1. Import Data [excel, sockets, SQL] and Analyze,
   1. specify dependent and independent variables [Y = Classification or Continuous]
   2. Check if it is a linear or non-linear model
   3. Peak into the Dataset, glimpse(), str(), nrow(), ncol()
2. Visually Analyzing Dataset for ‘variance of variables’ by Plots [ggplot2]
3. Data Purification
   1. N/A value and Missing value Treatment
   2. Deleting Unwanted Variables [Identity Number or 50% + empty values]
   3. Outlier Treatment
   4. Manipulation [(s|l|t)apply functions, dplyr, tidyr]
   5. Dummy Variable Creation [Created Dummy - 1]
4. Splitting Data into {Training, Testing} sets or K – Cross Validation
5. Training the Model [Applying the formula and getting the fit]
6. Analyzing the below factors and Evaluating the model
   1. Residuals
   2. Coefficients
      1. Estimate
      2. Std. Err [Lesser is better]
      3. T Value
      4. P-Value [Lesser than 0.05 reject H0, variable has significance]
   3. Residual Standard Error with ‘N - 1’ degrees of freedom [RMSE]
   4. Explained Variability or Coefficient of Variance ‘R^2’ [Greater is better]
      1. Multiple R^2
      2. Adjusted R^2 [ideally should be 76 – 85 %]
   5. F-Statistic
7. Eliminate Multicollinearity by removing variables with VIF>5
8. Final model = Run stepAIC() [‘Forward’ ‘Backward’ ‘Both’] to select most significant variables and prepare. [Lesser is better]
9. Predict based on Final Trained model
   1. Step 6 for Accuracy

Random Notes:

Formula for Linear regression: Y = x1 x1x

Y = Dependent variable

Constant Coefficient [Intercept]

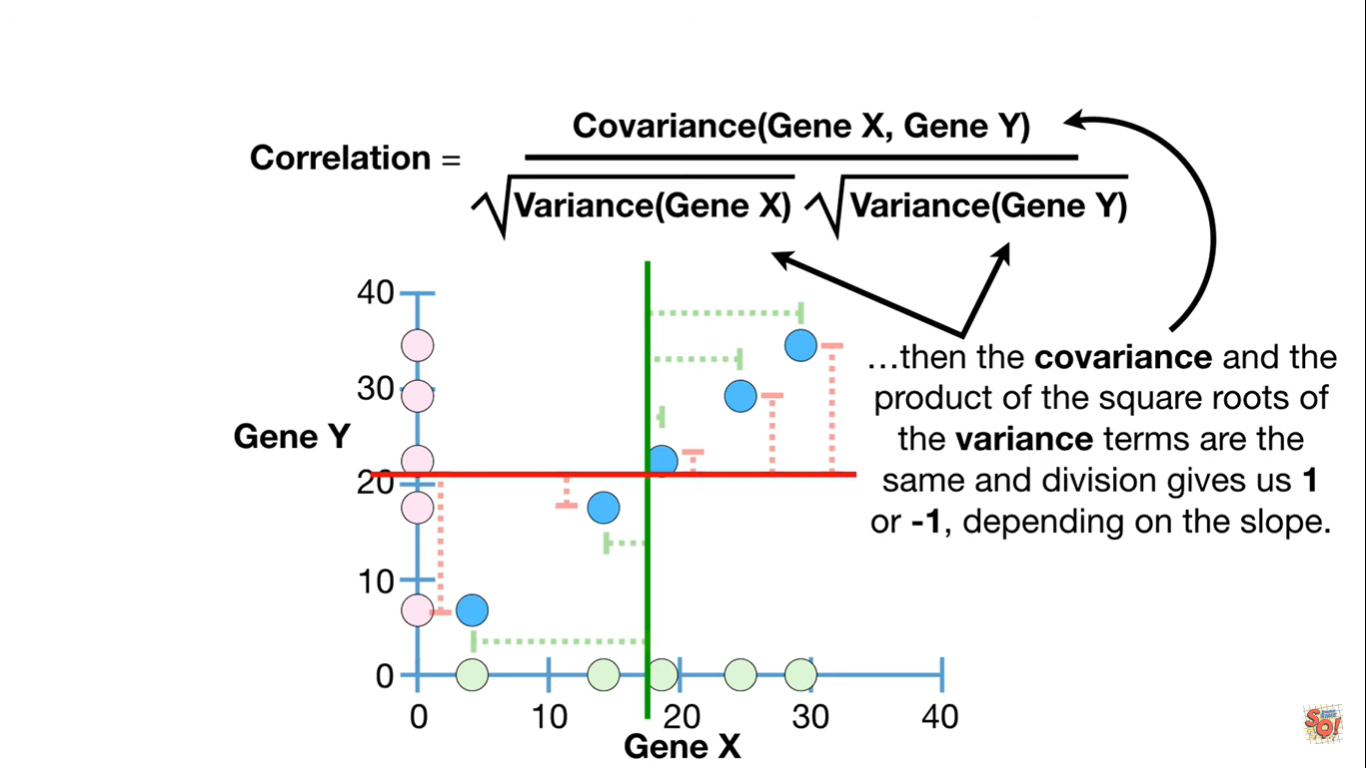
Coefficient of x

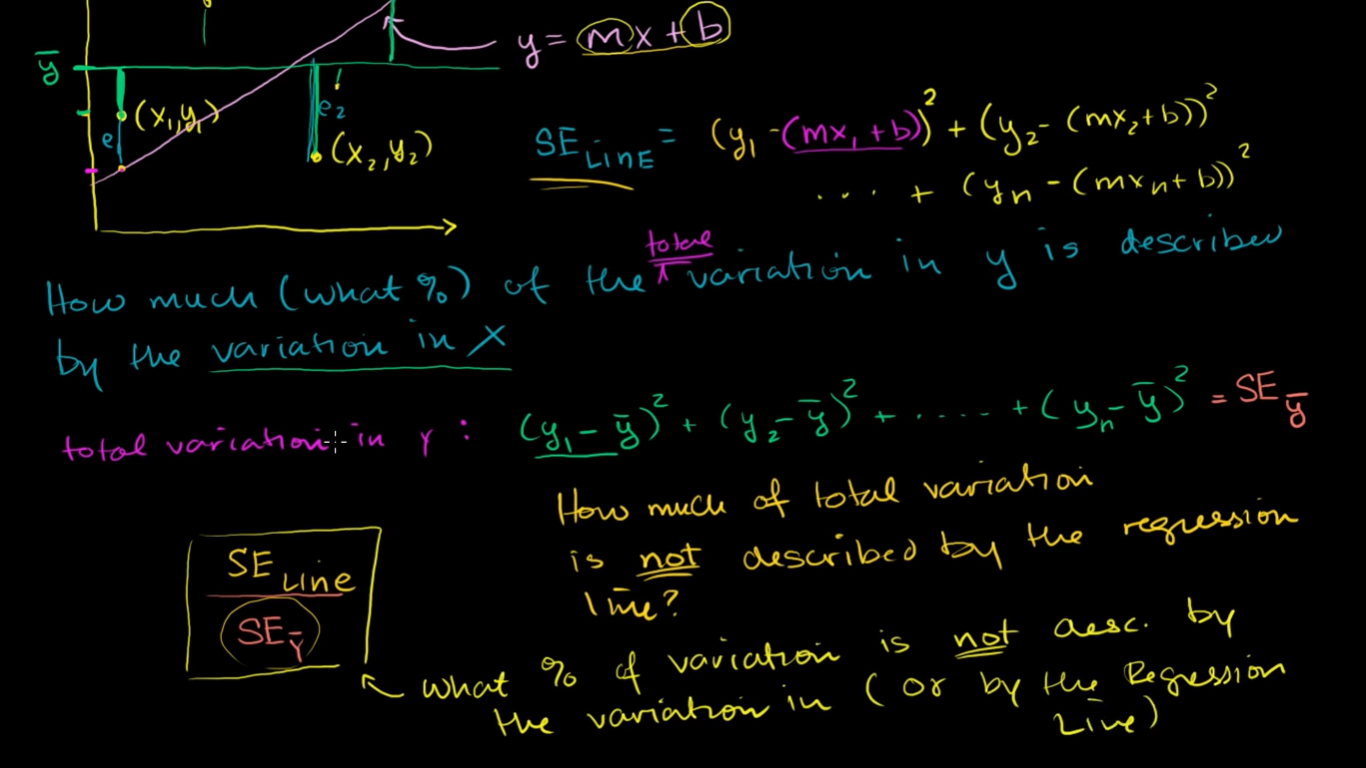
xIndependent variable

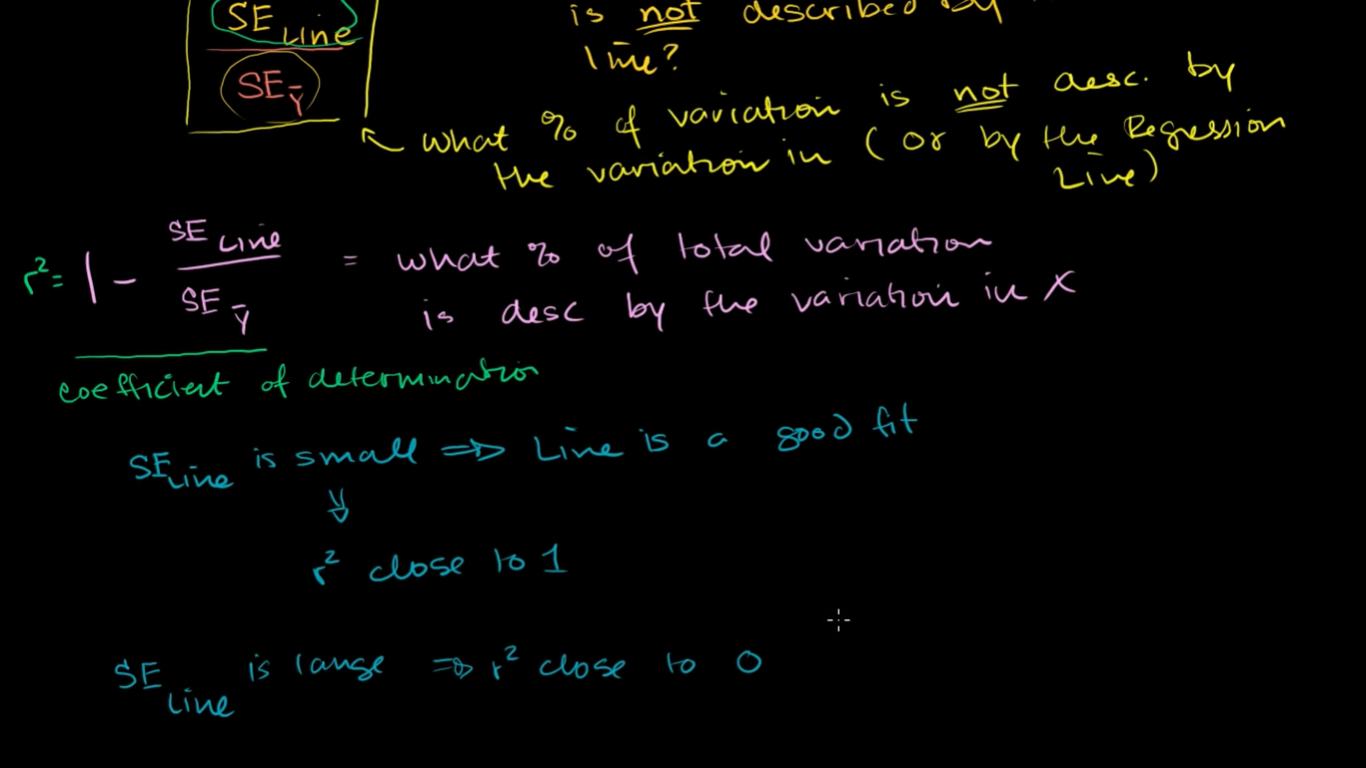
Residuals themselves have a distribution, min, max and quartiles. It is the distance b/w actual and predicted value.

Coefficients: Beta values in the equation, which act as units of change in the independent variable.

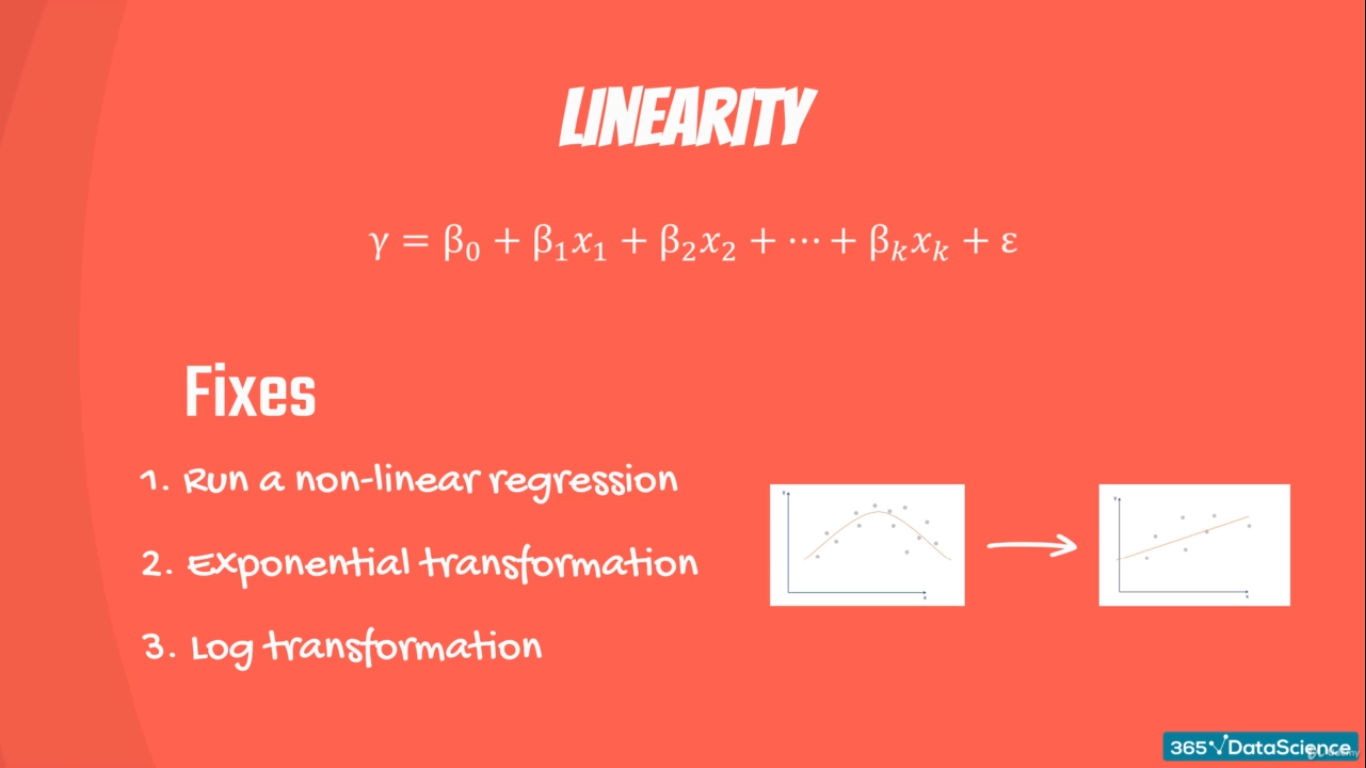
To eliminate Multi-collinearity we have to check the variables in corrplot(cor(<dataframe>)) 🡪 works only on numeric data.

Rejecting Null Hypothesis on a variable in LM means the variable has significant power over predicting the Y. To precisely find out the variables it is recommended to run step() function, which would try different combination in finding out the significance of them.





Linearity Assumption:

Only suitable if dependent variable [ Y ] can be predicted using an independent variable[x , x1], but sometime it might be a curved line instead of straight. Then the equations must be transformed appropriately, or algorithm should be changed to a non-linear one.

No Endogeneity:

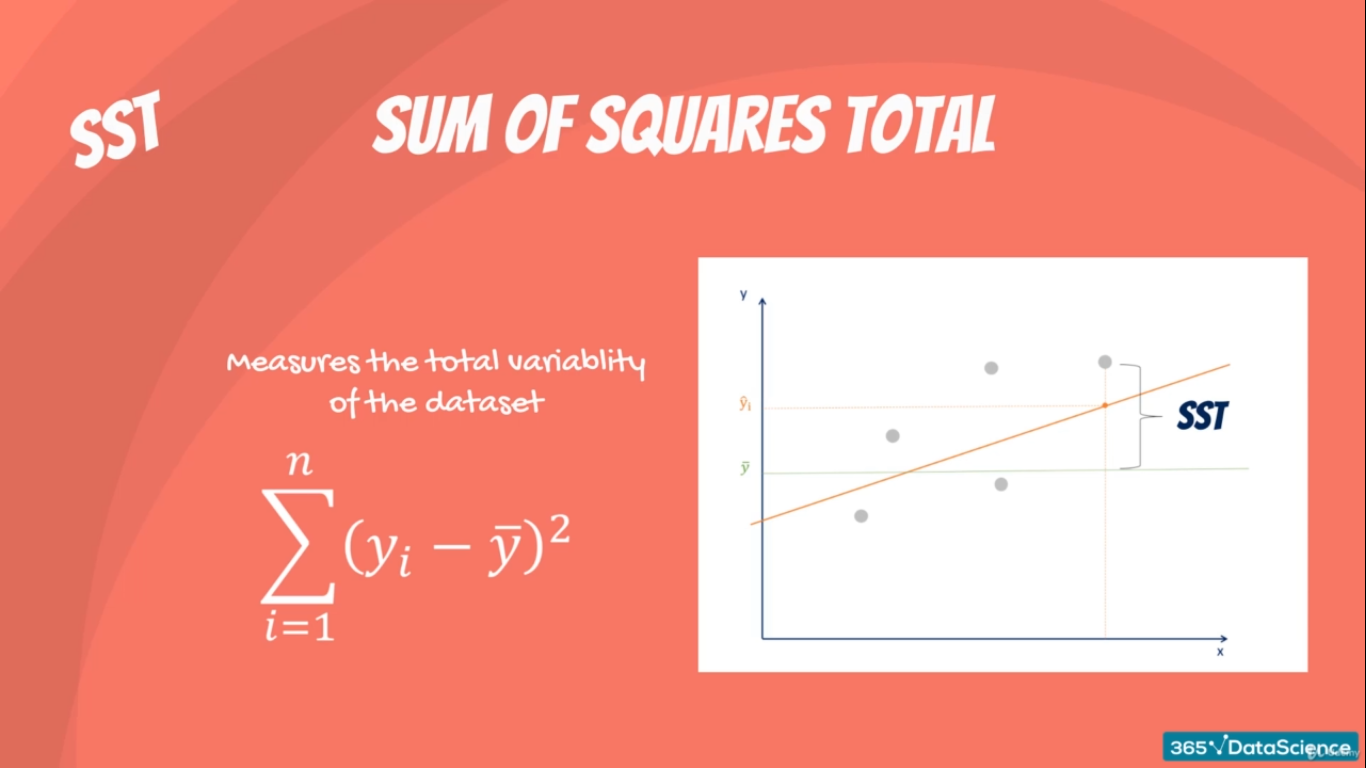
Prohibits from having a link between error and independent variables x,x1,x2… In terms Error and Independent Variables should not be corelated,

If they are correlated it leads to a problem ‘omitted variable bias’ 🡪[ occurs incase of a significant variable is not included in dataset, so all that is ignored will be added to the error]

Sum of Squares

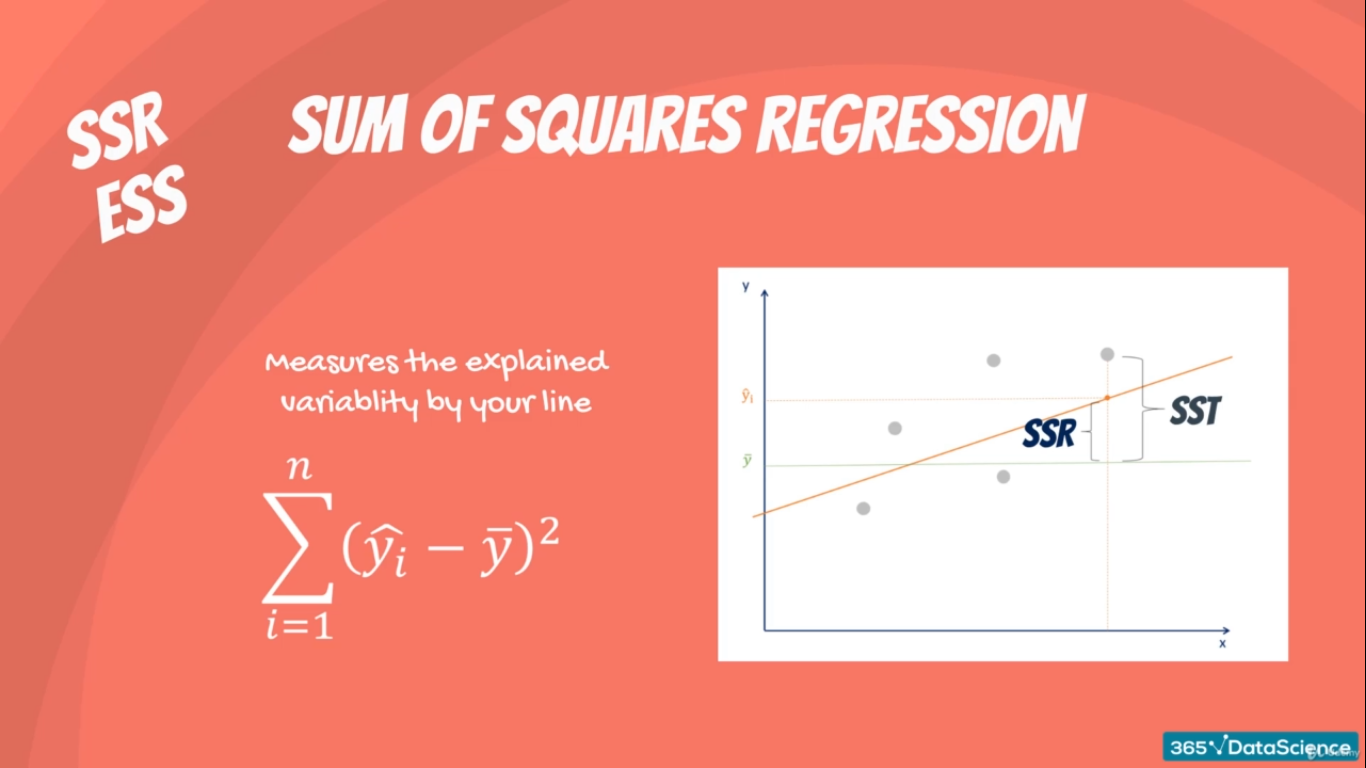
Total:

Sum of Squared difference b/w observed dependent variable and its mean [Dispersion of data around the mean]. Measures the total variability of Dataset



Regression:

Sum of Squared difference b/w predicted value and the mean of the dependent [Observed] value [How well the line fits data]. Measures explained variability by line.



Error:

Sum of Squared difference b/w predicted value and observed value. Measures unexplained variability by regression.

