REGRESSION

Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

The two basic types of regression are linear regression and multiple linear regressions, although there are non-linear regression methods for more complicated data and analysis. Linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple regressions use two or more independent variables to predict the outcome.

Regression can help finance and investment professionals as well as professionals in other businesses. Regression can also help predict sales for a company based on weather, previous sales, GDP growth or other types of conditions. The capital asset pricing model (CAPM) is an often-used regression model in finance for pricing assets and discovering costs of capital.

The general form of each type of regression is:

- Linear regression: Y = a + bX + u

- Multiple regression: Y = a + b1X1 + b2X2 + b3X3 + … + btXt + u

Where:

Y = the variable that you are trying to predict (dependent variable).

X = the variable that you are using to predict Y (independent variable).

a = the intercept.

b = the slope.

u = the regression residual.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using numbers with subscripts.

Assumptions in Regression

Regression is a parametric approach. ‘Parametric’ means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It fails to deliver good results with data sets which doesn’t fulfill its assumptions. Therefore, for a successful regression analysis, it’s essential to validate these assumptions.

So, how would you check (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

Let’s look at the important assumptions in regression analysis:

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.

2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.

3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.

5. The error terms must be normally distributed.

What if these assumptions get violated?

Let’s dive into specific assumptions and learn about their outcomes (if violated):

1. Linear and Additive: If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

How to check: Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms (X, X^2, X^3) in your model to capture the non-linear effect.

2. Autocorrelation: The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients. Let's understand narrow prediction intervals with an example:

For example, the least square coefficient of X^1 is 15.02 and its standard error is 2.08 (without autocorrelation). But in presence of autocorrelation, the standard error reduces to 1.20. As a result, the prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

How to check: Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, $0 < DW < 2$ implies positive autocorrelation while $2 < DW < 4$ indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

3. Multicollinearity: This phenomenon exists when the independent variables are found to be moderately or highly correlated. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

Also, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on which other predictors are available in the model. If this happens, you'll end up with an incorrect conclusion that a variable strongly / weakly affects target variable. Since, even if you drop one correlated variable from the model, its estimated regression coefficients would change. That's not good!

How to check: You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

4. Heteroskedasticity: The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

How to check: You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

5. Normal Distribution of error terms: If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares. Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.

How to check: You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

Limitations:

- It is assumed that the cause and effect relationship between the variables remains unchanged. This assumption may not always hold good and hence estimation of the values of a variable made on the basis of the regression equation may lead to erroneous and misleading results.

- The functional relationship that is established between any two or more variables on the basis of some limited data may not hold good if more and more data are taken into consideration. For example, in case of the Law of Return, the law of diminishing return may come to play, if too much of inputs are used with ca view to increasing the volume of output.

- It involves very lengthy and complicated procedure of calculations and analysis.
- It cannot be used in case of qualitative phenomenon viz. honesty, crime etc.

Regression Line

Definition: The Regression Line is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest. In other words, a line used to minimize the squared deviations of predictions is called as the regression line.

There are as many numbers of regression lines as variables. Suppose we take two variables, say X and Y, then there will be two regression lines:

- **Regression line of Y on X:** This gives the most probable values of Y from the given values of X.
- **Regression line of X on Y:** This gives the most probable values of X from the given values of Y.

The algebraic expression of these regression lines is called as Regression Equations. There will be two regression equations for the two regression lines.

The correlation between the variables depend on the distance between these two regression lines, such as the nearer the regression lines to each other the higher is the degree of correlation, and the farther the regression lines to each other the lesser is the degree of correlation.

The correlation is said to be either perfect positive or perfect negative when the two regression lines coincide, i.e. only one line exists. In case, the variables are independent; then the correlation will be zero, and the lines of regression will be at right angles, i.e. parallel to the X axis and Y axis.

Note: The regression lines cut each other at the point of average of X and Y. This means, from the point where the lines intersect each other the perpendicular is drawn on the X axis we will get the mean value of X. Similarly, if the horizontal line is drawn on the Y axis we will get the mean value of Y.

Advantages of Linear Regression

1. Linear Regression performs well when the dataset is **linearly separable**. We can use it to find the nature of the relationship among the variables.

2. Linear Regression is easier to implement, interpret and very efficient to train.

3. Linear Regression is prone to over-fitting but it can be easily avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages of Linear Regression

1. Main limitation of Linear Regression is the **assumption of linearity** between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.

2. **Prone to noise and over fitting:** If the number of observations are lesser than the number of features, Linear Regression should not be used, otherwise it may lead to over fit because is starts considering noise in this scenario while building the model.

3. **Prone to outliers:** Linear regression is very sensitive to outliers (anomalies). So, outliers should be analyzed and removed before applying Linear Regression to the dataset.

4. **Prone to multicollinearity:** Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables.

In summary, Linear Regression is great tool to analyze the relationships among the variables but it isn’t recommended for most practical applications because it over-simplifies real world problems by assuming linear relationship among the variables