Regression analysis

Regression analysis is a mathematically measured correlation of variables used as a predictive modelling method. You use regression modelling to predict numerical values depending on various inputs. For example, you can understand the relationship between an independent and dependent variable, allowing you to predict how the dependent variable changes along with its independent counterpart. In this case, the dependent variable is what you're measuring and the independent variable is the factor that causes change.

In business, regression analysis can help forecast trends, predict strengths and areas of weakness or establish cause-and-effect relationships to make informed business decisions and strategic plans. You often calculate regression analysis through machine learning or artificial intelligence, though there are also mathematical equations you can use. There are different analysis types that you can use based on the nature of the variables you're predicting and what information you'd like to gather from your analysis.

13 regression types:

There are several types of regression, and deciding which one to use depends on the number of factors involved. This can include the type of variables, the number of independent variables and the regression line shape. Here are 13 types of regression analysis to consider when performing data analysis

1. Simple regression

Simple regression methods help you estimate the relationship between a dependent variable and one independent variable. For example, you might use simple regression to compare the connection between umbrella sales (dependent variable) and the amount of rain a meteorologist forecasts (independent variable). Other examples of this variable relationship might be how much money someone earns based on their level of education or how high lumber prices get during labor shortages.

2. Multiple regression

Multiple regression analysis methods help you determine the relationship between a dependent variable and more than one independent variable. Adding more independent variables makes for a more complex regression analysis study, though it often generates more specific and realistic results. For example, you might evaluate if more umbrellas sell when the meteorologist forecasts rainy weather specifically during spring or comparatively across all seasons. Or you might review salary earnings for education, experience and proximity to a metropolitan area.

3. Linear regression

Linear regression analysis is a simple regression type that requires you to create a hypothetical line that best connects all data points. You determine the best fit line with linear regression and establish a predictor error between the predicted value based on the line and what's actually observed. The disadvantage of linear regression is the potential for outliers in the data so it's frequently used for small data pools of information or predictions. This is because some data points may not fit neatly into the regression line.

4. Multiple linear regression

Similar to linear regression, multiple linear regression shows the direct or linear correlation between variables. The difference is that it involves more than one dependent variable. Even though multiple linear regression may involve more dependent variables, it's also best used for smaller batches of data to prevent accuracy issues with outliers.

5. Logistic regression

Logistic regression helps measure the relationship between dependent and independent variables, though it doesn't correlate between independent variables. You often have a large data set when using logistic regression, and the dependent variable is usually discrete, meaning that you can count all values within a finite amount of time. With logic regression, the target variable typically only has two values, and a sigmoid curve shows the correlation.

6. Ridge regression

Ridge regression is another machine learning analysis you might use when there’s a strong correlation between independent variables. This means that as one independent variable changes, others can change with it. Typically, the least square estimates produce unbiased values, especially with data that's multi-collinear. If the collinear relation is extremely high, the analysis might produce a bias value, or a difference between the expected and true value of a variable.

7. Lasso regression

Lasso regression, or least absolute shrinkage and selection operator (LASSO), uses regularization and objective functions by prohibiting the size of the regression coefficient. It allows it to get closer to zero, unlike ridge regression, and you can pick a set of features from your database to build lasso regression models. Since only required features get used in lasso regression and all other features get marked zero, you can often avoid over fitting the model. Lasso regression also typically requires standardization.

8. Polynomial regression

Similar to multiple linear regression, polynomial regression analysis is modified, though it uses a linear model as an initial estimator. It shows the connection between independent and dependent variables, though they get recognized by the nth degree. Often the best fit line generated by polynomial regression is a curved line, rather than a straight one. Over fitting is a concern when using polynomial regression, so consider analyzing the generated curve toward the end because extrapolation, or the act of estimating a value by assuming that current trends may continue, can create different results.

9. Bayesian linear regression

The Bayes' theorem of probability and statistics is the foundation of the Bayesian linear regression technique. In this calculation, you determine the feature's posterior distribution rather than the least squares. Somewhat similar to the ridge regression and linear regression methods, Bayesian analysis often produces more consistent and stable results to analyze.

10. Jackknife regression

You can use jackknife regression when other methods are highly unlikely to produce estimates. It's a sampling method that omits an observation data point in each sample when calculating an estimate of bias or variance. You then find the average of all these estimated calculations through aggregation.

11. Elastic net regression

You often use elastic net regression when your data set has highly connected independent variables. It doesn't assume normality, similar to ridge and lasso regression analysis. The specific difference of elastic net regression is that it includes regularization penalties in the analysis.

12. Ecological regression

Ecological regression is specific to history or political science fields. It’s used to study predicted human behaviour within a population data set, like geographic locations, ages, ethnicities or levels of education or income. For example, ecological regression studies might focus on voting patterns and behaviour in specific demographics to help political candidates or parties prepare for upcoming campaigns or elections.

It's important to note this kind of regression analysis can sometimes create an ecological fallacy, where a researcher assumes that a phenomenon they witness or observe in groups holds true for all individuals. For example, if they find that populations with more cholesterol and fat in their diet have higher rates of diabetes, the researcher would assume the people who eat higher fat content foods may get a diabetes diagnosis in their lifetime. This is a fallacy because it can ignore individual factors and assume that all members of a group are the same.

13. Stepwise regression

Stepwise regression often gets constructed step by step, resulting in its name. You add or remove predictor variables one at a time depending on the criteria and test the significance for each version. This is different from multiple regression which considers all predictor variables at once. Within the stepwise regression method, there’s a backward or forward method, though both are often best used during exploratory research because predictions sometimes lack accuracy. More often, data analysts or statisticians might use stepwise regression techniques to test or confirm their suspicions before performing other regression tests.