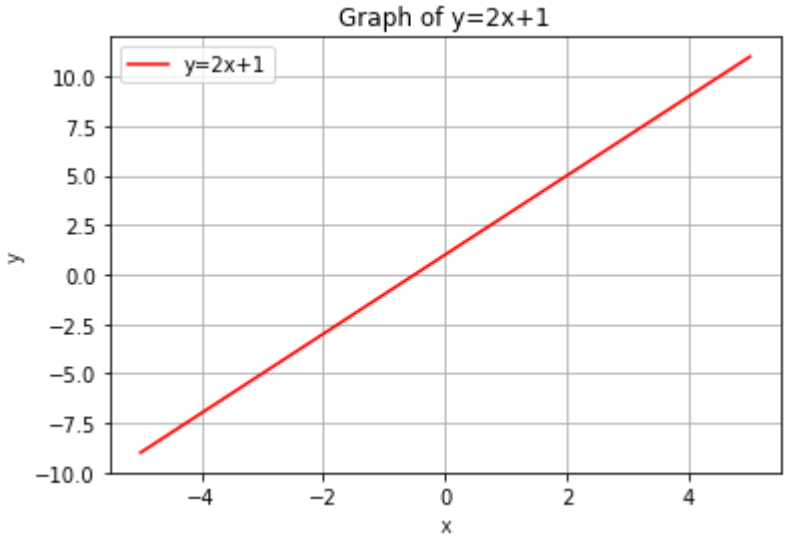
Assignment 17 Solutions

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

Ans: The equation `y=mx+c` represents a straight line graphically, where m is its slope/gradient and c its intercept. In this tutorial, you will learn how to plot y=mx+b in Python with Matplotlib.

In [1]:

```
import matplotlib.pyplot as plt
import numpy as np
x = np.linspace(-5,5,100)
y = 2*x+1
plt.plot(x, y, '-r', label='y=2x+1')
plt.title('Graph of y=2x+1')
plt.xlabel('x', color='#1C2833')
plt.ylabel('y', color='#1C2833')
plt.legend(loc='upper left')
plt.grid()
display(plt.show())
```



None

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

Ans: The slope of a line measures the steepness of the line. Most of you are probably familiar with associating slope with "Rise Over Run". Rise means how many units you move up or down from point to point. On the graph that would be a change in the y values. Run means how far left or right you move from point to point.

3. Use a graph to demonstrate slope, linear positive slope, and linear negative slope, as well as the different conditions that contribute to slope ?

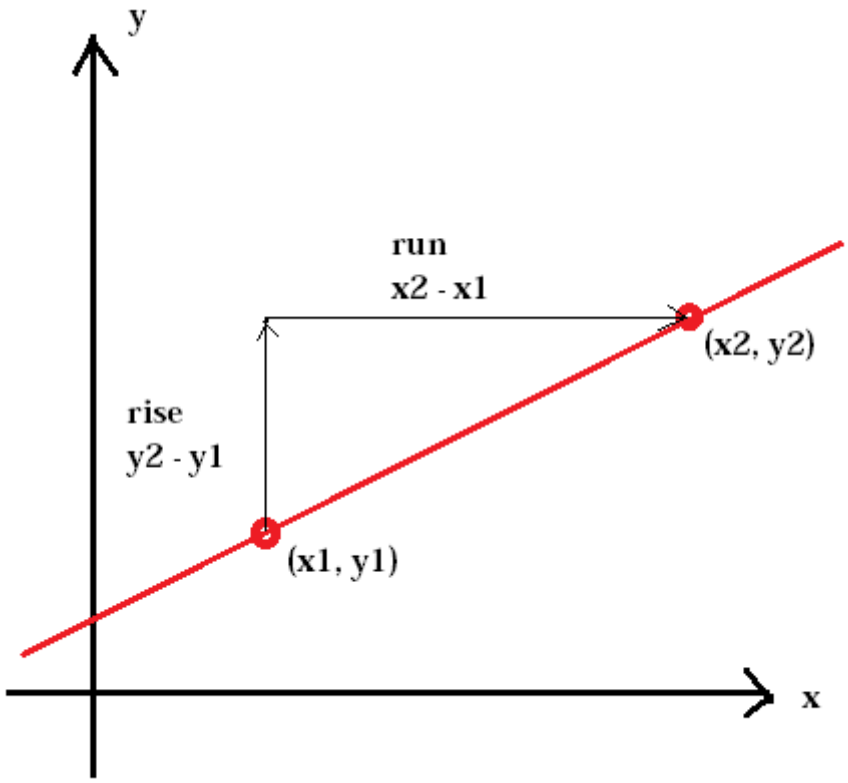
Ans: The steepness of a hill is called a slope. The same goes for the steepness of a line. The slope is defined as the ratio of the vertical change between two points, the rise, to the horizontal change between the same two points, the run.

```
Slope=Rise/Run=ChangeinY/changeinX
```

The slope of a line is usually represented by the letter m. (x1, y1) represents the first point whereas (x2, y2) represents the second point.

```
M=Y2-Y1/X2-X1
```

It is important to keep the x-and y-coordinates in the same order in both the numerator and the denominator otherwise you will get the wrong slope.



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

Ans: Image result for a graph to demonstrate curve linear negative slope and curve linear positive slope.If the signs are different then the answer is negative! If the slope is negative you can plot your next point by going down and right OR up and left. f the slope is positive you can plot your next point by going up and right OR down and left.

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

Ans: To find the maximum/minimum of a curve you must first differentiate the function and then equate it to zero. This gives you one coordinate. To find the other you must resubstitute the one already found into the original function.

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

Ans: This best line is the Least Squares Regression Line (abbreviated as LSRL). This is true where ˆy is the predicted y-value given x, a is the y intercept, b and is the slope. For every x-value, the Least Squares Regression Line makes a predicted y-value that is close to the observed y-value, but usually slightly off.

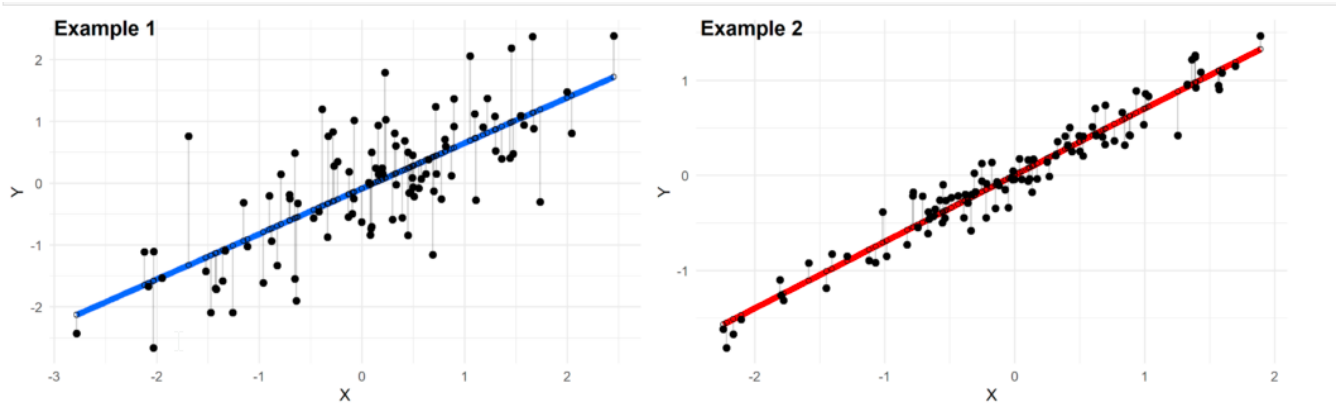
7. Provide a step-by-step explanation of the OLS algorithm ?

Ans: Ordinary Least Square Method :

- Set a difference between dependent variable and its estimation:
- Square the difference:
- Take summation for all data.
- To get the parameters that make the sum of square difference become minimum, take partial derivative for each parameter and equate it with zero.

8. What is the regression's standard error? To represent the same, make a graph ?

Ans: The standard error of the regression (S), also known as the standard error of the estimate, represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how wrong the regression model is on average using the units of the response variable.



9. Provide an example of multiple linear regression ?

Ans: Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable. Example: Prediction of CO2 emission based on engine size and number of cylinders in a car and predict

the price of a house based on its size

Let's say we want to predict the price of a house based on its size (in square feet), the number of bedrooms, and the age of the house (in years). We have a dataset with the following information for several houses:

```
House 1:
Size: 1500 sq ft
Bedrooms: 3
Age: 10 years
Price: $200,000

House 2:
Size: 1800 sq ft
Bedrooms: 4
Age: 5 years
Price: $250,000

House 3:
Size: 1200 sq ft
Bedrooms: 2
Age: 15 years
Price: $180,000

House 4:
Size: 2000 sq ft
Bedrooms: 3
Age: 8 years
Price: $220,000
```

To perform multiple linear regression, we'll formulate the model as follows:

Price = β0 + β1 * Size + β2 * Bedrooms + β3 * Age + ε

Where:

Price is the dependent variable (what we want to predict). Size, Bedrooms, and Age are the independent variables (input features). β0, β1, β2, and β3 are the regression coefficients to be estimated. ε is the error term.

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

Ans: There are four assumptions associated with a linear regression model:

- **Linearity:** The relationship between X and the mean of Y is linear.
- **Homoscedasticity:** The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **BLUE:** is an acronym for the following: Best Linear Unbiased Estimator. In this context, the definition of “best” refers to the minimum variance or the narrowest sampling distribution.

11. Describe two major issues with regression analysis ?

Ans: It involves very lengthy and complicated procedure of calculations and analysis. It cannot be used in case of qualitative phenomenon viz. honesty, crime etc.

The overall idea of regression is to examine two things:

1. Does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
2. Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta

12. How can the linear regression model's accuracy be improved ?

Ans: 8 Methods to Boost the Accuracy of a Model :

- Add more data. Having more data is always a good idea.
- Treat missing and Outlier values.
- Feature Engineering.
- Feature Selection.
- Multiple algorithms.
- Algorithm Tuning.
- Ensemble methods.

It's important to note that improving model accuracy is an iterative process. You may need to try different techniques and combinations to find the most effective approach for your specific problem and dataset.

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

Ans: In statistics, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.For this reason, polynomial regression is considered to be a special case of multiple linear regression.

Polynomial regression is one of the machine learning algorithms used for making predictions. For example, it is widely applied to predict the spread rate of COVID-19 and other infectious diseases.

14. Provide a detailed explanation of logistic regression ?

Ans: Short ans-Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set.Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.

Detailed explanation:-

Logistic regression is a statistical model used for binary classification tasks, where the goal is to predict the probability of an event occurring based on a set of input features. It is widely used in various fields, including healthcare, finance, marketing, and social sciences.

Here is a detailed explanation of how logistic regression works:

Problem Statement: In logistic regression, we have a dataset with input features (X) and corresponding binary labels (Y). The labels represent the presence or absence of an event of interest. For example, we might have data on various factors related to a disease and whether a patient has the disease or not.

Logistic Function (Sigmoid): Logistic regression uses a logistic function (also known as the sigmoid function) to transform a linear combination of the input features into a probability value. The logistic function maps any real-valued number to a value between 0 and 1, allowing us to interpret the output as a probability. The sigmoid function is defined as:

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

where z represents the linear combination of input features and coefficients.

Model Representation: The logistic regression model can be represented as follows:

p = sigmoid(β0 + β1X1 + β2X2 + ... + βn*Xn)

Here, p is the predicted probability of the event occurring, β0, β1, β2, ..., βn are the regression coefficients, and X1, X2, ..., Xn are the input features.

Model Training: The logistic regression model is trained by estimating the regression coefficients (β0, β1, β2, ..., βn) that best fit the data. The most common method for estimating these coefficients is maximum likelihood estimation (MLE). The goal is to find the coefficients that maximize the likelihood of the observed labels given the input features. This is typically done using optimization algorithms like gradient descent or numerical optimization techniques.

Decision Boundary: Once the model is trained and we have the coefficients, we can classify new instances by applying a decision threshold to the predicted probabilities. The threshold is usually set at 0.5. If the predicted probability is above the threshold, the instance is classified as one class (e.g., "positive" or "1"), and if it is below the threshold, it is classified as the other class (e.g., "negative" or "0").

Model Evaluation: The performance of a logistic regression model is typically evaluated using various metrics, including accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC). These metrics provide insights into how well the model classifies instances and its ability to discriminate between the two classes.

Regularization: To prevent overfitting and improve the model's generalization ability, logistic regression can be regularized. Regularization techniques like L1 regularization (Lasso) or L2 regularization (Ridge) can be applied to add a penalty term to the loss function. This encourages the model to shrink the coefficients towards zero, reducing complexity and potential overfitting.

Logistic regression is a powerful and interpretable model that allows us to predict probabilities and make binary classifications based on input features. While it is primarily used for binary classification, it can be extended to handle multi-class classification tasks using techniques like one-vs-rest or multinomial logistic regression.

15. What are the logistic regression assumptions ?

Ans: The assumptions of logistic regression can be summarized as follows:

Binary outcome: The dependent variable should be binary or dichotomous.

Independence of observations: The observations should be independent of each other.

Linearity in the logit: There should be a linear relationship between the independent variables and the logit of the dependent variable.

No multicollinearity: There should be no perfect multicollinearity among the independent variables.

Large sample size: Logistic regression performs well with a sufficiently large sample size. These assumptions help ensure the validity and reliability of the logistic regression model.

Basic assumptions that must be met for logistic regression include independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers

16. Go through the details of maximum likelihood estimation ?

Ans: Maximum likelihood estimation (MLE) is a statistical method used to estimate the parameters of a probability distribution or a statistical model by maximizing the likelihood function. It is a commonly used approach in various statistical models, including logistic regression.

Here are the details of maximum likelihood estimation:

Likelihood Function: The likelihood function represents the probability of observing the given data for a given set of parameter values. It is defined as the joint probability distribution of the observed data, treated as a function of the unknown parameters. In the case of logistic regression, the likelihood function represents the probability of observing the given set of binary outcomes (Y) for the given input features (X) and parameter values.

Log-Likelihood Function: Instead of working with the likelihood function directly, it is often more convenient to work with the log-likelihood function. Taking the logarithm of the likelihood function converts the product of probabilities into a sum of log-probabilities. The log-likelihood function is typically used because it simplifies calculations and allows for easier optimization.

Maximizing the Log-Likelihood: The goal of maximum likelihood estimation is to find the values of the model's parameters that maximize the log-likelihood function. This is achieved by taking derivatives of the log-likelihood function with respect to each parameter and setting them equal to zero. Solving these equations or using optimization algorithms (such as gradient descent) allows us to find the parameter values that maximize the log-likelihood function.

Iterative Optimization: In practice, finding the maximum likelihood estimates often involves iterative optimization algorithms. These algorithms iteratively update the parameter values to gradually approach the maximum of the log-likelihood function. Common optimization techniques include gradient descent, Newton-Raphson, and Fisher scoring.

Parameter Estimates: Once the optimization process converges, the estimated parameter values are obtained. These estimated parameters represent the maximum likelihood estimates, which are the values that maximize the likelihood function given the observed data.

Statistical Inference: Maximum likelihood estimation provides not only point estimates of the parameters but also information about their uncertainty. Standard errors, confidence intervals, and hypothesis tests can be derived based on the asymptotic properties of the maximum likelihood estimates. These statistical inferences allow us to make conclusions about the significance and precision of the estimated parameters.

Maximizing the likelihood function through maximum likelihood estimation is a fundamental principle in statistical modeling. It provides a powerful and widely used approach for estimating the parameters of various models, including logistic regression.

Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter values", are found such that they maximise the likelihood that the process described by the

In [ ]:

```
1
```