

**Cowles Foundation Paper 20**

Reprinted from

ECONOMETRICA, Journal of the Econometric Society, Vol. 15, No. 1, January, 1947  
The University of Chicago, Chicago 37, Illinois, U.S.A.

**NOTE ON THE DOOLITTLE SOLUTION\***

By NANCY BRUNER

SINCE correlation analysis is now so widely used, any short cut or time saver in the method is of importance. This note is to present a simple step which has proven of great aid in the Doolittle solution of simultaneous equations in correlation.

The Doolittle method employs a check column (hereafter referred to as the *c* column) which is most helpful in stopping the procedure when a mistake has been made. However, the *c* column frequently does not check, owing to rounding errors. The reasons for this are as follows:

1. In the Doolittle solution, equations are divided by their leading terms.

TABLE I  
COEFFICIENTS OF *m* CENTERED NORMAL EQUATIONS

Row	$x_1$	$x_2$	$x_3$	$x_4$	...	$x_m$	$x_0$	<i>c</i>
1	$a_{11}$	$a_{12}$	$a_{13}$	$a_{14}$	...	$a_{1m}$	$a_{10}$	$c_1$
2		$a_{22}$	$a_{23}$	$a_{24}$	...	$a_{2m}$	$a_{20}$	$c_2$
3			$a_{33}$	$a_{34}$	...	$a_{3m}$	$a_{30}$	$c_3$
4				$a_{44}$	...	$a_{4m}$	$a_{40}$	$c_4$
.					...	.	.	.
.					...	.	.	.
.					...	.	.	.
<i>m</i>						$a_{m,m}$	$a_{m,0}$	$c_m$

2. The greater the size of the leading term, the larger the error caused by this division step.

3. Further, the earlier such an inaccuracy comes into a solution, the greater will be its over-all effect on the checks.

A rearrangement of the block of normal equations removes this difficulty, and the *c* column check is greatly improved.

Assume that there are *m* independent variables,  $x_i$ , and  $i=1, 2, \dots, m$ , and one dependent variable,  $x_0$ . Table I shows the coefficients<sup>1</sup> of the centered normal equations and the *c* column, where

$$a_{ij} = n \sum x_i x_j, \quad c_i = n \sum x_i c,$$

and  $i = 0, 1, 2, \dots, m,$       and  $j = 0, 1, 2, \dots, m.$

The rearrangement of this block, so that the *c* column will check, is merely this: Arrange the main diagonal

\* This and the following article by Dickson H. Leavens will be reprinted as Cowles Commission Papers, New Series, No. 20.

<sup>1</sup> For simple machine computation, the matrix is of *n* times the coefficients of the centered normal equations.

$$a_{11}, a_{22}, a_{33}, a_{44}, \dots, a_{mm},$$

in an array from the smallest to largest, from top to bottom. Fill in the other coefficients in their appropriate places. This can be done since

$$a_{ij} = a_{ji}.$$

For example, Table A gives the coefficients of four centered normal equations,  $m=4$ .<sup>2</sup> At the completion of the Doolittle solution of these equations as they appear (rounding to six decimals), the  $c$  column checks to the third decimal place. However, if the suggested procedure is followed, as in Table B, the  $c$  column checks perfectly to the sixth place.

TABLE A  
COEFFICIENTS OF FOUR CENTERED NORMAL EQUATIONS

Row	$x_1$	$x_2$	$x_3$	$x_4$	$x_0$	$c$
1	207.7204	184.8486	14.7448	8.2056	6.4668	421.9862
2		3895.1755	8.8718	43.3372	204.2098	4336.4429
3			1.8340	1.1760	3.2732	29.8998
4				2.4240	4.9832	60.1260

TABLE B  
REARRANGEMENT OF COEFFICIENTS IN TABLE A

Row	$x_3$	$x_4$	$x_1$	$x_2$	$x_0$	$c$
3	1.8340	1.1760	14.7448	8.8718	3.2732	29.8998
4		2.4240	8.2056	43.3372	4.9832	60.1260
1			207.7204	184.8486	6.4668	421.9862
2				3895.1755	204.2098	4336.4429

Obviously the greater the number of equations to be solved, the more important is this step. In applied work, the routine calculations required in correlation are frequently done by clerical workers. The described rearrangement then is of particular assistance, since it saves hours of unnecessary rechecking.

*Western Auto Supply Company*

<sup>2</sup> Experience has shown that, in analyzing economic data, the coefficients often vary considerably in their respective sizes. Relative differences such as those in Table A are not uncommon.

## ACCURACY IN THE DOOLITTLE SOLUTION

By DICKSON H. LEAVENS

MISS BRUNER's note<sup>1</sup> arises from a difficulty that users of the Doolittle method of solving multiple-regression equations have often encountered, namely that of getting results to check to a satisfactory degree of accuracy. The fundamental problem is the difference between accuracy to a certain number of significant figures and accuracy to a certain number of decimal places. The data for any problem are usually given to a certain number of significant figures; for practical convenience of computation, however, it is preferable to work throughout to a fixed number of decimal places. It is desirable, therefore, to reconcile these two criteria as far as possible.

This note will discuss the relative accuracy and convenience of four methods of applying the Doolittle method, as follows:

Method A. Straight forward solution of the equations with the data of given magnitude and in given order.

Method B (Miss Bruner's). Solution with the data rearranged so that the terms in the principal diagonal increase in size.

Method C. Solution with the data in original order but adjusted so that the elements of the principal diagonal range between 0.1 and 10.0.

Method D. Solution with the data adjusted to represent coefficients of correlation between pairs of variables, thus making all the elements of the principal diagonal unity.

### METHOD A

Since in each step of the Doolittle method a whole line is divided by its first term, the larger this term, the smaller the quotient, and hence, when a fixed number of decimal places is retained, the more significant figures are lost. If the largest divisors happen to come in the earliest steps, the effect is aggravated, since each step depends on the preceding one. When the example given in Miss Bruner's Table A is worked by this method, carrying the computations to six decimal places, the terms that should be zero turn out to differ from it considerably. Since in practice these terms are not computed, the check sums that assume them to be zero do not agree with the result of operations on the *c* column, the last three decimal places being different in most steps. This is confusing to the computer, who cannot tell whether the discrepancies are due to errors or to the size of the divisors, and must waste time re-computing.

<sup>1</sup> ECONOMETRICA, Vol. 15, January, 1947, pp. 43-44.

## METHOD B

Miss Bruner's method gets around this difficulty by using the smallest divisors in the earliest steps. When the diagonal terms vary considerably in size this method is a great improvement on Method A. When Method B is applied to the same data, rearranged as in Miss Bruner's Table B, the check sums agree with the  $c$  column to six decimals and the computer is saved much needless checking.

This method has the great advantage of simplicity; the rearrangement of the data can be understood by any computer.

It has one disadvantage, however, in cases where it is desired to compute regressions with different numbers of independent variables. For example, there may be reason to believe that  $x_0$  depends primarily on  $x_1$  and  $x_2$ , but that the successive additions of  $x_3$  and  $x_4$  might improve the correlation. If the variables are arranged in the order 1, 2, 3, 4, then the results can be obtained for 1, 2; for 1, 2, 3; and for 1, 2, 3, 4; with a single forward solution, simply by starting the back solution at the proper point for each. In pioneer studies, this is frequently an important consideration, since it is impossible to foretell in advance whether additional variables will be fruitful. On the other hand, for routine problems where it is known just which variables are important, Method B, would seem to be an excellent device, giving greater accuracy than Method A and yet being much simpler for the computer than the other two methods described below.

## METHOD C

Another method attempts to get around the difficulty by adjusting the decimal points in the data by multiplying each row and column by such (positive or negative) powers of 10 as will make the diagonal terms range from 0.1 to 10.0. This method has been described by Duncan and Kenney<sup>2</sup> in matrix notation, but will here be treated in ordinary algebraic form which may be simpler for some computers.

If the regression equation is, for Miss Bruner's example of 4 independent variables,

$$(1) \quad b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 = x_0,$$

the columns of her Table A are multiplied by a set of adjustment factors, shown at the top of Table C, and the resulting rows are then multiplied by the same set, shown at the left of Table C, the final figures

<sup>2</sup> David B. Duncan and John F. Kenney, *On the Solution of Normal Equations and Related Topics*, published by John F. Kenney, University of Wisconsin in Milwaukee, 1946, 35 pp., esp. section 14, pp. 30-32.

being shown in the body of the same table. These new data may then be considered as representing a new equation,

$$(2) \quad b_1'x_1' + b_2'x_2' + b_3'x_3' + b_4'x_4' = x_0',$$

or

$$(3) \quad b_1'(0.1x_1) + b_2'(0.01x_2) + b_3'(1x_3) + b_4'(1x_4) = 0.1x_0,$$

where the  $x_i$  represent the adjusted data and the  $b_i$  the corresponding coefficients. Since the data are now adjusted to have approximately the

TABLE C

COEFFICIENTS OF TABLE A ADJUSTED TO APPROXIMATELY SAME MAGNITUDE

Adjust- ment factors		0.1	0.01	1.	1.	0.1	
		$x_1'$	$x_2'$	$x_3'$	$x_4'$	$x_0'$	$c'$
0.1	$x_1'$	2.077204	0.184849	1.474480	0.820560	0.064668	4.621761
0.01	$x_2'$		0.389518	0.088718	0.433372	0.204210	1.300667
1.	$x_3'$			1.834000	1.176000	0.327320	4.900518
1.	$x_4'$				2.424000	0.498320	5.352252
Readjust- ment fac- tors for coefficients		1	0.1	10.	10.	1	

same number of significant figures, working to a constant number of decimal places will give about the same results as working to a constant number of significant figures.

When the equations are solved for the  $b_i'$ , these must be converted back to the  $b_i$  to fit the decimal pointing of the original data. To find the conversion factors, equation (3) is multiplied by the power of 10, in this case  $10^1$ , that will make the coefficient of  $x_0$  equal to unity, giving

$$(4) \quad b_1'x_1 + 0.1b_2'x_2 + 10b_3'x_3 + 10b_4'x_4 = x_0.$$

Comparing (1) and (4), we see that:  $b_1=b_1'$ ,  $b_2=0.1b_2'$ ,  $b_3=10b_3'$ ,  $b_4=10b_4'$ . The readjustment factors are shown at the bottom of Table C.

When Method C is applied to the illustrative example with 6 decimal places, the cross sums check with the  $c$  column to within one or two units in the last decimal place, almost but not quite as well as in Method B.

Method C has the disadvantage of involving additional labor, in-

cluding the proper adjustment of the original data, the readjustment of the coefficients, and the computation of new check sums.

## METHOD D

While both Methods B and C have some advantages in giving greater accuracy, many users of multiple regressions now prefer to reduce the original matrix of Table A, whose terms are  $a_{ij} = n\sum x_i x_j$ , where  $x_i$  and  $x_j$  are deviations from the means of original variables  $X_i$  and  $X_j$ , by dividing each element by  $n^2\sigma_i\sigma_j$ , thus giving elements of the form  $r_{ij}$ ,

TABLE D  
COEFFICIENTS OF TABLE A DIVIDED BY  $n^2\sigma_i\sigma_j$  TO GIVE  
CORRELATION COEFFICIENTS

$n\sigma_j$		14.4125	62.4113	1.3543	1.5569	10.5651	
$n\sigma_i$		$x_1$	$x_2$	$x_3$	$x_4$	$x_0$	$c''$
14.4125	$x_1$	1.000000	0.205501	0.755413	0.365688	0.042469	2.369071
62.4113	$x_2$		1.000000	0.104963	0.446002	0.309699	2.066165
1.3543	$x_3$			1.000000	0.557740	0.228762	2.646878
1.5569	$x_4$				1.000000	0.302952	2.672382

that is, the coefficient of correlation between  $x_i$  and  $x_j$ , as shown in Table D. The diagonal elements are then all unity and the other elements are less than 1, and hence the number of significant figures will tend to be about the same as the number of decimal places throughout.

When this method is applied to the illustrative example, with computations carried out to six decimal places, the cross sums agree with column  $c''$  to within one or two units in the last decimal place.

This method is used by most writers who have published improved methods of Doolittle computations.<sup>3</sup> It is also useful if the solution is to include not only the column for the dependent variable,  $x_0$ , but also the columns of the identity matrix (unit diagonal and all other elements zero) used in finding the inverse of the independent-variable matrix and the standard errors of the  $b_i$ 's.

The solution of the adjusted data from Table D gives, not the  $b_i$  but the  $\beta_i$ , which must be multiplied by  $\sigma_0/\sigma_i$  to obtain the  $b_i$ .

Method D entails considerable additional labor, in dividing by

<sup>3</sup> See, for example: Harold Hotelling, "Some New Methods in Matrix Calculation," *Annals of Mathematical Statistics*, Vol. 14, March, 1943, pp. 1-34; Paul S. Dwyer, "Recent Developments in Correlation Technique," *Journal of the American Statistical Association*, Vol. 37, December, 1942, pp. 441-460, including a good bibliography.

$n^2\sigma_i\sigma_j$  in computing new check sums, and in converting the  $\beta_i$  into the  $b_i$ ; when the new information obtained is useful, the labor is justified.

## COMPARISON OF METHODS

In Table E are shown the coefficients,  $b_i$ , obtained by each method. The second section of the table shows the totals obtained by substituting the  $b_i$  in the original equations, the last column giving the original totals (the  $x_0$  column) for comparison. As a rough check, these

TABLE E  
RESULTS OF FOUR METHODS

Coefficients obtained	Method A	Method B	Method C	Method D	
$b_1$	-0.288257	-0.288257	-0.288260	-0.288207	
$b_2$	0.054666	0.054666	0.054666	0.054663	
$b_3$	3.658749	3.658773	3.658810	3.657929	
$b_4$	0.279201	0.279181	0.279160	0.279494	
Equations	Totals obtained by substituting the $b_0$ in original equations				Original totals ( $x_0$ )
I	6.466608	6.466798	6.466548	6.466753	6.4668
II	204.209240	204.208586	204.207450	204.212220	204.2098
III	3.273180	3.273201	3.273199	3.272731	3.2732
IV	4.983222	4.983202	4.983170	4.983248	4.9832
Sum	218.932250	218.931787	218.930367	218.934952	218.9330
Net discrepancy	-0.000750	-0.001213	-0.002633	0.001952	
Absolute discrepancy	0.000791	0.001819	0.002633	0.002984	

totals are summed and their net discrepancy from the original totals is shown. The absolute discrepancy (the sum of the individual discrepancies with sign neglected) is also shown.

It will be seen that all the methods give values of the  $b_i$  that agree to at least three and usually four or five significant figures. When the  $b_i$  are substituted in the original equations the net and gross discrepancies are of the order of 0.001. Method A here makes the best showing and Method D the worst; this seemingly paradoxical result may be due to the fact that some accuracy is lost in the adjustments to the original data carried out in Methods C and D; or it may simply be due to chance

in the figures of this particular example. In any case, Hotelling<sup>4</sup> does not consider the substitution test very helpful; he gives formulas for the upper limit of errors, but not for the distribution of errors.

The fact remains that, when applied to the present numerical example, Miss Bruner's Method B gives the best agreement between cross sums and the *c* column, and that its results are rough medians of those obtained by the four methods.

While the final results may not vary much, the agreement of the cross sums with the *c* column is an important practical timesaver for the computers. It would therefore seem to be good practice to use Method B rather than Method A whenever the original data vary much in size, provided the order in which the equations are arranged is immaterial and provided the extra information obtainable from Method D is not wanted. In cases where it is desired to test the effects of adding one variable at a time Method C may be useful.

*Cowles Commission for Research in Economics*  
*The University of Chicago*

<sup>4</sup> *Op. cit.*, section 3, pp. 6-8.