Linear Regression & Least Squares

- **Learning Problem :-**
  - Given input-output pairs, we need to learn a function, f, that maps input to the output.
  - **Components & Terminology :-**

    X : input matrix with multiple variables. $X_i \in \mathbb{R}^p$

    y : output vector, dependent variable.

    $Col(X)$: is the vector space that contains the input vectors

    B : Matrix of the trainable parameters

    f : the function that can be used to estimate the values of y based on X & B.

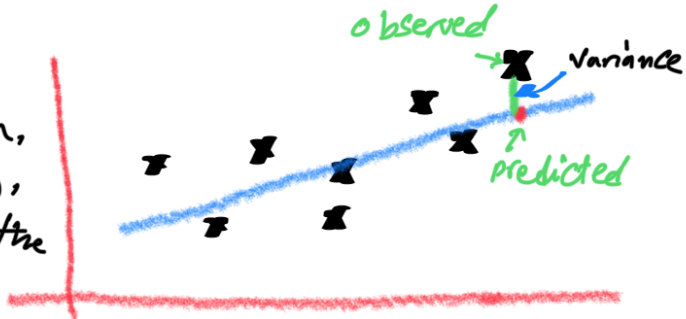
- **Linear Regression :-**
  - Fit a line (represented by f) on the input data.
  - is a common technique for many apps in econometrics, Genomics, Ecology, engineering, etc.
  - $y = mx + b$ ← bais / Intercept
    - y : predicted
    - m : slope
    - x : input vars
  - In the training time, (X, y) pairs are available, so we estimate parameters

- In the inference time, we estimate $\hat{y}$ based on input X & parameters $\{m, b\}$

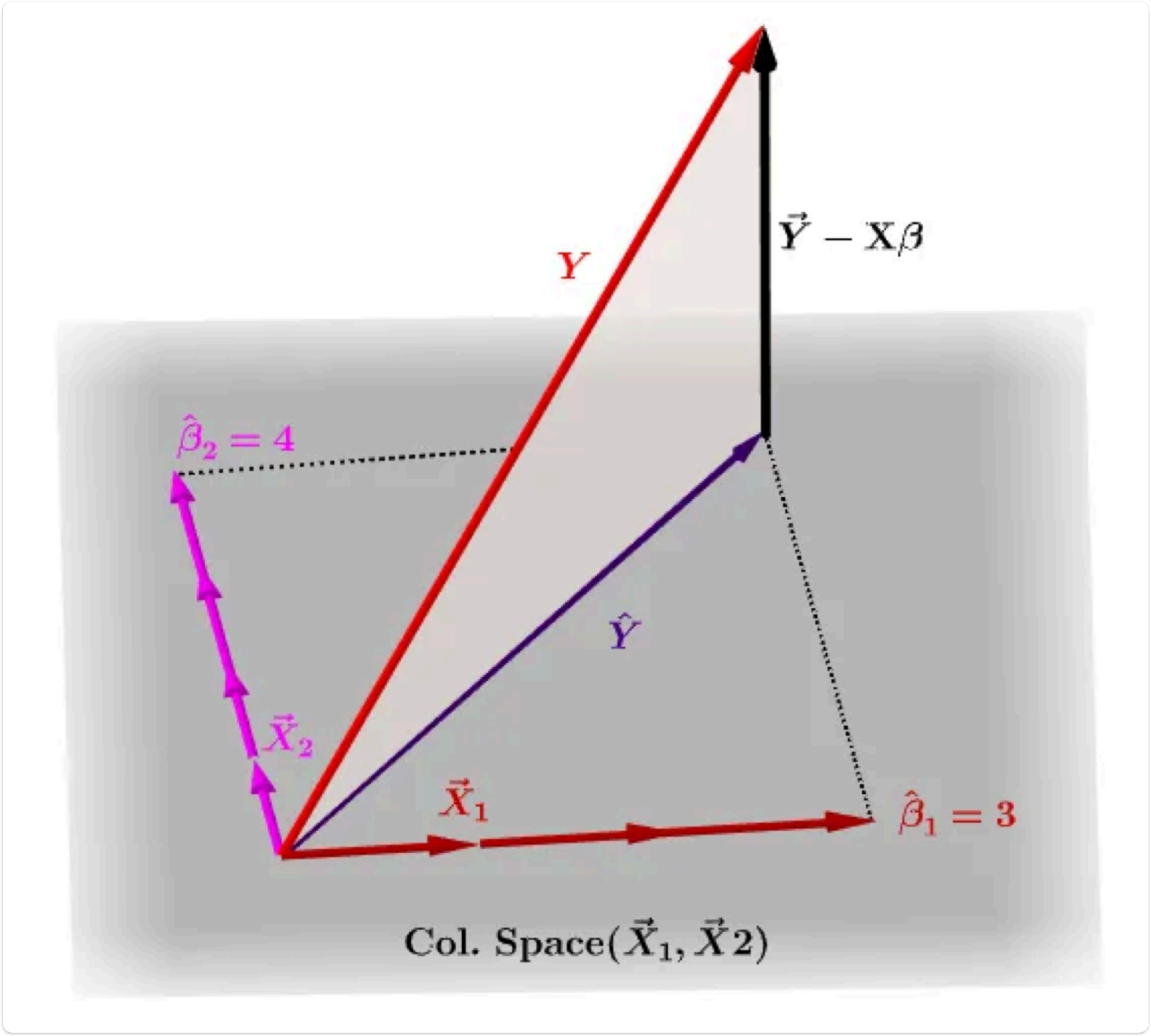
Q: How to estimate parameters B, where they represent the best model fitting data?

A: We use optimization techniques, such as Least squares or Gradient Descent.

- understanding Least squares Method
  - objective:- find the function, (in this case line), that best fits the input data, and enable good estimate of output variables. This can be done by minizing the error / variance between estimated / predicted values and observed y values.



- using "least squares"
  - Geometric Interpretation



The idea here is that the error vector is orthogonal onto the column space of the input.

Thus estimating the optimum values of $\hat{\beta}$ is possible. ← Good news !!!

- Mathmatical Representation.

Given input $X^T = (X_1, X_2, \ldots, X_N)$, and output $y = (y_1, y_2, \ldots, y_N)$, we predict $\hat{y}$ as:

$$\hat{y} = \hat{B}_0 + \sum_{j=1}^{N} X_j \hat{B}_j$$

$\hat{y}$: predicted; $\hat{B}_0, \hat{B}_j$: parameters; X_j: predictors

- By including $\hat{B}_0$ into the matrix of the learnt parameters, we have

$$\hat{y} = X^T \hat{B}$$

- Given pairs of X & y, we can estimate $\hat{B}$ as follows

$$RSS(\hat{B}) = \sum_{i=1}^{N} (y_i - X_i^T \hat{B})^2$$

(Residual sum of squares)

We minimize this function by taking the partial derivatives w.r.t. the individual parameters, $B_0, B_1, B_2, \ldots$

$$\hat{B} = (X^T X)^{-1} X^T y$$