The Omission or Addition of an Independent Variate in Multiple Linear Regression
Author(s): W. G. Cochran
Source: *Supplement to the Journal of the Royal Statistical Society*, Vol. 5, No. 2 (1938), pp. 171-176
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983654
Accessed: 25-06-2018 07:19 UTC

The Omission or Addition of an Independent Variate in Multiple Linear Regression

By W. G. Cochran

§ 1. *Introduction*

If y is the dependent variate and $x_1, x_2, \ldots x_r$ are the independent variates, the equations to determine the linear regression coefficients $b_1, b_2, \ldots b_r$ of y on $x_1, x_2, \ldots x_r$ are

$$\left.\begin{array}{l} b_1S(x_1^2) + b_2S(x_1x_2) + \ldots + b_rS(x_1x_r) = S(x_1y) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ b_1S(x_rx_1) + b_2S(x_rx_2) + \ldots + b_rS(x_r^2) = S(x_ry) \end{array}\right\} \quad . \quad (1)$$

In solving these equations, Fisher (1) has suggested that a set of auxiliary quantities $c_{pq}(p, q = 1, 2, \ldots r)$ should first be obtained. The quantities $c_{p1}, c_{p2}, \ldots c_{pr}$ are the solutions of the above equations with the right-hand side of the p^{th} equation replaced by 1, and the right-hand sides of the other equations by 0. The regression coefficients are obtained from the c's by means of the relations

$$b_i = \sum_{q=1}^{r} c_{iq}S(x_qy) \qquad i = 1, 2, \ldots r \quad . \quad . \quad . \quad (2)$$

To students carrying out a regression analysis for the first time, this procedure has sometimes seemed, as indeed it is, a somewhat roundabout method of determining the regression coefficients. The values of the b's alone, however, provide a very incomplete picture of the relationship between y and $x_1, \ldots x_r$; they do not show which of the independent variates are significantly related to the dependent variate, nor can limits be assigned from them within which the true values of the regression coefficients are likely to lie. When these points are realized, the convenience of Fisher's method may be appreciated, for the estimated standard error of b_i has been shown to be $s\sqrt{c_{ii}}$ (where s is the estimated standard error of a single observation), and is readily obtainable if the c's have been found.

Other properties of the c's which may sometimes be useful have been pointed out by Fisher. (1) The mean covariance of b_1 and b_2 is s^2c_{12}. Thus the standard error of the sum or difference of two regression coefficients may be obtained. This will be required if, for instance, independent variates such as maximum and minimum temperature are being replaced by mean temperature and range of temperature after the regression equations have been solved. (2) If the regressions of a number of dependent variates on the same set of independent variates are being examined, the c's remain the same throughout and serve for the determination of all regression coefficients. (3) It frequently happens that no apparent relation is found between the dependent variate and one or more of the independent variates. When this is the case, it is sometimes desirable to omit such variates from the regression equations. Knowing the c's, this may be done without the labour of re-solving the regression equations with the superfluous variates omitted. Fisher (1) has given formulæ for the adjustments required in the regression coefficients, and the corresponding adjustments in the c's are easily found. In this note the process will be reversed to show how to add a new independent variate to the equations without re-solving them.

§ 2. *The Omission of an Independent Variate*

The new regression coefficients, with the variate x_r omitted from the regression, will be denoted by $b'_1, b'_2, \ldots b'_{r-1}$.

The $(r - 1)$ equations satisfied by these are

$$\left.\begin{array}{l} b'_1S(x_1{}^2) + b'_2S(x_1x_2) + \ldots + b'_{r-1}S(x_1x_{r-1}) = S(x_1y) \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ b'_1S(x_{r-1}x_1) + b'_2S(x_{r-1}x_2) + \ldots + b'_{r-1}S(x_{r-1}{}^2) = S(x_{r-1}y) \end{array}\right\} \tag{3}$$

By subtracting the corresponding equation of set (1) from each of the above equations, the following equations are obtained:

$$\left.\begin{array}{l} \delta b_1S(x_1{}^2) + \delta b_2S(x_1x_2) + \ldots + \delta b_{r-1}S(x_1x_{r-1}) - b_rS(x_1x_r) = 0 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \delta b_1S(x_{r-1}x_1) + \delta b_2S(x_{r-1}x_2) + \ldots \\ \qquad\qquad + \delta b_{r-1}S(x_{r-1}{}^2) - b_rS(x_{r-1}x_r) = 0 \end{array}\right\} \tag{4}$$

where $\delta b_1 = b'_1 - b_1$ is the adjustment in b_1 produced by the elimination of x_r. From these equations we may determine the ratios $\delta b_1/b_r$, $\delta b_2/b_r \ldots$ The equations are, however, the same as the first $(r - 1)$ equations satisfied by $c_{1r}, c_{2r}, \ldots c_{rr}$, with δb_1 in place of c_{1r}, etc. and $-b_r$ in place of c_{rr}.

Hence $$\delta b_1/(-b_r) = c_{1r}/c_{rr} \quad . \quad . \quad . \quad . \quad . \tag{5}$$

that is $$\delta b_1 = b'_1 - b_1 = -(c_{1r}/c_{rr})b_r \quad . \quad . \quad . \tag{6}$$

as given by Fisher (1).

A similar treatment of the equations for the c''s and c's gives the results

$$\delta c_{11} = c'_{11} - c_{11} = -(c_{1r}{}^2/c_{rr}) \quad . \quad . \quad . \tag{7}$$

$$\delta c_{12} = c'_{12} - c_{12} = -c_{1r}c_{2r}/c_{rr} \quad . \quad . \quad . \tag{8}$$

Equations (6), (7) and (8) provide all the necessary adjustments to form the new b's and c's. If only the b''s and their standard errors are required, the non-diagonal c''s need not be found. The elimination of two variates is best carried out in two stages.

Where only a single independent variate is eliminated, this method is quicker than re-solving the regression equations, except when there are only two independent variates in the first instance. If two variates are being eliminated, the method is quicker if the original number of variates is six or more, and probably also with five variates.

§ 3. *The Addition of an Independent Variate*

If x_n is the new variate, the first step is to calculate $S(x_1 x_n)$, . . . $S(x_n^2)$ and $S(x_n y)$. Let dashes again denote the *new* coefficients. Regarding the original equations as obtained by eliminating x_n from the new equations, the adjustment equations (6), (7) and (8) become respectively

$$\delta b_1 = b'_1 - b_1 = + (c'_{1n}/c'_{nn})b'_n \quad . \quad . \quad . \quad . \quad . \quad . \quad (9)$$

$$\delta c_{11} = c'_{11} - c_{11} = + c'^2_{1n}/c'_{nn} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

$$\delta c_{12} = c'_{12} - c_{12} = + (c'_{1n}\, c'_{2n})/c'_{nn}, \text{ etc.} \quad . \quad . \quad . \quad . \quad (11)$$

These equations may be used to adjust all the existing coefficients; it is, however, first necessary to know the values of c'_{1n} . . . c'_{nn} and b'_n.

By writing down the equations satisfied by $c_{11}, c_{12}, \ldots c_{1r}$ and subtracting from each the corresponding equation satisfied by $c'_{11}, c'_{12} \ldots c'_{1n}$, we obtain the equations

$$\left.\begin{array}{r} \delta c_{11}S(x_1^2) + \delta c_{12}S(x_1x_2) + \ldots \qquad\qquad \\ + \delta c_{1r}S(x_1x_r) = -\, c'_{1n}S(x_1\, x_n) \\ . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ \delta c_{11}S(x_rx_1) + \delta c_{12}S(x_rx_2) + \ldots \qquad\qquad \\ + \delta c_{1r}S(x_r^2) = -\, c'_{1n}S(x_rx_n) \end{array}\right\} \quad (12)$$

These equations are, however, the same as the original equations for $b_1, \ldots b_r$ with $-\, c'_{1n}S(x_qx_n)$ in place of $S(x_qy)$ on the right-hand side. Hence

$$\delta c_{1p} = -\, c'_{1n}\sum_{q=1}^{r} c_{pq}S(x_qx_n) \qquad p = 1, 2, \ldots r \quad . \quad . \quad (13)$$

Hence by equations (11)

$$c'_{pn}/c'_{nn} = -\sum_{q=1}^{r} c_{pq}S(x_qx_n). \quad . \quad . \quad . \quad . \quad . \quad (14)$$

The last of the equations satisfied by c'_{nn} is

$$c'_{1n}S(x_nx_1) + \ldots + c'_{rn}S(x_nx_r) + c'_{nn}S(x_n^2) = 1 \quad . \quad . \quad . \quad (15)$$

By substituting from (14) for c'_{1n} . . . in terms of c'_{nn}, we get

$$c'_{nn}[S(x_n^2) - \sum_{p,q=1}^{r} c_{pq}S(x_px_n)S(x_qx_n)] = 1 \quad . \quad . \quad . \quad . \quad (16)$$

Equations (14) and (16) give c'_{1n} . . . c'_{nn}. We may then find b'_n from the usual relations between the b''s and the c''s, and hence adjust all the other b's and c's.

This process is in all cases more expeditious than re-solving the equations. The arrangement of the computations is best illustrated by a numerical example.

§ 4. *Example of the Addition of an Independent Variate*

In a study of the effects of weather factors on the numbers of noctuid moths per night caught in a light trap, regressions were worked out on the minimum night temperature, the maximum temperature of the previous day, the average speed of the wind during the night and the amount of rain during the night. The dependent variable was log (number of moths + 1). This was found to be roughly normally distributed, whereas the numbers themselves had an extremely skew distribution. Further, a change in one of the weather factors was likely to produce the same *percentage* change at different times in the numbers of moths rather than the same *actual* change. Three years' data were included. These were grouped in blocks of nine consecutive days, so as to eliminate as far as possible the effects of the lunar cycle. After the removal of differences between blocks, 72 degrees of freedom remained for the regressions.

The regression coefficients and their standard errors in convenient working units are as follows:

Min. Temp.	Max. Temp.	Wind	Rain
0·1981407 ±0·0650	0·0385284 ±0·0588	−0·5086492 ±0·1515	+0·0318482 ±0·0499

The analysis of variance is shown below:

TABLE I

	D.F.	Sums of Squares	Mean Squares
Regression	4	0·8274	0·2068
Deviations	68	2·7245	0·04007
Total	72	3·5519	0·04933

It was subsequently decided to investigate the effect of cloudiness, measured on a conventional scale as the percentage of starlight obscured by clouds in a night sky camera.

The calculations are shown in Table II. The original c's are first written down, and the corresponding sums of products of each variate with the new variate are placed in the right-hand column. The sum of products of each column with the right-hand column is placed at the foot of the column, *with the signs reversed.* By equations (14), these values are c_{15}'/c_{55}' . . . .

The sum of the products of these numbers with the corresponding numbers in the right-hand column is then calculated. The sum of

TABLE II

Addition of an Independent Variate

	Min. Temp. (1)	Max. Temp. (2)	Wind (3)	Rain (4)	Sums of Products with Cloud
		c_{pq}			$S(x_p x_5)$
(1)	+0·10542356	—0·04194620	—0·09606709	—0·01849096	—4·867
(2)	—0·04194620	+0·08603869	+0·03317271	+0·01290358	+0·206
(3)	—0·09606709	+0·03317271	+0·57265201	+0·00811662	—0·5446
(4)	—0·01849096	+0·01290358	+0·00811662	+0·06227532	—5·42
(5)	—	—	—	—	+7·87
		$c_{p5}'/c_{55}' = -\Sigma c_{pq} S(x_q x_5)$			c_{55}'
	+0·36919824	—0·13387286	—0·11853374	+0·24929891	+0·21013314
		$S(x_{py})$			
	+2·0744	+1·5747	—0·6440	+0·885	—1·933
	b_1	b_2	b_3	b_4	
	+0·1981407	+0·0385284	—0·5086492	+0·0318482	—
	b_1'	b_2'	b_3'	b_4'	b_5'
	+0·1142775 ±0·0704	+0·0689376 ±0·0576	—0·4817243 ±0·1459	—0·0247799 ±0·0528	—0·2271496 ±0·0882
		c_{pq}'			
(1)	+0·13406625	—0·05233216	—0·10526303	+0·00084984	+0·07758079
(2)	—	+0·08980468	+0·03650720	+0·00589052	—0·02813112
(3)	—	—	+0·57560443	+0·00190712	—0·02490787
(4)	—	—	—	+0·07533508	+0·05238596
(5)	—	—	—	—	+0·21013314

squares of the new variate (7·87) is added on the calculating machine. By equation (16) the reciprocal of the total is c_{55}' (0·21013314).

The regression coefficient b_5' may now be found. Since

$$b_5' = c_{15}'S(x_1y) + \ldots + c_{55}'S(x_5y) \quad . \quad . \quad (17)$$

$$b_5' = \{(c_{15}'/c_{55}')S(x_1y) + \ldots + (c_{45}'/c_{55}')S(x_4y) + S(x_5y)\} \times c_{55}' \quad (18)$$

$$= \{0{\cdot}36919824 \times 2{\cdot}0744 + \ldots \quad - 1{\cdot}933\} \times (0{\cdot}21013314)$$

$$= - {\cdot}2271496$$

which is obtained on the machine without any intermediate writing down.

At this stage the significance of the coefficient b_5' may be tested; if the new variate has no apparent effect, it may not be worth while to complete the calculations. The reduction in the sum of squares due to cloud is $b_5'^2/c_{55}' = 0{\cdot}2455$. From Table I the residual mean square (67 degrees of freedom) is found to be 0·03700, so that b_5' is definitely significant.

The calculations are completed by means of the adjustment equations (9), (10) and (11). In particular

$$b_1' = 0{\cdot}1981407 + (0{\cdot}36919824) \times (-0{\cdot}2271496) = 0{\cdot}1142775.$$

$$c_{15}' = (0{\cdot}36919824) \times (0{\cdot}21013314) = 0{\cdot}07758079.$$

$$c_{11}' = 0{\cdot}10542356 + (0{\cdot}36919824) \times (0{\cdot}07758079) = 0{\cdot}13406625.$$

$$c_{12}' = -0{\cdot}04194620 + (0{\cdot}36919824) \times (-0{\cdot}02813112) = -0{\cdot}05233216$$

In the last two cases the combined use of the ratios c_{15}'/c_{55}' . . . and the values c_{15}' . . . gives the adjustment terms in a single multiplication.

As a final check on the calculations, b_1', . . . b_5' should be substituted in the regression equations. The c''s may then be checked by verifying that the b''s obtained from the c''s in the usual way agree with the values already found. An intermediate check on the values c_{15}'/c_{55}' . . . may also be obtained by adding the four c's in each row and calculating the sum of the products of the totals with the values $S(x_1x_5)$. . . . This, with its sign reversed, is equal to

$$0{\cdot}36919824 - 0{\cdot}13387286 - 0{\cdot}11853374 + 0{\cdot}24929891.$$

The number of decimal places carried in the above calculation is excessive, though it facilitates the detection of errors when the final substitution in the regression equations is made. Six decimal places would have been sufficient in ordinary work.

Reference

Fisher, R. A., "Statistical Methods for Research Workers." Oliver and Boyd, Edinburgh, 6th Ed., 1936, §§ 29, 29.1.