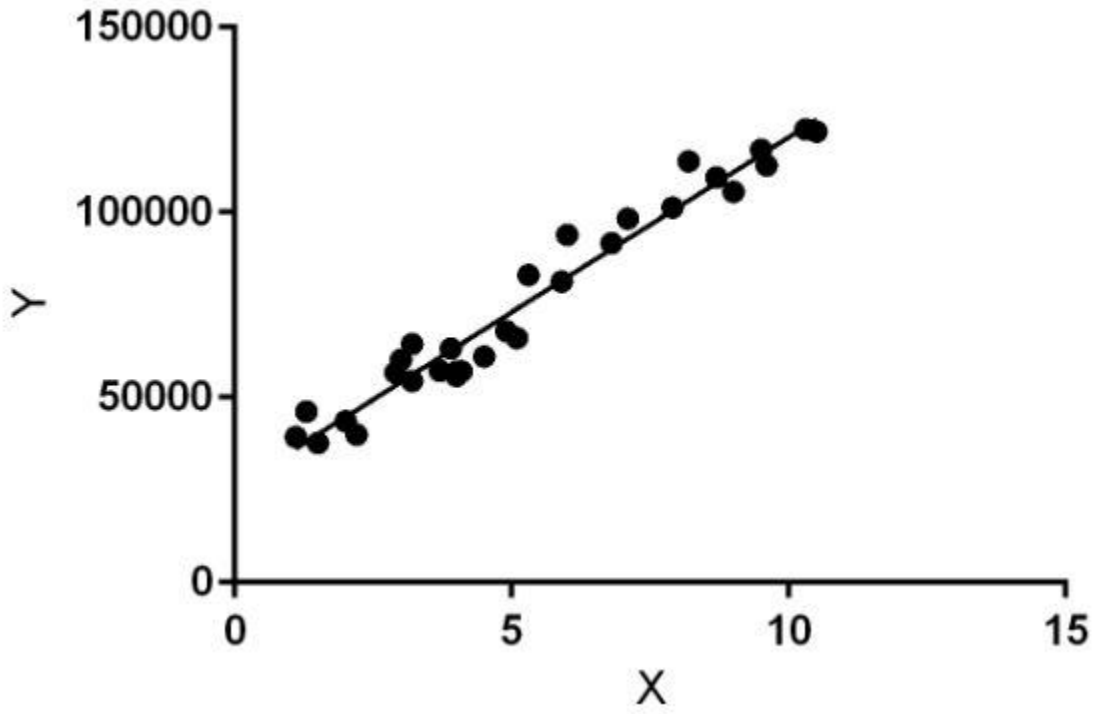
1. **Explain the linear regression algorithm in detail.**
   Answer:

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression:

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

How to update θ_1 and θ_2 values to get the best fit line ?

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

<u>Gradient Descent</u>:
To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

2. **What are the assumptions of linear regression regarding residuals?**
   Answer:

The mean of residuals is zero

No autocorrelation of residuals

The X variables and residuals are uncorrelated

The residuals should be normally distributed.

3. **What is the coefficient of correlation and the coefficient of determination?**
   Answer:

In simple linear regression analysis, the coefficient of correlation (or correlation coefficient) is a statistic which indicates an association between the independent variable and the dependent variable. The coefficient of correlation is represented by "r" and it has a range of -1.00 to +1.00.

When the coefficient of correlation is a *positive* amount, such as +0.80, it means the dependent variable is increasing when the independent variable is increasing. It also means that the dependent variable is decreasing when the independent variable is decreasing. However, a high positive correlation does not guarantee there is a cause and effect relationship. (A negative amount indicates an inverse association...the dependent variable is decreasing when the independent variable is increasing and vice versa.)

A coefficient of correlation of +0.8 or -0.8 indicates a *strong correlation* between the independent variable and the dependent variable. An *r* of +0.20 or -0.20 indicates a *weak correlation* between the variables. When the coefficient of correlation is 0.00 there is no correlation.

Relationship of Coefficient of Correlation to Coefficient of Determination

When the coefficient of correlation is squared, it becomes the coefficient of determination. This means that a coefficient of correlation of +0.80 will result in a *coefficient of determination* of 0.64 or 64%. (The coefficient of determination of 0.64 tells you that 64% of the change in the total of the dependent variable is associated with the change in the independent variable.) An *r* of +0.20 or -0.20 will result in an r-squared of only 4% (0.20 x 0.20), which means that only 4% of the change in the dependent variable is explained by the change in the independent variable.

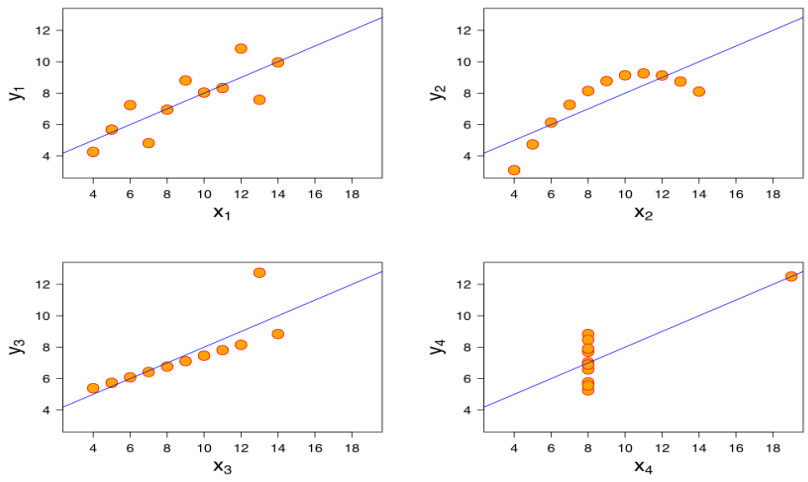
4. **Explain the Anscombe's quartet in detail.**
   Answer:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the

effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

For all four datasets:

Property	**Value**	**Accuracy**
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear Regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The *x* values are the same for the first three datasets.

Anscombe's quartet

I		**II**		**III**		**IV**	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed.

5. **What is Pearson's R?**
   Answer:
   Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Assumptions:

1. **Independent of case:** Cases should be independent to each other.
2. **Linear relationship:** Two variables should be linearly related to each other. This can be assessed with a scatterplot: plot the value of variables on a scatter diagram, and check if the plot yields a relatively straight line.
3. **Homoscedasticity:** the residuals scatterplot should be roughly rectangular-shaped.

Properties:

1. **Limit:** Coefficient values can range from +1 to -1, where +1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and a 0 indicates no relationship exists.
2. **Pure number:** It is independent of the unit of measurement. For example, if one variable's unit of measurement is in inches and the second variable is in quintals, even then, Pearson's correlation coefficient value does not change.
3. **Symmetric:** Correlation of the coefficient between two variables is symmetric. This means between X and Y or Y and X, the coefficient value of will remain the same.

Degree of correlation:

1. **Perfect:** If the value is near ± 1, then it said to be a perfect correlation: as one variable increases, the other variable tends to also increase (if positive) or decrease (if negative).
2. **High degree:** If the coefficient value lies between ± 0.50 and ± 1, then it is said to be a strong correlation.
3. **Moderate degree:** If the value lies between ± 0.30 and ± 0.49, then it is said to be a medium correlation.
4. **Low degree:** When the value lies below ± .29, then it is said to be a small correlation.
5. **No correlation:** When the value is zero.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
   Answer:

   Feature Scaling or Standardization: It is a step of Data Pre-Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

There are four common methods to perform Feature Scaling.

1. **Standardization:**

Standardization replaces the values by their Z scores.

$$x' = \frac{x - \bar{x}}{\sigma}$$

This redistributes the features with their mean **μ = 0** and standard deviation **σ =1**. sklearn.preprocessing.scale helps us implementing standardization in python.

2. Mean Normalization:

$$x' = \frac{x - \text{mean}(x)}{\max(x) - \min(x)}$$

This distribution will have values between **-1 and 1**with **μ=0**.

Standardization and **Mean Normalization** can be used for algorithms that assumes zero centric data like **Principal Component Analysis (PCA).**

3. Min-Max Scaling:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This scaling brings the value between 0 and 1.

4. Unit Vector:

$$x' = \frac{x}{||x||}$$

Scaling is done considering the whole feature vector to be of unit length.

Min-Max Scaling and **Unit Vector** techniques produces values of range [0,1]. When dealing with features with hard boundaries this is quite useful. For example, when dealing with image data, the colors can range from only 0 to 255.

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   Answer:

   These VIFs refers to that there is perfect collinearity: we have completely redundant variables. The very first thing we should do to address collinearity is to *think about what the variables mean. You should drop these kind of variables as it is of no meaning to be included in the model. For example:* coding for more than two genders, for instance, including an indicator of male is completely redundant with an indicator of female.

8. What is the Gauss-Markov theorem?
   Answer:

Gauss Markov Theorem

The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.

Gauss Markov Assumptions

There are five Gauss Markov assumptions (also called *conditions*):

1. **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
2. **Random**: our data must have been randomly sampled from the population.
3. **Non-Collinearity**: the regressors being calculated aren't perfectly correlated with each other.
4. **Exogeneity**: the regressors aren't correlated with the error term.
5. **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

Purpose of the Assumptions

The **Gauss Markov assumptions** guarantee the validity of ordinary least squares for estimating regression coefficients.

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are **rarely all met perfectly**, but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

The Gauss-Markov Assumptions in Algebra

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

$y_i = x_i' \beta + \varepsilon_i$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

- $E\{\varepsilon_i\} = 0, i = 1, \ldots, N$
- $\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots\ldots, x_N\}$ are independent
- $cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1, \ldots., N\ I \neq j.$
- $V\{\varepsilon_1 = \sigma^2, i= 1, \ldots.N$

The first of these assumptions can be read as "The expected value of the error term is zero.". The second assumption is collinearity, the third is exogeneity, and the fourth is homoscedasticity.

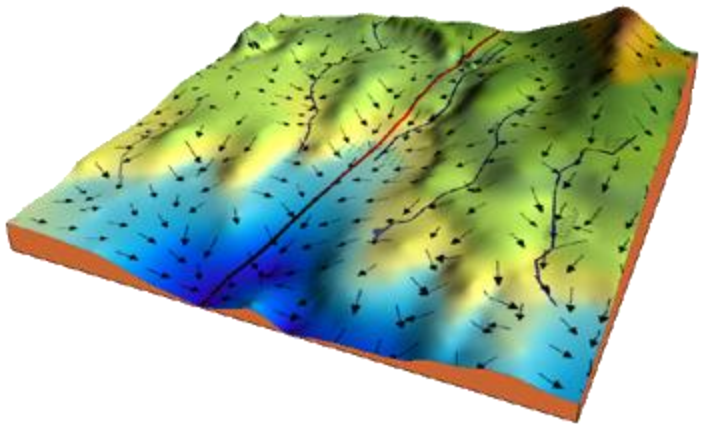
9. Explain the gradient descent algorithm in detail.
   Answer:

Gradient Descent

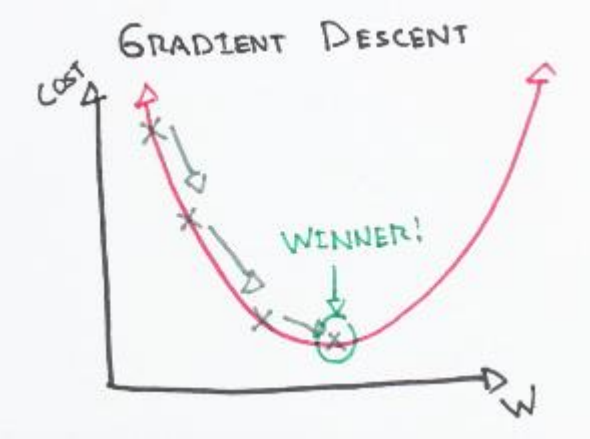
Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

Introduction

Consider the 3-dimensional graph below in the context of a cost function. Our goal is to move from the mountain in the top right corner (high cost) to the dark blue sea in the bottom left (low cost). The arrows represent the direction of steepest descent (negative gradient) from any given point–the direction that decreases the cost function as quickly as possible.



Starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next, we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.



Learning rate

The size of these steps is called the *learning rate*. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

Cost function

A Cost Function tells us “how good” our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.

Step-by-step

Now let’s run gradient descent using our new cost function. There are two parameters in our cost function we can control: *m*

(weight) and *b*

(bias). Since we need to consider the impact each one has on the final prediction, we need to use partial derivatives. We calculate the partial derivatives of the cost function with respect to each parameter and store the results in a gradient.

Math

Given the cost function:

$$f(m,b)=1N\sum i=1n(yi-(mxi+b))2$$

The gradient can be calculated as:

$$f'(m,b)=[[\lfloor dfdmdfdb \rceil]]=[1N\sum -2xi(yi-(mxi+b))1N\sum -2(yi-(mxi+b))]$$

To solve for the gradient, we iterate through our data points using our new *m* and *b* values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    Answer:

Purpose: Check If Two Data Sets Can Be Fit With the Same Distribution	The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the

0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

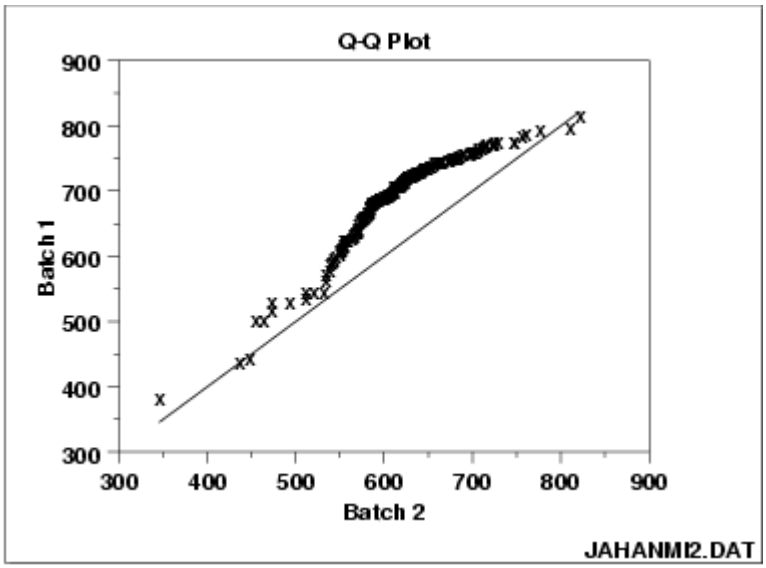
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

Sample Plot



This q-q plot of the JAHANMI2.DAT data set shows that

1. These 2 batches do not appear to have come from populations with a common distribution.
2. The batch 1 values are significantly higher than the corresponding batch 2 values.
3. The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

Definition: Quantiles for Data Set 1 Versus Quantiles of Data Set 2

The q-q plot is formed by:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points, but not what that quantile level actually is.

If the data sets have the same size, the q-q plot is essentially a plot of sorted data set 1 against sorted data set 2. If the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

Questions

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

Importance: Check for Common Distribution

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.