The Mathematics of Multiple Linear Regression

Regression is a subset of supervised machine learning. In this predictive analytics method, the target variable is of the continuous type, 20.4, 100.8, 1098.25, etc. This is in contrast to the classification type that yields a categorical output, (Yes, No), (Positive, Neutral, Negative), etc. Common use cases include: *Credit score prediction for banking domain*, *monthly sales in the marketing domain*, *house price prediction in the real estate domain, etc.*

Over time, many techniques have been developed to solve different regression problems, each having its unique intuition and statistical foundation. For the purpose of this article, the technique of focus is *Linear Regression*.

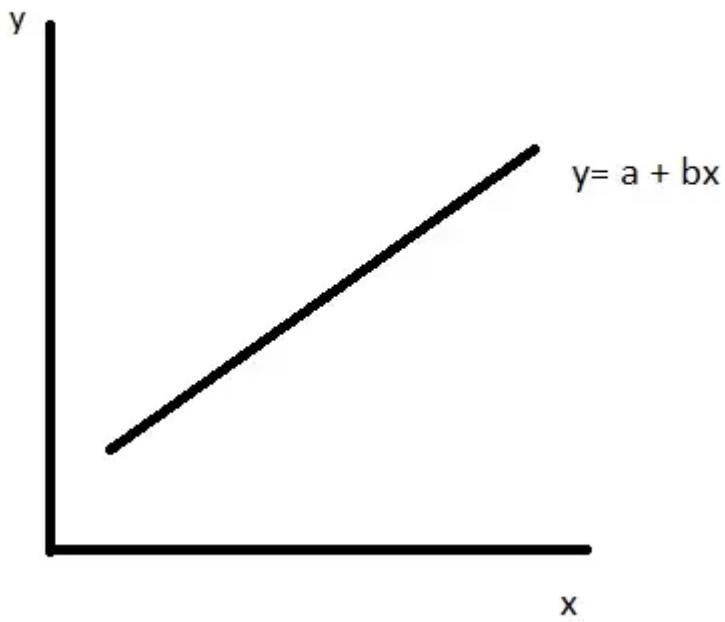
Linear regression is one of the common methods applied in solving regression problems. The idea of this method is developing a model using a linear equation to fit to the dataset.

The linear equation, the predictive power of Linear regression is the equation of a straight line, which is expressed as:

$y = a + bx$

Where:

- y = output variable
- a = intercept
- b = slope
- x = predictor/input variable



A simple linear regression graph

- The intercept(a) is the value of the output value when the input value is 0.
- The slope indicates the rate at which the output(y) changes with every unit increase in the input(x).

In a practical environment, regression problems deal with multiple predictor variables, this requires a more complex solution using a polynomial equation. Expressed as:

$y = a + b1x1 + b2x2 + b3x3 + b4x4 \ldots\ldots\ldots + bnxn$

where:

- n = ith number for predictor variables
- a = intercept
- b = feature coefficient/ feature rate/ feature weight / regression coefficient
- x = predictor variable
- y = output variable

A regression model is established when the intercept and feature coefficients are known. Using a 3 predictor feature problem for example, the solution is a model of the form y = a + b1x1 + b2x2 + b3x3.

To determine the regression coefficients

- **Step 1:** the normal equation is applied and expressed as below.

$$x_1 : \sum x_1^2 b_1 + \sum x_1 x_2 b_2 + \sum x_1 x_3 b_3 = \sum x_1 y$$

$$x_2 : \sum x_1 x_2 b_1 + \sum x_2^2 b_2 + \sum x_2 x_3 b_3 = \sum x_2 y$$

$$x_3 : \sum x_1 x_3 b_1 + \sum x_2 x_3 b_2 + \sum x_3^2 b_3 = \sum x_3 y$$

Normal Equation

- **Step 2: Determine the regression sums**

To determine the regression sums as expressed in the normal equation.

$$\sum x_i^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$\sum x_i y = \sum x_i y - \frac{(\sum x_1)(\sum y)}{n}$$

$$\sum x_i x_j = \sum x_i x_j - \frac{(\sum x_i)(\sum x_j)}{n}$$

Regression Sum

- **Step 3:** Create a matrix representation of the normal equation.

Applying the normal equation in determining the value of the regression coefficients, requires decomposing the normal equation into its matrix form.

$$A = \begin{matrix} \sum x_1^2 & \sum x_1 x_2 & \sum x_1 x_3 \\ \sum x_1 x_2 & \sum x_2^2 & \sum x_2 x_3 \\ \sum x_1 x_3 & \sum x_2 x_3 & \sum x_3^2 \end{matrix}$$

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$C = \begin{bmatrix} \sum x_1 y \\ \sum x_2 y \\ \sum x_3 y \end{bmatrix}$$

where ***A.B = C***

- **Step 4:** Determine the inverse of matrix A

Multiply both sides by the inverse of *matrix A*

$$A^{-1}.A.B = A^{-1}.C$$

Since $A^{-1}.A = I$

Where I = identity matrix

$$I.B = C.A^{-1}$$

$$B = C.A^{-1}$$

Step 5: Solve for the intercept “a”

Intercept “a”

Using $\bar{y} = a + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3$

Where:

$$\bar{y} = \frac{\sum y}{n}$$

$$\bar{x}_1 = \frac{\sum x_1}{n}$$

$$\bar{x}_2 = \frac{\sum x_2}{n}$$

$$\bar{x}_3 = \frac{\sum x_3}{n}$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$$

EXAMPLE

Using a dummy house price data to demonstrate the power of linear regression.

HOUSE PRICE DATA

	Room (x_1)	Bathroom (x_2)	Landsize (x_3)	Price in thousand (y)
	5	2	10	10
	8	3	12	15
	6	2	11	12
	10	3	12	20
	8	3	11	14
	7	3	10	13
	6	2	10	10
	12	4	15	25
	5	2	8	9
	10	4	13	23

House price data

From the above dataset, there are three predictor variables (x1, x2, x3). This indicates that the regression solution is a multi-variable regression model. Simply expressed as: $y = a + b1x1 + b2x2 + b3x3$

The challenge is to determine the regression coefficients ($b1$, $b2$, $b3$)and the intercept (a).

- **Step 1**: determine

$$x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3, x_1y, x_2y, x_3y$$

And their corresponding sums (∑)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		HOUSE PRICE DATA												
2		Room (x_1)	Bathroom (x_2)	Landsize (x_3)	Price in thousand (y)	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3	x_1y	x_2y	x_3y
3		5	2	10	10	25	4	100	10	50	20	50	20	100
4		8	3	12	15	64	9	144	24	96	36	120	45	180
5		6	2	11	12	36	4	121	12	66	22	72	24	132
6		10	3	12	20	100	9	144	30	120	36	200	60	240
7		8	3	11	14	64	9	121	24	88	33	112	42	154
8		7	3	10	13	49	9	100	21	70	30	91	39	130
9		6	2	10	10	36	4	100	12	60	20	60	20	100
10		12	4	15	25	144	16	225	48	180	60	300	100	375
11		5	2	8	9	25	4	64	10	40	16	45	18	72
12		10	4	13	23	100	16	169	40	130	52	230	92	299
13	∑	77	28	112	151	643	84	1288	231	900	325	1280	460	1782
14														

- **Step 2**: Calculate regression sums.

$$\sum x_1^2 = 643 - \frac{77^2}{10} = 50.1$$

$$\sum x_2^2 = 84 - \frac{28^2}{10} = 5.6$$

$$\sum x_3^2 = 1288 - \frac{112^2}{10} = 33.6$$

$$\sum x_1 y = 1280 - \frac{77*151}{10} = 117.3$$

$$\sum x_2 y = 460 - \frac{28*151}{10} = 37.2$$

$$\sum x_3 y = 1782 - \frac{112*151}{10} = 90.8$$

$$\sum x_1 x_2 = 231 - \frac{77*28}{10} = 15.4$$

$$\sum x_1 x_3 = 900 - \frac{77*112}{10} = 37.6$$

$$\sum x_2 x_3 = 325 - \frac{28*112}{10} = 11.4$$

- **Step 3**: Present regression sums in matrix order

$$A = \begin{bmatrix} 50.1 & 15.4 & 37.6 \\ 15.4 & 5.6 & 11.4 \\ 37.6 & 11.4 & 33.6 \end{bmatrix}$$

matrix of x^2 and xixj components

$$C = \begin{bmatrix} 117.3 \\ 37.2 \\ 90.8 \end{bmatrix}$$

xy components

- **Step 4**: Determine the inverse of matrix A

i. Find the determinant of A

Determinant of A

$|A| = 50.1((5.6 * 33.6) - (11.4 * 11.4)) - 15.4((15.4 * 33.6) - (11.4 * 37.6) + 37.6((15.4 * 15.40 - (5.6 * 37.6))$

$|A| = 2915.82 - 1367.52 - 1316$

$|A| = 232.3$

ii. Determine the matrix of minors

Matrix of minors

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$a_{11} = (5.6 * 33.6) - (1.4 * 11.4) = 58.2$

$a_{12} = (15.4 * 33.6) - (11.4 * 37.6) = 88.8$

$a_{13} = (15.4 * 11.4) - (5.6 * 37.6) = -35$

$a_{21} = (15.4 * 33.6) - (37.6 * 11.4) = 88.8$

$a_{22} = (50.1 * 33.6) - (37.6\ 8\ 37.6) = 269.6$

$a_{23} = (50.1 * 11.4) - (15.4 * 37.6) = -7.9$

$a_{31} = (15.4 * 11.4) - (37.6 * 5.6) = -35$

$a_{32} = (50.1 * 11.4) - (37.6 * 15.4) = -7.9$

$a_{33} = (50.1 * 5.6) - (15.4 * 15.4) = 43.4$

$$\begin{bmatrix} 58.2 & 88.8 & -35 \\ 88.8 & 269.6 & -7.9 \\ -35 & -7.9 & 43.4 \end{bmatrix}$$

iii. Determine the matrix cofactor

Cofactor matrix

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix} = \begin{bmatrix} 58.2 & -88.8 & -35 \\ -88.8 & 269.6 & 7.9 \\ -35 & 7.9 & 43.4 \end{bmatrix}$$

iv. Solve for the matrix Adjoint

Adjoint

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 58.2 & -88.8 & -35 \\ -88.8 & 269.6 & 7.9 \\ -35 & 7.9 & 43.4 \end{bmatrix}$$

v. Determine the inverse of matrix A

$$A^{-1} = \frac{1}{|A|} * \text{Adjoint} = \frac{1}{232.3} * \begin{bmatrix} 58.2 & -88.8 & -35 \\ -88.8 & 269.6 & 7.9 \\ -35 & 7.9 & 43.4 \end{bmatrix}$$

$$A^{-1} = \begin{bmatrix} 0.2505 & -0.38226 & -0.15067 \\ -0.38226 & 1.16057 & 0.034 \\ -0.15067 & 0.034 & 0.18683 \end{bmatrix}$$

- **Step 5**: Multiply inverse of matrix A by matrix C to determine the values of the feature coefficients b1, b2, b3

$$A^{-1} * C = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.2505 & -0.38226 & -0.15067 \\ -0.38226 & 1.16057 & 0.034 \\ -0.15067 & 0.034 & 0.18683 \end{bmatrix} * \begin{bmatrix} 117.3 \\ 37.2 \\ 90.8 \end{bmatrix}$$

$b_1 = (0.2505*117.3) + (-0.38226 * 37.2) + (-0.15067 * 90.8)$

$b_2 = (-0.38226 * 117.3) + (1.16057 * 37.2) + (0.034 * 90.8)$

$b3 = (-0.15067 * 117.3) + (0.034 * 37.2) + (0.18683 * 90.8)$

$b_1 = 1.483$

$b_2 = 1.421306$

$b_3 = 0.555373$

- **Step 6**: Determine the intercept “a”

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - b_3\bar{x}_3$$

$$a = \frac{151}{10} - (1.483 * \frac{77}{10}) - (1.421306 * \frac{28}{10}) - (0.555373 * \frac{112}{10})$$

$$a = -6.52$$

The model to solve the house price prediction problem can then be expressed as

$$y = -6.52 + 1.483x_1 + 1.421306x_2 + 0.555373x_3$$

at x1=0, x2 =0, x3=0, y = -6.52units

Model Evaluation

Using the model on the original data to evaluate its performance.

Making a sample prediction by determining the output value of the first row in the sample data.

$$y = -6.52 + 1.483x_1 + 1.421306x_2 + 0.555373x_3$$

$$\text{At } x_1 = 5, x_2 = 2, x_3 = 10$$

$$y = -6.52 + 1.483*5 + 1.421306*2 + 0.555373*10$$

$$y = 9.291342$$

The predicted values for all rows are

A	B	C	D	E	F
			HOUSE PRICE DATA		
	Room (x_1)	Bathroom (x_2)	Landsize (x_3)	Price in thousand (y)	Predicted Price in thousand (y')
	5	2	10	10	9.291342
	8	3	12	15	16.272394
	6	2	11	12	11.329715
	10	3	12	20	19.238394
	8	3	11	14	15.717021
	7	3	10	13	13.678648
	6	2	10	10	10.774342
	12	4	15	25	25.291819
	5	2	8	9	8.180596
	10	4	13	23	21.215073
∑	77	28	112	151	150.989344

Column F holds the predicted value for each data row

Mean Absolute Error (MAE)

This is the mean of absolute difference between the actual value and the predicted value.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - y'_i|$$

from the table, MAE can be established as:

```
mae = 1/10(|10–9.291342| + |15–16.272394| + |12–11.329715|…+|23–21.215073|)
```

E	F	G
	HOUSE PRICE DATA	
Price in thousand (y)	Predicted Price in thousand (y')	Absolute Error
10	9.291342	0.708658
15	16.272394	1.272394
12	11.329715	0.670285
20	19.238394	0.761606
14	15.717021	1.717021
13	13.678648	0.678648
10	10.774342	0.774342
25	25.291819	0.291819
9	8.180596	0.819404
23	21.215073	1.784927
151	150.989344	9.479104

mae = 9.479104/10

mae = 0.9479104

Root Mean Squared Error (RMSE)

This indicates how close the predicted values are to the actual value.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y'_i)^2}$$

RMSE for the model can be determined thus,

```
rmse = 1/10((10–9.291342)² + (15–16.272394)² + (12–11.329715)²…+(23–21.215073)²)
```

E	F	G	H
	HOUSE PRICE DATA		
Price in thousand (y)	Predicted Price in thousand (y')	Absolute Error	Square Error
10	9.291342	0.708658	0.502196161
15	16.272394	1.272394	1.618986491
12	11.329715	0.670285	0.449281981
20	19.238394	0.761606	0.580043699
14	15.717021	1.717021	2.948161114
13	13.678648	0.678648	0.460563108
10	10.774342	0.774342	0.599605533
25	25.291819	0.291819	0.085158329
9	8.180596	0.819404	0.671422915
23	21.215073	1.784927	3.185964395
151	150.989344	9.479104	11.10138373

rmse = (11.10138373/10)^0.5

rmse = 1.053631042units

This means that the predicted values are 1053.63units of currency away from the actual value. *NB: y is in thousands hence (1.05363 * 1000)*

R-Squared (R^2)

This measures how well the model fits the data. Value ranges from 0–1, a value closer to one denotes a good model while value closer to zero indicates a poorly fitted model.

R^2 = 1 - ((sum of square residuals)/(sum of square total))

$$R^2 = 1 - \frac{\sum(y_i - y'_i)^2}{\sum(y_i - \bar{y})^2}$$

Where:

y_i = i^{th} value of output variable

$\bar{y}$ = mean of the output variable "y"

y'_i = i^{th} value of the predicted output

To determine the model R^2

E	F	G	H
	HOUSE PRICE DATA		
Price in thousand (y)	Predicted Price in thousand (y')	SSR	SST
10	9.291342	0.502196161	26.01
15	16.272394	1.618986491	225
12	11.329715	0.449281981	144
20	19.238394	0.580043699	400
14	15.717021	2.948161114	196
13	13.678648	0.460563108	169
10	10.774342	0.599605533	100
25	25.291819	0.085158329	625
9	8.180596	0.671422915	81
23	21.215073	3.185964395	529
151	150.989344	11.10138373	2495.01

from the table above, the sum of square residuals (SSR) = 11.10138373 while the sum of square total (SST) = 2495.01

therefore R^2 = 1- (11.10138373/2495.01)

R^2 = 0.9955

R^2 expressed in percentage yields 99.55%

This means that 99.55% of the variation in the output values is accounted for by the input values.

In summary, a careful stride using the steps as applied in the solution above will solve regression problems.