  ε represents the error or residual terms

- To find beta coefficients, a method called the Ordinary Least Squares (OLS) algorithm is used.

- This algorithm aims to minimize the sum of squared differences between the predicted values and the actual observed values.



Process to Find the Beta Coefficients

- Once the regression equation is set up, create a design matrix. Organize independent variables into a matrix X, where each row represents an observation, and each column corresponds to a feature.

- Calculate the Coefficients: Use the OLS formula to calculate the coefficients:
  $\beta = (X^T X)^{-1} X^T y$
  Where:
  X^T is the transpose of the design matrix X.
  $(X^T X)^{-1}$ is the inverse of the matrix product of X^T and X.
  $X^T y$ is the matrix product of the transpose of X and the vector y (the dependent variable).

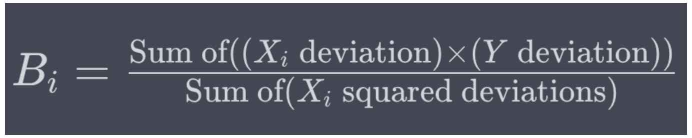
- Each β coefficient represents the change in y for a one-unit change in the corresponding independent variable while keeping the other variables constant.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Simple Formula to Find the Beta Coefficients

- The general formula for calculating a beta coefficient in multiple linear regression is as follows:



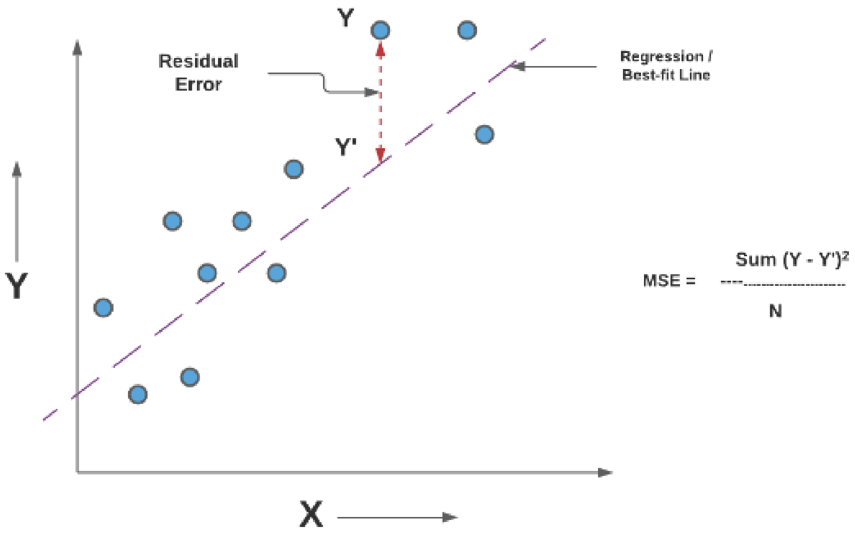
$$B_i = \frac{\text{Sum of}((X_i\text{ deviation})\times(Y\text{ deviation}))}{\text{Sum of}(X_i\text{ squared deviations})}$$

  - In the above formula, B_i represents the beta coefficient for the ith independent variable.
  - X_i deviation refers to difference between each value of ith independent variable and its mean.
  - Y deviation refers to difference between each value of dependent variable and its mean.
  - X_i squared deviations represent squared differences between each value of ith independent variable and its mean.

- Numerator represents covariance between independent variables and dependent variable. Covariance is a statistical measure that indicates extent to which two variables change together.

- Denominator represents variability or dispersion (spread from mean) of independent variables.

Cost Function and Gradient Descent



- To get the best fit line in linear regression, we are supposed to find the lowest total error of the line.

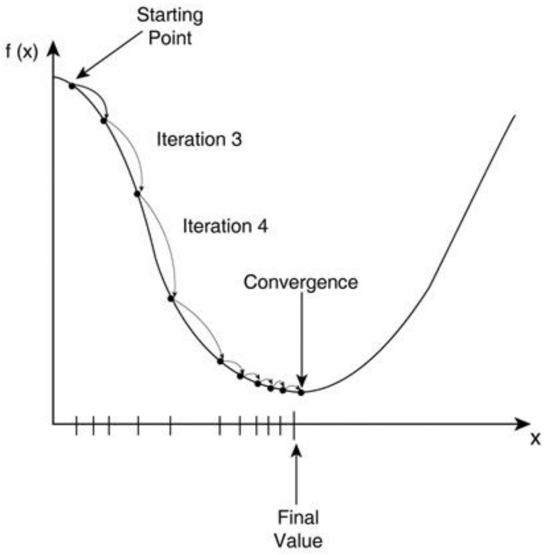
- Cost Function, which quantifies error between predicted value and expected values can be used to find correct best fit line. For linear regression, cost function is Mean Squared Error (MSE).

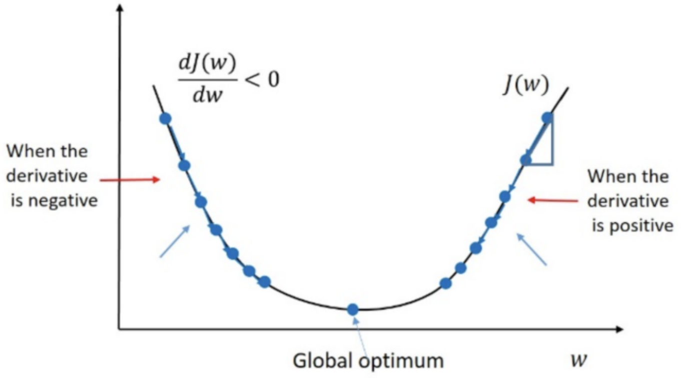
- MSE calculates the average squared difference between predicted values and actual values .

  Equation of Cost Function (Mean Squared Error):
  MSE = (1/n) * Σ(actual - predicted)^2
  Where n is the number of data points.

- Using the Cost function, we can draw a Gradient Descent Curve, which helps to find the correct gradient value for the best-fit line.

How is Gradient Descent Obtained?




$\frac{dJ(w)}{dw} < 0$
$J(w)$
When the derivative is negative
When the derivative is positive
Global optimum
w

- Intuitively, gradient means the slope of a curve at a given point in a specified direction. Here, x represents iterations, and f(x) represents cost function.

- The idea of gradient descent is to iteratively adjust the parameters in a way that reduces the difference between the predicted values and the actual observed values.

- As you repeat the adjustment of parameters, coefficients gradually adjust to minimize the cost function, leading to better predictions.

- It's like finding the path down a hill to reach the lowest point (minimum) of a cost landscape.

- The gradient of the curve where it's value is equal to 0, that is the best gradient (x) value for the best fit line.

Pop Quiz

Q. Which of the following best describes the concept of gradient descent?

a. A method to increase the complexity of a machine learning model
b. An algorithm used to find the maximum value of a cost function
c. An iterative optimization technique to minimize a cost function by adjusting model parameters
d. A process to visualize the relationship between features in a dataset

Pop Quiz

Q. Which of the following best describes the concept of gradient descent?

a. A method to increase the complexity of a machine learning model

b. An algorithm used to find the maximum value of a cost function

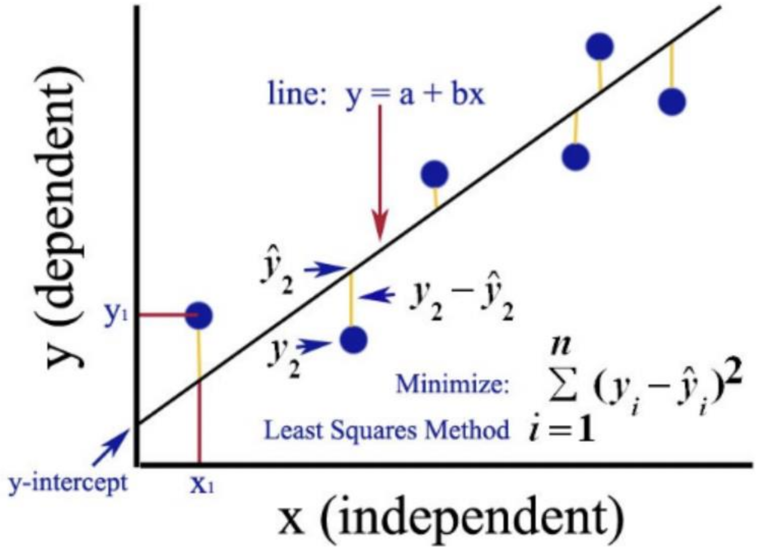
▌ **c. An iterative optimization technique to minimize a cost function by adjusting model parameters**

d. A process to visualize the relationship between features in a dataset

Q&A

OLS Method

OLS Method



- Ordinary Least Squares (OLS) is a linear regression technique used to find the best-fitting line for a set of data points by minimizing the residuals.
- It does so by estimating the coefficients of a linear regression model by minimizing cost function.
- OLS provides coefficients (slopes and intercepts) that directly relate to the impact of each independent variable on the dependent variable.
- This makes the model's results easy to interpret, especially when explaining relationships to non-technical stakeholders.

OLS Assumptions

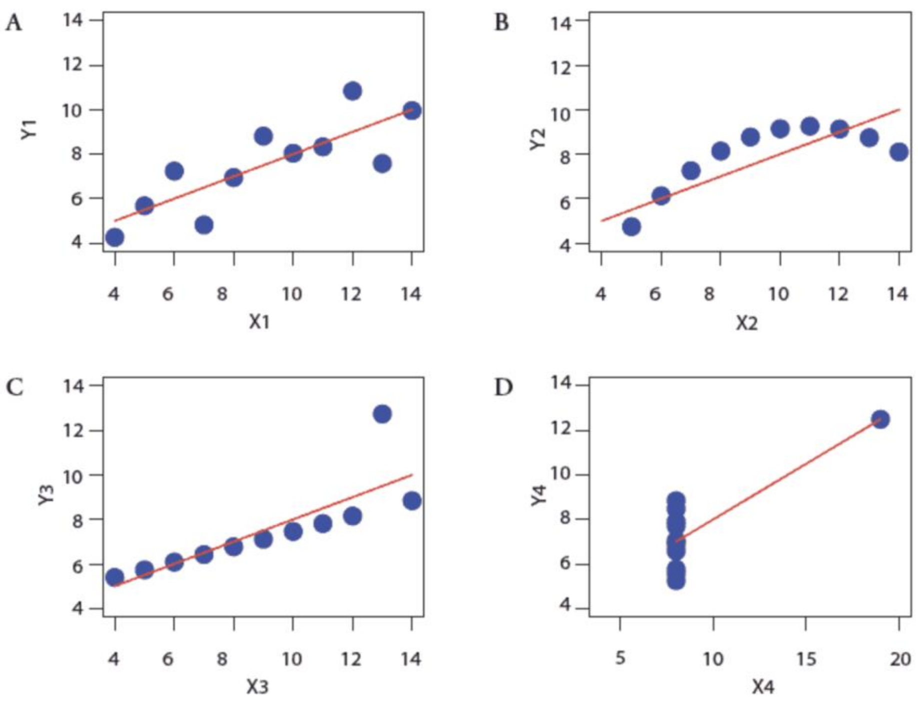
- The Ordinary Least Squares (OLS) method makes certain assumptions about the data and the model in order for its estimates to be valid and meaningful.

- These assumptions are important to ensure that OLS estimates are unbiased, efficient, and statistically reliable.

- Here are the key assumptions associated with OLS:

  - Linearity Assumption
  - Random Sampling of Observations
  - Homoscedasticity
  - No Autocorrelation
  - No Multi-collinearity
  - Normality of errors

OLS Assumptions : Linearity



- The relationship between independent variables and dependent variables should be linear. This means that changes in dependent variables are proportional to changes in independent variables.

- The linearity assumption ensures that the model's predictions accurately capture the underlying relationships in the data.

- In these cases, graph A depicts best linear relationship.

Poll Time

Q. If the linearity assumption is violated in regression analysis, it may lead to:

a. Underestimation of the model's predictive accuracy

b. Overestimation of the model's predictive accuracy

c. Inaccurate values for the model coefficients

d. Improved model performance due to increased flexibility

Poll Time

Q. If the linearity assumption is violated in regression analysis, it may lead to:

a. **Underestimation of the model's predictive accuracy**

b. Overestimation of the model's predictive accuracy

c. Inaccurate values for the model coefficients

d. Improved model performance due to increased flexibility

OLS Assumptions : Random Sampling of Observations

- This assumption states that the data points used in the analysis should be collected through a process of random sampling from the population of interest.

- This assumption ensures that the sample of observations is representative of the larger population and that each observation is independent of others.

- In other words, the assumption of random sampling helps prevent selection bias and ensures that the relationships observed in the sample can be generalized to the broader population.

- It's important to note that if the data are not collected through a process of random sampling, it could lead to biased and non-representative results.

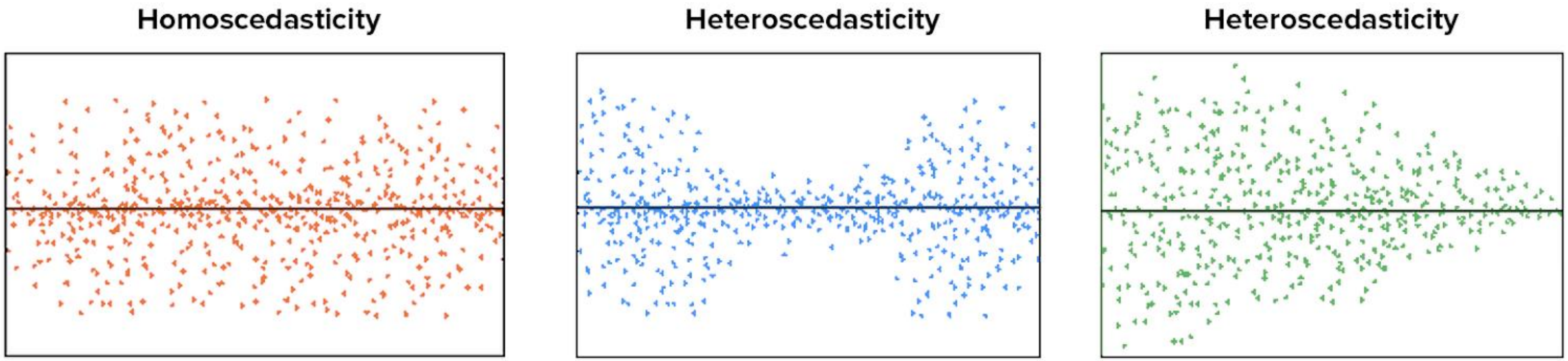
- For example, if certain observations are systematically included or excluded from the dataset, the relationships between variables might not accurately reflect the underlying population.

OLS Assumptions : Homoscedasticity

- This assumption states that the spread of the residuals (difference between actual values and predicted values of the dependent variable) should not change as the values of the independent variables change.

- It ensures that variance of residuals should be approximately constant across all levels of independent variables.

- Unequal variance (Heteroscedasticity) can lead to biased and inefficient coefficient estimates.



Homoscedasticity

Random Cloud (No Discernible Pattern)

Heteroscedasticity

Bow Tie Shape (Pattern)

Heteroscedasticity

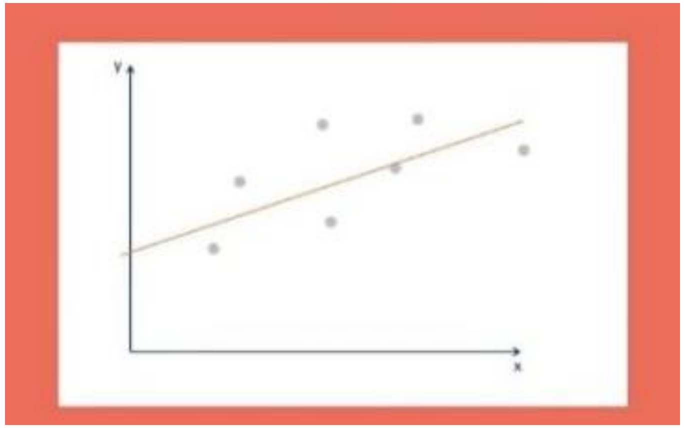
Fan Shape (Pattern)

OLS Assumptions : No Autocorrelation

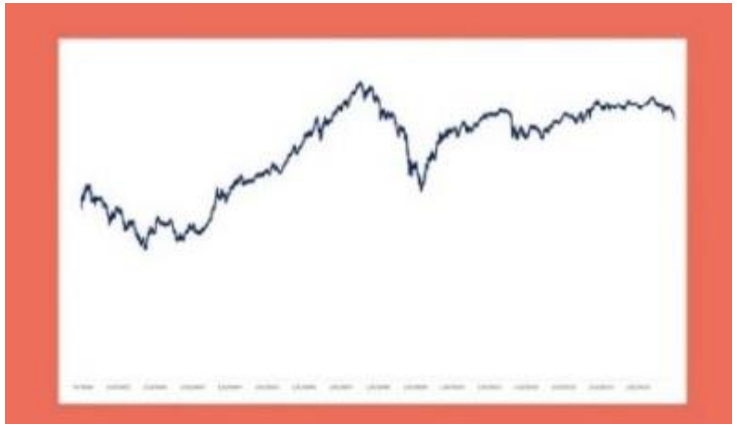
- This assumption, also known as serial correlation, occurs when the residuals (errors) of the model are correlated with each other over time or across observations.

- This assumption is also called independence of errors because it states that errors from one observation should not be systematically related to errors from previous or subsequent observations.

- Autocorrelation can be a concern in time-series data (sequence of observations collected at specific time intervals) or any situation where observations are ordered and might exhibit patterns or trends that persist over time.



Normal Data



Time series data

Pop Quiz

Q. Autocorrelation in linear regression refers to:

a. The correlation between two or more independent variables

b. The correlation between the dependent variable and the error term

c. The correlation between the residuals of a regression model

d. The correlation between the independent and dependent variables

Pop Quiz

Q. Autocorrelation in linear regression refers to:

a. The correlation between two or more independent variables

b. The correlation between the dependent variable and the error term

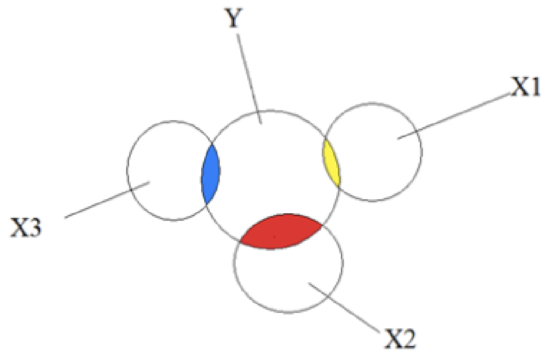
[✓ **c. The correlation between the residuals of a regression model**]

d. The correlation between the independent and dependent variables

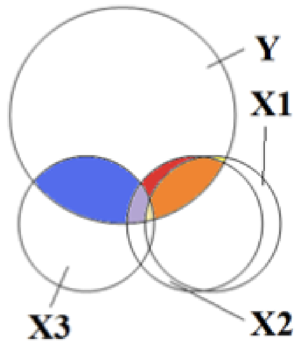
OLS Assumptions : No Multicollinearity

- This assumption states that there should be no strong linear relationships among the independent variables. In other words, they should not be highly correlated with each other.

- Multicollinearity occurs when two or more independent variables are highly correlated with each other. This can lead to challenges in accurately estimating the individual effects of these variables on the dependent variable.

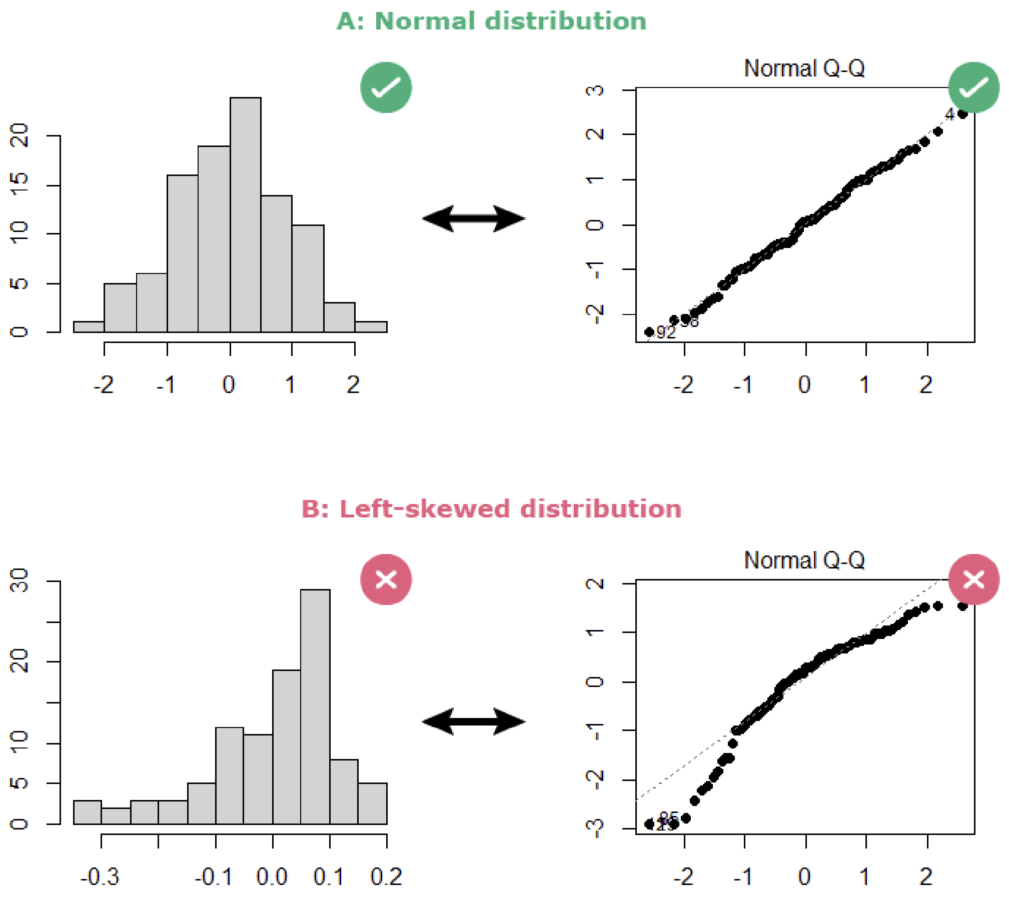


No Multicollinearity



Multicollinearity

OLS Assumptions : Normality of Errors



- This assumption states that the residuals should follow a roughly normal distribution.
- In other words, if you were to plot a histogram of the residuals, it should resemble a bell-shaped curve.
- If the plot deviates significantly from a straight line, it suggests non-normality.
- You can assess the normality assumption by examining a histogram or a Q-Q plot (quantile-quantile plot) of the residuals.

Poll Time

Q. What might be a consequence of violating the normality assumption in linear regression?

a. The model might not adequately capture the relationships between variables

b. The independent variables might not have a linear relationship with the dependent variable

c. The residuals might be correlated with each other

d. The coefficients might have inflated standard errors

Poll Time

Q. What might be a consequence of violating the normality assumption in linear regression?

a. The model might not adequately capture the relationships between variables

b. The independent variables might not have a linear relationship with the dependent variable

c. The residuals might be correlated with each other

d. **The coefficients might have inflated standard errors**

Q&A

Activity 1

Pre-requisites:

Familiarity with multiple linear regression and its components.

Scenario:

You are a HR analyst, and you have a dataset containing information about employees and their performance ratings. The dataset includes these features: Years of experience, Communication skills (rated on a scale of 1 to 10). You are interested in predicting the performance rating of employees based on these features.

Data:

Consider the table shown on the right side:

- Years of experience
- Communication skills
- Performance rating

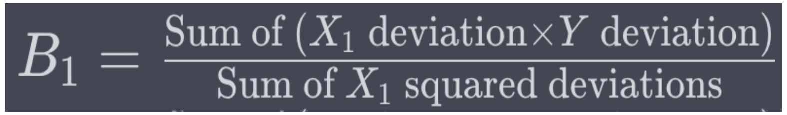
Years of Experience	Communication Skills	Performance Rating
5	7	85
3	6	70
7	8	95
10	9	110

Expected Outcome:

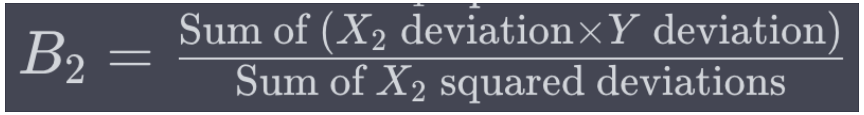
Calculate the beta coefficients for each independent variable in the multiple linear regression model. This will help to understand how to quantify the relationships between the variables and the impact of each feature on the predicted outcome.

Steps:

1) Calculate the means of the independent variables X1 (Years of Experience) and X2 (Communication Skills)
2) Calculate the means of the dependent variable Y (Performance Rating)
3) Calculate the deviations of each data point from the means for X1, X2, and Y
4) Calculate the sum of the product of deviations for each pair of variables : Sum of (X1 deviation * Y deviation), Sum of (X2 deviation * Y deviation)
5) Calculate the sum of squared deviations for X1 and X2
6) Calculate the beta coefficients for X1 and X2 using these formulas :



$$B_1 = \frac{\text{Sum of } (X_1 \text{ deviation} \times Y \text{ deviation})}{\text{Sum of } X_1 \text{ squared deviations}}$$



$$B_2 = \frac{\text{Sum of } (X_2 \text{ deviation} \times Y \text{ deviation})}{\text{Sum of } X_2 \text{ squared deviations}}$$

Activity 2

Pre-requisites:

Familiarity with OLS assumptions and scatter plots.

Scenario:

You're a sales manager in a retail company aiming to predict sales performance based on a crucial factor: the number of hours employees spend on training. You've collected data to establish this relationship and need to ensure that the linearity assumption behind your regression model are met. Accurate sales predictions are essential for optimizing resource allocation and achieving revenue targets.

Data:

Consider the table shown on the right side having hours of training and corresponding sales.

Hours of Training	Sales Performance
15	2000
20	2500
25	2800
18	2200
30	3200

Expected Outcome:

Gain practical experience in visually assessing the linearity assumption for the simple linear regression model.

Steps:

1) Create a scatter plot with Hours of Training on the x-axis and Sales Performance on the y-axis to visualize linearity.
2) Analyze the scatter plot to observe if there's a linear pattern between hours of training and sales performance.

Summary

- ✓ Multiple Linear Regression involves multiple independent variables and a continuous dependent variable. The target variable is influenced by weighted combinations of multiple predictors.
- ✓ The goal of linear regression is to minimize errors to best represent the relation between two or more variables.
- ✓ The idea of gradient descent is to iteratively adjust the parameters in a way that reduces the difference between the predicted values and the actual observed values.
- ✓ The Ordinary Least Squares (OLS) method makes certain assumptions about the data and the model for its estimates to be valid and meaningful.

Session Feedback

Next Session:
Deep Dive into Linear Regression

THANK YOU

Please complete your assessments and review the self-learning content for this session on the **PRISM** portal.

knowledgehut
upGrad