**Interview Questions and Answers**

**Topic: Linear Regression**

**Q1. 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.

**Q2. What are assumptions in a linear regression model?**

The assumptions of linear regression are:

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, i.e., the error terms are normally distributed around 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.

*Explanations:*

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.

**Q3. 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 the data should be homoscedastic, i.e., 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)](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem). 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.

**Q4. 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 the case of multivariate linear regression, two-dimensional pairwise scatter plots, rotating plots, and dynamic graphs can be plotted.

**Q5. 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 the 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.

The Null and Alternate Hypothesis used in the case of linear regression, respectively, are:

β1=0

β1≠0

Thus, if we reject the Null hypothesis, we can say that the coefficient β1 is not equal to zero and hence, is significant for the model. On the other hand, if we fail to reject the Null hypothesis, it is concluded that the coefficient is insignificant and should be dropped from the model.

**Q6. How do you interpret a linear regression model?**

A linear regression model is quite easy to interpret. The model is of the following form:

                                               y=β0+β1X1+β2X2+...+βnXn

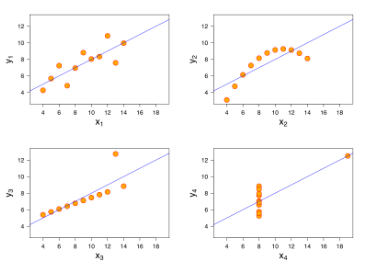
The significance of this model lies in the fact that one can easily interpret and understand the marginal changes and their consequences. For example, if the value of *x0* increases by 1 unit, keeping other variables constant, the total increase in the value of *y* will be *βi*. Mathematically, the intercept term (*β0*) is the response when all the predictor terms are set to zero or not considered.

**Q7. What are the shortcomings of linear regression?**

You should never just run a regression without having a good look at your data because simple linear regression has quite a few shortcomings:

1. It is sensitive to outliers
2. It models the linear relationships only
3. A few assumptions are required to make the inference

These **phenomena** can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So we should never ever run a regression without having a good look at our data.

**Q8. What parameters are used to check the significance of the model and the goodness of fit?**

To check if the overall model fit is significant or not, the primary parameter to be looked at is the **F-statistic**. While the t-test along with the p-values for betas test if each coefficient is significant or not individually, the F-statistic is a measure that can determine whether the overall model fit with all the coefficients is significant or not.

The basic idea behind the F-test is that it is a relative comparison between the model that you've built and the model without any of the coefficients except for β0. If the value of the F-statistic is high, it would mean that the Prob(F) would be low and hence, you can conclude that the model is significant. On the other hand, if the value of F-statistic is low, it might lead to the value of Prob(F) being higher than the significance level (taken 0.05, usually) which in turn would conclude that the overall model fit is insignificant and the intercept-only model can provide a better fit.

Apart from that, to test the goodness or the extent of fit, we look at a parameter called **R-squared** (for simple linear regression models) or **Adjusted R-squared** (for multiple linear regression models). If your overall model fit is deemed to be significant by the F-test, you can go ahead and look at the value of R-squared. This value lies between 0 and 1, with 1 meaning a perfect fit. A higher value of R-squared is indicative of the model being good with much of the variance in the data being explained by the straight line fitted. For example, an R-squared value of 0.75 means that 75% of the variance in the data is being explained by the model. But it is important to remember than R-squared only tells the extent of the fit and should not be used to determine whether the model fit is significant or not.

**Q9. If two variables are correlated, is it necessary that they have a linear relationship?**

No, not necessarily. If two variables are correlated, it is very much possible that they have some other sort of relationship and not just a linear one.

But the important point to note here is that there are two correlation coefficients that are widely used in regression. One is the Pearson's R correlation coefficient which is the correlation coefficient you've studied in the linear regression model. This correlation coefficient is designed for linear relationships and it might not be a good measure for if the relationship between the variables is non-linear. The other correlation coefficient is Spearman's R which is used to determine the correlation if the relationship between the variables is not linear. So even though, Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. For example, the correlation coefficients as given by both the techniques for the relationship y=X3 for 100 equally separated values between 1 and 100 were found out to be:

                                                                  Pearson′s R≈0.91

                                                                   Spearman′s R≈1

And as we keep on increasing the power, the Pearson's R value consistently drop whereas the Spearman's R remains robust at 1. For example, for the relationship y=X10 for the same data points, the coefficients were:

                                                                  Pearson′s R≈0.66

                                                                    Spearman′s R≈1

So the takeaway here is that if you have some sense of the relationship being non-linear, you should look at Spearman's R instead of Pearson's R. It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

**Q10. What is the difference between Least Square Error and Mean Square Error?**

Least Square Error is the method used to find the best-fit line through a set of data points. It is. The idea behind the least squared error method is to minimize the square of errors between the actual data points and the line fitted.

Mean Square Error, on the other hand, is used once you have fitted the model and want to evaluate it. So the mean squared error finds out the average of the difference between the actual and predicted values and hence, is a good parameter to compare various models on the same data set.

Thus, LSE is a method used to minimise the sum of squares and is used during model fitting, and MSE is a metric used to evaluate the model after fitting based on the average squared errors.

**Topic: Multiple Linear Regression**

**Q1. What is Multicollinearity? How does it affect the linear regression? How can you deal with it?**

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.

**Q2. How can you handle categorical variables present in the dataset?**

Many a time it might happen that your dataset has categorical variables that might be a potentially good predictor for the response variable. So handling them right is quite crucial.

One of the ways to handle categorical data with just two levels is to do a binary mapping of the variables wherein one of the levels will correspond to zero and the other to 1.

Another way of handling categorical variables with few levels is to perform a dummy encoding. The key idea behind dummy encoding is that for a variable with, say, 'N' levels, you create 'N-1' new indicator variables for each of these levels. So for a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', you would create a dummy table like the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Relationship Status | Single | In a relationship | Married |
| Single | 1 | 0 | 0 |
| In a relationship | 0 | 1 | 0 |
| Married | 0 | 0 | 1 |

But you can clearly see that there is no need of defining **three** different levels. If you drop a level, say 'Single', you would still be able to explain the three levels.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:

|  |  |  |
| --- | --- | --- |
| Relationship Status | In a relationship | Married |
| Single | 0 | 0 |
| In a relationship | 1 | 0 |
| Married | 0 | 1 |

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

Now, creating dummy variables might be useful when the number of levels in a categorical variable is small, but if a categorical variable has a hundred levels, it is clearly impossible to create 99 new variables. In such cases, grouping the variables might be useful. For example, for the variable “Cities in India”, you can use a geographical grouping, i.e.:

* Keep the 'n' largest cities, group the rest
* Geographical hierarchy
  + City < District < State < Zone
* Group cities with the similar value for the outcome variable
* Cluster cities with similar values for the predictor variables

Another way to deal with categorical variables is that you can perform a [One-hot encoding](https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f) which hasn't been covered in our syllabus.

**Q3. What is the major difference between R-squared and Adjusted R-squared/Why is it advised to use Adjusted R-squared in case of multiple linear regression?**

The major difference between R-squared and Adjusted R-squared is that R-squared doesn't penalise the model for having more number of variables. Thus, if you keep on adding variables to the model, the R-squared will always increase (or remain the same in the case when the value of correlation between that variable and the dependent variable is zero). Thus, R-squared assumes that any variable added to the model will increase the predictive power.

Adjusted R-squared on the other hand, penalises models based on the number of variables present in it. Its formula is given as:

                                                             Adj. R2=1−(1−R2)(N−1)N−k−1

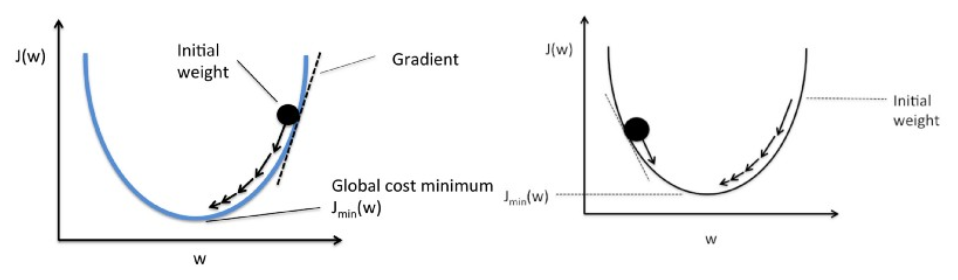
*where 'N' is the number of data points and 'k' is the number of features*

So if you add a variable and the Adjusted R-squared drops, you can be certain that that variable is insignificant to the model and shouldn't be used. So in the case of multiple linear regression, you should always look at the adjusted R-squared value in order to keep redundant variables out from your regression model.

**Q4. 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.

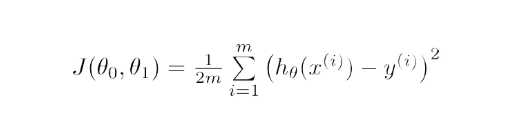
Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



**Gradient Descent**

Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin J(Θ*0*,Θ*1*), where J(Θ*0*,Θ*1*) is the cost function of the linear regression. It is given by —

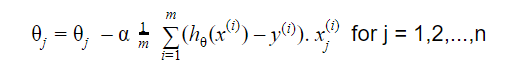


Here, *h* is the linear hypothesis model, h=Θ*0* + Θ1x, *y* is the true output, and *m* is the number of data points in the training set.

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



To read more about gradient descent, refer to the additional resources on linear regression here ([Link](https://learn.upgrad.com/v/course/1120/session/129888/segment/711313)).

**Q4. 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—

                                                                                    VIFi=11−Ri2

Here, VIFi  is the value of VIF for the ith variable,  Ri2  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.

**Q5. 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.