Assumptions in Linear Regression :

(i) Linear Model -> The relationship b/w the independent and dependent variables should be linear. The reason behind this relationship is that relationship will be non-linear ~~it'll~~ which is certainly is the case in the real world data then the predictions made by our linear regression model willn't be acurate & will vary from actual observations alot.

(ii) No Multicolinearity in the data -> If the predictor variables are corelated among themselves, then the data is said to have a multicollinearity problem.

High collinearity means that the two variables vary very similarily and contains the same kind of info. This will lead to redudancy in dataset. Due, to redudancy only the complexity of model'll increase and no new information or pattern is learned by the model.

We generally try to avoid highly correlated features even while using complex models

We can identify highly correlated dataset using scatterplots or heatmap.

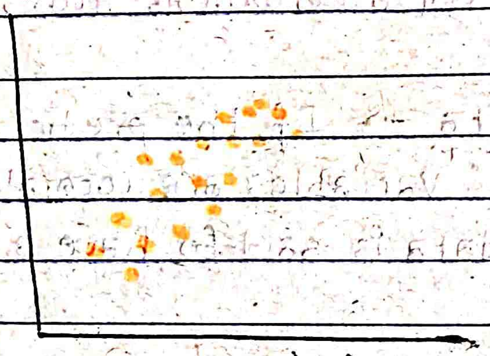
(iii) Homoescedasticity of Residuals or Equal Variances

Homoscedasticity is the term that states that the spread residuals which we are getting from the linear regression model should be homogenous or equal spaces.

If the spread of residuals is heterogenous then the model is called to be an unsatisfactory mo

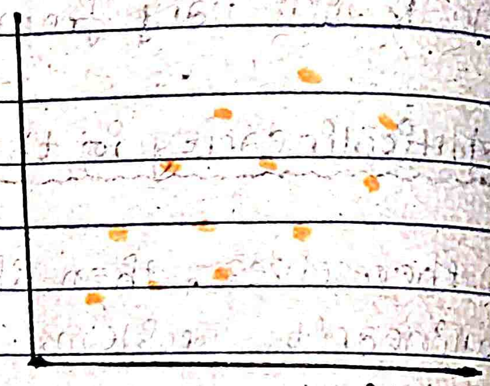
One can easily get an idea of the homoscedacity of the residuals by plotting a scatter plot of the residual data

Satisfactory model



Homoscedacity

Unsatisfactory model



Heteroscedacity

(iv) No Autocorrelation in Residuals ->

One of the critical assumptions of multiple L.R is that there should be no autocorrelation in the data. When the residuals are dependent on each other, there's autocorrelation.

ie, The factor is visible in the case of stock prices when the price of a stock is not independent of it's previous one

plotting the variables on a graph like a scatterplot or a line plot allows you to check for autocorrelatio if any

V) Nos of observations greater than the nos of predictors → for a better performing model, the number of trainning data or observations should be always greater than the number of test or prediction data.

However, greater the number of observations better the model performance

Therefore, to build a linear rego model you must have more observations than the number of independent variables (predictors) in the dataset. The reason behind this can be understood by the curse of dimensionality.

(Vi) Each observation is unique → Each observation in the dataset should be measured seperately on a unique occurence of the event that caused the observation.

i.e. if you want to include two observations to measure the density of liquid with 5kg mass and 5L volume, then you must experiment twice to measure the density for the two independent observations. Such observations are called replicate of each other. It would be wrong to use the same measurement for both observatio as you will disregard the random error.

(Vii) predictors are distributed Normally →
This assumption ensures that you have equally distributed ~~over~~ observations for a range of each predictor. So at the end of model trainning, the predicted values for each test data should be a normal distribution. One can get an idea of the distribution of the predicted values by plotting density, KDE or QQ plots of the predictions.