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

Linear Regression is a linear approach to modeling the relationship between a scalar variable (or dependent variable or y) and one or more explanatory variables (or independent variables or x). Mathematically speaking, linear regression equation:
y = mx + c, where y is the dependent variable, x is the independent variable, m is the slope of the line and c is the y-intercept (*value of **y** when x=0*)

Linear Regression Algorithm:

- Load and visualize the data in the dataset.
- Draw the scatterplot.
    (a) Observe for the linear and non-linear pattern of the data
    (b) Look for the outliers – If outliers are present try to treat the outliers.
    (c) If the data pattern is non-linear transform the columns as required.
- Fit the least-squares regression line to the data and check the assumptions of the model by looking at the Residual and normal probability plot (for normality assumption). If the assumptions of the model appear not to be met, a transformation or scaling of the data may be required.
- If necessary, transform the data and re-fit the least-squares regression line using the transformed data.
- Once a "good-fitting" model is determined, write the equation of the least-squares regression line. Include the standard errors of the estimates, the estimate of σ, and R-squared.
- Determine if the explanatory variable is a significant predictor of the response variable by performing a t-test or F-test. Include a confidence interval for the estimate of the regression coefficient (slope and intercept).
- Check for the error distribution and check for the performance.

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

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task to compute the regression coefficients. Regression models a target prediction based on independent variables.

Assumptions of Linear Regression

There are 5 basic assumptions of Linear Regression Algorithm:

1. **Linear Relationship between the features and target:**

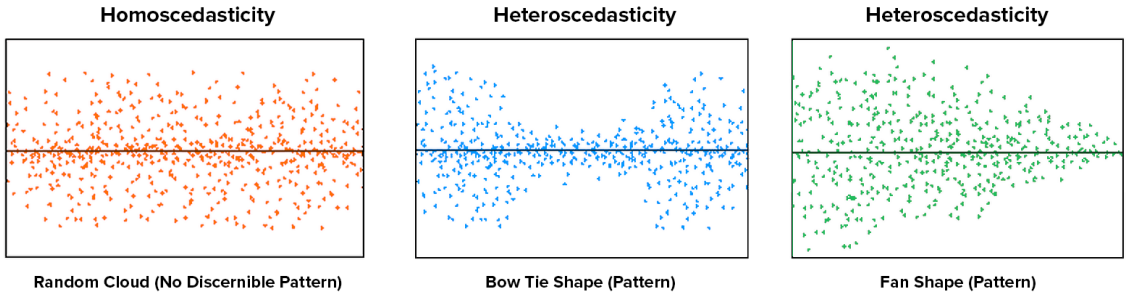
Linear Regression captures only linear relationships. This can be validated by drawing the scatter plot between the independent and dependent variables.

2.Little or no Multicollinearity between the features:

Multicollinearity is a state of very high inter-correlations or inter-association among the independent variables. Using the pair plots and heat maps we can observe the collinearity and if the high collinearity is exist, we can remove those features. If the correlation between the variables is stronger then it becomes difficult for the model to estimate the relationship between the independent and dependent variables. Hence it is suggested to remove the multi collinearity variables in the data frame.

3.Homoscedasticity Assumption:

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a goodway to check for homoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.



4.Normal distribution of error terms:

The residual errors follows the normal distribution , if the sample size increases, the normality distribution for the residuals is not needed, but in the linear regression model, we assume the residual terms(error) is always normal distribution. Normal distribution of the residuals can be validated by plotting a q-q plot.

5.Little or No autocorrelation in the residuals:

Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.

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

Coefficient of determination: R square is also called coefficient of determination, is the proportion of the variance in the dependent variable that is predictable from the independent variables. In other word Coefficient of Determination is the square of Coefficient of Correlation

Coefficient of Correlation: It is the degree of relationship between two variables say x and y. It is always between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two

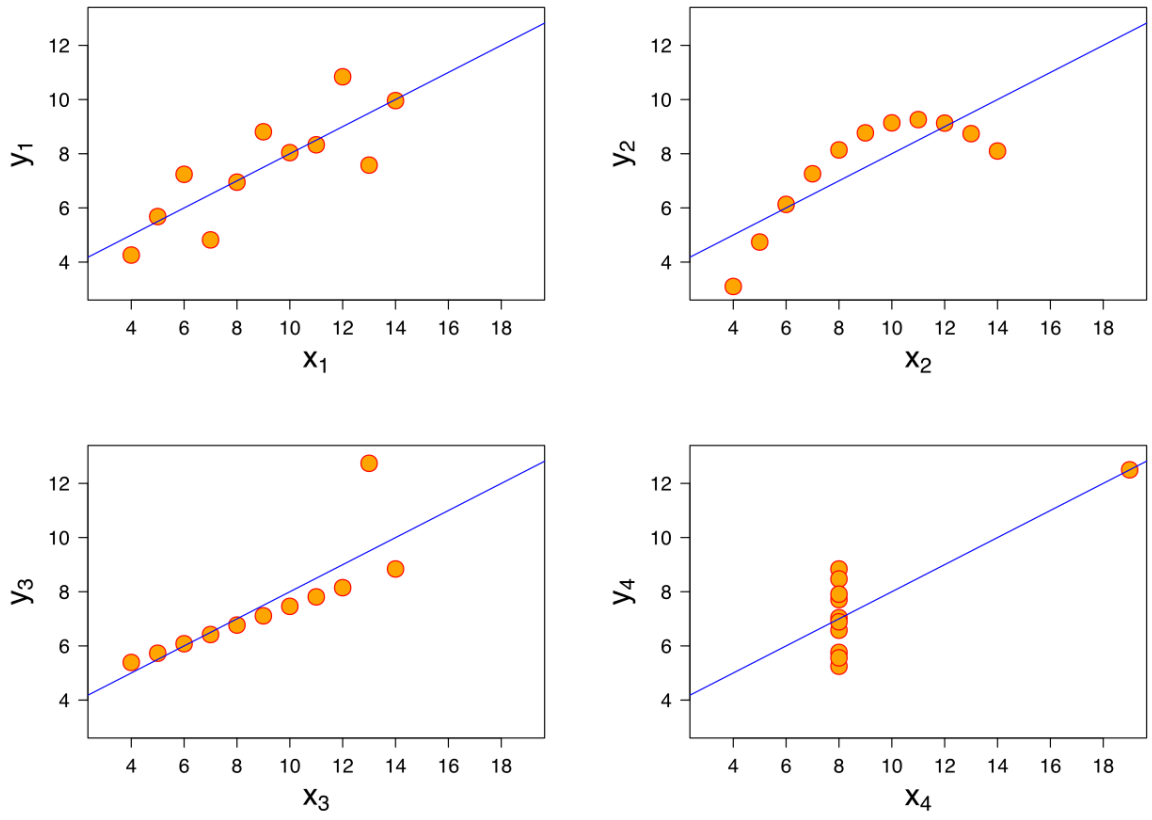
variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value.

R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.

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.



- Dataset I appears to be a best-fitting linear regression model.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

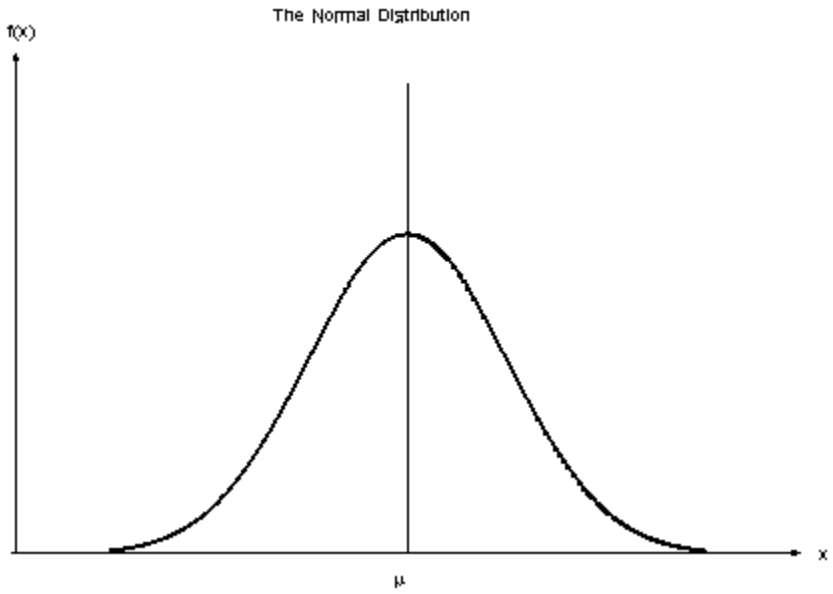
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

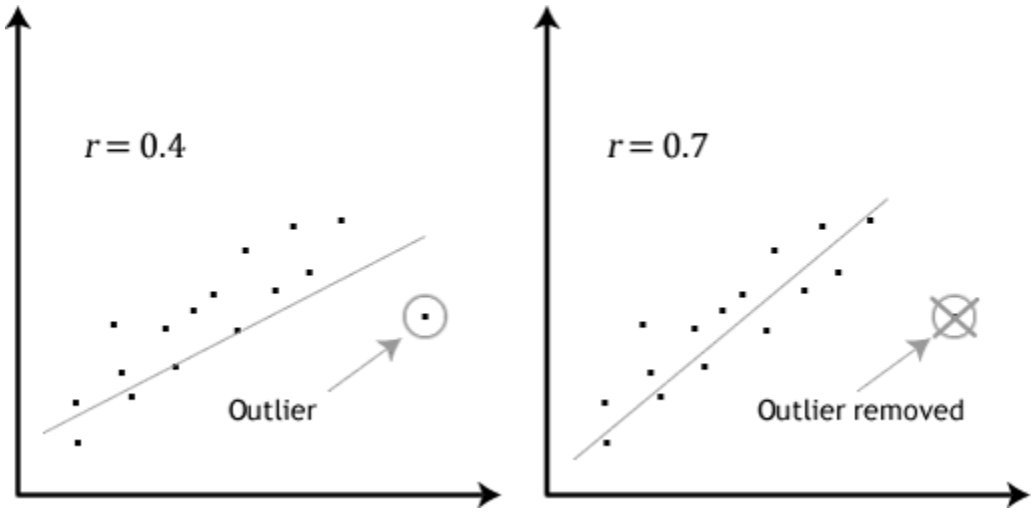
5. What is Pearson's R?

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r. You'll come across Pearson r correlation
Assumptions

1. For the Pearson r correlation, both variables should be **normally distributed**. i.e the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'.

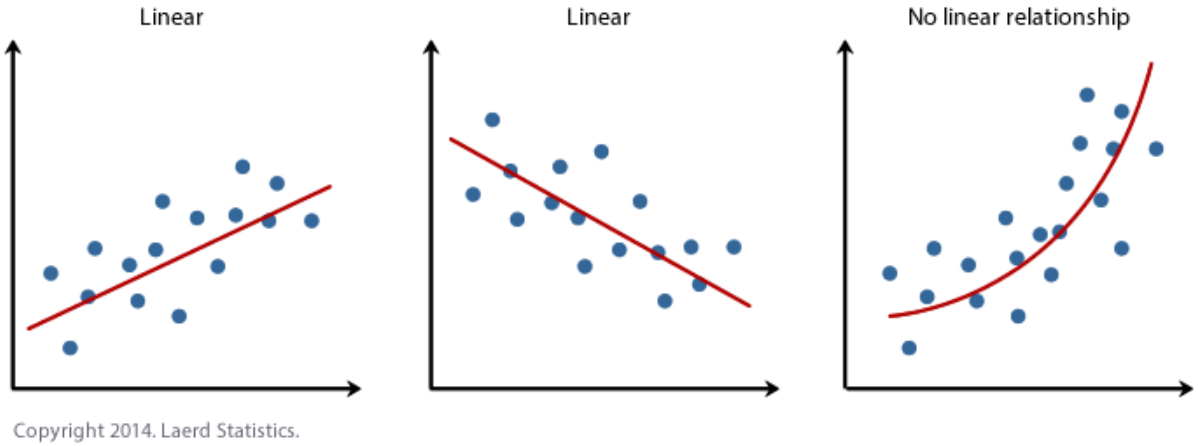


2. There should be **no significant outliers**. We all know what outliers are but we don't know the effect of outliers on Pearson's correlation coefficient, r. Pearson's correlation coefficient, r, is very sensitive to outliers, which can have a very large effect on the line of best fit and the Pearson correlation coefficient. This means — including outliers in your analysis can lead to misleading results.



3. Each variable should be **continuous** , If one or both of the variables are ordinal in measurement, then a Spearman correlation could be conducted instead.

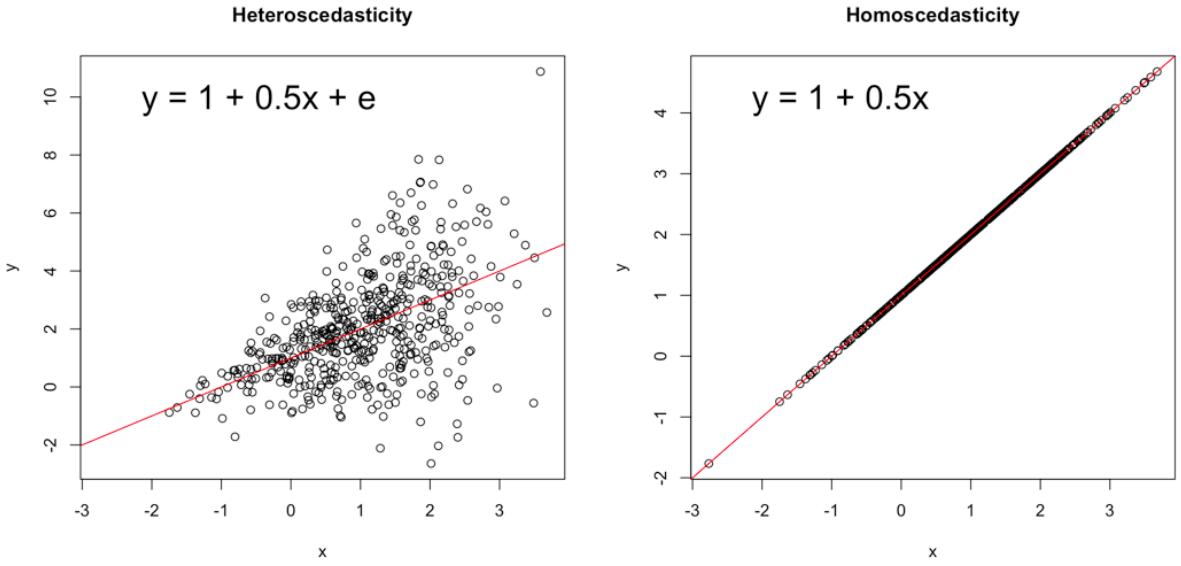
4. The two variables have a **linear relationship**. Scatter plots will help you tell whether the variables have a linear relationship. If the data points have a straight line (and not a curve), then the data satisfies the linearity assumption. If the data you have is not linearly related you might have to run a non-parametric .



5. The observations are **paired observations.** That is, for every observation of the independent variable, there must be a corresponding observation of the dependent variable.

6. **Homoscedascity**. Homoscedascity simply refers to '**equal variances**'. A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit,

then the data is homoscedastic. As a bonus — the opposite of homoscedascity is heteroscedascity which refers to refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it.



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

Scaling is the It is a step of Data Pre Processing which is applied to variables 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.

In regression, as long as you have a bias it does not matter if you normalize or not since you are discovering an affine map, and the composition of a scaling transformation and an affine map is still affine.

When there are learning rates involved, e.g. when you're doing gradient descent, the input scale effectively scales the gradients, which might require some kind of second order method to stabilize per-parameter learning rates. It's probably easier to normalize the inputs if it doesn't matter otherwise.

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.

Xchanged=X−Xmin / Xmax−Xmin

Standardization rescales data to have a mean (μμ) of 0 and standard deviation (σσ) of 1 (unit variance).
Xchanged=X−μσ

For most applications standardization is recommended.

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

VIF - the variance inflation factor.

If the two variables are perfectly correlated, then R-Squared value will become 1. In such a case, VIF becomes infinity.

The simple mathematical formulae for VIF is
VIF of X1 = 1/(1 - R-squared of X1).

In the above case, R-Squared is 1, as the variables are perfectly correlated, then,
VIF = 1/(1-1)= 1/0 = infinity.

The biggest factor that, VIF can become infinity, is the low sample size.

8. What is the Gauss-Markov theorem?

The Gauss Markov theorem says that, under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE), that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.

Assumptions

There are few assumptions are considered for the Gauss-Markov theorem, some of them are:

- The linear model is a vector of observations of the output variables or sample size.
- It is a matrix of inputs i.e., it is the number of inputs for each observation.
- It is a vector of regression coefficients and errors.

9. Explain the gradient descent algorithm in detail.

Gradient Descent is the most common optimization algorithm in *machine learning* and *deep learning*. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function *J(w)* w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

Gradient Descent algorithm:

Step 1: Initialize the weights (a & b) with random values and calculate Error (SSE):
In this step, we need to fit a line from the random variables of a& band calculate the prediction error(SSE).

Step 2: Calculate the gradient :
In this step, Change in SSE when the weights (a & b) are changed by a very small value from their original randomly initialized value. This helps us move the values of a & b in the direction in which SSE is minimized.

Step 3: Adjust the weights with the gradients to reach the optimal values where SSE is minimized
In this step, We need to update the random values of a, b so that we move in the direction of optimal a, b.

Step 4: Use the new weights for prediction and to calculate the new SSE

Step 5: Repeat steps 2 and 3 till further adjustments to weights doesn't significantly reduce the Error

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

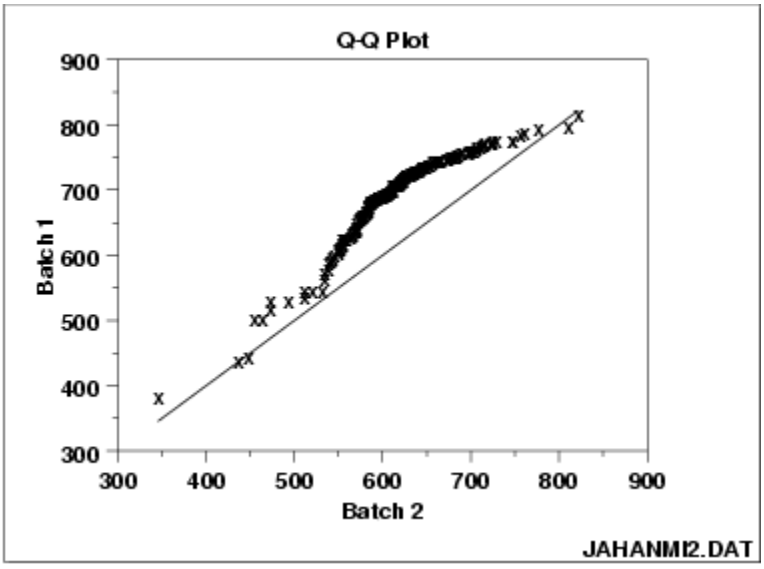
Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if 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.

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.



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.