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

- **Linear Regression** is a supervised machine learning algorithm. It performs the task to predict a dependent value (y) based on the given independent variables (x1, x2, … xn). The factor that is being predicted (the factor that the equation solves for) is called the dependent variable. The factors that are used to predict the value of the dependent variable are called the independent variables.

- **Hypothesis function** for Linear Regression:
    - y = β0 + β1*x1 + β2*x2 + ...
      where β0 = intercept,
      and β1, β2 … are coefficients of x also known as slope

- **In single linear regression**, there is only one independent variable and the hypothesis equation becomes as follows
    - y = β0 + β1*x1
      where β0 = intercept,
      and β1 the slope of the line

- **Multiple linear regression** is an extension to single linear regression, there are multiple independent variables and the model that fits the data is '**hyperplane**' instead of '**line**'.

- **Assumptions of Linear Regression**
    - Linear relation between errors and Y
    - Error terms are normally distributed.
    - Error terms are independent of each other.
    - Error terms have constant variance i.e. homoscedasticity.

- The **best fit line** is found by minimising 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 acual value of dependent variable.

- The **strength** of Linear Regression model can be assessed using R2 (R-square).
    - R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1.
    - In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model
    - Overall, the higher the R-squared, the better the model fits your data.
    - Mathematically, it is represented as: R2 = 1 - (RSS/TSS)

---

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

- **Normal distribution of residuals -** The error terms are normally distributed**.** The normal distribution of the residual terms is a very crucial assumption. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.

- **Linearity of residual -** The error terms are linear with y-values. It can be checked with scatter plot when Y values are taken on y axis and standardized residuals are taken on the x axis. If the scatter plot follows a linear pattern then this assumption is met.

- **Independence -** The error terms are independent of each other. This may create problem when we have longitudinal dataset (the observations from the same entity over time). In cross sectional datasets, independence assumption is assumed to be met. 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.

- **Equality of variance -** The errors terms should have constant variance i.e. homoscedasticity. Generally, non-constant variance arises in presence of outliers.

---

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

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. It has a range from -1 to 1 where:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
- A result of zero indicates no relationship at all.

Coefficient of determination (denoted by R2) is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination is the square of the correlation between predicted y scores and actual y scores; thus it ranges from 0 to 1.
- An R2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R2 of 1 means the dependent variable can be predicted without error from the independent variable.

- An R2 between 0 and 1 indicates the extent to which the dependent variable is predictable. For eg - An R2 of 0.50 means that the 50% of the variance in Y is predictable from X.

---

4. Explain the Anscombe's quartet in detail.

- **Anscombe's quartet** comprises 4 datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed.
- It was constructed by statistician Francis Anscombe to demonstrate both **the importance of graphing data before analyzing it** and the **effect of outliers** and other influential observations on statistical properties.

- Each dataset consists of eleven (x,y) pairs as follows:

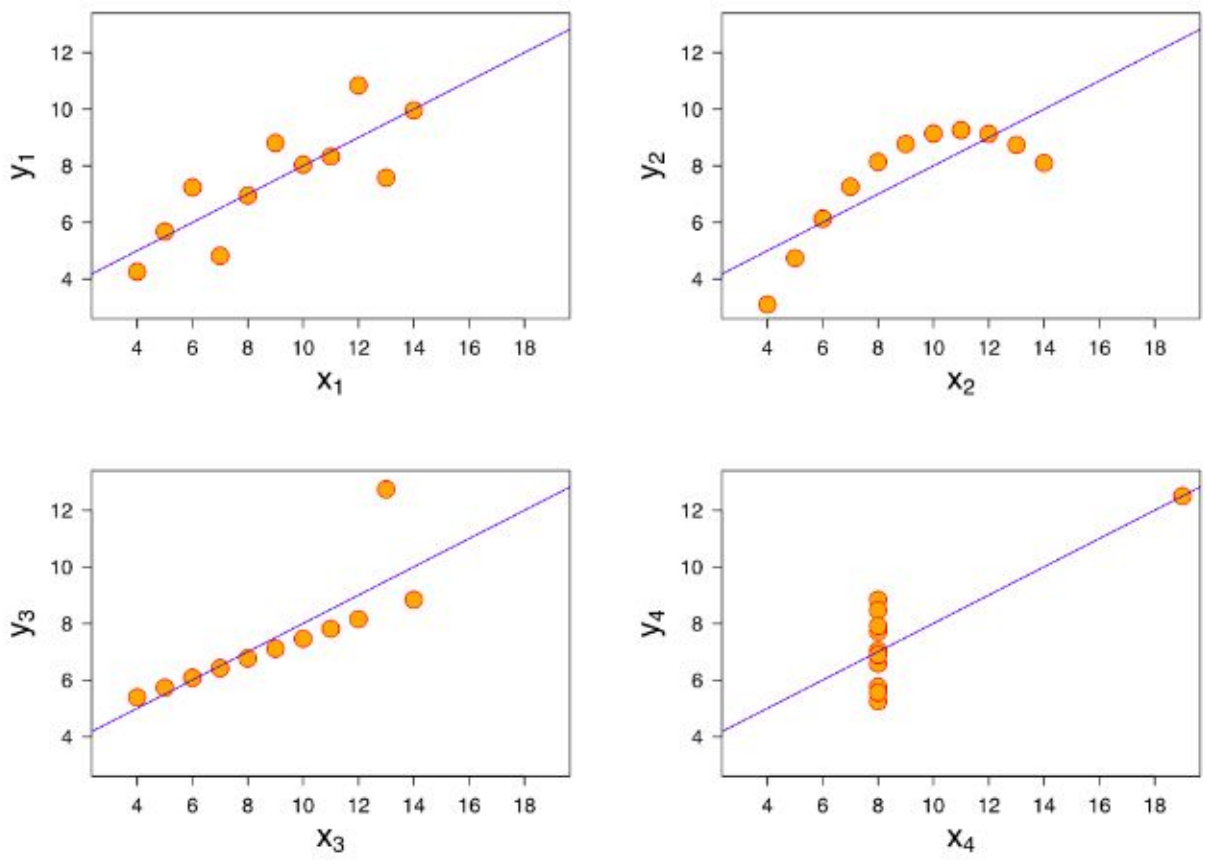
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

All the summary statistics are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12

- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

So far these four datasets appear to be pretty similar. But when plotted, we get the following results:



* He demonstrated that it's important to visualize the data to get a clear picture of what's going on.

---

5. What is Pearson's R ?

- A Pearson correlation is a number between -1 and 1 that indicates the extent to which two variables are linearly related.
- It is also known as the 'product moment correlation coefficient' (PMCC) and are suitable for metric variables only.
- Requirements for Pearson's correlation coefficient
    - Scale of measurement should be interval or ratio
    - Variables should be approximately normally distributed
    - The association should be linear
    - There should be no outliers in the data

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

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
- A result of zero indicates no relationship at all.

---

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

What is Scaling?

Feature Scaling is a method used to normalize the range of independent variables or features of data.

Why Scaling?

- Sometimes, our dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computation, this is a problem.
- If these algorithms only take in the magnitude of features neglecting the units, the result would vary greatly between different units.
- 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.

Normalized Scaling

Normalization rescales the values into a range of [0,1].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Normalizing the data is sensitive to outliers.

Standardized Scaling

Standardized Scaling replaces the values by their Z scores.

$$x' = \frac{x - \bar{x}}{\sigma}$$

This redistributes the features with their mean **μ = 0** and standard deviation **σ =1.**

---

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

- The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model.

- The extent to which a predictor is correlated with the other predictor variables in a linear regression can be quantified as the R-squared statistic of the regression where the predictor of interest is predicted by all the other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

- The VIF is **infinite,** it means that it has R-squared statistics as 1 i.e. this variable/feature can be completely explained by all the other predictor variables. It is highly multicollinear.

- The higher the VIF, the more the standard error is inflated, and the larger the confidence interval and the smaller the chance that a coefficient is determined to be statistically significant.

---

8. What is the Gauss-Markov theorem?

- 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 Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

---

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.

- The procedure starts off with initial values for the coefficients(thetas) for the function. These could be 0 or small random value.

- The cost of the coefficients is evaluated by plugging them into the function and calculating the cost. J(θ)

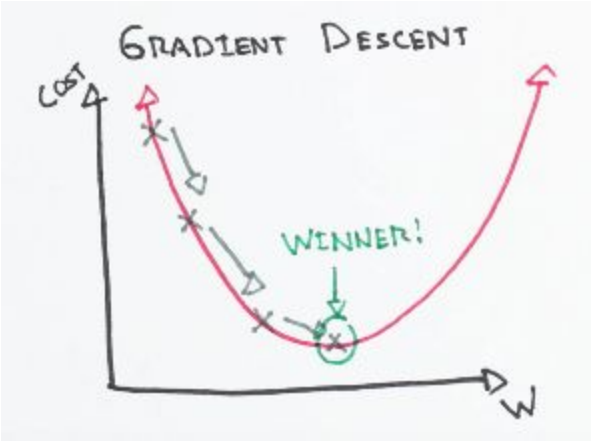
- The derivative of the cost is calculated so as to know the slope so that we know the direction to move the coefficient values in order to get a lower cost on the next iteration.

- The coefficient values can be updated using the formula

  $$\theta^1 = \theta^0 - \eta \frac{\partial}{\partial \theta} J(\theta)$$

  Where η is known as the **learning rate**, which defines the speed at which we want to move towards negative of the gradient.
  And J(θ) is the cost function and theta are the coefficients.

- The process of calculating coefficients using the above formula is repeated until the cost of the coefficients (cost) is 0 or close enough to zero to be good enough.

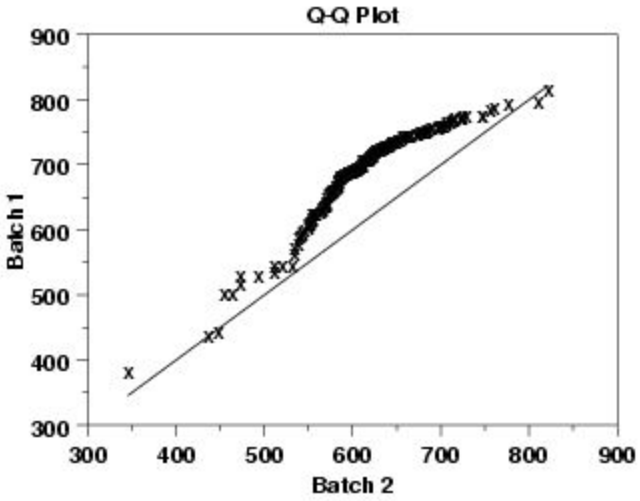


10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

- 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.

- 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:
    - The sample sizes do not need to be equal.
    - 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.



This q-q plot above shows that

- These 2 batches do not appear to have come from populations with a common distribution.
- The batch 1 values are significantly higher than the corresponding batch 2 values.

- In **Linear Regression**, this plot shows if residuals are normally distributed. It is used for checking whether residuals follow a straight line well or they deviate severely. It's good if residuals are lined well on the straight dashed line.