Chapter 4
Correlation Matrix and Partial Correlation: Explaining Relationships

Learning Objectives

After completing this chapter, you should be able to do the following:

- Learn the concept of linear correlation and partial correlation.
- Explore the research situations in which partial correlation can be effectively used.
- Understand the procedure in testing the significance of product moment correlation and partial correlation.
- Develop the hypothesis to test the significance of correlation coefficient.
- Formulate research problems where correlation matrix and partial correlation can be used to draw effective conclusion.
- Learn the application of correlation matrix and partial correlation through case study discussed in this chapter.
- Understand the procedure of using SPSS in computing correlation matrix and partial correlation.
- Interpret the output of correlation matrix and partial correlation generated in SPSS.

Introduction

One of the thrust areas in the management research is to find the ways and means to improve productivity. It is therefore important to know the variables that affect it. Once these variables are identified, an effective strategy may be adopted by prioritizing it to enhance the productivity in the organization. For instance, if a company needs to improve the sale of a product, then its first priority would be to ensure its quality and then to improve other variables like resources available to the marketing team, their incentive criteria, and dealer's scheme. It is because of the fact that the product quality is the most important parameter in enhancing sale.

J.P. Verma, *Data Analysis in Management with SPSS Software*,
DOI 10.1007/978-81-322-0786-3_4, © Springer India 2013

To find how strongly a given variable is associated with the performance of an employee, an index known as product moment correlation coefficient "r" may be computed. The product moment correlation coefficient is also known as correlation coefficient, and it measures only linear relation between two variables.

When we have two variables that covary, there are two possibilities. First, the change in a thing is concomitant with the change in another, as the change in a child's age covaries with his weight, that is, the older, the heavier. When higher magnitude on one variable occurs along with higher magnitude on another and the lower magnitudes on both also occur simultaneously, then the things vary together positively, and we denote this situation as positive correlation.

In the second situation, two things vary inversely. In other words, the higher magnitudes of one variable go along with the lower magnitudes of the other and vice versa. This situation is denoted as negative correlation.

The higher magnitude of correlation coefficient simply indicates that there is more likelihood that if the value of one variable increases, the value of other variable also increases or decreases. However, correlation coefficient does not reveal the real relationship between the two variables until the effects of other variables are eliminated.

This fact can be well explained with the following example. John and Philip work for the same company. John has a big villa costing $540,000, whereas Philip owns a three-room apartment, costing $160,000. Which person has a greater salary?

Here, one can reasonably assume that it must be John who earns more, as he has a more expensive house. As he earns a larger salary, the chances are that he can afford a more expensive house. One cannot be absolutely certain; of course, it may be that John's villa was a gift from his father, or he could have gotten it in a contest or it might be a result of any legal settlement. However, most of the time, an expensive house means a larger salary.

In this case, one may conclude that there is a correlation between someone's salary and the cost of the house that he/she possesses. This means that as one figure changes, one can expect the other to change in a fairly regular way.

In order to be confident that the relationship exists between any two variables, it must be exhibited across some cases. A case is a component of variation in a thing. For example, different levels of IQ that go along with different marks obtained in the final examination may be perceived across students. If the correlation between IQ and marks of the students is positive, it indicates that a student with high IQ has high marks and the one with low IQ has low marks.

The correlation coefficient gives fair estimate of the extent of relationship between any two variables if the subjects are chosen at random. But in most of the situations, samples are purposive, and, therefore, correlation coefficient in general may not give the correct picture of the real relationship. For example, in finding correlation coefficient between the age of customers and quantity of moisturizer purchased, if the sample is collected from the high socioeconomic population, the result may not be valid as in this section of society, people understand the importance of the product and can afford to invest on it. However, to establish the relationship between sales and age of the users, one should collect the sample from all the socioeconomic status groups.

Even if the sample is random, it is not possible to find the real relationship between any two variables as it might be affected by other variables. For instance, if the correlation computed between height and weight of the children belonging to age category 12–18 years is 0.85, it may not be considered as the real relationship. Here all the subjects are in the developmental age, and in this age category, if the height increases, weight also increases; therefore, the relationship exhibited between height and weight is due to the impact of age as well. To know the real relationship between the height and weight, one must eliminate the effect of age. This can be done in two ways. First, all the subjects can be taken in the same age category, but it is not possible in the experimental situation once the data collection is over. Even if an experimenter tries to control the effect of one or two variables manually, it may not be possible to control the effect of other variables; otherwise one might end up with getting one or two samples only for the study.

In the second approach, the effects of independent variables are eliminated statistically by partialing out their effects by computing partial correlation. Partial correlation provides the relationship between any two variables after partialing out the effect of other independent variables.

Although the correlation coefficient may not give the clear picture of the real relationship between any two variables, it provides the inputs for computing partial and multiple correlations, and, therefore, in most of the studies, it is important to compute the correlation matrix among the variables. This chapter discusses the procedure for computing correlation matrix and partial correlation using SPSS.

Details of Correlation Matrix and Partial Correlation

Matrix is an arrangement of scores in rows and column, and if its elements are correlation coefficients, it is known as correlation matrix. Usually in correlation matrix, upper diagonal values of the matrix are written. For instance, the correlation matrix with the variables X_1, X_2, X_3, and X_4 may look like as follows:

	X_1	X_2	X_3	X_4
X_1	1	0.5	0.3	0.6
X_2		1	0.7	0.8
X_3			1	0.4
X_4				1

The lower diagonal values in the matrix are not written because of the fact that the correlation between X_2 and X_4 is same as the correlation between X_4 and X_2.

Some authors prefer to write the above correlation matrix in the following form:

	X_1	X_2	X_3	X_4
X_1		0.5	0.3	0.6
X_2			0.7	0.8
X_3				0.4
X_4				

In this correlation matrix, diagonal values are not written as it is obvious that these values are 1 because correlation between the same two variables is always one.

In this section, we shall discuss the product moment correlation and partial correlation along with testing of their significance.

Product Moment Correlation Coefficient

Product moment correlation coefficient is an index which provides the magnitude of linear relationship between any two variables. When we refer to correlation matrix, it is usually a matrix of product moment correlation coefficients. It is represented by "r" and is given by the following formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}} \tag{4.1}$$

where N is the number of paired scores. The limits of r are from -1 to $+1$. The positive value of r means higher scores on one variable tend to be paired with higher scores on the other, or lower scores on one variable tend to be paired with lower scores on the other. On the other hand, negative value of r means higher scores on one variable tend to be paired with lower scores on the other and vice versa. Further, $r = +1$ indicates the perfect positive relationship between the two variables. This means that if there is an increase (decrease) in X by an amount "a," the Y will also be increased (decreased) by the same amount. Similarly $r = -1$ signifies the perfect negative linear correlation between the two variables. In this case, if the variable X is increased (decreased) by an amount "b," then the variable Y shall be decreased (increased) by the same amount. The three extreme values of the correlation coefficient r can be shown graphically in Fig. 4.1.

Example 4.1: Following are the scores on age and memory retention. Compute the correlation coefficient and test its significance at 5% level (Table 4.1).

Solution In order to compute the correlation coefficient, first of all the summation $\sum X$, $\sum Y$, $\sum X^2$, $\sum Y^2$, and $\sum XY$ shall be computed in Table 4.2.

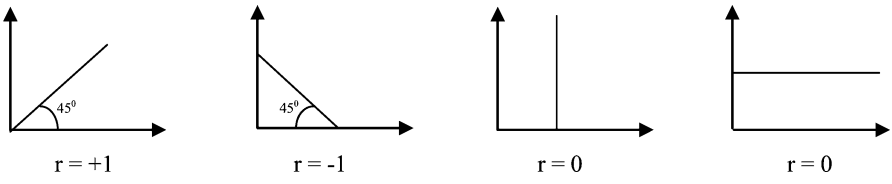


Fig. 4.1 Graphical presentations of the three extreme cases of correlation coefficient

Table 4.1 Data on age and memory retention

S.N.	Age	Memory retention
1	11	7
2	12	5
3	8	7
4	9	6
5	7	8
6	10	5
7	8	7
8	9	8
9	10	6
10	7	8

Table 4.2 Computation for correlation coefficient

S.N.	Age (X)	Memory retention (Y)	X^2	Y^2	XY
1	11	7	121	49	77
2	12	5	144	25	60
3	8	7	64	49	56
4	9	6	81	36	54
5	7	8	49	64	56
6	10	5	100	25	50
7	8	7	64	49	56
8	9	8	81	64	72
9	10	6	100	36	60
10	7	8	49	64	56
Total	91	67	853	461	597

Here N is 10:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

Substituting the values in the equation,

$$r = \frac{10 \times 597 - 91 \times 67}{\sqrt{[10 \times 853 - 91^2][10 \times 461 - 67^2]}}$$
$$= \frac{-127}{\sqrt{249 \times 121}} = -0.732$$

Testing the Significance

To test whether the correlation coefficient −0.732 is significant or not, the tabulated value of r required for significance at .05 level of significance and $N - 2(=8)$ degree of freedom can be seen from Table A.3 in the Appendix, which is 0.632. Hence, it may be concluded that there is a significant negative relationship between age and memory retention power. In other words, it may be inferred that as the age increases, the memory retention power decreases.

Properties of Coefficient of Correlation

1. The correlation coefficient is symmetrical with respect to the variables. In other words, correlation between height and weight is same as the correlation between weight and height. Mathematically $r_{xy} = r_{yx}$.
2. The correlation coefficient between any two variables lies in between -1 and $+1$. In other words, $-1 \leq r \leq 1$.
   Consider the following sum of the squares:

$$\sum \left[\frac{X-\bar{X}}{\sigma_x} \pm \frac{Y-\bar{Y}}{\sigma_y}\right]^2$$

$$\sigma_x^2 = \frac{1}{n}\sum (X-\bar{X})^2 \Rightarrow \sum (X-\bar{X})^2 = n\sigma_x^2$$

Since $\sigma_y^2 = \frac{1}{n}\sum (Y-\bar{Y})^2 \Rightarrow \sum (Y-\bar{Y})^2 = n\sigma_y^2$

and $r_{xy} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{n\sigma_x\sigma_y} \Rightarrow \sum (X-\bar{X})(Y-\bar{Y}) = n\sigma_x\sigma_y r_{xy}$

Now

$$\begin{aligned}\sum \left[\frac{X-\bar{X}}{\sigma_x} \pm \frac{Y-\bar{Y}}{\sigma_y}\right]^2 &= \frac{\sum (X-\bar{X})^2}{\sigma_x^2} \pm 2\frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sigma_x\sigma_y} + \frac{\sum (Y-\bar{Y})^2}{\sigma_y^2} \\ &= \frac{n\sigma_x^2}{\sigma_x^2} \pm \frac{2n\sigma_x\sigma_y r_{xy}}{\sigma_x\sigma_y} + \frac{n\sigma_y^2}{\sigma_y^2} = 2n \pm 2nr \\ &= 2n(1 \pm r)\end{aligned}$$

Since the expression in the left-hand side is always a positive quantity,

$$\therefore \qquad 2n(1 \pm r) \ \geq \ 0 \ (n{>}0)$$

Taking positive sign

$$1 + r \geq 0 \qquad \therefore\ r \geq -1 \tag{4.2}$$

And if the sign is negative,

$$1 - r \geq 0 \quad \therefore\ r \leq 1 \tag{4.3}$$

Combining (4.2) and (4.3),

$$-1 \leq r \leq 1$$

3. The correlation coefficient is independent of origin and unit of measurement (scale), that is,
if the two variables are denoted as X and Y and r_{xy}, the correlation coefficient, then

$$r_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{(X - \bar{X})^2}\sqrt{(Y - \bar{Y})^2}} \tag{4.4}$$

Let us apply the transformation by shifting the origin and scale of X and Y.
Let $U = \frac{X-a}{h}$ and $\mathrm{V} = \frac{\mathrm{Y-b}}{\mathrm{k}}$
where a, b, h, and k are constants.

$$\therefore \mathrm{X} = \mathrm{a} + \mathrm{hU} \Rightarrow \sum X = \sum a + h \sum U$$
$$\Rightarrow \bar{X} = a + h\bar{U}$$

Thus, $$X - \bar{\mathrm{X}} = h(U - \bar{\mathrm{U}}) \tag{4.5}$$

Similarly,

$$\mathrm{Y} = \mathrm{b} + \mathrm{kV} \Rightarrow \sum Y = \sum b + k \sum V$$
$$\Rightarrow \bar{Y} = b + k\bar{V}$$

Thus, $$Y - \bar{\mathrm{Y}} = k(V - \bar{\mathrm{V}}) \tag{4.6}$$

Substituting the values of $(X - \bar{X})$and $(Y - \bar{Y})$from The Equations (4.5) and (4.6) into (4.4),

$$r_{x,y} = \frac{\sum h(U - \bar{U}) \times k(V - \bar{V})}{\sqrt{h^2 \sum (U - \bar{U})^2}\sqrt{k^2 \sum (V - \bar{V})^2}}$$
$$= \frac{hk \sum (U - \bar{U})(V - \bar{V})}{hk\sqrt{\sum (U - \bar{U})^2}\sqrt{\sum (V - \bar{V})^2}}$$
$$= \mathrm{r_{u,v}}$$
$$\therefore \mathrm{r_{x,y}} = \mathrm{r_{u,v}}$$

Thus, it may be concluded that the coefficient of correlation between any two variables is independent of change of origin and scale.

4. Correlation coefficient is the geometrical mean between two regression coefficients. If b_{yx} and b_{xy} are the regression coefficients, then

$$r_{xy} = \pm\sqrt{b_{yx} \times b_{xy}}$$

Correlation Coefficient May Be Misleading

As per the definition, correlation coefficient indicates the linear relationship between the two variables. This value may be misleading at times. Look at the following three situations:

1. Researchers often conclude that a high degree of correlation implies a causal relationship between the two variables, but this is totally unjustified. For example, both events, represented by X_1 and X_2, might simply have a common cause. If in a study, X_1 represents the gross salary of the family per month and X_2 is the amount of money spent on the sports and leisure activities per month, then a strong positive correlation between X_1 and X_2 should not be concluded that the people spend more on sports and leisure activities if their family income is more. Now, if a third variable X_3, the socioeconomic status, is taken into account, it becomes clear that, despite the strong positive value of their correlation coefficient, there is no causal relationship between "sports and leisure expenditure" and "family income," and that both are in fact caused by the third variable "socioeconomic status." It is not the family income which encourages a person to spend more on sports and leisure activities but the socioeconomic status which is responsible for such a behavior. To know the causal relationship, partial correlation may be used with limitations.
2. A low or insignificant value of the correlation coefficient may not signify the lack of a strong link between the two variables under consideration. The lower value of correlation coefficient may be because of the other variables affecting the relationships in a negative manner. And, therefore, the effect of those variables eliminated may increase the magnitude of the correlation coefficient. Path Analysis may provide the insight in this direction where a correlation coefficient may be split into direct and indirect relationships.
3. The ecological fallacy is another source of misleading correlation coefficient. It occurs when a researcher makes an inference about the correlation in a particular situation based on correlation of aggregate data for a group. For instance, if a high degree of relationship exists between height and performance of athletes in the USA, it does not indicate that every tall athlete's performance is excellent in the USA. And if we conclude so, it will be an ecological fallacy.
4. Correlation does not explain causative relationship. High degree of correlation between two variables does not indicate that one variable causes another. In other words, correlation does not show cause and effect relationship. In a distance learning program, if there is a high degree of correlation between the student's performance and the number of contact classes attended, it does not necessarily indicate that one gets more marks because he learns more during contact classes. Neither does it necessarily imply that the more classes you attend, the more intelligent you become and get good marks. Some other explanation might also explain the correlation coefficient. The correlation means that the one who attends more contact classes gets higher marks and those who attend less classes get less marks. It does not explain why it is the case.

5. One must ensure that the result of correlation coefficient should be generalized only for that population from which the sample was drawn. Usually for a specific small sample, correlation may be high for any two variables, and if it is so, then it must be verified with the larger representative and relevant sample.

Limitations of Correlation Coefficients

One of the main limitations of the correlation coefficient is that it measures only linear relationship between the two variables. Thus, correlation coefficient should be computed only when the data are measured either on interval scale or ratio scale. The other limitation of the correlation coefficient is that it does not give the real relationship between the variables. To overcome this problem, partial correlation may be computed which explains the real relationship between the variables after controlling for other variables with certain limitations.

Testing the Significance of Correlation Coefficient

After computing the correlation coefficient, the next question is to find as to whether it actually explains some relationship or it is due to chance.

The following mutually exclusive hypotheses are tested by using the statistical test for testing the significance of correlation coefficient:

H_0: $\bar{r} = 0$ (There is no correlation between the two variables in the population.)
H_1: $\bar{r} \neq 0$ (There is a correlation between the two variables in the population.)

In fact, $\bar{r}$ indicates the population correlation coefficient, and we test its significance on the basis of the sample correlation coefficient. To test the above set of hypotheses, any of the following three approaches may be used.

First Approach

The easiest way to test the null hypothesis mentioned above is to look for the critical value of r with $n - 2$ degrees of freedom at any desired level of significance in Table A.3 in the Appendix. If the calculated value of r is less than or equal to the critical value of r, null hypothesis would fail to be rejected, and if calculated r is greater than critical value of r, null hypothesis may be rejected. For instance, if the correlation coefficient between height and self-esteem of 25 individuals is 0.45, then the critical value of r required for significance at .05 level of significance and $N - 2(=23)$ df from Table A.3 in the Appendix can be seen as 0.396. Since calculated value of r, that is, 0.45 is greater than the critical value of r (=0.396), the null hypothesis may be rejected at .05 level of significance, and we may conclude that there is a significant correlation between height and self-esteem.

Second Approach

Significance of correlation coefficient may be tested by using *t*-test as well. In this case, *t*-statistic is given by the following formula:

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2} \tag{4.7}$$

Here r is the observed correlation coefficient and n is the number of paired sets of data.

The calculated value of t is compared with that of tabulated value of t at .05 level and $n-2$ df ($=t_{.05}(n-2)$). The value of tabulated t can be obtained from Table A.2 in the Appendix.

Thus, if	Cal $t \leq t_{.05}(n-2)$,	null hypothesis is failed to be rejected at .05 level of significance
and if	Cal $t > t_{.05}(n-2)$,	null hypothesis may be rejected at .05 level of significance

Third Approach

In this approach, significance of correlation coefficient is tested on the basis of its p value. p value is the probability of wrongly rejecting the null hypothesis. If p value is .04 for a given correlation coefficient, it indicates that the chances of wrongly rejecting the null hypothesis are only 4%. Thus, so long p value is less than .05 the correlation coefficient is significant and the null hypothesis may be rejected at 5% level. On the other hand, if p value is more than or equal to .05, the correlation coefficient is not significant and the null hypothesis may not be rejected at 5% level.

Note: The SPSS output follows third approach and provides p values for each of the correlation coefficient in the correlation matrix.

Partial Correlation

Partial correlation is the measure of relationship between two variables after partialing out the effect of one or more independent variables. In computing partial correlation, the data must be measured either on interval or on ratio scale. For example, one may compute partial correlation if it is desired to see the relationship of age with stock portfolio after controlling the effect of income. Similarly to understand the relationship between price and demand would involve studying the relationship between price and demand after controlling the effect of money supply, exports, etc.

The partial correlation between X_1 and X_2 adjusted for X_3 is given by

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \tag{4.8}$$

The limits of partial correlation are −1 to +1.

The order of partial correlation refers to the number of independent variables whose effects are to be controlled. Thus, first-order partial correlation controls the effect of one variable, second-order partial correlation controls the effect of two variables, and so on.

The generalized formula for $(n - 2)$th order partial correlation is given by

$$r_{12.34.....n} = \frac{r_{12.345....(n-1)} - r_{1n.345...(n-1)}r_{2n.345....(n-1)}}{\sqrt{1 - r_{1n.345....(n-1)}^2}\sqrt{1 - r_{2n.345....(n-1)}^2}} \tag{4.9}$$

Limitations of Partial Correlation

1. Since partial correlation is computed by using product moment correlation coefficient, it also assumes the linear relationship. But generally, this assumption is not valid especially in social sciences, as linear relationship rarely exists in such parameters.
2. The reliability of partial correlation decreases if its order increases.
3. Large number of data is required to draw the valid conclusions from the partial correlations.
4. In spite of controlling the effect of many variables, one cannot be sure that the partial correlation explains the real relationship.

Testing the Significance of Partial Correlation

The significance of partial correlation is tested in a similar way as has been discussed above in case of product moment correlation.

In SPSS, significance of partial correlation is tested on the basis of p value. The partial correlation would be significant at 5% level if its p value is less than .05 and will be insignificant if the p value is equal to or more than .05.

Computation of Partial Correlation

Example 4.2: The following correlation matrix shows the correlation among different academic performance parameters. Compute partial correlations $r_{12.3}$ and $r_{12.34}$ and test their significance. Interpret the findings also (Table 4.3).

Table 4.3 Correlation matrix among different paramters

	X_1	X_2	X_3	X_4
X_1	1	0.7	0.6	0.4
X_2		1	0.65	0.3
X_3			1	0.5
X_4				1

X_1: GMAT scores, X_2: Mathematics marks in high school, X_3: IQ scores, X_4: GPA scores

Solution

(i) Computation of $r_{12.3}$

Since we know that $$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Substituting the values of correlation coefficients from the correlation matrix, we get

$$\begin{aligned} r_{12.3} &= \frac{0.7 - 0.6 \times 0.65}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.65^2}} \\ &= \frac{0.70 - 0.39}{\sqrt{0.64}\sqrt{0.5775}} \\ &= 0.51 \end{aligned}$$

(ii) Computation of $r_{12.34}$

Since we know that $r_{12.34} = \dfrac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}}$

We shall first compute the first-order partial correlations $r_{12.3}$, $r_{14.3,}$ and $r_{24.3}$ which are required to compute the second-order partial correlation $r_{12.34}$. Since $r_{12.3}$ has already been computed above, the remaining two shall be computed here.

Thus,

$$r_{14.3} = \frac{r_{14} - r_{13}r_{43}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{43}^2}} = \frac{0.4 - 0.6 \times 0.5}{\sqrt{1 - 0.6^2}\sqrt{1 - 0.5^2}} = \frac{0.1}{0.69} = 0.14$$

and

$$r_{24.3} = \frac{r_{24} - r_{23}r_{43}}{\sqrt{1 - r_{23}^2}\sqrt{1 - r_{43}^2}} = \frac{0.3 - 0.65 \times 0.5}{\sqrt{1 - 0.65^2}\sqrt{1 - 0.5^2}} = \frac{-0.025}{0.66} = -0.04$$

After substituting the values of $r_{12.3}$,$r_{14.3}$, and $r_{24.3}$, the second-order partial correlation becomes

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{1 - r_{14.3}^2}\sqrt{1 - r_{24.3}^2}} = \frac{0.51 - 0.14 \times (-0.04)}{\sqrt{1 - 0.14^2}\sqrt{1 - (-0.04)^2}}$$
$$= \frac{0.51 + 0.0056}{\sqrt{0.98}\sqrt{0.998}} = \frac{0.5156}{0.989}$$
$$= 0.521$$

Situation for Using Correlation Matrix and Partial Correlation

Employee's performance largely depends upon their work environment, and, therefore, organizations give more emphasis to improve the working environment of their employees by means of different programs and policies. In order to know the various parameters that are responsible for job satisfaction, statistical techniques like correlation coefficient and partial correlation can be used. With the help of these statistics, one can understand the extent of multicollinearity among independent variables besides understanding the pattern of relationship between job satisfaction and independent variables.

In order to develop an effective strategy to improve the level of job satisfaction of employees, one should know as to what parameters are significantly associated with it. These variables can be identified from the correlation matrix. All those variables which show significant relation with the job satisfaction may be identified for further investigation. Out of these identified variables, it may be pertinent to know as to which variable is the most important one. Simply looking to the magnitude of the correlation coefficient, it is not possible to identify the most important variable responsible for job satisfaction because high correlation does not necessarily mean real relationship as it may be due to other independent variables. Thus, in order to know as to which variable is the most important one, partial correlation may be computed by eliminating the effect of other variables so identified in the correlation matrix.

The application of correlation matrix and partial correlation can be understood by considering the following research study:

Consider a situation where an organization is interested in investigating the relationship of job satisfaction with certain environmental and motivational variables obtained on its employees. Besides finding the relationships of job satisfaction with environmental and motivational variables, it may be interesting to know the relationships among the environmental and motivational variables as well. The following variables may be taken in the study:

Dependent variable

1. Job satisfaction (X_1)

Independent variables

1. Autonomy (X_2)
2. Organizational culture (X_3)
3. Compensation (X_4)
4. Upward communications (X_5)
5. Job training opportunity (X_6)
6. Management style (X_7)
7. Performance appraisal (X_8)
8. Recognition (X_9)
9. Working atmosphere (X_{10})
10. Working relationships (X_{11})

The following computations may be done to fulifil the objectives of the study:

1. Compute product moment correlation coefficient between Job satisfaction and each of the environmental and motivational variables.
2. Identify few independent variables that show significant correlations with the Job satisfaction for further developing the regression model. Say these selected variables are X_3, X_9, X_6, and X_2.
3. Out of these identified variables in step 2, pick up the one having the highest correlation with the dependent variable (X_1), say it is X_6.
4. Then find the partial correlation between the variables X_1 and X_6 by eliminating the effect of variables X_3, X_9, and X_2 in steps. In other words, find the first-order partial correlation $r_{16.3}$, second-order partial correlation $r_{16.39}$, and third-order partial correlation $r_{16.392}$.
5. Similarly find the partial correlation between other identified variables X_3, X_9, and X_2 with that of dependent variable (X_1) in steps. In other words, compute the following three more sets of partial correlation:

(i) $r_{13.9}$, $r_{13.96}$, and $r_{13.962}$
(ii) $r_{19.3}$, $r_{19.36}$, and $r_{19.362}$
(iii) $r_{12.3}$, $r_{12.39}$, and $r_{12.396}$

Research Hypotheses to Be Tested

By computing product moment correlation and partial correlation, the hypotheses that can be tested are as follows:

(a) To test the significance of relationship between Job satisfaction and each of the environmental and motivational variables
(b) To test the significance of relationship among independent variables
(c) Whether few environmental and motivational variables are highly related with Job satisfaction

Statistical Test

To address the objectives of the study and to test the listed hypotheses, the following computations may be done:

1. Correlation matrix among all the independent variables and dependent variable
2. Partial correlations of different orders between the Job satisfaction and identified independent variables

Thus, we have seen how a research situation requires computing correlation matrix and partial correlations to fulfill the objectives.

Solved Example of Correlation Matrix and Partial Correlations by SPSS

Example 4.3 To understand the relationships between patient's loyalty and other variables, a study was conducted on 20 patients in a hospital. The following data was obtained. Construct the correlation matrix and compute different partial correlations using SPSS and interpret the findings (Table 4.4).

Table 4.4 Data on patient's loyalty and other determinants

S.N.	Trust of patient	Service quality	Customer satisfaction	Patient loyalty
1	37	69	52	25
2	35	69	50	20
3	41	94	70	26
4	33	50	41	7
5	54	91	66	25
6	41	69	56	20
7	44	68	53	23
8	45	95	71	32
9	49	95	68	21
10	42	75	57	28
11	35	82	70	28
12	37	80	57	22
13	47	82	61	23
14	44	74	59	26
15	54	100	78	29
16	35	82	54	24
17	39	63	36	16
18	32	57	38	15
19	53	99	74	32
20	49	98	63	25

Solution First of all, correlation matrix shall be computed using SPSS. Option shall be selected to show the significant correlation values. After selecting the variables that shows significant correlation with the customer loyalty, partial correlation shall be computed between customer loyalty and any of these selected variables after controlling the effect of the remaining variables. The correlation coefficients and partial correlations so obtained in the output using SPSS shall be tested for their significance by using the *p* value.

Computation of Correlation Matrix Using SPSS

(a) **Preparing Data File**
Before using the SPSS commands for computing correlation matrix, the data file is required to be prepared. The following steps will help you prepare the data file:

(i) *Starting the SPSS:* Use the following command sequence to start SPSS:

Start ⟶ All Programs ⟶ IBM SPSS Statistics ⟶ IBM SPSS Statistics 20

After checking the option **Type in Data** on the screen you will be taken to the **Variable View** option for defining the variables in the study.

(ii) *Defining variables:* There are four variables, namely, *Customer trust, Service quality, Customer satisfaction*, and *Customer loyalty* that need to be defined. Since these variables were measured on interval scale, they will be defined as Scale variable in SPSS. Any variable measured on interval or ratio scale is defined as Scale variable in SPSS. The procedure of defining the variable in SPSS is as follows:

1. Click **Variable View** to define variables and their properties.
2. Write short name of these variables, that is, *Trust, Service, Satisfaction,* and *Loyalty* under the column heading **Name.**
3. Full names of these variables may be defined as *Customer trust, Service quality, Customer satisfaction*, and *Customer loyalty* under the column heading **Label**.
4. Under the column heading **Measure**, select the option "Scale" for all these variables.
5. Use default entries in rest of the columns.

After defining all the variables in Variable View, the screen shall look like Fig. 4.2.

(iii) *Entering data* After defining these variables in the **Variable View**, click **Data View** on the left bottom of the screen to open the format for entering data. For each variable, enter the data column wise. After entering data, the screen will look like Fig. 4.3. Save the data file in the desired location before further processing.

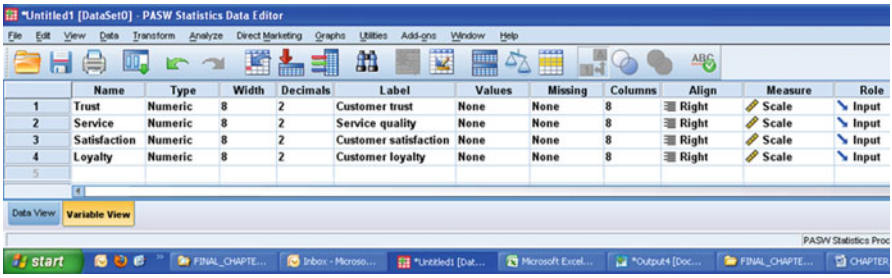


Fig. 4.2 Defining variables along with their characteristics

(b) **SPSS *Commands for Computing Correlation Coefficient***

After preparing the data file in data view, take the following steps to prepare the correlation matrix:

(i) *Initiating the SPSS commands to compute correlations:* In Data View, click the following commands in sequence:

Analyze ⟶ Correlate ⟶ Bivariate

The screen shall look like Fig. 4.4.

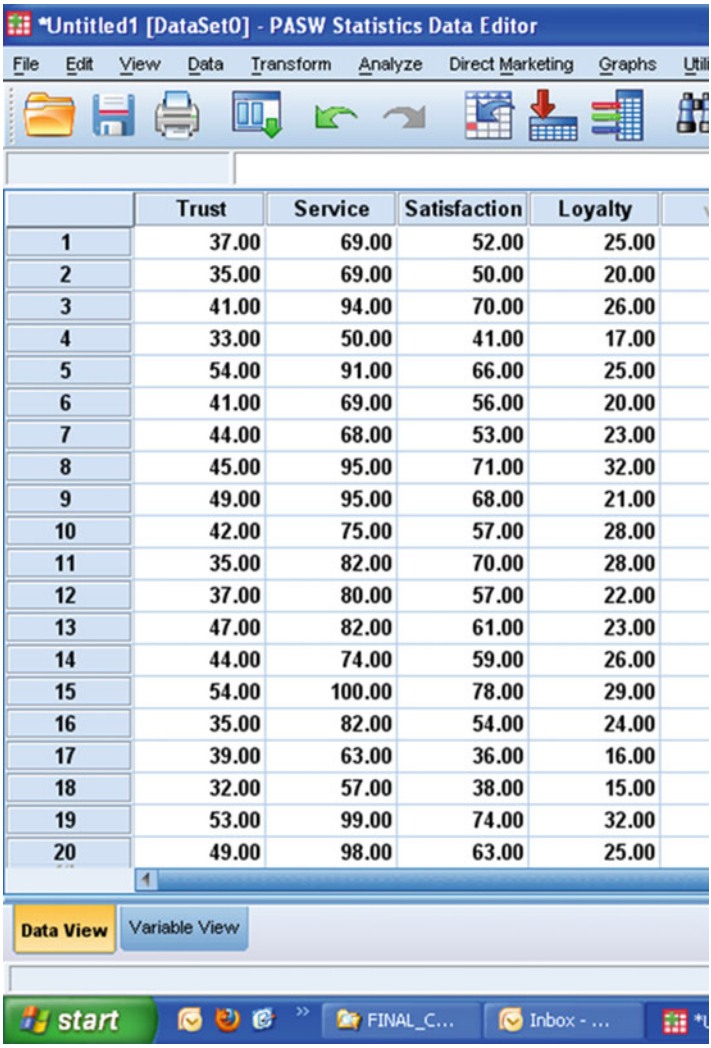
(ii) *Selecting variables for correlation matrix:* Clicking **Bivariate** option will take you to the next screen for selecting variables for the correlation matrix. Select all the variables from left panel to the right panel by using the arrow key. The variable selection may be made one by one or all at once. After selecting the variables, the screen shall look like Fig. 4.5.

(iii) *Selecting options for computation* After selecting the variables, option need to be defined for the correlation analysis. Take the following steps:

- In the screen shown in Fig. 4.5, ensure that the "Pearson," "Two-tailed," and "Flag significant correlations" options are checked. By default they are checked.
- Click the tag **Options**. This will take you to the screen shown in Fig. 4.6.
  - Check the option "Means and standard deviation."
  - Use default entries in other options. Readers are advised to try other options and see what changes they are getting in their output.
  - Click ***Continue***. This will take you back to the screen shown in Fig. 4.5.
- Click **OK.**

(c) ***Getting the Output***

After clicking **OK,** output shall be generated in the output windows. The two outputs generated in the form of descriptive statistics and correlation matrix are shown in Tables 4.5 and 4.6.



	Trust	Service	Satisfaction	Loyalty
1	37.00	69.00	52.00	25.00
2	35.00	69.00	50.00	20.00
3	41.00	94.00	70.00	26.00
4	33.00	50.00	41.00	17.00
5	54.00	91.00	66.00	25.00
6	41.00	69.00	56.00	20.00
7	44.00	68.00	53.00	23.00
8	45.00	95.00	71.00	32.00
9	49.00	95.00	68.00	21.00
10	42.00	75.00	57.00	28.00
11	35.00	82.00	70.00	28.00
12	37.00	80.00	57.00	22.00
13	47.00	82.00	61.00	23.00
14	44.00	74.00	59.00	26.00
15	54.00	100.00	78.00	29.00
16	35.00	82.00	54.00	24.00
17	39.00	63.00	36.00	16.00
18	32.00	57.00	38.00	15.00
19	53.00	99.00	74.00	32.00
20	49.00	98.00	63.00	25.00

Fig. 4.3 Scheme of data feeding for all the variables

Interpretation of the Outputs

The values of mean and standard deviation for all the variables are shown in Table 4.5. The user may draw the conclusions accordingly, and the findings may be used for further analysis in the study.

The actual output shows the full correlation matrix, but only upper diagonal values of the correlation matrix are shown in Table 4.6. This table shows the magnitude of correlation coefficients along with their p values and sample size. The product moment correlation coefficient is also known as Pearson correlation as

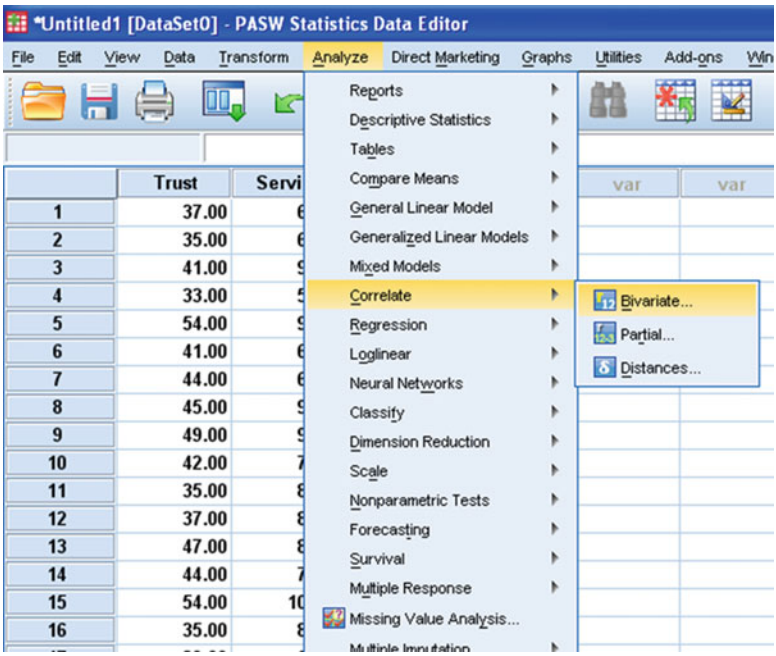


Fig. 4.4 Screen showing SPSS commands for computing correlation matrix

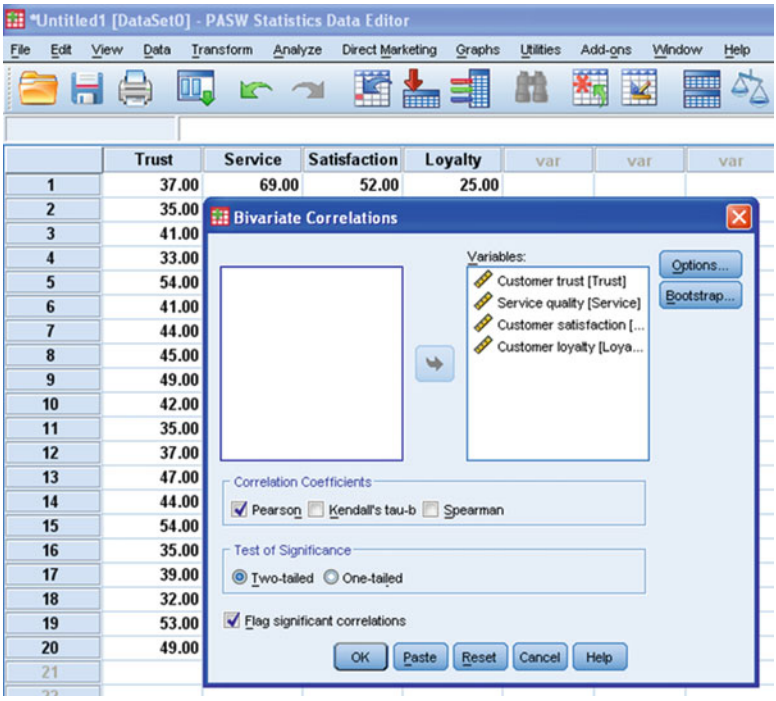


Fig. 4.5 Screen showing selection of variables for computing correlation matrix

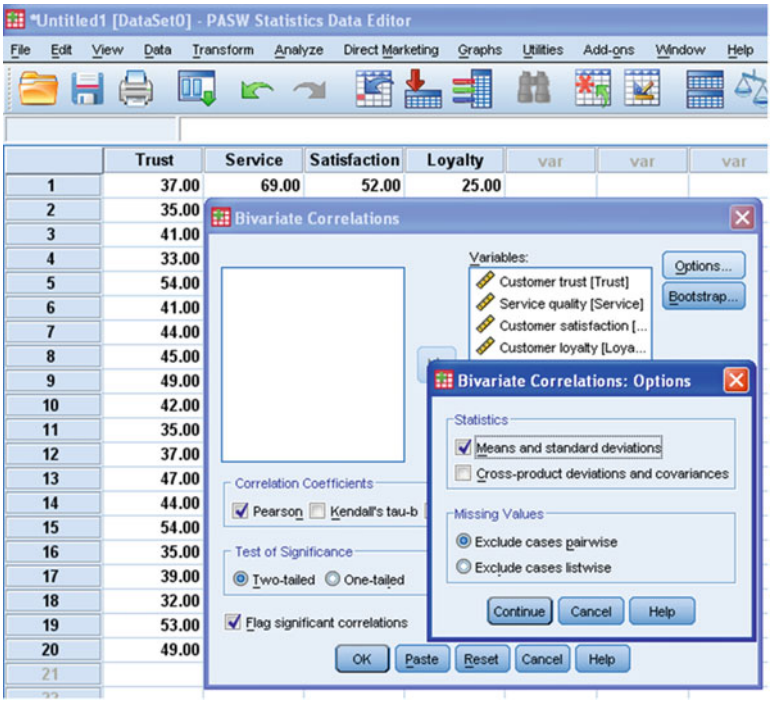


Fig. 4.6 Screen showing option for computing correlation matrix and other statistics

it was developed by the British mathematician Karl Pearson. The value of correlation coefficient required for significance (known as critical value) at 5% as well as at 1% level can be seen from Table A.3 in the Appendix. Thus, at 18 degrees of freedom, the critical values of r at 5 and 1% are 0.444 and 0.561, respectively. The correlation coefficient with one asterisk (*) mark is significant at 5% level, whereas the one with two asterisk (**) marks shows the significance at 1% level. In this example, the research hypothesis is two-tailed which states that "There is a significant correlation between the two variables." The following conclusions may be drawn from the results in Table 4.6:

(a) The Customer loyalty is significantly correlated with customer trust at 5% level, whereas it is significantly correlated with Service quality and Customer satisfaction at 1% level.
(b) Customer satisfaction is highly correlated with service quality. This is rightly so as only satisfied customers would be loyal to any hospital.
(c) All those correlation coefficients having p value less than .05 are significant at 5% level. This is shown by asterisk (*) mark by the side of the correlation coefficient. Similarly correlations having p value less than .01 are significant at 1% level, and this is indicated by two asterisk (**) marks by the side of correlation coefficient.

Table 4.5 Descriptive statistics for the data on customer's behavior

Variables	Mean	SD	*N*
Customer trust	42.3000	7.01952	20
Service quality	79.6000	14.77338	20
Customer satisfaction	58.7000	11.74779	20
Customer loyalty	23.8500	4.79336	20

Table 4.6 Correlation matrix for the data on customer's behavior along with p values

		Customer trust	Service quality	Customer satisfaction	Customer loyalty
		(X_1)	(X_2)	(X_3)	(X_4)
Customer trust	Pearson correlation	1	.754**	.704**	.550*
(X_1)	sig. (2-tailed)		.000	.001	.012
	N	20	20	20	20
Service quality	Pearson correlation		1	.910**	.742**
(X_2)	sig. (2-tailed)			.000	.000
	N		20	20	20
Customer satisfaction	Pearson correlation			1	.841**
(X_3)	sig. (2-tailed)				.000
	N			20	20
Customer loyalty	Pearson correlation				1
(X_4)	sig. (2-tailed)				
	N				20

**Correlation is significant at the 0.01 level (2-tailed); *Correlation is significant at the 0.05 level (2-tailed)

Computation of Partial Correlations Using SPSS

The decision of variables among which partial correlation needs to be computed depends upon objective of the study. In computing partial correlation, one of the variables is usually a criterion variable, and the other is the independent variable having the highest magnitude of correlation with the criterion variable. Criterion variable is the one in which the variation is studied as result of variation in other independent variables. Usually criterion variable is known as dependent variable. Here the criterion variable is the Customer loyalty because the effect of other variables needs to be investigated on it. Depending upon the situation, the researcher may choose any variable other than the highest correlated variable for computing partial correlation with that of dependent variable. In this example, partial correlation shall be computed between Customer loyalty (X_4) and Customer satisfaction (X_3) after eliminating the effect of Service quality (X_2) and Customer trust (X_1). This is because X_3 is highly correlated with the criterion variable X_4.

The decision of eliminating the effect of variables X_2 and X_1 has been taken because both these variables are significantly correlated with the criterion variable. However, one can investigate the relationship between X_4 vs. X_2 after eliminating the effect of the variables X_3 and X_1. Similarly partial correlation between X_4 vs. X_1 may also be investigated after eliminating the effect of the variables X_3 and X_2. The procedure of computing these partial correlations with SPSS has been discussed in the following sections:

(a) ***Data File for Computing Partial Correlation***
The data file which was prepared for computing correlation matrix shall be used for computing the partial correlations. Thus, procedure for defining the variables and entering the data for all the variables is exactly the same as was done in case of computing correlation matrix.

(b) **SPSS *Commands for Partial Correlation***
After entering all the data in the data view, take the following steps for computing partial correlation:

(i) *Initiating the SPSS commands for partial correlation:* In Data View, go to the following commands in sequence:

Analyze ⟶ Correlate ⟶ Partial

The screen shall look like Fig. 4.7.

(ii) *Selecting variables for partial correlation:* After clicking the **Partial** option, you will get the next screen for selecting variables for the partial correlation.

- Select the two variables *Customer loyalty* (X_4) and *Customer satisfaction* (X_3) from the left panel to the "Variables" section in the right panel. Here, relationship between the variables X_4 and X_3 needs to be computed after controlling the effects of *Service quality* (X_2) and *Customer trust* (X_1).
- Select the variables *Service quality* (X_2) and *Customer trust* (X_1) from the left panel to the "Controlling for" section in the right panel. X_2 and X_1 are the two variables whose effects are to be eliminated.

The selection of variables is made either one by one or all at once. To do so, the variable needs to be selected from the left panel, and by arrow command, it may be brought to the right panel. The screen shall look like Fig. 4.8.

(iii) *Selecting options for computation:* After selecting the variables for partial correlation and identifying controlling variables, option needs to be defined for the computation of partial correlation. Take the following steps:

- In the screen shown in Fig. 4.8, ensure that the options "Two-tailed" and "Display actual significance level" are checked. By default they are checked.
- Click the tag **Options**; you will get the screen as shown in Fig. 4.9. Take the following steps:
  - Check the box of "Means and standard deviations."

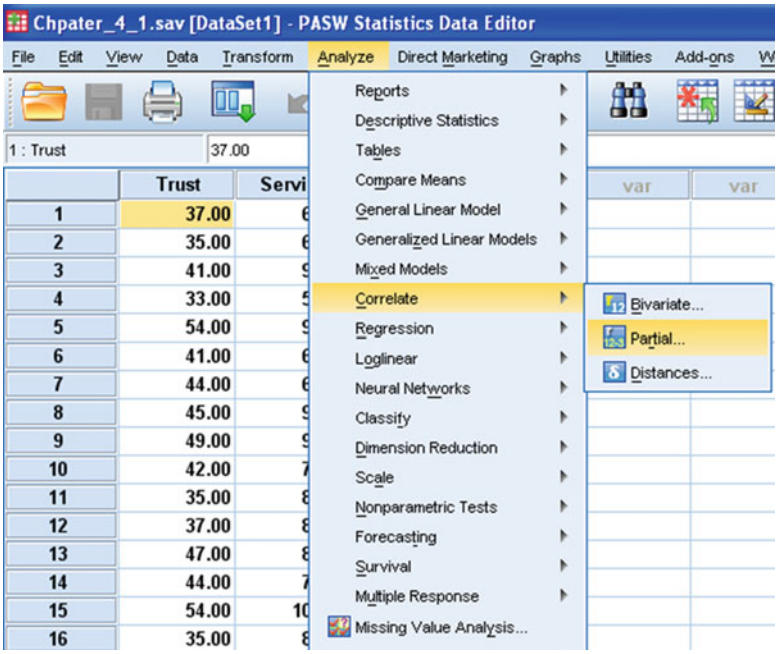


Fig. 4.7 Screen showing SPSS commands for computing partial correlations

- Use the default entries in other options. Readers are advised to try other options and see what changes they are getting in their outputs.
- Click ***Continue***.

– Click **OK.**

(c) ***Getting the Output***

After clicking **OK**, outputs shall be generated in the output panel. The output panel shall have two tables: one for descriptive statistics and the other for partial correlation. These outputs can be selected by right click of the mouse and may be pasted in the word file. In this example, the output so generated by the SPSS will look like as shown in Tables 4.7 and 4.8.

Interpretation of Partial Correlation

Table 4.7 shows the descriptive statistics of all the variables selected in the study. Values of mean and standard deviations may be utilized for further analysis. Readers may note that similar table of descriptive statistics was also obtained while computing correlation matrix by using SPSS.

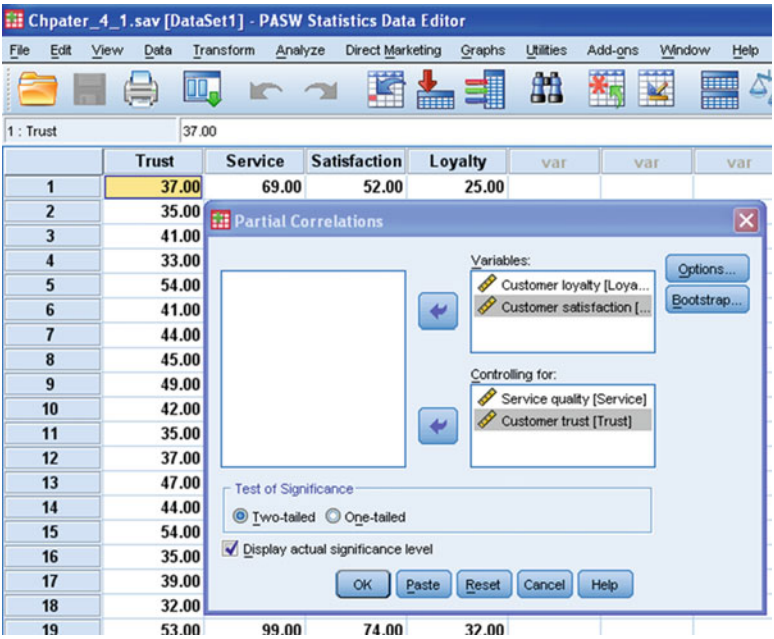


Fig. 4.8 Screen showing selection of variables for partial correlation

In Table 4.8, partial correlation between Customer loyalty (X_4) and Customer satisfaction (X_3) after controlling the effect of Service quality (X_2) and Customer trust (X_1) is shown as 0.600. Since p value for this partial correlation is .009, which is less than .01, it is significant at 1% level. It may be noted that the correlation coefficient between Customer loyalty and Customer satisfaction in Table 4.6 is 0.841 which is highly significant, but when the effects of Service quality and Customer trust are eliminated, the actual correlation dropped down to 0.600. But this partial correlation of 0.600 is still highly correlated in the given sample, and, hence, it may be concluded that within the framework of this study, there exists a real relationship between Customer loyalty and Customer satisfaction. One may draw the conclusion that at all cost, Customer satisfaction is the most important factor for maintaining patient's loyalty towards the hospital.

Summary of the SPSS Commands

(a) For Computing Correlation Matrix

1. Start the SPSS by using the following commands:

Start⟶ All Programs ⟶ IBM SPSS Statistics ⟶ IBM SPSS Statistics 20

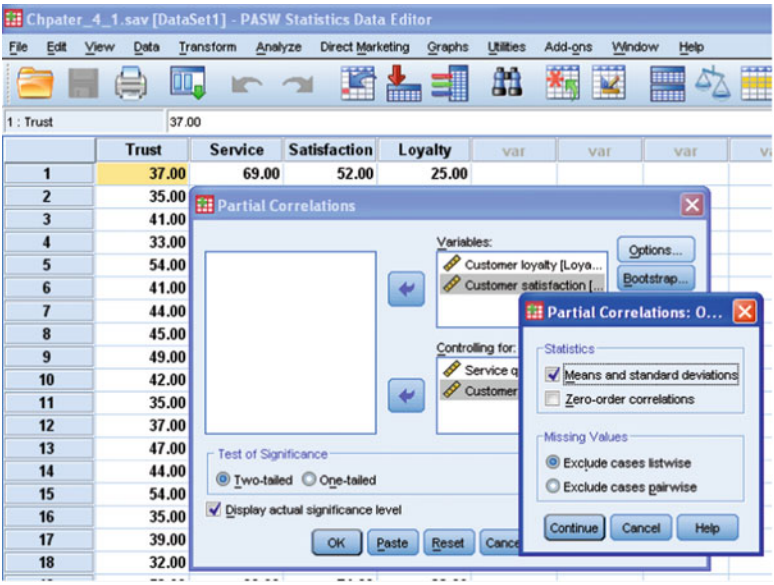


Fig. 4.9 Screen showing option for computing partial correlation and other statistics

Table 4.7 Descriptive statistics for the variables selected for partial correlations

Variables	Mean	SD	*N*
Customer loyalty	23.8500	4.79336	20
Customer satisfaction	58.7000	11.74779	20
Service quality	79.6000	14.77338	20
Customer trust	42.3000	7.01952	20

Table 4.8 Partial correlation between Customer loyalty (X_4) and Customer satisfaction (X_3) after controlling the effect of Service quality (X_3) and Customer trust (X_1)

Control variables			Customer loyalty (X_4)	Customer satisfaction (X_3)
Service quality (X_2) and Customer trust (X_1)	Customer loyalty (X_4)	Correlation	1.000	.600
		significance (2-tailed)		.009
		df	0	16
	Customer satisfaction (X_3)	Correlation	.600	1.000
		significance (2-tailed)	.009	
		df	16	0

Note: Readers are advised to compute partial correlations of different orders with the same data

2. Click **Variable View** tag and define the variables *Trust, Service, Satisfaction,* and *Loyalty* as Scale variables.
3. Once the variables are defined, type the data column wise for these variables by clicking **Data View**.
4. In Data View, click the following commands in sequence for correlation matrix:

Analyze ⟶ Correlate ⟶ Bivariate

5. Select all the variables from left panel to the "Variables" section of the right panel.
6. Ensure that the options "Pearson," "Two-tailed," and "Flag significant correlations" are checked by default.
7. Click the tag **Options** and check the box of "Means and standard deviations." Click ***Continue***.
8. Click **OK** for output.

(b) **For Computing Partial Correlation**

1. Follow steps 1–3 as discussed above.
2. With the same data file, follow the below-mentioned commands in sequence for computing partial correlations:

Analyze ⟶ Correlate ⟶ Partial

3. Select any two variables between which the partial correlation needs to be computed from left panel to the "Variables" section of the right panel.
   Select the variables whose effects are to be controlled, from left panel to the "Controlling for" section in the right panel.
4. After selecting the variables for computing partial correlation, click the caption **Options** on the screen. Check the box "Means and standard deviation" and press ***Continue***.
5. Click **OK** to get the output of the partial correlation and descriptive statistics.

Exercise

Short-Answer Questions

Note: Write the answer to each of the questions in not more than 200 words.

Q.1. "Product moment correlation coefficient is a deceptive measure of relationship, as it does not reveal anything about the real relationship between two variables." Comment on this statement.

Q.2. Describe a research situation in management where partial correlation can be used to draw some meaningful conclusions.

Q.3. Compute correlation coefficient between X and Y and interpret your findings considering that Y and X are perfectly related by equation $Y = X^2$.

X:	-3	-2	-1	0	1	2
Y:	9	4	1	0	1	4

Q.4. How will you test the significance of partial correlation using t-test?

Q.5. What does the p value refer to? How is it used in testing the significance of product moment correlation coefficient?

Multiple-Choice Questions

Note: For each of the question, there are four alternative answers. Tick mark the one that you consider the closest to the correct answer.

1. In testing the significance of product moment correlation coefficient, degree of freedom for t-test is

   (a) $N - 1$
   (b) $N + 2$
   (c) $N + 1$
   (d) $N - 2$

2. If the sample size increases, the value of correlation coefficient required for its significance

   (a) Increases
   (b) Decreases
   (c) Remains constant
   (d) May increase or decrease

3. Product moment correlation coefficient measures the relationship which is

   (a) Real
   (b) Linear
   (c) Curvilinear
   (d) None of the above

4. Given that $r_{12} = 0.7$ and $r_{12.3} = 0.28$, where X_1 is academic performance, X_2 is entrance test score, and X_3 is IQ, what interpretation can be drawn?

   (a) Entrance test score is an important contributory variable to the academic performance.
   (b) IQ affects the relationship between academic performance and entrance test score in a negative fashion.
   (c) IQ has got nothing to do with the academic performance.
   (d) It seems there is no real relationship between academic performance and entrance test score.

5. If p value for a partial correlation is 0.001, what conclusion can be drawn?

   (a) Partial correlation is not significant at 5% level.
   (b) Partial correlation is significant at 1% level.

(c) Partial correlation is not significant at 1% level.
(d) Partial correlation is not significant at 10% level.

6. Partial correlation is computed with the data that are measured in

(a) Interval scale
(b) Nominal scale
(c) Ordinal scale
(d) Any scale

7. In computing correlation matrix through SPSS, all variables are defined as

(a) Nominal
(b) Ordinal
(c) Scale
(d) Any of the nominal, ordinal, or scale option depending upon the nature of variable

8. In computing correlation matrix through SPSS, the following command sequence is used:

(a) Analyze -> Bivariate -> Correlate
(b) Analyze -> Correlate -> Bivariate
(c) Analyze -> Correlate -> Partial
(d) Analyze -> Partial -> Bivariate

9. While selecting variables for computing partial correlation in SPSS, in "Controlling for" section, the variables selected are

(a) All independent variables except the two between which partial correlation is computed.
(b) Any of the independent variables as it does not affect partial correlation.
(c) Only those variables whose effects need to be eliminated.
(d) None of the above is correct.

10. The limits of partial correlation are

(a) −1 to 0
(b) 0−1
(c) Sometimes more than 1
(d) −1 to +1

Assignments

1. In a study, Job satisfaction and other organizational variables as perceived by the employees were assessed. The data were obtained on interval scale and are shown in the below mentioned table. Compute the following:

(a) Correlation matrix with all the seven variables
(b) Partial correlations : $r_{12.3}$, $r_{12.35}$, and $r_{12.356}$
(c) Partial correlations : $r_{13.2}$, $r_{13.25}$, and $r_{13.256}$
(d) Partial correlations : $r_{16.2}$, $r_{16.23}$, and $r_{16.235}$

Data on Job satisfaction and other organizational variables as perceived by the employees

S. N.	Job satisfaction (X_1)	Autonomy (X_2)	Organizational culture (X_3)	Compensation (X_4)	Job training opportunity (X_5)	Recognition (X_6)	Working atmosphere (X_7)
1	75	45	34	23	35	44	23
2	65	41	31	24	32	34	32
3	34	27	25	14	28	38	25
4	54	38	28	25	37	32	27
5	47	32	26	26	28	37	31
6	33	28	32	14	25	23	32
7	68	41	38	23	38	32	28
8	76	42	45	28	29	42	42
9	68	37	42	26	29	37	24
10	45	29	35	27	19	30	22
11	36	32	23	31	18	26	31
12	66	33	44	28	38	39	43
13	72	41	45	24	35	36	27
14	58	41	38	25	26	36	28
15	26	23	25	26	24	18	19
16	61	45	42	22	36	42	39

2. The data in the following table shows the determinants of US domestic price of copper during 1966–1980. Compute the following and interpret your findings:

   (a) Correlation matrix with all the six variables
   (b) Partial correlations: $r_{12.3,}$ $r_{12.34}$, and $r_{12.346}$
   (c) Partial correlations: $r_{13.2,}$ $r_{13.24}$, and $r_{13.246}$

   Determinants of US domestic price of copper

Year	Avg. domestic copper price (X_1)	GNP (X_2)	Industrial production (X_3)	Exchange copper price (X_4)	No. of housing/year (X_5)	Aluminum price (X_6)
1966	36.60	753.00	97.8	555.0	1,195.8	24.50
1967	38.60	796.30	100.0	418.0	1,321.9	24.98
1968	42.20	868.50	106.3	525.2	1,545.4	25.58
1969	47.90	935.50	111.1	620.7	1,499.5	27.18
1970	58.20	982.40	107.8	588.6	1,469.0	28.72
1971	52.00	1,063.4	109.6	444.4	2,084.5	29.00
1972	51.20	1,171.1	119.7	427.8	2,378.5	26.67
1973	59.50	1,306.6	129.8	727.1	2,057.5	25.33
1974	77.30	1,412.9	129.3	877.6	1,352.5	34.06
1975	64.20	1,528.8	117.8	556.6	1,171.4	39.79
1976	69.60	1,700.1	129.8	780.6	1,547.6	44.49
1977	66.80	1,887.2	137.1	750.7	1,989.8	51.23
1978	66.50	2,127.6	145.2	709.8	2,023.3	54.42

(continued)

(continued)

Year	Avg. domestic copper price (X_1)	GNP (X_2)	Industrial production (X_3)	Exchange copper price (X_4)	No. of housing/year (X_5)	Aluminum price (X_6)
1979	98.30	2,628.8	152.5	935.7	1,749.2	61.01
1980	101.40	2,633.1	147.1	940.9	1,298.5	70.87

Note: The data were collected by Gary R. Smith from sources such as American Metal Market, Metals Week, and US Department of Commerce publications

Answers to Multiple-Choice Questions

Q.1	d	Q.2	b
Q.3	b	Q.4	d
Q.5	b	Q.6	a
Q.7	c	Q.8	b
Q.9	c	Q.10	d