Module 3: Multivariate Linear Regression

Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis. Multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables. Based on the number of independent variables, we try to predict the output.

Multivariate regression tries to find out a formula that can explain how factors in variables respond simultaneously to changes in others.

There are numerous areas where multivariate regression can be used. Let's look at some examples to understand multivariate regression better.

1. Praneeta wants to estimate the price of a house. She will collect details such as the location of the house, number of bedrooms, size in square feet, amenities available, or not. Basis these details price of the house can be predicted and how each variables are interrelated.
2. An agriculture scientist wants to predict the total crop yield expected for the summer. He collected details of the expected amount of rainfall, fertilizers to be used, and soil conditions. By building a Multivariate regression model scientists can predict his crop yield. With the crop yield, the scientist also tries to understand the relationship among the variables.
3. If an organization wants to know how much it has to pay to a new hire, they will take into account many details such as education level, number of experience, job location, has niche skill or not. Basis this information salary of an employee can be predicted, how these variables help in estimating the salary.
4. Economists can use Multivariate regression to predict the GDP growth of a state or a country based on parameters like total amount spent by consumers, import expenditure, total gains from exports, total savings, etc.
5. A company wants to predict the electricity bill of an apartment, the details needed here are the number of flats, the number of appliances in usage, the number of people at home, etc. With the help of these variables, the electricity bill can be predicted.

The above example uses Multivariate regression, where we have many independent variables and a single dependent variable.

Mathematical equation

The simple regression linear model represents a straight line meaning y is a function of x. When we have an extra dimension (z), the straight line becomes a plane.

Here, the plane is the function that expresses y as a function of x and z. The linear regression equation can now be expressed as:

y = m1.x + m2.z+ c

y is the dependent variable, that is, the variable that needs to be predicted.
x is the first independent variable. It is the first input.

m1 is the slope of x1. It lets us know the angle of the line (x).
z is the second independent variable. It is the second input.
m2 is the slope of z. It helps us to know the angle of the line (z).
c is the intercept. A constant that finds the value of y when x and z are 0.

The equation for a model with two input variables can be written as:

y = β0 + β1.x1 + β2.x2

What if there are three variables as inputs? Human visualizations can be only three dimensions. In the machine learning world, there can be n number of dimensions. The equation for a model with three input variables can be written as:

y = β0 + β1.x1 + β2.x2 + β3.x3

Below is the generalized equation for the multivariate regression model-

y = β0 + β1.x1 + β2.x2 +..... + βn.xn

Where n represents the number of independent variables, β0~ βn represents the coefficients, and x1~xn is the independent variable.

The multivariate model helps us in understanding and comparing coefficients across the output. Here, the small cost function makes Multivariate linear regression a better model.

Example: Multiple Linear Regression

Suppose we have the following dataset with one response variable *y* and two predictor variables X_1 and X_2:

y	X_1	X_2
140	60	22
155	62	25
159	67	24
179	70	20
192	71	15
200	72	14
212	75	14
215	78	11

Use the following steps to fit a multiple linear regression model to this dataset.

Step 1: Calculate X_1^2, X_2^2, X_1y, X_2y and X_1X_2.

	y	X_1	X_2		X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	140	60	22		3600	484	8400	3080	1320
	155	62	25		3844	625	9610	3875	1550
	159	67	24		4489	576	10653	3816	1608
	179	70	20		4900	400	12530	3580	1400
	192	71	15		5041	225	13632	2880	1065
	200	72	14		5184	196	14400	2800	1008
	212	75	14		5625	196	15900	2968	1050
	215	78	11		6084	121	16770	2365	858
Mean	181.5	69.375	18.125	Sum	38767	2823	101895	25364	9859
Sum	1452	555	145						

Step 2: Calculate Regression Sums.

Next, make the following regression sum calculations:

- $\Sigma x_1^2 = \Sigma X_1^2 - (\Sigma X_1)^2 / n = 38,767 - (555)^2 / 8 =$ **263.875**
- $\Sigma x_2^2 = \Sigma X_2^2 - (\Sigma X_2)^2 / n = 2,823 - (145)^2 / 8 =$ **194.875**
- $\Sigma x_1 y = \Sigma X_1 y - (\Sigma X_1 \Sigma y) / n = 101,895 - (555*1,452) / 8 =$ **1,162.5**
- $\Sigma x_2 y = \Sigma X_2 y - (\Sigma X_2 \Sigma y) / n = 25,364 - (145*1,452) / 8 =$ **-953.5**
- $\Sigma x_1 x_2 = \Sigma X_1 X_2 - (\Sigma X_1 \Sigma X_2) / n = 9,859 - (555*145) / 8 =$ **-200.375**

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

X_1^2	X_2^2	X_1y	X_2y	X_1X_2
3600	484	8400	3080	1320
3844	625	9610	3875	1550
4489	576	10653	3816	1608
4900	400	12530	3580	1400
5041	225	13632	2880	1065
5184	196	14400	2800	1008
5625	196	15900	2968	1050
6084	121	16770	2365	858
38767	2823	101895	25364	9859

Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 3: Calculate b_0, b_1, and b_2.

The formula to calculate b_1 is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_1} = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2] =$ **3.148**

The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_2} = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] =$ **-1.656**

The formula to calculate b_0 is: $y - b_1X_1 - b_2X_2$

Thus, $\mathbf{b_0} = 181.5 - 3.148(69.375) - (-1.656)(18.125) =$ **-6.867**

Step 5: Place b_0, b_1, and b_2 in the estimated linear regression equation.

The estimated linear regression equation is: $\hat{y} = b_0 + b_1*x_1 + b_2*x_2$

In our example, it is $\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation: $\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$

b_0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

b_1 = 3.148. A one unit increase in x_1 is associated with a 3.148 unit increase in y, on average, assuming x_2 is held constant.

b_2 = -1.656. A one unit increase in x_2 is associated with a 1.656 unit decrease in y, on average, assuming x_1 is held constant.