Int. J. Math. And Appl., 7(4)(2019), 153–159
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Deriving the Formula for Linear Regression

Nirbhay Narang[1,*]

1 Jayshree Periwal International School, Jaipur, Rajasthan, India.

Abstract: The goal of this research paper is to determine a generalised formula for the line and plane of best fit given a set of points in 2 and 3 dimensions respectively, as well as generalising a formula for a function of best fit in n dimensions. The applications are wide and varied, including statistical analysis and data science.

Keywords: Linear Regression, Matrices, Machine Learning, Planes of best fit, Linear Algebra.

1. Find the Equation of the Line of Best Fit

Given a set of n points $(x_1, y_1), (x_2, y_2), (x_2, y_2) \dots (x_n, y_n)$, our objective is to find a and b such that the straight line $y = ax + b$ is the line of best fit for the data set. To do this, let us first define a one-dimensional vector called y which is a vector of all the y-coordinates in the data set.

$$y = (y_1, y_2, ..., y_n)$$

Next, let us define another one-dimensional vector called y_p which denotes the output values of the function $y = ax + b$, or the expected values of the line of best fit. Since y_p is simply an expression for the output of $y = ax + b$, the vector y_p can be expressed as

$$y_p = (ax_1 + b), (ax_2 + b), ..., (ax_n + b)$$

In order to determine the equation of the line of best fit, we must minimise the distance between corresponding elements of y and y_p. Let us denote the sum of the squares of the distances between corresponding elements of the vectors y and y_p as J. Therefore,

$$J = (y_{p1} - y_1)^2 + (y_{p2} - y_2)^2 + ... + (y_{pn} - y_n)^2$$

J can also be written as

$$J = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 + ... + (ax_n + b - y_n)^2$$

Expanding J yields

$$J = a^2x_1^2 + 2(ax_1)(b - y_1) + (b - y_1)^2 + a^2x_2^2 + 2(ax_2)(b - y_2) + (b - y_2)^2 + ... + a^2x_n^2 + 2(ax_n)(b - y_n) + (b - y_n)^2$$

* *E-mail: me@nirbhay.co*

In sigma notation, J can be expressed as

$$J = \sum_{k=1}^{n} a^2x_k^2 + 2(ax_k)(b - y_n) + (b - y_k)^2$$

In order to obtain values for a and b, we must obtain equations for the same. Rearranging J as a quadratic expression in terms of a by determining the coefficients of a^2 and a yields

$$J = (x_1^2 + x_2^2 + ... + x_n^2)a^2 + 2((b - y_1)x_1 + (b - y_2)x_2 + ... + (b - y_n)x_n))a + (b - y_1)^2 + (b - y_2)^2 + ... + (b - y_n)^2$$

This can be further rearranged to yield

$$J = (x_1^2 + x_2^2 + ... + x_n^2)a^2 + 2((x_1 + x_2 + ... + x_n)b - (x_1y_1 + x_2y_2 + ... + x_ny_n)a + (b - y_1)^2 + (b - y_2)^2 + ... + (b - y_n)^2$$

This expression gives the distance or the expected error with a value for a Since it is a quadratic expression in the form $y = ax^2 + bx + c$, the function can be minimised using either differentiation to find the minima or via the use of the formula $x_{min} = \frac{-b}{2a}$, where a is the coefficient of the squared term and b is the coefficient of the term with power 1. Using the formula above,

$$a = \frac{-(2(x_1 + x_2 + ... + x_n)b - (x_1y_1 + x_2y_2 + ... + x_ny_n))}{2(x_1^2 + x_2^2 + ... + x_n^2)}$$

This can be rearranged to obtain:

$$a(x_1^2 + x_2^2 + ... + x_n^2) + (x_1 + x_2 + ... + x_n)b = x_1y_1 + x_2y_2 + ... + x_ny_n$$

This is the first equation in a system of two linear equations to find the values of a and b which minimise the error. In order to derive the second equation, we must repeat the process above but instead, derive the quadratic expression in terms of b. Recall that

$$J = a^2x_1^2 + 2(ax_1)(b - y_1) + (b - y_1)^2 + a^2x_2^2 + 2(ax_2)(b - y_2) + (b - y_2)^2 + ... + a^2x_n^2 + 2(ax_n)(b - y_n) + (b - y_n)^2$$

Rearranging J as a quadratic expression in terms of b by determining the coefficients of b^2 and b yields

$$J = (b^2 + b^2 + ...b^2) + 2((ax_1 - y_1) + (ax_2 - y_2) + ... + (ax_n - y_n))b + K$$

where K is a constant term. This can be further rearranged.

$$J = nb^2 + 2(ax_1 - y_1 + ax_2 - y_2 + ... + ax_n - y_n)b + K$$

Thus, the coefficient of b^2 is n and the coefficient of b is $2(a(x_1 + x_2 + ... + x_n) - (y_1 + y_2 + ... + y_n))$. Using the formula

$$x_{min} = \frac{-b}{2a}$$

The minimum value of b is given by

$$b = \frac{-(2(a(x_1 + x_2 + ... + x_n) - (y_1 + y_2 + ... + y_n)))}{2(n)}$$

This can be further simplified to

$$b = \frac{(y_1 + y_2 + ... + y_n) - a(x_1 + x_2 + ... + x_n)}{n}$$

From this, we know that

$$a(x_1 + x_2 + ... + x_n) + bn = y_1 + y_2 + ... + y_n$$

This is the second equation required to find the values for a and b which give the minimum error. Therefore, the system of equations is

$$a(x_1 + x_2 + ... + x_n) + nb = y_1 + y_2 + ... + y_n$$
$$a(x_1^2 + x_2^2 + ... + x_n^2) + (x_1 + x_2 + ... + x_n)b = x_1y_1 + x_2y_2 + ... + x_ny_n$$

This can be written in matrix form

$$\begin{bmatrix} (x_1 + x_2 + ... + x_n) & n \\ ({x_1}^2 + {x_2}^2 + ...{x_n}^2) & (x_1 + x_2 + ... + x_n) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} (y_1 + y_2 + ... + y_n) \\ (x_1y_1 + x_2y_2 + ... + x_ny_n) \end{bmatrix}$$

Solving the system of equations for a and b yields

$$a = \frac{((x_1 + x_2 + ...x_n) * (y_1 + y_2 + ... + y_n)) - (n * (x_1y_1 + x_2y_2 + ... + x_ny_n))}{(x_1 + x_2 + ... + x_n)^2 - ({x_1}^2 + x_2{}^2_+ ...{x_n}^2 *)n}$$
$$b = \frac{-(({x_1}^2 + {x_2}^2 + ... + {x_n}^2) * (y_1 + y_2 ... + y_n)) + (x_1 + x_2 + ... + x_n) * (x_1y_1 + x_2y_2 + ... + x_ny_n)}{(x_1 + x_2 + ... + x_n)^2 - ({x_1}^2 + {x_2}^2 + ... + {x_n}^2 * (n))}$$

This generalised formula can be used to find the equation of the line of regression for any n data points. This concludes Section 3 of this paper.

2. Find the Equation of the Plane of Best Fit

Given a set of n 3-dimensional points $(x_1, y_1, z_1), ...(x_n, y_n, z_n)$, our goal is to derive a generalised formula for a, b, c such that the error between the plane $z = ax + by + c$ and the given data set is at a minimum. Let the vector z denote all the z-coordinates in the data set.

$$z = (z_1, z_2, ..., z_n)$$

Now, let us denote a vector z_p which denotes the the output values of the equation of the plane of best fit $z = ax + by + c$. Therefore,

$$z_p = (ax_1 + by_1 + c), (ax_2 + by_2 + c), ..., (ax_n + by_n + c)$$

Let J denote the square of the distance between corresponding points of z and z_p. Therefore,

$$J = (ax_1 + by_1 + c - z_1)^2 + (ax_2 + by_2 + c - z_1)^2 + ... + (ax_n + by_n + c - z_n)^2$$

Since there are 3 unknowns, we will need 3 equations to find the values of a, b, c. First, let us find a quadratic equation in terms of a. Since

$$J = (ax_1 + by_1 + c - z_1)^2 + (ax_2 + by_2 + c - z_1)^2 + ... + (ax_n + by_n + c - z_n)^2$$

Expanding J yields

$$J = a^2({x_1}^2 + {x_2}^2 + ... + {x_n}^2) + b^2({y_1}^2 + {y_2}^2 + ... + {y_n}^2) + 2a(x_1y_1 + x_2y_2 + ... + x_ny_n) + c^2 + {z_1}^2 - 2c(z_1 + z_2 + ... + z_n)$$
$$+ 2c(z_1 + z_2 + ... + z_n) + 2ac(x_1 + x_2 + .. + x_n) - 2a(x_1z_1 + x_2z_2 + ... + x_nz_n) + 2bc(y_1 + y_2 + .. + y_n)$$
$$- 2b(y_1z_1 + y_2z_2 + ... + y_nz_n)$$

In order to determine a quadratic equation in terms of a, we must determine the coefficients of a^2 and a. From the equation above, the coefficient of a^2 is $({x_1}^2 + {x_2}^2 + ... + {x_n}^2)$. The coefficient of a is $2(b(x_1y_1 + x_2y_2 + ... + x_ny_n) + c(x_1 + x_2 + ... + x_n) - (x_1z_1 + x_2z_2 + ... + x_nz_n))$. Using the formula

$$x_{min} = \frac{-b}{2a}$$

The value for a which minimises the error is given by

$$a = \frac{-(2(b(x_1y_1 + x_2y_2 + ... + x_ny_n) + c(x_1 + x_2 + ... + x_n) - (x_1z_1 + x_2z_2 + ... + x_nz_n)))}{2({x_1}^2 + {x_2}^2 + ... + {x_n}^2)}$$

This can be simplified to obtain

$$a({x_1}^2 + {x_2}^2 + ... + {x_n}^2) + b(x_1y_1 + x_2y_2 + ... + x_ny_n) + (x_1 + x_2 + ... + x_n)c = x_1z_1 + x_2z_2 + ... + x_nz_n$$

This is the first in a system of 3 linear equations in terms of a, b, c. To obtain the second equation, we will rearrange J as a quadratic expression in terms of b. Recall that J is given by

$$J = a^2({x_1}^2 + {x_2}^2 + ... + {x_n}^2) + b^2({y_1}^2 + {y_2}^2 + ... + {y_n}^2) + 2a(x_1y_1 + x_2y_2 + ... + x_ny_n) + c^2 + {z_1}^2 - 2c(z_1 + z_2 + ... + z_n)$$
$$+ 2c(z_1 + z_2 + ... + z_n) + 2ac(x_1 + x_2 + .. + x_n) - 2a(x_1z_1 + x_2z_2 + ... + x_nz_n) + 2bc(y_1 + y_2 + .. + y_n)$$
$$- 2b(y_1z_1 + y_2z_2 + ... + y_nz_n)$$

From the equation above, the coefficient of b^2 is $({y_1}^2 + {y_2}^2 + ... + {y_n}^2)$. The coefficient of b is $2(y_1 + y_2 + ... + y_n)(a(x_1 + x_2 + ... + x_n) - (z_1 + z_2 + ... + z_n) + c)$. Using the formula

$$x_{min} = \frac{-b}{2a}$$

The value for b which minimises the error is given by

$$b = \frac{-2(y_1 + y_2 + ... + y_n)(a(x_1 + x_2 + ... + x_n) - (z_1 + z_2 + ... + z_n) + c)}{2({y_1}^2 + {y_2}^2 + ... + {y_n}^2)}$$

This can be rearranged to obtain

$$b({y_1}^2 + {y_2}^2 + ...{y_n}^2) + a(x_1y_1 + x_2y_2 + ... + x_ny_n) + c(y_1 + y_2 + ... + y_n) = y_1z_1 + y_2z_2 + ... + y_nz_n$$

This is the second equation in a system of 3 linear equations. To determine the third and final equation, we must rearrange J as a a quadratic expression in terms of c. Recall that J is given by

$$J = a^2({x_1}^2 + {x_2}^2 + ... + {x_n}^2) + b^2({y_1}^2 + {y_2}^2 + ... + {y_n}^2) + 2a(x_1y_1 + x_2y_2 + ... + x_ny_n) + c^2 + {z_1}^2 - 2c(z_1 + z_2 + ... + z_n)$$

$$+2c(z_1+z_2+...+z_n)+2ac(x_1+x_2+..+x_n)-2a(x_1z_1+x_2z_2+...+x_nz_n)+2bc(y_1+y_2+..+y_n)$$
$$-2b(y_1z_1+y_2z_2+...+y_nz_n)$$

From the equation above, the coefficient of c^2 is $\sum_{k=1}^{n} 1 = n$. The coefficient of c is $2(a(x_1+x_2+...+x_n)+b(y_1+y_2+...+y_n)-(z_1+z_2+...+z_n))$. Using the formula

$$x_{min} = \frac{-b}{2a}$$

The value for c which minimises the error is given by

$$c = \frac{-(2(a(x_1+x_2+...+x_n)+b(y_1+y_2+...+y_n)-(z_1+z_2+...+z_n))}{2n}$$

This can be rearranged to obtain

$$cn + a(x_1+x_2+...+x_n)+b(y_1+y_2+...+y_n) = z_1+z_2+...+z_n$$

This is the third equation in a system of 3 linear equations.

$$cn + a(x_1+x_2+...+x_n)+b(y_1+y_2+...+y_n) = z_1+z_2+...+z_n$$
$$b({y_1}^2+{y_2}^2+...{y_n}^2)+a(x_1y_1+x_2y_2+...+x_ny_n)+c(y_1+y_2+...+y_n) = y_1z_1+y_2z_2+...+y_nz_n$$
$$a({x_1}^2+{x_2}^2+...+{x_n}^2)+b(x_1y_1+x_2y_2+...+x_ny_n)+(x_1+x_2+...+x_n)c = x_1z_1+x_2z_2+...+x_nz_n$$

The solutions of these three linear equations will give the values of a, b, c which will minimise the error between the plane of best fit $z = ax + by + c$ and the data set.

3. Extending Linear Regression into an n Dimensional Space

Consider k points in a $n-1$ dimensional space. Our goal is to generalise a method to find the equation of a $n-1$th dimensional surface which minimises the distance between the set of k points and itself. In order to do this, consider a matrix of k points in $n-1$ dimensions called A which consists of our input variables denoted by x_n^k.

$$A = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_n^1 \\ x_1^2 & x_2^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^k & x_2^k & \dots & x_n^k \end{bmatrix}$$

Let us also consider a matrix B which consists of k points which are our output variables denoted by y_k

$$B = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

In this case, our unknown variables are the coefficients of $x_1, x_2, \ldots, x_k$ denoted by $b_1, b_2, \ldots, b_k$ and the constant term b_0 respectively. Let us now create a matrix called x to denote these unknowns

$$x = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}$$

Since the constant term b_0 is present, we must modify the matrix A to reflect these changes.

$$A = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \ldots & x_n^1 \\ 1 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x_1^k & x_2^k & \ldots & x_n^k \end{bmatrix}$$

In order to solve for the matrix x, we must solve $Ax = B$. However, in this case, the matrix A is not invertible. Attempting to solve for x will lead to no solution. In order to overcome this problem, we must multiply both sides of the equality by A^T. The transpose of A times A will always be square and symmetrical, so the matrix obtained will always be invertible. This will allow us to solve for x. Multiplying both sides by A^T yields $A^T Ax = A^T B \rightarrow x = (A^T A)^{-1} A^T B$. This gives us the result we need. An example is outlined below.

Let A_{points} denote a set of 4 points in 3 dimensional space

$$A_{points} = \{(1,3,2),(4,5,1),(7,6,4),(9,8,1)\}$$

Let A_{mat} denote a matrix containing the input variables of these points.

$$A_{mat} = \begin{bmatrix} 1 & 1 & 3 \\ 1 & 4 & 5 \\ 1 & 7 & 6 \\ 1 & 9 & 8 \end{bmatrix}$$

Let us now denote a matrix B containing the output variables of these points.

$$B = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 1 \end{bmatrix}$$

Let us now denote a matrix x to hold the values of the coefficients of x_1,x_2 (b_1,b_2) since there are only two input variables and a constant term b_0.

$$x = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

We know that $A_{mat} * x = B$. However, since the matrix A_{mat} is not invertible, we must follow the steps outlined above.

$$A_{mat}^T A_{mat} x = A_{mat}^T B$$

$$\rightarrow x = (A_{mat}^T A_{mat})^{-1} A_{mat}^T B$$

$$\rightarrow x = \left[\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 7 & 9 \\ 3 & 5 & 6 & 8 \end{bmatrix} * \begin{bmatrix} 1 & 1 & 3 \\ 1 & 4 & 5 \\ 1 & 7 & 6 \\ 1 & 9 & 8 \end{bmatrix} \right]^{-1} * \left[\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 4 & 7 & 9 \\ 3 & 5 & 6 & 8 \end{bmatrix} * \begin{bmatrix} 2 \\ 1 \\ 4 \\ 1 \end{bmatrix} \right]$$

Computation of the above yields

$$x = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \frac{681}{62} \\ \frac{69}{31} \\ \frac{-233}{62} \end{bmatrix}$$

Therefore, $y = \frac{69}{31}x_1 - \frac{233}{62}x_2 + \frac{681}{62}$. These results can be corroborated by using software.

4. Conclusion

In this paper, we first derived the formula for linear regression in the 2D and 3D planes algebraically, using a system of linear equations to generalise a formula to obtain the values of the coefficients of the terms in the equation. We also derived a generalised formula for linear regression in n dimensions using linear algebra and matrices. This was done by pre-multiplying both sides by the transpose of the matrix.

References

[1] Shalabh, *Improving the Predictions in Linear Regression Models*, Journal of Statistical Research, 33(1)(1999).

[2] Shalabh, H. Toutenburg and C. Heumann, *Mean squared error matrix comparison of least squares and Stein-rule estimators for regression coefficients under non-normal disturbances*, Discussion paper 496, University of Munich, Munich, Germany, (2006).

[3] A. Schneider, G. Hommel and M. Blettner, *Linear regression analysis: part 14 of a series on evaluation of scientific publications.* Deutsches Arzteblatt International, 107(44)(2010), 776782.