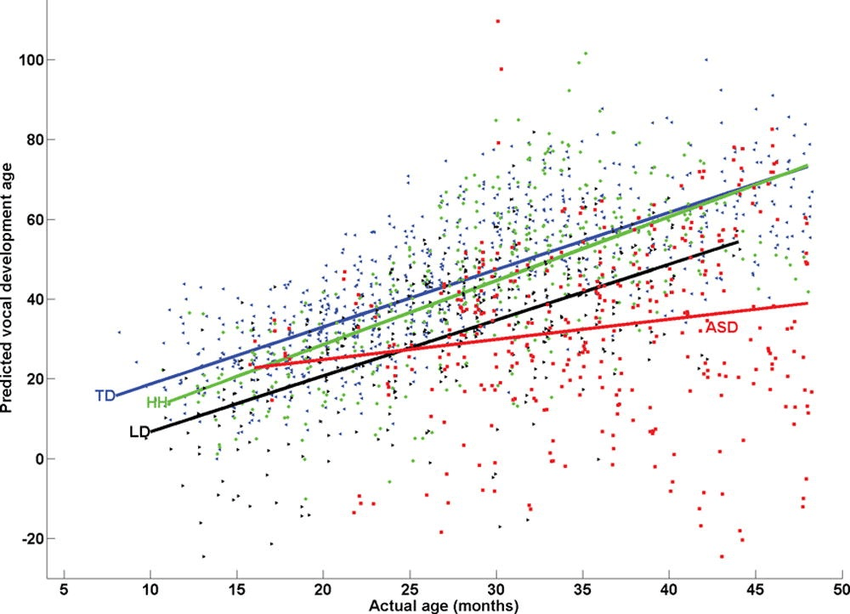
**What is Multiple Linear Regression?**

Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while the variables we use to predict the value of the dependent variable are known as independent or explanatory variables.



**Summary**

Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable.

The technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance.

Multiple regression can take two forms, i.e., linear regression and non-linear regression.

**Multiple Linear Regression Formula**



Where:

yi​ is the dependent or predicted variable

β0 is the y-intercept, i.e., the value of y when both xi and x2 are 0.

β1 and β2 are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.

βp is the slope coefficient for each independent variable

ϵ is the model’s random error (residual) term.

**Understanding Multiple Linear Regression**

Simple linear regression enables statisticians to predict the value of one variable using the available information about another variable. Linear regression attempts to establish the relationship between the two variables along a straight line.

Multiple regression is a type of regression where the dependent variable shows a linear relationship with two or more independent variables. It can also be non-linear, where the dependent and independent variables do not follow a straight line.

Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

**Assumptions of Multiple Linear Regression**

Multiple linear regression is based on the following assumptions:

**1. A linear relationship between the dependent and independent variables**

The first assumption of multiple linear regression is that there is a linear relationship between the dependent variable and each of the independent variables. The best way to check the linear relationships is to create scatterplots and then visually inspect the scatterplots for linearity. If the relationship displayed in the scatterplot is not linear, then the analyst will need to run a non-linear regression or transform the data using statistical software, such as SPSS.

**2. The independent variables are not highly correlated with each other**

The data should not show multicollinearity, which occurs when the independent variables (explanatory variables) are highly correlated. When independent variables show multicollinearity, there will be problems figuring out the specific variable that contributes to the variance in the dependent variable. The best method to test for the assumption is the Variance Inflation Factor method.

**3. The variance of the residuals is constant**

Multiple linear regression assumes that the amount of error in the residuals is similar at each point of the linear model. This scenario is known as homoscedasticity. When analysing the data, the analyst should plot the standardized residuals against the predicted values to determine if the points are distributed fairly across all the values of independent variables. To test the assumption, the data can be plotted on a scatterplot or by using statistical software to produce a scatterplot that includes the entire model.

**4. Independence of observation**

The model assumes that the observations should be independent of one another. Simply put, the model assumes that the values of residuals are independent. To test for this assumption, we use the Durbin Watson statistic.

The test will show values from 0 to 4, where a value of 0 to 2 shows positive autocorrelation, and values from 2 to 4 show negative autocorrelation. The mid-point, i.e., a value of 2, shows that there is no autocorrelation.

**5. Multivariate normality**

Multivariate normality occurs when residuals are normally distributed. To test this assumption, look at how the values of residuals are distributed. It can also be tested using two main methods, i.e., a histogram with a superimposed normal curve or the Normal Probability Plot method.