Linear Regression Pitfalls

Comprehension

When we fit a linear regression model to a particular data set, many problems may arise. Most common among these are the following:

1. Non-constant variance
2. Autocorrelation and time series issue
3. Multicollinearity
4. Overfitting
5. Extrapolation

In this segment, you will learn about each of these problems, and we will also discuss the methods for overcoming some of these pitfalls.

1. **Non-constant variance**

Constant variance of error terms is one of the assumptions of linear regression. Unfortunately, many times, we observe non-constant error terms. As discussed earlier, as we move from left to right on the residual plots, the variances of the error terms may show a steady increase or decrease. This is also termed as heteroscedasticity.

When faced with this problem, one possible solution is to transform the response Y using a function such as log or the square root of the response value. Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.

2. **Autocorrelation**

This happens when data is collected over time and the model fails to detect any time trends. Due to this, errors in the model are correlated positively over time, such that each error point is more similar to the previous error. This is known as autocorrelation, and it can sometimes be detected by plotting the model residuals versus time. Such correlations frequently occur in the context of time series data,

which consists of observations for which measurements are obtained at discrete points in time.

In order to determine whether this is the case for a given data set, we can plot the residuals from our model as a function of time. If the errors are uncorrelated, then there should be no observable pattern. However, on the other hand, if the consecutive values appear to follow each other closely, then we may want to try an autoregression model.

3. **Multicollinearity**

If two or more of the predictors are linearly related to each other when building a model, then these variables are considered multicollinear. A simple method to detect collinearity is to look at the correlation matrix of the predictors. In this correlation matrix, if we have a high absolute value for any two variables, then they can be considered highly correlated. A better method to detect multicollinearity is to calculate the variance inflation factor (VIF), which you studied in the Linear Regression module.

When faced with the problem of collinearity, we can try a few different approaches. One is to drop one of the problematic variables from the regression model. The other is to combine the collinear variables together into a single predictor. Regularization (which we will discuss in the next session) helps here as well.

4. **Overfitting**

When a model is too complex, it may lead to overfitting. It means the model may produce good training results but would fail to perform well on the test data. One possible solution for overfitting is to increase the amount and diversity of the training data. Another solution is regularization, which we will cover in the next session.

5. **Extrapolation**

Extrapolation occurs when we use a linear regression model to make predictions for predictor values that are not present in the range of data used to build the model. For instance, suppose we have built a model to predict the weight of a child given its height, which ranges from 3 to 5 feet. If we now make predictions for a child with height greater than 5 feet or less than 3 feet, then we may get incorrect predictions. The predictions are valid only within the range of values that are used for building the model. Hence, we should not extrapolate beyond the scope of the model.