Parul® University

Data Visualization and Data Analytics

Prof. Khushbu Chauhan, Assistant Professor
Information Technology

CHAPTER 4

Regression

Outlines

- Application of Regression for Analytics
- Introduction to Regression
- Simple and Multiple Linear Regression
- Correlation vs. Regression
- SST (Sum of Squares Total)
- SSR (Sum of Squares Regression)
- SSE (Sum of Squares Error)
- R-Square
- Adjusted R-Squared
- Logistic Regression

Application of Regression for Analytics

- Regression analysis is a reliable method of identifying which variables have impact on a topic of interest.
- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.
- It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.
- For example, if you've been putting on weight over the last few years, it can predict how much you'll weigh in ten years time if you continue to put on weight at the same rate.

FORECASTING

- The most common use of regression analysis in business is for forecasting future opportunities and threats, demand analysis, for example, forecasts the amount of things a customer is likely to buy.

- When it comes to business, though, demand is not the only dependent variable. Regressive analysis can anticipate significantly more than just direct income.

- For example, we may predict the highest bid for an advertising by forecasting the number of consumers who would pass in front of a specific billboard.

- Insurance firms depend extensively on regression analysis to forecast policyholder creditworthiness and the amount of claims that might be filed in a particular time period.

CAPM

The Capital Asset Pricing Model (CAPM), which establishes the link between an asset's projected return and the related market risk premium, relies on the linear regression model.

It is also frequently used in financial analysis by financial analysts to anticipate corporate returns and operational performance.

The beta coefficient of a stock is calculated using regression analysis. Beta is a measure of return volatility in relation to total market risk.

Because it reflects the slope of the CAPM regression, we can rapidly calculate it in Excel using the SLOPE tool.

Introduction to Regression

Regression analysis is a statistical method that helps us to analyze and understand the relationship between two or more variables of interest.

The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored, and how they are influencing each other.

Regression analysis is used for prediction and forecasting. This has a substantial overlap to the field of machine learning. This statistical method is used across different industries such as

1) Financial Industry- Understand the trend in the stock prices, forecast the prices, evaluate risks in the insurance domain

2) Marketing- Understand the effectiveness of market campaigns, forecast pricing and sales of the product.

3) Manufacturing- Evaluate the relationship of variables that determine to define a better engine to provide better performance

4) Medicine- Forecast the different combination of medicines to prepare generic medicines for diseases.

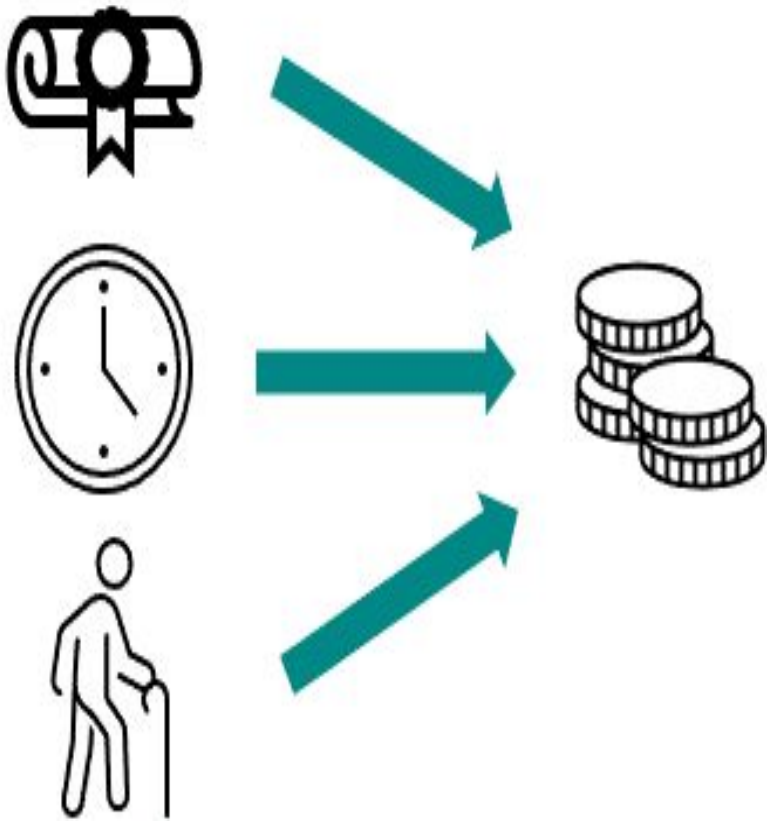
SIMPLE AND MULTIPLE

1) Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables.

2) Simple linear regression is a linear approach to model the relationship between a dependent variable and one independent variable. Multiple linear regression uses a linear function to predict the value of a dependent variable containing the function n independent variables.

Simple Linear Regression

Multiple Linear Regression



SIMPLE LINEAR REGRESSION

The goal of a **simple linear regression** is to predict the value of a dependent variable based on an independent variable. The greater the linear relationship between the independent variable and the dependent variable, the more accurate is the prediction.

Visually, the relationship between the variables can be shown in a scatter plot. The greater the linear relationship between the dependent and independent variables, the more the data points lie on a straight line.

MULTIPLE LINEAR REGRESSION

Unlike simple linear regression, multiple linear regression allows more than two independent variables to be considered. The goal is to estimate a variable based on several other variables. The variable to be estimated is called the dependent variable (criterion). The variables that are used for the prediction are called independent variables (predictors).

Multiple linear regression is frequently used in empirical social research as well as in market research. In both areas it is of interest to find out what influence different factors have on a variable.

CORRELATION VS. REGRESSION

There are some key differences between correlation and regression that are important in understanding the two.

Regression establishes how x causes y to change, and the results will change if x and y are swapped. With correlation, x and y are variables that can be interchanged and get the same result.

Correlation is a single statistic or data point, whereas regression is the entire equation with all the data points that are represented with a line.

Correlation shows the relationship between the two variables, while regression allows us to see how one affects the other.

The data shown with regression establishes a cause and effect, when one changes, so does the other, and not always in the same direction. With correlation, the variables move together.

Differences Between Correlation and Regression

Correlation	Regression
1 Relationship	1 One affects the other
2 Variables move together	2 Cause and effect
3 x and y can be interchanged	3 x and y cannot be interchanged
4 Data represented in single point	4 Data represented by line

CORRELATION VS. REGRESSION

	Correlation	Regression
When to use	When summarizing direct relationship between two variables	To predict or explain numeric response
Able to quantify direction of relationship?	Yes	No
Able to quantify strength of relationship?	Yes	Yes
Able to show cause and effect?	No	Yes
Able to predict and optimize?	No	Yes
X and Y are interchangeable?	Yes	No
Uses a mathematical equation?	No	$y = a + b\,(x)$

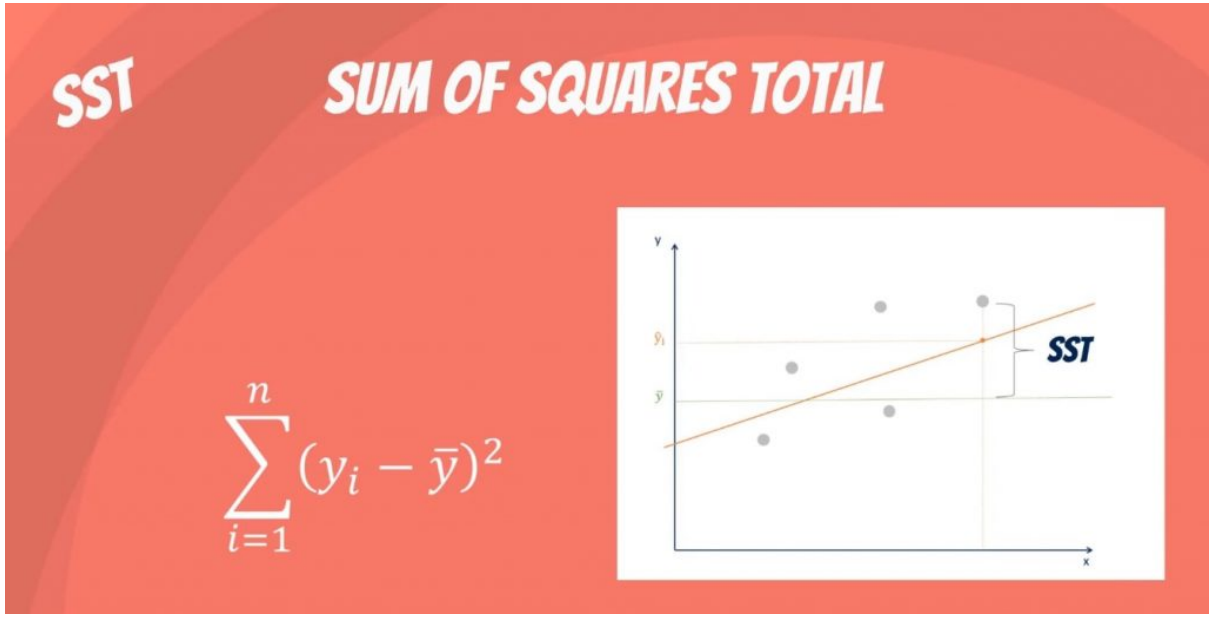
SST (SUM OF SQUARES TOTAL)

What is the SST?

The sum of squares total, denoted by SST, is the squared differences between the observed *dependent variable* and its mean. You can think of this as the dispersion of the observed variables around the mean – much like the variance in descriptive statistics.

It is a measure of the total variability of the dataset.

Side note: There is another notation for the SST. It is TSS or total sum of squares.



SSR (Sum of Squares Regression)

The second term is the **sum of squares due to regression**, or **SSR**. It is the sum of the differences between the *predicted* value and the **mean** of the *dependent variable*. Think of it as a measure that describes how well our line fits the data.

SSR SUM OF SQUARES REGRESSION

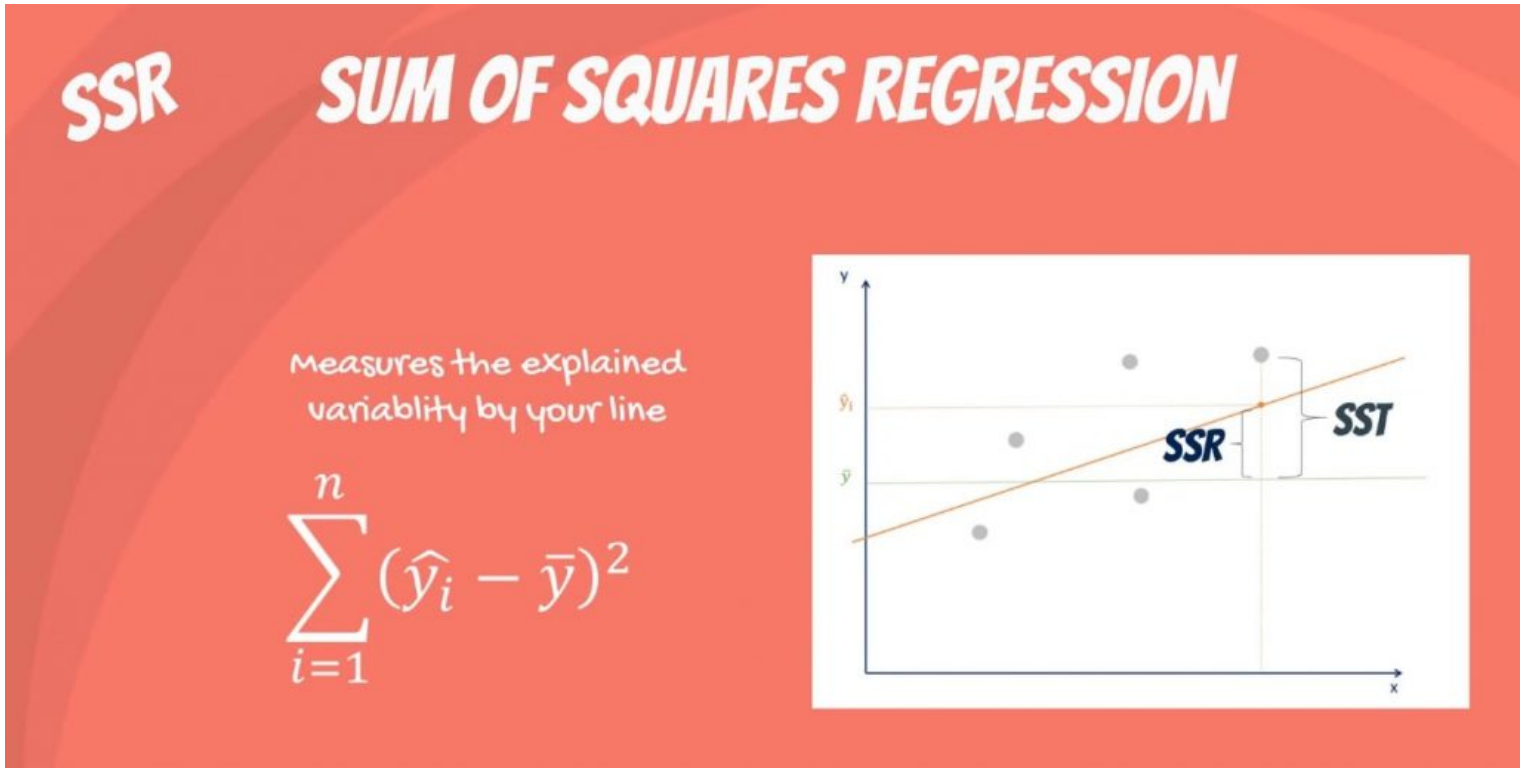
Measures the explained variablity by your line

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SSE (SUM OF SQUARES ERROR) R-SQUARE

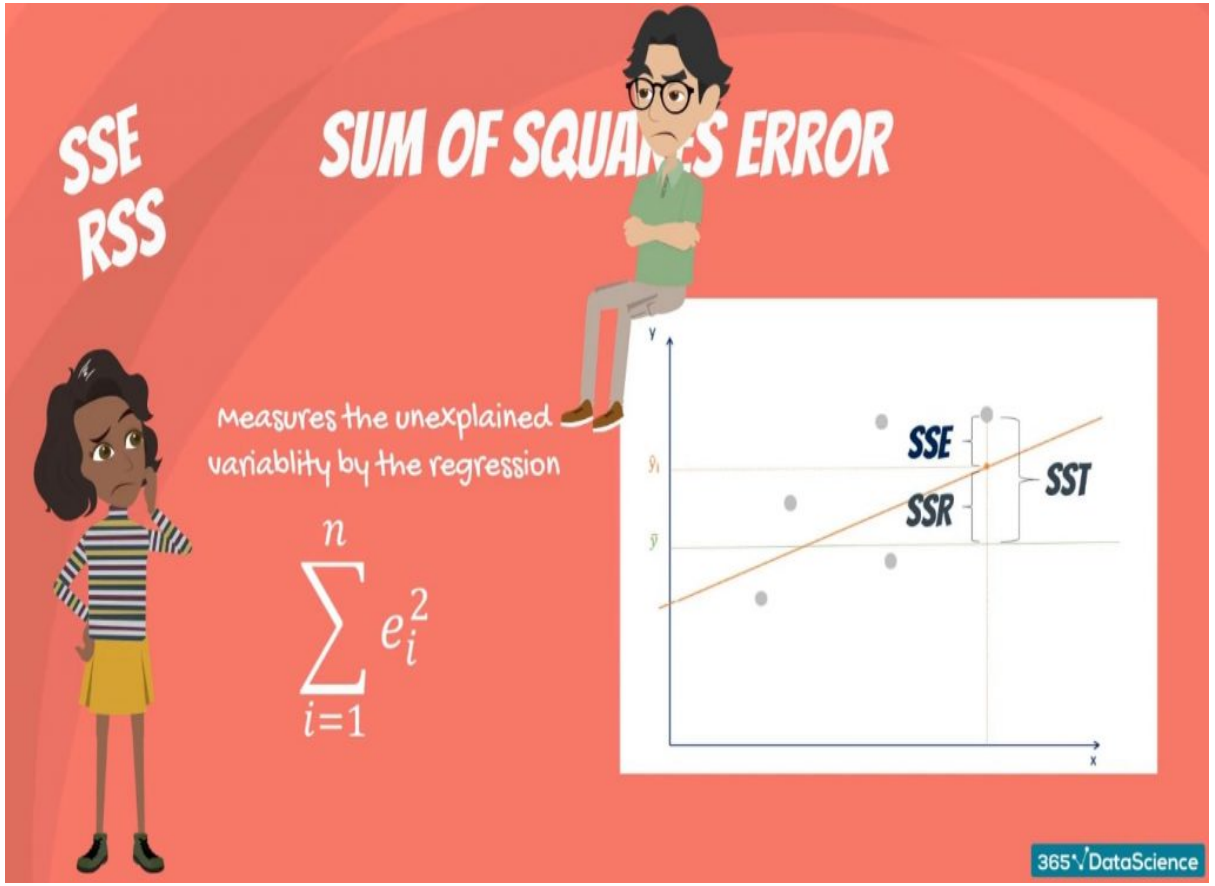
What is the SSE?

The last term is the sum of squares error, or SSE. The error is the difference between the *observed* value and the *predicted* value.



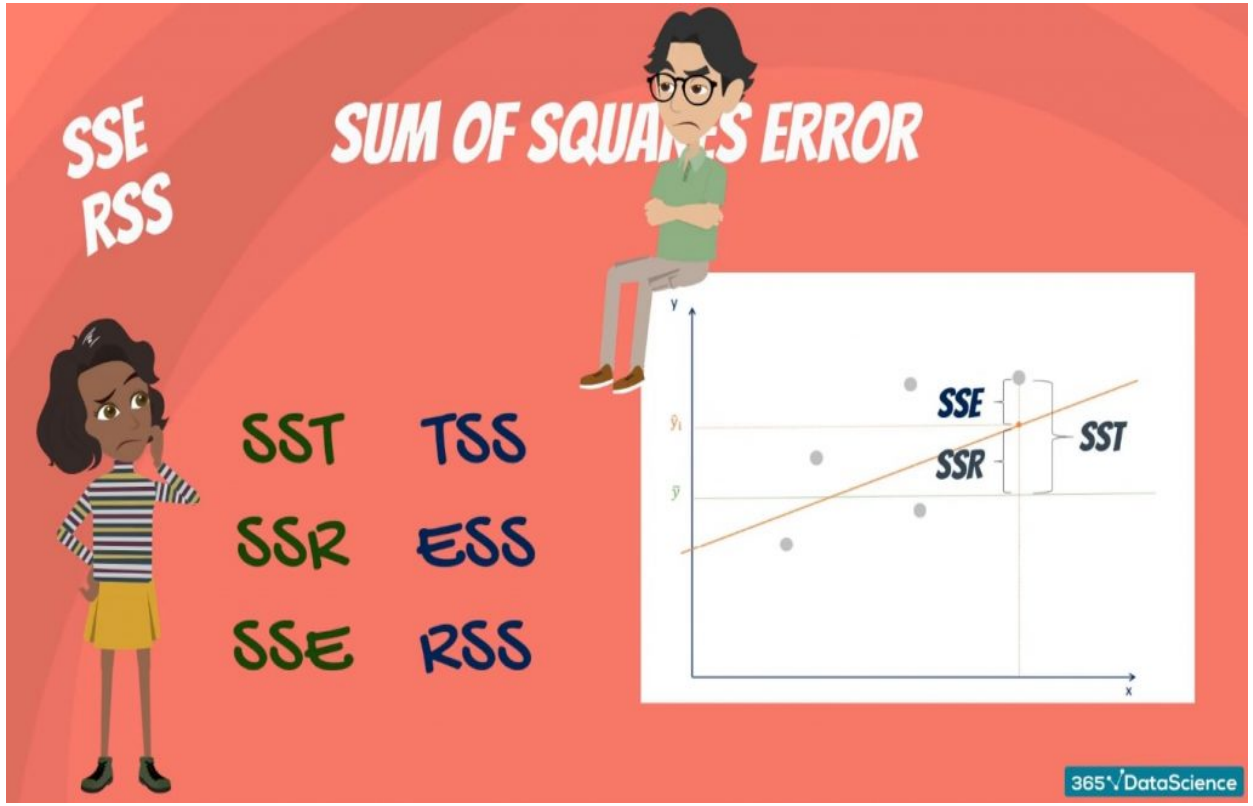
The Confusion between the Different Abbreviations

It becomes really confusing because some people denote it as SSR. This makes it unclear whether we are talking about the sum of squares due to regression or sum of squared residuals.



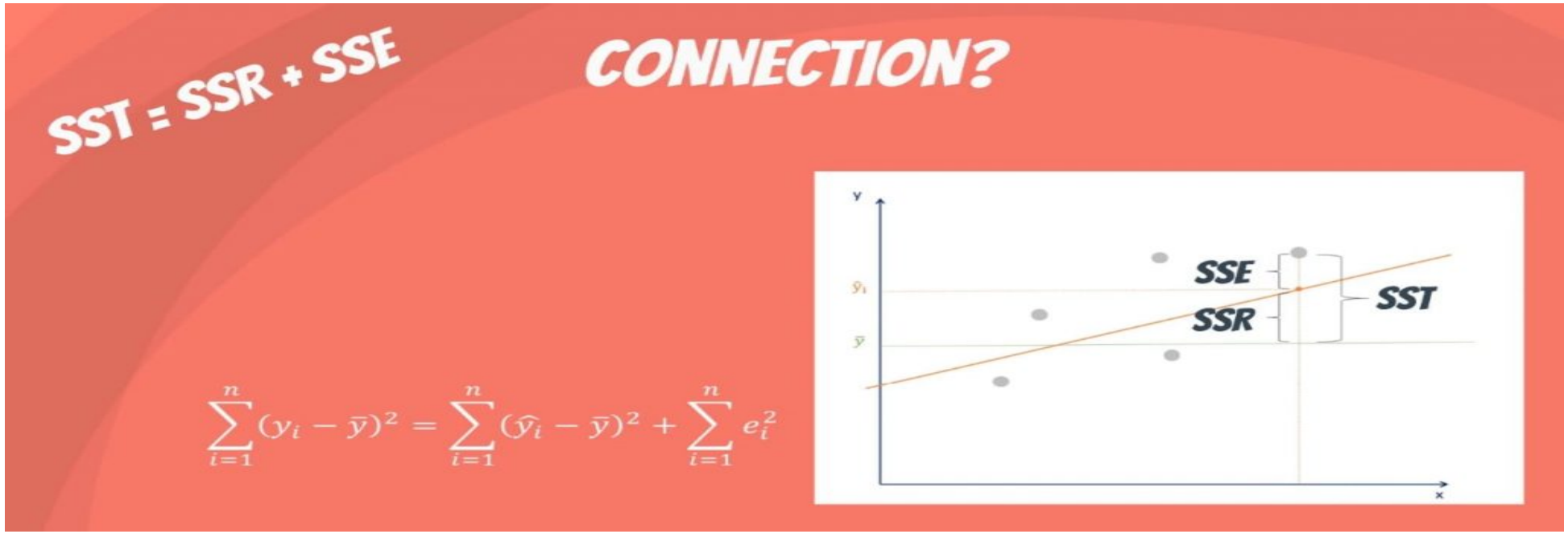
In any case, neither of these are universally adopted, so the confusion remains and we'll have to live with it.

Simply remember that the two notations are SST, SSR, SSE, or TSS, ESS, RSS.



How Are They Related?
Mathematically, SST = SSR + SSE.

The rationale is the following: the total variability of the data set is equal to the variability explained by the regression line plus the unexplained variability, known as error.



The rationale is the following: the total variability of the data set is equal to the variability explained by the regression line plus the unexplained variability, known as error.

Given a constant total variability, a lower error will cause a better regression. Conversely, a higher error will cause a less powerful regression. And that's what you must remember, no matter the notation.

R-SQUARE AND ADJUSTED R-SQUARED

- Goal : find a line with SSres as low as possible
- Value of R^2 generally lies between 0 and 1
- The more closer the R^2 is to 1 the better our line is.
- Why we square ? To deal with positive values and to deal with outliers , however we can use power of 4 or 6 or etc. but square is convention and that is widely accepted hence we will use that for now.
- R^2 can take negative values (if our data fits worstly to our linear regression)

PROBLEMS WITH R^2

As number of independent values increase the value of R^2 will either increase but will never descrease.

Hence we will not know how good of a influence does this newly added independent variable has on our dependent variable.

Reason for this is that any independent variable has tendency of slightly correalation with the dependent variable. This might help reducing the SSres value hence the value of R^2 increases.

To overcome this we use adjusted R^2

Adjusted R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

$y = b_0 + b_1 * x_1 + b_2 * x_2$

Problem:

$+ b_3 * x_3$

SS_{res} -> Min → R^2 will never decrease

ADJUSTED R^2

Adjusted R^2 deals with additional independent variables.

This r squared value tend to penalize the value of the rsquare if our choice of independent variable wasn't good (i.e independent variable had no effect on dependent variable)

Also the bias of R SQUARE to not to decrease is handled pretty well in this adjusted R SQUARED method.

DIFFERENCE

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

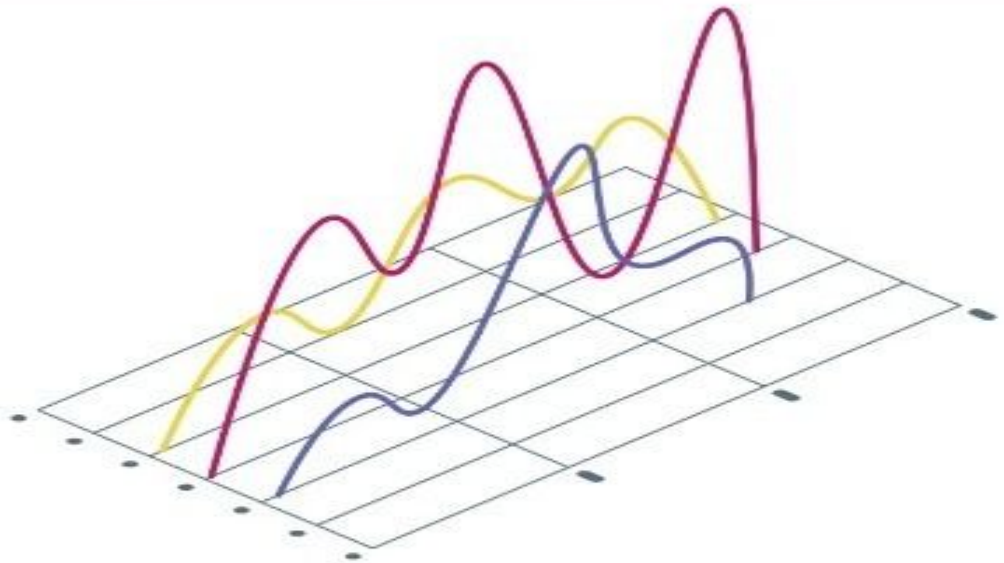
R-squared or R2 explains the degree to which your input variables explain the variation of your output / predicted variable. So, if R-square is 0.8, it means 80% of the variation in the output variable is explained by the input variables. So, in simple terms, higher the R squared, the more variation is explained by your input variables and hence better is your model.

However, the problem with R-squared is that it will either stay the same or increase with addition of more variables, even if they do not have any relationship with the output variables. This is where "Adjusted R square" comes to help. Adjusted R-square penalizes you for adding variables which do not improve your existing model.

Hence, if you are building Linear regression on multiple variable, it is always suggested that you use Adjusted R-squared to judge goodness of model. In case you only have one input variable, R-square and Adjusted R squared would be exactly same.

Typically, the more non-significant variables you add into the model, the gap in R-squared and Adjusted R-squared increases.

What is Logistic Regression?



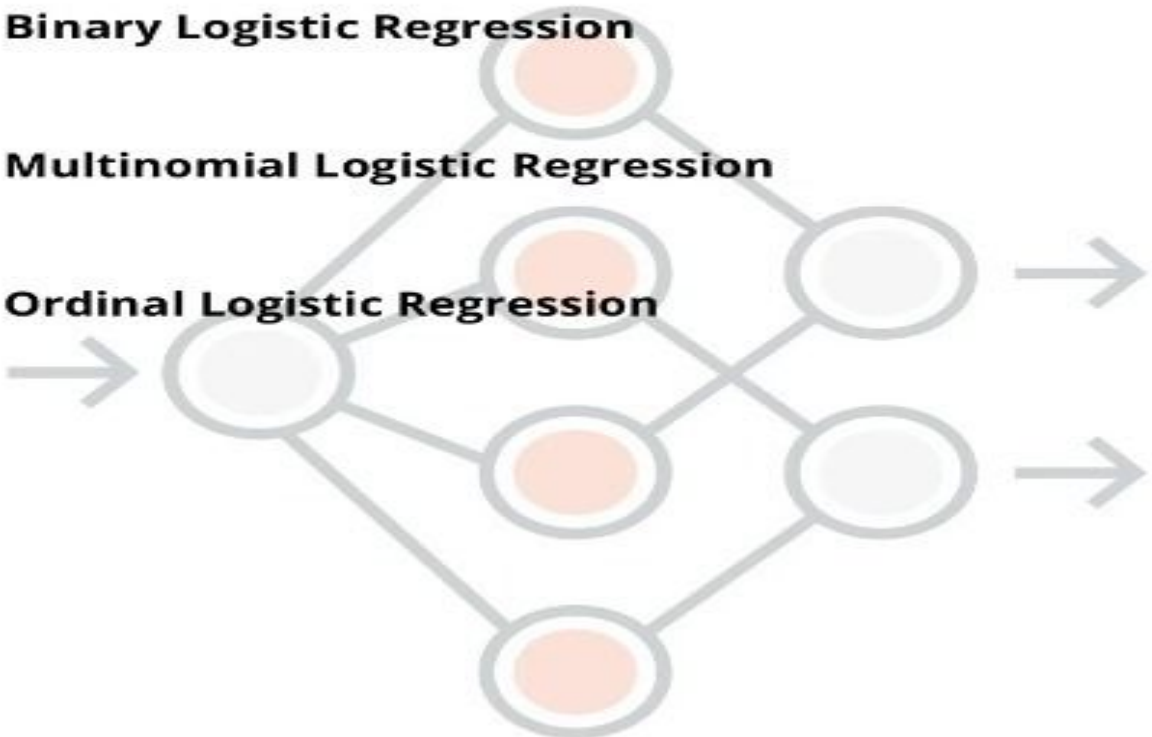
Logistic regression is a statistical technique for describing and explaining the connection between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

Assumptions of Logistic Regression

- Adequate sample size (too few participants for too many predictors is bad).
- Absence of multicollinearity (multicollinearity = high intercorrelations among the predictors).
- No outliers

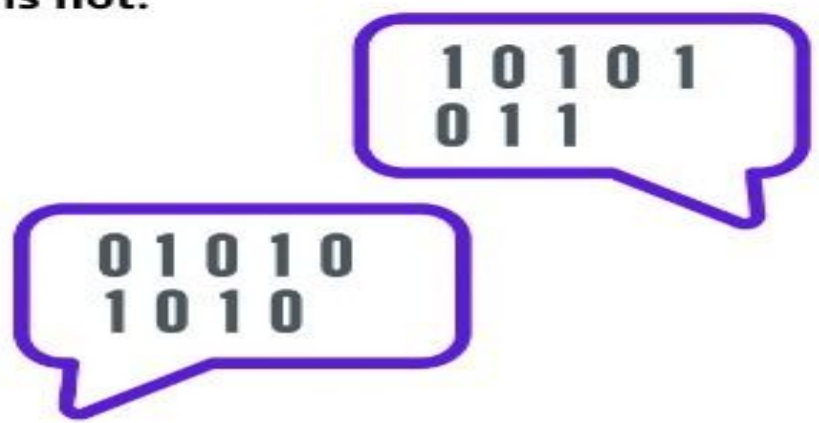
Types of Logistic Regression

- Binary Logistic Regression
- Multinomial Logistic Regression
- Ordinal Logistic Regression



Binary Logistic Regression

- Based on the values of the independent variables, binary logistic regression is used to estimate the likelihood of being a case (predictors).
- The odds are calculated by dividing the chance that a given result is a case by the probability that it is not.



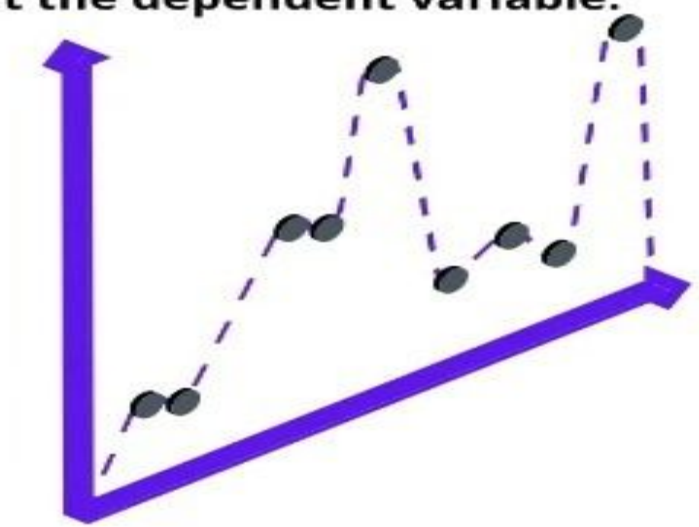
Multinomial Logistic Regression

- Multinomial logistic regression is a classification technique that extends logistic regression to situations with more than two discrete outcomes.

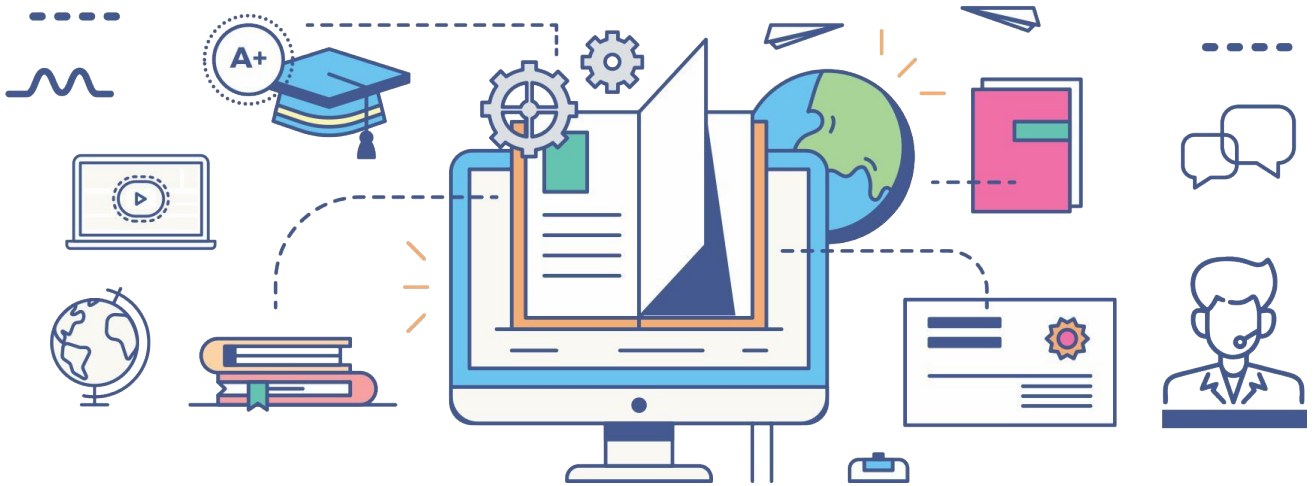
- Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

Ordinal Logistic Regression

- Ordinal Regression (sometimes called Ordinal Logistic Regression) is a binomial logistic regression extension.
- With 'ordered' multiple categories and independent variables, ordinal regression is used to predict the dependent variable.



DIGITAL LEARNING CONTENT



Parul® University

www.paruluniversity.ac.