Q: What are the assumptions associated with a linear regression model?

A1: There are four assumptions associated with a linear regression model:

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

Q: What is Ordinary Least Squares in regression?

A1: Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent variables and a dependent variable. In other words, you can consider the OLS as a strategy to obtain, from your model, a ‘straight line’ which is as close as possible to your data points. Even though OLS is not the only optimization strategy, it is the most popular for regression tasks.

A2: Ordinary least squares (OLS), estimates the parameters in a regression model by minimizing the sum of the squared residuals. This method draws a line through the data points that minimizes the sum of the squared differences between the observed values and the corresponding fitted values.

A3: Ordinary Least Squares regression (OLS) often called “Linear Regression”. It’s a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent variables and a dependent variable.

A4: Bir tahmin yaparken eldeki verilerin doğrusal bir çizgi etrafında olmasını isteriz. Eldeki verilerin çizilecek olan bu doğruya olan uzaklıklarının karelerinin toplamının minimum olması daha sonra yapılacak tahminlerin daha doğru sonuç vermesini sağlayacaktır.

A5: Error is the difference between prediction and reality: the vertical distance between a real data point and the regression line. OLS is concerned with the squares of the errors. It tries to find the line going through the sample data that minimizes the sum of the squared errors.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Q: Günün Sorusu: What is bias in machine learning and why is it important?

A1: Bias can be best described as the difference between the actual prediction of our model to the correct value trying to predict. A model with high bias might pay very little attention to the training data and hence it rather oversimplifies the model. Such a model will always lead to high errors in training and test data. Bias is a phenomenon that completely skews the result of an algorithm in favor of or against an idea. Bias is considered to be a systematic error that occurs in the machine learning model automatically due to incorrect assumptions in the ML process.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

***If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to “discover” it. In general, data science becomes useful when we have a large number of data examples and when the patterns are too complex for humans to discover and extract manually.*** As a lower bound, we can take a large number of data examples to be defined as more than a human expert can check easily. With regard to the complexity of the patterns, again, we can define it relative to human abilities. We humans are reasonably good at defining rules that check one, two, or even three attributes (also commonly referred to as features or variables), but when we go higher than three attributes, we can start to struggle to handle the interactions between them. By contrast, data science is often applied in contexts where we want to look for patterns among tens, hundreds, thousands, and, in extreme cases, millions of attributes.

*-- Kelleher, John D., and Brendan Tierney. Data science. MIT Press, 2018.*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Q: What is bias variance trade-off in machine learning?

A1: If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.