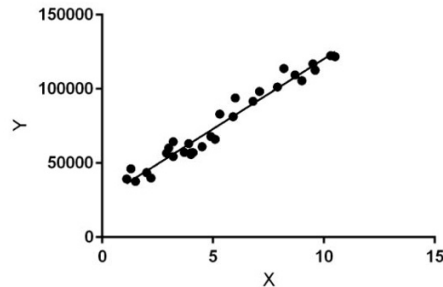


1. Explain the linear regression algorithm in detail.

Ans: **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 :

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

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).

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

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

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

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

Ans: Building a linear regression is only half of the work. In order to actually predict something, the model should conform to the assumptions :

Assumption 1. - The regression model is linear in pattern.

$$Y = a + (\beta_1 * X_1) + (\beta_2 * X_2^2)$$

Though, the X_2 is raised to power 2, the equation is still linear in beta parameters. So the assumption is satisfied in this case.

Assumption 2. – The mean of residuals is zero or very close to zero.

Assumption 3. – Homoscedasticity of residuals or equal variance

Assumption 4. – No autocorrelation of residuals.

This is applicable especially for time series data. Auto-correlation is the correlation of time series with lags of itself. When residuals are auto correlated, it means that the current callus is dependent of the previous values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.

To check autocorrelation residuals, there are 3 ways to do:

1. using acf plot
2. using run test
3. using Durbin-Watson test

Assumption 5. – The X-variables and residuals are uncorelated.

Assumption 6. – The number of observations must be greater than number of Xs which can be directly observed by looking data.

Assumption 7. – The variability of X values is positive. This means X values in given sample must not all be the same or even nearly same.

Assumption 8. – The regression model is correctly specified. This means that if Y and X variable has an inverse relationship, the model equation should be specified appropriately:

$$Y = \beta_1 + \beta_2 * \left(\frac{1}{X} \right)$$

Assumption 9. – No perfect multicollinearity. There is no perfect linear relationship between explanatory variables.

Assumption 10. – Normality of residuals. The residuals should be normally distributed. If the maximum likelihood method (not OLS) is used to compute the estimates, this also implies the Y and the Xs are also normally distributed.

3. What is the coefficient of correlation and the coefficient of determination?

Ans: How well does your regression equation truly represent set of data? To answer this, one of the ways to determine the answer to this equation is to exam the correlation coefficient and the coefficient of determination.

Coefficient of correlation (r):

- The quantity r , called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honour of its developer Karl Pearson.
- The mathematical formula for computing r is:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$

where n is the number of pairs of data.

- The value of r is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
- **Positive Correlation** : If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- **Negative Correlation** : If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- **No Correlation** : If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables
- A **perfect correlation** of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*.

Coefficient of determination(r^2 or R^2):

- The *coefficient of determination*, r^2 , is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
- The *coefficient of determination* is the ratio of the explained variation to the total variation.
- The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y .
- The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

4. Explain the Anscombe's quartet in detail.

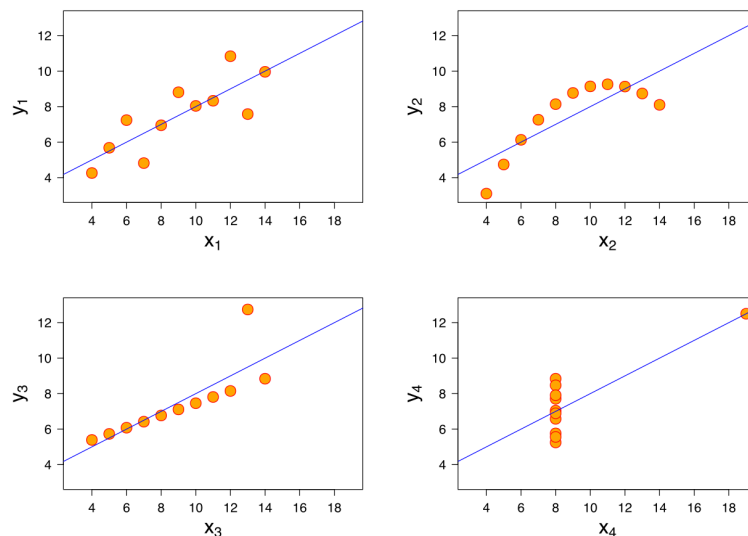
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x , y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY**, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

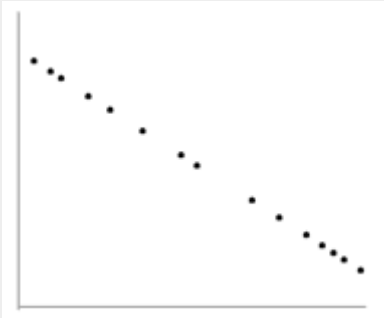

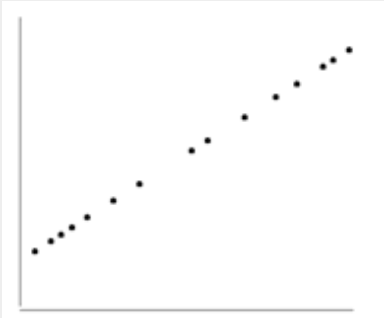
5. What is Pearson's R?

Correlation is a technique for investigating the relationship between two quantitative, continuous variables. E.g. Age and blood pressure. Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

- This we can get by drawing scatter plot of the variables to check the linearity.
- The correlation coefficient should not be calculated if the relationship is not linear.
- Formula to get r:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

r = -1		data lie on a perfect straight line with a negative slope
r = 0		no linear relationship between the variables
r = +1		data lie on a perfect straight line with a positive slope

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

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In scaling (*also called min-max scaling*), you transform the data such that the features are within a specific range e.g. [0, 1].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x' is the normalized value.

Most of the time, your dataset will contain features highly varying in magnitude, units and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem. Eg. If left alone, these algorithms only take in the magnitude of features neglecting the units. The results would vary greatly between different units, 5kg and 5000gms. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes.

To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

For most applications standardization is recommended.

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

An infinite VIF will be returned for two variables that are exactly collinear, variables that are exactly the same or linear transformations of each other. Perhaps check your variables for exact collinearity, using the `pairs` function or something similar.

In VIF, each feature is regression against all other features. If R^2 is more which means this feature is correlated with other features. [0]

- $VIF = 1 / (1 - R^2)$
- When R^2 reaches 1, VIF reaches infinity

8. What is the Gauss-Markov theorem?

A theorem that proves that if the error terms in a multiple regression have the same variance and are uncorrelated, then the estimators of the parameters in the model produced by least squares estimation are better (in the sense of having lower dispersion about the mean) than any other unbiased linear estimator.

The proof for this theorem goes way beyond the scope of this blog post. However, the critical point is that when you satisfy the classical assumptions, you can be confident that you are obtaining the best possible coefficient estimates. The Gauss-Markov theorem does not state that these are just the best possible estimates for the OLS procedure, but the best possible estimates for *any* linear model estimator.

The classical assumptions of OLS linear regression, I explain those assumptions and how to verify them. In this post, I take a closer look at the nature of OLS estimates. What does the Gauss-Markov theorem mean exactly when it states that OLS estimates are the best estimates when the assumptions hold true?

The Gauss-Markov theorem famously states that OLS is BLUE. 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. More specifically, when your model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

9. Explain the gradient descent algorithm in detail.

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

Batch Gradient Descent: This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.

Let $h_{\theta}(x)$ be the hypothesis for linear regression. Then, the cost function is given by:

Let Σ represents the sum of all training examples from $i=1$ to m .

```
J_train(θ) = (1/2m) Σ( h_θ(x(i)) - y(i))2

Repeat {
  θ_j = θ_j - (learning rate/m) * Σ( h_θ(x(i)) - y(i))x_j(i)
  For every j = 0 ... n
}
```

Where $x_j^{(i)}$ Represents the j^{th} feature of the i^{th} training example. So if m is very large, then the derivative term fails to converge at the global minimum.

Stochastic Gradient Descent: This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.

- 1) Randomly shuffle the data set so that the parameters can be trained evenly for each type of data.
- 2) As mentioned above, it takes into consideration one example per iteration.

Hence,

Let $(x^{(i)}, y^{(i)})$ be the training example

$\text{Cost}(\theta, (x^{(i)}, y^{(i)})) = (1/2) \Sigma (h_{\theta}(x^{(i)}) - y^{(i)})^2$

$J_{\text{train}}(\theta) = (1/m) \Sigma \text{Cost}(\theta, (x^{(i)}, y^{(i)}))$

Repeat {

For $i=1$ to m {

$\theta_j = \theta_j - (\text{learning rate}) * \Sigma (h_{\theta}(x^{(i)}) - y^{(i)})x_j^{(i)}$

For every $j = 0 \dots n$

}

}

Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where $b < m$ are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

Say b be the no of examples in one batch, where $b < m$. Assume $b = 10$, $m = 100$; **Note:** However we can adjust the batch size. It is generally kept as power of 2. The reason behind it is because some hardware such as GPUs achieve better run time with common batch sizes such as power of 2.

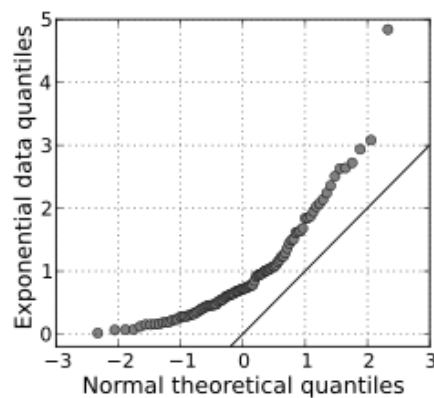
```
Repeat {  
  For  $i=1, 11, 21, \dots, 91$   
  
    Let  $\Sigma$  be the summation from  $i$  to  $i+9$  represented by  $k$ .  
  
     $\theta_j = \theta_j - (\text{learning rate/size of } (b)) * \Sigma (h_{\theta}(x^{(k)}) - y^{(k)})x_j^{(k)}$   
    For every  $j = 0 \dots n$   
}
```

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

The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. For example, if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line as below:



Q-Q plots are ubiquitous in statistics. Most people use them in a single, simple way: fit a linear regression model, check if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator $\hat{\beta}$ is Gaussian either, so the standard confidence intervals and significance tests are invalid.