Linear Regression

Dr. Muhammed Jamshed Alam Patwary

Assistant Professor, CSE

International Islamic University Chittagong

Topics

- **Introduction to the concept of Regression.**
- **Manually assumption of regression.**
- **Mathematical analysis of single variable regression with a numerical example.**
- **Regression goodness of fit (R^2 value)**
- **Mathematical analysis of multiple variable regression with a numerical example.**

What is Regression?

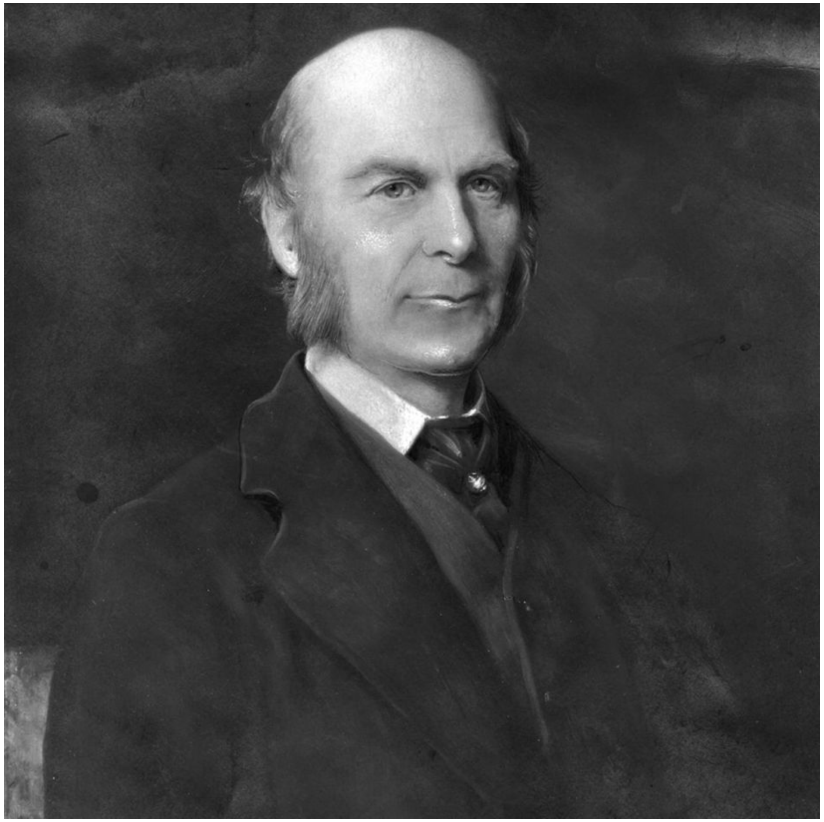
The probable movement of one variable in terms of the other variables is called regression.

In other words the statistical technique by which we can estimate the unknown value of one variable (dependent) from the known value of another variable is called regression.

Example: The productions of paddy of amount y is dependent on rainfall of amount x. Here x is independent variable and y is dependent variable.

Historical Note

- The term "regression" was used by a famous Biometrician Sir. F. Galton (1822-1911) in 1877.
- To explain the relationship between the heights of fathers and their sons.



Linear Regression with single variable

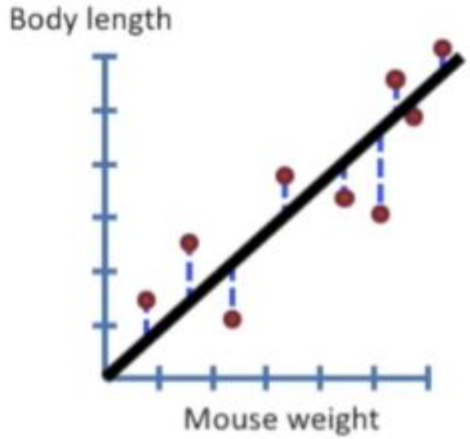
Regression analysis is used to predict the value of one variable (the dependent variable) on the basis of other variables (the independent variables).

<u>The linear model will be :</u>

Y = mx + c ;

where,

- **Y = Dependent Variable**
- **m = Regression slope**
- **x = Independent Variable**
- **c = Intercept Value**



Simple regression

y = y-intercept + slope *x*

Derivation of equation

The equation of the line of dependence of y on variable x is:

$$y = a + bx$$

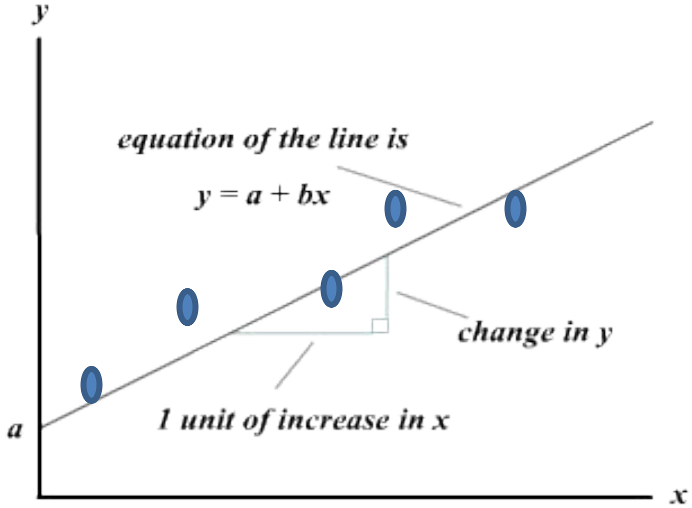
Here y is the Dependent variable and x is the Independent variable. And a and b are two constants. "a" is the distance from the point where the line crosses the y-axis to the point of origin. "b" is the slope of the line. This is called the coefficient of regression of y on the variable.

This indicates the change in the y variable for every single change in the x variable.

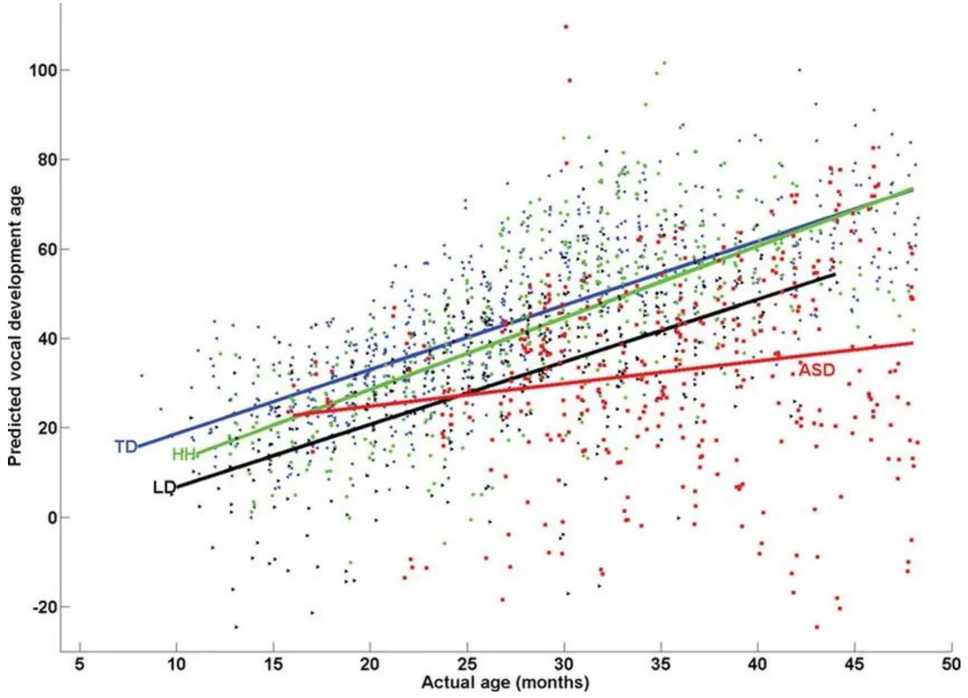
Constants a and b are two unknown parameters. Once their value is determined, the dependence line will be determined. The equation of the dependence line is determined

$$Y = a + bx \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ (1)$$

Below is a diagram of the dependency line.



Different dependence lines are available for different values of a and b. So, the values of a and b are calculated based on the observed values of the two variables using the least square method.



Suppose $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ are the observed values of the n pairs of variables x and y. The principle of the least square method is to calculate the values of constant a and b in such a way that the sum of the squares of the deviation of the observed value y = a + bx + e from Y is the minimum. That is, the values of a and b have to be determined in such a way that

$$S = \Sigma e_i^2 = \Sigma(y_i - Y_i)^2 = \Sigma(y_i - a - bx_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Here y_i and Y_i are the i-th observed and determined values of the variable y, respectively. That is

$$y_i = a + bx_i + e_i \quad \text{and} \quad Y_i = a + bx_i$$

Here the observed values of the variables x and y are specified. So different values of S are found for different values of a and b. That is, s is a function of a and b.

Now, according to Calculus's principle of largest and smallest values, the value of S will be minimal only when the value of the partial differentiation of S, relative to a and b, is zero.

That means,

$$\frac{\delta s}{\delta a} = 0 \text{ and } \frac{\vartheta s}{\vartheta b} = 0$$

We get,

$$\frac{\vartheta s}{\vartheta a} = \frac{\vartheta}{\vartheta a}(\Sigma(y_i\text{-}a\text{-}bx_i)^2 = 0$$

$$\Rightarrow\ 2\Sigma(y_i\text{-}a\text{-}bx_i)(-1) = 0$$

$$\Rightarrow \Sigma(y_i\text{-}a\text{-}bx_i) = 0$$

$$\Rightarrow \Sigma y_i = na + b\Sigma x_i\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Now divide the equation number 3 by n on both sides

$$\frac{\Sigma y_i}{n} = \frac{na}{n} + \frac{b\Sigma x_i}{n}\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

$$\Rightarrow \overline{y} = a + b\overline{x}$$

$$\therefore\ a = \overline{y} - b\overline{x}$$

Again,

$$\frac{\vartheta s}{\vartheta b} = \frac{\vartheta}{\vartheta b}[\Sigma(y_i\text{-}a\text{-}bx_i)^2] = 0$$

$$\Rightarrow 2\Sigma(y_i\text{-}a\text{-}bx_i)(\text{-}x_i) = 0$$

$$\Rightarrow \Sigma(y_ix_i\text{-}ax_i\text{-}bx_i^2) = 0$$

$$\Rightarrow \Sigma y_i\, x_i = a\Sigma x_i + b\Sigma x_i^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Here, $\underline{x}$ and $\underline{y}$ are the mean of x and y respectively.

Putting the value of a = $\bar{y}$ - b$\bar{x}$ in equation (4),

$$\Sigma y_ix_i = (\underline{y} - b\underline{x})\Sigma x_i + b\Sigma x_i^2$$

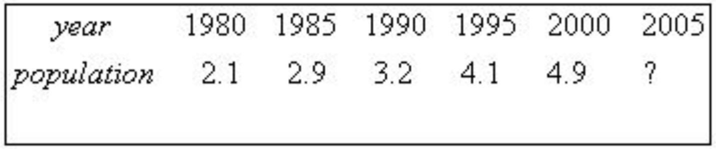
$$\therefore b = \frac{\Sigma(x_i-\underline{x})(y_i-\underline{y})}{\Sigma(x_i-\underline{x})^2}$$

Here , a & b are regression coefficients.

Regression Model For Prediction (Single Variable)

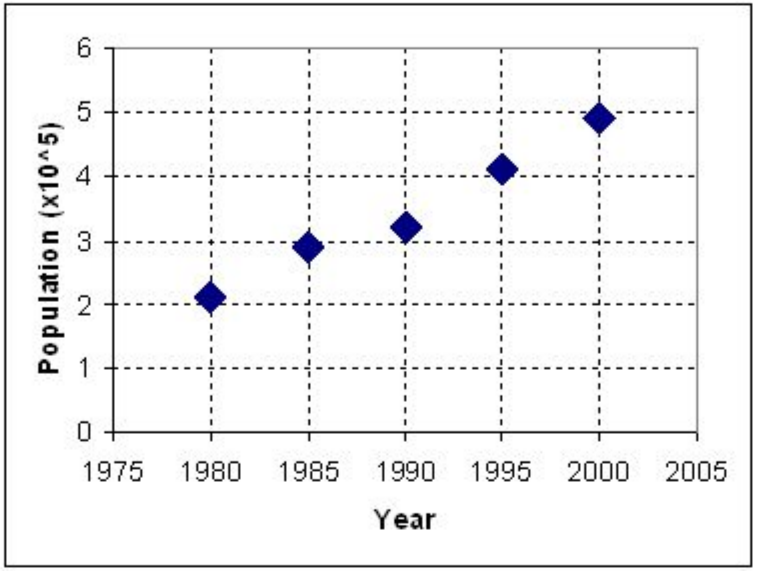
Suppose you have data of population (in hundred thousand) of a medium size city over 20 years (based on every 5-year census) as shown table 1. You want to predict the population in the year 2005.

Table 1 data for regression analysis



year	1980	1985	1990	1995	2000	2005
population	2.1	2.9	3.2	4.1	4.9	?

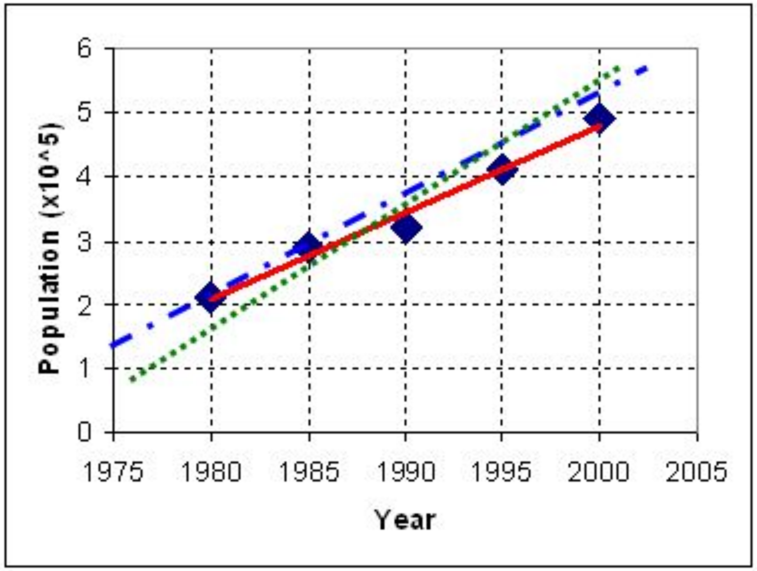
Let us plot the data above into a graph. One point in the graph represents data of one year. Since we have 5 data, thus we have five points.



Regression Model For Prediction

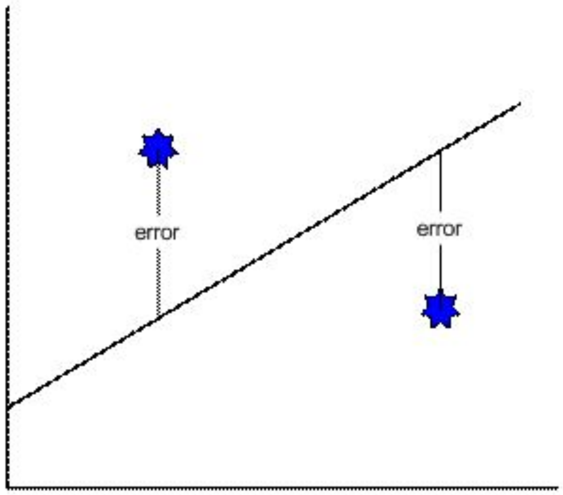
We have several proposals here. I will plot three seems best proposal:

- **Blue line (slash dot line)**
- **Red line (solid line)**
- **Green line (small dot line)**



Regression Model For Prediction

The diagram below is showing how we measure the error. When the point data is above the line model, we say that the error is positive, while if the line model is above the data, we say the error is negative.



Square error = $(Y_i - \hat{y}_i)^2$
Y_i = observe value
$\hat{y}_i$ = calculated value

It does not work because some error is positive and some of the error is negative. The sum of error may be zero. If we sum all the error, we may get many lines.

When we square the error, regardless it is positive or negative, the number become positive.

Sum Square Error

Year (y)	Population (X)	Blue Line (sq error)	Red Line(sq error)	Green Line(sq error)
1980	2.1	2.10(0.00)	2.08(0.00)	1.66(0.19)
1985	2.9	2.92(0.0004)	2.76(0.02)	2.60(0.09)
1990	3.2	3.70(0.25)	3.44(0.06)	3.54(0.12)
1995	4.1	4.50(0.16)	4.12(0.00)	4.48(0.14)
2000	4.9	5.30(0.16)	4.80(0.01)	5.42(0.27)
	Σ(SSE) =	0.57	0.09	0.81

We may obtain that the red line give the minimum sum of square error (=0.09) among the three proposals.

Numerical Example

The best line model can be computed using formula of linear regression.

Y= mx +c ;

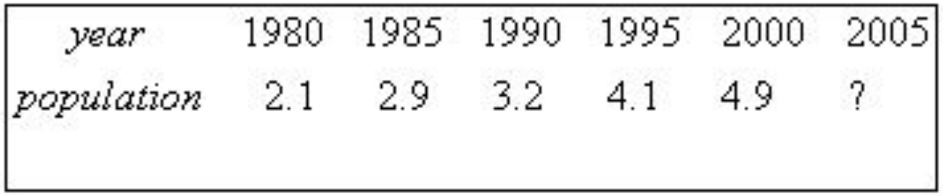
where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis),m is the slope of the line and c is the y-intercept.

$$slope = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$intercept = \bar{y} - slope \cdot \bar{x}$$

Regression Numerical Example

Suppose we have the following 5 data points and we want to predict the population data for the year 2005 using linear regression model.



year	1980	1985	1990	1995	2000	2005
population	2.1	2.9	3.2	4.1	4.9	?

In this section, we will use hand calculation using linear regression formula.

Regression Numerical Example

Year (x)	Population (y)	xy	x^2
1980	2.1	4158	3920400
1985	2.9	5756.5	3940225
1990	3.2	6368	3960100
1995	4.1	8179.5	3980025
2000	4.9	9800	4000000
Σ(x) = 9950	Σ(y) = 17.2	Σ(xy) = 34262	Σ(x^2) = 19800750

Here , $\overline{x} = 1990$ and $\overline{y} = 3.44$ **Total ,n =5**

$$slope = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2} = \frac{5 \cdot 34262 - 9950 \cdot 17.2}{5 \cdot 19800750 - (9950)^2} = 0.136$$

$$intercept = \overline{y} - slope \cdot \overline{x} = 3.44 - 0.136 \cdot 1990 = -267.2$$

Thus, we get the regression line

$Population = 0.136 * Year - 267.2$ **[y = mx + c]**

Regression Numerical Example

Using this regression line, we can predict the number of Population in the city for year 2005.

Population = 0.136 * year – 267.2 = (0.136*2005) – 267.2 = 5.48

So , 5.48 Million is the prediction population for the year 2005.

Regression Goodness Of Fit

Suppose you have a regression formula y = mx+c; as the best line model.

How fit is the data to our model?

There are unlimited numbers of model combination aside from linear model. Our data may be represented by curvilinear or non- linear model.

Most common indices are:

- **R-squared, or coefficient of determination**
- **Adjusted R-squared**
- **Standard Error**
- **F statistics**
- **t statistics**

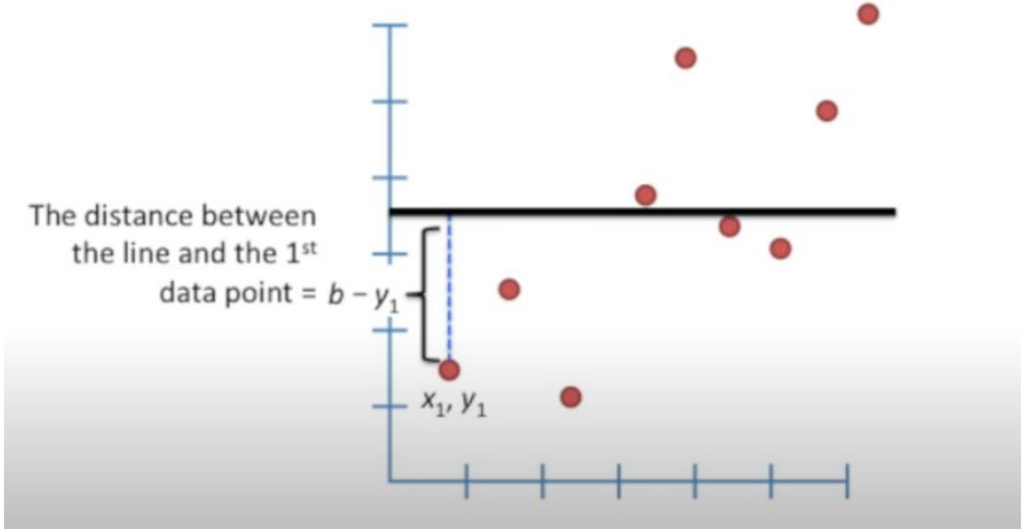
Regression Goodness Of Fit (R^2)

It is the ratio of sum of square error (SSE) from the regression model and the sum of squares difference around the mean (SST = sum of square total)

$$R^2 = 1 - \frac{SSE}{SST}$$

where $SSE = \sum_i (y_i - \hat{y}_i)^2$ and $SST = \sum_i (y_i - \bar{y})^2$.

Where y_i = observant y and $\hat{y}_i$ = calculated y from equation ,$\bar{y}$ = mean of y


The distance between the line and the 1st data point = $b - y_1$
x_1, y_1

R^2 Calculation

Y_i					
2.1	(0.136* 1980) – 267.2 = 2.08	0.02	0.0004	-1.34	1.79
2.9	(0.136* 1985) – 267.2 = 2.76	0.14	0.0196	-0.54	0.29
3.2	(0.136* 1990) – 267.2 = 3.44	-0.24	0.0576	-0.24	0.0576
4.1	(0.136* 1995) – 267.2 = 4.12	-0.02	0.0004	0.66	0.435
4.9	(0.136* 2000) – 267.2 = 4.8	0.1	0.01	1.46	2.13

Now , putting the values of SSE and SST on R^2 equation ,

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.088}{4.71} = 0.981 \text{ (98\%)}$$

Linear Regression With Multiple Variable

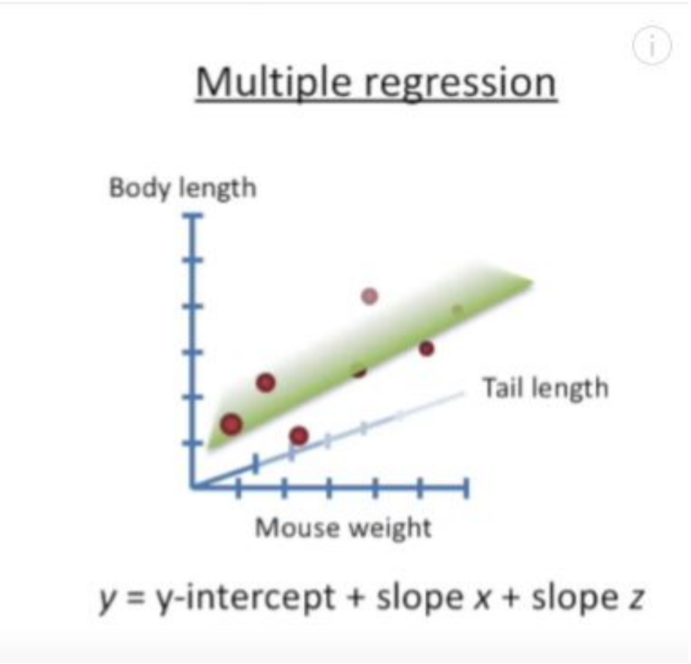
It measure that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variable known as independent variable.

$Y = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots b_nx_n$;

Where , a = intercept = $\bar{y} - b_1\overline{x_1} - b_2\overline{x_2}$

$x_1, x_2 .. x_n$ = multiple variable

$b_1 , b_2 ..b_n$ = correspondence slope



y = y-intercept + slope x + slope z

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

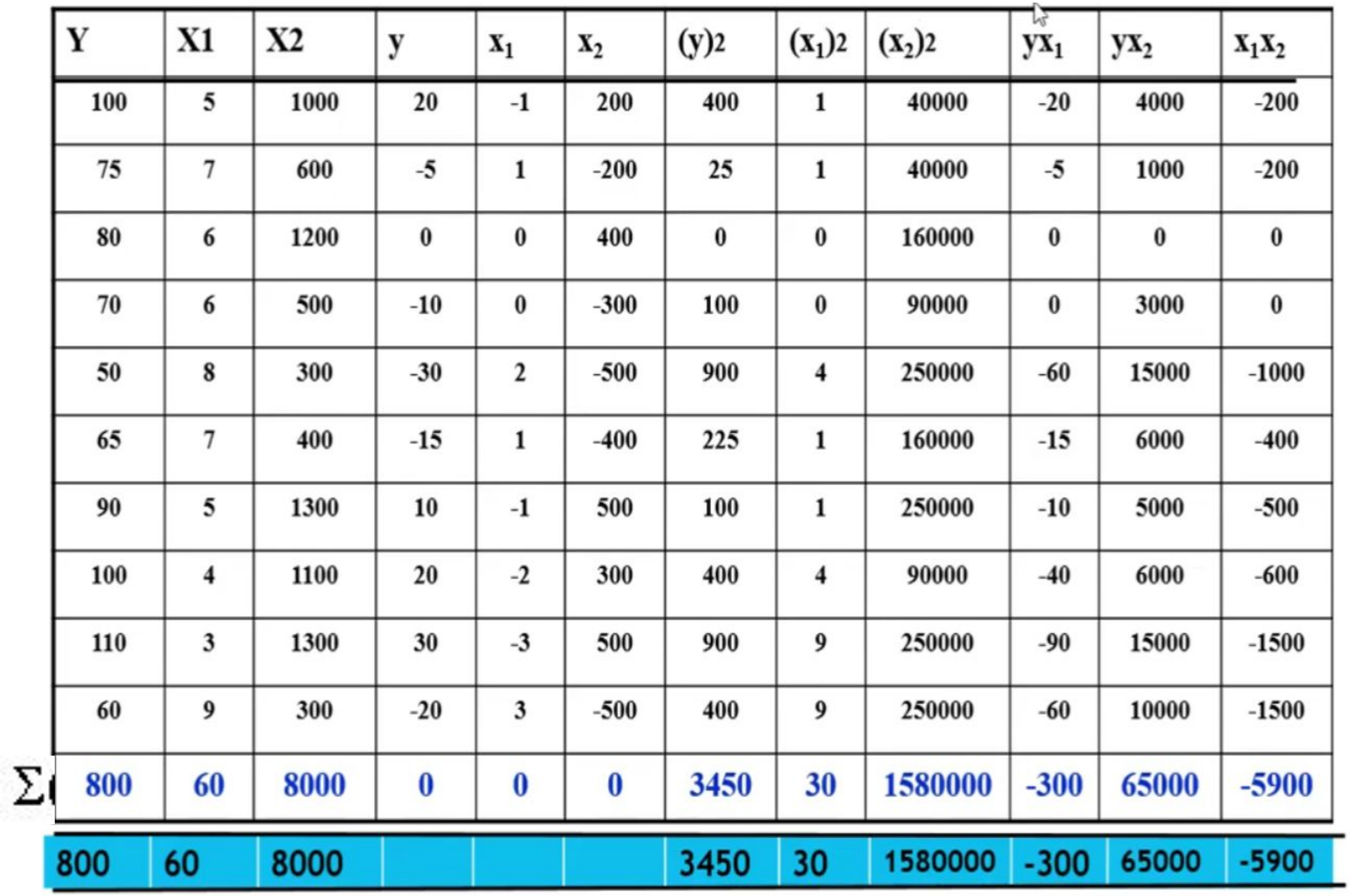
$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Numerical Example

A given example where the quantity demand depends on the price and income of the product.

Quantity Demand (Y)	Price (X_1)	Income (X_2)
100	5	1000
75	7	600
80	6	1200
70	6	500
50	8	300
65	7	400
90	5	1300
100	4	1100
110	3	1300
60	9	300
?	6	600

Calculation



	Y	X1	X2	y	x_1	x_2	(y)2	$(x_1)2$	$(x_2)2$	yx_1	yx_2	x_1x_2
	100	5	1000	20	-1	200	400	1	40000	-20	4000	-200
	75	7	600	-5	1	-200	25	1	40000	-5	1000	-200
	80	6	1200	0	0	400	0	0	160000	0	0	0
	70	6	500	-10	0	-300	100	0	90000	0	3000	0
	50	8	300	-30	2	-500	900	4	250000	-60	15000	-1000
	65	7	400	-15	1	-400	225	1	160000	-15	6000	-400
	90	5	1300	10	-1	500	100	1	250000	-10	5000	-500
	100	4	1100	20	-2	300	400	4	90000	-40	6000	-600
	110	3	1300	30	-3	500	900	9	250000	-90	15000	-1500
	60	9	300	-20	3	-500	400	9	250000	-60	10000	-1500
Σ(	800	60	8000	0	0	0	3450	30	1580000	-300	65000	-5900
	800	60	8000				3450	30	1580000	-300	65000	-5900

Here,
Y= Qd.

X1=Price.

X2=Income.

$y=Y-\overline{Y}$

$x_1=X_1-\overline{X_1}$

$x_2=X_2-\overline{X_2}$

$\overline{Y} = \Sigma Y/n$
$= 800/10$
$= 80$

$\overline{X1} = \Sigma Y/n$
$= 60/10$
$= 6$

$\overline{X2} = \Sigma Y/n$
$= 8000/10$
$= 800$

Calculation

Now,

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} \qquad = \frac{1580000*(-300)-(-5900)*65000}{30*1580000-(-5900)^2} = -7.18$$

b1 = - 7.18 ;

b2 = 0.014 ;

& a = 111.8 ;

So estimated regression line will be –

Y = 111.8 – 7.18 * x_1+ 0.014* x_2

where , x_1 = price , x_2 = income and Y =Quantity Demand.

We want to predict the quantity demand for a product price 6 with income 600 using linear regression model.

QD = 111.8 – (7.18 * 6) + (0.014 * 600) = 77.12

R^2 Calculation

Calculating R^2 is the same for both simple and multiple regression.

References

- (4) Linear Regression and Linear Models – YouTube
- Microsoft Excel Tutorials: Regression (revoledu.com)
- (4) What is Multiple Regression | numerical explanation AND interpretation of Multiple regression – YouTube

Thank You.