Linear Regression in ML

**Linear Regression in Machine learning**

Linear regression is a type of [**supervised machine-learning algorithm**](https://www.geeksforgeeks.org/supervised-machine-learning/) that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It assumes that there is a linear relationship between the input and output, meaning the output changes at a constant rate as the input changes. This relationship is represented by a straight line.

**For example** we want to predict a student's exam score based on how many hours they studied. We observe that as students study more hours, their scores go up. In the example of predicting exam scores based on hours studied. Here

* **Independent variable (input):** Hours studied because it's the factor we control or observe.
* **Dependent variable (output):**Exam score because it depends on how many hours were studied.

We use the independent variable to predict the dependent variable.

**Best Fit Line in Linear Regression**

In linear regression, the best-fit line is the straight line that most accurately represents the relationship between the independent variable (input) and the dependent variable (output). It is the line that minimizes the difference between the actual data points and the predicted values from the model.

**1. Goal of the Best-Fit Line**

The goal of linear regression is to find a straight line that minimizes the error (the difference) between the observed data points and the predicted values. This line helps us predict the dependent variable for new, unseen data.



Linear Regression

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

**2. Equation of the Best-Fit Line**

For simple linear regression (with one independent variable), the best-fit line is represented by the equation

*y=mx+b*

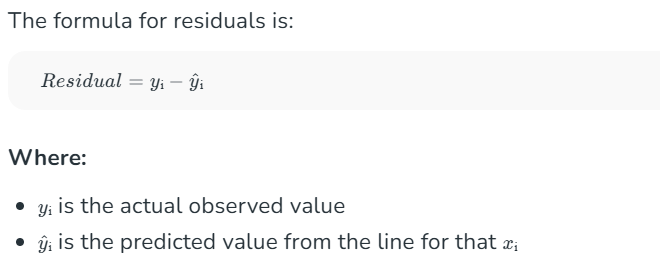
**Where:**

* **y** is the predicted value (dependent variable)
* **x** is the input (independent variable)
* **m** is the slope of the line (how much y changes when x changes)
* **b** is the intercept (the value of y when x = 0)

The best-fit line will be the one that optimizes the values of m (slope) and b (intercept) so that the predicted y values are as close as possible to the actual data points.

**3. Minimizing the Error: The Least Squares Method**

To find the best-fit line, we use a method called [**Least Squares**](https://www.geeksforgeeks.org/least-square-method/). The idea behind this method is to minimize the sum of squared differences between the actual values (data points) and the predicted values from the line. These differences are called residuals.

**

The least squares method minimizes the sum of the squared residuals:

**

This method ensures that the line best represents the data where the sum of the squared differences between the predicted values and actual values is as small as possible.

**4. Interpretation of the Best-Fit Line**

* **Slope (m):** The slope of the best-fit line indicates how much the dependent variable (y) changes with each unit change in the independent variable (x). For example if the slope is 5, it means that for every 1-unit increase in x, the value of y increases by 5 units.
* **Intercept (b):** The intercept represents the predicted value of y when x = 0. It’s the point where the line crosses the y-axis.

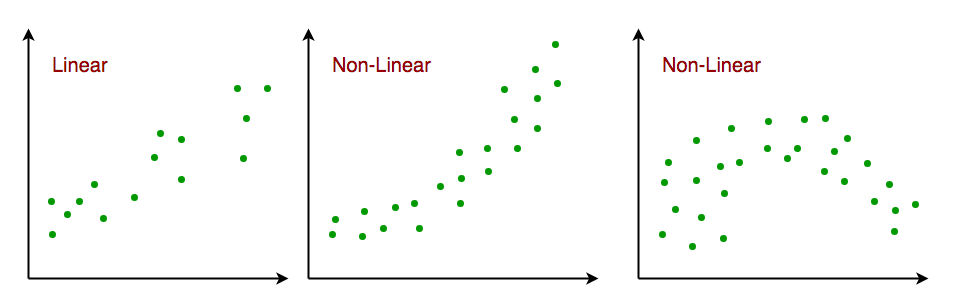
In linear regression some hypothesis are made to ensure reliability of the model's results.

***Limitations***

* ***Assumes Linearity:*** *The method assumes the relationship between the variables is linear. If the relationship is non-linear, linear regression might not work well.*
* ***Sensitivity to Outliers:*** *Outliers can significantly affect the slope and intercept, skewing the best-fit line.*

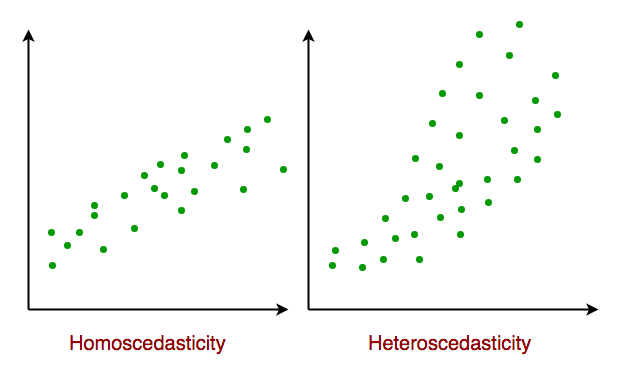
**Assumptions of the Linear Regression**

**1. Linearity**: The relationship between inputs (X) and the output (Y) is a straight line.

Linearity

**2. Independence of Errors**: The errors in predictions should not affect each other.

**3. Constant Variance (Homoscedasticity):**The errors should have equal spread across all values of the input. If the spread changes (like fans out or shrinks), it's called heteroscedasticity and it's a problem for the model.

Homoscedasticity

**4. Normality of Errors**: The errors should follow a normal (bell-shaped) distribution.

**5. No Multicollinearity*(*for multiple regression)**: Input variables shouldn’t be too closely related to each other.

**6. No Autocorrelation**: Errors shouldn't show repeating patterns, especially in time-based data.

**7. Additivity**: The total effect on Y is just the sum of effects from each X, no mixing or interaction between them.'

*To understand Multicollinearityin detail refer to article:* [***Multicollinearity***](https://www.geeksforgeeks.org/multicollinearity-in-regression-analysis/)*.*

[*https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/*](https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/) *- types of regressions*

**Types of Linear Regression**

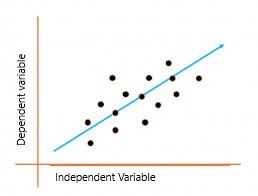
When there is only one independent feature it is known as Simple Linear Regression or [Univariate Linear Regression](https://www.geeksforgeeks.org/univariate-linear-regression-in-python/) and when there are more than one feature it is known as Multiple Linear Regression or [Multivariate Regression](https://www.geeksforgeeks.org/multivariate-regression/).

**1. Simple Linear Regression**

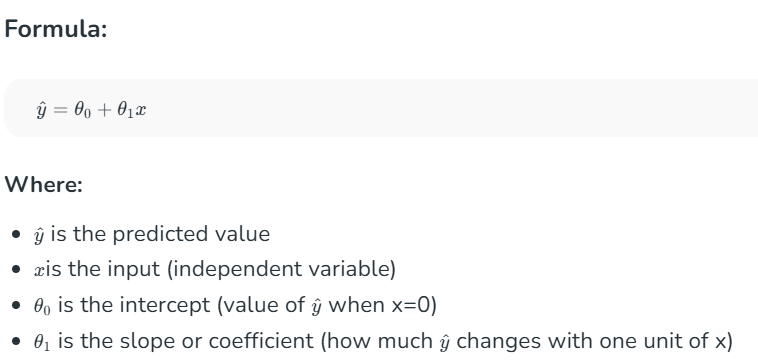
[Simple linear regression](https://www.geeksforgeeks.org/simple-linear-regression-in-python/) is used when we want to predict a target value (dependent variable) using only one input feature (independent variable). It assumes a straight-line relationship between the two.

In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.

Linear regression is a quiet and the simplest statistical regression technique used for predictive analysis in [**machine learning**](https://www.analyticsvidhya.com/blog/2022/01/machine-learning-algorithms/). It shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent (output) variable i.e. Y-axis, called linear regression. If there is a single input variable X (independent variable), such linear regression is ***simple linear regression***.



The graph above presents the linear relationship between the output(y) and predictor(X) variables. The blue line is referred to as the best-fit straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

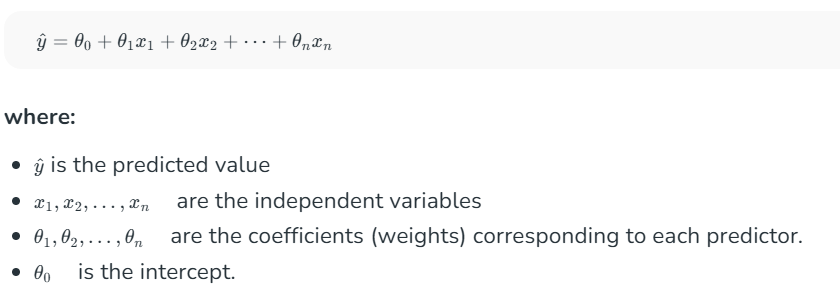
****

**Example:**

Predicting a person’s salary (y) based on their years of experience (x).

**2. Multiple Linear Regression**

[Multiple linear regression](https://www.geeksforgeeks.org/ml-multiple-linear-regression-using-python/) involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

**

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.

**Use Case of Multiple Linear Regression**

Multiple linear regression allows us to analyze relationship between multiple independent variables and a single dependent variable. Here are some use cases:

* **Real Estate Pricing:** In real estate MLR is used to predict property prices based on multiple factors such as location, size, number of bedrooms, etc. This helps buyers and sellers understand market trends and set competitive prices.
* **Financial Forecasting:** Financial analysts use MLR to predict stock prices or economic indicators based on multiple influencing factors such as interest rates, inflation rates and market trends. This enables better investment strategies and risk management24.
* **Agricultural Yield Prediction:** Farmers can use MLR to estimate crop yields based on several variables like rainfall, temperature, soil quality and fertilizer usage. This information helps in planning agricultural practices for optimal productivity
* **E-commerce Sales Analysis:** An e-commerce company can utilize MLR to assess how various factors such as product price, marketing promotions and seasonal trends impact sales.

Now that we have understood about linear regression, its assumption and its type now we will learn how to make a linear regression model.

**Cost function for Linear Regression**

As we have discussed earlier about best fit line in linear regression, its not easy to get it easily in real life cases so we need to calculate errors that affects it. These errors need to be calculated to mitigate them. The difference between the predicted value Y hat (y^)     and the true value Y and it is called [cost function](https://www.geeksforgeeks.org/what-is-cost-function/) or the[loss function](https://www.geeksforgeeks.org/ml-common-loss-functions/).

In Linear Regression, the Mean Squared Error (MSE) cost function is employed, which calculates the average of the squared errors between the predicted values y*i hat*​ and the actual values yi. The purpose is to determine the optimal values for the intercept θ1*θ*1​ and the coefficient of the input feature θ2*θ*2​ providing the best-fit line for the given data points. The linear equation expressing this relationship is .

MSE function can be calculated as:

**

Utilizing the MSE function, the iterative process of gradient descent is applied to update the values of \θ1&θ2*θ*1​&*θ*2​. This ensures that the MSE value converges to the global minima, signifying the most accurate fit of the linear regression line to the dataset.

This process involves continuously adjusting the parameters \(\theta\_1\) and \(\theta\_2\) based on the gradients calculated from the MSE. The final result is a linear regression line that minimizes the overall squared differences between the predicted and actual values, providing an optimal representation of the underlying relationship in the data.

Now we have calculated loss function we need to optimize model to mtigate this error and it is done through gradient descent.

**Considerations of Multiple Linear Regression**

All four assumptions made for Simple Linear Regression still hold for Multiple Linear Regression along with a few new additional assumptions.

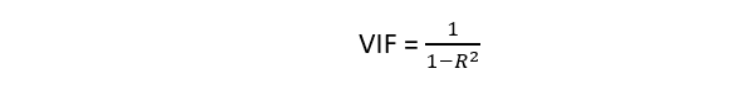
* **Overfitting**: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the overfitting of a model. This usually leads to high training accuracy and very low test accuracy.
* **Multicollinearity**: It is the phenomenon where a model with several independent variables, may have some variables interrelated.
* **Feature Selection:** With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

**Multicollinearity**

As [**multicollinearity**](https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/) makes it difficult to find out which variable is contributing towards the prediction of the response variable, it leads one to conclude incorrectly, the effects of a variable on the target variable.  Though it does not affect the precision of the model predictions, it is essential to properly detect and deal with the multicollinearity present in the model, as random removal of any of these correlated variables from the model causes the coefficient values to swing wildly and even change signs.

Multicollinearity can be detected using the following methods.

* **Pairwise Correlations:** Checking the pairwise correlations between different pairs of independent variables can throw useful insights into detecting multicollinearity.
* **Variance Inflation Factor (VIF):** Pairwise correlations may not always be useful as it is possible that just one variable might not be able to completely explain some other variable but some of the variables combined could be ready to do this. Thus, to check these sorts of relations between variables, one can use VIF. VIF explains the relationship of one independent variable with all the other independent variables. VIF is given by,



*where i* refers to the*ith* variable which is being represented as a linear combination of the rest of the independent variables.

The common heuristic followed for the VIF values is if VIF > 10 then the value is high and it should be dropped. And if the VIF=5 then it may be valid but should be inspected first. If VIF < 5, then it is considered a good VIF value.

**Overfitting and Underfitting in Linear Regression**

There have always been situations where a model performs well on training data but not on the test data. While training models on a dataset, overfitting, and underfitting are the most common problems faced by people.

Before understanding overfitting and underfitting one must know about bias and variance.

**Bias**

Bias is a measure to determine how accurate a model’s predictions are likely to be on future unseen data. Complex models, assuming there is enough training data available, can make accurate model predictions. Whereas the models that are too naive, are very likely to perform badly concerning model predictions. Simply, Bias is errors made by training data.

Generally, linear algorithms have a high bias which makes them fast to learn and easier to understand but in general, are less flexible. Implying lower predictive performance on complex problems that fail to meet the expected outcomes.

**Variance**

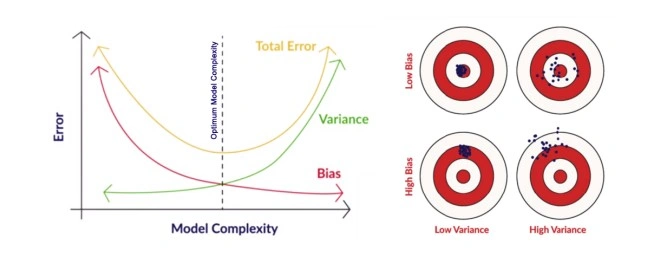
Variance is the sensitivity of the model towards training data, that is it quantifies how much the model will react when input data is changed.

Ideally, the model shouldn’t change too much from one training dataset to the next training data, which means that the algorithm is good at picking out the hidden underlying patterns between the inputs and the output variables.

Ideally, a model should have lower variance which means that the model doesn’t change drastically after changing the training data(it is generalizable). Having higher variance will make a model change drastically even on a small change in the training dataset.

Let’s understand what a bias-variance tradeoff is.

**Bias Variance Tradeoff**



In the pursuit of optimal performance, a supervised machine learning algorithm seeks to strike a balance between low bias and low variance for increased robustness.

In the realm of machine learning, there exists an inherent relationship between bias and variance, characterized by an inverse correlation.

* Increased bias leads to reduced variance.
* Conversely, heightened variance results in diminished bias.

Finding an equilibrium between bias and variance is crucial, and algorithms must navigate this trade-off for optimal outcomes.

In practice, calculating precise bias and variance error terms is challenging due to the unknown nature of the underlying target function.

Now, let’s delve into the nuances of [**overfitting and underfitting**](https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/).

**Overfitting**

When a model learns every pattern and noise in the data to such an extent that it affects the performance of the model on the unseen future dataset, it is referred to as ***overfitting***. The model fits the data so well that it interprets noise as patterns in the data.

When a model has low bias and higher variance it ends up memorizing the data and causing overfitting.  Overfitting causes the model to become specific rather than generic. This usually leads to high training accuracy and very low test accuracy.

Detecting overfitting is useful, but it doesn’t solve the actual problem. There are several ways to prevent overfitting, which are stated below:

* Cross-validation
* If the training data is too small to train add more relevant and clean data.
* If the training data is too large, do some feature selection and remove unnecessary features.
* Regularization

**Underfitting**

Underfitting is not often discussed as often as overfitting is discussed. When the model fails to learn from the training dataset and is also not able to generalize the test dataset, is referred to as ***underfitting***. This type of problem can be very easily detected by the performance metrics.

When a model has high bias and low variance it ends up not generalizing the data and causing underfitting. It is unable to find the hidden underlying patterns in the data. This usually leads to low training accuracy and very low test accuracy. The ways to prevent underfitting are stated below,

* Increase the model complexity
* Increase the number of features in the training data
* Remove noise from the data.

**Assumptions of Linear Regression:**

What is **The equal variance of residuals (Homoscedasticity, Heteroscedasticity)**

**Normal distribution of residuals**

**Independence of residuals:**

**Linearity of residuals**:

**Full Derivation of simple linear regression**

**What is the Best Fit Line?**

In simple terms, the best-fit line is a line that best fits the given scatter plot. Mathematically, you obtain the best-fit line by minimizing the Residual Sum of Squares (RSS).

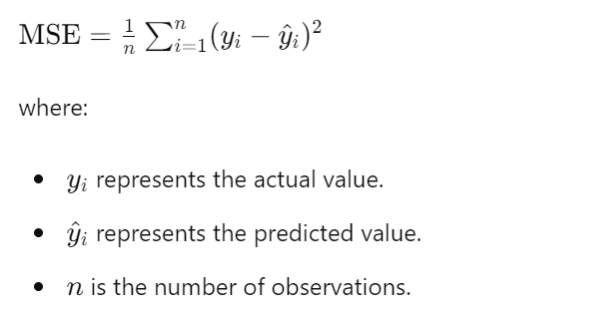
**Cost Function for Linear Regression**

The [**cost function**](https://www.analyticsvidhya.com/blog/2021/03/data-science-101-introduction-to-cost-function/) helps to work out the optimal values for B 0  and B 1 , which provides the best-fit line for the data points.

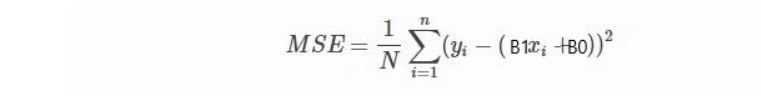
In Linear Regression, generally [**Mean Squared Erro**](https://www.analyticsvidhya.com/blog/2024/07/mean-squared-error/)[**r (MSE)**](https://www.analyticsvidhya.com/blog/2024/07/mean-squared-error/) cost function is used, which is the average squared error that occurred between the **y predicted**and **y i**.

We calculate MSE using the simple linear equation

y=mx+b



Where y^ = B1xi + B0;



Using the MSE function, we’ll update the values of B 0  and B 1  such that the MSE value settles at the minima.  These parameters can be determined using the gradient descent method such that the value for the cost function is minimum.

**Regularisation in linear regression**

<https://www.analyticsvidhya.com/blog/2021/11/study-of-regularization-techniques-of-linear-model-and-its-roles/>

**derivation of simple linear regression:**

**What is Linear Regression?**  
Linear Regression is a supervised Machine learning algorithm. Linear regression is a type of regression analysis where there is a linear relationship between the independent(x) and dependent(y) variables. There can be positive Linear Relationship or Negative Linear Relationship

**The objective of Linear Regression :**  
To fit a **best-fit line** in such a manner that the differences between the distance of the actual data points from the plotted curve/line are minimum

**Types of Linear Regression**  
1 Simple Linear Regression  
2 Multiple Linear Regression  
3 Polynomial Linear Regression

**Simple Linear Regression**

It is an approach for predicting Y (dependent variable) on the basis of the single independent variable  
It assumes that there is approximately a linear relationship between X and Y

this is given by **Y ≈ β1X + β0**  
where,  
X is the independent variable  
Y is the dependent variable  
β1 is the slop (or coefficient or weight)  
β0 is y-intercept when x = 0, (or offset)

**β0 and β1** are two unknown constants that represent intercept and slop terms in the linear model.  
This **β0 and β1** are known as model coefficients or **model parameters**

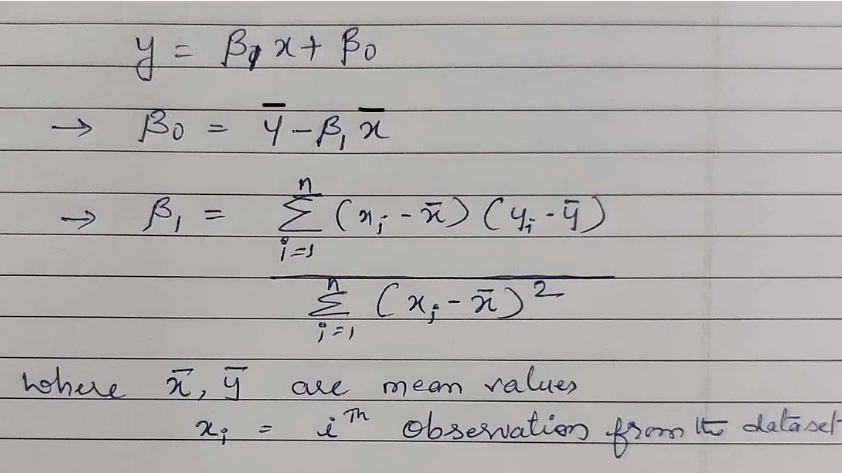
Once we train the model with our training data to produce estimates β0 and β1 we can predict the future values on the basis of particular x value by computing **yhat = β0hat + β1hat \* x**

**Mathematical Intuition**  
*How to find the values of Coefficients (β0 and β1)*

The coefficients can be found in two different ways  
**Closed-form**  
1. Uses direct formula — [Wikipedia](https://en.wikipedia.org/wiki/Closed-form_expression)  
2. also called as **Ordinary least square** ( LinearRegression() in sci-kit learn uses OLS method by default)  
**Non-Closed-form** — uses differentiation  
— Solved by Gradient decent (SGDRegressor() in sci-kit learn uses gradient descent)

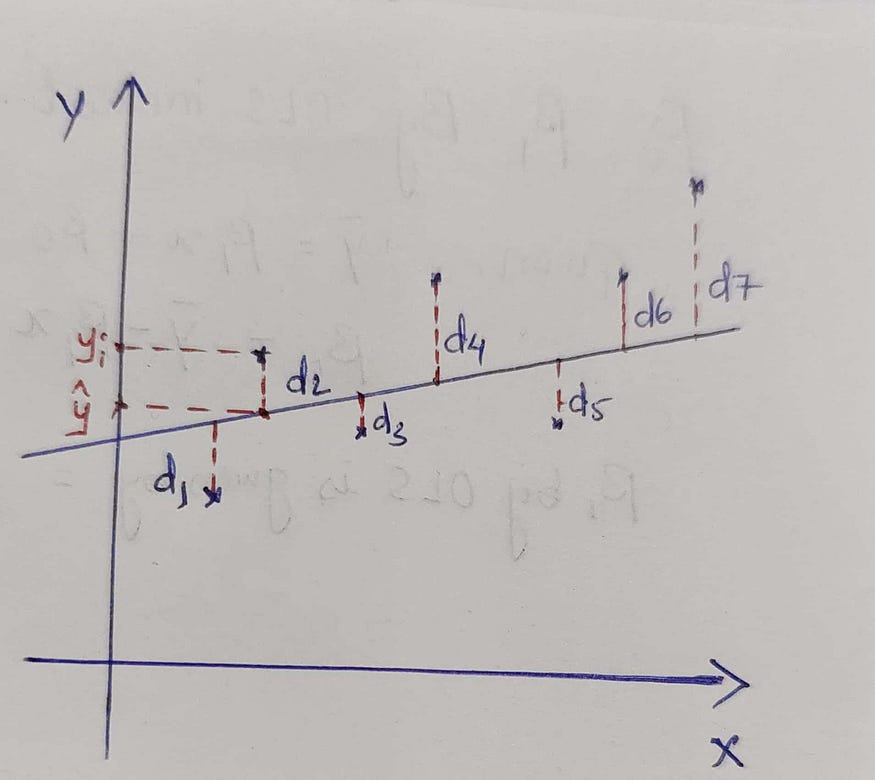
***Why are we using Gradient Decent instead of the direct formula from OLS****?*  
— When the dimensionality increases the complexity of calculation increases with the OLS formula. OLS is considered when the dataset is very small.

**Ordinary Least Square Method (Closed Form)**  
OLS method tries to find the β0 and β1 which minimizes the sum of squared error



*Let's build this equation from scratch*

Assuming that independent and dependent variables are linear in nature, we try to fit “**best line of fit”**that tries to pass through all the data points as close as possible to reduce the error (also called residuals)



Here d1,d2,d3,d4….dn represents errors(residuals)  
then, Error can be written as

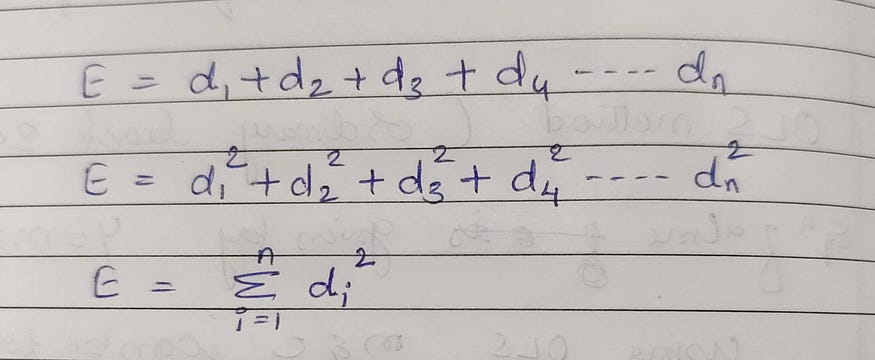
𝑑1+𝑑2+𝑑3+𝑑4+….+𝑑𝑛

Errors can be negative or positive values, this makes errors cancel out at some point. To overcome this issue and find the total errors, we square them

***Why modulus is not used to convert the errors to positive values?***

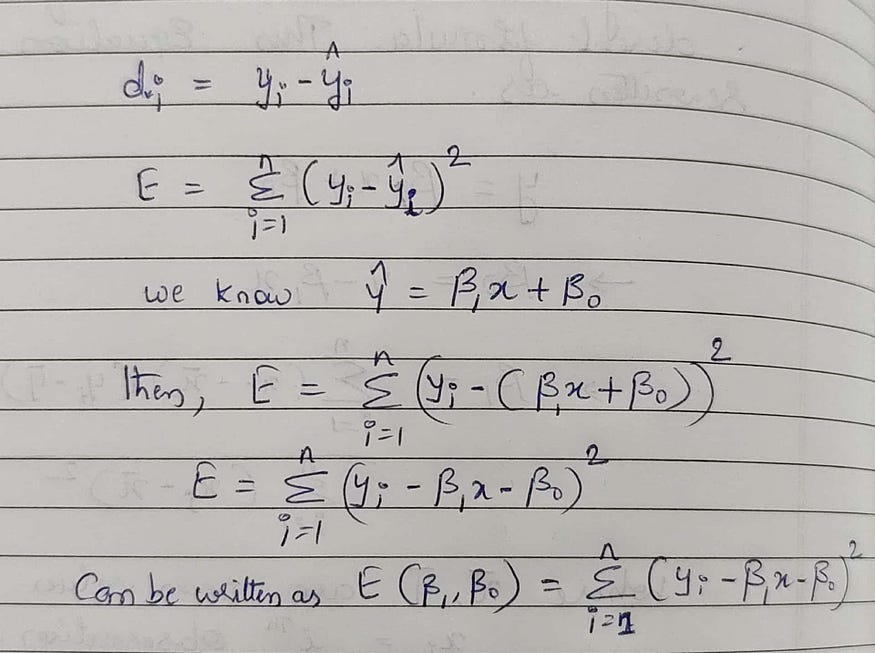
There are two reasons,  
1. we want to penalize outliers  
2. Mod can not be differentiable at the origin

now the Error term can be written as,



This E is now referred to as **Error Function** or **Cost Function**(You will see Error term is also represented by **J** in some books)

Decomposing the error term furthermore,

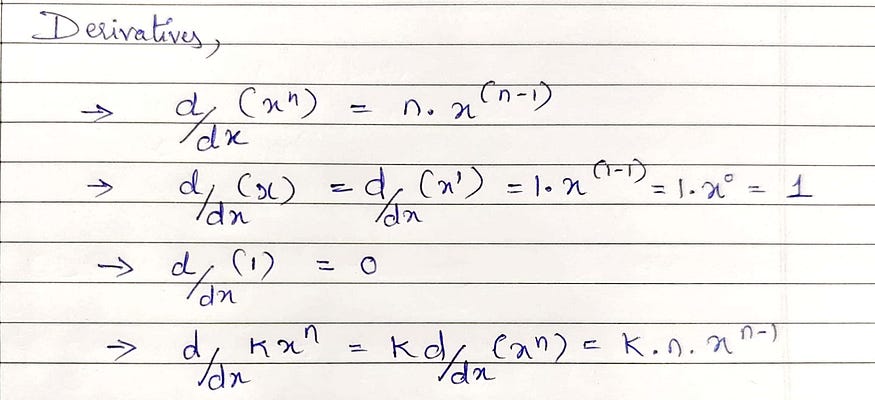


Now we have the Error function decomposed, we have to now find the values of **β0**and **β1** which minimizes the error.

From theory we know, y = f(x), meaning, y is a function of x, that is, a change in x will change the values of y.

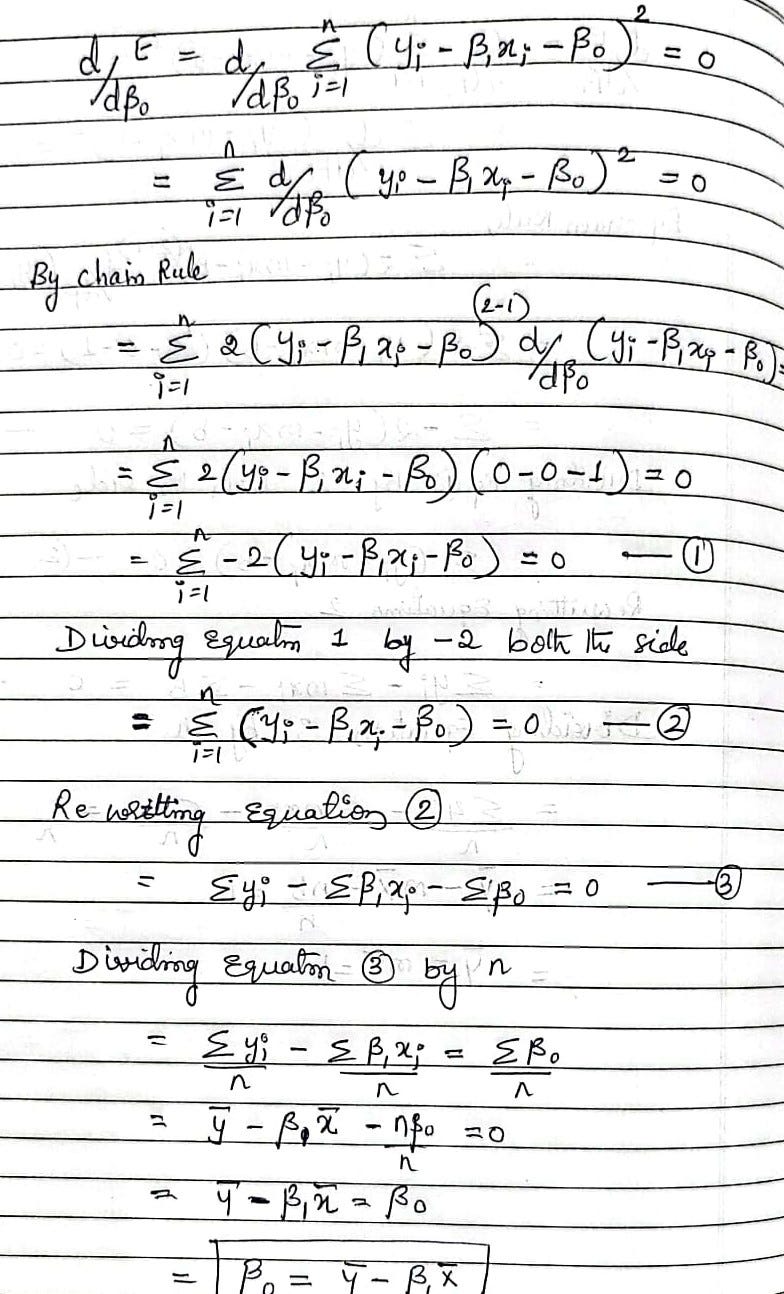
Similarly, a change in β0 and β1 will change the values of error and finds the minimum errors.

From maths, we know,

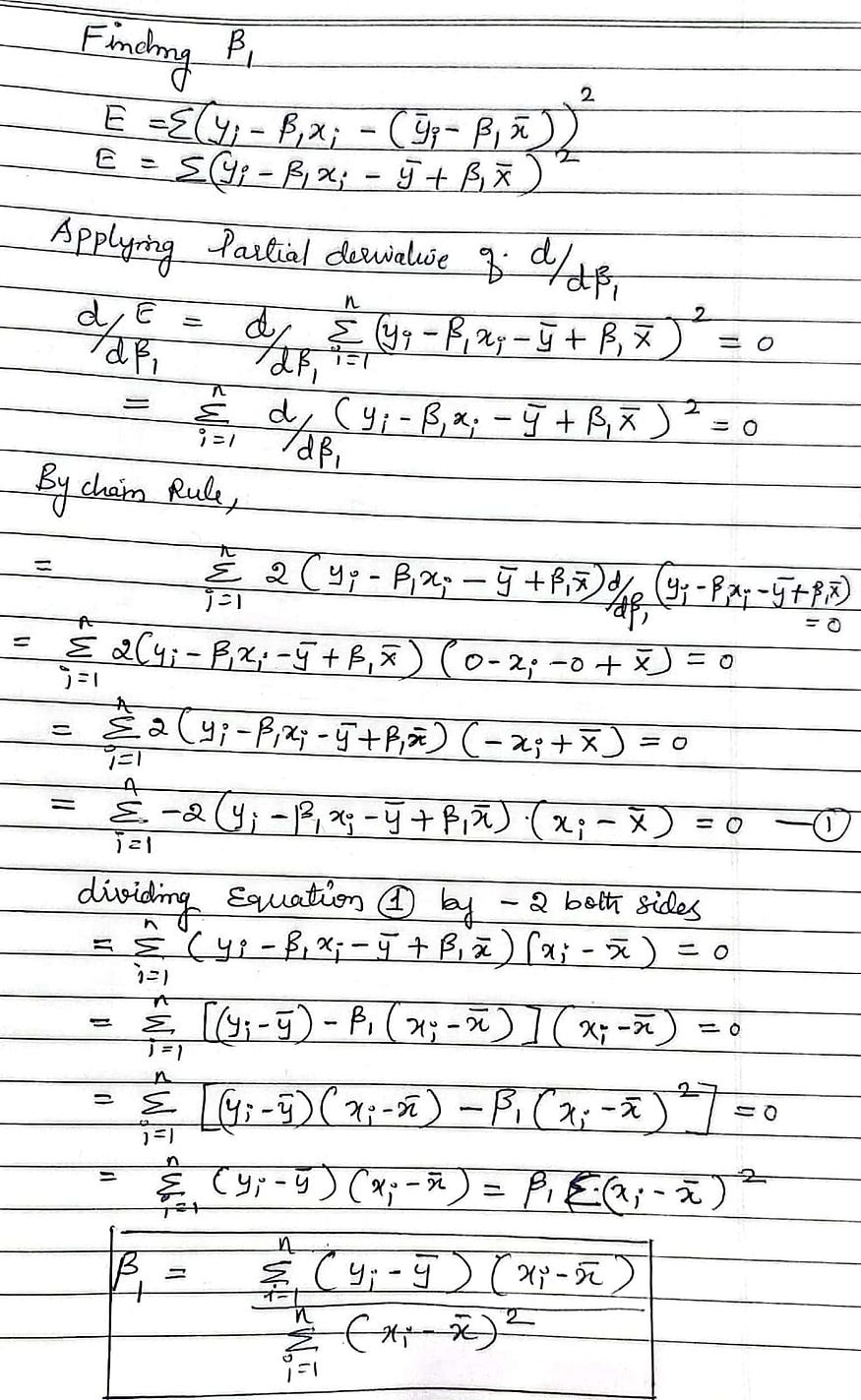


To find the minimum value of the error function, we have to carry out two steps,  
1. Find the partial derivatives of errors with respect to β0 and β1  
2. set up the derivatives to 0 and find the critical points, where the slop of the tangent line becomes 0

*Finding β0*



*Finding β1*



Link- <https://git.io/JK6sS>

**Full Derivation of multiple linear regression**

***What is Multiple Linear Regression?***

Multiple Linear Regression is a supervised machine learning algorithm. This algorithm uses more than one independent variable to predict a target variable. The below table is an example of a multiple linear regression problem.

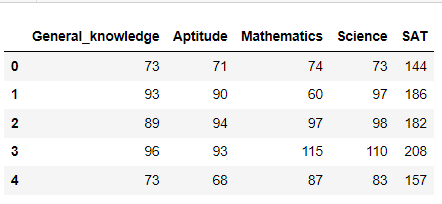
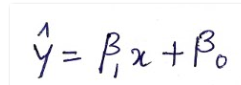


Table with 4 independent variables and 1 dependent variable

The above table has marks details and SAT score of 5 students. Each subject is a independent variable( as called predictors,explanatory variable or regressor (in regression problem)) and one dependent variable (also called label,target,response variable). Since we have more than one independent variable we have to find the **coefficients**for each independent variable. In my [Simple Linear Regression](https://medium.com/@bhanu0925/deriving-equation-for-simple-linear-regression-ols-method-44d4f7d38dc1) article, we have discussed how to find the coefficients. Now we will extent simple linear to multiple linear and find the coefficients.

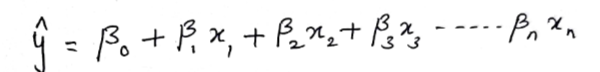
**Mathematical Derivation**

We know the equation of a line for simple linear regression is



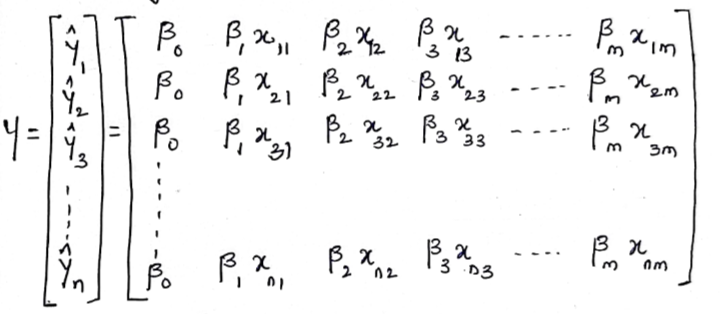
Equation of a line

Similarly, now for the multiple linear regression, the formula will be,



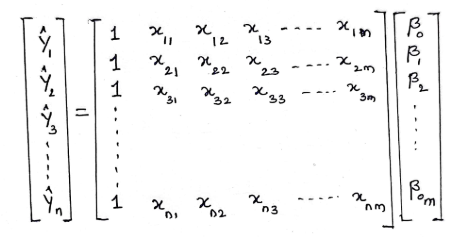
Equation of a plane with multiple independent variables

The above formula can be rewritten in matrix form as,



Matrix representation for n records

This can be further rewritten as,

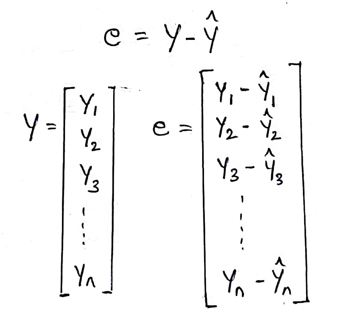


Now the above matrix can be represented as (all capital letters represents Matix in further equations)

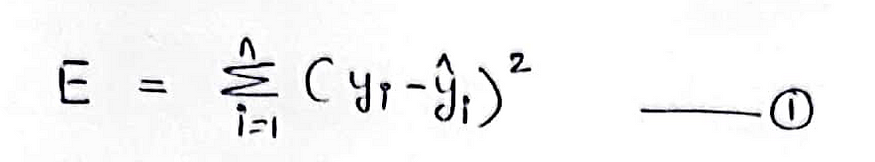


matrix representation

we know, the error is given by (Actual — Predicted), So the error can be represented in matrix form as below.



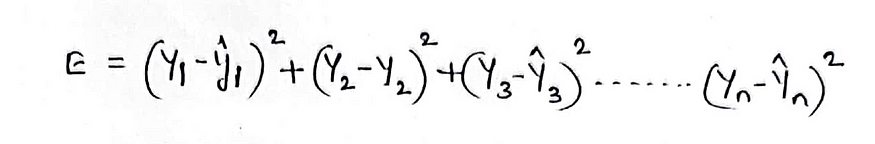
we know, for linear regression, for n observations, the **error/cost function** is given by,



Error/cost function

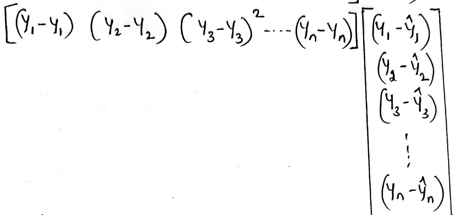
Lets, derive an equation for the error function to represent in matrix form

Expanding the equation 1, we can rewrite it as,



Expanding the error function

The above equation can be rewritten in matrix form as,



Representing error function in matrix form

Hence, the Error/Cost function in matrix notation is given by,

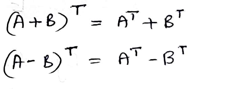


Error Function in Matrix notation

Now the error equation becomes,



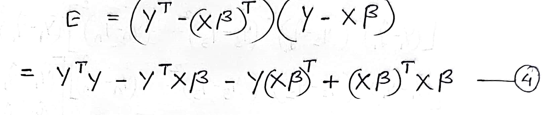
By Linear algebra,



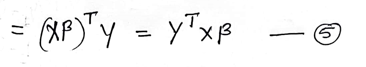
Substituting the values in equation 2, equation 2 can be rewritten as,



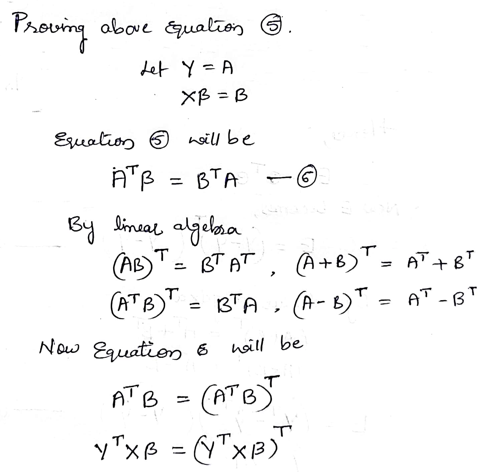
Replacing, Yhat = Xβ in equation 3,



To simplify equation 4, we have to prove the following term as equal



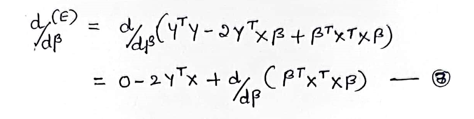
solving equation 5,



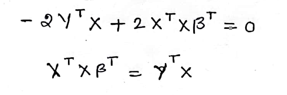
after simplifying, equation 4 becomes,



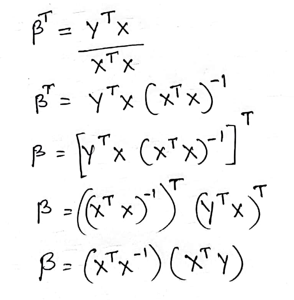
Applying partial derivatives concerning β,



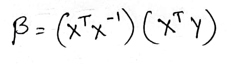
By applying matrix differentiation on equation 8,



simplifying the equation further,



After simplifying we have got a β coefficient matrix.



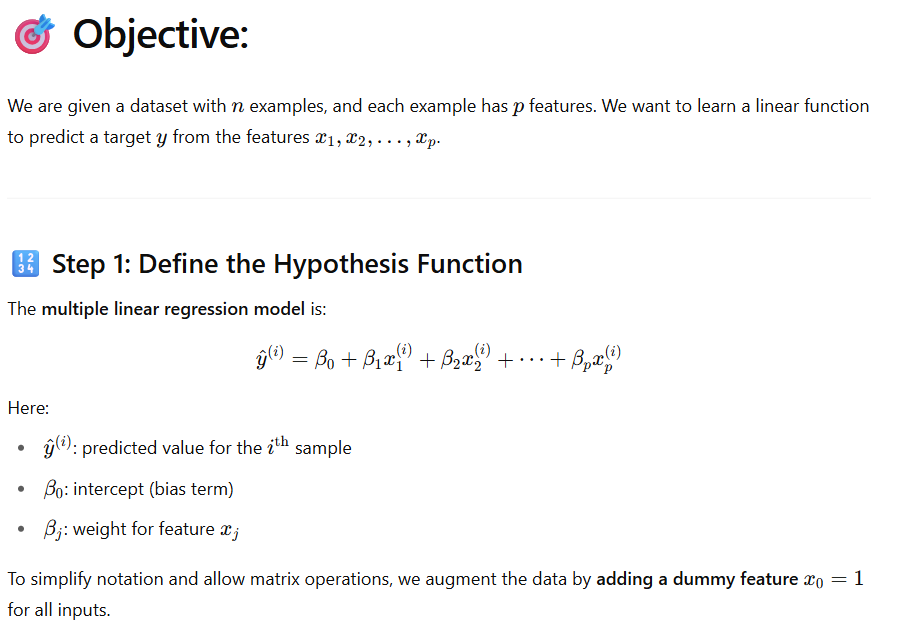
In the above equation X represents X\_train, Y represents y\_train from the dataset. Plugging those values into the above equations will give us the β coefficient matrix. This matrix will have (independent variables + 1) different values. Suppose if your data set has 5 independent variables, your coefficients will be 6.

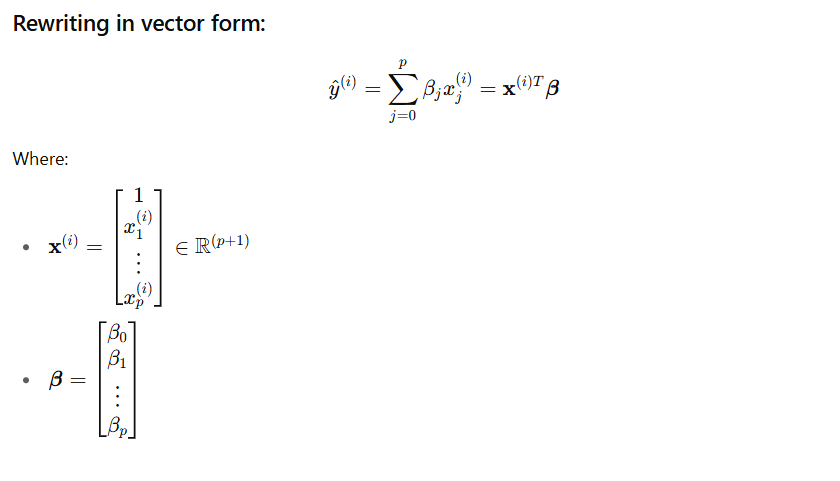
Predictions can be made using X\_Test and β coefficient. That will be the dot product of X\_test and β coefficient.

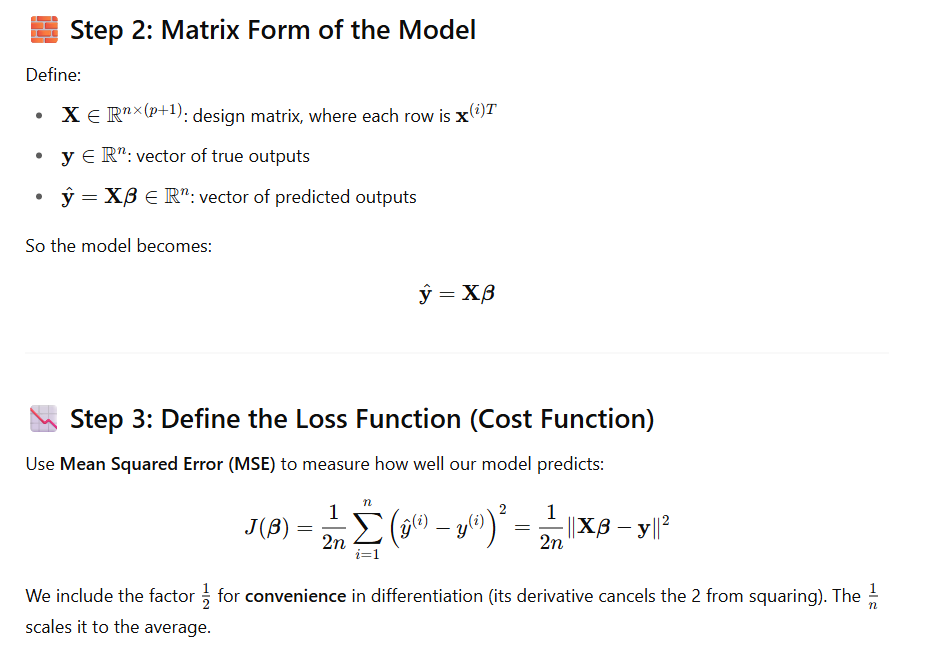
**Conclusion**

This normal equation is used in Sci-kit learn’s LinearRegression() which works well for small datasets. LinearRegression()’s computational complexity is too high since it involves calculating matrix inverse operation and complexity is o(n³). Please refer [complexity of matrix inverse operation](https://en.wikipedia.org/wiki/Computational_complexity_of_mathematical_operations). To overcome these complexity issues, [SGDRegressor](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html" \t "_blank)() method is used, which uses gradient descent to find the coefficients.

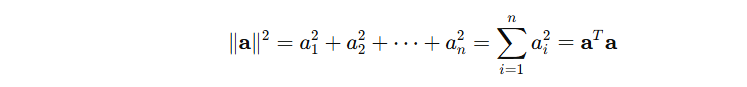
**Full derivation of multiple linear regression**

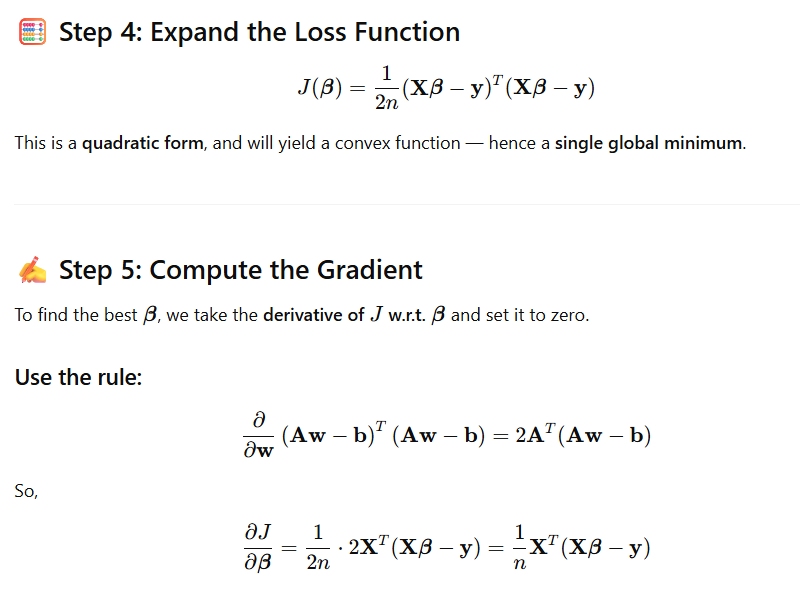


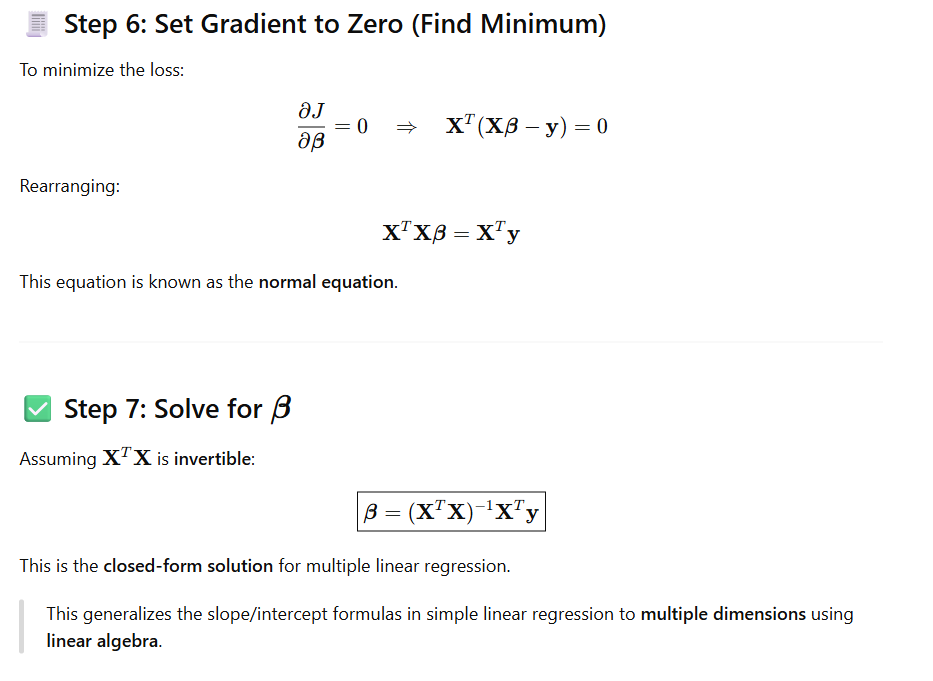


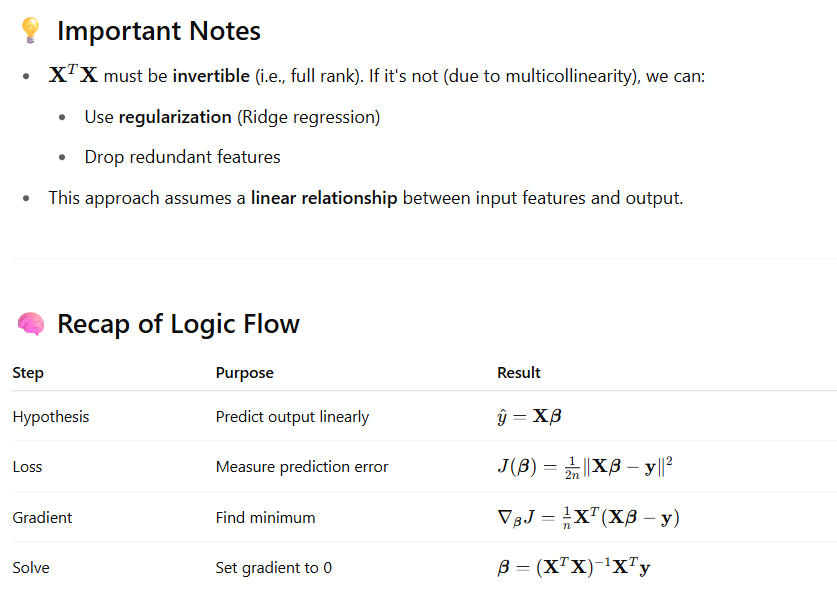


Use this below formula and convert above formula

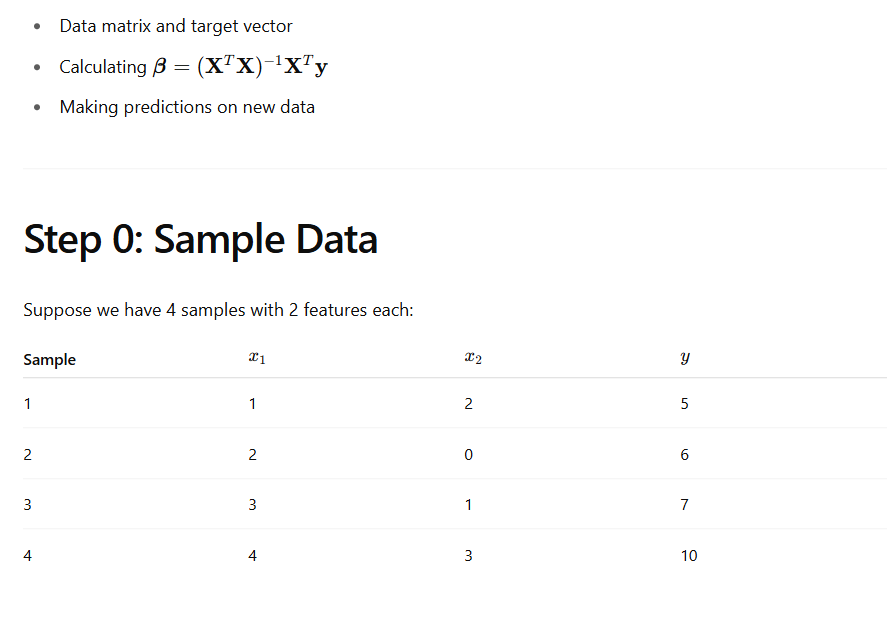


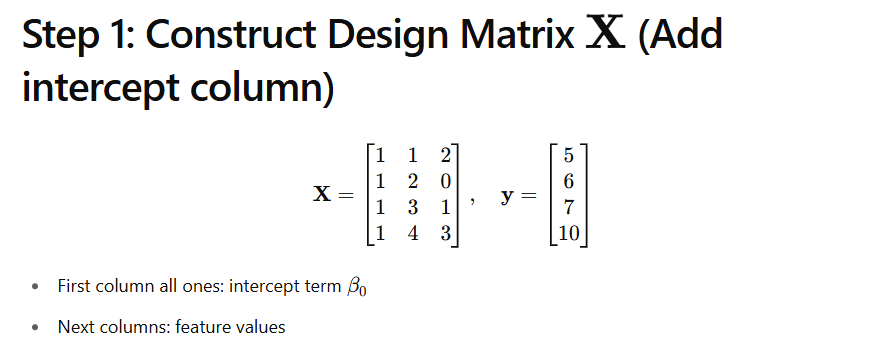


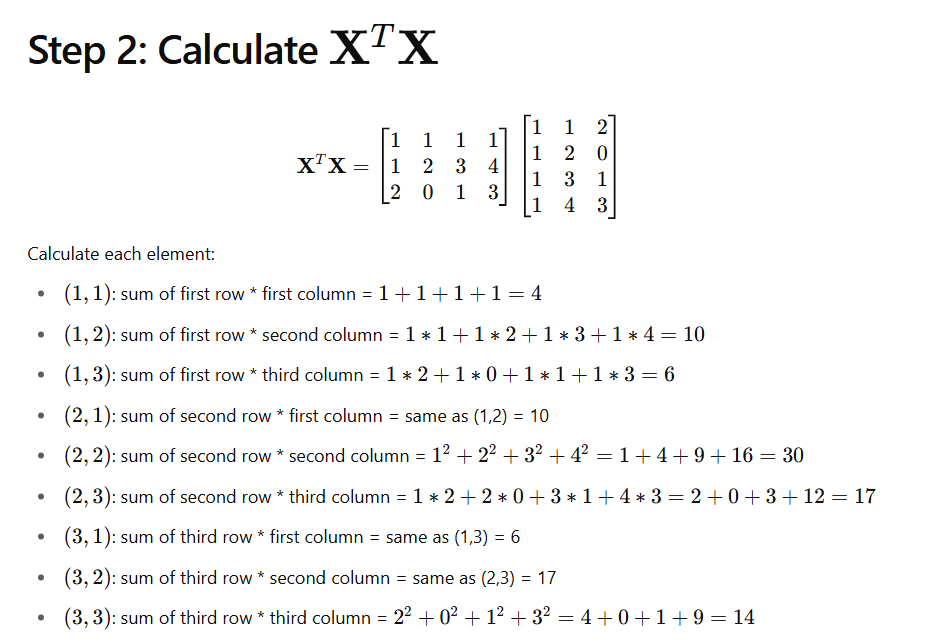


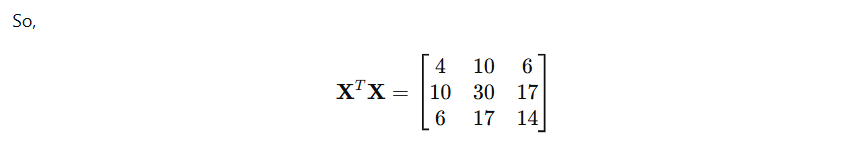


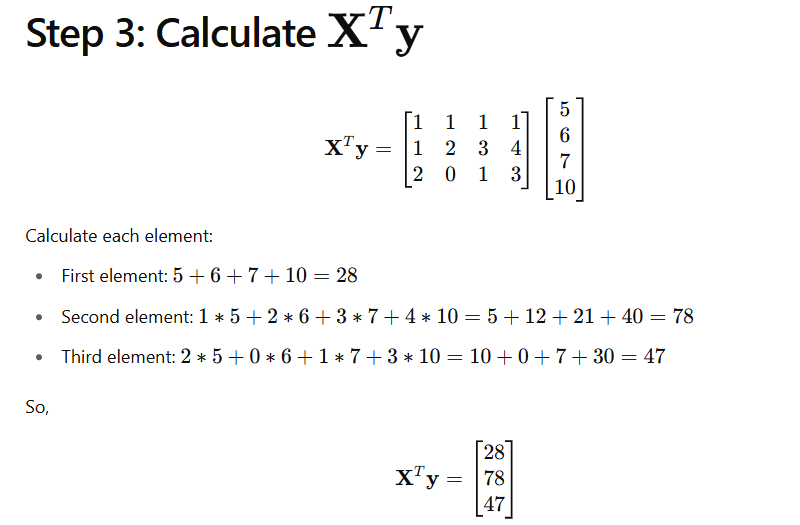
**Let’s do a full, step-by-step example of multiple linear regression with 2 features and 1 target, showing:**

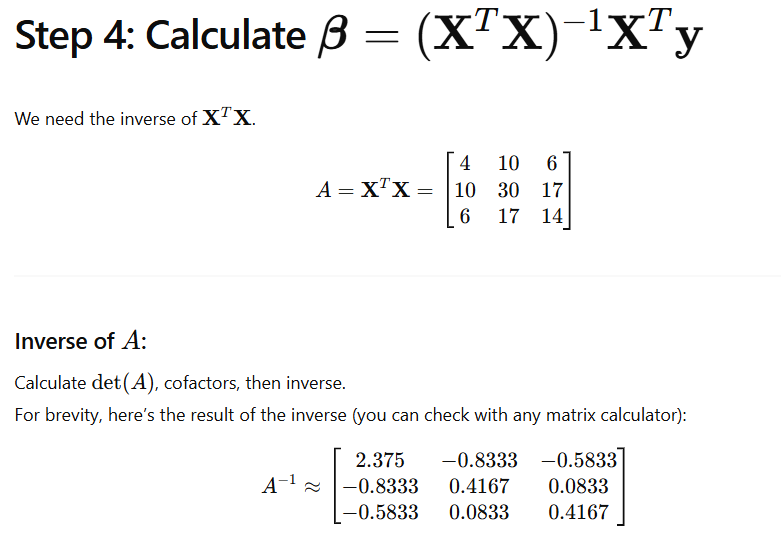


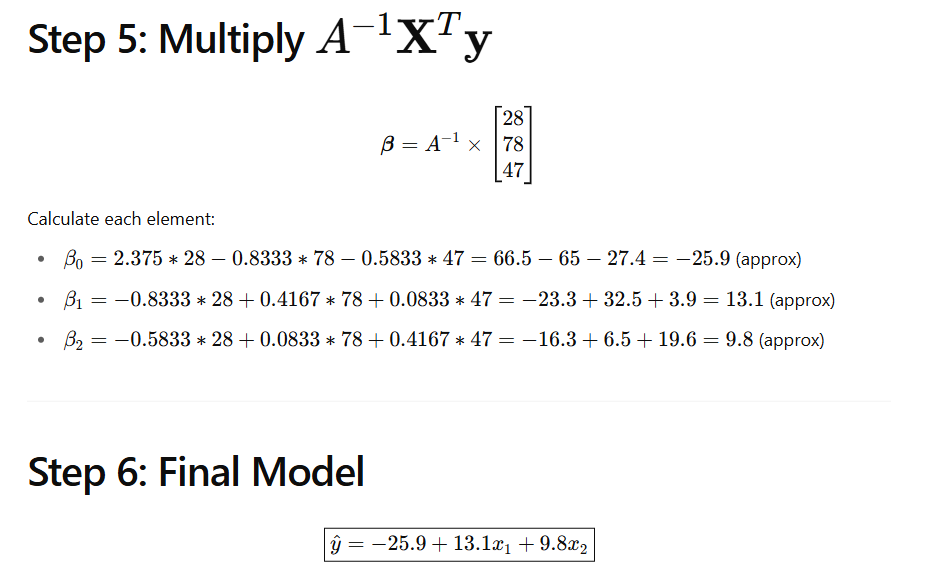


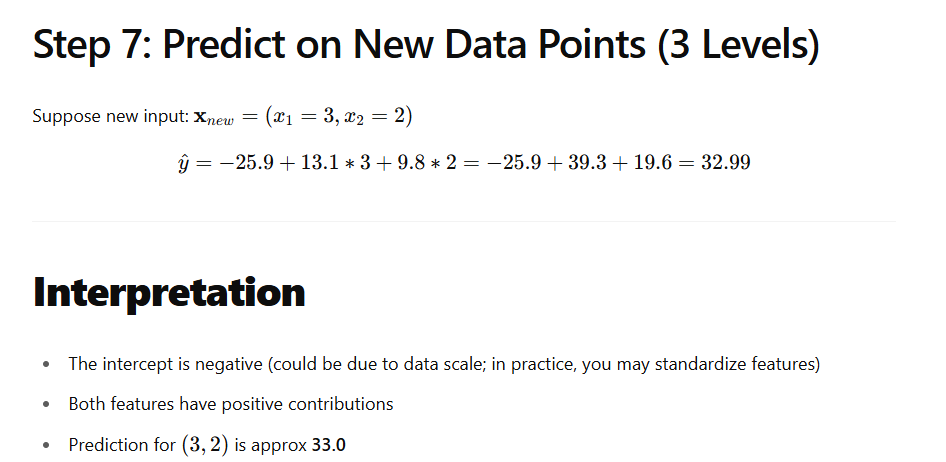




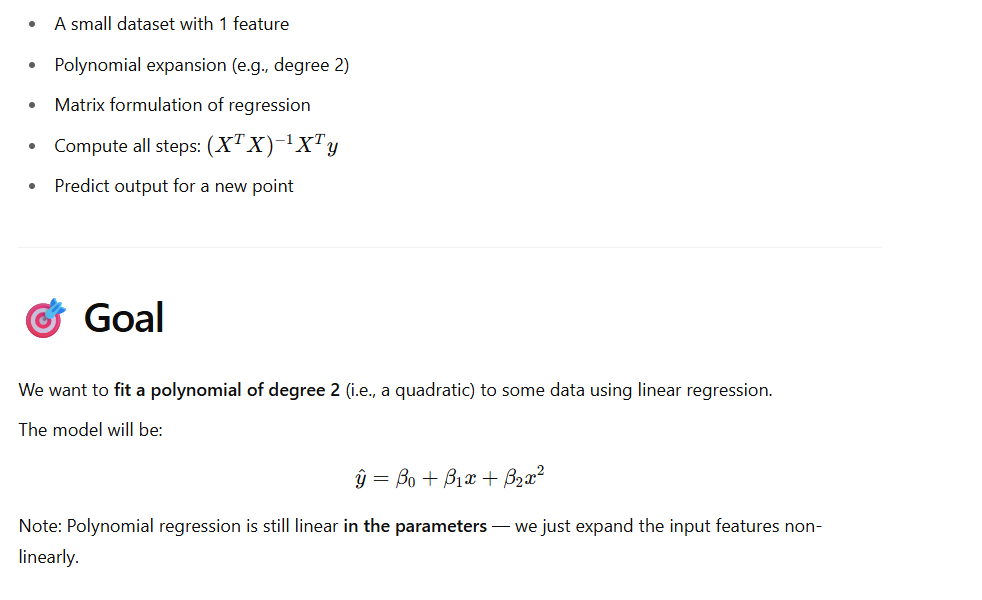


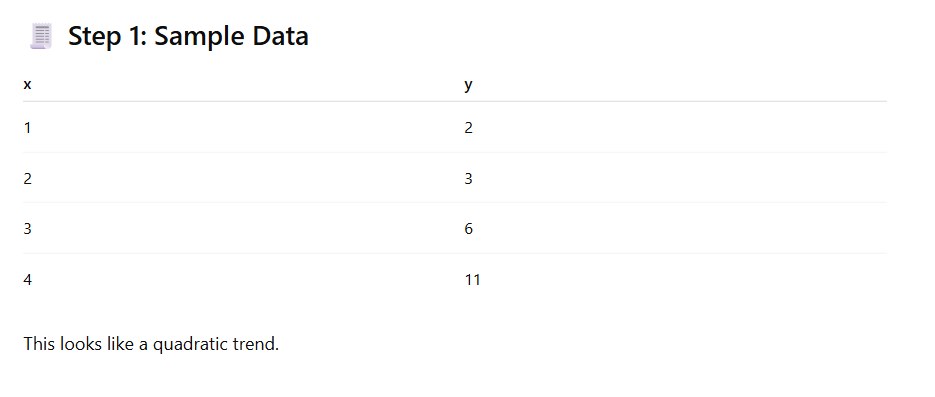


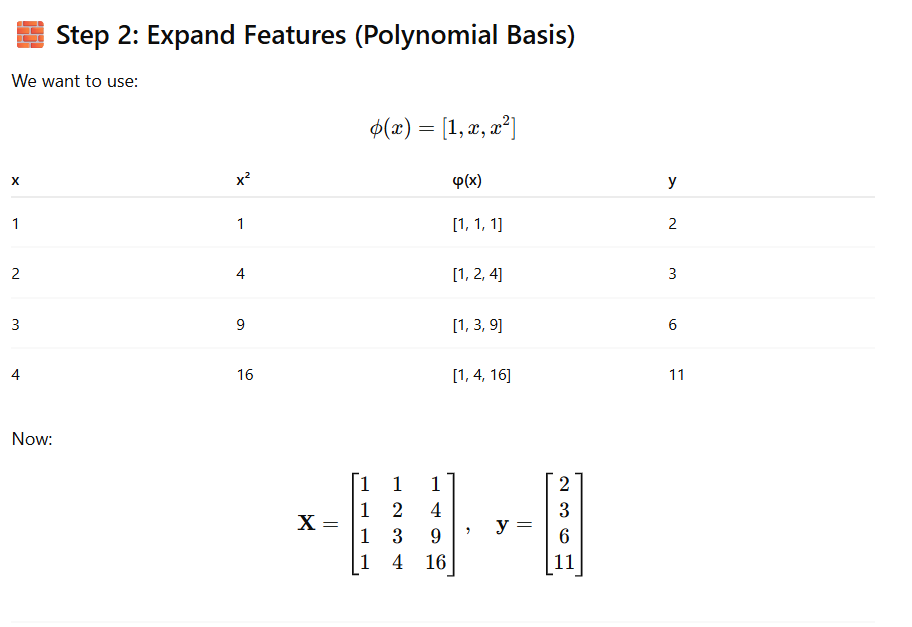


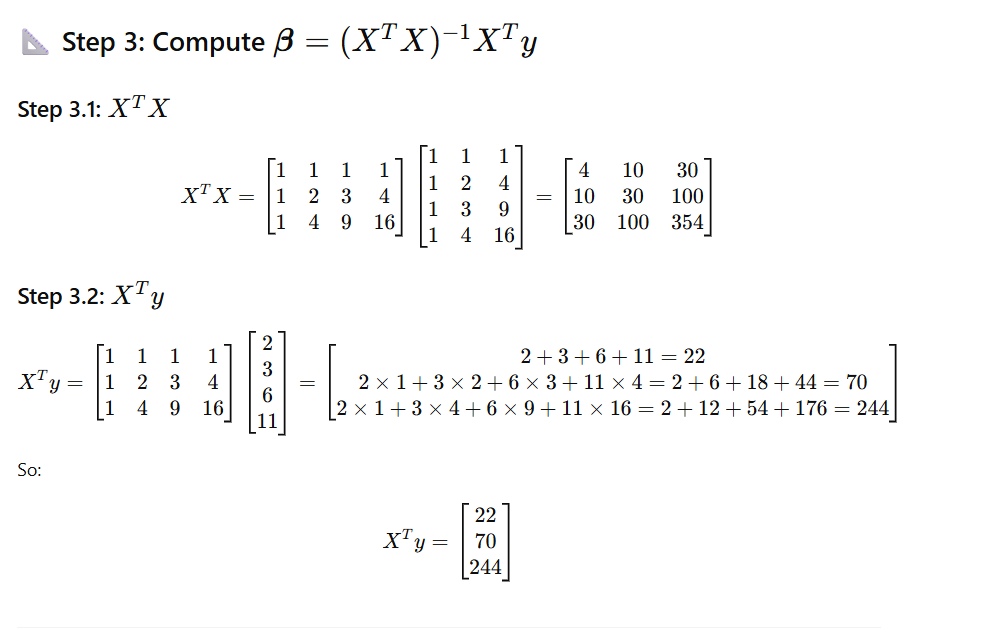


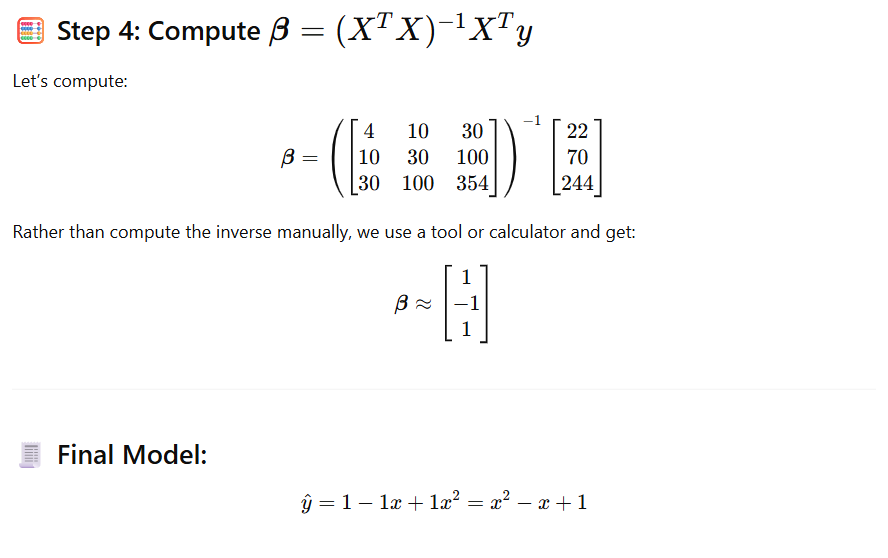
**Let’s now do a detailed worked-out numerical example for polynomial regression, including:**

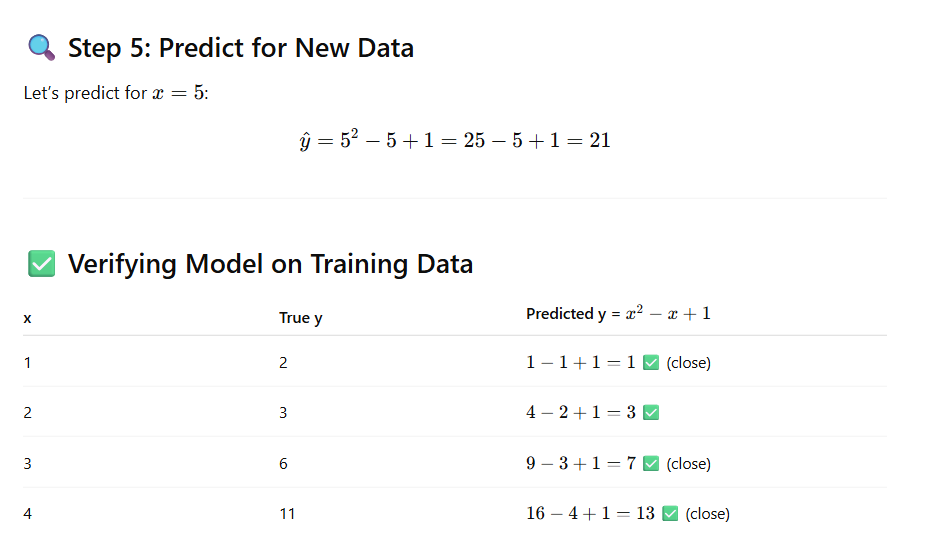










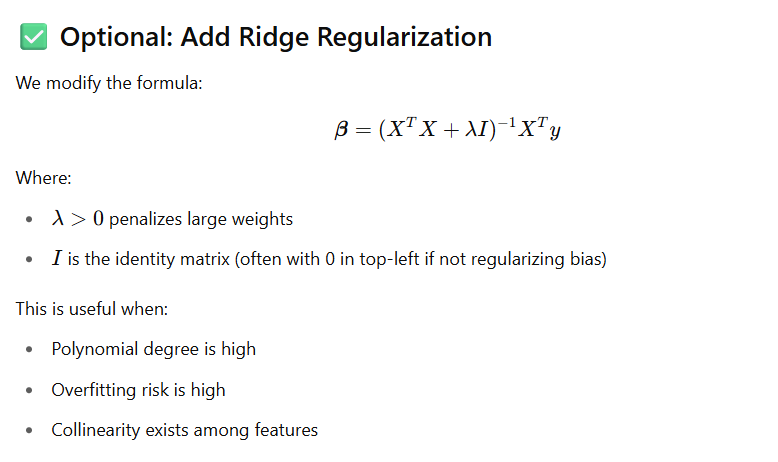


**📌 Summary**

We:

* Expanded the input features using a degree-2 polynomial
* Used standard linear regression steps on the transformed data
* Derived the coefficients
* Predicted on unseen input

Polynomial regression is **linear regression on a non-linear transformation** of features.

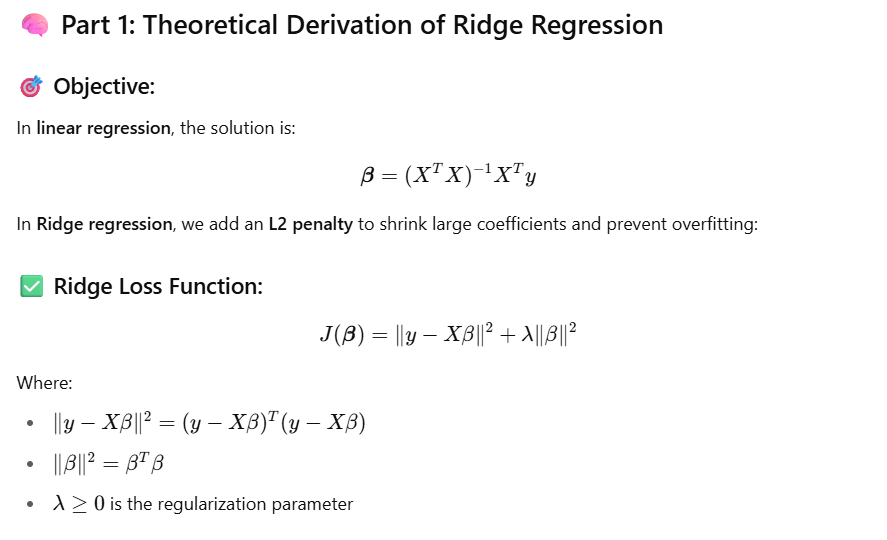


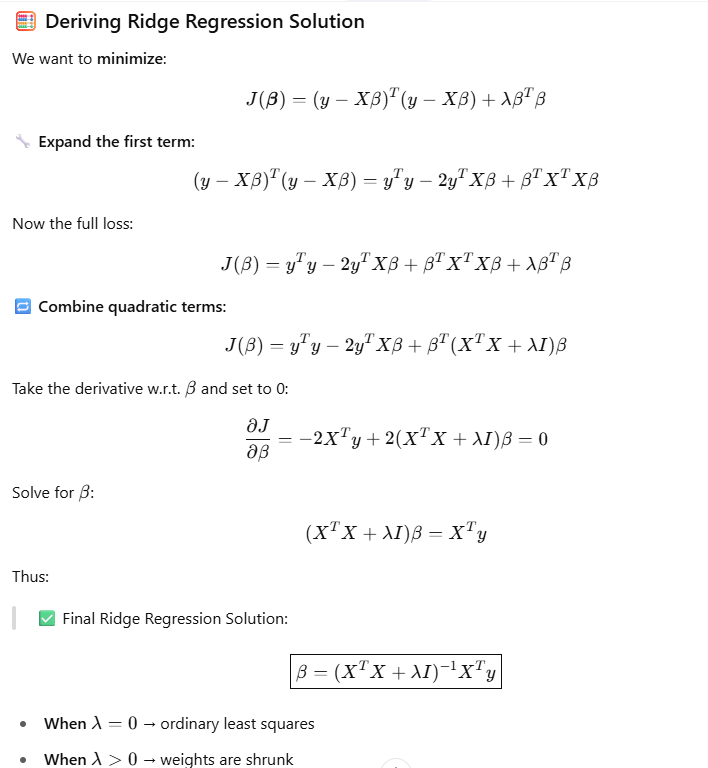
Try a range of λ values (e.g., 0.001, 0.01, 0.1, 1, 10, 100)

**Balance: Bias vs Variance**

* **Too small λ** → overfitting (high variance)
* **Too large λ** → underfitting (high bias)

**Bayesian ridge regression**





import numpy as np

# Feature matrix (with bias term)

X = np.array([

[1, 1],

[1, 2],

[1, 3],

[1, 4]

]) # Shape: (4, 2) -> 1 bias + 1 feature

# Target values

y = np.array([2, 3, 6, 11]).reshape(-1, 1) # Shape: (4, 1)

# Regularization parameter

lmbda = 1.0

# Identity matrix (don’t regularize bias term)

I = np.eye(X.shape[1])

I[0, 0] = 0 # Don’t penalize bias term

# Ridge Regression formula

XtX = X.T @ X

ridge\_term = XtX + lmbda \* I

Xty = X.T @ y

beta = np.linalg.inv(ridge\_term) @ Xty

print("Ridge coefficients:", beta.ravel())

**What all hyperparameters are used in linear regression**

In **Linear Regression**, especially as implemented in libraries like **scikit-learn**, there are a few key **hyperparameters** — although basic linear regression has none in the mathematical sense, various **variants** like **Ridge**, **Lasso**, and **ElasticNet** do include hyperparameters.

Here's a breakdown by model type:

**✅ 1. Ordinary Linear Regression (OLS)**

In **pure OLS** (e.g., LinearRegression() in scikit-learn), there are **no core hyperparameters** for model complexity control, but a few **optional parameters** exist:

| **Hyperparameter** | **Purpose** |
| --- | --- |
| fit\_intercept | Whether to calculate the intercept bbb or assume it is zero |
| normalize *(deprecated)* | Previously normalized the input data (now handled in pipeline) |
| copy\_X | Whether to copy input X or overwrite |
| n\_jobs | Number of cores used for computation |
| positive | If True, forces coefficients to be positive |

**✅ 2. Ridge Regression (L2 Regularization)**

python

CopyEdit

from sklearn.linear\_model import Ridge

| **Hyperparameter** | **Purpose** |
| --- | --- |
| alpha | Controls the regularization strength (λ\lambdaλ); higher → more penalty |
| fit\_intercept | Whether to estimate the intercept |
| solver | Algorithm to use: "auto", "svd", "cholesky", "saga", etc. |
| normalize | Deprecated (use StandardScaler instead) |
| max\_iter | Maximum number of iterations (for solvers like saga) |
| tol | Tolerance for convergence |
| random\_state | For stochastic solvers |

**✅ 3. Lasso Regression (L1 Regularization)**

python

CopyEdit

from sklearn.linear\_model import Lasso

| **Hyperparameter** | **Purpose** |
| --- | --- |
| alpha | Regularization parameter; larger → more sparsity (zero coefficients) |
| max\_iter | Max number of iterations (important for convergence) |
| tol | Tolerance for stopping criterion |
| selection | "cyclic" or "random" (coordinate descent update rule) |

**✅ 4. ElasticNet (Combination of L1 and L2)**

python

CopyEdit

from sklearn.linear\_model import ElasticNet

| **Hyperparameter** | **Purpose** |
| --- | --- |
| alpha | Overall regularization strength |
| l1\_ratio | Mix between Lasso and Ridge: |

markdown

CopyEdit

- 1 = Lasso

- 0 = Ridge

- 0.5 = Equal mix |

| max\_iter | Max iterations for convergence |  
| tol | Stopping tolerance |

**🧪 Hyperparameter Tuning Tips**

* Use **cross-validation** to choose alpha, l1\_ratio, etc.
* Use RidgeCV, LassoCV, or ElasticNetCV for automated tuning.

python

CopyEdit

from sklearn.linear\_model import RidgeCV

model = RidgeCV(alphas=[0.01, 0.1, 1.0, 10.0], cv=5)

**📌 Summary Table**

| **Model** | **Key Hyperparameters** |
| --- | --- |
| OLS | fit\_intercept, copy\_X, positive |
| Ridge | alpha, solver, tol, max\_iter |
| Lasso | alpha, selection, tol, max\_iter |
| ElasticNet | alpha, l1\_ratio, tol, max\_iter |

**Advantages of Linear Regression**

| **Advantage** | **Explanation** |
| --- | --- |
| 💡 **Simplicity & Interpretability** | The model is easy to understand and interpret. Coefficients directly show how features influence the output. |
| 🧮 **Fast Training & Prediction** | It’s computationally efficient even with large datasets — can be solved analytically using normal equations or optimized with gradient descent. |
| 📈 **Works Well with Linearly Separable Data** | If a true linear relationship exists, linear regression can model it accurately. |
| 📊 **Statistical Foundations** | Comes with strong statistical tools: confidence intervals, hypothesis testing, R², etc. |
| 🔍 **Feature Importance** | Coefficients can indicate which features are most impactful, making it useful for feature selection. |
| 🧪 **Basis for Many Models** | Linear regression is the foundation for more advanced models (logistic regression, GLMs, regularized models). |

**❌ Disadvantages of Linear Regression**

| **Disadvantage** | **Explanation** |
| --- | --- |
| 🔄 **Assumes Linearity** | Can't capture complex or non-linear relationships without feature engineering or transformation (e.g., polynomial terms). |
| 🔺 **Sensitive to Outliers** | Since it minimizes squared error, extreme values can disproportionately affect the model. |
| 🔀 **Multicollinearity Issues** | When features are highly correlated, coefficient estimates become unstable and unreliable. |
| 📈 **Assumes Homoscedasticity** | Assumes equal variance of errors — heteroscedasticity can lead to inefficiency and misleading results. |
| 🔄 **Assumes Independence of Errors** | Not suitable for time-series or sequential data without modifications. |
| ⛔ **Does Not Handle Missing Data Automatically** | Requires imputation or data cleaning before modeling. |
| 🎯 **Prone to Underfitting** | Tends to have high bias and may not capture complex patterns in the data. |

**📌 Summary Table**

| **✅ Pros** | **❌ Cons** |
| --- | --- |
| Easy to interpret | Assumes linear relationships only |
| Fast to train and predict | Sensitive to outliers |
| Good with small to medium data | Poor performance with complex data |
| Helps understand feature impact | Suffers from multicollinearity |
| Useful as a baseline model | Requires preprocessed and clean data |

**🧠 When to Use Linear Regression**

Use it when:

* The relationship between features and target is mostly linear
* You want a fast and interpretable model
* Data is not too noisy or full of outliers
* You need a strong baseline to compare with more complex models

Avoid it when:

* The relationship is complex or non-linear
* There are many categorical or highly correlated features
* You're working with high-dimensional, sparse, or noisy data without regularization

**Linear Regression Real Use Case**

**📘 Goal: Predict median house value based on average number of rooms per household.**

# 📌 Step 1: Import Libraries

import numpy as np

import matplotlib.pyplot as plt

from sklearn.datasets import fetch\_california\_housing

from sklearn.linear\_model import LinearRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import mean\_squared\_error, r2\_score

from sklearn.preprocessing import StandardScaler

from sklearn.pipeline import make\_pipeline

# 📌 Step 2: Load California Housing Dataset

data = fetch\_california\_housing()

X = data.data # Features

y = data.target # Target: Median house value

# Let's use only 1 feature for visualization: 'AveRooms' (index 3)

X = X[:, [3]] # Average number of rooms per household

# 📌 Step 3: Train/Test Split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# 📌 Step 4: Define and Train Linear Regression Model

model = make\_pipeline(

StandardScaler(), # Normalize input features

LinearRegression()

)

model.fit(X\_train, y\_train)

# 📌 Step 5: Predict on Test Data

y\_pred = model.predict(X\_test)

# 📌 Step 6: Evaluate Model

mse = mean\_squared\_error(y\_test, y\_pred)

r2 = r2\_score(y\_test, y\_pred)

print("✅ Mean Squared Error (MSE):", mse)

print("✅ R^2 Score:", r2)

# 📌 Step 7: Visualize Results

plt.figure(figsize=(10,6))

plt.scatter(X\_test, y\_test, color='blue', label='Actual', alpha=0.5)

plt.plot(X\_test, y\_pred, color='red', label='Predicted Line', linewidth=2)

plt.xlabel('Average Rooms per Household')

plt.ylabel('Median House Value')

plt.title('Linear Regression - California Housing')

plt.legend()

plt.grid(True)

plt.show()

📊 Model Evaluation

| Metric | Meaning |
| --- | --- |
| MSE | Lower MSE means better prediction performance |
| R² Score | Closer to 1 is better (1 = perfect fit) |

**import numpy as np**

**import pandas as pd**

**from sklearn.datasets import make\_regression**

**from matplotlib import pyplot**

**from sklearn.model\_selection import train\_test\_split**

**X\_train,X\_test,y\_train,y\_test = train\_test\_split(X,y,test\_size=0.2,random\_state=2)**

**from sklearn.linear\_model import LinearRegression**

**class OLSFormulaSLR:**

**def \_\_init\_\_(self):**

**self.beta1 = None**

**self.beta0 = None**

**def fit(self,X\_train,y\_train):**

**num = 0**

**den = 0**

**for i in range(X\_train.shape[0]):**

**num = num + ((X\_train[i] - X\_train.mean())\*(y\_train[i] - y\_train.mean()))**

**den = den + ((X\_train[i] - X\_train.mean())\*(X\_train[i] - X\_train.mean()))**

**self.beta1 = num/den**

**self.beta0 = y\_train.mean() - (self.beta1 \* X\_train.mean())**

**print("Coefficient - ",self.beta1)**

**print("Intercept - ",self.beta0)**

**def predict(self,X\_test):**

***#print(X\_test)***

**return self.beta1 \* X\_test + self.beta0**

***# generate regression dataset***

**X, y = make\_regression(n\_samples=100, n\_features=1, noise=100,random\_state=2021)**

***# plot regression dataset***

**pyplot.scatter(X,y)**

**pyplot.show()**

**slr = OLSFormulaSLR()**

**slr.fit(X\_train,y\_train)**

**print(slr.predict(X\_test[0]))**

**lr = LinearRegression()**

**lr.fit(X\_train,y\_train)**

**X\_test[0]**

**pred = lr.predict(X\_test[0].reshape(-1,1))**

**pred**

**lr.intercept\_**

**lr.coef\_**

**Evaluation Metrics for Linear Regression**

A variety of [evaluation measures](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/) can be used to determine the strength of any linear regression model. These assessment metrics often give an indication of how well the model is producing the observed outputs.

The most common measurements are:

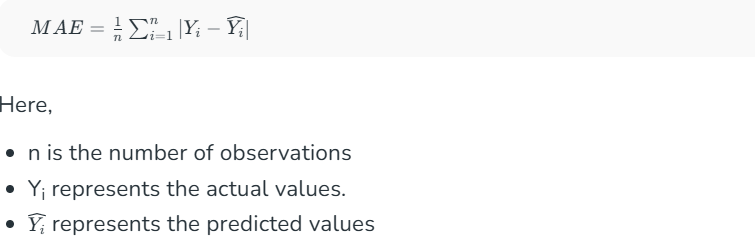
**Regression Evaluation Metrics**

In the regression task, we are supposed to predict the target variable which is in the form of continuous values. To evaluate the performance of such a model below mentioned evaluation metrics are used:

* [**Mean Absolute Error**](https://www.geeksforgeeks.org/how-to-calculate-mean-absolute-error-in-python)
* [**Mean Squared Error**](https://www.geeksforgeeks.org/python-mean-squared-error)
* [**Root Mean Square Error**](https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming)
* **Root Mean Square Logarithmic Error**
* [**R2 - Score**](https://www.geeksforgeeks.org/python-coefficient-of-determination-r2-score)

**Mean Absolute Error (MAE)**

Mean Absolute Error(MAE) is the average distance between predicted and original values. Basically, it gives how we have predicted from the actual output. However, there is one limitation i.e. it doesn't give any idea about the direction of the error which is whether we are under-predicting or over-predicting our data. It can be represented mathematically in this way:



Lower MAE value indicates better model performance. It is not sensitive to the outliers as we consider absolute differences.

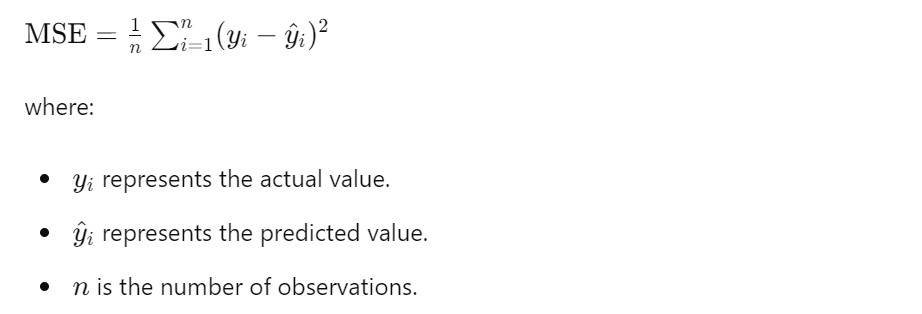
**Mean Squared Error (MSE)**

MSE is similar to mean absolute error but the difference is it takes the square of the average of between predicted and original values. The main advantage to take this metric is here, it is easier to calculate the gradient whereas, in the case of mean absolute error, it takes complicated programming tools to calculate the gradient. By taking the square of errors it pronounces larger errors more than smaller errors, we can focus more on larger errors. It can be expressed mathematically in this way.



What is Mean Squared Error?

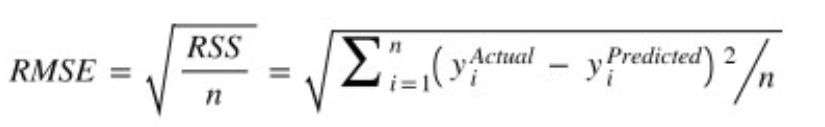
The mean squared error is the average of the squared differences between the expected and actual values. The mathematical notation for it is as follows:



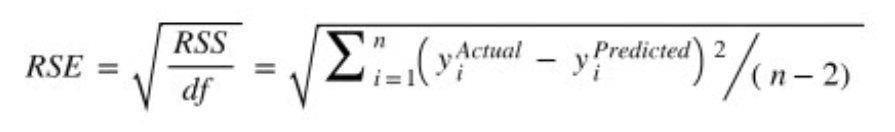
The squaring of errors ensures that positive and negative differences do not cancel each other out. Additionally, squaring emphasizes larger errors, making MSE sensitive to outliers.

**Root Mean Square Error (RMSE)**

The Root Mean Squared Error is the square root of the variance of the residuals. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values. Mathematically it can be represented as,



To make this estimate unbiased, one has to divide the sum of the squared residuals by the **degrees of freedom** rather than the total number of data points in the model. This term is then called the **Residual Standard Error(RSE)**. Mathematically it can be represented as,



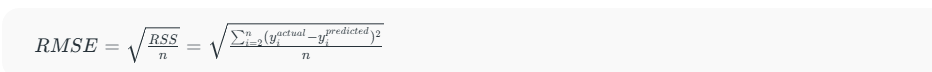
R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables (i.e. it is not a normalized measure), it can change with the change in the unit of the variables.

**Root Mean Squared Logarithmic Error (RMSLE)**

There are times when the target variable varies in a wide range of values. And hence we do not want to penalize the overestimation of the target values but penalize the underestimation of the target values. For such cases, RMSLE is used as an evaluation metric which helps us to achieve the above objective.

Some changes in the original formula of the RMSE code will give us the RMSLE formula that is as shown below:

The square root of the residuals' variance is the [Root Mean Squared Error](https://www.geeksforgeeks.org/root-mean-square-error-in-r-programming/). It describes how well the observed data points match the expected values or the model's absolute fit to the data.  
In mathematical notation, it can be expressed as:

**

Rather than dividing the entire number of data points in the model by the number of degrees of freedom, one must divide the sum of the squared residuals to obtain an unbiased estimate. Then, this figure is referred to as the Residual Standard Error (RSE).

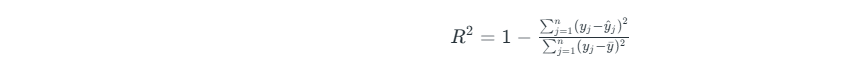
In mathematical notation, it can be expressed as:

**

RSME is not as good of a metric as R-squared. Root Mean Squared Error can fluctuate when the units of the variables vary since its value is dependent on the variables' units (it is not a normalized measure).

**R2 - Score**

The coefficient of determination also called the R2 score is used to evaluate the performance of a linear regression model. It is the amount of variation in the output-dependent attribute which is predictable from the input independent variable(s). It is used to check how well-observed results are reproduced by the model, depending on the ratio of total deviation of results described by the model.



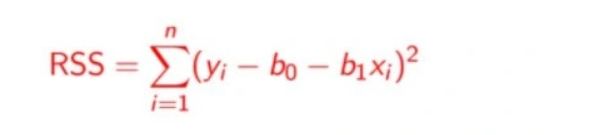
**Coefficient of Determination or R-Squared (R2)**

R-squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.

Mathematically it can be represented as,

**R2 = 1 – ( RSS/TSS )**

* **Residual sum of Squares (RSS)** is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.



* **Total Sum of Squares (TSS)** is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

Total Sum of Squares

where y hat is the mean of the sample data points.

The significance of R-squared is shown by the following figures,

