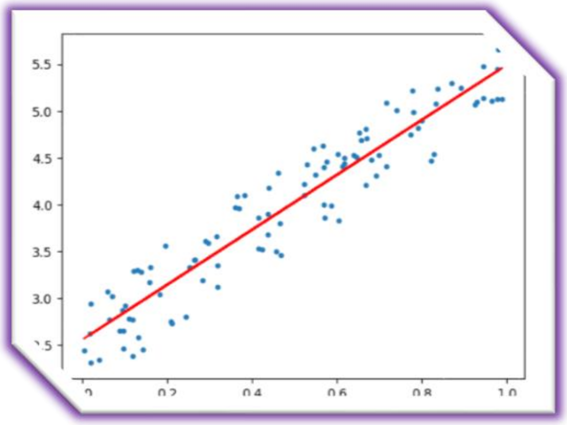
Linear Regression Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression – Machine Learning Algorithm



In simple terms, linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

Two types of linear regression under this module:

- Simple linear regression
  - It explains the relationship between a dependent variable and one independent variable using a straight line
  - $Y = \beta_0 + \beta_1.X$
    - β_0 -> Intercept
    - β_1 -> Slope
- Multiple linear regression
  - It explains the relationship between one dependent variable and several independent variables
  - $Y = \beta_0 + \beta_1.X1 + \beta2.X2+ \beta3. Xn……… \beta n.Xn$

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

The assumptions of linear regression are:

The assumption about the form of the model:

It is assumed that there is a linear relationship between the dependent and independent variables. It is known as the ‘linearity assumption’.

Assumptions about the residuals:

Normality assumption: It is assumed that the error terms, $\varepsilon(i)$, are normally distributed.

Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, $\sigma2$. This assumption is also known as the assumption of homogeneity or homoscedasticity.

Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

Assumptions about the estimators:

The independent variables are measured without error.

The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

Explanations:

This is self-explanatory.

If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data.
Also, the mean of the residuals should be zero.
$Y(i)i= \beta0+ \beta1x(i) + \varepsilon(i)$

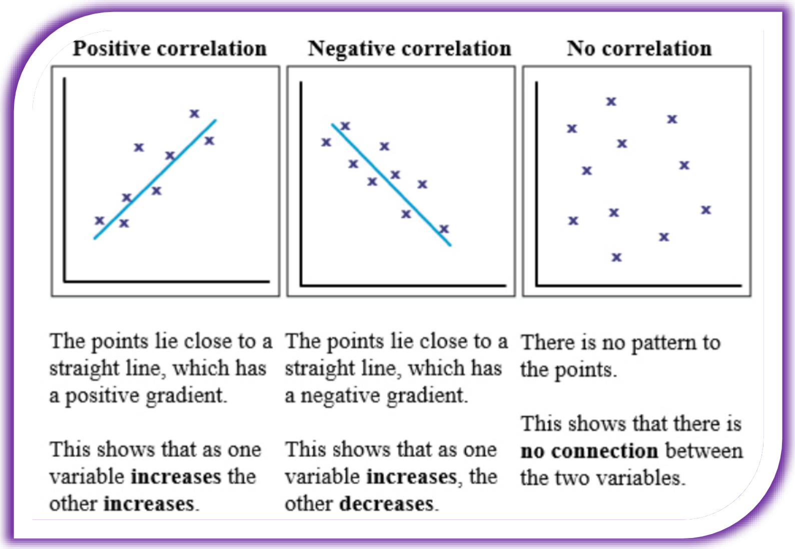
This is the assumed linear model, where ε is the residual term.
$E(Y) = E(\beta0+ \beta1x(i) + \varepsilon(i))$
$\quad = E(\beta0+ \beta1x(i) + \varepsilon(i))$
If the expectation(mean) of residuals, $E(\varepsilon(i))$, is zero, the expectations of the target variable and the model become the same, which is one of the targets of the model.
The residuals (also known as error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it implies that there is some relation that the regression model is not able to identify.

If the independent variables are not linearly independent of each other, the uniqueness of the least squares solution (or normal equation solution) is lost.

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

Correlation of coefficients:

- It is used in statistics to measure how powerful a relationship between two variables and Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1
    - 1 indicates a strong positive relationship.
    - -1 indicates a strong negative relationship.
    - A result of zero indicates no relationship at all.



Positive correlation

The points lie close to a straight line, which has a positive gradient.

This shows that as one variable **increases** the other **increases**.

Negative correlation

The points lie close to a straight line, which has a negative gradient.

This shows that as one variable **increases**, the other **decreases**.

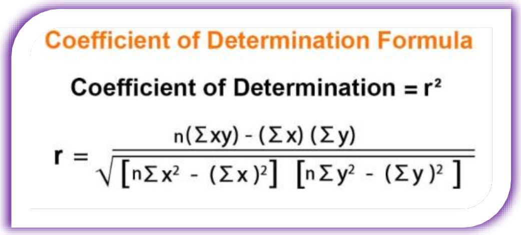
No correlation

There is no pattern to the points.

This shows that there is **no connection** between the two variables.

Correlation of determination:

- It is used in statistical analysis that evaluates how strong a model describes and predicts future results. It is commonly denoted as "R-Squared"
- The coefficient of determination is the square of the correlation (r) between predicted y scores and actual y scores; so, it ranges from 0 to 1.
- R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- R2 of 1 means the dependent variable can be predicted without error from the independent variable.



Coefficient of Determination Formula

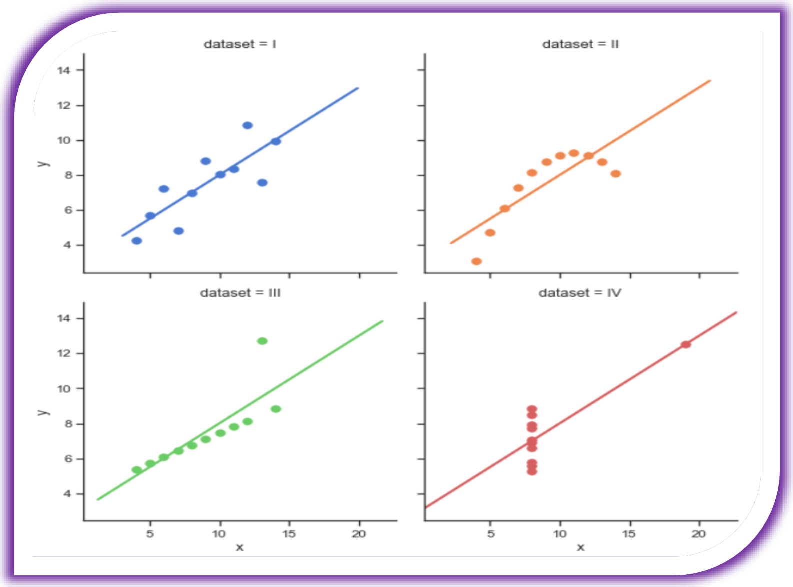
Coefficient of Determination = r^2

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

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet contains 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. It is used to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties

- 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?

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

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

An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables

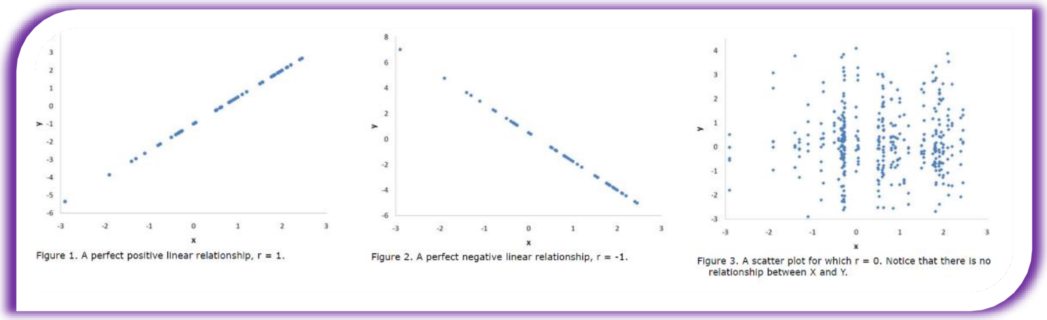


Figure 1. A perfect positive linear relationship, r = 1.

Figure 2. A perfect negative linear relationship, r = -1.

Figure 3. A scatter plot for which r = 0. Notice that there is no relationship between X and Y.

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

Scaling:

Scaling 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 range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Difference between normalized scaling and standardized scaling:

Normalized 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.

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardized Scaling:

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

$$X_{changed} = \frac{X - \mu}{\sigma}$$

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. This is referred to as the problem of multicollinearity. The problem is that, as the Xs become more highly correlated, it becomes more and more difficult to determine which X is producing the effect on Y.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

8. What is the Gauss-Markov theorem?

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.

When estimating regression models, we know that the results of the estimation procedure are random. However, when using unbiased estimators, at least on average, we estimate the true parameter. When comparing different unbiased estimators, it is therefore interesting to know which one has the highest precision: being aware that the likelihood of estimating the exact value of the parameter of interest is 00 in an empirical application, we want to make sure that the likelihood of obtaining an estimate very close to the true value is as high as possible. This means we want to use the estimator with the lowest variance of all unbiased estimators, provided we care about unbiasedness. The Gauss-Markov theorem states that, in the class of conditionally unbiased linear estimators, the OLS estimator has this property under certain conditions.

9. Explain the gradient descent algorithm in detail.

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.

In Other words, this optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).

Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.

Gradient Descent Procedure:

The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

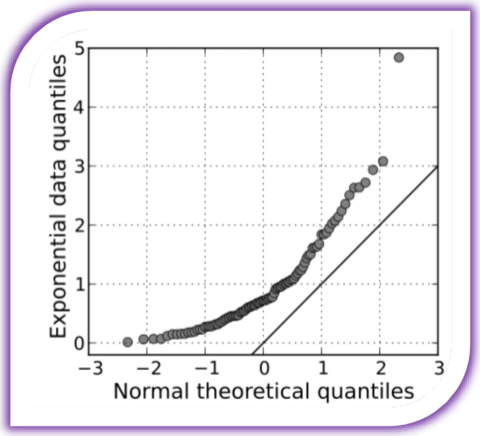
Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

You can see how simple gradient descent is. It does require you to know the gradient of your cost function or the function you are optimizing, but besides that, it's very straightforward. Next we will see how we can use this in machine learning algorithms.

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



Any quantile-to-quantile plot will plot on the x-axis the quantiles of one variable and on the y-axis the quantiles of the other variable

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.First, the set of intervals for the quantiles is chosen. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). Thus the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions

This Q–Q plot compares a sample of data on the vertical axis to a statistical population on the horizontal axis

Reference link:

https://machinelearningmastery.com/gradient-descent-for-machine-learning/
https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html
https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem
http://www.imagelab.at/help/vif_descriptors.htm
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
https://www.statisticssolutions.com/pearsons-correlation-coefficient/
https://medium.com/datadriveninvestor/anscombes-quartet-12649db7eac0
https://en.wikipedia.org/wiki/Anscombe%27s_quartet
https://seaborn.pydata.org/examples/anscombes_quartet.html
https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
https://www.investopedia.com/terms/c/correlationcoefficient.asp
https://boostedml.com/2019/03/linear-regression-plots-how-to-read-a-qq-plot.html
https://data.library.virginia.edu/understanding-q-q-plots/