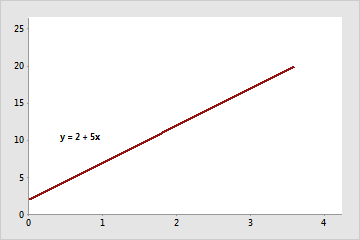
**Machine Learning Assignment 17**

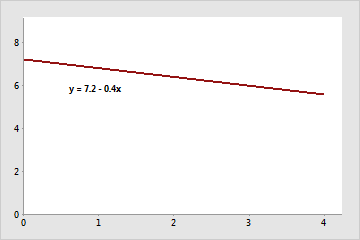
1. Using a graph to illustrate slope and intercept, define basic linear regression.

Ans-) The slope indicates the steepness of a line and the intercept indicates the location where it intersects an axis. The slope and the intercept define the linear relationship between two variables, and can be used to estimate an average rate of change. The greater the magnitude of the slope, the steeper the line and the greater the rate of change.

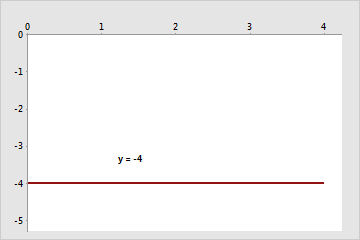
By examining the equation of a line, you quickly can discern its slope and y-intercept (where the line crosses the y-axis).



The slope is positive 5. When x increases by 1, y increases by 5. The y-intercept is 2.



The slope is negative 0.4. When x increases by 1, y decreases by 0.4. The y-intercept is 7.2.



The slope is 0. When x increases by 1, y neither increases or decreases. The y-intercept is -4.

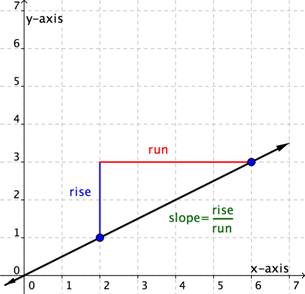
Usually, this relationship can be represented by the equation y = b0 + b1x, where b0 is the y-intercept and b1 is the slope.

For example, a company determines that job performance for employees in a production department can be predicted using the regression model y = 130 + 4.3x, where x is the hours of in-house training they receive (from 0 to 20) and y is their score on a job skills test. The value of the y-intercept (130) indicates the average job skill score for an employee with no training. The value of the slope (4.3) indicates that for each hour of training, the job skill score increases, on average, by 4.3 points.

2. In a graph, explain the terms rise, run, and slope.

Ans-)  The vertical change between two points is called the rise, and the horizontal change is called the run. The slope equals the rise divided by the run: Slope =rise/run

Slope =rise/run.

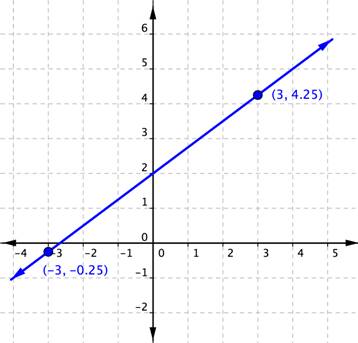


You can determine the slope of a line from its graph by looking at the rise and run. One characteristic of a line is that its slope is constant all the way along it. So, you can choose any 2 points along the graph of the line to figure out the slope.

3. Use a graph to demonstrate slope, linear positive slope, and linear negative slope, as well as the

different conditions that contribute to the slope.

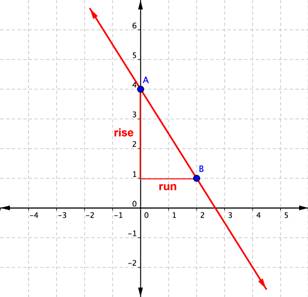
Ans-) the following two examples, you will see a slope that is positive and one that is negative.

The positive slope of the line graphed below.

The next example shows a line with a negative slope.

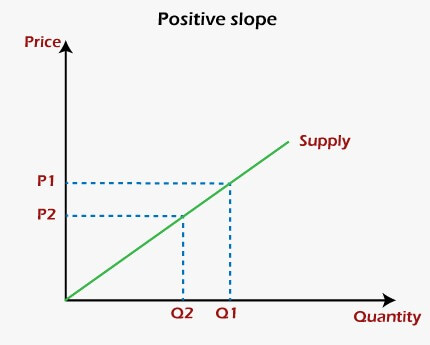
EXAMPLE

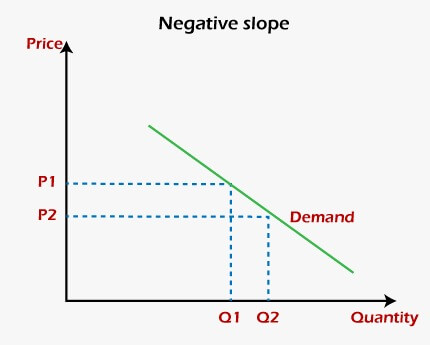
Find the slope of the line graphed below.



4. Use a graph to demonstrate curve linear negative slope and curve linear positive slope.

Ans-) There are two main types of slopes which are given below:

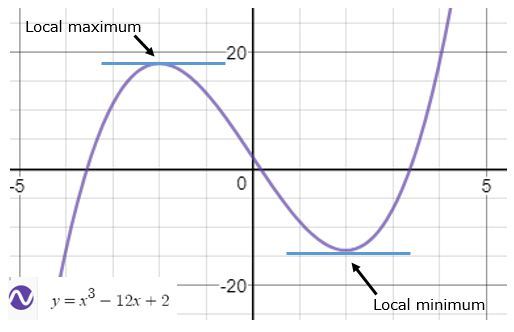
**Positive Slope:** A slope in which two variables, i.e., variable at x-axis and variable at the y-axis, shows a positive relationship is known as positive slope. In simpler words, a positive slope is one in which the variable x increases with the increase in variable y and/or variable y increases with the increase in variable x. Similarly, the variable x decreases with the decrease in variable y, and/or variable y decreases with the decrease in variable x. It means both the variables are complements to each other. A positive slope moves in the upward direction or is upward sloping.  
In graphical terms, a positive slope is one in which the line on the graph rises when it moves from left to right. The concept of positive slope can be clearly understood with the help of the supply curve of a producer or firm in economics. The two variables of the curve are price at the y-axis and quantity of goods at the x-axis. Let us assume the firm is producing the goods for profit maximization. Therefore, when the prices of the goods increase, the quantity supplied by the firm of those goods will also increase, while when the prices decrease, the quantity supplied by the firm will decrease. In other words, at higher prices, the firm or producer will increase the quantity supplied to earn more profit, while at lower prices, they will reduce the quantity supplied to reduce the loss. Hence, it shows the prices and quantity supplied are positively related to each other, which can be cleared from the diagram given below:  


**Negative Slope:** A slope in which two variables, i.e., variable at x-axis and variable at the y-axis, shows a negative relationship is known as negative slope. In other words, a negative slope is one in which the variable x increases with the decrease in variable y and/or variable y increases with the decrease in variable x. In the same manner, the variable x decreases with the increase in variable y, and/or variable y decreases with the increase in variable x. This represents an inverse relationship between these two variables. A negative slope moves in the downward direction or is downward sloping.  
Graphically, a negative slope is one in which the line on the graph falls when it moves from left to right. One of the best examples of the negative slope of the graph is the demand curve in economics. The two variables of the curve are price at the y-axis and quantity of goods at the x-axis. As we know, the consumers buy a large quantity of a good at a lower price than at a higher price. Therefore, the quantity demanded by the consumers of goods will decrease with an increase in the prices of those goods. On the other hand, when prices of the goods will decrease, the quantity demand will increase. Hence, it shows a negative relationship between the prices and quantity supplied of those goods. It can be cleared from the diagram given below:  


5. Use a graph to show the maximum and low points of curves.

Ans-) In this graph of the function y = x^3 - 12x+ 2 there is a local maximum (at x= -2) and local minimum (at x= 2).

The blue horizontal line shows that the gradient at these points is zero i.e. f'(x) = 0



6. Use the formulas for a and b to explain ordinary least squares.

**Ans-)** In linear regression, Ordinary Least Squares (OLS) is a method to estimate the parameters of a linear equation, which is used to model the relationship between the dependent variable and one or more independent variables. The OLS method is based on minimizing the sum of the squares of the differences between the predicted values and the actual values.

The linear equation can be represented as:

y = a + bx

Where y is the dependent variable, x is the independent variable, a is the intercept, and b is the slope.

The formula for calculating the slope (b) is:

b = Σ((xi - x̄)(yi - ȳ))/ Σ(xi - x̄)²

Where Σ represents the sum, xi is the ith value of the independent variable, x̄ is the mean of the independent variable, yi is the ith value of the dependent variable, and ȳ is the mean of the dependent variable.

The formula for calculating the intercept (a) is:

a = ȳ - bx̄

Where ȳ is the mean of the dependent variable, and x̄ is the mean of the independent variable.

Once the values of a and b are determined using the OLS method, they can be used to predict the value of the dependent variable for a given value of the independent variable.

The OLS method is widely used because it is simple and provides accurate estimates of the parameters of the linear equation. However, it assumes that the errors in the model are normally distributed and have a constant variance, which may not always be the case in real-world applications.

7. Provide a step-by-step explanation of the OLS algorithm.

**Ans-)** The OLS algorithm is a method for finding the best fit line through a set of data points. Here is a step-by-step explanation of the algorithm:

* Gather data: Collect data on the dependent and independent variables of interest.
* Calculate the means: Calculate the mean of the dependent variable (Y) and the mean of the independent variable (X).
* Calculate the deviations: For each data point, calculate the deviation of Y and X from their respective means.
* Calculate the sum of products: Multiply the deviation of Y and X for each data point and sum these products.
* Calculate the sum of squares: Square the deviation of X for each data point and sum these squares.
* Calculate the slope: Divide the sum of products by the sum of squares to obtain the slope of the line (b).
* Calculate the intercept: Calculate the intercept (a) by subtracting the product of the slope and the mean of X from the mean of Y.
* Calculate the predicted values: Using the equation of the line (Y = a + bX), calculate the predicted value of Y for each data point.
* Calculate the residuals: For each data point, calculate the difference between the predicted value and the actual value of Y.
* Calculate the sum of squares of residuals: Square the residuals for each data point and sum these squares.
* Calculate the variance: Divide the sum of squares of residuals by the degrees of freedom to obtain the variance.
* Calculate the standard error of the estimate: Take the square root of the variance to obtain the standard error of the estimate.
* Evaluate the model: Assess the goodness of fit of the model using measures such as R-squared and adjusted R-squared.
* Use the model: Use the model to make predictions or draw conclusions about the relationship between the dependent and independent variables.

8. What is the regression’s standard error?

Ans-)The regression's standard error is a measure of the variability of the residuals or errors in a regression model. It indicates how closely the actual data points are clustered around the predicted values by the regression model.

To represent the regression's standard error, we can plot the residuals against the predicted values on a scatter plot. The residuals are the differences between the actual data points and the predicted values by the regression model. If the residuals are randomly scattered around the horizontal line at zero, it indicates that the regression model is a good fit for the data and has a low standard error.

9. Provide an example of multiple linear regression.

**Ans-)** Multiple linear regression is used when there is more than one predictor variable that affects the response variable. For example, consider a scenario where we want to predict the salary of an employee based on their years of experience, level of education, and age. Here, salary is the response variable, while years of experience, level of education, and age are the predictor variables. The multiple linear regression model can be represented as:

Salary = β0 + β1(Years of experience) + β2(Level of education) + β3(Age) + ε

where β0, β1, β2, and β3 are the regression coefficients, and ε is the error term.

10. Describe the regression analysis assumptions and the BLUE principle.

**Ans-)** Regression analysis assumptions include:

* Linearity: The relationship between the response variable and predictor variables is linear.
* Independence: The errors of the regression model are independent.
* Homoscedasticity: The variance of the error term is constant across all levels of the predictor variables.
* Normality: The error term follows a normal distribution.

The BLUE principle (Best Linear Unbiased Estimator) is a statistical principle that states that the best linear unbiased estimator of the regression coefficients is obtained using the OLS method.

11. Describe two major issues with regression analysis.

Ans-) Two major issues with regression analysis are:

1. **Overfitting:** Overfitting occurs when the regression model is too complex and fits the noise in the data, rather than the underlying signal. This results in poor performance on new, unseen data.
2. **Multicollinearity:** Multicollinearity occurs when there is a high correlation between predictor variables. This makes it difficult to determine the effect of each predictor variable on the response variable, and can lead to unstable estimates of the regression coefficients.

12. How can the linear regression models accuracy be improved?

Ans-)The accuracy of the linear regression model can be improved by:

* Adding more relevant predictor variables to the model.
* Removing irrelevant predictor variables from the model.
* Transforming the predictor variables to better fit the assumptions of the model.
* Regularizing the model to prevent overfitting.
* Using a more powerful regression algorithm, such as polynomial regression or decision tree regression.

13. Using an example, describe the polynomial regression model in detail.

**Ans-)** Polynomial regression is a form of linear regression in which the relationship between the response variable and the predictor variable is modeled as an nth-degree polynomial. For example, consider a scenario where we want to predict the temperature based on the time of day. A simple linear regression model may not capture the underlying relationship, as the temperature may vary in a non-linear manner throughout the day. In this case, we can use a polynomial regression model with a higher degree to better capture the relationship between the two variables. The polynomial regression model can be represented as:

Temperature = β0 + β1(Time) + β2(Time^2) + β3(Time^3) + … + βn(Time^n) + ε

where β0, β1, β2, β3, and βn are the regression coefficients, and ε is the error term. The degree of the polynomial (n) can be chosen based on the complexity of the relationship between the response and predictor variables.

14. Provide a detailed explanation of logistic regression.

**Ans-)** Logistic regression is a statistical method used to predict the probability of an event occurring based on the relationship between a dependent variable and one or more independent variables. It is a popular technique used in binary classification problems, where the outcome is either a "success" or "failure" or "yes" or "no."

In logistic regression, the dependent variable is a binary variable that represents the occurrence or non-occurrence of an event. The independent variables, on the other hand, can be continuous, categorical, or dichotomous. The objective of logistic regression is to find the best-fitting model that explains the relationship between the independent variables and the probability of the dependent variable occurring.

The logistic regression model is based on the logistic function, also known as the sigmoid function. The sigmoid function takes any input value and returns a value between 0 and 1, representing the probability of an event occurring. The formula for the logistic function is:

p = 1 / (1 + e^(-z))

where p is the probability of the event occurring, e is the base of the natural logarithm, and z is a linear combination of the independent variables.

z = β0 + β1x1 + β2x2 + ... + βnxn

where β0 is the intercept or the constant term, β1 to βn are the coefficients for the independent variables x1 to xn, respectively.

The coefficients in logistic regression represent the change in the log odds of the dependent variable for a one-unit change in the corresponding independent variable. In other words, they indicate how much each independent variable contributes to the probability of the dependent variable occurring.

The logistic regression model is fitted using a maximum likelihood estimation method. The goal is to maximize the likelihood function, which measures the likelihood of observing the data given the model parameters. The maximum likelihood estimates for the model parameters are obtained by iteratively adjusting the coefficients until the likelihood function is maximized.

Once the logistic regression model is fitted, it can be used to make predictions for new observations. The model predicts the probability of the dependent variable occurring, which can be converted into a binary outcome by setting a threshold value. If the predicted probability is above the threshold value, the outcome is classified as a success, otherwise, it is classified as a failure.

Logistic regression is a powerful tool for analyzing binary classification problems. However, it is important to note that it makes several assumptions, including linearity, independence of errors, and absence of multicollinearity. Violations of these assumptions can affect the validity and reliability of the model. Therefore, it is important to carefully evaluate the model and its assumptions before making any conclusions or predictions.

15. What are the logistic regression assumptions?

**Ans-)** Logistic regression, like any other regression analysis, makes certain assumptions that must be met to obtain reliable results. Some of the key assumptions in logistic regression include:

* Linearity of the Logit: The relationship between the independent variables and the logit of the dependent variable should be linear. This is checked by plotting the logit of the dependent variable against the independent variables, and checking for a linear relationship.
* Independence of observations: The observations in the dataset must be independent of each other. This means that there should be no correlation or pattern between the residuals of the model.
* Absence of multicollinearity: There should be no perfect or high correlation between the independent variables in the dataset. This is because high multicollinearity can lead to unstable and unreliable estimates of the regression coefficients.
* Large sample size: A large sample size is recommended to obtain reliable estimates of the regression coefficients and to ensure that the results are statistically significant.
* Normality of residuals: The residuals of the model should be normally distributed. This is checked by plotting a histogram of the residuals and checking for a normal distribution.

16. Go through the details of maximum likelihood estimation.

**Ans-)** Maximum likelihood estimation (MLE) is a method used to estimate the parameters of a statistical model. It involves finding the values of the parameters that maximize the likelihood function of the model, given the observed data. In logistic regression, MLE is used to estimate the parameters of the logistic regression model.

The likelihood function in logistic regression is defined as the probability of observing the data given the values of the parameters. The goal of MLE is to find the values of the parameters that maximize this function. The most common method used to find the maximum likelihood estimates is the Newton-Raphson algorithm, which involves iteratively finding the values of the parameters that maximize the likelihood function.

To estimate the parameters of a logistic regression model using MLE, we start with an initial guess for the parameters and then iteratively update the parameters using the Newton-Raphson algorithm until convergence is achieved. The convergence criterion is usually based on a specified tolerance level or the number of iterations.

Once the maximum likelihood estimates of the parameters are obtained, we can use them to make predictions about the probability of the dependent variable based on the values of the independent variables. These predictions can be used to classify new observations into different categories based on their predicted probabilities.