Linear Regression Questions Answers

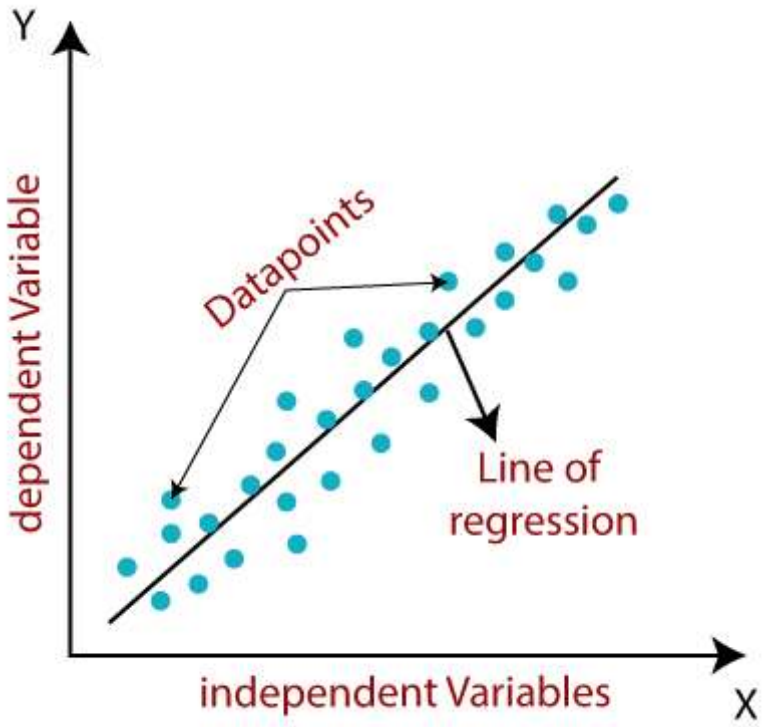
Submitted by : Rajan Kumar

1. What is Linear Regression

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting.

Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

x: input training data (univariate – one input variable(parameter)) y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values. θ1: intercept θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

2. How can you calculate error in Linear Regression?

Ans:

Linear regression most often uses mean-square error (MSE) to calculate the error of the model.

MSE is calculated by:

1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. calculating the mean of each of the squared distances.

Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE.

3. Difference between Loss and Cost function?

Ans:

There is no major difference.

1. When calculating loss we consider only a single data point, then we use the **loss function**.
2. Whereas, when calculating the sum of error for multiple data then we use the **cost function**.

The Most commonly used loss functions are **Mean-squared error** and **Hinge loss**.

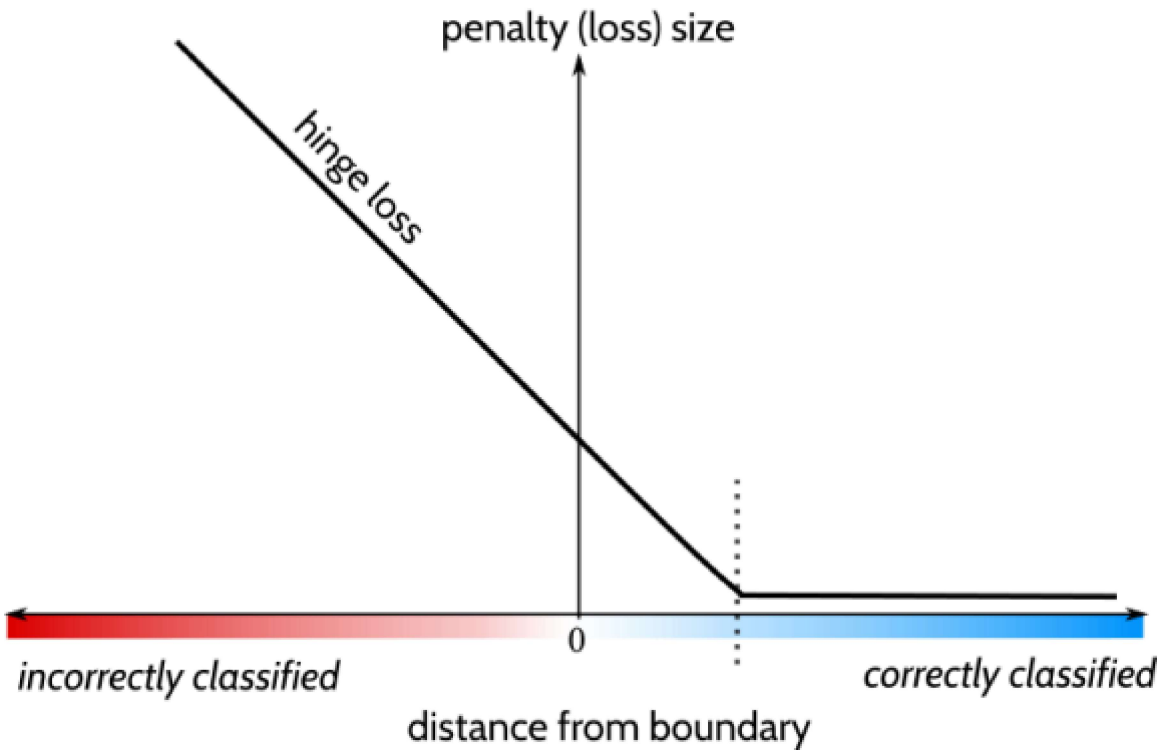
Mean-Squared Error(MSE): we can say how our model predicted values against the actual values.

MSE = √(predicted value - actual value)2

Hinge loss: It is used to train the machine learning classifier, which is

L(y) = max(0,1- yy)

Where y = -1 or 1 indicating two classes and y represents the output form of the classifier. The most common cost function represents the total cost as the sum of the fixed costs and the variable costs in the equation y = mx + b



4. MAE , MSE and RMSE?

Ans:

MAE :

MAE evaluates the absolute distance of the observations (the entries of the dataset) to the predictions on a regression, taking the average over all observations. We use the absolute value of the distances so that negative errors are accounted properly. This is exactly the situation described on the image above.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i^{real} - y_i^{pred}|$$

MSE :

Another way to do so is by squaring the distance, so that the results are positive. This is done by the MSE, and higher errors (or distances) weigh more in the metric than lower ones, due to the nature of the power function.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{real} - y_i^{pred})^2$$

RMSE :

A backlash in MSE is the fact that the unit of the metric is also squared, so if the model tries to predict price in US, $the MSE will yield a number with unit (US)^2$ which does not make sense. RMSE is used then to return the MSE error to the original unit by taking the square root of it, while maintaining the property of penalizing higher errors.

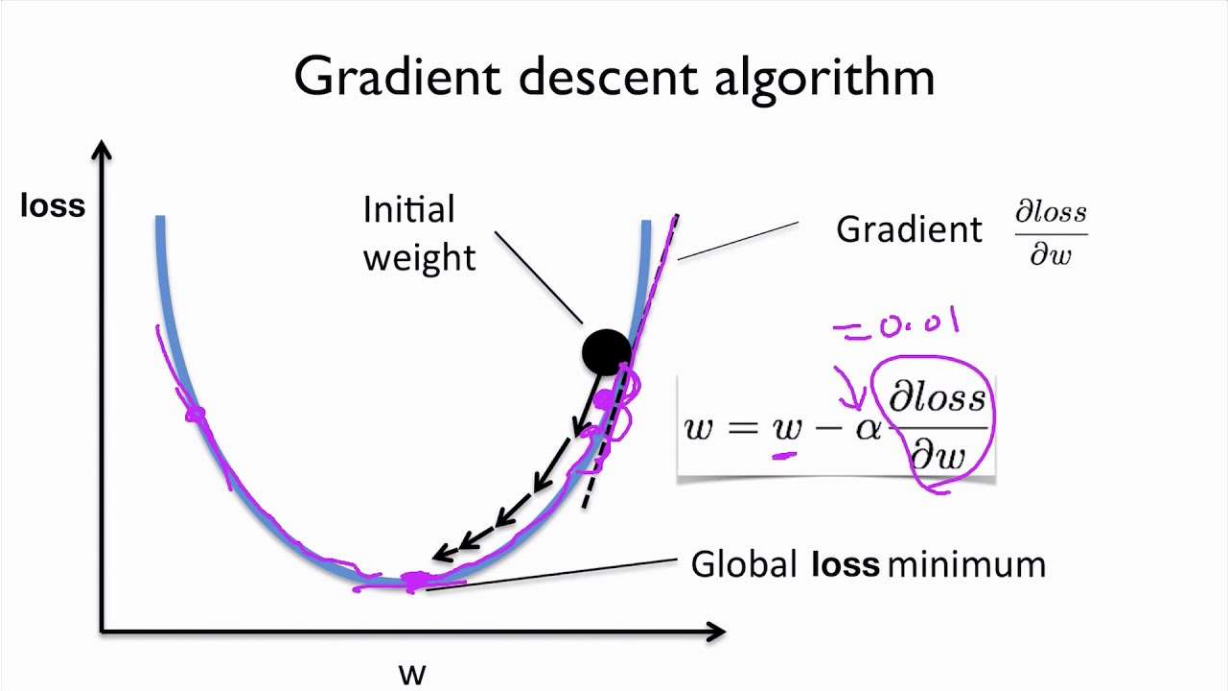
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{real} - y_i^{pred})^2}$$

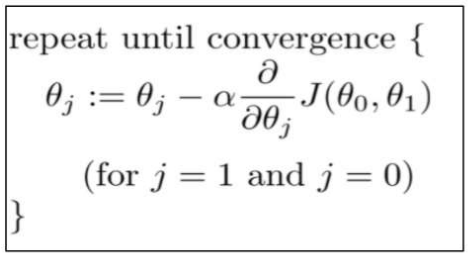
In [ ]:

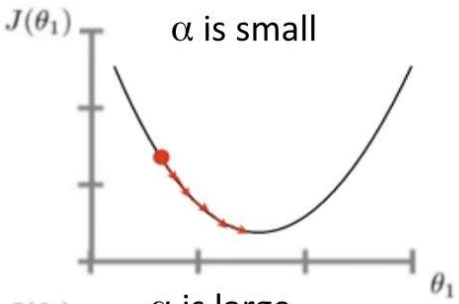
5. Explain how Gradient descent work in Linear Regression ?

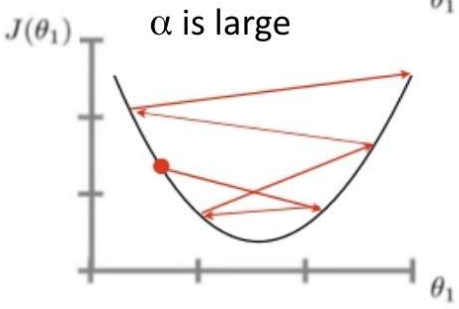
Ans:

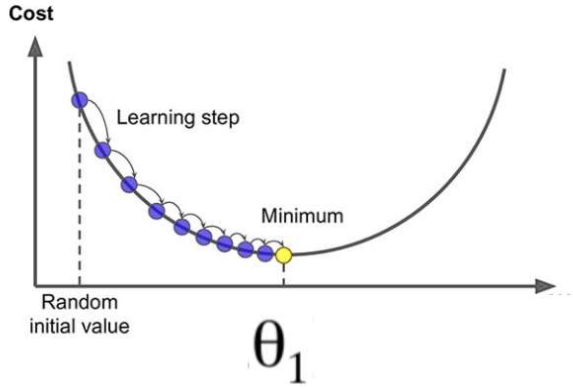
Gradient descent is an algorithm that approaches the least squared regression line via minimizing sum of squared errors through multiple iterations. So far, I've talked about simple linear regression, where you only have 1 independent variable (i.e. one set of x values).











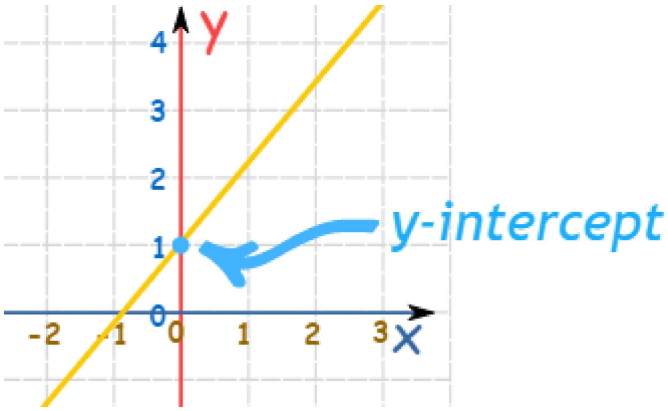
In [3]:
```
def update_weights(m, b, X, Y, learning_rate):
    m_deriv = 0
    b_deriv = 0
    N = len(X)
    for i in range(N):
        # Calculate partial derivatives
        # -2x(y - (mx + b))
        m_deriv += -2*X[i] * (Y[i] - (m*X[i] + b))

        # -2(y - (mx + b))
        b_deriv += -2*(Y[i] - (m*X[i] + b))

    # We subtract because the derivatives point in direction of steepest ascent
    m -= (m_deriv / float(N)) * learning_rate
    b -= (b_deriv / float(N)) * learning_rate

    return m, b
```

In [ ]:
```
```

6. Explain the intercept term means?

Ans:

The intercept (often labeled as constant) is the point where the function crosses the y-axis. In some analysis, the regression model only becomes significant when we remove the intercept, and the regression line reduces to Y = bX + error



7. Write all the assumption for Linear Regression?

Ans:

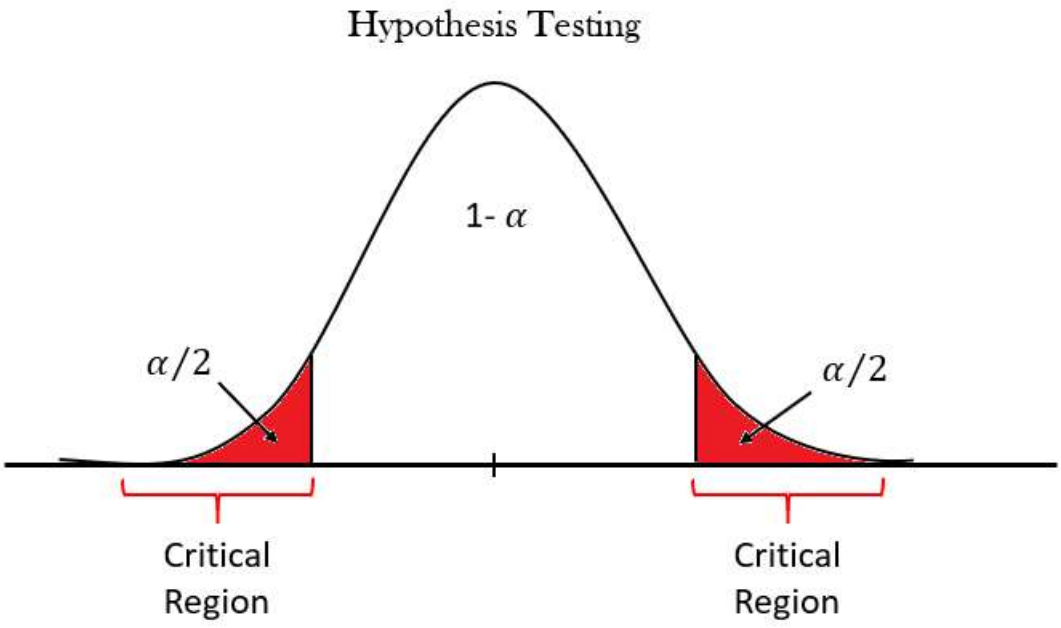
Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

1. Linear relationship
2. Multivariate normality
3. No or little multicollinearity
4. No auto-correlation
5. Homoscedasticity

8. How is Hypothesis testing used in Linear Regression?

Ans:

Hypothesis testing is used to confirm if our beta coefficients are significant in a linear regression model. Every time we run the linear regression model, we test if the line is significant or not by checking if the coefficient is significant.



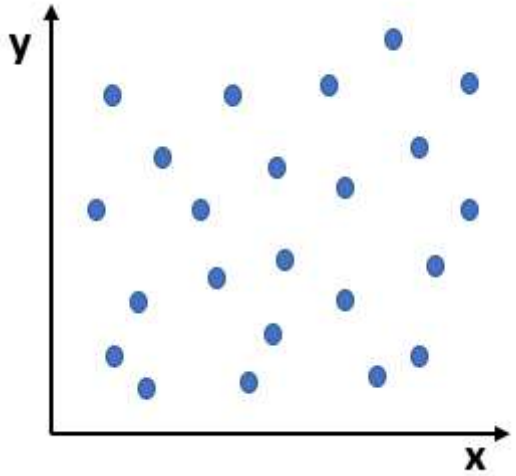
Let us understand it in Simple Linear Regression first.

When we fit a straight line through the data, we get two parameters i.e., the intercept (β_0) and the slope (β_1).

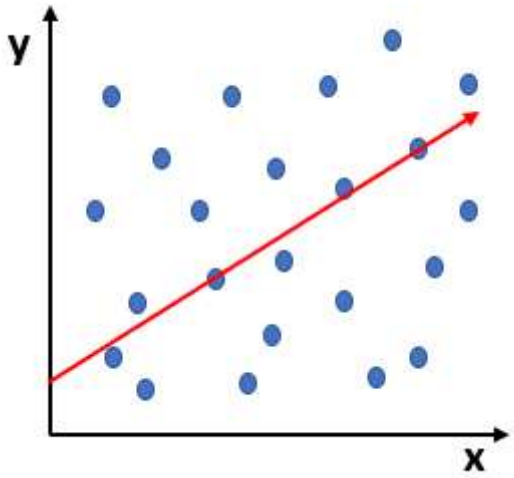
$$y = \beta_0 + \beta_1 x$$

Now, β_0 is not of much importance right now, but there are a few aspects around β_1 which needs to be checked and verified. Suppose we have a dataset for which the scatter plot looks like the following:

Scatter Plot



When we run a linear regression on this dataset in Python, Python will fit a line on the data which looks like the following:



We can clearly see that the data in randomly scattered and doesn't seem to follow linear trend. Python will anyway fit a line through the data using the least squared method. We can see that the fitted line is of no use in this case. Hence, every time we perform linear regression, we need to test whether the fitted line is a significant one or not (in other terms, test whether β_1 is significant or not). We will use Hypothesis Testing on β_1 for the same.

Steps to Perform Hypothesis testing:

1.Set the Hypothesis

2.Set the Significance Level, Criteria for a decision

3.Compute the test statistics

4.Make a decision

9. How would you decide the importance of variable for multivariate regression?

Ans:

1. Variables that are already proven in the literature to be related to the outcome.
2. Variables that can either be considered the cause of the exposure, the outcome, or both.
3. Interaction terms of variables that have large main effects.

10. What is the difference between R^2 vs adjusted R^2?

Ans:

However, there is one main difference between R^2 and the adjusted R^2:
R^2 assumes that every single variable explains the variation in the dependent variable.
The adjusted R^2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable.

In [ ]: