*Linear Regression Assumptions*

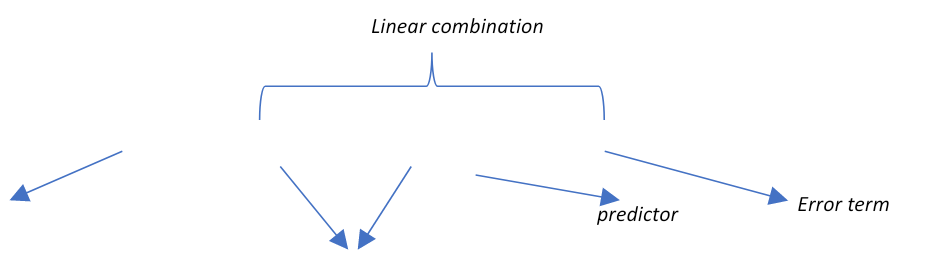
Regression is one of the very first algorithms that people come across when they take a deep dive into the world of Data Science and Machine Learning. If you try hard to remember, we have been acquainted with it long before and long enough in our high school days. Yes!! You heard that right, and I am going to tell you a secret. It is the same straight-line equation that we came across in Coordinate Geometry which we dearly hated it because of the vast number of equations. Kudos!! You got it. The only nuance in it is that the same straight-line becomes a plane and other higher dimensional object when we consider a higher dimension space.

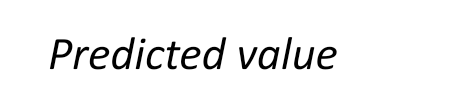
This simple to understand yet most advanced algorithm(or equation) was built long ago by mathematicians and statisticians based on a few underlying assumptions. Here I will be going to highlight those assumptions. The assumptions play a key role in modelling the data in hand, as it tends to fail if it does not satisfy the assumptions. It is of utmost importance that these assumptions are taken care of.

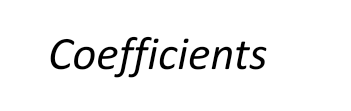
A close up of a map

Description automatically generated

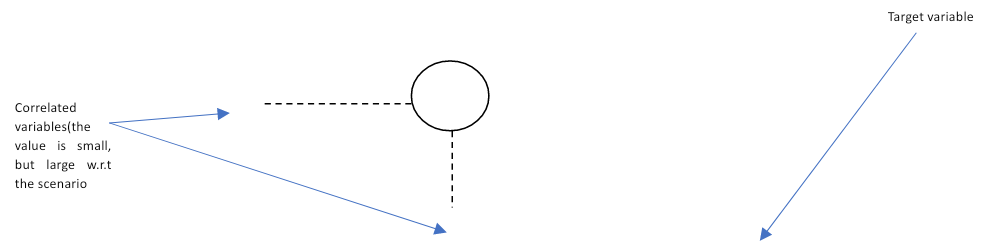
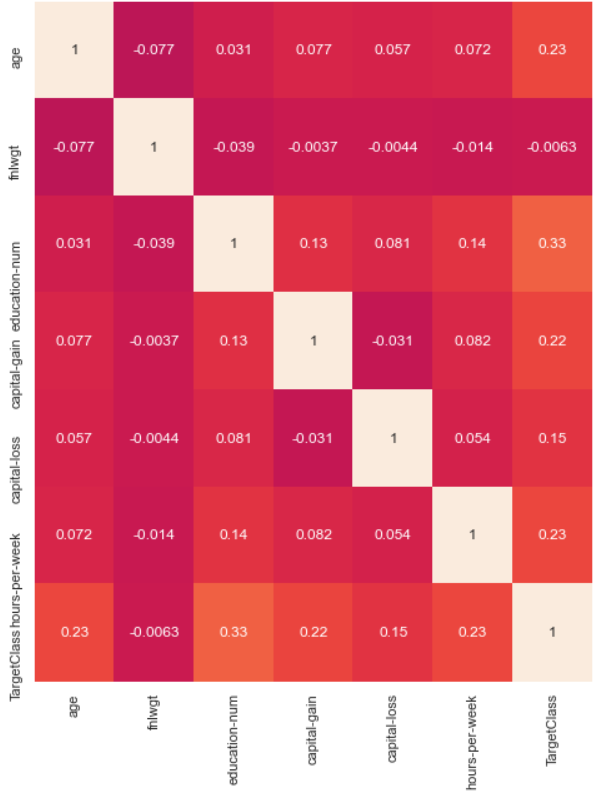
*The coefficients of this fitted line are b0 = -10.460 and b\_1 = 1.1696*

1. Linearity: Exactly as it sounds. The equation should be linear, i.e. additive terms only. Now all “*data enthusiasts*” will say, ‘but there are multiplicative terms too’. Yes, there are multiplicative terms, and those are just multiplication of the coefficients and the data points matrix. But to determine the coefficients, the equation should be in a *linear* fashion. I shall keep the determination of coefficients ,by techniques like gradient descent and maximum likelihood estimation, for another blog post.



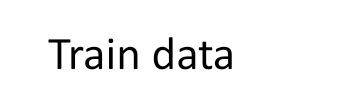
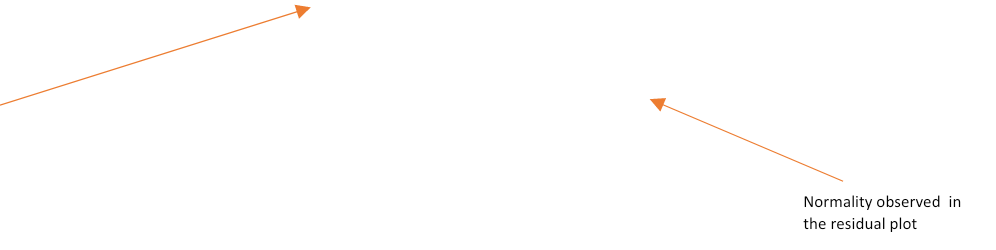


1. Multicollinearity: Multicollinearity occurs when several *x* terms or independent terms are closely related to each other. We tend to forget this essential concept as we can only visualize a single dimension equation where there is a single dependent and independent variable, but if there are multiple independent variables as it happens to be in most cases, there should be nearly *no multicollinearity between the independent terms*. The problem of multicollinearity is, it inflates the variance of the coefficients. We can measure this inflation by a factor called VIF(variance inflation factor). Let us keep aside VIF and the techniques to remove multicollinearity as of now and jump to the next assumption.



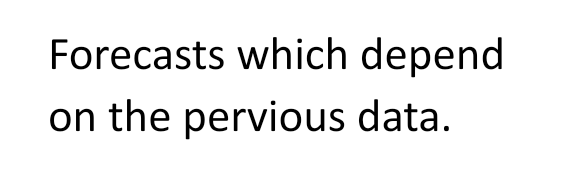
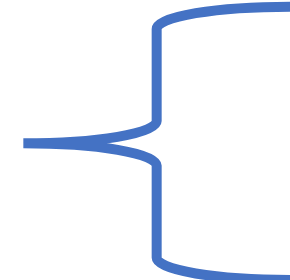
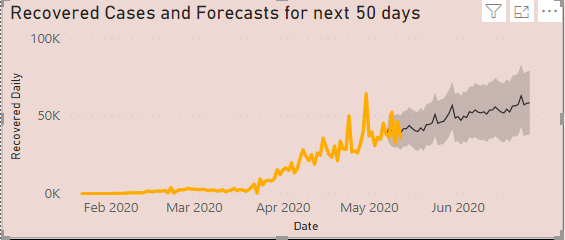
1. Heteroscedasticity: Ok a wired big term. It simply means when you plot the individual error against the predicted value, the variance of the error values tends to be variable. In case of linear regression, there should be *no heteroscedasticity* observed, which means, that the variance among the error terms should neither be increasing nor decreasing. It should be centered around the mean as shown below. This gives a sense of understanding that whether the predicted values will be consistent across all types of fluctuations in the independent terms.

A close up of a map

Description automatically generated

*Homoscedasticity is seen for the train data*

1. Normality of error terms: This is quite an important assumption when considering finite sample space. It is more of practical proof to show then theoretical. This normality approach states that *residuals ,which is the variation of the dependent variable which are not explained by the predictors, tend to follow a normal distribution with zero as the mean*. And, inferencing out the coefficients lies largely on this assumption while considering maximum likelihood estimation.
2. Autoregressive: This is quite simple to understand. So, in regression analysis the prediction *should not be a linear combination of past predictions(autoregression)*, instead, the predicted value should be a linear combination of the indepesndent variables. If you try to relate the above statement, dependency on previous predictions is seen in *time series,* where we consider the past values.



*The assumption fails in such cases and regression cannot be applied.*

These assumptions play a key role when we want to train a model. It can either reject a model completely or improve the accuracy tremendously. The real-world data comes with its nuances and making it conformable for the assumptions (with various transformations and statistical tests ) is the priority when handling regression.

*P.S: One quick advice: Always code to obtain any kind of inferences, and that is what I have done above to gain some insights 😊)*