**Skill test Questions and Answers**

**1) True-False: Linear Regression is a supervised machine learning algorithm.**

A) TRUE  
B) FALSE

**Solution: (A)**

Yes, Linear regression is a supervised learning algorithm because it uses true labels for training. Supervised learning algorithm should have input variable (x) and an output variable (Y) for each example.

**2) True-False: Linear Regression is mainly used for Regression.**

A) TRUE  
B) FALSE

**Solution: (A)**

**Linear Regression** has dependent variables that have continuous values.

**3) True-False: It is possible to design a Linear regression algorithm using a neural network?**

A) TRUE  
B) FALSE

**Solution: (A)**

True. A Neural network can be used as a *universal* approximator, so it can definitely implement a linear regression algorithm.

**4) Which of the following methods do we use to find the best fit line for data in Linear Regression?**

A) Least Square Error  
B) Maximum Likelihood  
C) Logarithmic Loss  
D) Both A and B

**Solution: (A)**

In linear regression, we try to minimize the least square errors of the model to identify the line of best fit.

**5) Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?**

A) AUC-ROC  
B) Accuracy  
C) Logloss  
D) Mean-Squared-Error

**Solution: (D)**

Since linear regression gives output as continuous values, so in such case we use mean squared error metric to evaluate the model performance. Remaining options are use in case of a classification problem.

**6) True-False: Lasso Regularization can be used for variable selection in Linear Regression.**

A) TRUE  
B) FALSE

**Solution: (A)**

True, In case of lasso regression we apply absolute penalty which makes some of the coefficients zero.

**7) Which of the following is true about Residuals ?**

A) Lower is better  
B) Higher is better  
C) A or B depend on the situation  
D) None of these

**Solution: (A)**

Residuals refer to the error values of the model. Therefore lower residuals are desired.

**8)**Suppose that we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this data.

You found that correlation coefficient for one of it’s variable(Say X1) with Y is -0.95.

**Which of the following is true for X1?**

A) Relation between the X1 and Y is weak  
B) Relation between the X1 and Y is strong  
C) Relation between the X1 and Y is neutral  
D) Correlation can’t judge the relationship

**Solution: (B)**

The absolute value of the correlation coefficient denotes the strength of the relationship. Since  absolute correlation is very high it means that the relationship is strong between X1 and Y.

**9) Looking at above two characteristics, which of the following option is the correct for Pearson correlation between V1 and V2?**

If you are given the two variables V1 and V2 and they are following below two characteristics.

1. If V1 increases then V2 also increases

2. If V1 decreases then V2 behavior is unknown

A) Pearson correlation will be close to 1  
B) Pearson correlation will be close to -1  
C) Pearson correlation will be close to 0  
D) None of these

**Solution: (D)**

We cannot comment on the correlation coefficient by using only statement 1.  We need to consider the both of these two statements. Consider V1 as x and V2 as |x|. The correlation coefficient would not be close to 1 in such a case.

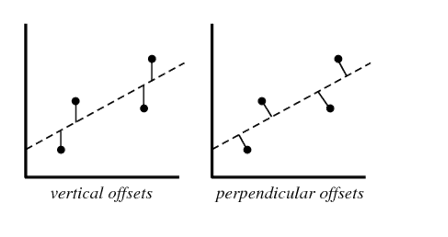
**10) Suppose Pearson correlation between V1 and V2 is zero. In such case, is it right to conclude that V1 and V2 do not have any relation between them?**

A) TRUE  
B) FALSE

**Solution: (B)**

Pearson correlation coefficient between 2 variables might be zero even when they have a relationship between them. If the correlation coefficient is zero, it just means that that they don’t move together. We can take examples like y=|x| or y=x^2.

**11) Which of the following offsets, do we use in linear regression’s least square line fit? Suppose horizontal axis is independent variable and vertical axis is dependent variable.**



A) Vertical offset  
B) Perpendicular offset  
C) Both, depending on the situation  
D) None of above

**Solution: (A)**

We always consider residuals as vertical offsets. We calculate the direct differences between actual value and the Y labels. Perpendicular offset are useful in case of PCA.

**12) True- False: Overfitting is more likely when you have huge amount of data to train?**

A) TRUE  
B) FALSE

**Solution: (B)**

With a small training dataset, it’s easier to find a hypothesis to fit the training data exactly i.e. overfitting.

**13) We can also compute the coefficient of linear regression with the help of an analytical method called “Normal Equation”. Which of the following is/are true about Normal Equation?**

1. We don’t have to choose the learning rate
2. It becomes slow when number of features is very large
3. Thers is no need to iterate

A) 1 and 2  
B) 1 and 3  
C) 2 and 3  
D) 1,2 and 3

**Solution: (D)**

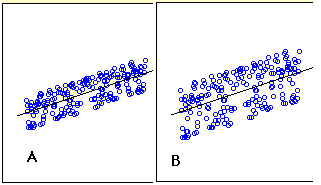
Instead of gradient descent, Normal Equation can also be used to find coefficients. Refer this [article](http://eli.thegreenplace.net/2014/derivation-of-the-normal-equation-for-linear-regression/) for read more about normal equation.

**14) Which of the following statement is true about sum of residuals of A and B?**

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases A and B.

**Note:**

1. Scale is same in both graphs for both axis.
2. X axis is independent variable and Y-axis is dependent variable.



A) A has higher sum of residuals than B  
B) A has lower sum of residual than B  
C) Both have same sum of residuals  
D) None of these

**Solution: (C)**

Sum of residuals will always be zero, therefore both have same sum of residuals

**Question Context 15-17:**

Suppose you have fitted a complex regression model on a dataset. Now, you are using Ridge regression with penality x.

**15) Choose the option which describes bias in best manner.**  
A) In case of very large x; bias is low  
B) In case of very large x; bias is high  
C) We can’t say about bias  
D) None of these

**Solution: (B)**

If the penalty is very large it means model is less complex, therefore the bias would be high.

**16) What will happen when you apply very large penalty?**

A) Some of the coefficient will become absolute zero  
B) Some of the coefficient will approach zero but not absolute zero  
C) Both A and B depending on the situation  
D) None of these

**Solution: (B)**

In lasso some of the coefficient value become zero, but in case of Ridge, the coefficients become close to zero but not zero.

**17) What will happen when you apply very large penalty in case of Lasso?**  
A) Some of the coefficient will become zero  
B) Some of the coefficient will be approaching to zero but not absolute zero  
C) Both A and B depending on the situation  
D) None of these

**Solution: (A)**

As already discussed, lasso applies absolute penalty, so some of the coefficients will become zero.

**18) Which of the following statement is true about outliers in Linear regression?**

A) Linear regression is sensitive to outliers  
B) Linear regression is not sensitive to outliers  
C) Can’t say  
D) None of these

**Solution: (A)**

The slope of the regression line will change due to outliers in most of the cases. So Linear Regression is sensitive to outliers.

**19) Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?**

A) Since the there is a relationship means our model is not good  
B) Since the there is a relationship means our model is good  
C) Can’t say  
D) None of these

**Solution: (A)**

There should not be any relationship between predicted values and residuals. If there exists any relationship between them,it means that the model has not perfectly captured the information in the data.

**Question Context 20-22:**

Suppose that you have a dataset D1 and you design a linear regression model of degree 3 polynomial and you found that the training and testing error is “0” or in another terms it perfectly fits the data.

**20) What will happen when you fit degree 4 polynomial in linear regression?**  
A) There are high chances that degree 4 polynomial will over fit the data  
B) There are high chances that degree 4 polynomial will under fit the data  
C) Can’t say  
D) None of these

**Solution: (A)**

Since is more degree 4 will be more complex(overfit the data) than the degree 3 model so it will again perfectly fit the data. In such case training error will be zero but test error may not be zero.

**21) What will happen when you fit degree 2 polynomial in linear regression?**  
A) It is high chances that degree 2 polynomial will over fit the data  
B) It is high chances that degree 2 polynomial will under fit the data  
C) Can’t say  
D) None of these

**Solution: (B)**

If a degree 3 polynomial fits the data perfectly, it’s highly likely that a simpler model(degree 2 polynomial) might under fit the data.

**22) In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?**

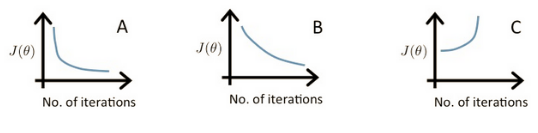
A) Bias will be high, variance will be high  
B) Bias will be low, variance will be high  
C) Bias will be high, variance will be low  
D) Bias will be low, variance will be low

**Solution: (C)**

Since a degree 2 polynomial will be less complex as compared to degree 3, the bias will be high and variance will be low.

**Question Context 23:**

Which of the following is true about below graphs(A,B, C left to right) between the cost function and Number of iterations?



**23) Suppose l1, l2 and l3 are the three learning rates for A,B,C respectively. Which of the following is true about l1,l2 and l3?**

A) l2 < l1 < l3

B) l1 > l2 > l3  
C) l1 = l2 = l3  
D) None of these

**Solution: (A)**

In case of high learning rate, step will be high, the objective function will decrease quickly initially, but it will not find the global minima and objective function starts increasing after a few iterations.

In case of low learning rate, the step will be small. So the objective function will decrease slowly

**Question Context 24-25:**

We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly.

**24) Now we increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training error?**

A) Increase  
B) Decrease  
C) Remain constant  
D) Can’t Say

**Solution: (D)**

Training error may increase or decrease depending on the values that are used to fit the model. If the values used to train contain more outliers gradually, then the error might just increase.

**25) What do you expect will happen with bias and variance as you increase the size of training data?**

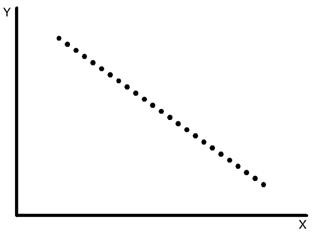
A) Bias increases and Variance increases  
B) Bias decreases and Variance increases  
C) Bias decreases and Variance decreases  
D) Bias increases and Variance decreases  
E) Can’t Say False

**Solution: (D)**

As we increase the size of the training data, the bias would increase while the variance would decrease.

**Question Context 26:**

Consider the following data where one input(X) and one output(Y) is given.



**26) What would be the root mean square training error for this data if you run a Linear Regression model of the form (Y = A0+A1X)?**

A) Less than 0  
B) Greater than zero  
C) Equal to 0  
D) None of these

**Solution: (C)**

We can perfectly fit the line on the following data so mean error will be zero.

**Question Context 27-28:**

Suppose you have been given the following scenario for training and validation error for Linear Regression.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Scenario** | **Learning Rate** | **Number of iterations** | **Training Error** | **Validation Error** |
| 1 | 0.1 | 1000 | 100 | 110 |
| 2 | 0.2 | 600 | 90 | 105 |
| 3 | 0.3 | 400 | 110 | 110 |
| 4 | 0.4 | 300 | 120 | 130 |
| 5 | 0.4 | 250 | 130 | 150 |

**27) Which of the following scenario would give you the right hyper parameter?**

A) 1  
B) 2  
C) 3  
D) 4

**Solution: (B)**

Option B would be the better option because it leads to less training as well as validation error.

**28) Suppose you got the tuned hyper parameters from the previous question. Now, Imagine you want to add a variable in variable space such that this added feature is important. Which of the following thing would you observe in such case?**

A) Training Error will decrease and Validation error will increase

B) Training Error will increase and Validation error will increase  
C) Training Error will increase and Validation error will decrease  
D) Training Error will decrease and Validation error will decrease  
E) None of the above

**Solution: (D)**

If the added feature is important, the training and validation error would decrease.

**Question Context 29-30:**

Suppose, you got a situation where you find that your linear regression model is under fitting the data.

**29) In such situation which of the following options would you consider?**

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

A) 1 and 2  
B) 2 and 3  
C) 1 and 3  
D) 1, 2 and 3

**Solution: (A)**

In case of under fitting, you need to induce more variables in variable space or you can add some polynomial degree variables to make the model more complex to be able to fir the data better.

**30) Now situation is same as written in previous question(under fitting).Which of following regularization algorithm would you prefer?**

A) L1  
B) L2  
C) Any  
D) None of these

**Solution: (D)**

I won’t use any regularization methods because regularization is used in case of overfitting.

## ****Let’s get started with linear regression!****

### What is linear regression?

In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.  
In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

### 2. State the assumptions in a linear regression model.

**There are three main assumptions in a linear regression model:**

1. The 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, ε(i), are normally distributed.
   2. Zero mean assumption: It is assumed that the residuals have a mean value of zero.
   3. Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 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.

**Explanation:**

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= β0+ β1x(i) + ε(i)  
   This is the assumed linear model, where ε is the residual term.  
   E(Y) = E(β0+ β1x(i) + ε(i))  
           = E(β0+ β1x(i) + ε(i))  
   If the expectation(mean) of residuals, E(ε(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 error terms) should be independent. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves. If some correlation is present, it 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 squares solution (or normal equation solution) is lost.

### What is feature engineering? How do you apply it in the process of modelling?

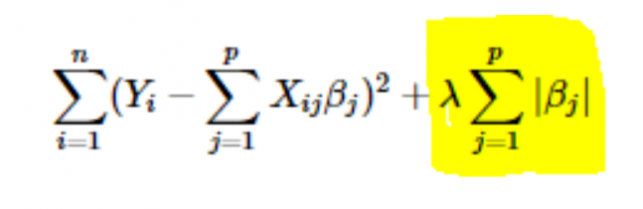
Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.  
In layman terms, feature engineering means the development of new features that may help you understand and model the problem in a better way. Feature engineering is of two kinds — business driven and data-driven. Business-driven feature engineering revolves around the inclusion of features from a business point of view. The job here is to transform the business variables into features of the problem. In case of data-driven feature engineering, the features you add do not have any significant physical interpretation, but they help the model in the prediction of the target variable.  
To apply feature engineering, one must be fully acquainted with the dataset. This involves knowing what the given data is, what it signifies, what the raw features are, etc. You must also have a crystal clear idea of the problem, such as what factors affect the target variable, what the physical interpretation of the variable is, etc.

[5 Breakthrough Applications of Machine Learning](https://upgrad.com/blog/machine-learning-and-its-breakthrough-applications/)

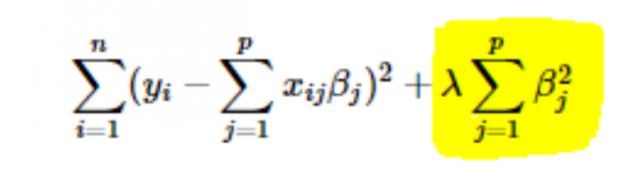
### What is the use of regularisation? Explain L1 and L2 regularisations.

Regularisation is a technique that is used to tackle the problem of overfitting of the model. When a very complex model is implemented on the training data, it overfits. At times, the simple model might not be able to generalise the data and the complex model overfits. To address this problem, regularisation is used.  
Regularisation is nothing but adding the coefficient terms (betas) to the cost function so that the terms are penalised and are small in magnitude. This essentially helps in capturing the trends in the data and at the same time prevents overfitting by not letting the model become too complex.

* L1 or LASSO regularisation: Here, the absolute values of the coefficients are added to the cost function. This can be seen in the following equation; the highlighted part corresponds to the L1 or LASSO regularisation. This regularisation technique gives sparse results, which lead to feature selection as well.

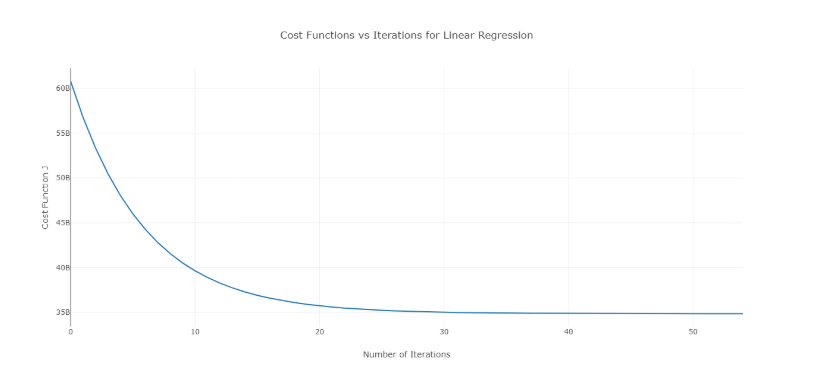


* L2 or Ridge regularisation: Here, the squares of the coefficients are added to the cost function. This can be seen in the following equation, where the highlighted part corresponds to the L2 or Ridge regularisation.



### 5. How to choose the value of the parameter learning rate (α)?

Selecting the value of learning rate is a tricky business. If the value is too small, the gradient descent algorithm takes ages to converge to the optimal solution. On the other hand, if the value of the learning rate is high, the gradient descent will overshoot the optimal solution and most likely never converge to the optimal solution.  
To overcome this problem, you can try different values of alpha over a range of values and plot the cost vs the number of iterations. Then, based on the graphs, the value corresponding to the graph showing the rapid decrease can be chosen.

  
The aforementioned graph is an ideal cost vs the number of iterations curve. Note that the cost initially decreases as the number of iterations increases, but after certain iterations, the gradient descent converges and the cost does not decrease anymore.  
If you see that the cost is increasing with the number of iterations, your learning rate parameter is high and it needs to be decreased.

### 6. How to choose the value of the regularisation parameter (λ)?

Selecting the regularisation parameter is a tricky business. If the value of λ is too high, it will lead to extremely small values of the regression coefficient β, which will lead to the model underfitting (high bias – low variance). On the other hand, if the value of λ is 0 (very small), the model will tend to overfit the training data (low bias – high variance).  
There is no proper way to select the value of λ. What you can do is have a sub-sample of data and run the algorithm multiple times on different sets. Here, the person has to decide how much variance can be tolerated. Once the user is satisfied with the variance, that value of λ can be chosen for the full dataset.  
One thing to be noted is that the value of λ selected here was optimal for that subset, not for the entire training data.

### 7. Can we use linear regression for time series analysis?

One can use linear regression for time series analysis, but the results are not promising. So, it is generally not advisable to do so. The reasons behind this are —

1. Time series data is mostly used for the prediction of the future, but linear regression seldom gives good results for future prediction as it is not meant for extrapolation.
2. Mostly, time series data have a pattern, such as during peak hours, festive seasons, etc., which would most likely be treated as outliers in the linear regression analysis.

### 8. What value is the sum of the residuals of a linear regression close to? Justify.

**Ans** The sum of the residuals of a linear regression is 0. Linear regression works on the assumption that the errors (residuals) are normally distributed with a mean of 0, i.e.

***Y = βT X + ε***

Here, Y is the target or dependent variable,  
*β*is the vector of the regression coefficient,  
X is the feature matrix containing all the features as the columns,  
ε is the residual term such that*ε*~ N(0,σ2).  
So, the sum of all the residuals is the expected value of the residuals times the total number of data points. Since the expectation of residuals is 0, the sum of all the residual terms is zero.  
**Note**: N(μ,σ2) is the standard notation for a normal distribution having mean μ and standard deviation σ2.

### 9. How does multicollinearity affect the linear regression?

**Ans** Multicollinearity occurs when some of the independent variables are highly correlated (positively or negatively) with each other. This multicollinearity causes a problem as it is against the basic assumption of linear regression. The presence of multicollinearity does not affect the predictive capability of the model. So, if you just want predictions, the presence of multicollinearity does not affect your output. However, if you want to draw some insights from the model and apply them in, let’s say, some business model, it may cause problems.  
One of the major problems caused by multicollinearity is that it leads to incorrect interpretations and provides wrong insights. The coefficients of linear regression suggest the mean change in the target value if a feature is changed by one unit. So, if multicollinearity exists, this does not hold true as changing one feature will lead to changes in the correlated variable and consequent changes in the target variable. This leads to wrong insights and can produce hazardous results for a business.  
A highly effective way of dealing with multicollinearity is the use of VIF (Variance Inflation Factor). Higher the value of VIF for a feature, more linearly correlated is that feature. Simply remove the feature with very high VIF value and re-train the model on the remaining dataset.

### 10. What is the normal form (equation) of linear regression? When should it be preferred to the gradient descent method?

**The normal equation for linear regression is —**

β=(XTX)-1.XTY

Here, Y=βTX is the model for the linear regression,  
Y is the target or dependent variable,  
β is the vector of the regression coefficient, which is arrived at using the normal equation,  
X is the feature matrix containing all the features as the columns.  
Note here that the first column in the X matrix consists of all 1s. This is to incorporate the offset value for the regression line.  
Comparison between gradient descent and normal equation:

|  |  |
| --- | --- |
| **Gradient Descent** | **Normal Equation** |
| Needs hyper-parameter tuning for alpha (learning parameter) | No such need |
| It is an iterative process | It is a non-iterative process |
| O(kn2) time complexity | O(n3) time complexity due to evaluation of *X*T*X* |
| Prefered when n is extremely large | Becomes quite slow for large values of n |

Here, ‘k’ is the maximum number of iterations for gradient descent, and ‘n’ is the total number of data points in the training set.  
Clearly, if we have large training data, normal equation is not prefered for use. For small values of ‘n’, normal equation is faster than gradient descent.

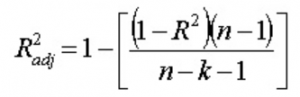
### 11. You run your regression on different subsets of your data, and in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

This case implies that the dataset is heterogeneous. So, to overcome this problem, the dataset should be clustered into different subsets, and then separate models should be built for each cluster. Another way to deal with this problem is to use non-parametric models, such as decision trees, which can deal with heterogeneous data quite efficiently.

### 12. Your linear regression doesn’t run and communicates that there is an infinite number of best estimates for the regression coefficients. What could be wrong?

This condition arises when there is a perfect correlation (positive or negative) between some variables. In this case, there is no unique value for the coefficients, and hence, the given condition arises.

### 13. What do you mean by adjusted R2? How is it different from R2?

Adjusted R2, just like R2, is a representative of the number of points lying around the regression line. That is, it shows how well the model is fitting the training data. The formula for adjusted R2  is —  
  
Here, n is the number of data points, and k is the number of features.  
One drawback of R2 is that it will always increase with the addition of a new feature, whether the new feature is useful or not. The adjusted R2 overcomes this drawback. The value of the adjusted R2 increases only if the newly added feature plays a significant role in the model.

### 14. How do you interpret the residual vs fitted value curve?

The residual vs fitted value plot is used to see whether the predicted values and residuals have a correlation or not. If the residuals are distributed normally, with a mean around the fitted value and a constant variance, our model is working fine; otherwise, there is some issue with the model.  
The most common problem that can be found when training the model over a large range of a dataset is heteroscedasticity(this is explained in the answer below). The presence of heteroscedasticity can be easily seen by plotting the residual vs fitted value curve.

### 15. What is heteroscedasticity? What are the consequences, and how can you overcome it?

A random variable is said to be heteroscedastic when different subpopulations have different variabilities (standard deviation).  
The existence of heteroscedasticity gives rise to certain problems in the regression analysis as the assumption says that error terms are uncorrelated and, hence, the variance is constant. The presence of heteroscedasticity can often be seen in the form of a cone-like scatter plot for residual vs fitted values.  
One of the basic assumptions of linear regression is that heteroscedasticity is not present in the data. Due to the violation of assumptions, the Ordinary Least Squares (OLS) estimators are not the Best Linear Unbiased Estimators (BLUE). Hence, they do not give the least variance than other Linear Unbiased Estimators (LUEs).  
There is no fixed procedure to overcome heteroscedasticity. However, there are some ways that may lead to a reduction of heteroscedasticity. They are —

1. Logarithmising the data: A series that is increasing exponentially often results in increased variability. This can be overcome using the log transformation.
2. Using weighted linear regression: Here, the OLS method is applied to the weighted values of X and Y. One way is to attach weights directly related to the magnitude of the dependent variable.

[How does Unsupervised Machine Learning Work?](https://upgrad.com/blog/how-does-unsupervised-machine-learning-work/)

### 16. What is VIF? How do you calculate it?

Variance Inflation Factor (VIF) is used to check the presence of multicollinearity in a dataset. It is calculated as—   
  
Here, VIFj  is the value of VIF for the jth variable,  
Rj2 is the R2 value of the model when that variable is regressed against all the other independent variables.  
If the value of VIF is high for a variable, it implies that the R2  value of the corresponding model is high, i.e. other independent variables are able to explain that variable. In simple terms, the variable is linearly dependent on some other variables.

### 17. How do you know that linear regression is suitable for any given data?

To see if linear regression is suitable for any given data, a scatter plot can be used. If the relationship looks linear, we can go for a linear model. But if it is not the case, we have to apply some transformations to make the relationship linear. Plotting the scatter plots is easy in case of simple or univariate linear regression. But in case of multivariate linear regression, two-dimensional pairwise scatter plots, rotating plots, and dynamic graphs can be plotted.

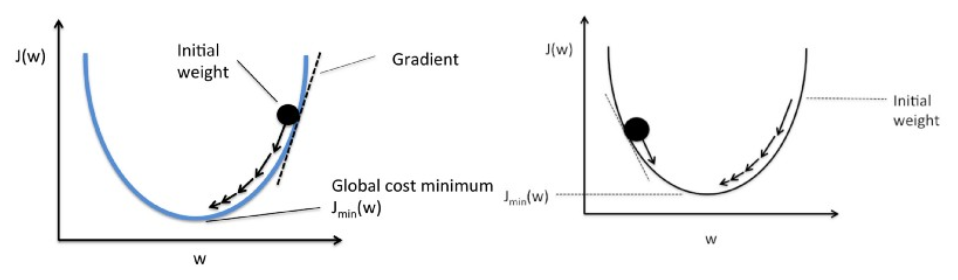
### 18. How is hypothesis testing used in linear regression?

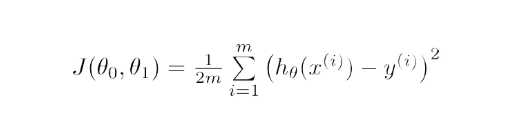
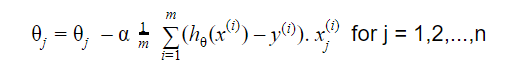
Hypothesis testing can be carried out in linear regression for the following purposes:

1. To check whether a predictor is significant for the prediction of the target variable. Two common methods for this are —
   1. By the use of p-values:  
      If the p-value of a variable is greater than a certain limit (usually 0.05), the variable is insignificant in the prediction of the target variable.
   2. By checking the values of the regression coefficient:  
      If the value of regression coefficient corresponding to a predictor is zero, that variable is insignificant in the prediction of the target variable and has no linear relationship with it.
2. To check whether the calculated regression coefficients are good estimators of the actual coefficients.

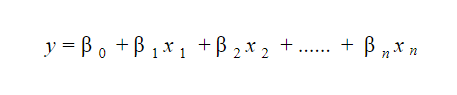
### 19. Explain gradient descent with respect to linear regression.

Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.  
Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Mathematically, the aim of gradient descent for linear regression is to find the solution of  
ArgMin J(Θ0,Θ*1*), where J(Θ*0*,Θ1) is the cost function of the linear regression. It is given by —    
  
Here, h is the linear hypothesis model, h=Θ*0* + Θ1x, y is the true output, and m is the number of the data points in the training set.  
Gradient Descent starts with a random solution, and then based on the direction of the gradient, the solution is updated to the new value where the cost function has a lower value.  
The update is:  
Repeat until convergence  


### 20. How do you interpret a linear regression model?

A linear regression model is quite easy to interpret. The model is of the following form:  
  
The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of x0 increases by 1 unit, keeping other variables constant, the total increase in the value of y will be βi. Mathematically, the intercept term (β0) is the response when all the predictor terms are set to zero or not considered.  
[These 6 Machine Learning Techniques are Improving Healthcare](https://upgrad.com/blog/machine-learning-applications-in-healthcare-2018/)

### 21. What is robust regression?

A regression model should be robust in nature. This means that with changes in a few observations, the model should not change drastically. Also, it should not be much affected by the outliers.  
A regression model with OLS (Ordinary Least Squares) is quite sensitive to the outliers. To overcome this problem, we can use the WLS (Weighted Least Squares) method to determine the estimators of the regression coefficients. Here, less weights are given to the outliers or high leverage points in the fitting, making these points less impactful.

### 22. Which graphs are suggested to be observed before model fitting?

Before fitting the model, one must be well aware of the data, such as what the trends, distribution, skewness, etc. in the variables are. Graphs such as histograms, box plots, and dot plots can be used to observe the distribution of the variables. Apart from this, one must also analyse what the relationship between dependent and independent variables is. This can be done by scatter plots (in case of univariate problems), rotating plots, dynamic plots, etc.

### 23. What is the generalized linear model?

The generalized linear model is the derivative of the ordinary linear regression model. GLM is more flexible in terms of residuals and can be used where linear regression does not seem appropriate. GLM allows the distribution of residuals to be other than a normal distribution. It generalizes the linear regression by allowing the linear model to link to the target variable using the linking function. Model estimation is done using the method of maximum likelihood estimation.

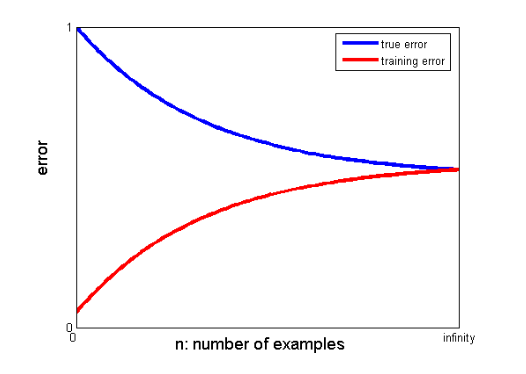
### 24. Explain the bias-variance trade-off.

Bias refers to the difference between the values predicted by the model and the real values. It is an error. One of the goals of an ML algorithm is to have a low bias.  
Variance refers to the sensitivity of the model to small fluctuations in the training dataset. Another goal of an ML algorithm is to have low variance.  
For a dataset that is not exactly linear, it is not possible to have both bias and variance low at the same time. A straight line model will have low variance but high bias, whereas a high-degree polynomial will have low bias but high variance.  
There is no escaping the relationship between bias and variance in machine learning.

1. Decreasing the bias increases the variance.
2. Decreasing the variance increases the bias.

So, there is a trade-off between the two; the ML specialist has to decide, based on the assigned problem, how much bias and variance can be tolerated. Based on this, the final model is built.

### 25. How can learning curves help create a better model?

Learning curves give the indication of the presence of overfitting or underfitting.  
In a learning curve, the training error and cross-validating error are plotted against the number of training data points. A typical learning curve looks like this:  
  
If the training error and true error (cross-validating error) converge to the same value and the corresponding value of the error is high, it indicates that the model is underfitting and is suffering from high bias.  
If there is a significant gap between the converging values of the training and cross-validating errors, i.e. the cross-validating error is significantly higher than the training error, it suggests that the model is overfitting the training data and is suffering from a high variance.

## Linear Regression Interview Questions – Fundamental Questions

Let us begin with a fundamental Linear Regression Interview Questions

### 1. What is a Linear Regression?

In simple terms, linear regression is adopting a linear approach to modeling the relationship between a dependent variable (scalar response) and one or more independent variables (explanatory variables). In case you have one explanatory variable, you call it a simple linear regression. In case you have more than one independent variable, you refer to the process as multiple linear regressions.

### 2. Can you list out the critical [assumptions of linear regression](https://www.digitalvidya.com/blog/assumptions-of-linear-regression/)?

There are three crucial assumptions one has to make in linear regression. They are,

1. It is imperative to have a linear relationship between the dependent and independent A scatter plot can prove handy to check out this fact.
2. The independent variables in the dataset should not exhibit any multi-collinearity. In case they do, it should be at the barest minimum. There should be a restriction on their value depending on the domain requirement.
3. Homoscedasticity is one of the most critical It states that there should be an equal distribution of errors.

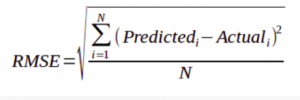
### 3.    What is Heteroscedasticity?

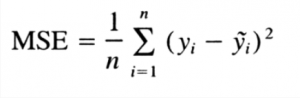
Heteroscedasticity is the exact opposite of homoscedasticity. It entails that there is no equal distribution of the error terms. You use a log function to rectify this phenomenon.

### 4.    What is the primary difference between R square and adjusted R square?

In [linear regression](https://www.digitalvidya.com/blog/linear-regression/), you use both these values for model validation. However, there is a clear distinction between the two. R square accounts for the variation of all independent variables on the dependent variable. In other words, it considers each independent variable for explaining the variation. In the case of Adjusted R square, it accounts for the significant variables alone for indicating the percentage of variation in the model. By significant, we refer to the P values less than 0.05.

### 5.    Can you list out the formulas to find RMSE and MSE?





The most common measures of accuracy for any linear regression are RMSE and MSE. MSE stands for Mean Square Error whereas RMSE stands for Root Mean Square Error. The formulas of RMSE and MSE are as hereunder.

## Linear Regression Interview Questions – Complex Questions

We have seen some of the basic interview questions on linear regression. As we move further into the article, we shall look at some of the complex linear regression interview questions as well.

Bottom of Form

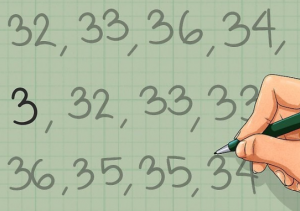
### 6.    Can you name a possible method of improving the accuracy of a linear regression model?

You can do so in many ways. One of the most common ways is ‘The Outlier Treatment.’

Outliers have great significance in [linear regression](https://www.digitalvidya.com/blog/what-is-regression/) because regression is very sensitive to outliers. Therefore, it becomes critical to treat outliers with appropriate values. It can also prove useful if you replace the values with mean, median, mode or percentile depending on the distribution.

### 7.    What are outliers? How do you detect and treat them?

An outlier is an observation point distant from other observations. It might be due to a variance in the measurement. It can also indicate an experimental error. Under such circumstances, you need to exclude the same from the data set. If you do not detect and treat them, they can cause problems in statistical analysis.



You can see that 3 is the outlier in this example.

There is no strict mathematical calculation of how to determine an outlier. Deciding whether an observation is an outlier or not, is itself a subjective exercise. However, you can detect outliers through various methods. Some of them are graphical and are known as normal probability plots whereas some are model-based. You have some hybrid techniques such as Boxplots.

Once you have detected the outlier, you should either remove them or correct them to ensure accurate analysis. Some of the methods of eliminating outliers are the Z-Score and the IQR Score methods.

### 8.    How do you interpret a Q-Q plot in a linear regression model?

As the name suggests, the Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words, you plot quantiles against quantiles.

Whenever you interpret a Q-Q plot, you should concentrate on the ‘y = x’ line. You also call it the 45-degree line in statistics. It entails that each of your distributions has the same quantiles. In case you witness a deviation from this line, one of the distributions could be skewed when compared to the other.

### 9.    What is the importance of the F-test in a linear model?

The F-test is a crucial one in the sense that it tests the goodness of the model. When you reiterate the model to improve the accuracy with the changes, the F-test proves its utility in understanding the effect of the overall regression.

### 10.  What are the disadvantages of the linear regression model?

One of the most significant demerits of the linear model is that it is sensitive and dependent on the outliers. It can affect the overall result. Another notable demerit of the linear model is overfitting. Similarly, underfitting is also a significant disadvantage of the linear model.

### 11.  What is the curse of dimensionality? Can you give an example?

When you analyze and organize data in high-dimensional spaces (usually in thousands), various situations can arise that usually do not do so when you analyze data in low-dimensional settings (3-dimensional physical space). The curse of dimensionality refers to such phenomena.

Here is an example.

All kids love to eat chocolates. Now, you bring a truckload of chocolates in front of the kid. These chocolates come in different colors, shapes, tastes, and price. Consider the following scenario.

The kid has to choose one chocolate from the truck depending on the following factors.

1. Only taste – There are usually four tastes, sweet, salty, sour, and bitter. Hence, the child will have to try out only four chocolates before choosing one to its liking.
2. Taste and Color – Assume there are only four colors. Hence, the child will now have to taste a minimum of 16 (4 X 4) before making the right choice.
3. Taste, color, and shape – Let us assume that there are five shapes. Therefore, the child will now have to eat a minimum of 80 chocolates (4 X 4 X 5).

What will happen to the child if it tries out 80 chocolates at a time? It will naturally become sick. Hence, it will not be in a position to try out the chocolates. This example is the perfect one to explain the curse of dimensionality. The more the options you have, the more the problems you encounter.

## Linear Regression Interview Questions – Multiple Choice Questions

Let us now look at some multiple-choice linear regression interview questions.

### 1.    In regression analysis, which of the statements is true?

1. The mean of residuals is always equal to Zero
2. The Mean of residuals is less than Zero at all times
3. The Mean of residuals is more than Zero at all times
4. You do not have any such rule for residuals.

The correct answer is A. In [regression analysis](https://www.digitalvidya.com/blog/regression-analysis/), the sum of the residuals in regression is always equal to Zero. Thus, it implies that the mean will also be Zero if the sum of the residuals is Zero.

Top of Form

### 2.    Which of the statements is correct about Heteroscedasticity?

1. Linear regression with different error terms
2. Linear regression with constant error terms
3. Linear regression with no error terms
4. None of the above

The solution is the option A. When you have a non-constant variance in the error terms, it results in Heteroscedasticity. Such non-constant variance occurs because you have outliers.

### 3.    Which of the following plots is best suited to test the linear relationship of independent and dependent continuous variables?

1. Scatter Plot
2. Bar Chart
3. Histograms
4. None of the above options

The answer is A. The Scatter plot is the best way to determine the relationship between continuous variables. You can find out how one variable changes with respect to the other.

### 4.    If you have only one independent variable, how many coefficients will you require to estimate in a simple linear regression model?

1. One
2. Two
3. No idea

The answer is B. Consider the simple linear regression with one independent variable. Y = a + bx. You can see that you need two coefficients.

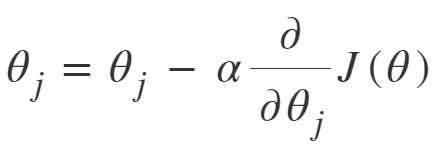
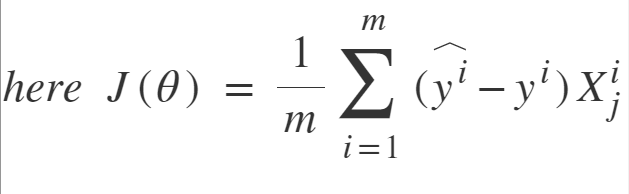
* **What is linear regression?**
* A linear regression is a linear approximation of a causal relationship between two or more variables.
* It falls under the supervised machine learning algorithms.
* **What is process of carrying out a linear regression?**
* Get sample data
* Design a model that works on that sample
* Make predictions for the whole population
* **How do you represent a simple linear regression?**
* Y = b0 +b1 x1 + e
* Y – dependent variable
* X1 – independent variable
* e – Error term = Y – Y(hat)
* **What is the difference between correlation and regression?**
* Correlation does not apply causation. Regression is done to understand the impact of independent variable on the dependent variable.
* Correlation is symmetric regrading both the variables p(x,y) = p(y,x). Regression is one way.
* Correlation does not capture the direction of causal relationship. Regression captures the cause and effect.
* **What are the columns in the coefficient table?**
* The coefficient table contains the variable name, coefficient, standard error and p-value.
* **What is standard error?**
* Standard error shows the accuracy for each variable
* **What is p-value?**
* The p-value shows the significance of the variable. It tells us if the variable is useful or not.
* The H0 is coefficient = 0 and the H1 is coefficient ≠ 0
* If p-value < 0.05 (in most of the cases) we reject H0
* **What is OLS?**
* OLS stands for ordinary least square
* It measures the error between the actual Y and predicted Y
* Lower the error, better is the model
* **What are the other regression methods?**
* Generalized least squares
* Maximum likelihood estimates
* Bayesian regression
* Kernel regression
* Gaussian regression
* **What is TSS, ESS and RSS?**
* TSS stands for Total Sum of Squares. It measures the total variability.
* TSS = ∑(y – y(mean))2
* ESS stands for Explained Sum of Squares. It measures the variability that is explained.
* ESS = ∑(y(pred) – y(mean))2
* RSS stands for Residual Sum of Squares. It measures the difference between the observed Y and predicted Y.
* RSS = ∑(y – y(pred))2
* **What is the relationship between TSS, ESS and RSS?**
* TSS = ESS + RSS
* Total variability = Explained variability + Unexplained variability
* **What is R-Squared?**
* R-Squared is also known as goodness of fit
* Smaller the RSS, better is the model
* R-Sq = ESS / TSS = 1 – (RSS / TSS)
* R-Squared takes a value between 0 and 1.
* If R-Sq = 0 then the model does not explain any variability
* If R-Sq = 1 then the model explains entire variability
* **What is adjusted R-Squared?**
* Adjusted R-Squared is a step on R-Squared and adjusts for the number of variables included in the model
* As we add more variables the explanatory power of the model may increase.
* Adjusted R-Squared penalizes the model for the number of variables that are used in the model.
* **What is the relationship between R-Squared and Adjusted R-Squared?**
* Adj R-Sq is always lower than the R-Sq
* Adj R-Sq = 1 – ((1-RSq) \* (n-1) / (n-p-1))
* Where n is the number of observations and p is the number of variables
* **What happens when we add a variable and it increases the R-Sq but decreases the Adj R-Sq?**
* The variable can be omitted since it holds no predictive power
* We should also look at the p-value of the added variable and confirm our decision
* **What is feature selection?**
* It is a method to simplify the model and improves the speed
* It is done to avoid too many features
* p-value in regression coefficient table can be used to drop insignificant variables
* **What is feature scaling?**
* Different variables have different magnitude
* Feature scaling is done to bring the variables to the same magnitude
* Standardization is one of the methods used for feature scaling
* **What is standardization?**
* It is also called normalization
* X (std) = (x – µ) / σ
* Regardless of the data we will get data with mean 0 and standard deviation of 1
* **What is the interpretation of the weights?**
* In ML coefficients are called weights.
* A positive weight shows that as feature increases in value, so does Y
* A negative weight shows that as feature decreases in value, so does Y
* **What is the difference between overfitting and underfitting?**
* Underfitting happens when the model has not captured the underlying logic of the data.
* Overfitting happens when the model has focused too much on the training dataset that it cannot understand test dataset
* **How to identify if the model is overfitting or underfitting?**
* Underfit model performs bad (low accuracy) on training and bad (low accuracy) on test.
* Overfit model performs good (high accuracy) on training and bad (low accuracy) on test.
* A good model performs good (high accuracy) on training and good (high accuracy) on test.
* **What is multiple linear regression?**
* In multiple linear regression that are more than one predictor.
* Good models require multiple independent variables in order to address the higher complexity of the problem.
* Y = b0 +b1 x1 + b2 x2 + … + bk xk + e
* **What are the assumptions of linear regression?**
* Linearity
* No endogeneity
* Normality and homoscedasticity
* No autocorrelation
* No multi-collinearity
* **What happens if the linear regression violates any of its assumptions?**
* The biggest mistake you can make is to perform a regression that violates one of its assumptions.
* If the regression assumptions are violated, then performing regression analysis will yield incorrect results.
* **What does linearity mean?**
* It means a linear relationship
* To check if there is linear relationship between x and y the simplest thing to do is plot a scatter plot between x and y
* **What are the fixes of linearity?**
* If linearity assumption is violated, then we can use non-linear regression
* We can also transform the x (exponential transformation / log transformation)
* **What does no endogeneity mean?**
* No endogeneity means no relationship between x and ε
* It may be because we have omitted an important predictor from the model
* **What is omitted variable bias?**
* If the modeler forgets to include an important predictor in the model
* It may lead to counter-intuitive coefficient signs
* Once the important variable is included rest of the coefficients fall into place
* **What is the assumption of normality?**
* It means the normal distribution of the error term
* The mean of the residuals should be zero
* The standard deviation of the residuals should be constant
* **What is the assumption of homoscedasticity?**
* In simple terms it means the equal variance
* There is no relationship between the error term and the predicted Y
* **How to prevent heteroscedasticity?**
* It may be due to outliers
* It may be due to omitted variable bias
* Log transformation
* **What does autocorrelation mean?**
* It is common in time series modeling
* It means that Y(t) is dependent on historical values, Y(t-1) or Y(t-2) or … Y(t-k)
* **How to detect autocorrelation?**
* DW test is used to detect autocorrelation
* If DW test statistics is less than 1 then there is strong autocorrelation
* If DW test statistics is close to 2 then there is no autocorrelation
* If DW test statistics is more then 3 then there is strong autocorrelation
* **What are the remedies to remove autocorrelation?**
* There is no remedy in linear regression
* The modelers can try different models like AR, MA, ARMA or ARIMA
* **What does multicollinearity mean?**
* When two or more variables have high correlation
* If there is perfect multicollinearity then standard error will be infinite
* Imperfect multicollinearity means that the correlation is slightly less than 1 or slightly more than -1. However imperfect multicollinearity also causes serious issues in the model
* **What are the fixes of multicollinearity?**
* Find the correlation between each pair of independent variables
* If two variables are highly correlated, then either drop one of them or transform them into a single variable

# **Difference between Batch Gradient Descent and Stochastic Gradient Descent**

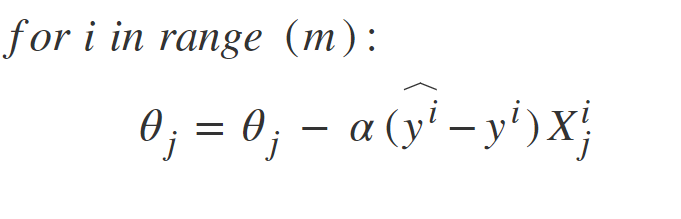
Last Updated: 02-07-2020

In order to train a Linear Regression model, we have to learn some model parameters such as feature weights and bias terms. An approach to do the same is Gradient Descent which is an iterative optimization algorithm capable of tweaking the model parameters by minimizing the cost function over the train data. It is a complete algorithm i.e it is guaranteed to find the global minimum (optimal solution) given there is enough time and the learning rate is not very high. Two Important variants of Gradient Descent which are widely used in Linear Regression as well as Neural networks are Batch Gradient Descent and Stochastic Gradient Descent(SGD).

**Batch Gradient Descent:** Batch Gradient Descent involves calculations over the full training set at each step as a result of which it is very slow on very large training data. Thus, it becomes very computationally expensive to do Batch GD. However, this is great for convex or relatively smooth error manifolds. Also, Batch GD scales well with the number of features.

[**Stochastic Gradient Descent:**](https://www.geeksforgeeks.org/ml-stochastic-gradient-descent-sgd/) SGD tries to solve the main problem in Batch Gradient descent which is the usage of whole training data to calculate gradients as each step. SGD is stochastic in nature i.e it picks up a “random” instance of training data at each step and then computes the gradient making it much faster as there is much fewer data to manipulate at a single time, unlike Batch GD.



There is a downside of the Stochastic nature of SGD i.e once it reaches close to the minimum value then it doesn’t settle down, instead bounces around which gives us a good value for model parameters but not optimal which can ve solved by reducing the learning rate at each step which can reduce the bouncing and SGD might settle down at global minimum after some time.

## Difference between Batch Gradient Descent and Stochastic Gradient Descent

| **S.NO.** | **BATCH GRADIENT DESCENT** | **STOCHASTIC GRADIENT DESCENT** |
| --- | --- | --- |
| 1. | Computes gradient using the whole Training sample | Computes gradient using a single Training sample |
| 2. | Slow and computationally expensive algorithm | Faster and less computationally expensive than Batch GD |
| 3. | Not suggested for huge training samples. | Can be used for large training samples. |
| 4. | Deterministic in nature. | Stochastic in nature. |
| 5. | Gives optimal solution given sufficient time to converge. | Gives good solution but not optimal. |
| 6. | No random shuffling of points are required. | The data sample should be in a random order, and this is why we want to shuffle the training set for every epoch. |
| 7. | Can’t escape shallow local minima easily. | SGD can escape shallow local minima more easily. |
| 8. | Convergence is slow. | Reaches rthe convergence much faster. |

##### *1. What Are the Basic Assumption?(favourite)*

There are four assumptions associated with a linear regression model:

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

##### *2. Advantages*

1. Linear regression performs exceptionally well for linearly separable data
2. Easy to implement and train the model
3. It can handle overfitting using dimensionlity reduction techniques and cross validation and regularization

##### *3. Disadvantages*

1. Sometimes Lot of Feature Engineering Is required
2. If the independent features are correlated it may affect performance
3. It is often quite prone to noise and overfitting

##### *4. Whether Feature Scaling is required?*

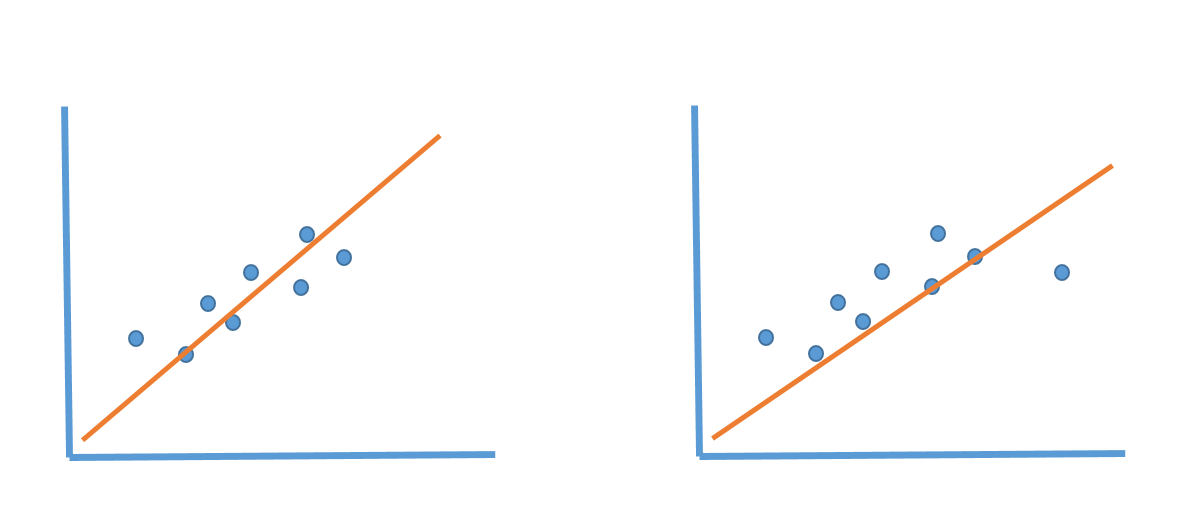
Yes

##### *5. Impact of Missing Values?*

It is sensitive to missing values

##### *6. Impact of outliers?*

linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects.

****

##### *Types of Problems it can solve(Supervised)*

1. Regression

##### *Overfitting And Underfitting*

HomeWork?

##### *Different Problem statement you can solve using Linear Regression*

1. Advance House Price Prediction
2. Flight Price Prediction