Linear Regression Subjective Questions

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

Answer.

Linear Regression is a Machine Learning Algorithm based on Supervision Learning. In Supervision Learning, Supervised Learning is a task of inferring a function from labeled training data.

Linear Regression is comes from regress, which means to return to a less developed state. In Linear Regression, we try to find/ predict the value of a variable (independent) variables by regressing one or more variables(dependent).

The word linear means that we try to find the relationship between independent and dependent variables as a Straight Line.

Equation of St Line – Y – Mx + C

Now, to further analyze, let's break the model into Simple Linear Regression and Multiple Linear Regression Based on number of dependent variables.

Simple Linear Regression Model:

The model is given by equation: $Y = B0 + B1X + E$

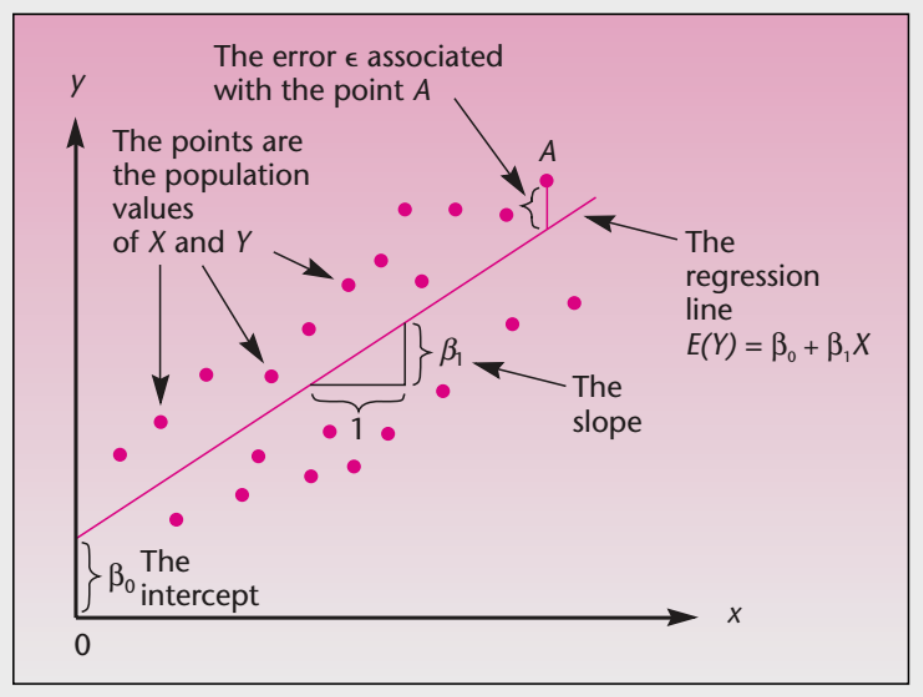
Here, B0 is The intercept on the Y axis.

B1 – Coefficient of X.

E – Error Terms

B0 is the Y intercept of the straight line given by $Y = B0 + B1X$ (the line does not contain the error term).
B1 is the slope of the line $Y = B0 + B1X$

Below figure explains the Simple Linear Model:



We notice that that the Line consists of two parts – Dependent and independent variables. These constitute the nonrandom term of the equation. However, the error component is a Random component.

We determine the line using Method of Least Squares.

The method of least squares gives us the best linear unbiased estimators (BLUE) of the regression parameters B0 and B1. These estimators both are unbiased and have the lowest variance of all possible unbiased estimators of the regression parameters. These properties of the least-squares estimators are specified by a well-known theorem, the Gauss-Markov theorem. We denote the least-squares estimators by b0 and b1.

With the help of Method of Least Squares, we are able to determine a regression line with minimum errors for the Simple Linear Regression Model.

Assumptions: For the Above model we have below assumptions:

- The relationship between X and Y is a straight-line relationship.
- The values of the independent variable X are assumed fixed (not random); the only randomness in the values of Y comes from the error term e.
- The errors e is normally distributed with mean 0 and a constant variance sigma2. The errors are uncorrelated (not related) with one another in successive observations.

Multiple Linear Regression Model:

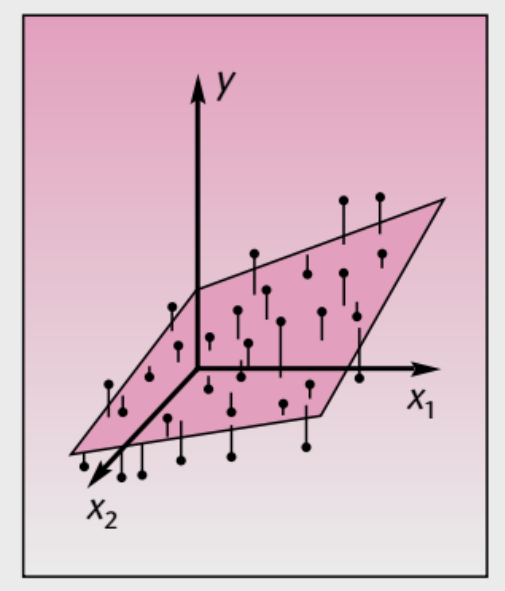
The multiple Linear Regression Model considers effect of many dependent variables X on Y.

The population regression model of a dependent variable Y on a set of k independent variables X1, X2, . . . , Xk is given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Where B0 is length of the intercept. B1….Bk are the intercepts.

Here, instead of a Regression line, we have a regression plane. (fig Below)



The Assumptions of the model are:

- For each observation, the error term e is normally distributed with mean zero and standard deviation and is independent of the error terms associated with all other observations.
- In the context of regression analysis, the variables Xj are considered fixed quantities, although in the context of correlational analysis, they are random variables. In any case, Xj are independent of the error term e. When we assume that Xj are fixed quantities, we are assuming that we have realizations of k variables Xj and that the only randomness in Y comes from the error term e.

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

Answer:

The below assumptions are made regarding the residuals in Linear Regression:

- The Mean of residuals is 0.

We Assume that the residual has a mean of 0. i.e they are equally spread around the regression line.

- Homoscedasticity of residuals or equal variance.

We Assume that the residual terms are equal variance from the regression line. The variance does not increase of decrease along the line.

- No autocorrelation of residuals.

This is applicable especially for time series data. Autocorrelation is the correlation of a time Series with lags of itself. When the residuals are autocorrelated, it means that the current value is dependent of the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.

- The X variables and residuals are uncorrelated

The Residuals are believed to be perfectly random. They should not be correlated with the dependent variables. If they are, it would mean a flaw in our model.

- Normality of residuals

The residuals should be normally distributed. If the maximum likelihood method (not OLS) is used to compute the estimates, this also implies the Y and the Xs are also normally distributed.

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

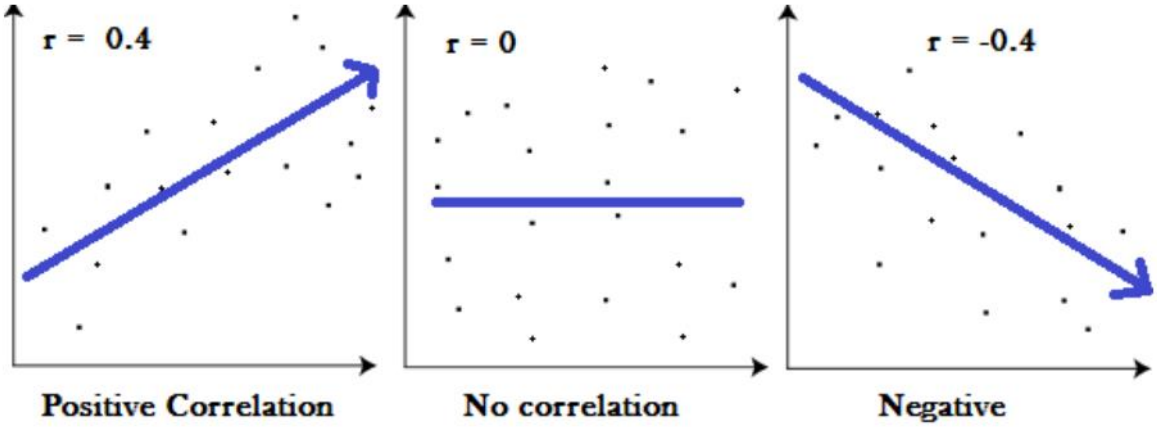
Answer:

The correlation coefficient (R) is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the correlation measurement. A correlation of -1.0 shows a perfect negative correlation, while a correlation of 1.0 shows a perfect positive correlation. A correlation of 0.0 shows no relationship between the movement of the two variables.

Coefficient of Relation is given as:

$$\rho_{xy} = \frac{\mathrm{Cov}(r_x, r_y)}{\sigma_x \sigma_y}$$

Below graph shows the correlation coefficient:



The coefficient of determination (R2) is a measure used in statistical analysis that assesses how well a model explains and predicts future outcomes. It tells us the level of variance that is explained by the statistical model. The value lie between 0 and 1 where 1 means every value is explained in the statistical model.

One way of interpreting this figure is to say that the variables included in a given model explain approximately x% of the observed variation. So, if the R2 = 0.50, then approximately half of the observed variation can be explained by the model.

However, R-squared is unable to determine whether the data points or predictions are biased. It also doesn't tell the analyst or user whether the coefficient of determination value is good or not. A low R-squared is not bad, for example, and it's up to the person to make a decision based on the R-squared number.

The coefficient of determination should not be interpreted naively. For example, if a model's R-squared is reported at 75%, the variance of its errors is 75% less than the variance of the dependent variable, and the standard deviation of its errors is 50% less than the standard deviation of the dependent variable. The standard deviation of the model's errors is about one-third the size of the standard deviation of the errors that you would get with a constant-only model.

Finally, even if an R-squared value is large, there may be no statistical significance of the explanatory variables in a model, or the effective size of these variables may be very small in practical terms.

4. Explain the Anscombe's quartet in detail.

Answer:

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, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

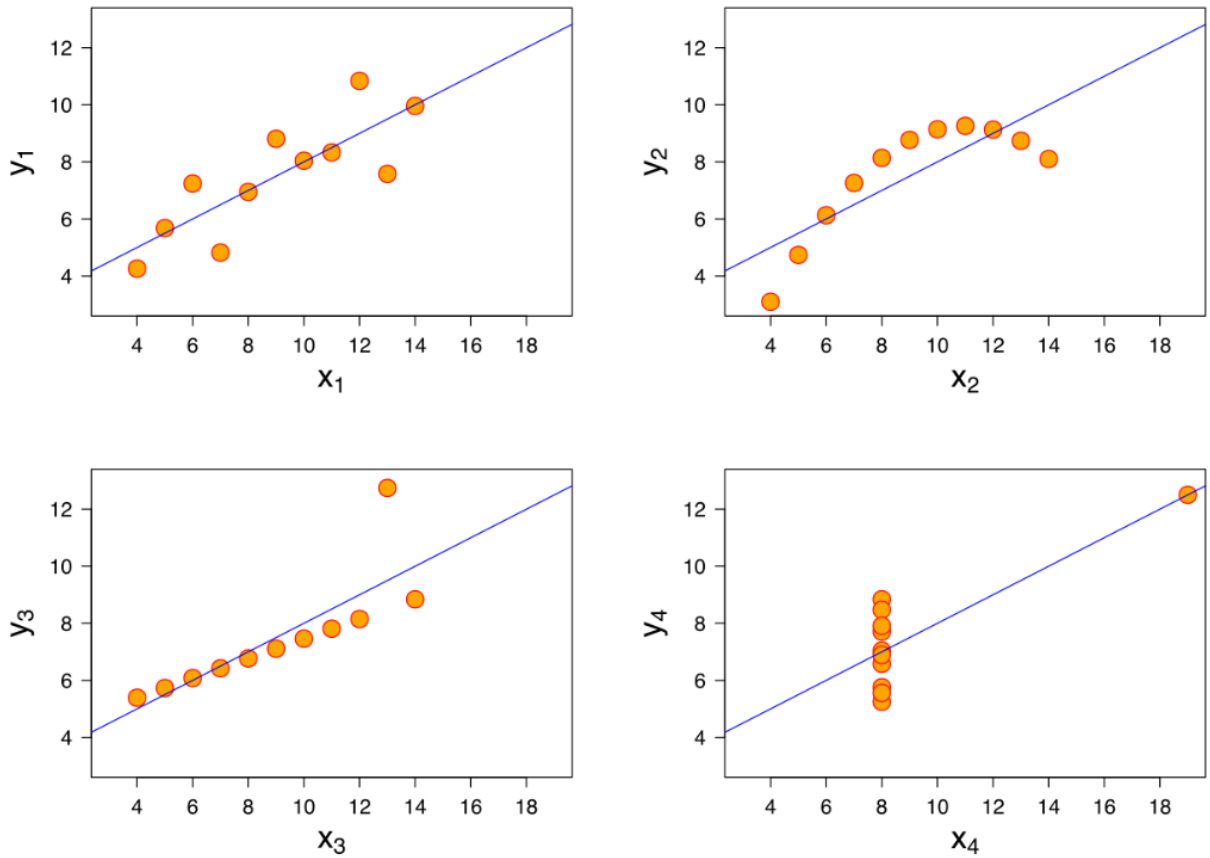
Let's have a look at below example:

	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

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- 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

Hence, we notice that the descriptive statistics is similar for all four datasets.

Now, let's plot the fraps for the above datasets:



Here we notice that all four datasets have **same line of regression**, however they are very different.

- 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.

Inference:

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?

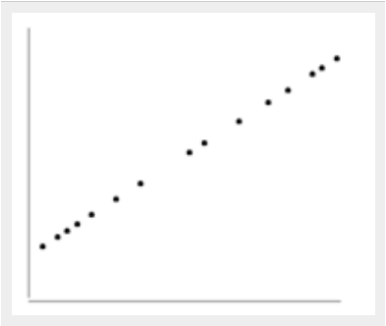
Answer:

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.

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

R = 1: Data lie on a perfect straight line with a negative slope

R = 0: No linear relationship between the variables



R= +1: Data lie on a perfect straight line with a positive slope

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

Answer:

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

We perform Scaling, when features are of different ranges. For example, consider a data set containing two features, age(x1), and income(x2). Where age ranges from 0–100, while income ranges from 0–20,000 and higher. Income is about 1,000 times larger than age and ranges from 20,000–500,000. So, these two features are in very different ranges. When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn't necessarily mean it is more important as a predictor.

Hence, it's good to scale the variables as it will help to determine a better variable.

Normalized Scaling vs Standardized Scaling:

In normalized scaling, the features are scaled between 0 and 1, while in standardized scaling they are scaled around 0.

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

Answer:

Variance inflation factors (VIF) show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

For any predictor orthogonal (independent) to all other predictors, the variance inflation factor is 1.0. VIFi thus provides us with a measure of how many times larger the variance of the ith regression coefficient will be for multicollinear data than for orthogonal data (where each VIF is 1.0). If the VIF's are not unusually larger than 1.0, multicollinearity is not a problem. An advantage of knowing the VIF for each variable is that it gives a tangible idea of how much of the variances of the estimated coefficients are degraded by the multicollinearity.

In case of infinite VIF, it means there is perfect multicollinearity between two variables. In such case, we must address the data for multicollinearity.

8. What is the Gauss-Markov theorem?

Answer:

Consider a Dataset (for OLS model) with below assumptions:

Linearity: The parameters we are estimating using the OLS method must be themselves linear.

Random: Data must be randomly sampled from the population.

Non-Collinearity: The regressors being calculated aren't perfectly correlated with each other.

Exogeneity: The regressors aren't correlated with the error term.

Homoscedasticity: No matter what the values of our regressors might be, the error of the variance is constant.

Gauss-Markov theorem tells us that if above assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

In practice, the Gauss Markov assumptions are rarely all met perfectly, but they are still useful as a benchmark, and because they show us what 'ideal' conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

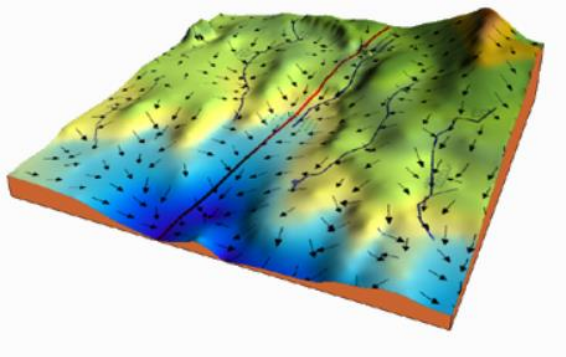
9. Explain the gradient descent algorithm in detail.

Answer:

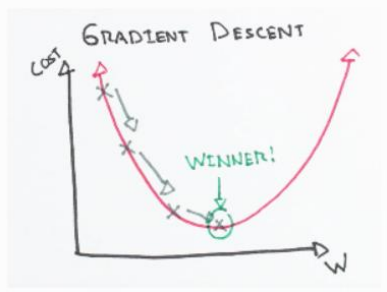
Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

To understand the algorithm, let's understand the below analogy:

Consider the 3-dimensional graph below in the context of a cost function. Our goal is to move from the mountain in the top right corner (high cost) to the dark blue sea in the bottom left (low cost). The arrows represent the direction of steepest descent (negative gradient) from any given point–the direction that decreases the cost function as quickly as possible.



Starting at the top of the mountain, we take our first step downhill in the direction specified by the negative gradient. Next we recalculate the negative gradient (passing in the coordinates of our new point) and take another step in the direction it specifies. We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.



The size of these steps is called the **learning rate**. With a high learning rate we can cover more ground each step, but we risk overshooting the lowest point since the slope of the hill is constantly changing. With a very low learning rate, we can confidently move in the direction of the negative gradient

since we are recalculating it so frequently. A low learning rate is more precise, but calculating the gradient is time-consuming, so it will take us a very long time to get to the bottom.

A **Cost Functions** tells us “how good” our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.

Now let’s run gradient descent using our new cost function. There are two parameters in our cost function we can control: m (weight) and b (bias). Since we need to consider the impact each one has on the final prediction, we need to use partial derivatives. We calculate the partial derivatives of the cost function with respect to each parameter and store the results in a gradient.

The Cost Function:

$$f(m,b) = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Gradient Can be calculated as:

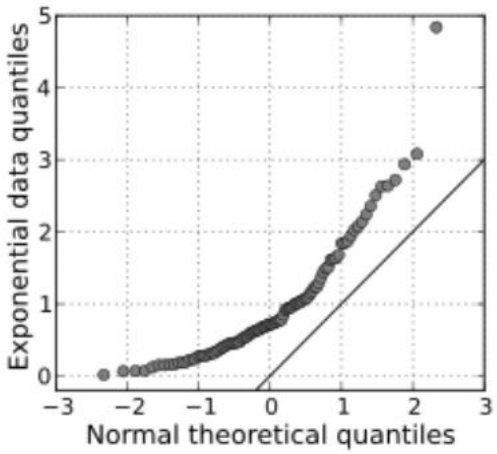
$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

To solve for the gradient, we iterate through our data points using our new m and b values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

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

Answer:

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.



QQ plots are used to fit a Linear Regression Model and check if points lie fairly on the regression line.

The importance is, that this helps us in testing our model, heck if the points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either. This implies that for small sample sizes, you can't assume your estimator B is Gaussian either, so the standard confidence intervals and significance tests are invalid.