10 Linear Regression

In business, financial markets or policy formulation, the key to decision making lies in understanding the relationship between two or more variables. For instance, firms would be interested in the relationship between the demand for a product at different price points.; policymakers would be interested in the link between say, GDP growth rates and unemployment rates in an economy; in finance, investors or traders would want to identify the factors that drive capital appreciation of stocks.
Further, these firms, policymakers and investors/traders would want to forecast the future paths of the phenomena they are observing with a reasonable degree of confidence. Which brings us to regression.
In regression, a function is created which helps in analyzing the relationship between variables and also in predicting one variable based on its relationship with other variables. The variable which is predicted is known as the **dependent variable** and is generally denoted by Y, and the variables which have an impact on the dependent variable are known as **independent variables** and are generally denoted by X_1, X_2... and so on.

For the scope of this primer we will only look at **linear regression**, where we seek to model the linear relationship between a dependent variable Y and independent variables (X_is).

10.1 Simple Linear Regression

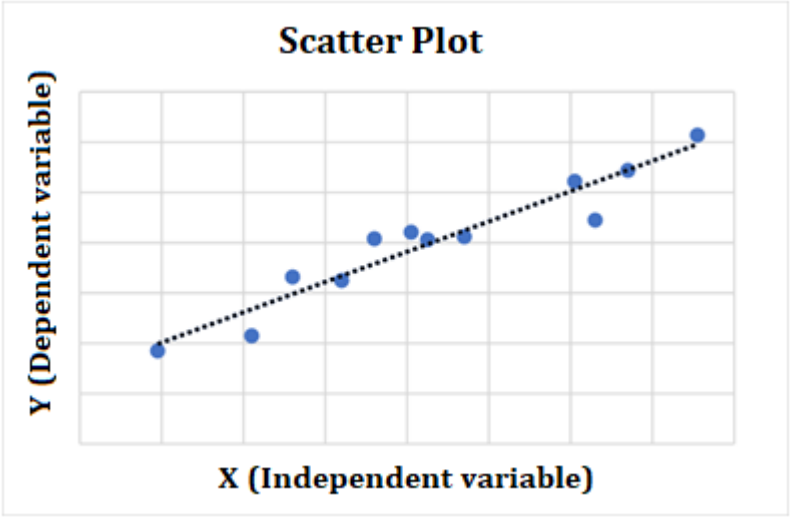
Simple linear regression is a special case of linear regression when we have only one independent variable. In this case, we work with the assumption that a linear function satisfies and best depicts the relationship between two variables.
"Linear" means that the function we are looking for is a straight line (so our function f will be of the form $f(x) = mx + c$ for constants m and c). Of course, the points might not fit the function exactly but the aim is to get as close as possible.

Using simple linear regression, we aim to evaluate the extent to which changes in one variable (referred to as the dependent variable) can be explained by variations in one or more additional variables (known as the independent variables).

Important point : The term 'linear' has implications for both independent variable(s) and the unknown parameters (i.e, the coefficients). A linear regression function assumes that the relation being modeled must be linear in the coefficients but not necessarily the variables (Example : Polynomial Regression is a Linear Regression) .
You can read more about it here .

10.2 Scatter Plot for a Simple Linear Regression

One of the best ways to visually depict the relationship between two variables is by using a **scatter plot**. Scatter plot is used to visually check for the presence or absence of a linear relationship between two variables. Hence, it is quite often used in preliminary analysis before we actually build a linear regression model.
A simple scatter plot is shown in the following figure:

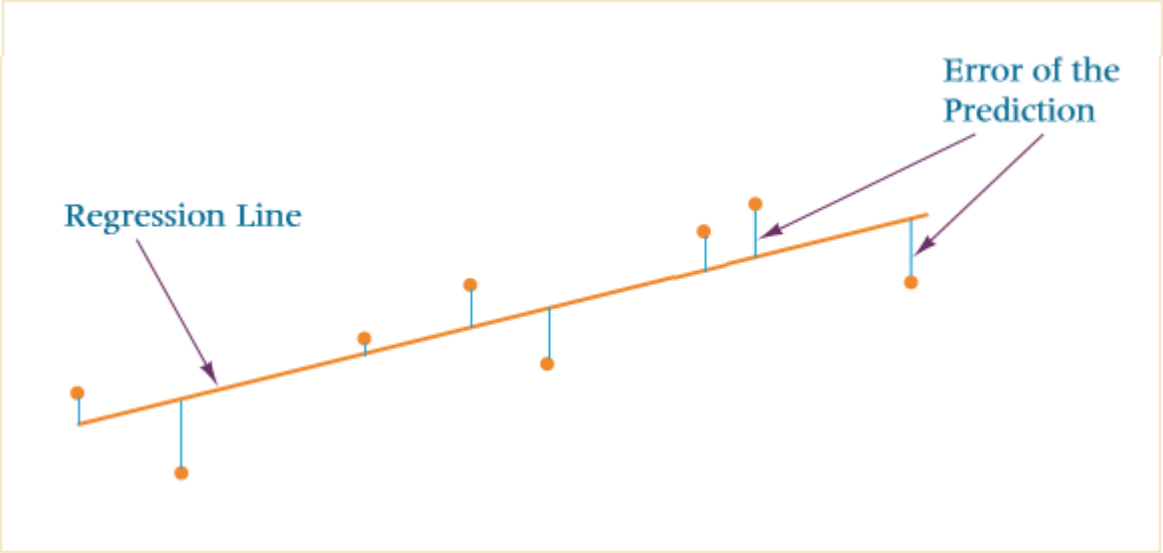


In this plot, blue dots represent the different points (X, Y) where Y is the dependent variable and X is the independent variable. As most points lie around the black line, it is an indication that a linear relationship might indeed exist between X and Y. The black line is also called the **regression line** and is estimated using a method called ordinary least squares (OLS), which we discuss in the next section.

10.3 Ordinary Least Squares (OLS) Method

The linear model we are looking for is of the form:
$Y_i = \alpha + \beta x_i + \varepsilon_i$
where, $\alpha + \beta$ xi is the "regression function "and εi is the "error" term.
β = the regression or slope coefficient; sensitivity of Y to changes in X
α = value of Y when X = 0
ε = random error or shock ; unexplained component of Y (by X)

This error may be reduced by using more independent variables or by using different, more appropriate independent variables.
The regression function contains all information regarding the two variables while the error term is the error in prediction which is represented by blue lines in the following figure:



The regression line mentioned in the previous section is obtained by using the method of ordinary least squares i.e. we find the optimal values for **α** and **β** ,that minimize the sum of all the squared errors i.e. $\sum \varepsilon i^2$. The good news is that these days all of this is automatically done by computers using Excel, Python and many others.

Consider the following dataset

X	2	4	6	8
Y	3	7	5	10

You want to find the relationship between X and Y. We estimate the coefficients (not going into details of formula and derivation as this can be done directly using Excel or Python) $\beta = 0.95$ and $\alpha = 1.5$. So the relationship can be expressed as : $Y = 1.5 + 0.95X$

Now, if we plug in the values of X in the above equation, we will get predicted values of Y. These values of Y will differ from the actual values of Y and this difference between actual values of Y and predicted values of Y, **ε** is known as the error term.
Using OLS Method , **α** and **β** are predicted in such a way that sum of these squared residuals(i.e, error terms) is minimized.

10.4 Assessing the Fit of the Regression Model

Once the values for **α** and **β** are obtained, we have the model. However, any software that we use to build the regression model will also give us a few metrics as output. These tell us about how well the model fits our data. Some of these metrics are:
1) R^2 **(Coefficient of Determination)**: R^2 is the percentage of variation in the data that is explained by our model. Its value lies in the range of 0 to 1. Higher the value of R^2, higher the accuracy of the model. But, the problem with $R^2$is that its value always increases on addition of new independent variables to the model, which may lead to a complicated and lengthy model.

Word of caution : Though a higher value of R^2 indicates that the model properly explains the variation in independent variable, a very high value of R^2 is also possible due to overfitting of the model on the training data(i.e, when we use this model for prediction on a new dataset the predictions may be worse).

2) **Adjusted** R^2: Adjusted R^2 can be used in place of R^2 to determine the best fitting model as its value does not increase on the addition of new variables.

3) **Standard errors for α and β:** The standard error determines level of variability associated with the estimated values of variables (**α** and **β** in this case). Accuracy of the model is more when the standard errors are less.

4). **F statistic**: F statistic can evaluate the overall significance of the model in one go. It is represented as the ratio of explained variance to unexplained variance. Value of F statistics can range from zero to arbitrarily any number. Higher the value of F statistic, better the model.

5) **t-stats and p-values**: t-stats here are the statistic values for a hypothesis tests conducted using the t-statistic for each estimated coefficient and the intercept. Here the test is to find whether the estimated values of **α** and **β** are significantly different from zero on an individual level. The corresponding p-values for the tests are also reported. Generally, a model with p-values less than 0.05 is desirable as it indicates that the estimated values for **α** and **β** are reliable.

10.5 Multiple Linear Regression

The ideas discussed above can be extended to the case where we have more than one explanatory/independent variable. But in that case we have to estimate multiple coefficients for each explanatory variable.
The regression equation for multiple linear regression is of the form:
$Y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \ldots\ldots\ldots\beta nXn + e$
Where,
Y = Dependent variable
X1, X2, X3…….Xn = Independent variables
β0 = Y-intercept
β1, β2…..βn = coefficients of respective independent variables
βi is interpreted as the increment in Y for a unit change in Xi, holding constant the impact of all other independent variables. This is why slope coefficients in multiple regression are sometimes called partial slope coefficients.

10.6 Assumptions of Linear Regression

Regression is a technique, that makes certain assumptions about the data for its analysis. Without fulfilling the assumptions, regression will make biased or erratic predictions. Some of the assumptions of regression and the ways to test them are explained as follows:

a). Linear relationship: dependent and independent variables should be linearly related to each other.

b). No multicollinearity: Correlations among independent variables must be negligible. If the independent variables will be correlated, then the model would not be able to identify the true effect of independent variables on the dependent variable.

- ***Test for multicollinearity***: Variance Inflation Factor (VIF) is used to check multicollinearity. VIF value ≤ 4 shows no multicollinearity, while VIF ≥ 10 implies serious multicollinearity.

c). No heteroskedasticity: Error terms in regression must have a constant variance, which is known as homoskedasticity. Errors with the presence of non-constant variance lead to heteroskedasticity. Non-constant variance is mainly due to the presence of outliers.

- ***Test for heteroskedasticity***: Cook-Weisberg test, White general test

d). **Normal distribution of errors**: Error terms must be normally distributed. Non-normal distribution of errors shows that the model has few unusual data points which need to be rectified to make a better performing model.

- ***Test for normality of errors:*** Kolmogorov-Smirnov test, Shapiro-Wilk test

e). **No autocorrelation in error terms**: Error terms or residuals should not be correlated. Accuracy of the model is reduced significantly due to the presence of correlation in error terms. Absence of correlation among errors is termed as autocorrelation.

- ***Test for autocorrelation***: Durbin-Watson statistic (DW). Positive autocorrelation means value of DW in the range of 0 to 2. If DW $=2$, then no autocorrelation, while $2<DW<4$ implies negative autocorrelation.

10.7 An Example

Sophie is looking to design a trading strategy to trade in stock XYZ. She has observed that stock XYZ always moves when the price of oil fluctuates. She wants to understand and quantify this relationship as she wants to find out whether the price of oil is a factor that influences the movement of stock XYZ. She obtains the data for the past one year (252 data points) for both the variables.

Sophie uses a computer program to build a simple linear regression model using the data she has obtained and gets the following output:

SUMMARY OUTPUT					
Regression Statistics					
R Square	0.5766				
Adjusted R Square	0.5071				
Standard Error	3.24728502				
Observations	252				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	385.7503885	385.7504	36.58184	5.30726E-09
Residual	250	2636.215006	10.54486		
Total	251	3021.965395			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	34.493735	3.79463175	9.09014	3.14E-17	
X Variable 1	0.37185923	0.061481689	6.048292	5.31E-09	

Let us interpret this result.

As the R square value is 0.5766, the model is able to explain almost 58% of the variation i.e. the price of oil (X) is able to explain a high degree of variation in the price of stock XYZ.
In the last table, we can see the actual estimated values for the intercept and the slope. Thus, the model is: **(Price of XYZ) = 34.49373 + 0.3718 *(Price of Oil)**

Also, the p-values for both the intercept and the slope are very small (<0.05), leading to the statistical significance of the values obtained. The small 'Significance F' value in the second table also indicates to the same conclusion i.e. the overall model is significant.

Thus, being a data driven trader, Sophie decides to use the price of oil as an important variable while designing a strategy to trade stock XYZ.