1. Explain the linear regression algorithm in detail.

Answer:

Linear Regression is a machine learning algorithm based on the technique of **supervised learning**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the **relationship** between variables and forecasting.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

where $\hat{Y}$ is the predicted or expected value of the dependent variable, X_1 through X_p are p distinct independent or predictor variables, b_0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, and b_1 through b_p are the estimated regression coefficients. Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable. In the multiple regression situation, b_1, for example, is the change in Y relative to a one unit change in X_1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed). Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

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

Answer:

1. Residuals are normally distributed (not X, Y)
2. Residual terms are independent of each other
3. Residuals have constant variance (homoscedasticity)

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

Answer:

coefficient of correlation

It is measure to shoe **how strong a relationship** is between two variables.

- 1 indicates a strong **positive** relationship.
  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. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- -1 indicates a strong **negative** relationship.
  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. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- A result of zero indicates **no relationship** at all.
  Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

coefficient of determination

It 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 (r) between predicted y scores and actual y scores; thus, it ranges from 0 to 1.
- With linear regression, the coefficient of determination is also equal to the square of the correlation between x and y scores.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in *Y* is predictable from *X*; an R^2 of 0.20 means that 20 percent is predictable; and so on.

The formula for computing the coefficient of determination for a linear regression model with one independent variable is given below.

$$R^2 = \{ (1 / N) * \Sigma [(x_i - x) * (y_i - y)] / (\sigma_x * \sigma_y)$$

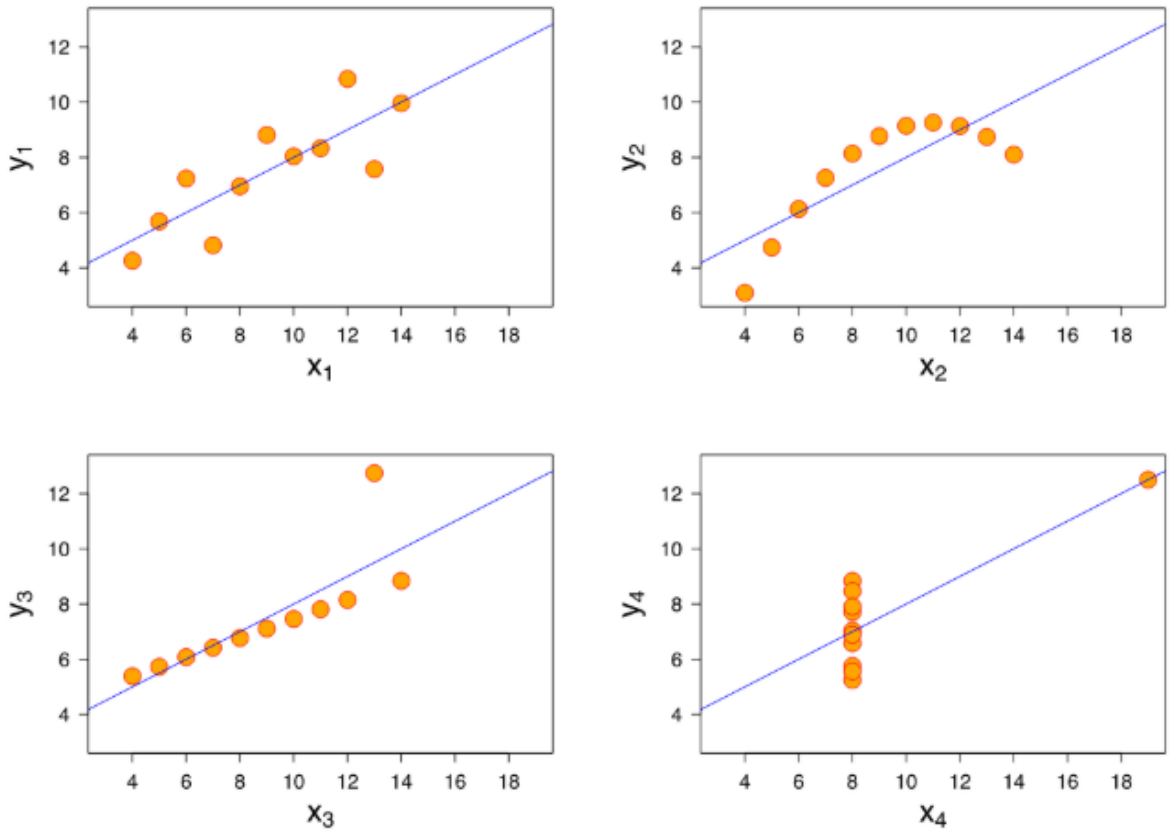
where N is the number of observations used to fit the model, Σ is the summation symbol, x_i is the x value for observation i, x is the mean x value, y_i is the y value for observation i, y is the mean y value, σ_x is the standard deviation of x, and σ_y is the standard deviation of y

4. Explain the Anscombe's quartet in detail.

Answer:

It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed.

Anscombe's quartet is a classic example of the drawback to just reporting correlation. It explains how a set of four different pairs of variables can deliver the same correlation coefficient, while the relationships between each pair are completely different.



- 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.

Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier. Dataset IV looks like *x* remains constant, except for one outlier as well.
Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.Visualizing the data helped in two ways. It gave us a better picture of what realistic starting law salaries look like, and also allowed us to ask a follow-up question that exposed a potential flaw in our data.

5. What is Pearson's R?

Answer:

It is a **statistic** that measures linear **correlation** between two variables *X* and *Y*.

Pearson's Correlation Coefficient is a linear correlation coefficient that returns a value of **between -1 and +1**. A -1 means there is a strong negative correlation and +1 means that there is a strong positive correlation. A 0 means that there is no correlation (this is also called **zero correlation**).

The symbol for Pearson's correlation is "ρ" when it is measured in the *population* and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's correlation coefficient when applied to a sample is commonly represented by r_{xy} and may be referred to as the *sample correlation coefficient* or the *sample Pearson correlation coefficient*. We can obtain a formula for r_{xy} by substituting estimates of the covariances and variances based on a sample data, it Is given as,

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
-

Statistical inference based on Pearson's correlation coefficient often focuses on one of the following two aims:

- One aim is to test the null hypothesis that the true correlation coefficient *ρ* is equal to 0, based on the value of the sample correlation coefficient *r*.
- The other aim is to derive a confidence interval that, on repeated sampling, has a given probability of containing *ρ*

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

Answer:

It is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

- When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0
- On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X' is 1
- If the value of X is between the minimum and the maximum value, then the value of X' is between 0 and 1

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here's the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

- μ is the mean of the feature values and σ is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

Answer:

VIF is infinite, If there is perfect correlation. A large **value of VIF** indicates that there is a correlation between the variables. An **infinite VIF value** happens if the corresponding variable may be expressed exactly by a linear combination of other variables.

8. What is the Gauss-Markov theorem?

Answer:

The **Gauss Markov theorem** tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the *best linear unbiased estimate* possible.

There are five Gauss Markov assumptions (also called *conditions*):

1. **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
2. **Random**: our data must have been randomly sampled from the population.
3. **Non-Collinearity**: the regressors being calculated aren't perfectly correlated with each other.
4. **Exogeneity**: the regressors aren't correlated with the error term.

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

9. Explain the gradient descent algorithm in detail.

Answer:

Gradient Descent is an optimization algorithm which we use to minimize cost function by repeating until convergence.

Cost Function

It tells us how good are model is at making predictions of the dependent variable(price of house) for the given set of parameters(theta1 and theta0). Cost function is basically the sum of squared difference of actual and predicted value. Minimum the value of cost function is better is our model at predicting values. So our goal is to minimize cost function.

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Now we can minimize cost function(get the local optima) with the help of gradient descent.We have two functions here h(x) and J(theta0,theta1), for different values of theta0,theta1 h(x) will have different values and there will be a graph for J(theta0,theta1) plotting different values. We want the global minima of that graph.When we plot for both theta0 and theta1 ,J(theta0,theta1) it forms a 3-D figure.

The evaluation of how close a fit a machine learning model estimates the target function can be calculated a number of different ways, often specific to the machine learning algorithm. The cost function involves evaluating the coefficients in the machine learning model by calculating a prediction for the model for each training instance in the dataset and comparing the predictions to the actual output values and calculating a sum or average error (such as the Sum of Squared Residuals or SSR in the case of linear regression).

From the cost function a derivative can be calculated for each coefficient so that it can be updated using exactly the update equation described above.

The cost is calculated for a machine learning algorithm over the entire training dataset for each iteration of the gradient descent algorithm. One iteration of the algorithm is called one batch and this form of gradient descent is referred to as batch gradient descent.

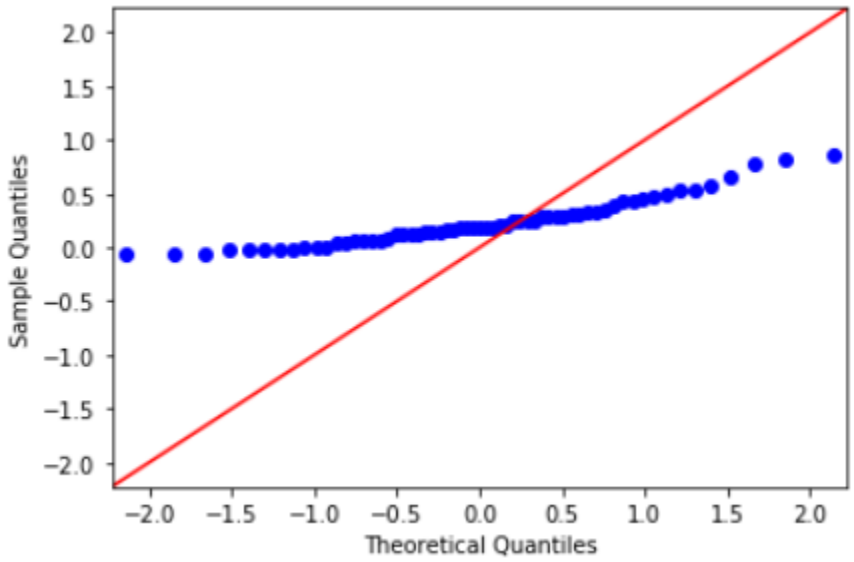
Batch gradient descent is the most common form of gradient descent described in machine learning.

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

Answer:

The “QQ” in QQ plot means quantile-quantile — that is, the QQ plot compares the quantiles of our data against the quantiles of the desired distribution (defaults to the normal distribution, but it can be other distributions too as long as we supply the proper quantiles).

Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets.



The slope tells us whether the steps in our data are too big or too small (or just right)

A steeply sloping section of the QQ plot means that in this part of our data, the observations are more spread out than we would expect them to be if they were normally distributed.

A flat QQ plot means that our data is more bunched together than we would expect from a normal distribution.