CHAPTER 15

Multiple and Partial Correlation

MULTIPLE CORRELATION

The degree of relationship existing between three or more variables is called *multiple correlation*. The fundamental principles involved in problems of multiple correlation are analogous to those of simple correlation, as treated in Chapter 14.

SUBSCRIPT NOTATION

To allow for generalizations to large numbers of variables, it is convenient to adopt a notation involving subscripts.

We shall let $X_1, X_2, X_3, \ldots$ denote the variables under consideration. Then we can let $X_{11}, X_{12}, X_{13}, \ldots$ denote the values assumed by the variable X_1, and X_{21}, X_{22}, $X_{23}, \ldots$ denote the values assumed by the variable X_2, and so on. With this notation, a sum such as $X_{21} + X_{22} + X_{23} + \cdots + X_{2N}$ could be written $\sum_{j=1}^{N} X_{2j}$, $\sum_j X_{2j}$, or simply $\sum X_2$. When no ambiguity can result, we use the last notation. In such case the mean of X_2 is written $\bar{X}_2 = \sum X_2/N$.

REGRESSION EQUATIONS AND REGRESSION PLANES

A *regression equation* is an equation for estimating a dependent variable, say X_1, from the independent variables $X_2, X_3, \ldots$ and is called a *regression equation of* X_1 *on* $X_2, X_3, \ldots$. In functional notation this is sometimes written briefly as $X_1 = F(X_2, X_3, \ldots)$ (read "X_1 is a function of X_2, X_3, and so on").

For the case of three variables, the simplest regression equation of X_1 on X_2 and X_3 has the form

$$X_1 = b_{1.23} + b_{12.3}X_2 + b_{13.2}X_3 \qquad (1)$$

where $b_{1.23}, b_{12.3}$, and $b_{13.2}$ are constants. If we keep X_3 constant in equation (*1*), the graph of X_1 versus X_2 is a straight line with slope $b_{12.3}$. If we keep X_2 constant, the graph of X_1 versus X_3 is a straight line with slope $b_{13.2}$. It is clear that the subscripts after the dot indicate the variables held constant in each case.

Due to the fact that X_1 varies partially because of variation in X_2 and partially because of variation in X_3, we call $b_{12.3}$ and $b_{13.2}$ the *partial regression coefficients* of X_1 on X_2 keeping X_3 constant and of X_1 on X_3 keeping X_2 constant, respectively.

Equation (*1*) is called a *linear regression equation* of X_1 on X_2 and X_3. In a three-dimensional rectangular coordinate system it represents a plane called a *regression plane* and is a generalization of the regression line for two variables, as considered in Chapter 13.

NORMAL EQUATIONS FOR THE LEAST-SQUARES REGRESSION PLANE

Just as there exist least-squares regression lines approximating a set of N data points (X, Y) in a two-dimensional scatter diagram, so also there exist *least-squares regression planes* fitting a set of N data points (X_1, X_2, X_3) in a three-dimensional scatter diagram.

The least-squares regression plane of X_1 on X_2 and X_3 has the equation (*1*) where $b_{1.23}, b_{12.3}$, and $b_{13.2}$ are determined by solving simultaneously the *normal equations*

$$\begin{aligned}
\sum X_1 &= b_{1.23}N + b_{12.3}\sum X_2 + b_{13.2}\sum X_3 \\
\sum X_1X_2 &= b_{1.23}\sum X_2 + b_{12.3}\sum X_2^2 + b_{13.2}\sum X_2X_3 \\
\sum X_1X_3 &= b_{1.23}\sum X_3 + b_{12.3}\sum X_2X_3 + b_{13.2}\sum X_3^2
\end{aligned} \tag{2}$$

These can be obtained formally by multiplying both sides of equation (*1*) by 1, X_2, and X_3 successively and summing on both sides.

Unless otherwise specified, whenever we refer to a regression equation it will be assumed that the least-squares regression equation is meant.

If $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, and $x_3 = X_3 - \bar{X}_3$, the regression equation of X_1 on X_2 and X_3 can be written more simply as

$$x_1 = b_{12.3}x_2 + b_{13.2}x_3 \tag{3}$$

where $b_{12.3}$ and $b_{13.2}$ are obtained by solving simultaneously the equations

$$\begin{aligned}
\sum x_1x_2 &= b_{12.3}\sum x_2^2 + b_{13.2}\sum x_2x_3 \\
\sum x_1x_3 &= b_{12.3}\sum x_2x_3 + b_{13.2}\sum x_3^2
\end{aligned} \tag{4}$$

These equations which are equivalent to the normal equations (*2*) can be obtained formally by multiplying both sides of equation (*3*) by x_2 and x_3 successively and summing on both sides (see Problem 15.8).

REGRESSION PLANES AND CORRELATION COEFFICIENTS

If the linear correlation coefficients between variables X_1 and X_2, X_1 and X_3, and X_2 and X_3, as computed in Chapter 14, are denoted respectively by r_{12}, r_{13}, and r_{23} (sometimes called *zero-order correlation coefficients*), then the least-squares regression plane has the equation

$$\frac{x_1}{s_1} = \left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right)\frac{x_3}{s_3} \tag{5}$$

where $x_1 = X - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, and $x_3 = X_3 - \bar{X}_3$ and where s_1, s_2, and s_3 are the standard deviations of X_1, X_2, and X_3, respectively (see Problem 15.9).

Note that if the variable X_3 is nonexistent and if $X_1 = Y$ and $X_2 = X$, then equation (*5*) reduces to equation (*25*) of Chapter 14.

STANDARD ERROR OF ESTIMATE

By an obvious generalization of equation (*8*) of Chapter 14, we can define the *standard error of estimate* of X_1 on X_2 and X_3 by

$$s_{1.23} = \sqrt{\frac{\sum (X_1 - X_{1,\text{est}})^2}{N}} \tag{6}$$

where $X_{1,\text{est}}$ indicates the estimated values of X_1 as calculated from the regression equations (*1*) or (*5*).

In terms of the correlation coefficients r_{12}, r_{13}, and r_{23}, the standard error of estimate can also be computed from the result

$$s_{1.23} = s_1 \sqrt{\frac{1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{7}$$

The sampling interpretation of the standard error of estimate for two variables as given on page 313 for the case when N is large can be extended to three dimensions by replacing the lines parallel to the regression line with planes parallel to the regression plane. A better estimate of the population standard error of estimate is given by $\hat{s}_{1.23} = \sqrt{N/(N-3)}\, s_{1.23}$.

COEFFICIENT OF MULTIPLE CORRELATION

The coefficient of multiple correlation is defined by an extension of equation (*12*) or (*14*) of Chapter 14. In the case of two independent variables, for example, the *coefficient of multiple correlation* is given by

$$R_{1.23} = \sqrt{1 - \frac{s_{1.23}^2}{s_1^2}} \tag{8}$$

where s_1 is the standard deviation of the variable X_1 and $s_{1.23}$ is given by equation (*6*) or (*7*). The quantity $R_{1.23}^2$ is called the *coefficient of multiple determination*.

When a linear regression equation is used, the coefficient of multiple correlation is called the *coefficient of linear multiple correlation*. Unless otherwise specified, whenever we refer to multiple correlation, we shall imply linear multiple correlation.

In terms of r_{12}, r_{13}, and r_{23}, equation (*8*) can also be written

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}} \tag{9}$$

A coefficient of multiple correlation, such as $R_{1.23}$, lies between 0 and 1. The closer it is to 1, the better is the linear relationship between the variables. The closer it is to 0, the worse is the linear relationship. If the coefficient of multiple correlation is 1, the correlation is called *perfect*. Although a correlation coefficient of 0 indicates no linear relationship between the variables, it is possible that a *nonlinear relationship* may exist.

CHANGE OF DEPENDENT VARIABLE

The above results hold when X_1 is considered the dependent variable. However, if we want to consider X_3 (for example) to be the dependent variable instead of X_1, we would only have to replace the subscripts 1 with 3, and 3 with 1, in the formulas already obtained. For example, the regression equation of X_3 on X_1 and X_2 would be

$$\frac{x_3}{s_3} = \left(\frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}\right)\frac{x_2}{s_2} + \left(\frac{r_{13} - r_{23}r_{12}}{1 - r_{12}^2}\right)\frac{x_1}{s_1} \tag{10}$$

as obtained from equation (*5*), using the results $r_{32} = r_{23}$, $r_{31} = r_{13}$, and $r_{21} = r_{12}$.

GENERALIZATIONS TO MORE THAN THREE VARIABLES

These are obtained by analogy with the above results. For example, the linear regression equations of X_1 on X_2, X_3, and X_4 can be written

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{11}$$

and represents a *hyperplane in four-dimensional space*. By formally multiplying both sides of equation (*11*) by 1, X_2, X_3, and X_4 successively and then summing on both sides, we obtain the normal equations for determining $b_{1.234}$, $b_{12.34}$, $b_{13.24}$, and $b_{14.23}$; substituting these in equation (*11*) then gives us the *least-squares regression equation of X_1 on X_2, X_3, and X_4*. This least-squares regression equation can be written in a form similar to that of equation (*5*). (See Problem 15.41.)

PARTIAL CORRELATION

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables involved are kept constant; that is, when the effects of all other variables are removed (often indicated by the phrase "other things being equal"). This can be obtained by defining a *coefficient of partial correlation*, as in equation (*12*) of Chapter 14, except that we must consider the explained and unexplained variations that arise both with and without the particular independent variable.

If we denote by $r_{12.3}$ the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant, we find that

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

Similarly, if $r_{12.34}$ is the coefficient of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant, then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}r_{23.4}}{\sqrt{(1 - r_{13.4}^2)(1 - r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}r_{24.3}}{\sqrt{(1 - r_{14.3}^2)(1 - r_{24.3}^2)}} \tag{13}$$

These results are useful since by means of them any partial correlation coefficient can ultimately be made to depend on the correlation coefficients r_{12}, r_{23}, etc. (i.e., the *zero-order correlation coefficients*).

In the case of two variables, X and Y, if the two regression lines have equations $Y = a_0 + a_1X$ and $X = b_0 + b_1Y$, we have seen that $r^2 = a_1b_1$ (see Problem 14.22). This result can be generalized. For example, if

$$X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4 \tag{14}$$

and

$$X_4 = b_{4.123} + b_{41.23}X_1 + b_{42.13}X_2 + b_{43.12}X_3 \tag{15}$$

are linear regression equations of X_1 on X_2, X_3, and X_4 and of X_4 on X_1, X_2, and X_3, respectively, then

$$r^2_{14.23} = b_{14.23}b_{41.23} \tag{16}$$

(see Problem 15.18). This can be taken as the starting point for a definition of linear partial correlation coefficients.

RELATIONSHIPS BETWEEN MULTIPLE AND PARTIAL CORRELATION COEFFICIENTS

Interesting results connecting the multiple correlation coefficients can be found. For example, we find that

$$1 - R^2_{1.23} = (1 - r^2_{12})(1 - r^2_{13.2}) \tag{17}$$

$$1 - R^2_{1.234} = (1 - r^2_{12})(1 - r^2_{13.2})(1 - r^2_{14.23}) \tag{18}$$

Generalizations of these results are easily made.

NONLINEAR MULTIPLE REGRESSION

The above results for linear multiple regression can be extended to nonlinear multiple regression. Coefficients of multiple and partial correlation can then be defined by methods similar to those given above.

Solved Problems

REGRESSION EQUATIONS INVOLVING THREE VARIABLES

15.1 Using an appropriate subscript notation, write the regression equations of (*a*), X_2 on X_1 and X_3; (*b*) X_3 on X_1, X_2, and X_4; and (*c*) X_5 on X_1, X_2, X_3, and X_4.

SOLUTION

(*a*) $X_2 = b_{2.13} + b_{21.3}X_1 + b_{23.1}X_3$

(*b*) $X_3 = b_{3.124} + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.12}X_4$

(*c*) $X_5 = b_{5.1234} + b_{51.234}X_1 + b_{52.134}X_2 + b_{53.124}X_3 + b_{54.123}X_4$

15.2 Write the normal equations corresponding to the regression equations (*a*) $X_3 = b_{3.12} + b_{31.2}X_1 + b_{32.1}X_2$ and (*b*) $X_1 = b_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$.

SOLUTION

(*a*) Multiply the equation successively by 1, X_1, and X_2, and sum on both sides. The normal equations are

$$\begin{aligned}
\sum X_3 &= b_{3.12}N + b_{31.2}X_1 + b_{32.1}\sum X_2 \\
\sum X_1X_3 &= b_{3.12}\sum X_1 + b_{31.2}\sum X_1^2 + b_{32.1}\sum X_1X_2 \\
\sum X_2X_3 &= b_{3.12}\sum X_2 + b_{31.2}\sum X_1X_2 + b_{32.1}\sum X_2^2
\end{aligned}$$