UNIT -2

SIMPLE LINEAR REGRESSION ANALYSIS -1

INTRODUCTION TO PARAMETRIC ML MODEL:

PARAMETRIC ML MODEL - *A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.*

Some examples of parametric machine learning algorithms include:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

Benefits of Parametric Machine Learning Algorithms:

- **Simpler**: These methods are easier to understand and interpret results.
- **Speed**: Parametric models are very fast to learn from data.
- **Less Data**: They do not require as much training data and can work well even if the fit to the data is not perfect.

Limitations of Parametric Machine Learning Algorithms:

- **Constrained**: By choosing a functional form these methods are highly constrained to the specified form.
- **Limited Complexity**: The methods are more suited to simpler problems.
- **Poor Fit**: In practice the methods are unlikely to match the underlying mapping function.

NON PARAMETRIC ML MODEL:

Nonparametric methods : *are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.*

Some more examples of popular nonparametric machine learning algorithms are:

- k-Nearest Neighbors
- Decision Trees like CART and C4.5
- Support Vector Machines

Benefits of Nonparametric Machine Learning Algorithms:

- **Flexibility**: Capable of fitting a large number of functional forms.
- **Power**: No assumptions (or weak assumptions) about the underlying function.
- **Performance**: Can result in higher performance models for prediction.

Limitations of Nonparametric Machine Learning Algorithms:

- **More data**: Require a lot more training data to estimate the mapping function.
- **Slower**: A lot slower to train as they often have far more parameters to train.
- **Overfitting**: More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

LINEAR MODEL & IT'S ASSUMPTIONS :

The algorithms involve two steps:

1. Select a form for the function.

2. Learn the coefficients for the function from the training data.

- An easy-to-understand functional form for the mapping function is a line, as is used in linear regression:

$$b0 + b1*x1 + b2*x2 = 0$$

- Where b0, b1 and b2 are the coefficients of the line that control the intercept and slope, and x1 and x2 are two input variables.
- Assuming the functional form of a line greatly simplifies the learning process. Now, all we need to do is estimate the coefficients of the line equation and we have a predictive model for the problem.
- Often the assumed functional form is a linear combination of the input variables and as such parametric machine learning algorithms are often also called "linear machine learning algorithms".
- The problem is, the actual unknown underlying function may not be a linear function like a line. It could be almost a line and require some minor transformation of the input data to work right. Or it could be nothing like a line in which case the assumption is wrong and the approach will produce poor results.

SIMPLE LINEAR REGRESSION:

Simple linear regression is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

“**Regression models** describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.”

- The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

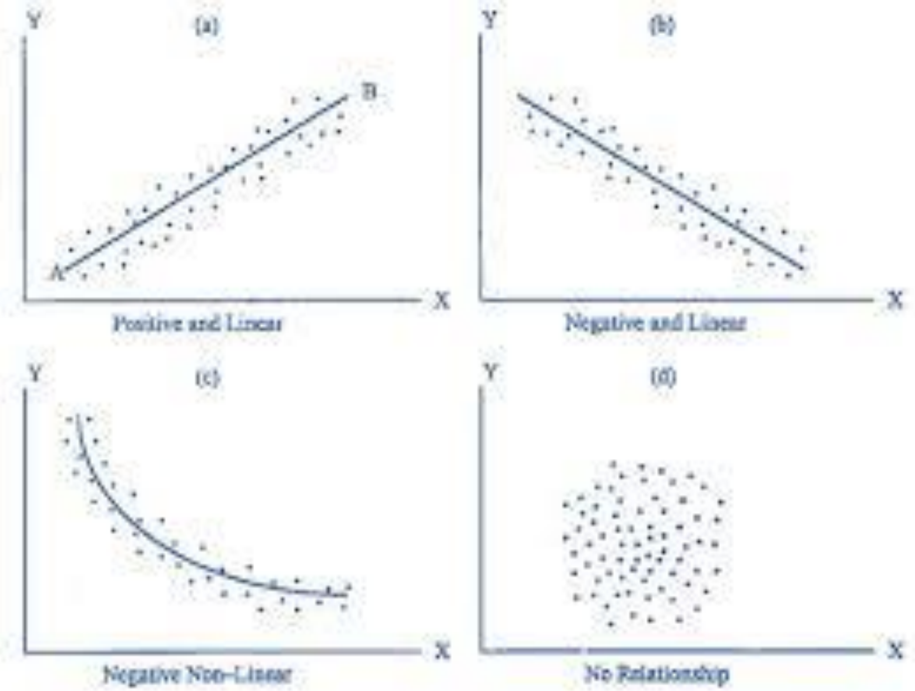
- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- $\mathbf{B_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.
- $\mathbf{B_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model.

GRADIENT DESCENT ALGORITHM:

- An algorithm to minimize a function by optimizing its parameters.
- *“Gradient Descent is an optimization algorithm used in machine learning to minimize the cost function by iteratively updating the model parameters. It is widely used in linear regression, logistic regression, and neural networks.”*

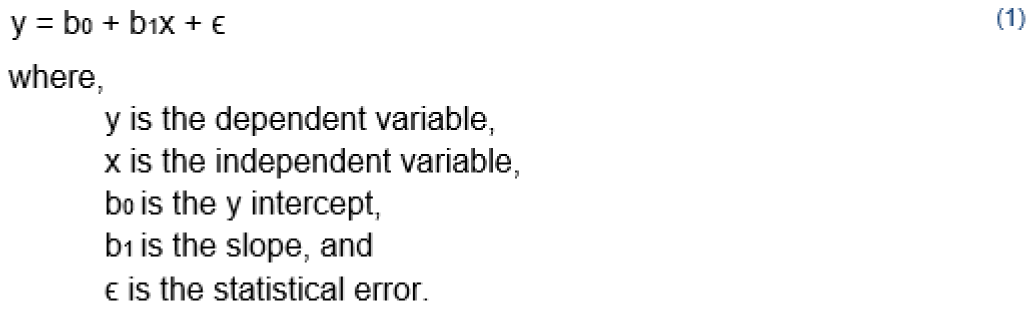
SCATTER DIAGRAM : is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.



Simple Linear regression parameter estimation:

The statistical error accounts for error that results from the inaccuracy in modeling and measuring the relationship between x and y.

It is important to note that the parameters b0 and b1, and the error term ϵ are unknown. These are population parameters which are theoretical values and cannot be determined



$$y = b_0 + b_1x + \epsilon \qquad (1)$$

where,

y is the dependent variable,
x is the independent variable,
b_0 is the y intercept,
b_1 is the slope, and
ϵ is the statistical error.

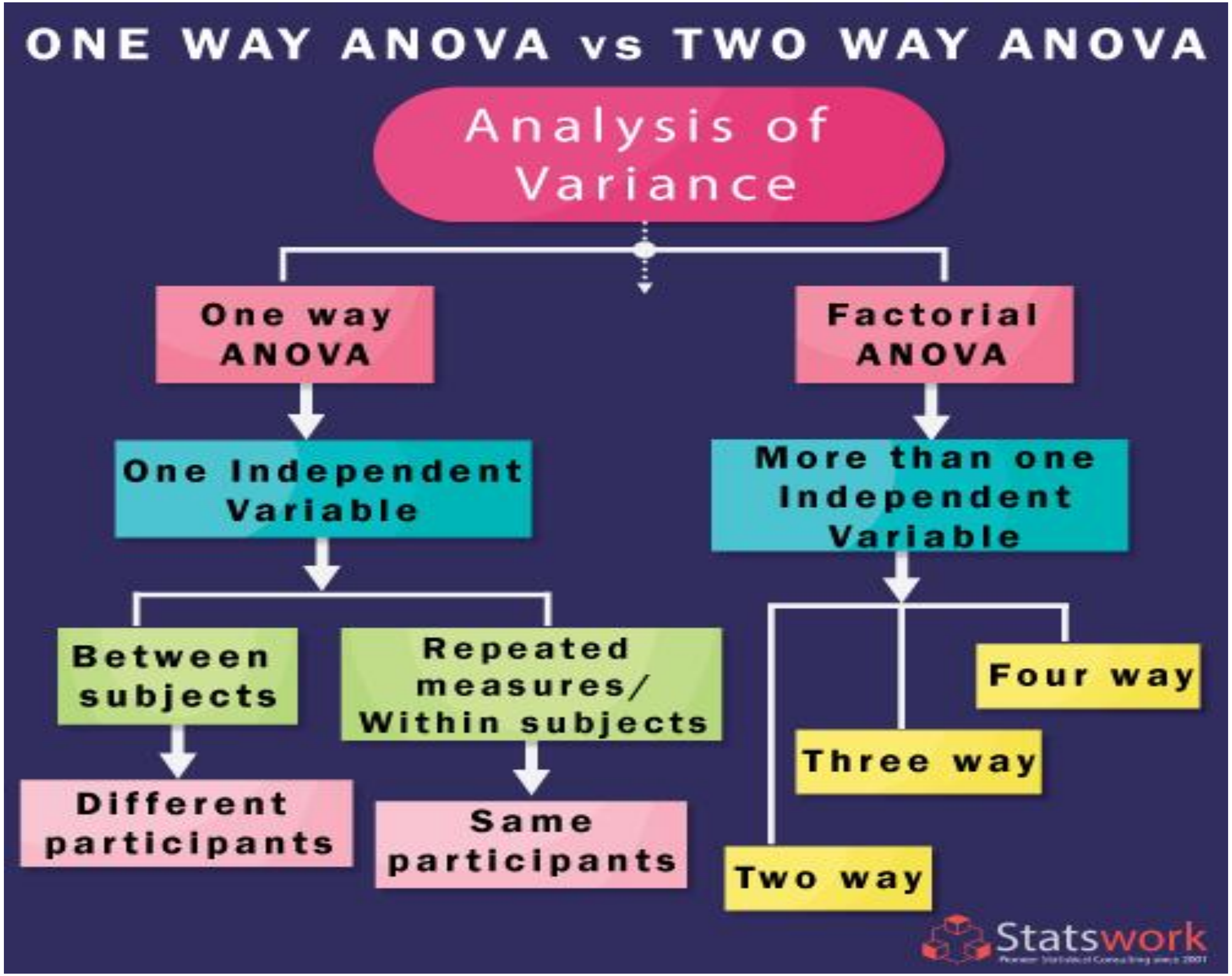
The most common method used to estimate the parameters b0 and b1 is the method of ***least squares***.

Please refer for more: https://medium.com/@devraj.agarwal/simple-linear-regression-parameter-estimates-explained-c8da2466bed6

ANALYSIS OF VARIANCE (ANOVA) & PARTIAL T-TEST, ESTIMATION OF SIGMA SQUARED.

What is an analysis of variance?

- An analysis of variance (ANOVA) tests whether statistically significant differences exist between more than two samples. For this purpose, the means and variances of the respective groups are compared with each other. In contrast to the t-test, which tests whether there is a difference between two samples, the ANOVA tests whether there is a difference between more than two groups.



ONE WAY ANOVA vs TWO WAY ANOVA

Analysis of Variance

One way ANOVA

One Independent Variable

Between subjects

Different participants

Repeated measures/ Within subjects

Same participants

Factorial ANOVA

More than one Independent Variable

Two way

Three way

Four way

Statswork

Pioneer Statistical Consulting since 2001

T-TEST

A t-test is a statistical hypothesis test used in machine learning to compare the means of two groups. It's used to determine if the difference between the groups is statistically significant.

How is a t-test used in machine learning?

- **Feature selection**: Compare the performance of a model with and without a feature
- **Model validation**: Compare different models
- **Linear regression**: Determine the significance of coefficients in a regression model

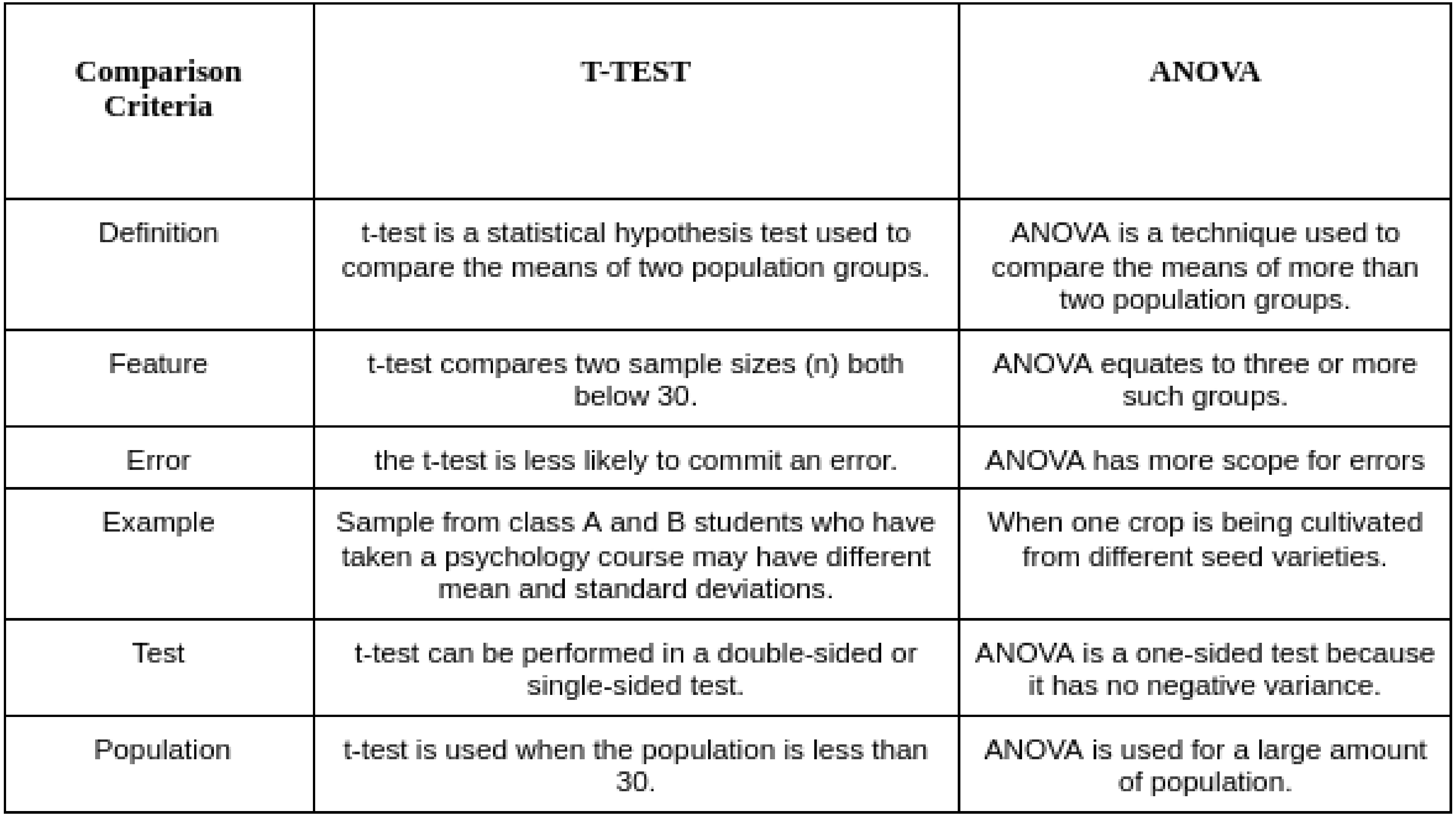
Types of t-tests

- **Independent samples t-test**: Compares the means of two independent groups
- **Paired samples t-test**: Compares the means of two related groups
- **One sample t-test**: Compares the mean of a single group to a known average

How is a t-test performed?

1. Calculate the t-statistic, which measures the difference between the means of the two groups
2. Determine the degrees of freedom
3. Use the t-distribution to estimate how likely it is that the true mean is in a given range
4. Determine if the difference between the groups is statistically significant

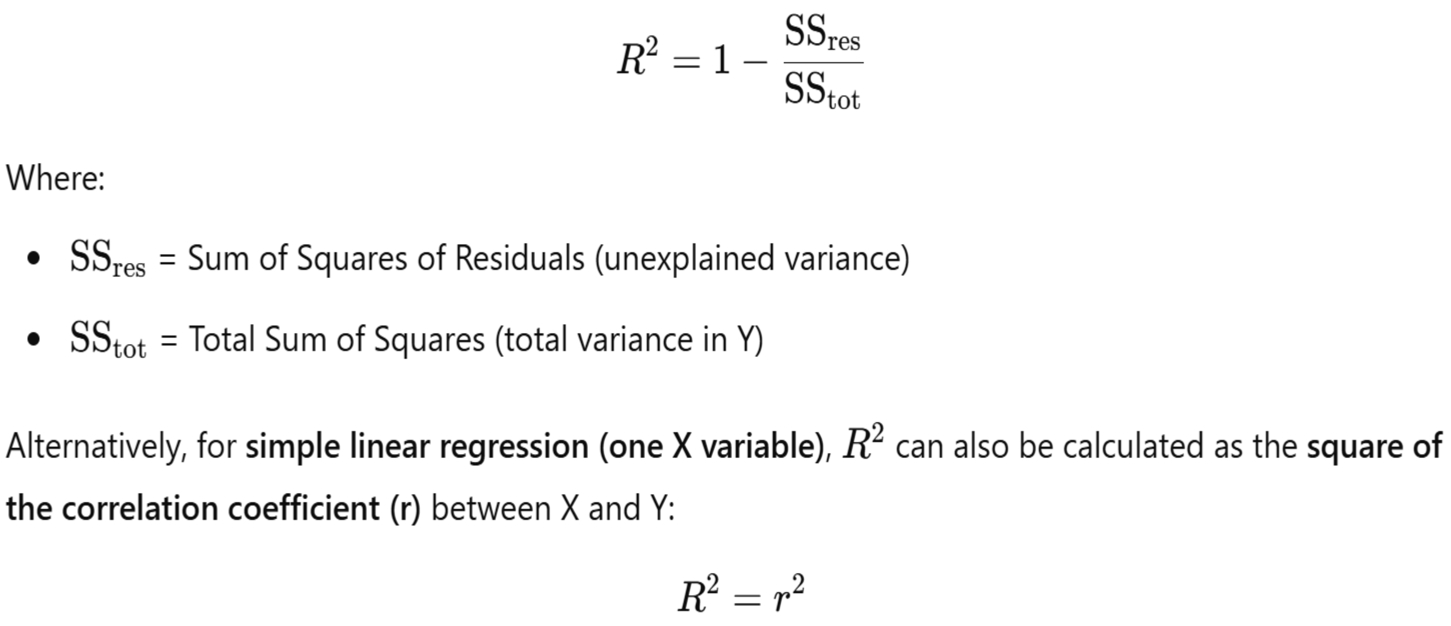
Assumptions of a t-test the data is normally distributed, the samples are independent of each other, and the paired observations are matched.



Comparison Criteria	T-TEST	ANOVA
Definition	t-test is a statistical hypothesis test used to compare the means of two population groups.	ANOVA is a technique used to compare the means of more than two population groups.
Feature	t-test compares two sample sizes (n) both below 30.	ANOVA equates to three or more such groups.
Error	the t-test is less likely to commit an error.	ANOVA has more scope for errors
Example	Sample from class A and B students who have taken a psychology course may have different mean and standard deviations.	When one crop is being cultivated from different seed varieties.
Test	t-test can be performed in a double-sided or single-sided test.	ANOVA is a one-sided test because it has no negative variance.
Population	t-test is used when the population is less than 30.	ANOVA is used for a large amount of population.

R SQUARE

- **R-squared** in **simple linear regression** is a statistical measure that represents the **proportion of the variance in the dependent variable (Y)** that is explained by the independent variable (X) using the linear regression model.
- **Formula for R^2:**



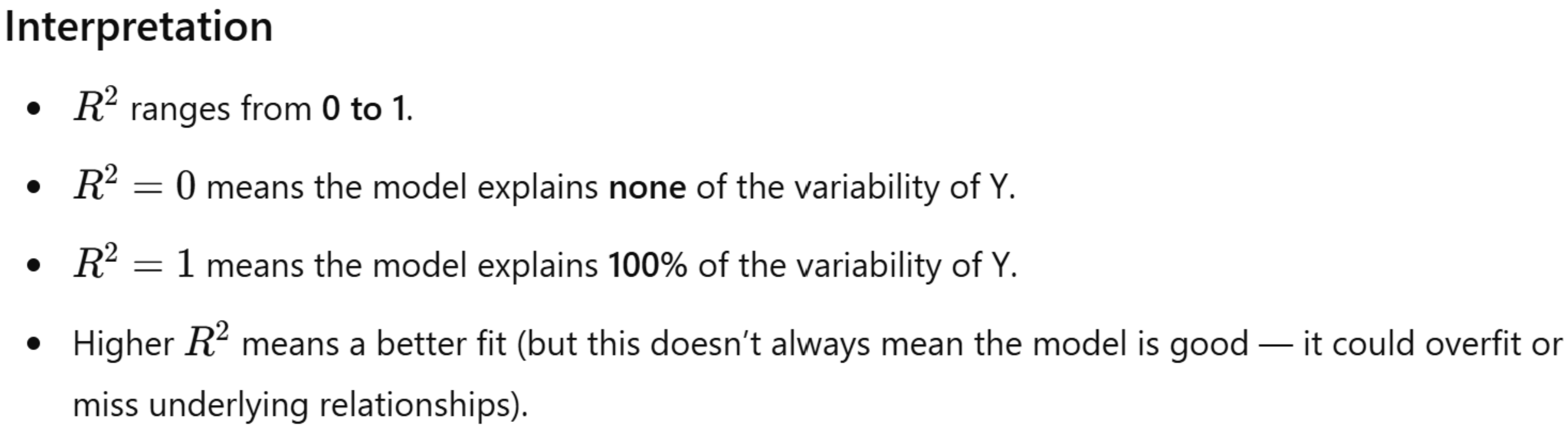
$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

Where:

- SS_{res} = Sum of Squares of Residuals (unexplained variance)
- SS_{tot} = Total Sum of Squares (total variance in Y)

Alternatively, for **simple linear regression (one X variable)**, R^2 can also be calculated as the **square of the correlation coefficient (r)** between X and Y:

$$R^2 = r^2$$

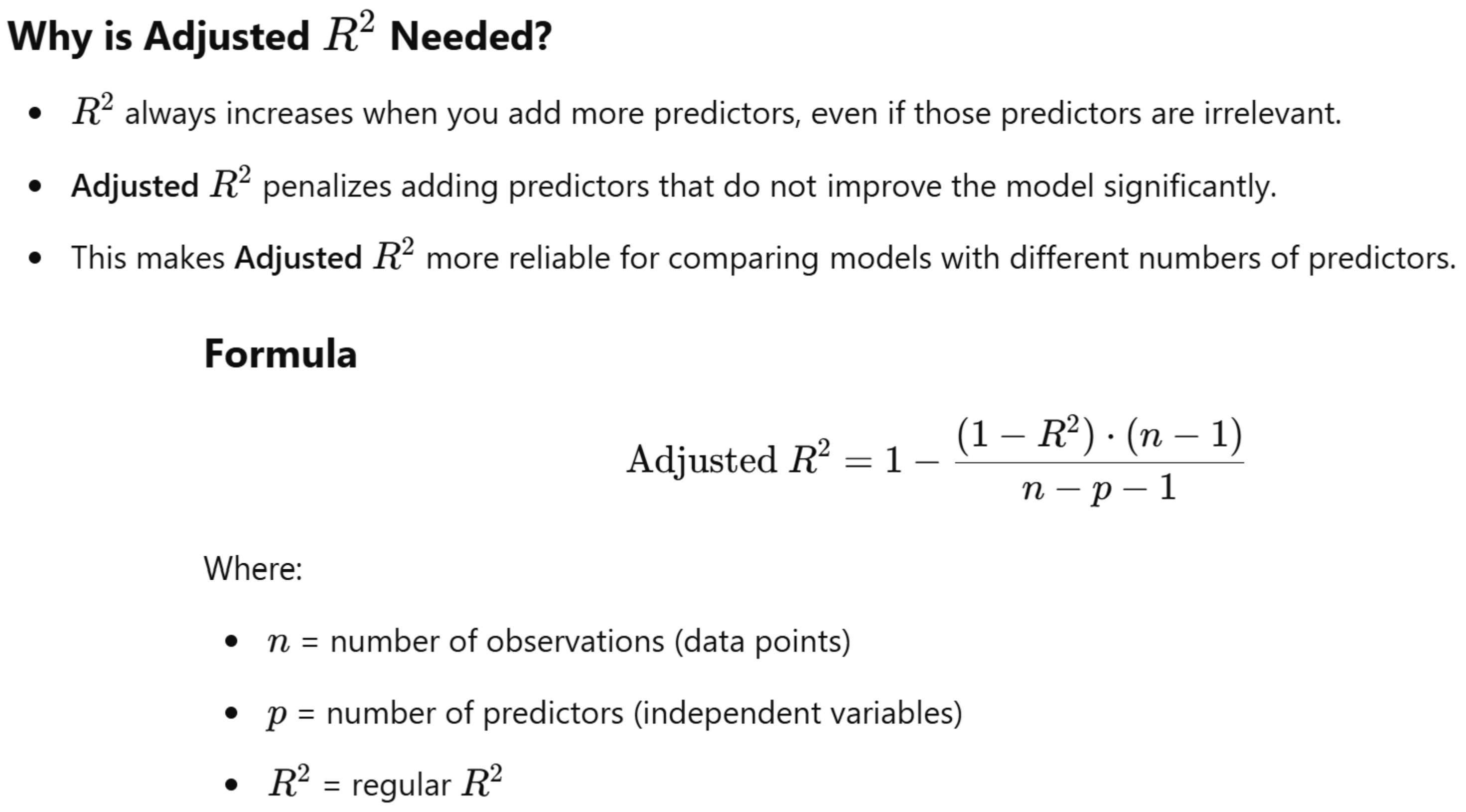


Interpretation

- R^2 ranges from **0 to 1**.
- $R^2 = 0$ means the model explains **none** of the variability of Y.
- $R^2 = 1$ means the model explains **100%** of the variability of Y.
- Higher R^2 means a better fit (but this doesn't always mean the model is good — it could overfit or miss underlying relationships).

ADJUSTED R-SQUARED

- **Adjusted R2R^2R2** is a modified version of **R2R^2R2** that adjusts for the number of predictors (independent variables) in a regression model.



Why is Adjusted R^2 Needed?

- R^2 always increases when you add more predictors, even if those predictors are irrelevant.
- **Adjusted** R^2 penalizes adding predictors that do not improve the model significantly.
- This makes **Adjusted** R^2 more reliable for comparing models with different numbers of predictors.

Formula

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2) \cdot (n - 1)}{n - p - 1}$$

Where:

- n = number of observations (data points)
- p = number of predictors (independent variables)
- R^2 = regular R^2