ASSIGNMENT

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

I. Linear regression is basically a commonly used type of predictive analysis. Two things examine in a linear regression.
   I. Does a set of predictor variables do a good job in predicting an outcome(dependent) variable.
   II. Which variables are significant predictors of the outcome variables, and in what way do they indicated by the sign of beta estimates- impact the outcome variables?

These regression estimates are used to explain the relationship between one dependent and one or more independent variable.

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

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

The assumptions of linear regression are:

1. 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'.
2. Assumptions about the residuals:
   1. Normality assumption: It is assumed that the error terms, $\varepsilon^{(i)}$, are normally distributed.

2. 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.
3. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ^2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

3. Assumptions about the estimators:
    1. The independent variables are measured without error.
    2. The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.

Explanations:

1. This is self-explanatory.
2. 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.
   $Y^{(i)i} = \beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)}$
   This is the assumed linear model, where ε is the residual term.
   $E(Y) = E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$
   $= E(\beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)})$
   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 the error terms) should be independent, meaning there is no correlation between the residuals and the predicted values, or among the residuals. Any correlation implies that there is some relation that the regression model is not able to identify.

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

4. **What is the coefficient of correlation and the coefficient of determination?**

Correlation Coefficient : A correlation coefficient is a statistical measure of the degree to which changes to the value of one variable predict change to the value of another. In positively correlated variables, the value increases or decreases in tandem. In negatively correlated variables, the value of one increase as the value of the other decreases.

Correlation coefficients are expressed as values between +1 and -1. A coefficient of +1 indicates a perfect positive correlation: A change in the value of one variable will predict a change in the same direction in the second variable. A coefficient of -1 indicates a perfect negative correlation: A change in the value of one variable predicts a change in the opposite direction in the second variable. Lesser degrees of correlation are expressed as non-zero decimals. A coefficient of zero indicates there is no discernible relationship between fluctuations of the variables.

Determination Coefficient : The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor. It is relied on heavily in trend analysis and is represented as a value between 0 and 1.

The closer the value is to 1, the better the fit, or relationship, between the two factors. The coefficient of determination is the square of the correlation coefficient, also known as "R," which allows it to display the degree of linear correlation between two variables.

This correlation is known as the "goodness of fit." A value of 1.0 indicates a perfect fit, and it is thus a very reliable model for future forecasts, indicating that the model explains all the variations observed. A value of 0, on the other hand, would indicate that the model fails to accurately model the data at all. For a model with several variables, such as a multiple regression model, the adjusted R^2 is a better coefficient of determination. In economics, an R^2 value above 0.60 is seen as worthwhile.

5. Explain the Anscombe's quartet in detail?

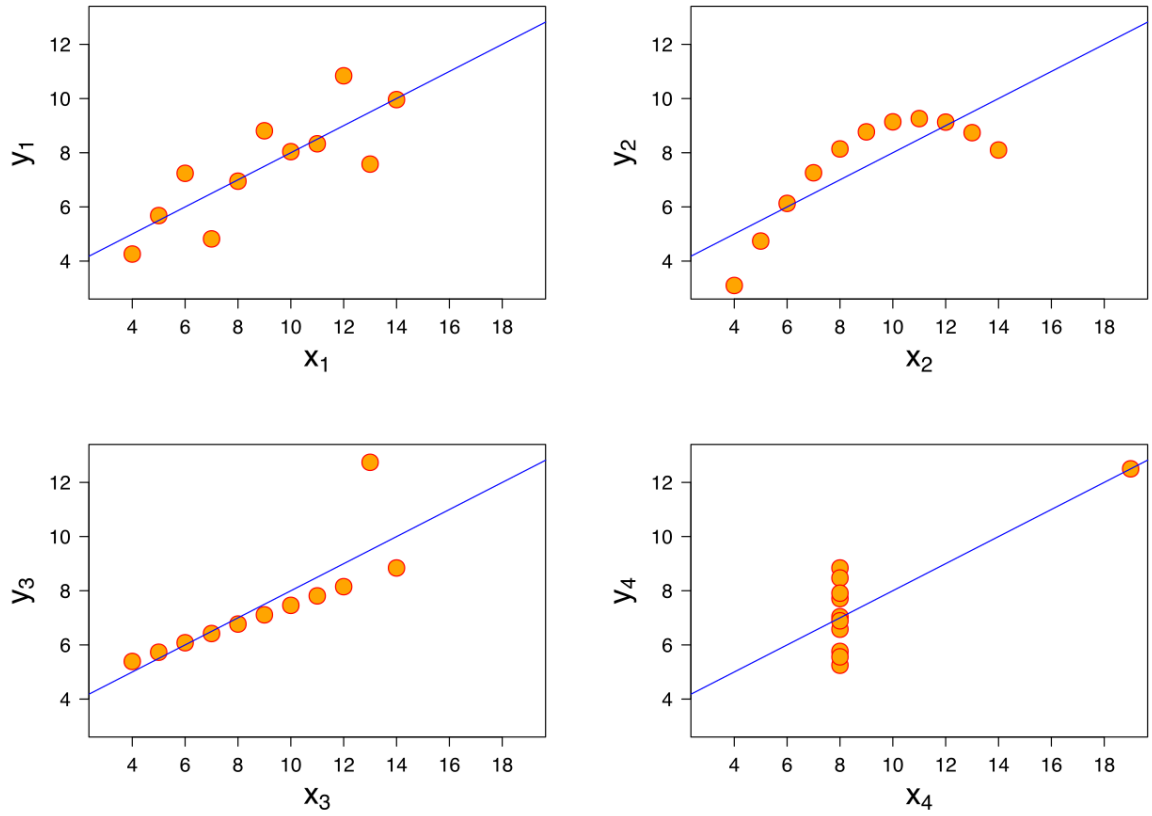
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups :

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story :



- Dataset I appear 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.

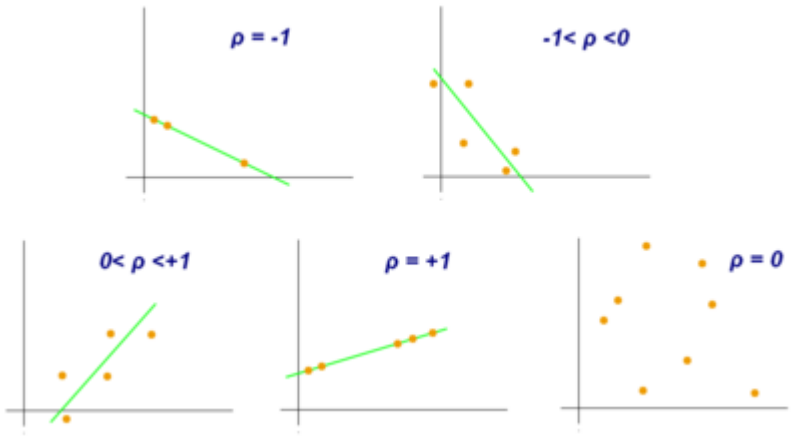
6. What is Pearson's R?

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by **r**.

Correlation 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.

A Pearson's correlation is used when you want to find a linear relationship between two variables. It can be used in a causal as well as a associative research hypothesis but it can't be used with a attributive RH because it is univariate.



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

- 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.
- Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data.
-
- For example, suppose that we have the students' weight data, and the students' weights span [160 pounds, 200 pounds]. To rescale this data, we first subtract 160 from each student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

Standardization

Standardization (or **Z-score normalization**) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with

μ=0 and σ=1

where μ is the mean and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows:

$$z = \frac{x - \mu}{\sigma}$$

$$z = \frac{x - \mu}{\sigma}$$

Normalization

Normalization often also simply called **Min-Max scaling** basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values). It works better for cases in which the standardization might not work so well. If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better.

8. **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. 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.

9. 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 requirement that the estimator be unbiased cannot be dropped, since biased estimators exist with lower variance.

10. Explain the gradient descent algorithm in detail?

- Gradient descent is an iterative optimization to find the minimum of a function.
- Gradient Descent Algorithm in Linear Regression:

Cost Function

$$J(\Theta_0, \Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y_i]^2$$

$h_\Theta(x_i)$: Predicted Value; y_i: True Value

Gradient Descent

$$\Theta_j = \Theta_j - \alpha\frac{\partial}{\partial\Theta_j}J(\Theta_0, \Theta_1)$$

α: Learning Rate

Now,

$$\frac{\partial}{\partial\Theta}J_\Theta = \frac{\partial}{\partial\Theta}\frac{1}{2m}\sum_{i=1}^{m}[h_\Theta(x_i) - y]^2$$

$$= \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x_i) - y)\frac{\partial}{\partial\Theta_j}(\Theta x_i - y)$$

$$= \frac{1}{m}(h_\Theta(x_i) - y)x_i$$

Therefore,

$$\Theta_j := \Theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\Theta(x_i) - y)x_i]$$

θ_j : Weights of the hypothesis.
-> $h_\theta(x_i)$: predicted y value for i^{th} input.
-> **j** : Feature index number (can be 0, 1, 2, ..., n).
-> **α** : Learning Rate of Gradient Descent.

To understand in a simpler way, let’s us take the example Suppose you are at the top of a mountain, and you must reach a lake which is at the lowest point of the mountain. A twist is that you are blindfolded, and you have zero visibility to see where you are headed. So, what approach will you take to reach the lake?
The best way is to check the ground near you and observe where the land tends to descend. This will give an idea in what direction you should take your first step. If you follow the descending path, it is very likely you would reach the lake.

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

- **Q-Q plot** is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.
- The quantile-quantile or **q-q** plot is an exploratory graphical device used to check the validity of a distributional assumption for a data set. In general, the basic idea is to compute the theoretically expected value for each data point based on the distribution in question.
- Purpose: Check If Two Data Sets Can Be Fit with the Same Distribution. The quantile-quantile (**q-q**) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.