Example 6.8.1.

Work out the simple linear regression of O_2 consumption (Y ml/minute) on tracheal ventilation (X ml/minute) using the data of Example 6.3.1.

Solution :

1. *Computation of b_{YX} from sum of products :*

From the data of Example 6.3.1, the scores of tracheal ventilation (X) and O_2 consumption (Y) are entered in Table 6.5 for further treatment.

Table 6.5. Computation of sum of products and sum of squares.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
66.0	3.3	– 12.5	– 0.7	156.25	+ 8.75
89.1	4.9	+ 10.6	+ 0.9	112.36	+ 9.54
72.0	3.5	– 6.5	– 0.5	42.25	+ 3.25
87.5	4.7	+ 9.0	+ 0.7	81.00	+ 6.30
75.2	3.7	– 3.3	– 0.3	10.89	+ 0.99
78.2	4.0	– 0.3	0	0.09	0
83.5	4.3	+ 5.0	+ 0.3	25.00	+ 1.50
71.6	3.4	– 6.9	– 0.6	47.61	+ 4.14
85.6	4.4	+ 7.1	+ 0.4	50.41	+ 2.84
76.3	3.8	– 2.2	– 0.2	4.84	+ 0.44
Σ 785.0	40.0	—	—	530.70	+ 37.75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{785.0}{10} = 78.5. \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40.0}{10} = 4.0.$$

$$b_{YX} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{37.75}{530.70} = 0.071.$$

2. *Computation of b_{YX} from raw scores (alternative method) :*

From the data of Example 6.3.1, X and Y scores are entered in Table 6.6 for further treatment.

scores. (ii) Scores of each variable should be distributed in the population in a unimodal, bilaterally symmetric or almost symmetric, *normal or nearly normal distribution* with not much skewness of its tails. (iii) There should exist a *linear association* between the variations of the two variables. (iv) The pair of scores of the two variables for each individual or case should have occurred in the sample *at random*, obeying the laws of probability and *independent* of all other similar pairs of scores ; this last assumption ensures that the sample may be a representative of the population, enabling the inference made from the sample to be generalized for the corresponding population.

It follows from these assumptions that the product moment r cannot be used in correlating such variables as are not associated linearly, or are discontinuous in nature (e.g., heart rate, cell count and litter size) or are ordinal variables (e.g., ferocity) or nonmeasurable qualitative variables (e.g., sex and race), or have prominently skewed or non-normal distributions in the population.

6.3.2. Properties of product-moment r

(a) The magnitude of the computed r is a measure of the *strength of association* between the variables while its algebraic sign indicates whether the variables vary in the same direction *(positive)* or in opposite directions *(negative)*. Thus, + 0.80 indicates a high positive correlation, – 0.72 shows a high negative correlation, +0.14 indicates a low positive correlation, while 0.00 means the absence of any linear correlation.

(b) If every score of any or each variable is multiplied, divided, added or subtracted by a constant number, it does not result in any change in the r value between the two variables.

(c) Correlation depends on that proportion of total variance of each variable which is associated with the variance of the other. This makes the value of r directly proportional to the *covariance* of the two variables. Where X and Y are the scores, $\bar{X}$ and $\bar{Y}$ are the means, s_x and s_y are the unbiased standard deviations, and $Cov(X,Y)$ is the covariance of two variables, and n is the sample size,

$$s_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}};\ s_y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n-1}};\ Cov(X,Y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1};$$

$$r = \frac{Cov(X,Y)}{s_x s_y} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)\,s_x s_y}.$$

(d) The r values, computed between two given variables in different samples from the same population, lie dispersed forming a *sampling distribution* of r around the population correlation coefficient (ρ) because of their varying differences, called

As the computed t exceeds the critical $t_{0.05}$, P is too low ($P < 0.05$). The H_O is rejected and the computed $R_{1.23}$ is significant.

6.7 Regression

Some variables cannot be measured directly or easily, with sufficient precision, or without errors. For predicting the very likely score of such a variable in any given individual or case, a statistical method of prediction, called *regression*, may be applied. The latter depends on the already known or measured scores of one or more variables correlated significantly with the variable to be predicted. In any regression, the variable to be predicted is called the dependent variable or *criterion* for that regression ; the variable(s) whose known or measured score(s) may form the basis of the prediction, should be called the independent variable(s) or *predictor(s)*. In every regression, there is a single criterion ; but there may be one or more predictors in a regression.

6.7.1. Types of regression

Regression may be broadly classified into simple and multiple regressions depending on the *number of predictors* used. In all types of regression, scores of a single variable would be predicted; but where the known or measured score of a single predictor is used in working out the regression, the latter is called a *simple regression,* while the scores of more than one predictor are used in predicting the score of a criterion in a *multiple regression.* It should be understood that in predicting the score of a criterion in any individual, the predictor scores of the same individual must be used ; moreover, there must exist significant simple correlations between the scores of the criterion and those of each predictor. An example of simple regression is the regression of the blood insulin concentraion (X_1) in a patient on his/her blood sugar concentraion (X_2), X_1 and X_2 being respectively the criterion and the predictor. A multiple regression of the O_2 consumption (X_1) may be worked out in a locust on the combination of scores of its tracheal ventilation (X_2) and the O_2 tension (X_3) in its inspired air— X_1 is the criterion here while X_2 and X_3 are two predictors.

Regressions may again be classified into *linear and nonlinear regressions,* depending on whether there is respectively a linear association or a nonlinear association between the criterion and the predictor. For example, if the criterion is linearly correlated with the predictor, the scores of the former are predicted by working out an equation for a straight line, depending on the linear association between the two. On the contrary, if the criterion has a nonlinear correlation with the predictor, scores of the criterion have to be predicted in terms of a curved line like a sigmoid or hyperbolic or exponential curve, according to the form of their association.

moment *r*;

- compute Kendall's *tau* between two variables and find its significance,
- know when and how partial and multiple linear correaltions are worked out and their significances are found out,
- understand what is meant by regression and know its different types and models,
- describe the assumptions and properties of simple linear regression,
- work out simple linear regression for predicting the score of one variable on the measured score of another, and
- know when and how to work out multiple linear regression of one variable on the combination of observed scores of two or more other variables.

6.2 Correlation

Correlation is the quantitative estimation and numerical expression of the magnitude or strength as well as the algebraic sign or direction of the association between two or more variables in a system. The correlation coefficient serves basically as a measure of the intensity or degree of association between the variances (Sub-sections 2.6.3 and 5.2.1) of two or more variables in the cases of the sample, while its algebraic sign is the indicator of whether those variables vary in the same direction or in opposite directions. The correlation coefficient is the sample statistic for correlation and ranges from −1.00 to +1.00 in value.

6.2.1. Types of correlation

Correlation may be *simple* or *multiple,* according as it is computed between two variables or more than two variables. For example, there may be a *simple correlation* between trunk length and wing length in a sample of cockroaches ; on the contrary, there may be a *multiple correlation* between oxygen consumption and the combination of atmospheric oxygen tension and tracheal ventilation volume in a sample of locusts.

Correlation may again be *linear* and *nonlinear,* according as the relation between the variables conforms to a straight line equation and a linear graph, or follows the equation of a curved line and a nonlinear graph. For example, there may be a *simple linear correlation* between body weight and gill weight in a sample of fishes ; but the initial velocity of an enzyme action and the corresponding substrate concentration may have a *simple nonlinear correlation,* conforming to a rectangular hyperbola.

Correlation may also be either *positive* or *negative*. If high scores of a variable are mostly accompanied by high scores of another variable while low scroes of one are usually associated with low scores of the other, the two variables are varying in the same direction and are said to bear a *positive correlation* with each other ; an example is the positive correlation between body height and body weight in many

(g) Between a pair of variables correlated with one another, regression can be worked out in *two ways*, viz., a regression of variable X as criterion on variable Y as predictor, and another regression of variable Y as criterion on variable X as predictor. However, it is sensible to compute the regression of that variable of the pair as criterion, whose direct measurement is less precise or/and more complicated than that of the other variable, while the latter is used as the predictor.

(h) A *regression equation* is worked out using a statistic called the *regression coefficient* (e.g., b_{YX} for the regression of Y on X) which is a measure of the average rate of change of criterion scores (Y) with each unit change in the scores (X) of the predictor.

6.8 Simple linear regression

This is the regression for predicting the probable score of a criterion (say, Y) on the measured, known or given score of a *single predictor* (say, X) where the two variables are *linearly correlated*. It may belong to *either model I or model II*, depending on whether the predictor is a fixed experimental treatment free from random errors, or a classification variable suffering from random errors. The regression of the criterion Y on the predictor X and that of the criterion X on the predictor Y require the working out of *two* respective and separate regression equations. For linear regression, each *regression equation* is an equation for a straight line (*regression line*) expressing the scores of the criterion as the *linear function* of the scores of the predictor. The *slope* of the regression line is given by the *regression coefficient* which is the measure of the average rate of change of criterion scores for unit changes in predictor scores. The regression coefficient is used in working out the *y-intercept* of the regression line, i.e., its point of intersection with the ordinate scale for criterion scores.

6.8.1. Assumptions for simple linear regression

Following assumptions should be justifiable for working out a simple linear regression of a criterion (Y) on a single predictor (X). (i) The criterion as well as the predictor should be *continuous measurement variable*, scores of both being quantitatively measurable and occurring even in infinitely small fractions of units. (ii) Scores of each variable should form a *normal or near-normal distribution* in the population from which the sample has been drawn. (iii) There should exist a significant *linear corelation* between the criterion and the predictor. (iv) The actual score of criterion for each individual should occur in the sample obeying the *laws of probability*, thus ensuring the representative nature of the sample with respect to the population. (v) The actual criterion scores (Y) of a large number of cases, having an identical

score ($\hat{Y}$) may be estimated by the *SE of estimate* (s_{YX} for $\hat{Y}$ on X) using the *SD of* the criterion scores. Thus, for the respective regression equations,

$$s_{YX} = s_Y\sqrt{1-r_{YX}^2} \; ; \quad s_{XY} = s_X\sqrt{1-r_{XY.}^2} \; .$$

6.8.3. Computation of simple linear regression

1. *Regression of Y on X :*

(i) Computation of b_{YX} *using sum of products :*

$$\bar{X} = \frac{\Sigma X}{n}; \quad \bar{Y} = \frac{\Sigma Y}{n}; \quad b_{YX} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}.$$

(ii) Computation of b_{YX} *using raw scores :*

$$b_{YX} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}.$$

(iii) Computation of a_{YX} and *regression equation :*

$$a_{YX} = \bar{Y} - b_{YX}\bar{X}; \quad \hat{Y} = a_{YX} + b_{YX}X.$$

2. *Regression of X on Y :*

(i) Computation of b_{XY} *using sum of products :*

$$\bar{X} = \frac{\Sigma X}{n}; \quad \bar{Y} = \frac{\Sigma Y}{n}; \quad b_{XY} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(Y-\bar{Y})^2}.$$

(ii) Computation of b_{XY} *using raw scores :*

$$b_{XY} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}.$$

(iii) Computation of a_{XY} and *regression equation :*

$$a_{XY} = \bar{X} - b_{XY}\bar{Y}; \quad \hat{X} = a_{XY} + b_{XY}Y.$$

3. *Drawing of the regression line :*

Several predictor scores are chosen from within their range in the sample, for computing the corresponding criterion scores with the help of the regression equation. The computed criterion scores are plotted against the respective predictor scores on a graph paper and the plotted points are used in drawing the regression line. (See Example 6.8.1.)

samples. But if high scores of one variable are mostly accompanied by [illegible] scores of another variable while high scores of the latter are usually associated with low scores of the former, the scores of the variables are usually varying in opposite directions and the variables bear a *negative correlation* with each other ; for example, a negative correlation exists between blood sugar level and blood insulin level in animal samples.

While the magnitude of correlation is expressed numerically, ranging from −1.00 to +1.00, the algebraic +/− sign preceding the numerical value indicates whether the correlation is positive or negative.

6.2.2. Properties of correlations

General properties of correlations are summarized below. (i) A correlation coefficient worked out with a sample drawn from a population would hold good only within the limits of the particular stratum or class of the population from which the sample has been drawn, and would also be confined within other conditions and situations prevailing during the work. Thus, a correlation coefficient worked out with a sample of adults may not hold good for children of the same population, or that worked out with a sample of females may not apply to males. (ii) A correlation coefficient between two variables does not necessarily indicate that variations of one of them may be either the cause or the effect of variations of the other ; their correlation may very well have arisen from the association of some other variable in common with both of them. (iii) A correlation coefficient *cannot directly predict* the score of one of the variables from that of the other in the same individual. (iv) The correlation coefficient between two variables varies from sample to sample even when they have been drawn from the same population ; so, the sample correlation coefficients (r) lie dispersed around the population correlation coefficient (ρ) to form a *sampling distribution* of r values, owing to their respective *sampling errors*.

6.3 Product-moment correlation

Karl Pearson's *product-moment correlation coefficient* or Pearson's r is a *simple linear correlation coefficient*, used in correlating two variables which have a linear association with each other.

6.3.1. Assumptions for product-moment r

Product-moment r can be applied for correlating two variables, only if it can be *logically assumed* that the following conditions or criteria are fulfilled in the case under investigation. (i) Both the variables being correlated should be *continuous measurement variables*, with their scores quantitatively measurable and occurring even in infinitely small fractional units, with no gaps in the respective scales of

Table 6.6. Computation of b_{YX} from raw scores.

X	Y	X^2	XY
66.0	3.3	4356.00	217.80
89.1	4.9	7938.81	436.59
72.0	3.5	5184.00	252.00
87.5	4.7	7656.25	411.25
75.2	3.7	5655.04	278.24
78.2	4.0	6115.24	312.80
83.5	4.3	6972.25	359.05
71.6	3.4	5126.56	243.44
85.6	4.4	7327.36	376.64
76.3	3.8	5821.69	289.94
Σ 785.0	40.0	62153.20	3177.75

$$\bar{X} = \frac{\Sigma X}{n} = \frac{785.0}{10} = 78.5. \qquad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40.0}{10} = 4.0.$$

$$b_{YX} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{10 \times 3177.75 - 785.0 \times 40.0}{10 \times 62153.20 - 785.0^2} = 0.071.$$

3. *Computation of a_{YX} and regression equation :*

$$a_{YX} = \bar{Y} - b_{YX}\,\bar{X} = 4.0 - 0.071 \times 78.5 = -1.57.$$

$$\hat{Y} = a_{YX} + b_{YX} X, \quad or \quad \hat{Y} = -1.57 + 0.071X.$$

4. *Drawing of regression line :*

Four X scores are chosen at random from within the range of X in the data and used in computing the respective $\hat{Y}$ scores.

(i) Where $X = 70$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 70 = 3.4$.

(ii) Where $X = 75$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 75 = 3.8$.

(iii) Where $X = 80$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 80 = 4.1$.

(iv) Where $X = 85$, $\hat{Y} = a_{YX} + b_{YX} X = -1.57 + 0.071 \times 85 = 4.5$.

Each $\hat{Y}$ score is plotted against the corresponding X score on a graph paper and the plotted points are used to draw the regression line of Y on X (Fig 6.1).