Multiple Linear regression

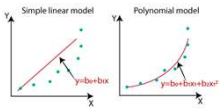
Assumptions

- **Independence**: the scores of any particular subject are independent of the scores of all other subjects
- **Normality**: in the population, the scores on the dependent variable are normally distributed for each of the possible combinations of the level of the X variables; each of the variables is normally distributed
- **Homoscedasticity**: in the population, the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.

  *(In statistics, **homoscedasticity** is when a sequence of random variables has the same finite variance, or homogeneity of variance. It can also refer to when the variance in scores on one variable is similar at all values of another variable.)*
- **Linearity**: In the population, the relation between the dependent variable and the independent variable is linear when all the other independent variables are held constant.

Simple vs. Multiple Regression

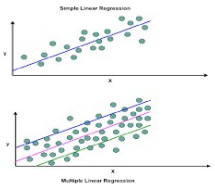
- One dependent variable Y predicted from one independent variable X
- One regression coefficient
- r^2: proportion of variation in dependent variable Y predictable from X



- One dependent variable Y predicted from a set of independent variables (X1, X2 ….Xk)
- One regression coefficient for each independent variable
- R^2: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

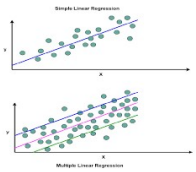
Multiple Regression Analysis (MRA)

- Method for studying the relationship between a dependent variable and two or more independent variables.
- Purposes:
  - Prediction
  - Explanation
  - Theory building



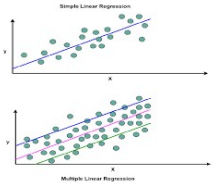
Operation?

- Uses the ordinary least squares solution (as does simple linear or bi-variable regression)
- Describes a line for which the (sum of squared) differences between the predicted and the actual values of the dependent variable are at a minimum.
- Represents the "function" that minimizes the sum of the squared errors.
- $Y_{pred} = a + a_1X_1 + a_2X_2 \ldots + a_nX_n$



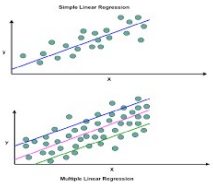
Operation?

- MLR produces a model that identifies the best weighted combination of independent variables to predict the dependent (or criterion) variable.
- **Ypred = $a + a_1X_1 + a_2X_2 \ldots + a_nX_n$**
  - MLR estimates the relative importance of several hypothesized predictors.
  - MLR assess the contribution of the combined variables to change the dependent variable.



Design Requirements

- One dependent variable (criterion)
- Two or more independent variables (predictor variables).
- Sample size: >= 50 (at least 10 times as many cases as independent variables)



Intuitive Method

For the one variable case, the calculation of b and a was:

$$a_1 = \frac{\sum xy}{\sum x^2}$$

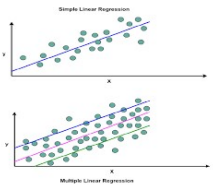
$$a = \bar{Y} - b\bar{X}$$

For the two variable case:

$$a_1 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

and

$$a_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$



Matrix Formulation

Example: Multiple Linear Regression

Find the Multiple Linear Regression model for the following given data points.

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

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

Example: Multiple Linear Regression

Find the Multiple Linear Regression model for the following given data points.

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

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						

Example: Multiple Linear Regression

Find the Multiple Linear Regression model for the following given data points.

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

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
Sum	38767	2823	101895	25364	9859
Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Example: Multiple Linear Regression

Find the Multiple Linear Regression model for the following given data points.

*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

Step 3: Calculate a_0, a_1, and a_2.

The formula to calculate a_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, $\mathbf{a_1}$ = [(194.875)(1162.5) – (-200.375)(-953.5)] / [(263.875) (194.875) – (-200.375)^2] = **3.148**

The formula to calculate a_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, $\mathbf{a_2}$ = [(263.875)(-953.5) – (-200.375)(1152.5)] / [(263.875) (194.875) – (-200.375)^2] = **-1.656**

The formula to calculate a_0 is: $y - a_1X_1 - a_2X_2$
Thus, a_0 = 181.5 – 3.148(69.375) – (-1.656)(18.125) = **-6.867**

Example: Multiple Linear Regression

Find the Multiple Linear Regression model for the following given data points.

*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

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

The estimated linear regression equation is: $\hat{y} = a_0 + a_1{*}x_1 + a_2{*}x_2$
In our example, it is $\hat{\mathbf{y}} = \mathbf{-6.867 + 3.148x_1 - 1.656x_2}$

How to Interpret a Multiple Linear Regression Equation

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

a_0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

a_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.

a_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.

Example 1: Pastry Sweetness Data- A designed experiment is done to assess how moisture content and sweetness of a pastry product affect a taster's rating of the product (pastry.txt). In a designed experiment, the eight possible combinations of four moisture levels and two sweetness levels are studied. Two pastries are prepared and rated for each of the eight combinations, so the total sample size is $n = 16$. The y-variable is the rating of the pastry. The two x-variables are moisture and sweetness. The values (and sample sizes) of the x-variables were designed so that the x-variables were not correlated.

Rating	Moisture	Sweetness
64	4	2
73	4	4
61	4	2
76	4	4
72	6	2
80	6	4
71	6	2
83	6	4
83	8	2
89	8	4
86	8	2
93	8	4
88	10	2
95	10	4
94	10	2
100	10	4

There is a linear relationship between rating and moisture and there is also a sweetness difference. The results given in the following output are for three different regressions - separate simple regressions for each x-variable and a multiple regression that incorporates both x-variables.

Regression Analysis: Rating versus Moisture

```
Regression Equation

Rating = 50.77 + 4.425 Moisture
```

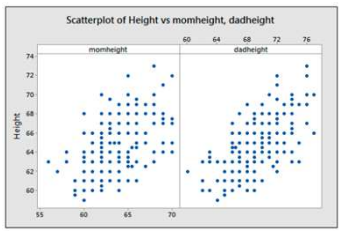
Regression Analysis: Rating versus Sweetness

```
Regression Equation

Rating = 68.63 + 4.38 Sweetness
```

Regression Analysis: Rating versus Moisture, Sweetness

```
Regression Equation

Rating = 37.65 + 4.425 Moisture + 4.375 Sweetness
```

Example 2: Female Stat Students
The data are from n = 214 females in statistics classes at the University of California at Davis (stat_females.txt). The variables are y = student's self-reported height, x1 = student's guess at her mother's height, and x2 = student's guess at her father's height. All heights are in inches. The scatterplots below are of each student's height versus mother's height and student's height against father's height.




```
Regression Equation

Height = 18.55 + 0.3035 momheight + 0.3879 dadheight
```

Example 2: Female Stat Students

The data are from n = 214 females in statistics classes at the University of California at Davis (stat_females.txt). The variables are y = student's self-reported height, x1 = student's guess at her mother's height, and x2 = student's guess at her father's height. All heights are in inches. The scatterplots below are of each student's height versus mother's height and student's height against father's height.

Interpretations

•The sample multiple regression equation is predicted student height = 18.55 + 0.3035 × mother's height + 0.3879 × father's height. To use this equation for prediction, we substitute specified values for the two parents' heights.

•We can interpret the "slopes" in the same way that we do for a simple linear regression model but we have to add the constraint that values of other variables remain constant. For example:

- When father's height is held constant, the average student height increases 0.3035 inches for each one-inch increase in mother's height.
- When mother's height is held constant, the average student height increases 0.3879 inches for each one-inch increase in father's height.

https://www.ques10.com/p/66524/multiple-linear-regression-problem-1/