**Linear Regression Questions**

1. Basic Assumptions of a Linear Regression Model.
2. How do we test these assumptions?
3. What is multi-collinearity? How does it affect a regression model? How to

spot ?

1. multicollinearity in a regression model?
2. How do we solve for multi-collienarity?
3. how to deal with outliers (e.g., RANSAC)
4. What is regression analysis?
5. What do coefficient estimates mean?
6. How do you measure fit of the model? What do R and D mean?
7. What is Ordinary Least Squares?
8. What are some possible problems with regression models? How do you

avoid or compensate for them?

1. Why do the residuals from a linear regression add up to 0?
2. Is this still true if you fit a regression without intercept?
3. We first regress YY on X1X1 and X2X2, then regress YY on X1X1 and ZZ,

where Z=X1−X2Z=X1−X2. How are the coefficients in the two regressions

related?

1. What are the conditions for Omitted Variable Bias and how does it affect the

coefficient estimates? Why? What are some fixes for OVB?

1. How do you interpret the coefficients in a log-log model? Why?
2. What are the downfalls of using too many or too few variables?
3. What is overfitting a regression model? What are ways to avoid it?  
    In linear regression, under what condition R^2 always equals a perfect 1?
4. once given a regression output and asked to explain it in technical and lay

terms.

1. **Why data cleaning plays a vital role in analysis?**
2. **Differentiate between univariate, bivariate and multivariate analysis.**
3. **What do you understand by the term Normal Distribution?**

**ANSWER** : Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up. However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell shaped curve. The random variables are distributed in the form of an symmetrical bell shaped curve.

1. **Do gradient descent methods always converge to same point?**
2. **What is Gradient Descent?**
3. **What are various steps involved in an analytics project?**

**ANSWER:**

* 1. Understand the business problem
  2. Explore the data and become familiar with it.
  3. Prepare the data for modelling by detecting outliers, treating missing values,

transforming variables, etc.

* 1. After data preparation, start running the model, analyse the result and tweak

the approach. This is an iterative step till the best possible outcome is achieved.

* 1. Validate the model using a new data set.
  2. Start implementing the model and track the result to analyse the performance

of the model over the period of time.

1. **During analysis, how do you treat missing values?**

**ANSWER:** The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.There are various factors to be considered when answering this question-

Understand the problem statement, understand the data and then give the answer.Assigning a default value which can be mean, minimum or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.

If you have a distribution of data coming, for normal distribution give the mean value.

Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

1. **What is the difference between squared error and absolute error?**
2. **How do you decide whether your linear regression model fits the data?**
3. **How are confidence intervals constructed and how will you interpret them?**
4. On a linear regression, if one new variable is added, and that variable is

highly correlated to another, will the t-test statistic will go up or down?