Chapter 9
Statistical Inference and The Relationship between two variables

Prepared By : Dr. Shuhrat Khan

REGRESSION
CORRELATION
ANALYSIS OF VARIANCE

EQUATION OF REGRESSION

- Regression, Correlation and Analysis of Covariance are all statistical techniques that use the idea that one variable say, may be related to one or more variables through an equation. Here we consider the relationship of two variables only in a linear form, which is called linear regression and linear correlation; or simple regression and correlation. The relationships between more than two variables, called multiple regression and correlation will be considered later.
- Simple regression uses the relationship between the two variables to obtain information about one variable by knowing the values of the other. The equation showing this type of relationship is called simple linear regression equation. The related method of correlation is used to measure how strong the relationship is between the two variables is.

٢٠١

Line of Regression

DEPENDENT VARIABLE

INDEPENDENT VARIABLE

TWO RANDOM VARIABLE

OR

BIVARIATE

RANDOM

VARIABLE

Simple Linear Regression:

Suppose that we are interested in a variable Y, but we want to know about its relationship to another variable X or we want to use X to predict (or estimate) the value of Y that might be obtained without actually measuring it, provided the relationship between the two can be expressed by a line.' X' is usually called the **independent variable** and 'Y' is called the **dependent variable.**

We assume that the values of variable X are either fixed or random. By fixed, we mean that the values are chosen by researcher--- either an experimental unit (patient) is given this value of X (such as the dosage of drug or a unit (patient) is chosen which is known to have this value of X.

By random, we mean that units (patients) are chosen at random from all the possible units,, and both variables X and Y are measured.

We also assume that for each value of x of X, there is a whole range or population of possible Y values and that the mean of the Y population at X = x, denoted by $\mu_{y/x}$, is a linear function of x. That is,

$$\mu_{y/x} = \alpha + \beta x$$

ESTIMATION

We select a sample of n observations (x_i,y_i) from the population, WITH the goals

- Estimate α and β.
- Predict the value of Y at a given value x of X.
- Make tests to draw conclusions about the model and its usefulness.

- We estimate the parameters α and β by 'a' and 'b' respectively by using sample regression line:
- $\hat{Y} = a + bx$
- Where we calculate
-

ESTIMATION AND CALCULATION OF CONSTANTS , "a" AND "b"

$$a = \bar{y} - b\bar{x}$$

$$B = \frac{(\sum xiyi - n\bar{x}\,\bar{y})}{(\sum x_i{}^2 - n\bar{x}^2)}$$

EXAMPLE

- investigators at a sports health centre are interested in the relationship between oxygen consumption and exercise time in athletes recovering from injury. Appropriate mechanics for exercising and measuring oxygen consumption are set up, and the results are presented below:

x variable –

exercise time (min)	y variable oxygen consumption
0.5	620
1.0	630
1.5	800
2.0	840
2.5	840
3.0	870
3.5	1010
4.0	940
4.5	950
5.0	1130

calculations

- $\overline{x} = 2.75$ $\bar{y} = 863$ $N = 10$

  $\Sigma x = 27.5$ $\Sigma y = 8630$ $(\sum x)^2 = 756.25$ $(\Sigma y)^2 = 74476900$ $\Sigma xy = 25750$

  $\Sigma x^2 = 96.25$ $\Sigma y^2 = 7672500$

$$b = \frac{(25750 - 10x2.75x863)}{(96.25 - 10x2.75^2)} = 97.82$$

$a = \bar{y} - b\bar{x}$ or $a = 863 - (97.82\,X2.75) = 594$

$\hat{y}\ for\ given\ x = 2.8 = 594 + (97.82 \times 2.8) = 868$ units

Pearson's Correlation Coefficient

- With the aid of Pearson's correlation coefficient (r), we can determine the strength and the direction of the relationship between X and Y variables,
- both of which have been measured and they must be quantitative.
- For example, we might be interested in examining the association between height and weight for the following sample of eight children:

Height and weights of 8 children

Child	Height(inches)X	Weight(pounds)Y
A	49	81
B	50	88
C	53	87
D	55	99
E	60	91
F	55	89
G	60	95
H	50	90
Average	(= 54 inches)	(= 90 pounds)

Scatter plot for 8 babies

height	weight
49	81
50	88
53	
55	
60	
55	
60	
50	

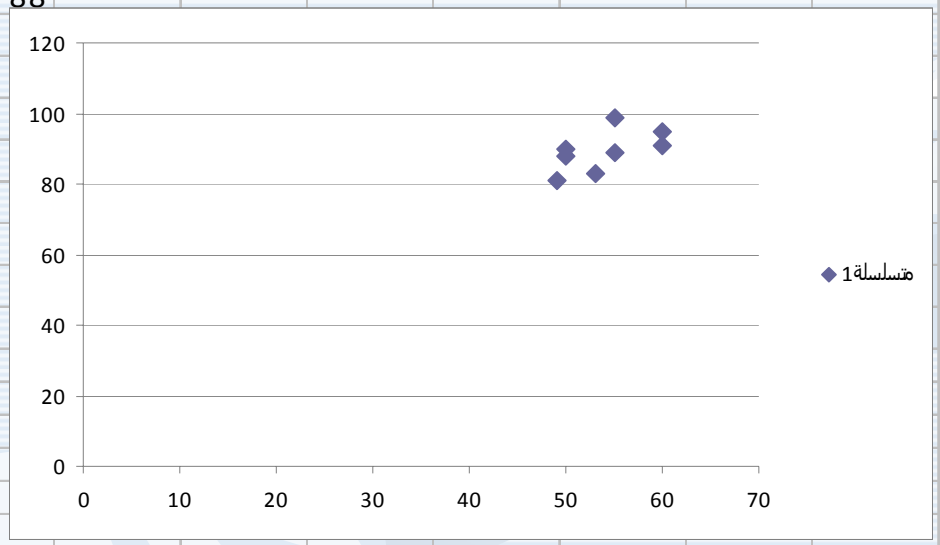

Table : The Strength of a Correlation

Value of r (positive or negative)	Meaning
0.00 to 0.19	A very weak correlation
0.20 to 0.39	A weak correlation
0.40 to 0.69	A modest correlation
0.70 to 0.89	A strong correlation
0.90 to 1.00	A very strong correlation

FORMULA FOR CORRELATION COEFFECIENT (r)

$$r = (\Sigma(X - \bar{x})(Y - \bar{y}\ )/\sqrt{(\Sigma(X - \bar{x})^2\ \Sigma(Y - \bar{y})^2\ )}$$

- With Pearson's *r*, $\Sigma(X - \bar{X})\ (Y - \bar{Y})$
- means that we add the products of the deviations to see if the positive products or negative products are more abundant and sizable. Positive products indicate cases in which the variables go in the same direction (that is, both taller or heavier than average or both shorter and lighter than average);
- negative products indicate cases in which the variables go in opposite directions (that is, taller but lighter than average or shorter but heavier than average).
-

Computational Formula for Pearsons's Correlation Coefficient r

Where SP (sum of the product), SSx (Sum of the squares for x) and SSy (sum of the **squares for y) can be computed as follows:**

$$SP = \sum XY - N\bar{X}\bar{Y} = (\sum(X - \bar{x})(Y - \bar{y})$$

$$SSx = \sum X^2 - N\bar{X}^2 = (\sum(X - \bar{x})^2$$

$$SSy = \sum Y^2 - N\bar{Y}^2 = \sum(Y - \bar{y})^2$$

$$r = \frac{SP}{\sqrt{SSx \times SSy}} = \frac{(\sum(X - \bar{x})(Y - \bar{y})}{\sqrt{SSx \times SSy}}$$

$$= \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}}$$

$$r = \frac{\sum XY - N\bar{X}\bar{Y}}{\sqrt{(\sum X^2 - N\bar{X}^2)(\sum Y^2 - N\bar{Y}^2)}} = \frac{981 - 8(10.5)(11.5)}{\sqrt{[946 - (10.5)^2][1118 - (11.5)^2]}}$$

$$\frac{15}{\sqrt{(64)(60)}} = (15)/(61.97) = +.24$$

Child	X	Y	X^2	Y^2	XY
A	12	144	12	144	144
B	10	8	100	64	80
C	6	12	36	144	72
D	16	11	256	121	176
E	8	10	100	64	80
F	9	8	81	64	72
G	12	16	144	256	192
H	11	15	121	225	165
	1118	**946**	**92**	**84**	∑

Table 2 : Chest circumference and Birth Weight of 10 babies

X(cm)	y(kg)	x^2	y^2	xy
22.4	2.00	501.76	4.00	44.8
27.5	2.25	756.25	5.06	61.88
28.5	2.10	812.25	4.41	59.85
28.5	2.35	812.25	5.52	66.98
29.4	2.45	864.36	6.00	72.03
29.4	2.50	864.36	6.25	73.5
30.5	2.80	930.25	7.84	85.4
32.0	2.80	1024.0	7.84	89.6
31.4	2.55	985.96	6.50	80.07
32.5	3.00	1056.25	9.00	97.5
TOTAL				
292.1	24.8	8607.69	62.42	731.61

Checking for significance

$$r = \frac{71.92}{\sqrt{754.45 \times 9.16}} = 0.86$$

- There appears to be a strong between chest circumference and birth weight in babies.
- We need to check that such a correlation is unlikely to have arisen by in a sample of ten babies.
- Tables are available that gives the significant values of this correlation ratio at two probability levels.
- First we need to work out degrees of freedom. They are the number of pair of observations less two, that is (n – 2)= 8.
- Looking at the table we find that our calculated value of 0.86 exceeds the tabulated value at 8 df of 0.765 at p= 0.01. Our correlation is therefore statistically highly significant.