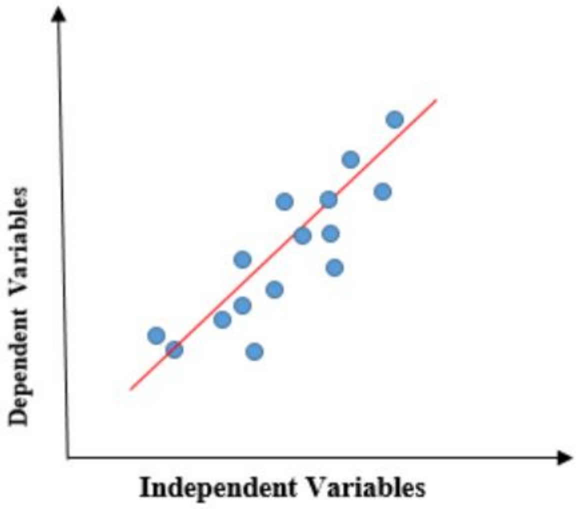
Q1: Explain the linear regression algorithm in detail.

Linear Regression explains the relationship between a dependant variable and independent variable. It's a technique that supports in finding the correlation among variables using a straight line. Linear regression is statistical method used for predictive analysis.

If there is a single input variable than it is considered as **Simple Linear Regression** and if there is a more than one input variables than it is known as **Multiple Linear Regression**.

The linear regression model gives a sloped straight line describing the best fit relationship among the variables.



The above graph shows the linear relationship between independent (x-axis) and dependent (y-axis) variable. We can state it as if the value of x increases than the value of y likewise increases.

Based on the above data points we can calculate the best fit line from the following equation:

$$y = mx + b \rightarrow y = b_0 + b_1x$$

Where,

b_0 = Intercept

b_1 = Slope

A more complex, multiple linear regression equation becomes

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots.$$

Where,

b_0, b_1, b_2, b_3 are the coefficients or weights which our model try to learn and the variables x_1 ,x_2 ,x_3 represent the attributes, or distinct pieces of information, which we have about for each observations.

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

The strength of the linear regression model can be assessed using ***R²*** *and* ***RSE***

R² or **Coefficient of Determination** is a measure that assesses the ability of a model to predict or explain an outcome in the linear regression setting. More specifically, R2 indicates the proportion of the variance in the dependent variable (Y) that is predicted or explained by linear regression and the predictor variable (X, also known as the independent variable).

In general, a high R2 value indicates that the model is a good fit for the data, although interpretations of fit depend on the context of analysis.

RSE or **Residual Standard Error** is a way to measure the standard deviation of the residuals in a regression model. The lower the value for RSE, the more closely a model is able to fit the data (but be careful of overfitting). This can be a useful metric to use when comparing two or more models to determine which model best fits the data.

$$R^2 = 1 - (RSS / TSS)$$

Where,

RSS = Residual sum of squares

TSS = Sum of error of the data from mean

Linear regression is a vast topic which provides the relationship among variables through the best fit line and according to this fit line predict the next event.

Q2: Explain the Anscombe’s quartet in detail.

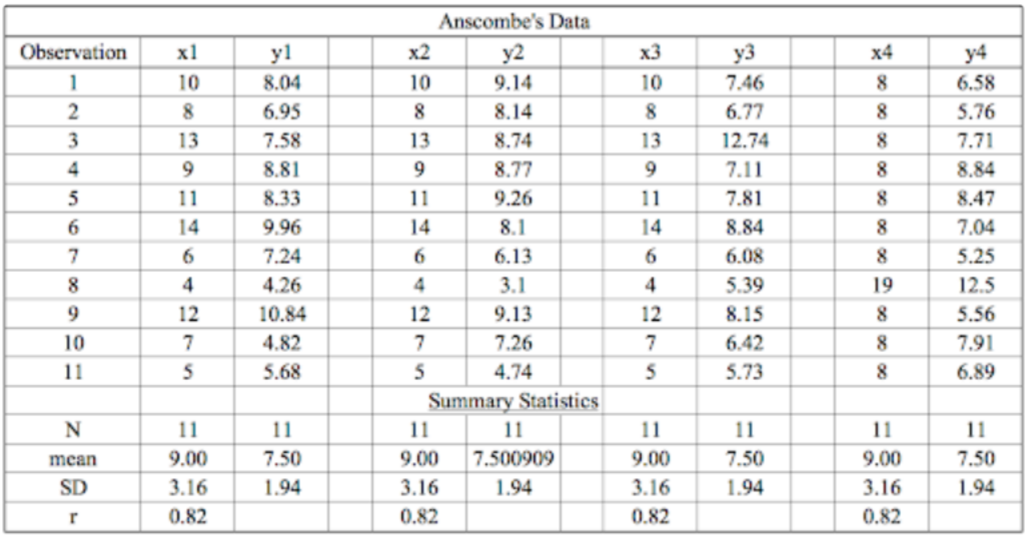
Anscombe’s quartet is a group of four data sets that are nearly identical in simple descriptive statistics, but yet appear very different when graphed.

Perhaps the most elegant demonstration of the dangers of summary statistics is Anscombe’s Quartet. This group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when plotted on scatter plots. Each dataset consists of eleven (x,y) pairs.

Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business. But there’s a danger in relying only on summary statistics and ignoring the overall distribution Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models.

Anscombe’s quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

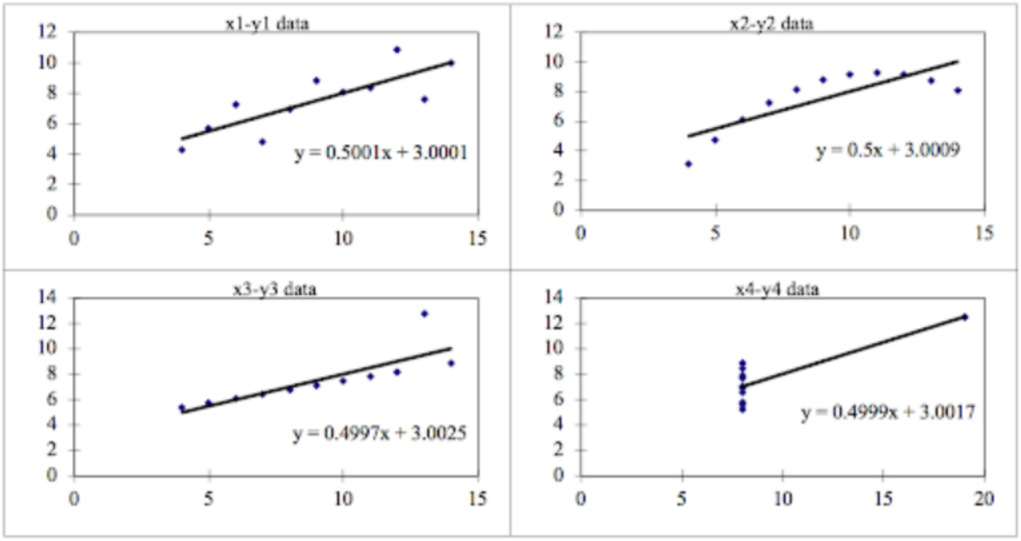
The statistical information for these four data sets are approximately similar. We can compute them as follows:



Anscombe's Data

Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



We can describe the four data sets as:

Data Set 1: fits the linear regression model pretty well

Data Set 2: cannot fit the linear regression model because the data is non-linear

Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model

Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model

As shown above, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set in order to help build a well-fit model.

Q3: What is Pearson’s R?

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.

It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

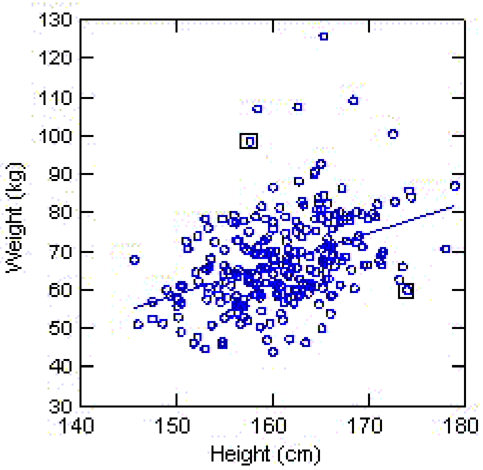
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

"Tends to" means the association holds "on average", not for any arbitrary pair of observations, as the following scatterplot of weight against height for a sample of older women shows.



The correlation coefficient is positive and height and weight tend to go up and down together. Yet, it is easy to find pairs of people where the taller individual weighs less, as the points in the two boxes illustrate.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. It also helps in speeding up the calculations in an algorithm.

Why Scaling

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

For **Example,** if an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions.

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1. Min-Max can be calculated as:

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

One disadvantage of **normalization** over **standardization** is that it loses some information in the data, especially about outliers.

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

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).

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

Usage:

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

Advantages of Q-Q plot

- Since Q-Q plot is like probability plot. So, while comparing two datasets the sample size need not to be equal.
- Since we need to normalize the dataset, so we don’t need to care about the dimensions of values.