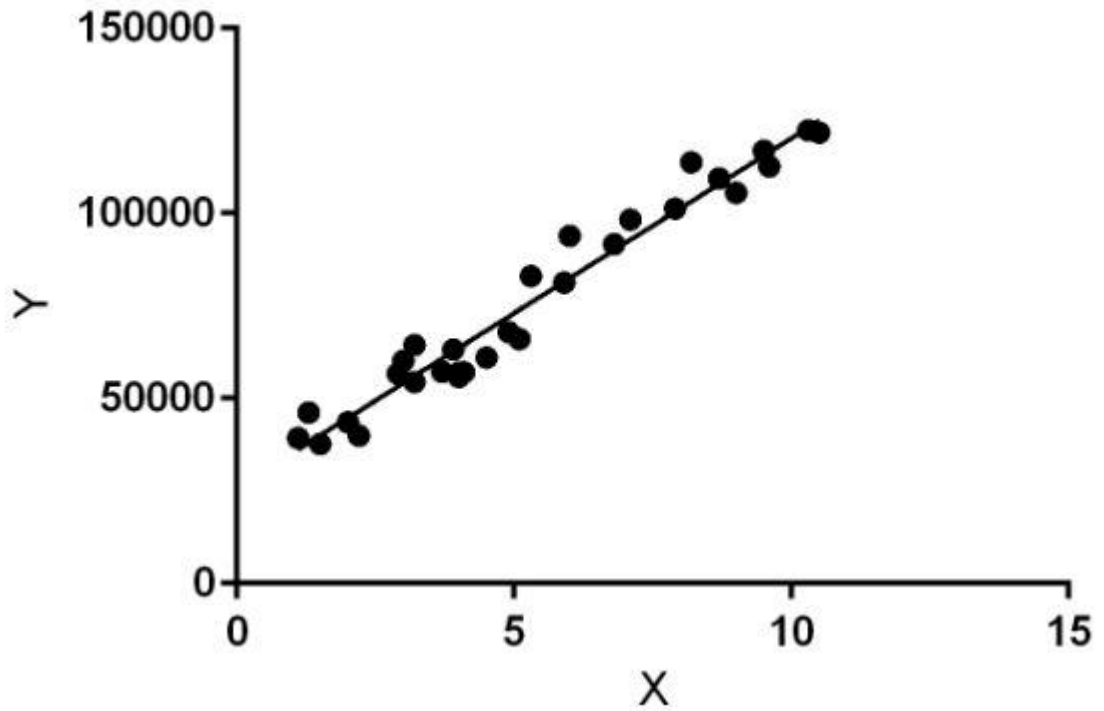
Linear Regression Subjective Questions

BY

NAME : VENKATESH . P

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 :

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

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).

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

Ans : There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
No Multicollinearity—Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.

Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.
Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis.

Assumptions of Multiple Linear Regression

Multiple linear regression analysis makes several key assumptions:

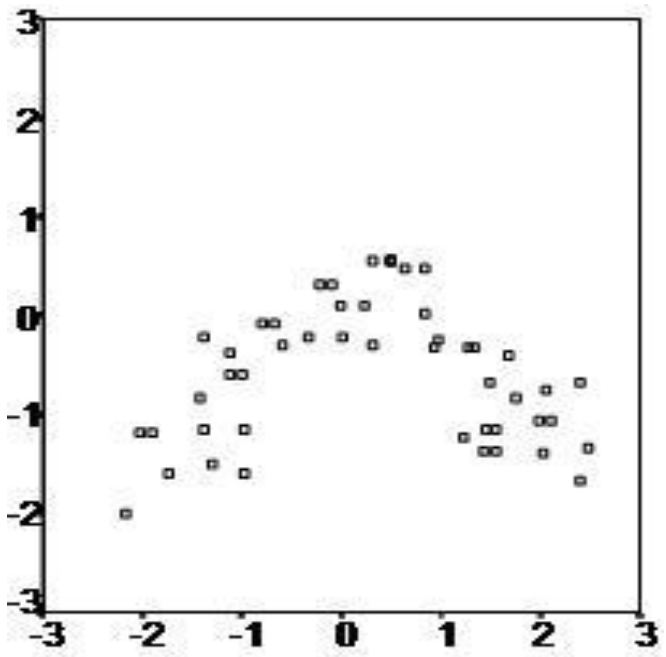
- There must be a linear relationship between the outcome variable and the independent variables. Scatterplots can show whether there is a linear or curvilinear relationship.
- Multivariate Normality–Multiple regression assumes that the residuals are normally distributed.
- No Multicollinearity— Multiple regression assumes that the independent variables are not highly correlated with each other. This assumption is tested using Variance Inflation Factor (VIF) values.
- Homoscedasticity–This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

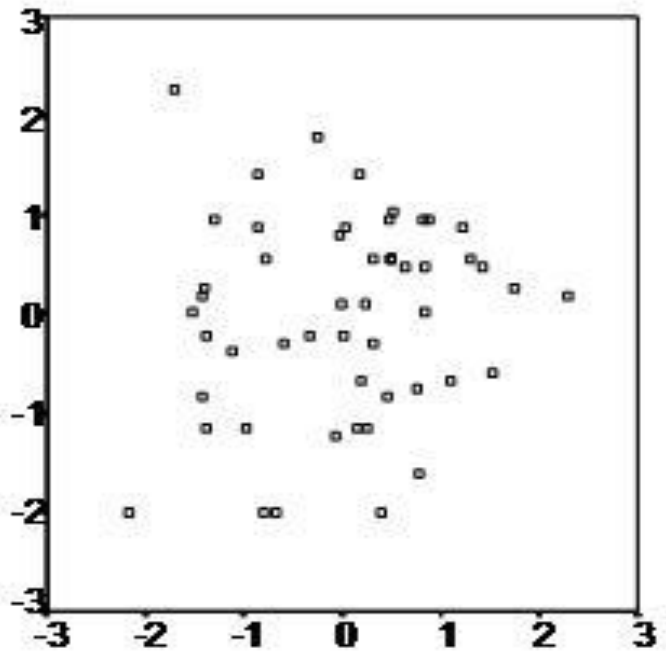
Intellectus Statistics automatically includes the assumption tests and plots when conducting a regression.

Multiple linear regression requires at least two independent variables, which can be nominal, ordinal, or interval/ratio level variables. A rule of thumb for the sample size is that regression analysis requires at least 20 cases per independent variable in the analysis.

Multiple Linear Regression Assumptions

First, multiple linear regression requires the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatterplots. The following two examples depict a curvilinear relationship (left) and a linear relationship (right).

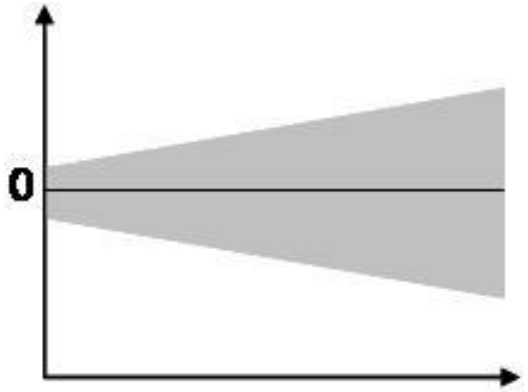


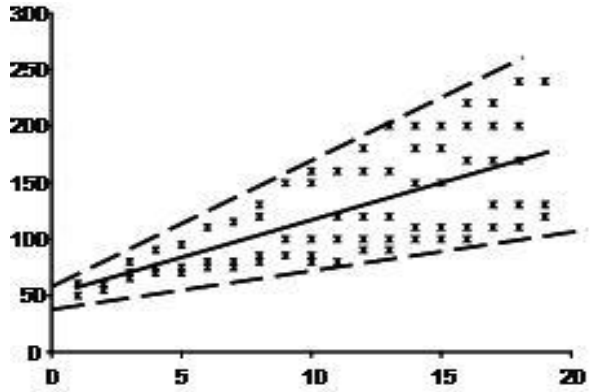


Second, the multiple linear regression analysis requires that the errors between observed and predicted values (i.e., the residuals of the regression) should be normally distributed. This assumption may be checked by looking at a histogram or a Q-Q-Plot. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves.

Third, multiple linear regression assumes that there is no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be checked multiple ways:
1) Correlation matrix – When computing a matrix of Pearson's bivariate correlations among all independent variables, the magnitude of the correlation coefficients should be less than .80.
2) Variance Inflation Factor (VIF) – The VIFs of the linear regression indicate the degree that the variances in the regression estimates are increased due to multicollinearity. VIF values higher than 10 indicate that multicollinearity is a problem.
If multicollinearity is found in the data, one possible solution is to center the data. To center the data, subtract the mean score from each observation for each independent variable. However, the simplest solution is to identify the variables causing multicollinearity issues (i.e., through correlations or VIF values) and removing those variables from the regression.
The last assumption of multiple linear regression is homoscedasticity. A scatterplot of residuals versus predicted values is good way to check for homoscedasticity. There should be no clear pattern in the distribution; if there is a cone-shaped pattern (as shown below), the data is heteroscedastic.





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

Ans : 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:

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

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*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger
correlation than a study using social science data.

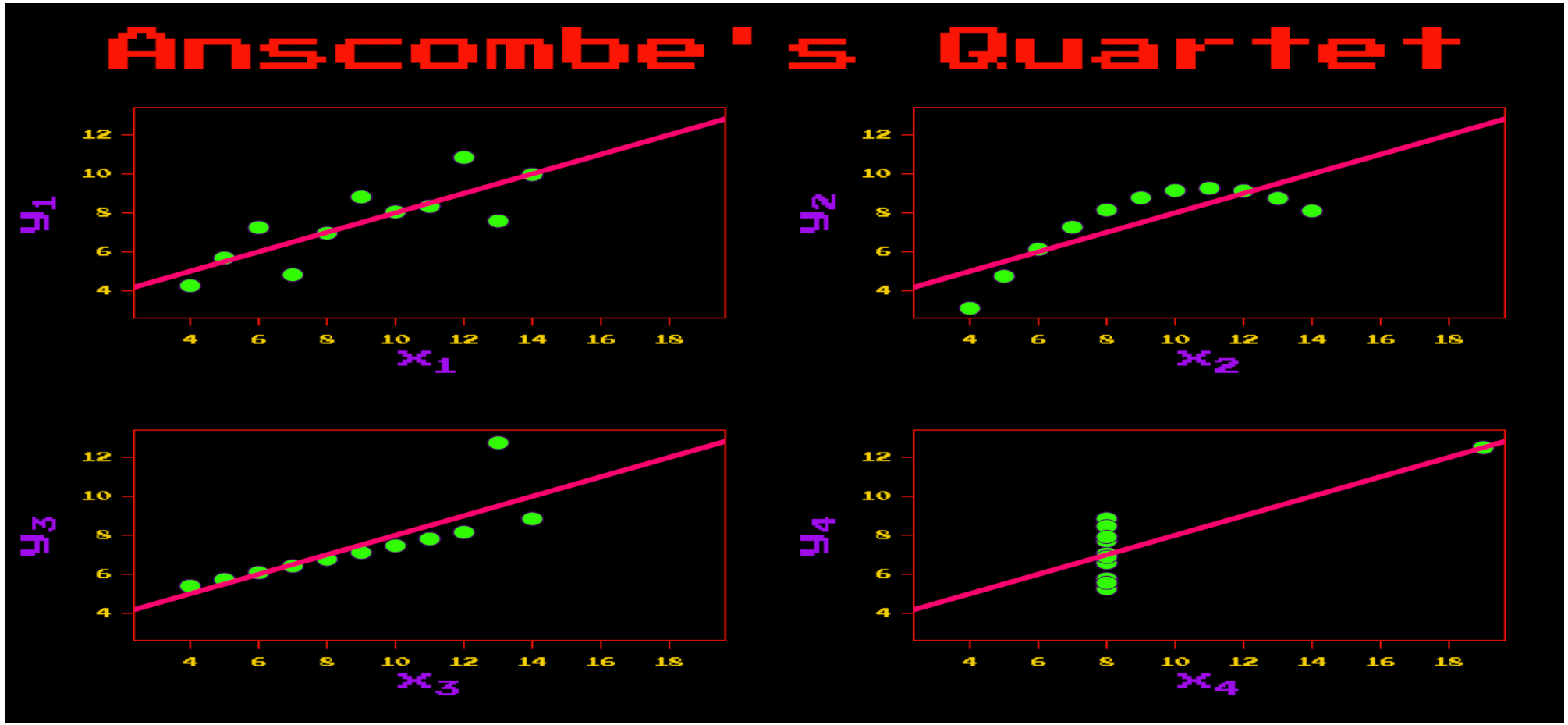
The coefficient of determination, R^2, is used to analyse how differences in one variable can be explained by a difference in a second variable. For example, *when* a person gets pregnant has a direct relation to when they give birth.
More specifically, R-squared gives you the percentage variation in y explained by x-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables.

The coefficient of determination can be thought of as a percent. It gives you an idea of how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted. If the coefficient is 0.80, then 80% of the points should fall within the regression line. Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coefficient is an indicator of a better goodness of fit for the observations.

The CoD can be **negative**, although this usually means that your model is a poor fit for your data. It can also become negative if you didn't set an intercept.

4. Explain the Anscombe’s quartet in detail.

Ans : **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.



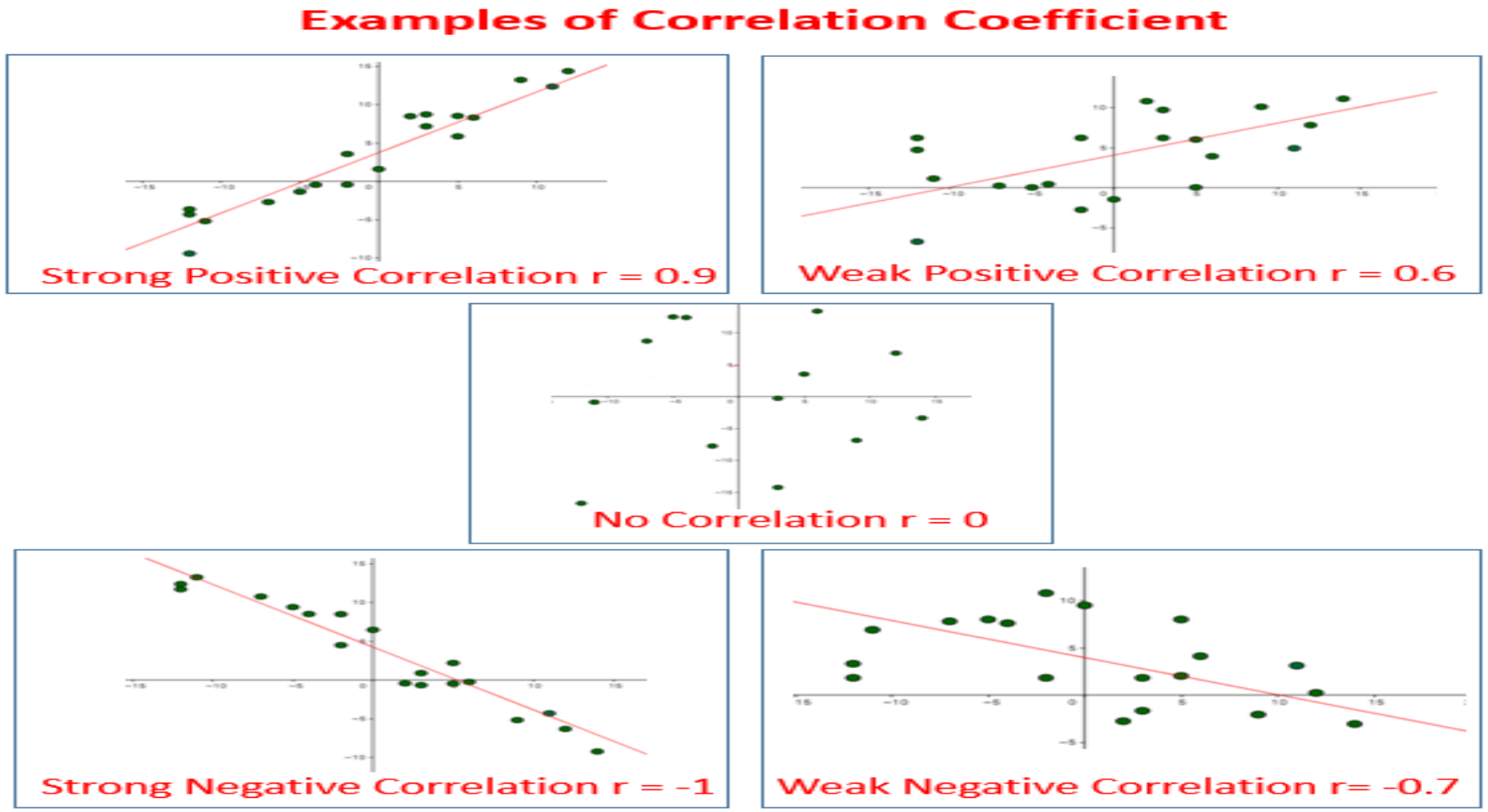
- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as Gaussian with mean linearly dependent on x.
- The second graph (top right) is not distributed normally; while a relationship between the two variables 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 analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

5. What is Pearson’s R?

Ans : The **Pearson** product-moment **correlation** coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as **Pearson's correlation** or simply as the **correlation** coefficient.

Pearson's r can range from -1 to 1.



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

Ans : Feature **scaling** is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally **performed** during the data pre-processing step.

If **feature scaling** is not done, then a **machine learning** algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

It basically helps to normalise **the** data within a particular range. Sometimes, it also helps in speeding up **the** calculations in an algorithm.

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.

$$X_{changed} = (X - X_{min}) / (X_{max} - X_{min})$$

Standardization rescales data to have a mean ($\mu\mu$) of 0 and standard deviation ($\sigma\sigma$) of 1 (unit variance).

$$X_{changed} = (X - \mu) / \sigma$$

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

Ans : If there is perfect correlation, then **VIF** = **infinity**. A large **value** of **VIF indicates** that there is a correlation between the variables. If the **VIF** is 4, this **means** that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity

8. What is the Gauss-Markov theorem?

Ans : In statistics, the **Gauss–Markov theorem** states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the **best linear unbiased estimator** (**BLUE**) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists. Here "best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed (only uncorrelated with mean zero and homoscedastic with finite variance). The requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

9. Explain the gradient descent algorithm in detail.

Ans : **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.

Now there are many types of gradient descent algorithms. They can be classified by two methods mainly:

- **On the basis of data ingestion**
  1. Full Batch Gradient Descent Algorithm
  2. Stochastic Gradient Descent Algorithm

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

- **On the basis of differentiation techniques**
  1. First order Differentiation
  2. Second order Differentiation

Gradient descent requires calculation of gradient by differentiation of cost function. We can either use first order differentiation or second order differentiation.

Vanilla Gradient Descent

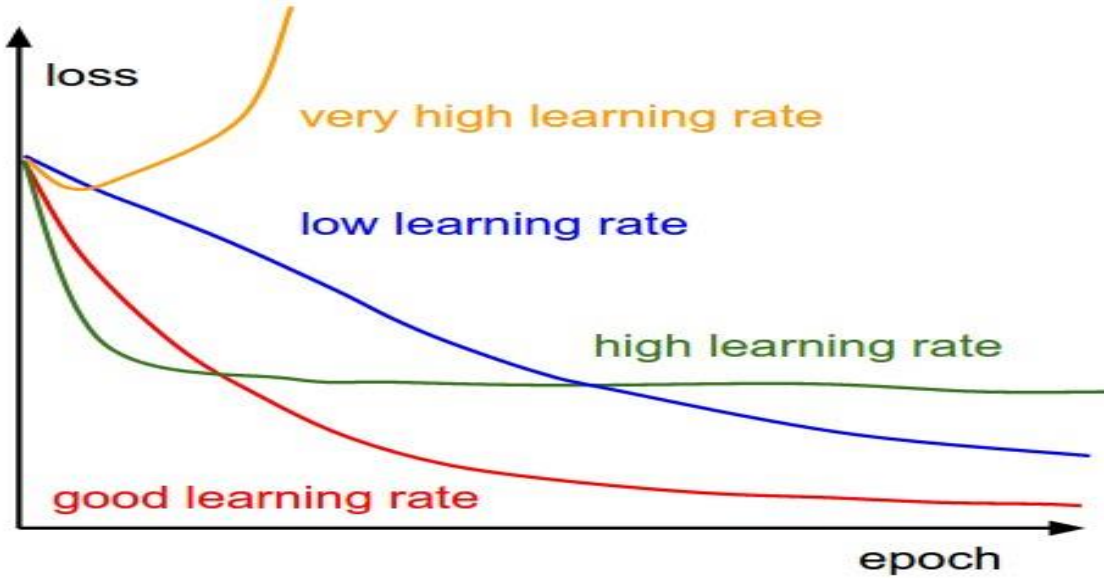
This is the simplest form of gradient descent technique. Here, vanilla means pure / without any adulteration. Its main feature is that we take small steps in the direction of the minima by taking gradient of the cost function.

Let's look at its pseudocode.

update = learning_rate * gradient_of_parameters

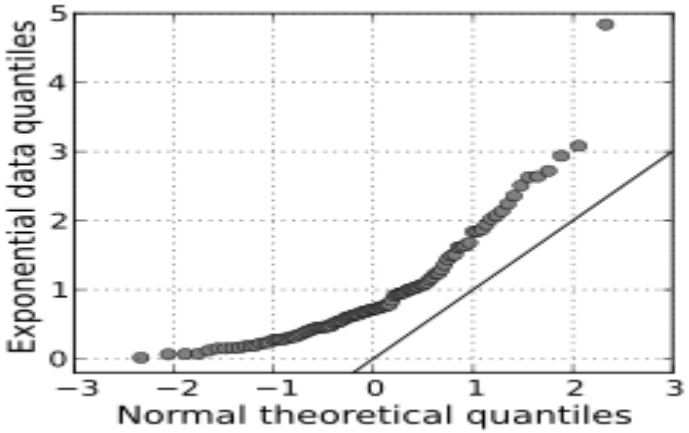
parameters = parameters - update

Here, we see that we make an update to the parameters by taking gradient of the parameters. And multiplying it by a learning rate, which is essentially a constant number suggesting how fast we want to go the minimum. Learning rate is a hyper-parameter and should be treated with care when choosing its value.



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

Ans : **Q-Q Plots** (**Quantile-Quantile plots**) are **plots** of two quantiles against each other. 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.
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 image above shows quantiles from a theoretical normal distribution on the horizontal axis. It's being compared to a set of data on the y-axis. This particular type of Q Q plot is called a **normal quantile-quantile (QQ) plot.** The points are not clustered on the 45 degree line, and in fact follow a curve, suggesting that the sample data is not normally distributed.