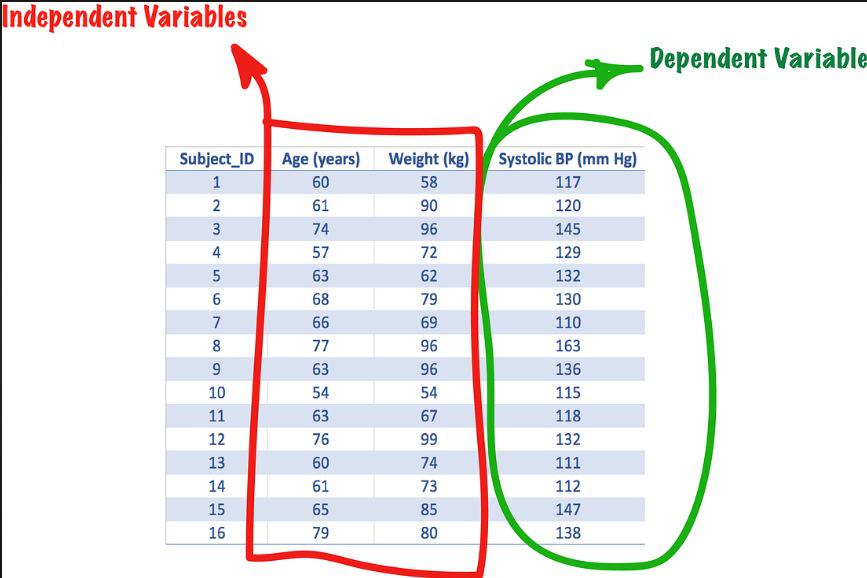
Introduction to Linear Regression

Linear regression is a statistical method that aims to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. This widely-used technique serves as a fundamental building block in both statistics and machine learning, providing valuable insights into the relationships between variables.

Suppose below is the observed data :



**Basic Concept:**

At its core, linear regression assumes a linear relationship between the dependent variable (BP) and the independent variable(s) (Age, Years). The relationship is expressed through a linear equation:

BP = β0 + β1(Age)+ β2(years) + ????

where (BP) is dependent variable,

(Age, years) are independent variable,

β0is the y-intercept

β1 , β2 are coefficients

ε is the error term.

**Objective**

The primary goal of linear regression is to determine the best-fitting line, minimizing the sum of squared differences between the observed and predicted values. In other words, the model aims to capture the underlying linear relationship between variables while accounting for variability.

**Estimation of Coefficients:**

The process of finding the optimal coefficients involves using statistical methods such as the least squares method. The coefficients β0,β1 ,...,βn are estimated to create a model that accurately represents the data. This fitting process provides a mathematical representation of how changes in the independent variables correlate with changes in the dependent variable.

**Advantages of Linear Regression**

* Linear regression is a relatively simple algorithm, making it easy to understand and implement.
* Linear regression is computationally efficient and can handle large datasets effectively. It can be trained quickly on large datasets, making it suitable for real-time applications.
* Linear regression often serves as a good baseline model for comparison with more complex machine learning algorithms.
* Linear regression is a well-established algorithm with a rich history and is widely available in various machine learning libraries and software packages.

**Disadvantages of Linear Regression**

* Linear regression assumes a linear relationship between the dependent and independent variables. If the relationship is not linear, the model may not perform well.
* Linear regression is sensitive to multicollinearity, which occurs when there is a high correlation between independent variables. Multicollinearity can inflate the variance of the coefficients and lead to unstable model predictions.
* Linear regression assumes that the features are already in a suitable form for the model. Feature engineering may be required to transform features into a format that can be effectively used by the model.
* Linear regression is susceptible to both overfitting and underfitting. Overfitting occurs when the model learns the training data too well and fails to generalize to unseen data. Underfitting occurs when the model is too simple to capture the underlying relationships in the data.

Linear Regression and Gradient Descent

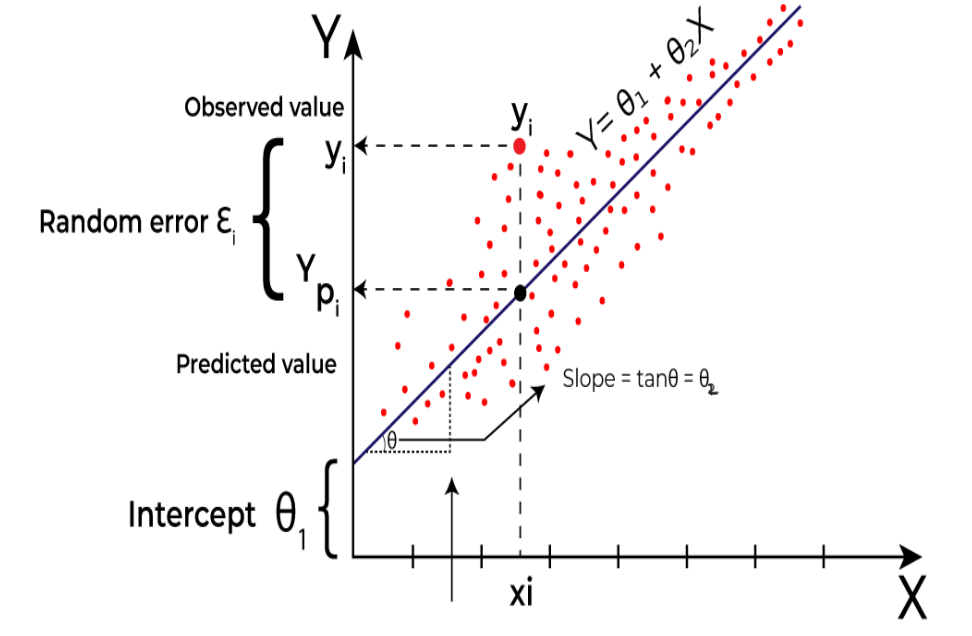
The goal of the Linear regression algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

**Best Fit Line**

Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line. The best fit line is determined by finding the coefficients ( β0 and β1) that minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the regression equation. This process is often achieved through the method of least squares.

The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



**Use of Best Fit Line**

**Prediction:** Once the best fit line is established, it can be used to make predictions. Given a value of the independent variable(s), the equation allows you to estimate the corresponding value of the dependent variable. This is particularly useful for forecasting and understanding trends in data.

**Understanding Relationships:** The slope (β1) of the best fit line indicates the strength and direction of the relationship between the variables. A positive slope suggests a positive correlation, while a negative slope indicates a negative correlation.

**Visual Representation:**The best fit line is often plotted on a scatterplot of the data points. It visually represents the linear trend in the data and helps assess how well the model fits the observed data.

**Cost Function of Linear Regression**

In regression, the difference between the observed value of the dependent variable(yi) and the predicted value(predicted) is called the residuals.

εi = ypredicted – yi = Ŷ - yi

where ypredicted = Ŷ = ϴ1 + ϴ2 Xi

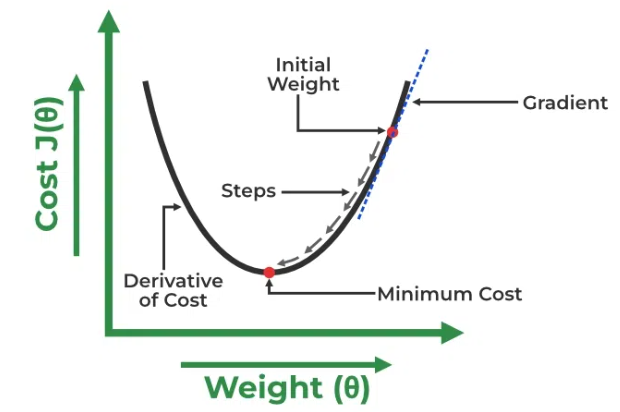
The cost function helps to work out the optimal values for B0 and B1, which provides the best fit line for the data points.

In Linear Regression, generally Mean Squared Error (MSE) cost function is used, which is the average of squared error that occurred between the ypredicted and yi. This is the specified cost function.

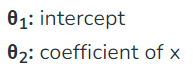
cost_fun

**Gradient Descent**

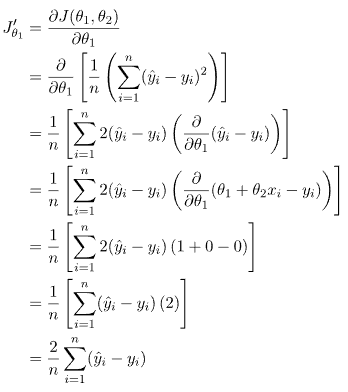
A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the values to reach the minimum cost function.



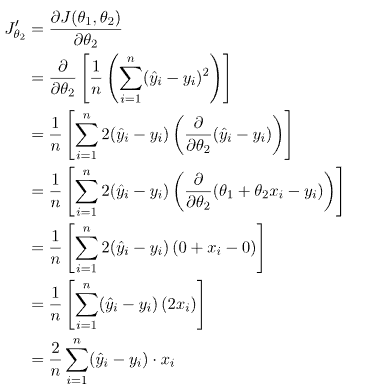
To update ϴ1 and ϴ2 values in order to reduce the Cost function (minimizing MSE value) and achieve the best-fit line the model uses Gradient Descent.



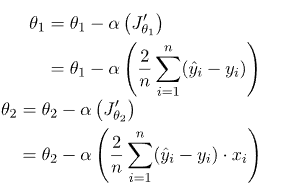
Let's differentiate the cost function(J) with respect to ϴ1 :



Let's differentiate the cost function(J) with respect to ϴ1 :



Now for the new best fit line the new ϴ1 and ϴ2 are :



where, alpha is the Learning rate

MSE vs MAE

**MSE (Mean Squared Error)**

Mean Squared Error (MSE) is the average squared error between actual and predicted values. The mean squared error is calculated by -

cost_fun

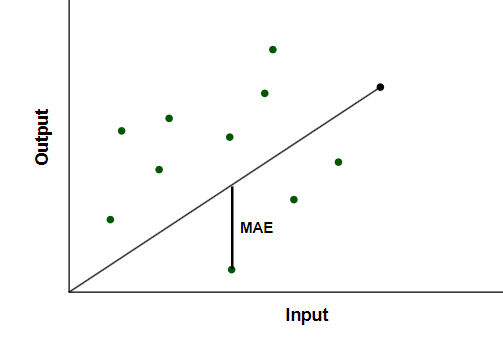
MSE should be interpreted as an error metric where the closer your value is to 0, the more accurate your model is. However, MSE is simply the average of the squared errors, meaning the resulting value will unfortunately not be understood within the context of your model target.  
There is no general rule for how to interpret given MSE values. It is an absolute value which is unique to each dataset and can only be used to say whether the model has become more or less accurate than a previous run.

**MAE (Mean Absolute Error)**

MAE is the average of absolute value between predicted and actual values. The mean absolute error is calculated by -

cost_fun_MAE

The Mean Absolute Error (MAE) serves as an indicator of the accuracy of a predictive model. A lower MAE suggests a more accurate model. However, it's important to note that the interpretation of MAE is specific to the scale of the target variable being predicted. Unlike some other metrics, MAE is returned in the same units as the target variable, making its interpretation dataset-dependent.



**Choosing Between MSE and MAE in Specific Scenarios**

The key difference between squared error and absolute error is that squared error punishes large errors to a greater extent than absolute error, as the errors are squared instead of just calculating the difference.

Let's explore situations where Mean Squared Error (MSE) is more suitable than Mean Absolute Error (MAE), and vice versa:

Use MSE instead of MAE when :

1. MSE penalizes larger errors more heavily due to the squaring of differences. If your project is particularly concerned about minimizing the impact of large errors and is more tolerant of small errors, MSE may be more appropriate.

2. When using optimization algorithms to train machine learning models, MSE can offer better numerical stability in certain cases. The squared term often leads to smoother and more well-behaved optimization landscapes.

3. MSE amplifies the differences between small and large errors. This can be beneficial when you want a metric that reflects and magnifies the variations in performance, making it easier to distinguish between models with subtle differences.

Use MAE instead of MSE when :

1. If your dataset contains outliers and you want the metric to be less influenced by extreme values, MAE is a better choice at that situation.

2. MAE provides error values in the same units as the target variable, making it more interpretable. If clear communication of the error in a way that stakeholders can easily understand is crucial, MAE is often preferred.

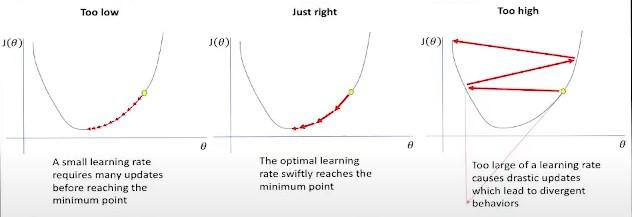
Learning Rate

Learning rate (η) is one such hyper-parameter that defines the adjustment in the weights of our network with respect to the loss gradient descent. It determines how fast or slow we will move towards the optimal weights.

When training a model, the goal is to minimize a loss function, which measures how well the model's predictions match the actual data. Gradient descent achieves this by iteratively adjusting the model's parameters in the direction that reduces the loss function. The learning rate controls how large these adjustments are.

As we know, to update the weights, we use the following function:

mnew= mold- η 𝛿E/𝛿m



**The Impact of Learning Rate on Training**

**Convergence Speed:** A well-chosen learning rate accelerates the convergence of the algorithm to the optimal solution. If the learning rate is too high, the algorithm may overshoot the minimum, causing the loss function to diverge or oscillate. Conversely, if the learning rate is too low, the convergence process will be slow, requiring many iterations to reach the optimal solution.

**Stability and Accuracy:**The stability of the training process is directly influenced by the learning rate. A high learning rate can lead to erratic updates, making the training process unstable and the model less accurate. A low learning rate, while stable, might cause the model to get stuck in local minima, potentially leading to suboptimal performance.

**Generalization:** The learning rate also affects the model's ability to generalize to new data. A well-tuned learning rate helps the model find a balance between fitting the training data and maintaining the ability to perform well on unseen data.

**Choosing the Right Learning Rate**

Selecting an appropriate learning rate is often a matter of experimentation and experience. Here are some common strategies:

**Grid Search:** This involves testing a range of learning rates, often on a logarithmic scale, to identify the best performing value. While this can be computationally expensive, it provides a systematic approach to finding a suitable learning rate.

**Learning Rate Schedulers:**Adaptive learning rate methods adjust the learning rate during training. Techniques like learning rate decay, where the learning rate is reduced over time, or more sophisticated approaches like ReduceLROnPlateau, which reduces the learning rate when a metric has stopped improving, can help maintain an optimal learning rate throughout training.

**Cyclical Learning Rates:**This method involves varying the learning rate cyclically between a minimum and maximum value during training. This approach can help the model escape local minima and explore the loss surface more effectively.

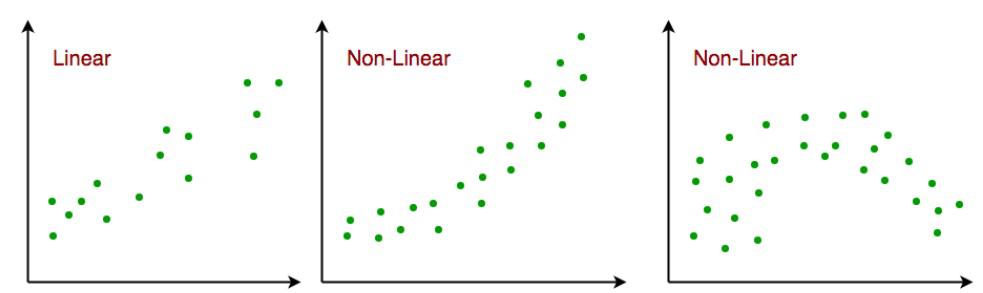
**Adaptive Optimizers:**Optimizers such as Adam, RMSprop, and AdaGrad adjust the learning rate for each parameter dynamically based on the gradients. These optimizers often require less manual tuning of the learning rate and can lead to faster convergence.

Crucial Assumptions in Linear Regression: What You Need to Know

Linear regression is a powerful statistical method widely used for modeling the relationship between a dependent variable and one or more independent variables. However, like any statistical technique, linear regression comes with a set of assumptions that must be met for the results to be valid and reliable. These assumptions play a crucial role in the interpretation of the regression analysis. The key assumptions of linear regression are as follows:

**Linearity:**

The fundamental assumption of linear regression is that there exists a linear relationship between the independent variables and the dependent variable. This means that the change in the mean of the dependent variable is proportional to a change in the independent variable(s). It is essential to check for linearity by examining scatter plots and ensuring that the relationship follows a straight line.



**Concept:**The relationship between the dependent variable Y and the independent variable(s) X is represented as a linear combination.

Formula:

Y=β0+β1X1+β2X2+…+βnXn+ϵ

where Y is the dependent variable, β0 is the intercept, β1,β2,…,βnare the coefficients, X1,X2,…,Xn are the independent variables, and ϵ represents the error term.

**Independence:**

Another critical assumption is the independence of observations. Each data point should be independent of others, meaning that the value of the dependent variable for one observation should not be influenced by the values of the independent variables for other observations. This assumption is often violated in time series data or repeated measures studies, requiring special attention and techniques.

**Concept:**Each observation is independent of others.

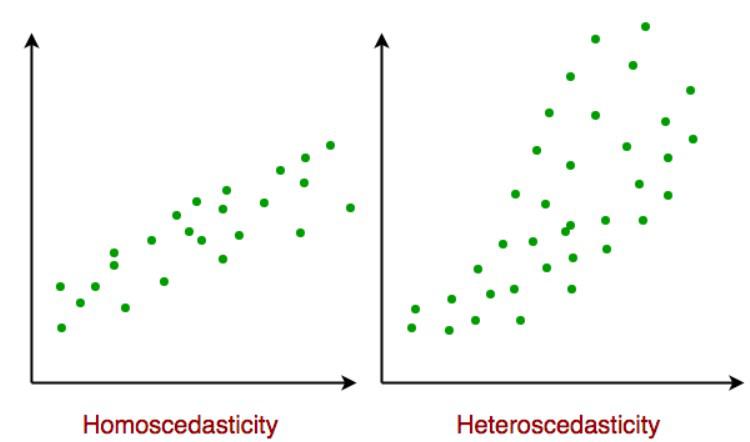
Formula:

Cov(ϵi,ϵj)=0

for all i ≠ j, where Cov denotes the covariance between two error terms.

**Homoscedasticity:**

Homoscedasticity refers to the constant variance of errors across all levels of the independent variable(s). In other words, the spread of residuals should be consistent throughout the range of predictor values. Heteroscedasticity, where the variance of errors is not constant, can lead to inefficient parameter estimates and affect the reliability of hypothesis tests. Residual plots are commonly used to assess homoscedasticity.



**Concept:**The variance of errors is constant across all levels of the independent variable(s).

Formula:

Var(ϵi)=σ2

for all i, where Var denotes the variance, and σ2 is a constant.

**Normality of Residuals:**

The assumption of normality pertains to the distribution of the residuals, which should ideally be normally distributed. While normality is not critical for large sample sizes due to the Central Limit Theorem, deviations from normality in small samples may impact the validity of statistical inferences. Techniques like the Shapiro-Wilk test or normal probability plots can be employed to assess normality.

**Concept:** The residuals follow a normal distribution.

Formula:

This can be checked using statistical tests like the Shapiro-Wilk test or by examining a normal probability plot.

**No Perfect Multicollinearity:**

Multicollinearity occurs when two or more independent variables in the regression model are highly correlated. This can create issues in estimating the individual contributions of each variable to the dependent variable. Variance inflation factor (VIF) and correlation matrices are common tools to detect multicollinearity, and corrective measures may involve excluding variables or combining them.

**Concept:** The independent variables are not perfectly correlated.

Formula:

Check for high correlation coefficients between independent variables and calculate the Variance Inflation Factor (VIF).

VIF= 1 / (1−R2)

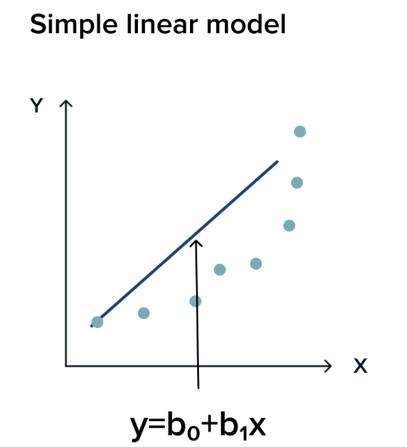
where R2 is the coefficient of determination from the regression of one independent variable against the others.

Types of Linear Regression

Linear regression is a versatile statistical technique used for modeling the relationship between a dependent variable and one or more independent variables. There are various types of linear regression models, each designed to address specific scenarios and data characteristics.

**1. Simple Linear Regression:**

Simple linear regression involves predicting a dependent variable based on a single independent variable.



Formula:

Y=β0 +β1 X+ϵ

where,

* Y is the dependent variable
* X is the independent variable
* β0 is the intercept
* β1 is the slope
* ϵ is the error term.

**Example:** predicting a student's test score (Y) based on the number of hours spent studying (X).

**Multiple Linear Regression:**

Multiple linear regression extends simple linear regression by incorporating multiple independent variables.

Formula:

Y=β0 +β1 X1 +β2X2 +…+βn Xn +ϵ

where,

* X1 ,X2 ,…,Xn are the independent variables
* β1 ,…,βn are their respective coefficients.
* β0 is the intercept

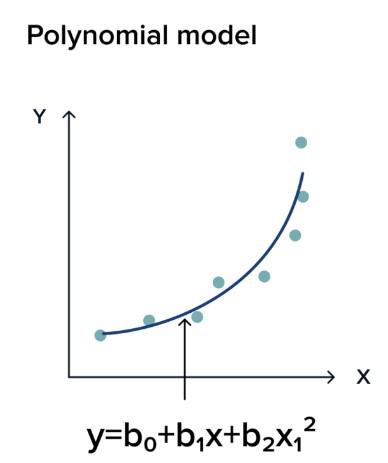
The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

**Example:** Predicting house prices (Y) based on variables like square footage (X1), number of bedrooms (X2), and location (X3)

**Polynomial Regression:**

Polynomial regression captures non-linear relationships by including polynomial terms of the independent variable.



Formula:

Y=β0+β1X+β2X2+…+βnXn+ϵ

where,

* X is the independent variable
* n determines the degree of the polynomial.
* β1 ,…,βn are their respective coefficients.

**Example:** Modeling the trajectory of a projectile (Y) based on time (X) using a quadratic equation.

Model Evaluation Metric: R-square vs Adjusted R-square

When assessing the performance of regression models, R-square (R²) and Adjusted R-square are commonly used metrics. Both provide insights into how well the model fits the data, but they have distinct purposes. Let's delve into the details of these metrics to understand their strengths and limitations.

**Understanding R-square (R²)**

Definition:

R-square measures the proportion of the variance in the dependent variable (let's say, house prices) that is explained by the independent variables (features like square footage, bedrooms, and neighborhood characteristics).

Formula:

R2= 1− (Sum of Squared Residuals)/(Total Sum of Squares)

* **Sum of Squared Residuals:**The sum of the squared differences between predicted and actual values.
* **Total Sum of Squares:**The sum of the squared differences between actual values and the mean of the dependent variable.

Interpretation:

* R2 ranges from 0 to 1.
* A higher R2 value indicates a better fit.
* However, R2 has limitations, particularly when it comes to model complexity and the addition of unnecessary predictors.

**Residuals and R2**

Residuals (e) represent the differences between actual and predicted values. In the context of house price prediction, residuals help us understand how well the model captures the true variability in house prices.

e=Actual Value−Predicted Value

**Introduction to Adjusted R-square**

While R2 provides a measure of fit, it doesn't account for the number of predictors in the model. Adjusted R-square addresses this limitation by penalizing the inclusion of unnecessary predictors.

Formula:

Adjusted R2 = 1 − (1−R2)⋅(n−1)/(n−k−1)

* **n:** Number of observations
* **k:** Number of predictors

**Practical Application - House Price Prediction**

Let's consider a house price prediction project where we use features like square footage, bedrooms, and neighborhood characteristics to predict house prices. After building a regression model, we calculate R2 and Adjusted R2 to evaluate its performance.

**Choosing Between R2 and Adjusted R2**

* R2 is useful for comparing models with the same predictors.
* Adjusted R2 is preferred when comparing models with different numbers of predictors, helping to identify the most parsimonious model.

**Conclusion**

In house price prediction projects, the choice between R2 and Adjusted R2 depends on the specific goals of the analysis. While R2 gives a general sense of fit, Adjusted R2 considers model complexity.

Understanding these metrics, along with residuals and graphical representations, provides a holistic approach to model evaluation. This knowledge is invaluable in making informed decisions when building regression models for predicting house prices or any other dependent variable.