Multiple Linear Regression

When we want to understand the relationship between a single predictor variable and a response variable, we often use simple linear regression.

However, if we'd like to understand the relationship between *multiple* predictor variables and a response variable then we can instead use **multiple linear regression**.

If we have *p* predictor variables, then a multiple linear regression model takes the form:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$$

where:

- **Y**: The response variable
- **X_j**: The jth predictor variable
- **β_j**: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- **ε**: The error term

The values for β_0, β_1, B_2, … , β_p are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):

$$\mathbf{RSS = \Sigma(y_i - \hat{y}_i)^2}$$

where:

- **Σ**: A greek symbol that means *sum*
- **y_i**: The actual response value for the ith observation
- **$\hat{y}_i$**: The predicted response value based on the multiple linear regression model

$$b_1 = \frac{\left(\sum x_2^2\right)\left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$b_2 = \frac{\left(\sum x_1^2\right)\left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_1 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$a = b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

Example: Multiple Linear Regression by Hand

Suppose we have the following dataset with one response variable *y* and two predictor variables X1 and X2:

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 X12, X22, X1y, X2y and X1X2.

	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
Sum	38767	2823	101895	25364	9859

Step 2: Calculate Regression Sums.

Next, make the following regression sum calculations:

- $\Sigma x12 = \Sigma X12 - (\Sigma X1)2 / n = 38{,}767 - (555)2 / 8 =$ **263.875**
- $\Sigma x22 = \Sigma X22 - (\Sigma X2)2 / n = 2{,}823 - (145)2 / 8 =$ **194.875**
- $\Sigma x1y = \Sigma X1y - (\Sigma X1\Sigma y) / n = 101{,}895 - (555*1{,}452) / 8 =$ **1,162.5**
- $\Sigma x2y = \Sigma X2y - (\Sigma X2\Sigma y) / n = 25{,}364 - (145*1{,}452) / 8 =$ **-953.5**
- $\Sigma x1x2 = \Sigma X1X2 - (\Sigma X1\Sigma X2) / n = 9{,}859 - (555*145) / 8 =$ **-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
Sum	38767	2823	101895	25364	9859
Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 3: Calculate b0, b1, and b2.

The formula to calculate b1 is: [(Σx22)(Σx1y) – (Σx1x2)(Σx2y)] / [(Σx12) (Σx22) – (Σx1x2)2]

Thus, **b1** = [(194.875)(1162.5) – (-200.375)(-953.5)] / [(263.875) (194.875) – (-200.375)2] = **3.148**

The formula to calculate b2 is: [(Σx12)(Σx2y) – (Σx1x2)(Σx1y)] / [(Σx12) (Σx22) – (Σx1x2)2]

Thus, **b2** = [(263.875)(-953.5) – (-200.375)(1152.5)] / [(263.875) (194.875) – (-200.375)2] = **-1.656**

The formula to calculate b0 is: y – b1X1 – b2X2

Thus, **b0** = 181.5 – 3.148(69.375) – (-1.656)(18.125) = **-6.867**

Step 5: Place b0, b1, and b2 in the estimated linear regression equation.

The estimated linear regression equation is: ŷ = b0 + b1*x1 + b2*x2

In our example, it is **ŷ = -6.867 + 3.148x1 – 1.656x2**

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation: ŷ = -6.867 + 3.148x1 – 1.656x2

b0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

b1 = 3.148. A one unit increase in x1 is associated with a 3.148 unit increase in y, on average, assuming x2 is held constant.

b2 = -1.656. A one unit increase in x2 is associated with a 1.656 unit decrease in y, on average, assuming x1 is held constant.