Mathematics of Multiple Regression

https://www.statology.org/multiple-linear-regression-by-hand/

Multiple linear regression is a method used to quantify the relationship between two or more predictor variables and a response variable. Suppose we have the following dataset with one response variable, y, and two predictor variables, X_1 and X_2:

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

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
	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

<u>Step 2</u>: Calculate Regression Sums:

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

	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

<u>Step 3</u>: Calculate b_0, b_1, and b_2:

$$\mathbf{b_0} = \bar{y} - b_1\bar{X}_1 - b_2\bar{X}_2$$
$$= 181.5 - 3.148(69.375) - (-1.656)(18.125) \ = \mathbf{-6.867}$$

$$\mathbf{b_1} = [(\Sigma x_2^2)(\Sigma x_1 y) \ - (\Sigma x_1 x_2)(\Sigma x_2 y)] \ / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$$
$$= [(194.875)(1162.5) \ - (-200.375)(-953.5)] \ / [(263.875)(194.875) - (-200.375)^2]$$
$$= \mathbf{3.148}$$

$$\mathbf{b_2} = [(\Sigma x_1^2)(\Sigma x_2 y) \ - (\Sigma x_1 x_2)(\Sigma x_1 y)] \ / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$$
$$= [(263.875)(-953.5) \ - (-200.375)(1152.5)] \ / [(263.875)(194.875) - (-200.375)^2]$$
$$= \mathbf{-1.656}$$

<u>Step 5</u>: Place b_0, b_1, and b_2 in the estimated linear regression equation:

$$\hat{y} = b_0 + b_1{*}x_1 + b_2{*}x_2$$
$$= \mathbf{-6.867 + 3.148x_1 - 1.656x_2}$$

Where $\mathbf{X_1}$ and $\mathbf{X_2}$ are the inputs for the prediction.

Home work

- Predict the value of **BMI** (Body Mass Index) from the ***Height*** and ***Weight*** of a person using logistic regression.
- Download "**Gender-Height-Weight-BMI**" CSV dataset from https://www.kaggle.com/datasets/yersever/500-person-gender-height-weight-bodymassindex/
- Rename the dataset 'bmi.csv'

About Dataset

Gender: Male / Female (***no need to include this column in your program***)

Height: Number (cm)

Weight: Number (Kg)

Prediction BMI Index (your simulation should include the following text result with each BMI prediction):

0 - Extremely Weak

1 - Weak

2 - Normal

3 - Overweight

4 - Obesity

5 - Extreme Obesity