# THE SIMULTANEOUS COMPUTATION OF GROUPS OF REGRESSION EQUATIONS AND ASSOCIATED MULTIPLE CORRELATION COEFFICIENTS

By Paul S. Dwyer

1. **Introduction.** The need sometimes arises for the prediction of a number of different variables from a given group of so-called fundamental variables. In the work of college prediction, for example, one might desire regression equations predicting certain measures of college achievement (e.g., first semester average, first semester English grade, first semester mathematics grade, number of hours of *A* received during first semester, etc.) on the basis of a number of other factors (e.g., high school record, score on American Council on Education Psychological Examination, score on some standard English achievement test, score on some standard mathematics achievement test, etc.). It is the purpose of this paper to show how the regression coefficients and the associated multiple correlation coefficients can be obtained simultaneously. The essence of the method is a simple device by which one solution of general normal equations may be made to serve for all cases.

2. **The normal equations.** Let $x_1, x_2, x_3, \cdots x_n$, be the so-called fundamental variables and let $x_k$ be the predicted variable. The normal equations are computed by standard methods which result in one of the three types.

Type I. Normal equations for determining $b_0, b_1, b_2, b_3, \cdots, b_n$.

$$b_0 n + b_1\Sigma x_1 + b_2\Sigma x_2 + b_3\Sigma x_3 + \cdots\cdots\cdots + b_n\Sigma x_n - \Sigma x_k = 0$$

$$b_0\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + b_3\Sigma x_1x_3 + \cdots\cdots\cdots + b_n\Sigma x_1x_n - \Sigma x_1x_k = 0$$

$$b_0\Sigma x_2 + b_1\Sigma x_1x_2 + b_2\Sigma x_2^2 + b_3\Sigma x_2x_3 + \cdots\cdots\cdots + b_n\Sigma x_2x_n - \Sigma x_2x_k = 0$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$b_0\Sigma x_n + b_1\Sigma x_nx_1 + b_2\Sigma x_nx_2 + b_3\Sigma x_nx_3 + \cdots\cdots\cdots + b_n\Sigma x_n^2 - \Sigma x_nx_k = 0$$

Type II. Normal equations for determining $b_1, b_2, b_3, \cdots, b_n$.

$$\bar{x}_i = x_i - M_{x_i}$$

$$b_1\Sigma\bar{x}_1^2 + b_2\Sigma\bar{x}_1\bar{x}_2 + b_3\Sigma\bar{x}_1\bar{x}_3 + \cdots\cdots\cdots\cdots\cdots + b_n\Sigma\bar{x}_1\bar{x}_n - \Sigma\bar{x}_1\bar{x}_k = 0$$

$$b_1\Sigma\bar{x}_2\bar{x}_1 + b_2\Sigma\bar{x}_2^2 + b_3\Sigma\bar{x}_2\bar{x}_3 + \cdots\cdots\cdots\cdots\cdots + b_n\Sigma\bar{x}_2\bar{x}_n - \Sigma\bar{x}_2\bar{x}_k = 0$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$b_1\Sigma\bar{x}_n\bar{x}_1 + b_2\Sigma\bar{x}_n\bar{x}_2 + b_3\Sigma\bar{x}_n\bar{x}_3 + \cdots\cdots\cdots\cdots\cdots + b_n\Sigma\bar{x}_n^2 - \Sigma\bar{x}_n\bar{x}_k = 0$$

Type III. Normal equations for determining $\beta_1, \beta_2, \beta_3, \cdots, \beta_n$.

$$\begin{aligned}
&\beta_1 + r_{12}\beta_2 + r_{13}\beta_3 + \cdots + r_{1n}\beta_n - r_{1k} = 0\\
&r_{21}\beta_1 + \beta_2 + r_{23}\beta_3 + \cdots + r_{2n}\beta_n - r_{2k} = 0\\
&\cdots\cdots\cdots\cdots\\
&r_{n1}\beta_1 + r_{n2}\beta_2 + r_{n3}\beta_3 + \cdots + r_{nn}\beta_n - r_{nk} = 0
\end{aligned}$$

The three types are special cases of the general

$$\begin{aligned}
&d_{11}y_1 + d_{12}y_2 + d_{13}y_3 + \cdots + d_{1j}y_j + \cdots + d_{1n}y_n - d_{1k} = 0\\
&d_{21}y_1 + d_{22}y_2 + d_{23}y_3 + \cdots + d_{2j}y_j + \cdots + d_{2n}y_n - d_{2k} = 0\\
&d_{31}y_1 + d_{32}y_2 + d_{33}y_3 + \cdots + d_{3j}y_j + \cdots + d_{3n}y_n - d_{3k} = 0\\
&\cdots\cdots\cdots\cdots\\
&d_{i1}y_1 + d_{i2}y_2 + d_{i3}y_3 + \cdots + d_{ij}y_j + \cdots + d_{in}y_n - d_{ik} = 0\\
&\cdots\cdots\cdots\cdots\\
&d_{n1}y_1 + d_{n2}y_2 + d_{n3}y_3 + \cdots + d_{nj}y_j + \cdots + d_{nn}y_n - d_{nk} = 0
\end{aligned}$$

where $y_j$ are the regression coefficients and $d_{ij} = d_{ji}$.

The methods described in this paper are applicable to the general case and hence to each of the three particular types.

In examining the normal equations, it is noticed that the first $n$ terms of each equation are completely determined by the $n$ fundamental variables. The equations, aside from the last terms, are identical no matter what variable is predicted. It is only necessary to devise a technique for separating the contributions of the $d_{ik}$ terms.

3. **Solution by determinants.** One method utilizes determinants. The value $y_j$ is expressed in terms of a determinant involving a column with entries $d_{1k}, d_{2k}, d_{3k}, \cdots, d_{nk}$. The determinant is expanded in terms of this column.

Specifically, let $D$ be the determinant of the coefficients of the $y_j$ and let $D_{ij}$ be the cofactor of any element $d_{ij}$ of $D$. Then

$$D = \sum_{i=1}^{n} D_{ij}\, d_{ij}$$

and

$$\begin{aligned}
y_1 &= \frac{1}{D}\,(D_{11}\,d_{1k} + D_{21}\,d_{2k} + D_{31}\,d_{3k} + \cdots + D_{j1}\,d_{jk} + \cdots + D_{n1}\,d_{nk}.)\\
y_2 &= \frac{1}{D}\,(D_{12}\,d_{1k} + D_{22}\,d_{2k} + D_{32}\,d_{3k} + \cdots + D_{j2}\,d_{jk} + \cdots + D_{n2}\,d_{nk}.)\\
&\cdots\cdots\cdots\cdots\\
y_i &= \frac{1}{D}\,(D_{1i}\,d_{1k} + D_{2i}\,d_{2k} + D_{3i}\,d_{3k} + \cdots + D_{ji}\,d_{jk} + \cdots + D_{ni}\,d_{nk}.)\\
&\cdots\cdots\cdots\cdots\\
y_n &= \frac{1}{D}\,(D_{1n}\,d_{1k} + D_{2n}\,d_{2k} + D_{3n}\,d_{3k} + \cdots + D_{jn}\,d_{jk} + \cdots + D_{nn}\,d_{nk}.)
\end{aligned}$$

It is only necessary to compute $\frac{D_{ji}}{D}$ to find the coefficient of $d_{jk}$ in the expansion of $y_i$.

An illustration is given. The normal equations are

$$\beta_1 + .3300\,\beta_2 + .2100\,\beta_3 - r_{1k} = 0$$
$$.3300\,\beta_1 + \beta_2 - .4800\,\beta_3 - r_{2k} = 0$$
$$.2100\,\beta_1 - .4800\,\beta_2 + \beta_3 - r_{3k} = 0$$

from which at once

$$\beta_1 = \frac{1}{D}(.7696\,r_{1k} - .4308\,r_{2k} - .3684\,r_{3k})$$

$$\beta_2 = \frac{1}{D}(-.4308\,r_{1k} + .9559\,r_{2k} + .5493\,r_{3k})$$

$$\beta_3 = \frac{1}{D}(-.3684\,r_{1k} + .5493\,r_{2k} + .8911\,r_{3k})$$

and also

$$\begin{aligned} D = .550072 &= (1.00)(.7696) + (.33)(-.4308) + (.21)(-.3684) \\ &= (.33)(-.4308) + (\ 1.00)(.9559) + (-.48)(.5493) \\ &= (.21)(-.3684) + (-.48)(.5493) + (\ 1.00)(.8911) \end{aligned}$$

so that

$$\beta_1 = \ 1.3991\,r_{1k} - \ .7832\,r_{2k} - \ .6697\,r_{3k}.$$
$$\beta_2 = -.7832\,r_{1k} + 1.7378\,r_{2k} + \ .9986\,r_{3k}.$$
$$\beta_3 = -.6697\,r_{1k} + \ .9986\,r_{2k} + 1.6200\,r_{3k}.$$

It is only necessary to insert any given values $r_{1k}$, $r_{2k}$, $r_{3k}$, to obtain the coefficients of any specific regression equation.

4. **Solutions without determinants.** Theoretically the solution by determinants is excellent but as the number of variables increases the work of computing the $n^2$ cofactors $\left[\text{or the } \frac{n(n+1)}{2} \textit{ different} \text{ cofactors}\right]$ becomes enormous. We desire a technique for separating the contributions of the last terms when determinants are not used. This can be accomplished by using a separate column for each $d_{ik}$. Before algebraic manipulation, the value $d_{ik}$ is factored from the column and, after manipulative solution is complete, the multiplication by $d_{ik}$ is carried out.

As an example consider the normal equations

$$\beta_1 + r_{12}\beta_2 - r_{1k} = 0$$

$$r_1\beta_{21} + \beta_2 - r_{2k} = 0$$

where $r_{12} = r_{21} = .3300$. Then the normal equations may be represented by rows (1) and (2) of Table I.

TABLE I

| Row | Operation | $\beta_1$ | $\beta_2$ | $r_{1k}$ | $r_{2k}$ |
|---|---|---|---|---|---|
| (1) | | 1.0000 | .3300 | −1.0000 | |
| (2) | | .3300 | 1.0000 | | −1.0000 |
| (3) | − .3300 times (2) | − .1089 | − .3300 | | .3300 |
| (4) | (1) + (3) | .8911 | | −1.0000 | .3300 |
| (5) | − (4) divided by .8911 | −1.0000 | | 1.1222 | − .3703 |
| (6) | − .3300 times (5) | .3300 | | − .3703 | .1222 |
| (7) | − (2) + (6) | | −1.0000 | − .3703 | 1.1222 |

The four decimal place solution, whose steps are indicated by (3) (4) (5) (6)(7), is from (5) and (7)

$$\beta_1 = 1.1222\, r_{1k} - .3703\, r_{2k}$$

$$\beta_2 = -.3703\, r_{1k} + 1.1222\, r_{2k}$$

This device may be combined with most of the standard methods of solving normal equations.

5. **Combination with Doolittle method.** Especially to be recommended is a combination of this device with the Doolittle method which is recognized as a most efficient method of solving normal equations in from five to ten variables [1] [2]. One of the advantages of the Doolittle method is that related multiple regression coefficients may be obtained from the same forward solution, though additional back solutions are necessary [3].

The problem which led to the development of this technique was the simultaneous prediction of scores on various occupations covered by the Strong Vocational Interest Blank from the scores on a few fundamental occupations. A multiple factor analysis revealed that five basic factors account for most of the scores. Five occupational scores, serving as approximations to the five basic factors, were used as the fundamental variables and the other scores were predicted from them.

As an illustration of this prediction technique combined with the Doolittle method, I have selected three test scores as fundamental since the solution based on them shows all the steps of the Doolittle method and is shorter than the five

variable problem. Actually, solution by determinants (section 3) is advised for problems involving three variables. The steps of the Doolittle solution are presented in Table II. The results should be compared with those of the determinant solution of section 3.

The first column indicates the row and the second the description of the algebraic operation. The next three columns are the standard columns of a Doolittle presentation with the conventional elimination of the lower left entries. The next three columns carry through the Doolittle method with the values $r_{1k}$, $r_{2k}$, $r_{3k}$ kept in separate columns. The last column is an adaptation of the conventional summary check column of the Doolittle solution.

TABLE II

*Generalized Doolittle Presentation*

| Row | Operation | $\beta_1$ | $\beta_2$ | $\beta_3$ | $r_{1k}$ | $r_{2k}$ | $r_{3k}$ | $S$ |
|---|---|---|---|---|---|---|---|---|
| (1) | | 1.0000 | .3300 | .2100 | −1.0000 | | | .5400 |
| (2) | | .3300 | 1.0000 | −.4800 | | −1.0000 | | −.1500 |
| (3) | | .2100 | −.4800 | 1.0000 | | | −1.0000 | −.2700 |
| (4) | Repeat (1) | 1.0000 | .3300 | .2100 | −1.0000 | | | .5400 |
| (5) | Negative of (4) | −1.0000 | −.3300 | −.2100 | 1.0000 | | | −.5400 |
| (6) | Repeat (2) | | 1.0000 | −.4800 | | −1.0000 | | −.1500 |
| (7) | −.3300 times (4) | | −.1089 | −.0693 | .3300 | | | −.1782 |
| (8) | (6) + (7) | | .8911 | −.5493 | .3300 | −1.0000 | | −.3282 |
| (9) | −(8) divided by .8911 | | −1.0000 | .6164 | −.3703 | 1.1222 | | .3683 |
| (10) | Repeat (3) | | | 1.0000 | | | −1.0000 | −.2700 |
| (11) | −.2100 times (4) | | | −.0441 | .2100 | | | −.1134 |
| (12) | .6164 times (8) | | | −.3386 | .2034 | −.6164 | | −.2023 |
| (13) | (10) + (11) + (12) | | | .6173 | .4134 | −.6164 | −1.0000 | −.5857 |
| (14) | −(13) divided by .6173 | | | −1.0000 | −.6697 | .9985 | 1.6200 | .9488 |
| (15) | .6164 times (14) | | | −.6164 | −.4128 | .6155 | *.9986* | .5848 |
| (16) | (9) + (15) | | −1.0000 | | −.7831 | 1.7377 | *.9986* | .9531 |
| (17) | −.2100 times (14) | | | .2100 | .1406 | *−.2097* | *−.3402* | −.1992 |
| (18) | −.3000 times (16) | | .3300 | | .2584 | *−.5734* | *−.3295* | −.3145 |
| (19) | (5) + (17) + (18) | −1.0000 | | | 1.3990 | *−.7831* | *−.6697* | −1.0537 |

The general solution is read from rows (19) (16) (14) and is

$$\beta_1 = 1.3990\, r_{1k} - .7831\, r_{2k} - .6697\, r_{3k}.$$

$$\beta_2 = -.7831\, r_{1k} + 1.7377\, r_{2k} + .9986\, r_{3k}.$$

$$\beta_3 = -.6697\, r_{1k} + .9985\, r_{2k} + 1.6200\, r_{3k}.$$

which agrees, aside from the last place, with the result of the solution by determinants.

It is wise to check in the original equations (1), (2), (3) as soon as any $\beta_i$ is found. Row (14), for example, should be checked by showing

$$(-.6697)(1.00) + (.9985)(\ .33) + (1.6200)(\ .21) = \ .0000$$

$$(-.6697)(\ .33) + (.9985)(\ 1.00) + (1.6200)(-.48) = -.0001$$

$$(-.6697)(\ .21) + (.9985)(-.48) + (1.6200)(\ 1.00) = \ 1.0001$$

The same should be done with row (16) as soon as it is computed. Row (19) should be treated similarly.

6. **Many regression equations.** If large numbers of regression equations are to be generated (the Strong Vocational Interest Study had 29 dependent variables), the following technique is suggested. Make a table with columns $r_{1k}$, $r_{2k}$, etc. and use the rows to indicate the different values of $k$. On another slip of paper insert the general values $\beta_1$, $\beta_2$, $\beta_3$, $\cdots$ $\beta_n$ in successive rows so that a folding of the paper will bring any general $\beta$ expansion in conjunction with the $r$'s of any test, $k$. The scheme is illustrated in Table III.

TABLE III

| No. | Occupation | $r_{1k}$ | $r_{2k}$ | $r_{3k}$ | | $\beta_{1k}$ | $\beta_{2k}$ | $\beta_{3k}$ | | $r$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Teacher | 1.00 | .33 | .21 | | 1.00 | .00 | .00 | | 1.00 |
| 2 | Physicist | .33 | 1.00 | −.48 | | .00 | 1.00 | .00 | | 1.00 |
| 3 | Office Worker | .21 | −.48 | 1.00 | | .00 | .00 | 1.00 | | 1.00 |
| 4 | Doctor | .17 | .79 | −.52 | | −.03 | .72 | −.17 | | .81 |
| 5 | Lawyer | −.02 | .16 | −.59 | | .24 | −.30 | −.78 | | .64 |
| 6 | Engineer | .16 | .78 | −.02 | | −.37 | 1.21 | .64 | | .93 |
| | $\beta_1$ | 1.3990 | −.7831 | −.6697 | | ↑ 1.0000 | | | | |
| | $\beta_2$ | −.7831 | 1.7377 | .9986 | | | ↑ 1.0000 | | | |
| | $\beta_3$ | −.6697 | .9985 | 1.6200 | | | | ↑ 1.0000 | | |
| 10 | Mathematician etc. | .46 | .96 | −.49 | | .19 | .82 | −.14 | | .97 |

Thus, for the occupation of Engineer,

$$\beta_1 = \ 1.3990\ (.16) + (-.7831)(.78) + (-.6697)(-.02) = -.37$$

$$\beta_2 = -.7831\ (.16) + (\ 1.7377)(.78) + (\ .9996)(-.02) = \ 1.21$$

$$\beta_3 = -.6697\ (.16) + (\ .9985)(.78) + (\ 1.6200)(-.02) = \ .64$$

The value of the multiple correlation coefficient is then computed from the formula

$$r_{k.123\cdots n} = \sqrt{\beta_{1k} r_{1k} + \beta_{2k} r_{2k} + \cdots + \beta_{nk} r_{nk}}$$

In the illustration above

$$r_{k.123} = \sqrt{(-.37)(.16) + (1.21)(.78) + (.64)(-.02)}$$
$$= .93$$

7. **Regression equations by deletion.** The method of getting related regression coefficients and correlation coefficients, described by Kurtz [3], is also applicable. Again, a problem involving more than three variables is needed to show the real value of the scheme but the technique may be illustrated in the three variable case. We wish to find, from the forward solution of Table II, the regression equation and the multiple correlation coefficient when the first two fundamental variables only are used. We delete all columns involving test 3 and complete the back solution as indicated in Table IV, which may be viewed as a substitute for the last ten rows of Table II.

TABLE IV
(See Table II)

| Row | Operation | $\beta_1$ | $\beta_2$ | $\beta_3$ | $r_{1k}$ | $r_{2k}$ | $r_{3k}$ | $S$ |
|---|---|---|---|---|---|---|---|---|
| (20) | Repeat (9) | | −1.0000 | | −.3703 | 1.1222 | | |
| (21) | −.3300 times (20) | | .3300 | | .1222 | −.3703 | | |
| (22) | (5) + (21) | −1.0000 | | | 1.1222 | −.3703 | | |

The results are

$$\beta_1 = 1.1222\, r_{1k} - .3703\, r_{2k}\,.$$

$$\beta_2 = -.3703\, r_{1k} + 1.1222\, r_{2k}\,.$$

and these agree with the results of section 4.

8. **The simplified back solution.** In every case in which the $\beta$'s have been given in terms of $r$'s the matrix of the coefficients is symmetric (sections 3, 4, 5, 7). One wonders if this symmetry is generally true and if it holds for normal equations of Type I or Type II.

Determinants are much more useful in establishing general properties, such as the one under discussion, than they are in computing the values of regression coefficients in the case of a problem involving many variables. We return to the determinant notation of section 3.

In each of the three types, and hence in the general case $d_{ij} = d_{ji}$ so that $D$ is a symmetric determinant, $D_{ij} = D_{ji}$ and $\frac{D_{ij}}{D} = \frac{D_{ji}}{D}$. Hence the matrix of the coefficients of the solution is symmetric.

This result may be used (1) to check the expanded results or (2) to eliminate some of the work of the back solution. The $n$ coefficients must be recorded for $\beta_n$ after which the column indicated by $r_{nk}$ may be dropped. The first $n - 1$ coefficients must be computed for $\beta_{n-1}$ after which the column indicated by $r_{n-1,k}$ may be dropped, etc. The italicized entries in Table II are the ones which are eliminated in this way. The remaining coefficients are sufficient to completely determine the symmetric matrix.

The summary right hand check column can not be readily used in the simplified back solution but it is hardly to be recommended anyway. Kurtz [3] argues against it on the ground that it is not necessary. The essential check is to see that each $\beta$ solution satisfies all of the original equations.

9. **Conclusion.** This paper provides a technique for the computation of general regression equations and shows how the technique may be combined with the Doolittle method in providing a practical means of mass prediction.

UNIVERSITY OF MICHIGAN.

## REFERENCES

[1] TOLLEY, H. R. AND EZEKIAL, MORDECAI. *The Doolittle method for solving multiple correlation equations versus the Kelley-Salisbury Iteration method.* Journal of American Statistical Association, 1927(22), pp. 497–500.

[2] KELLEY, T. L. AND MCNEMAR, Q. *Doolittle versus Kelley-Salisbury iteration methods for computing multiple regression coefficients.* Journal of American Statistical Association, 1929 (24), pp. 164–169.

[3] KURTZ, A. K. *The use of the Doolittle method in obtaining related multiple correlation coefficients.* Psychometrika, Vol. I, no. 1, March 1936, pp. 45–51.

For historical and bibliographical references the reader is advised to consult:

GRIFFIN, H. D. *On partial correlation versus partial regression for obtaining the multiple regression equations.* Jour. of Ed. Psy. 1939(22), pp. 35–44.

GRIFFIN, H. D. *Simplified Schema for multiple linear correlation.* Jour. of Ex. Ed. Vol. I, no. 1, March 1933, pp. 239–254.