Day 9: Multiple Linear Regression

Problem | Submissions | Leaderboard | Editorial | Tutorial

If Y is linearly dependent only on X, then we can use the ordinary least square regression line, $\hat{Y} = a + b \cdot X$. However, if Y shows linear dependency on m variables $X_1, X_2, \ldots, X_m$, then we need to find the values of a and m other constants ($b_1, b_2, \ldots, b_m$). We can then write the regression equation as:

$$\hat{Y} = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ldots + b_m \cdot X_m$$

Matrix Form of the Regression Equation

Let's consider that Y depends on two variables, X_1 and X_2. We write the regression relation as $\hat{Y} = a + b_1 \cdot X_1 + b_2 \cdot X_2$. Consider the following matrix operation:

$$\begin{bmatrix} 1 & X_1 & X_2 \end{bmatrix} \times \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = a + b_1 \cdot X_1 + b_2 \cdot X_2$$

We define two matrices, X and B:

- $X = \begin{bmatrix} 1 & X_1 & X_2 \end{bmatrix}$
- $B = \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix}$

Now, we rewrite the regression relation as $\hat{Y} = X \cdot B$. This transforms the regression relation into matrix form.

Generalized Matrix Form

We will consider that Y shows a linear relationship with m variables, X_1, X_2, ..., X_m. Let's say that we made n observations on n different tuples $(x_1, x_2, \ldots, x_m)$:

$y_1 = a + b_1 \cdot x_{1,1} + b_2 \cdot x_{2,1} + b_3 \cdot x_{3,1} + \ldots + b_m \cdot x_{m,1}$
$y_2 = a + b_1 \cdot x_{1,2} + b_2 \cdot x_{2,2} + b_3 \cdot x_{3,2} + \ldots + b_m \cdot x_{m,2}$
$y_3 = a + b_1 \cdot x_{1,3} + b_2 \cdot x_{2,3} + b_3 \cdot x_{3,3} + \ldots + b_m \cdot x_{m,3}$
$\ldots$
$y_n = a + b_1 \cdot x_{1,n} + b_2 \cdot x_{2,n} + b_3 \cdot x_{3,n} + \ldots + b_m \cdot x_{m,n}$

Now, we can find the matrices:

- $X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & x_{3,1} & \ldots & x_{m,1} \\ 1 & x_{1,2} & x_{2,2} & x_{3,2} & \ldots & x_{m,2} \\ 1 & x_{1,3} & x_{2,3} & x_{3,3} & \ldots & x_{m,3} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1,n} & x_{2,n} & x_{3,n} & \ldots & x_{m,n} \end{bmatrix}$
- $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \ldots \\ y_n \end{bmatrix}$

Finding the Matrix B

We know that $Y = X \cdot B$
$\Rightarrow X^T \cdot Y = X^T \cdot X \cdot B$

$\Rightarrow (X^T \cdot X)^{-1} \cdot X^T \cdot Y = I \cdot B$

$\Rightarrow B = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$

Note: M^T is the transpose matrix of M, M^{-1} is the inverse matrix of M, and I is the identity matrix.

Finding the Value of Y

Suppose we want to find the value of Y for some tuple $(x_1, x_2, x_3, \ldots, x_m)$, then,

$$Y = \begin{bmatrix} 1 & x_1 & x_2 & \ldots & x_m \end{bmatrix} \times B$$

Example

Consider Y shows a linear relationship with X_1 and X_2:

$X_1 = \{5, 6, 7, 8, 9\}$

$X_2 = \{7, 6, 4, 5, 6\}$

$Y = \{10, 20, 60, 40, 50\}$

Now, we can define the matrices:

- $X = \begin{bmatrix} 1 & 5 & 7 \\ 1 & 6 & 6 \\ 1 & 7 & 4 \\ 1 & 8 & 5 \\ 1 & 9 & 6 \end{bmatrix}$
- $Y = \begin{bmatrix} 10 \\ 20 \\ 60 \\ 40 \\ 50 \end{bmatrix}$

Now, find the value of B:

- $X^T X = \begin{bmatrix} 5 & 35 & 28 \\ 35 & 255 & 193 \\ 28 & 193 & 162 \end{bmatrix}$
- $(X^T X)^{-1} = \begin{bmatrix} 18.8884 & -1.23721 & -1.7907 \\ -1.23721 & 0.12093 & 0.0697674 \\ -1.7907 & 0.0697674 & 0.232558 \end{bmatrix}$
- $(X^T X)^{-1} X^T = \begin{bmatrix} 0.167442 & 0.72093 & 3.06512 & 0.0372093 & -2.9907 \\ -0.144186 & -0.0930233 & -0.111628 & 0.0790698 & 0.269767 \\ 0.186047 & 0.0232558 & -0.372093 & -0.0697674 & 0.232558 \end{bmatrix}$
- $(X^T X)^{-1} X^T Y = \begin{bmatrix} 51.9535 \\ 6.65116 \\ -11.1628 \end{bmatrix}$

So, $B = \begin{bmatrix} 51.9535 \\ 6.65116 \\ -11.1628 \end{bmatrix}$, which means $a = 51.9535$, $b_1 = 6.65116$, and $b_2 = -11.1628$.

Let's find the value of Y at $(x_1 = 5, x_2 = 5)$

$$Y = \begin{bmatrix} 1 & 5 & 5 \end{bmatrix} \times \begin{bmatrix} 51.9535 \\ 6.65116 \\ -11.1628 \end{bmatrix} = 29.39535$$

Multiple Regression in R

```
x1 = c(5, 6, 7, 8, 9)
x2 = c(7, 6, 4, 5, 6)
y = c(10, 20, 60, 40, 50)
m = lm(y ~ x1 + x2)
show(m)
```

Running the above code produces the following output:

```
Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     51.953        6.651      -11.163
```

Multiple Regression in Python

```
from sklearn import linear_model
x = [[5, 7], [6, 6], [7, 4], [8, 5], [9, 6]]
y = [10, 20, 60, 40, 50]
lm = linear_model.LinearRegression()
lm.fit(x, y)
a = lm.intercept_
b = lm.coef_
print a, b[0], b[1]
```

Running the above code produces the following output:

```
51.9534883721 6.6511627907 -11.1627906977
```

Solve Problem