Linear Algebra for Machine Learning

Contents

[Linear Regression in Linear Algebra 2](#__RefHeading___Toc2615_3270521293)

[Example of Simple regression problem: 8](#__RefHeading___Toc2617_3270521293)

[Residual vector: 10](#__RefHeading___Toc2619_3270521293)

[Multiple Linear Regression: 10](#__RefHeading___Toc2621_3270521293)

[Full Rank Design Matrix: 11](#__RefHeading___Toc2623_3270521293)

[When n>>p, 11](#__RefHeading___Toc2625_3270521293)

[When n<<p, 12](#__RefHeading___Toc2627_3270521293)

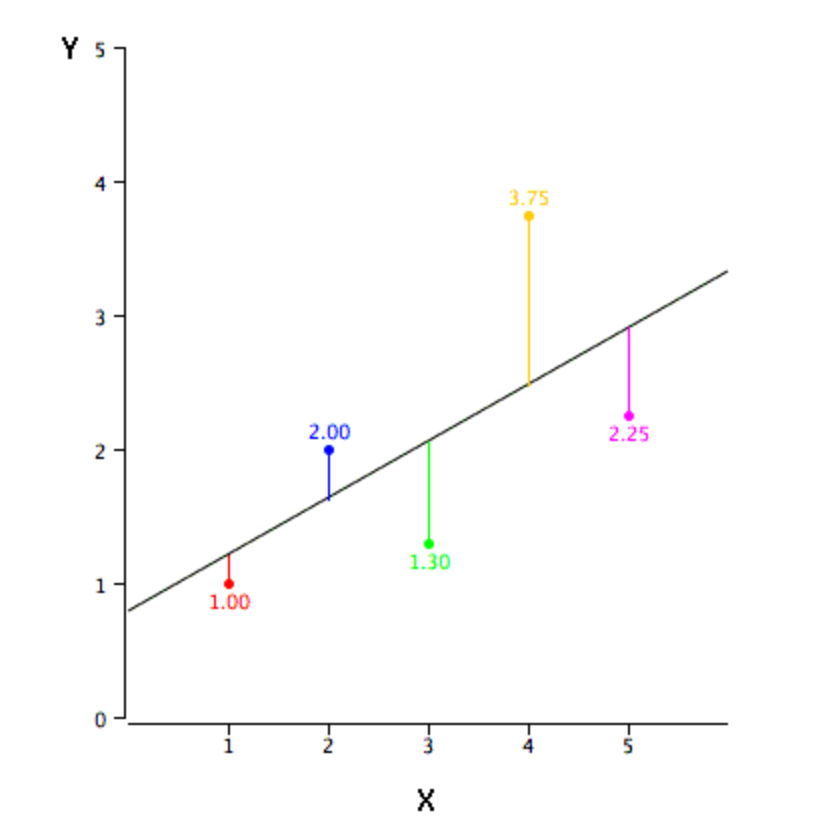
[Weighted Linear Regression: 12](#__RefHeading___Toc2629_3270521293)

# Linear Regression in Linear Algebra

Linear regression is all about finding a line (eqn of line) that best defines the relationship between factors (predictors) and one output – response variable.

Let us take a simple linear regression to begin with. We want to find the best fit line through a set of data points: (x1, y1), (x2, y2), … (xn, yn). But what does the best fit mean?

If we can find a slope and an intercept for a single line that passes through all the possible data points, then that is the best fit line. But in most of the cases, such a line does not exist! So we resolve to finding a line such that when a connecting line is drawn parallel to the y-axis from the data points to the regression line, which measures the error of each data point, the sum of all such errors should be minimum.

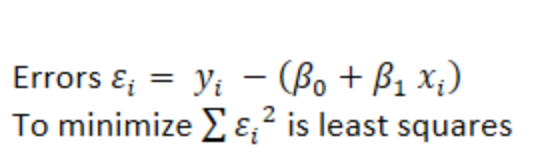


In the diagram, errors are represented by red, blue, green, yellow, and the purple line.

Let the line be:

Y = B0 +B1x1 + e here B0 +B1x1 defines the line while e defines the average of error i.e. average of the lines sizes (red,blue,gree etc lines in graph above).

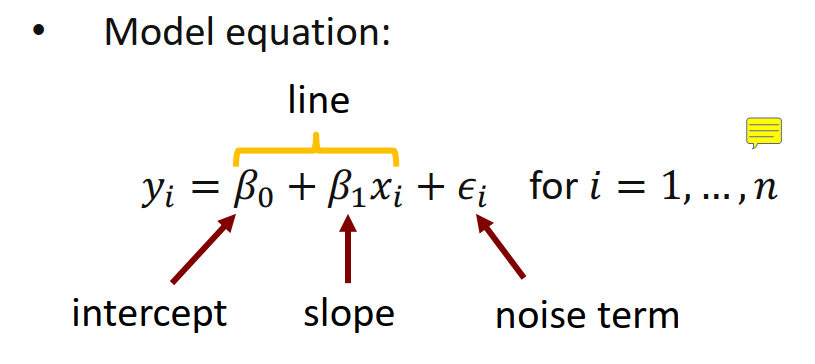
The mathematical way to find the "best fit" one.

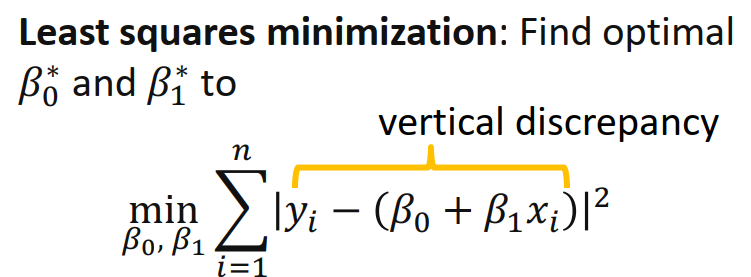


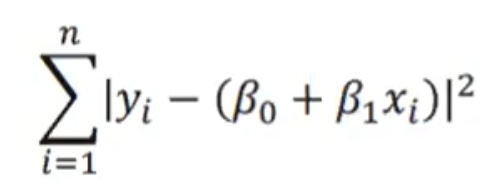
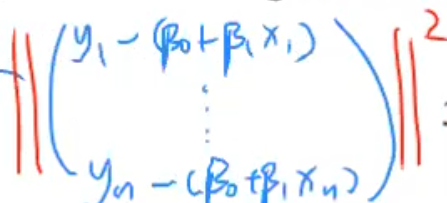
Let's call that number β. E, the sum of the squared errors (SSE) is a function of β since for each choice of β can calculate the amount each estimate is off, square it, and sum them together.

What β minimizes the total sum of the squared errors? This is just a calculus problem. Take the derivative of E by β and set it equal to zero. This gives an equation for β. Check the second derivative is positive to know that it's a minimium. Thus you get an equation for β which minimizes the error.

In matrix form:

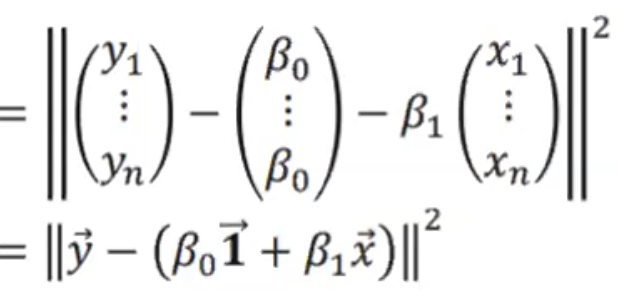




 is actually just 

L2norm squared

Further conversion to vector form:



This gives the projection of y on two vectors B0 and B1. B0+B1 is a linear combination and two lines define a plane so in other words, this is the projection of y on the 1,x plane. AKA, projection of y on the span of 1 and x vectors.

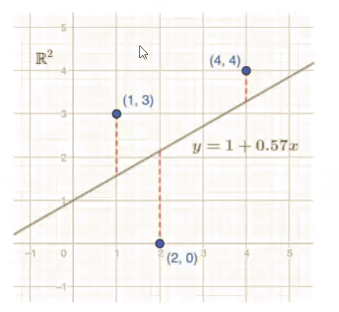
Linear regression is all about finding the two B0 and B1 such that it is closest (thus best fit) to 1 and x vectors. Linear Regression is a find the “orthogonal projection” exercise such that the projection vector is closest to the y end point and on the 1,x plane.

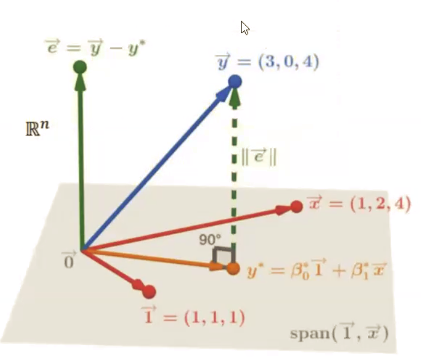
Basically minimization of below



Example of Linear Regression:

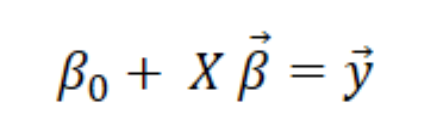
Three points A(1 3), B(4 4) C(2,0) are given and we need to fit the fitted line.



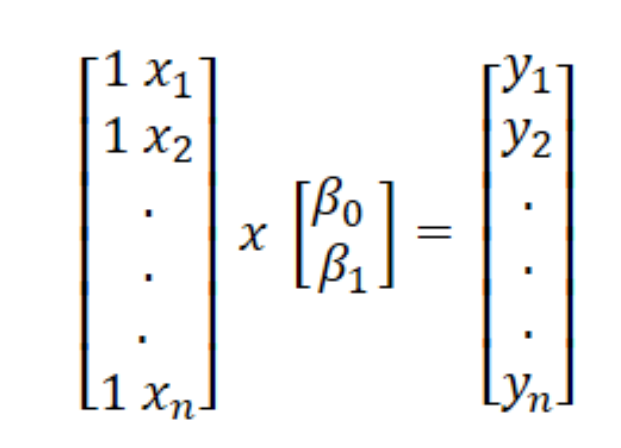


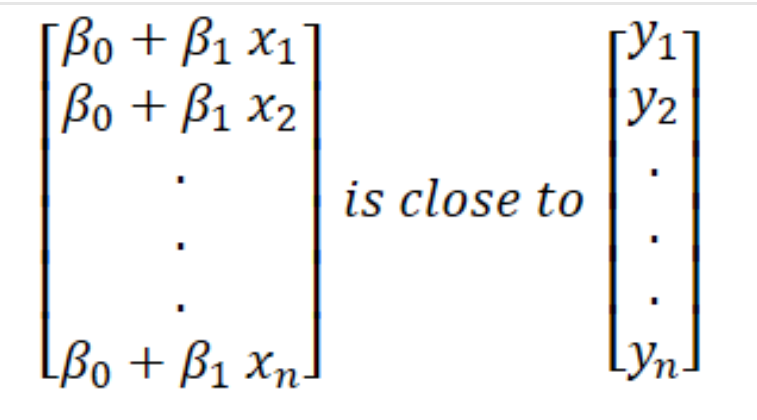
Note that y = (3,0,4) the y coordinates of the orginal points and same for x

To formulate this as a matrix solving problem, consider linear equation  
is given below, where Beta 0 is the intercept and Beta is the slope.

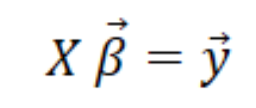


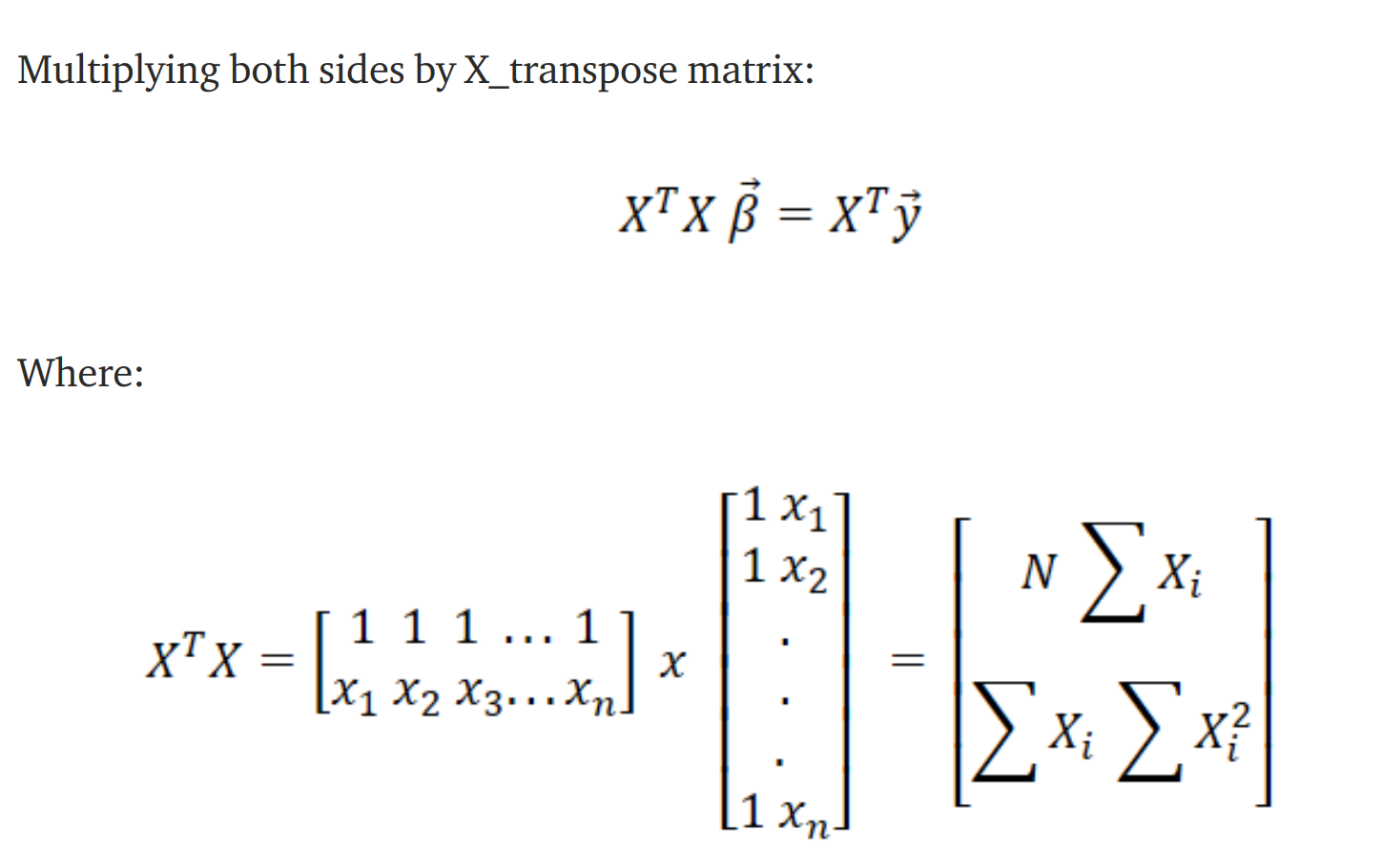
To simplify this notation, we will add Beta 0 to the Beta vector. This is done by adding an  
extra column with 1’s in X matrix and adding an extra variable in the Beta vector.  
Consequently, the matrix form will be:



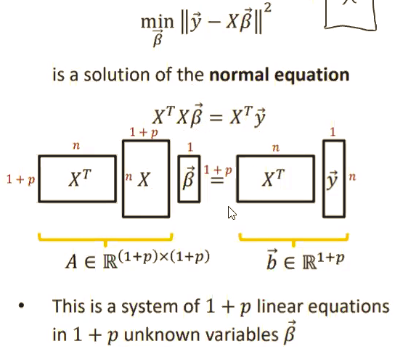


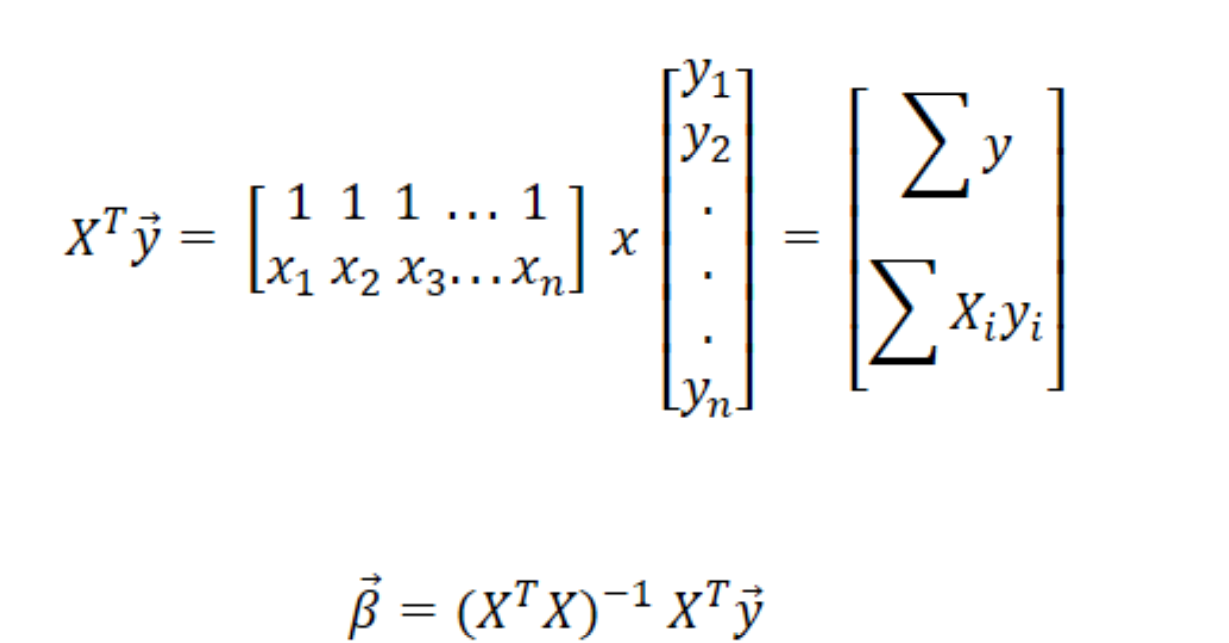
Let us consider our initial equation:

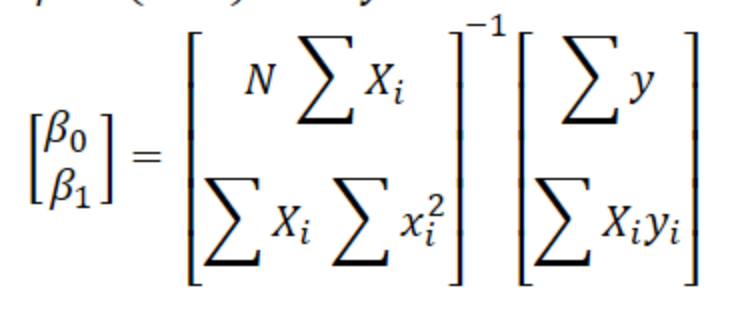


above eqn is callled the **Normal Equation**

XT is mulitplied both sides only so to make X an invertible matrix and thus can be taken to other side. XTX always becomes a square martix which is needed for a matrix to be invertible. This way XTX always becomes a square matrix for any application/usage. XTX becomes invertible and so taken to other side which allows us to solve and find B matrix.



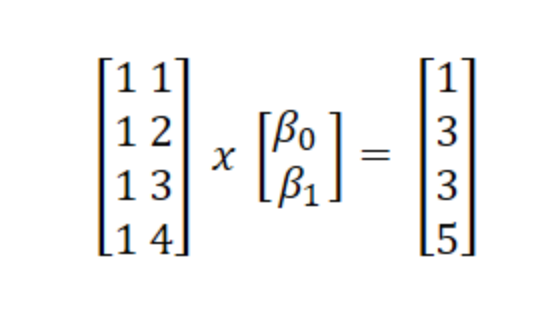




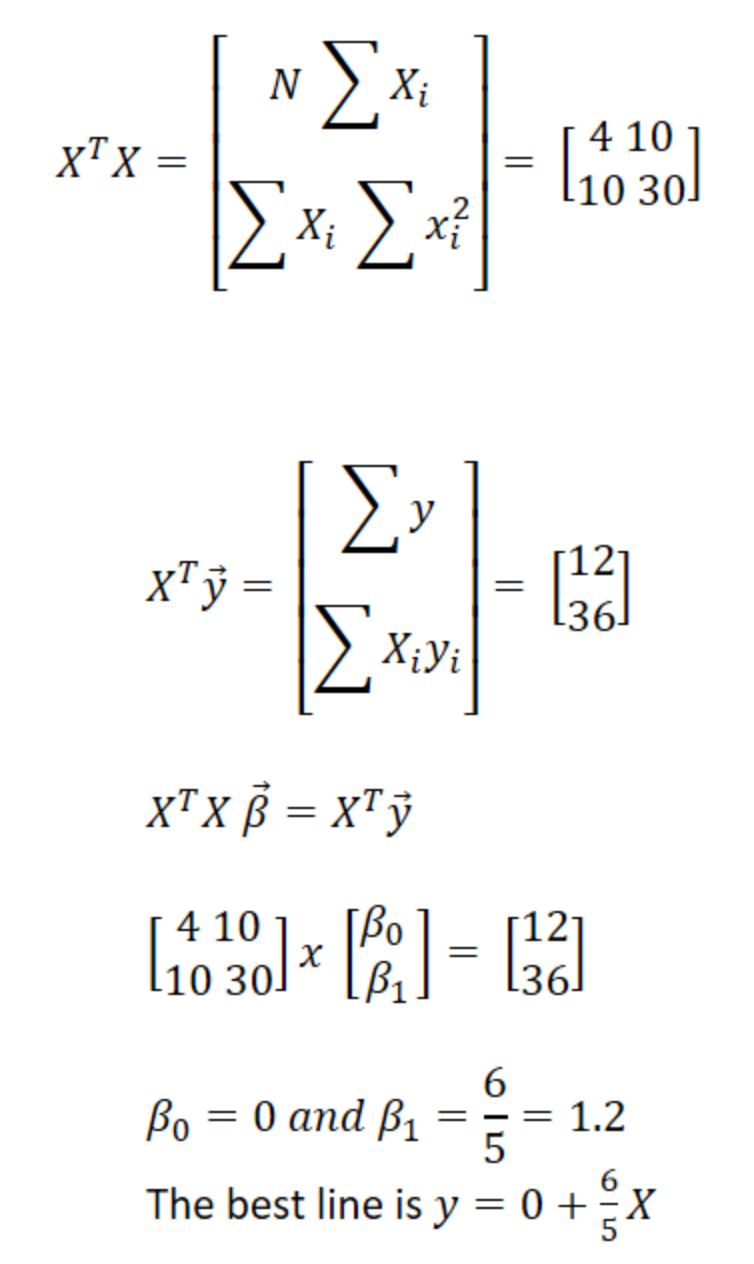
Is the eqn used in programming to find the Beta matrix. Usually, Y which are the response/output (eg: price of house) and x, the predictors (eg: area of house, locality of house) are known. XT, XTX is found multiplied with X and y to get the Beta values.

## Example of Simple regression problem:

For simplicity, we will start with a simple linear regression problem which has 4 data points (1, 1), (2, 3), (3, 3) and (4, 5). X = [1, 2, 3, 4] and y = [1, 3, 3, 5]. When we convert into matrix form as described above, we get:



|  |  |  |
| --- | --- | --- |
|  |  |  |



import numpy as np

X = np.matrix([[1, 1],

[1, 2],

[1, 3],

[1, 4]])

print(X)

XT = np.matrix.transpose(X)

print(XT)

y = np.matrix([[1],

[3],

[3],

[5]])

print(y)

XT\_X = np.matmul(XT, X)

print(XT\_X)

XT\_y = np.matmul(XT, y)

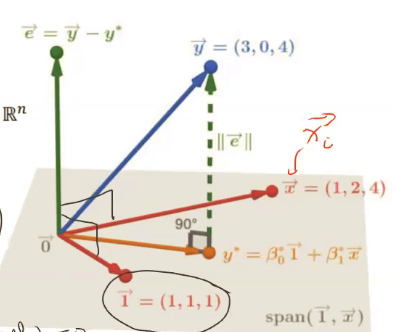
print(XT\_y)

betas = np.matmul(np.linalg.inv(XT\_X), XT\_y)

print(betas)

### Residual vector:

Residual vector e is perpendicular to all the X vectors/predictor data points.



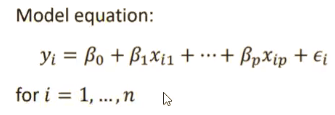
In a fitted linear regression system, . This will appear strange that with all the complexities, . This is because due to orthogonal nature, cos 90 = 0. Thus dot product of 1 vector and e vector = 0.

[e1,e2,e3...] [1]

[1]

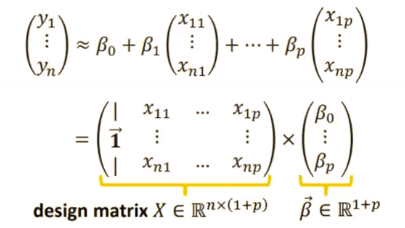
[1] = e1+e2+e3... = 0 since cos 90 is zero. e is perpendicular to 1 vector.

## Multiple Linear Regression:



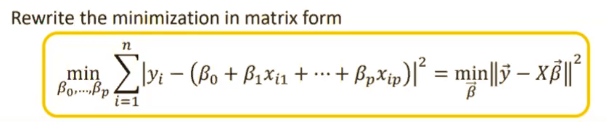
More than one predictor factors.

Vector form:



Do note that above a new vector 1 is included so that the constant B0 is included in the solution to be found.

X matrix is called Design matrix and B matrix is called Beta matrix



## Full Rank Design Matrix:

(XTX)-1 inversion is possible only if the X design matrix is full rank. Full rank as in it has 1+p independent vectors (p is the total number of predictor factors).

### When n>>p,

Let there be large number of data points thus becomes the number of rows in the X matrix. n is the number of data points and p is the number of predictors. X has 1+p columns.

Since n >> p, solution with unique B values is usually possible. In this situation, the max rank of X will usually be 1+p which is the max it can be. Any thing more than 1+p are a redundant vector information.

### When n<<p,

When n <<p, this situation is referred to as “high dimensional vector space”. “high dimensional vector space” aka more predictors than data points. You wont get unique Beta values as solution in such scenarios.

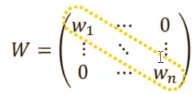
In a “high dimensional” situation, you will not be able to get unique Beta values and thus the fit of the linear regression model is lesser than optimal. This is the case of infinitely possible solutions. In such a case, you cant write one equation where the B0, B1,,etc are unique.

Such Linear model might predict well but lack model explainability.

### Weighted Linear Regression:

In some cases we would want to increase the effect of few predictors and reduce that of the others. To do this we include a w matrix in the normal equation.



W is a diagonal matrix with the amplificaiton/reduction factor along the diagonal. 

fit of the linear regression model is lesser than optimal. This is the case of inf