MISSING VALUES IN COVARIANCE IN THE CASE OF THE RANDOMIZED BLOCK

by

Catherine Shannon

A Thesis Submitted to the Graduate Committee

For the Degree of

MASTER OF SCIENCE

in

Statistics

Virginia Polytechnic Institute

1948

Acknowledgments

This author wishes to thank Dr. D. B. DeLury for suggesting this research problem and to express gratitude to Dr. Boyd Harshbarger for his suggestions as to the procedure to follow in arriving at the results contained in this paper.

Table of Contents

Introduction

In most experimental work, the results of one or more observations are occasionally lost or distorted by some disturbing factor in such a way as to make particular observations useless. Crops may be destroyed, animals may die, or errors may be made in recording the data. In the laboratory it may be possible to repeat a portion of the experiment and obtain new values for those that are missing, but often that is impossible and one has to make the best of the results available. Of course, observations should be rejected in the analysis of results only under extreme circumstances, when it is quite obvious that the treatment being studied is not responsible for the apparently anomalous results. In this event, it would be quite helpful to be able to insert an estimate of the missing value in its place and to proceed with the analysis in the customary fashion.

The analysis of the data in the case in which a single variable is being studied and one or more observations are missing has been fairly well covered. One way in the randomized block layout has been to ignore all the treatment yields in the block in which the missing plot was located and to carry out a straightforward analysis of variance of the data from the remaining n-1 blocks. However, by doing this, much of the information in the experiment was not taken into account in the analysis.

In order to retain all of the information and still use the ordinary methods of analysis, Allan and Wishart proposed a formula by which one missing item could be supplied and not change the value of the analysis. In the randomized block, they minimized the error sum of squares and solved the resulting normal equations for a quantity "k" which was to be constant throughout the blocks. Their value for "k" was

$$k = \frac{(n+s-1)S_y - nS_{yb} - sS_{yt}}{(n-1)(s-1)}$$

where n is the number of blocks,

s is the number of treatments,

S_y is the sum of all known observations in the data,

S_{yb} is the sum of the known observations in the block containing the missing value,

and S_{yt} is the sum of the known observations in the treatment containing the missing value.

This idea was simplified somewhat by Yates and extended to include more than one missing item. He set up the regular analysis of variance table using symbols for the block, treatment, and total sum of squares from which he obtained the error sum of squares, minimized the error sum of squares, and solved for "x", the missing value. This gave him the following formula which corresponds to the one above:

$$x = \frac{tT + bB - S}{(t-1)(b-1)}$$

where b is the number of blocks,

t is the number of treatments,

B is the sum of the items in the same block as the missing

item,

T is the sum of the items with the same treatment as the missing item,

and S is the sum of all observed items.

To obtain the values for more than one unit, the reiterative method is commonly used. An approximation is given to all of the missing values except the one which is being calculated. The above formula is used then to determine the missing value. The calculated value is inserted in its position and used to determine the rest of the numbers. This process is continued through several cycles until the values so derived are fairly constant for each position. However, some efficiency is lost with each number that is calculated, and it is well not to try to fill in a very large number of missing values as it would tend to ruin the analysis.

Since the beginning of the study of missing values, many people have added to the knowledge and theory concerning their estimation. It is now possible to find theory and formulae for missing values for many of the experimental designs. The mathematical basis for the method of estimating missing values is the substitution of a value for the one missing that will make the sum of squares of deviations from the mean a minimum. This means that the supplied item is its own expected value, its deviation from the expected being zero. This value may be entered in the table as the estimate of the missing item. The analysis of variance would then proceed as usual, with one modification: the number of degrees of freedom for total and error sums of squares would be decreased by unity for each value supplied.

One of the cases which has not been studied is that of missing values in covariance. Since apparently no theory has been developed for this instance, it is proposed in this paper to present a formula that may be used in a randomized block having two variates when the dependent variable is missing and the concomitant variable is present. It will be shown that the missing value may be estimated by

$$y_{\alpha\alpha} = \left[\frac{nB_{\alpha y} + sT_{\alpha y} - G_y}{(n-1)(s-1)}\right] - r\left[\frac{nB_{\alpha x} + sT_{\alpha x} - G_x}{(n-1)(s-1)} - x_{\alpha\alpha}\right]$$

where n and s are the number of blocks and treatments respectively in the experiment, r is the regression coefficient attached to the variable x, and the other symbols are block, treatment, and grand totals for y and x from the data table.

The General Case

Let us first examine the general case in which we have n blocks and s treatments in our randomised block. Let us assume that every cell of our table except one contains two values: 1) $x_{\alpha\beta}$ the concomitant variable, and 2) $y_{\alpha\beta}$ the dependent variable. In that one cell, assume that only the $x_{\alpha\beta}$ value is present. This may be shown in the following table.

Table 1.

n by s Randomized Block

Blocks	Treatments							
	1	2	. . . . .	α	. . . . .	s-1	s	Totals
1	h_{11}	h_{12}		$h_{1\alpha}$		$h_{1,s-1}$	h_{1s}	B_{1h}
2	h_{21}	h_{22}		$h_{2\alpha}$		$h_{2,s-1}$	h_{2s}	B_{2h}
.								
.								
.								
λ	$h_{\lambda 1}$	$h_{\lambda 2}$		$x_{\lambda\alpha}$		$h_{\lambda,s-1}$	$h_{\lambda s}$	$B_{\lambda h}$
.								
.								
.								
n	h_{n1}	h_{n2}		$h_{n\alpha}$		$h_{n,s-1}$	h_{ns}	B_{nh}
Totals	T_{1h}	T_{2h}		$T_{\alpha h}$		$T_{\alpha,s-1}$	$T_{\alpha s}$	G_h

where h is x and y

$\alpha = 1 \ldots\ldots\ldots\ldots\ldots\ldots s$

$\nu = 1 \ldots\ldots\ldots\ldots\ldots\ldots n.$

Assume that observation $y_{\nu\alpha}$ is missing in the above table and that we wish to find an acceptable estimate of that value.

The usual procedure is to minimise the error sum of squares

1) $\sum_{\nu\alpha}(y_{\nu\alpha} - Y_{\nu\alpha})^2$

where

$$y_{\nu\alpha} = \mu + \beta_\nu + \tau_\alpha + \rho x_{\nu\alpha} + \epsilon_{\nu\alpha}$$

and

$$Y_{\nu\alpha} = \mu + \beta_\nu + \tau_\alpha + \rho x_{\nu\alpha}$$

Here μ is the population mean,

β_ν is the effect of block ν,

τ_α is the effect of treatment α,

ρ is the regression coefficient,

and $\epsilon_{\nu\alpha}$ is normally and independently distributed.

This means minimizing

$$\sum_{\nu\alpha}(y_{\nu\alpha} - \beta_\nu - \tau_\alpha - \mu - \rho x_{\nu\alpha})^2$$

or, by replacing the population parameters by their statistical estimates,

2) $\sum_{\nu\alpha}(y_{\nu\alpha} - m - b_\nu - t_\alpha - r x_{\nu\alpha})^2$

subject to the restrictions

$$\sum_\nu b_\nu = \sum_\alpha t_\alpha = 0.$$

By introducing the Lagrange multipliers, we may find the unrestricted minimum of

3) $\sum_{\nu\alpha}(y_{\nu\alpha} - m - b_\nu - t_\alpha - r x_{\nu\alpha})^2 + 2\lambda_1 \sum_\nu b_\nu + 2\lambda_2 \sum_\alpha t_\alpha$

To bring this into conformity with the standard form of regression equation, this may be written

4) $F = \sum_{ik}(y_{ik} - m - b_i\delta_{ij} - t_k\delta_{ik} - rx_{ik})^2 + 2\lambda_1\sum b_i + 2\lambda_2\sum t_k$

where $\delta_{pq} = 1$ or 0 according as $p = q$ or $p \neq q$ with $j = 1.....n$ and $k = 1.....s$. The δ's are the independent variables and the b's, t's, m, and r are the regression coefficients, chosen to minimize the error sum of squares.

The derivatives of F with respect to m, the b's, t's, λ's, and r, equated to zero, provide a set of normal equations which determine the values of the regression coefficients. In the general case, we have

5)
$$\frac{\partial F}{\partial m} = \sum_{ik}(y_{ik} - m - b_i\delta_{ij} - t_k\delta_{ik} - rx_{ik}) = 0$$
$$\frac{\partial F}{\partial b_i} = \sum_{ik}(y_{ik} - m - b_i\delta_{ij} - t_k\delta_{ik} - rx_{ik})\delta_{ij} + \lambda_1 = 0$$
$$\frac{\partial F}{\partial t_k} = \sum_{ik}(y_{ik} - m - b_i\delta_{ij} - t_k\delta_{ik} - rx_{ik})\delta_{ik} + \lambda_1 = 0$$
$$\frac{\partial F}{\partial r} = \sum_{ik}(y_{ik} - m - b_i\delta_{ij} - t_k\delta_{ik} - rx_{ik})x_{ik} = 0$$
$$\frac{\partial F}{\partial \lambda_1} = \sum_i b_i = 0$$
$$\frac{\partial F}{\partial \lambda_2} = \sum_k t_k = 0$$

as our normal equations. These may in turn be written as

6) $\sum_i\sum_k y_{ik} = (ns-1)m + s\sum b_i + n\sum t_k + r\sum_i\sum_k x_{ik}$

7) $\sum_i\sum_k y_{ij} = s(n-1)m + s\sum b_i + (n-1)\sum t_k + r\sum_i\sum_k x_{ik} + \lambda_1$

8) $\sum_i\sum_k y_{ik} = (s-1)nm + (s-1)\sum b_i + n\sum t_k + r\sum_i\sum_k x_{ik} + \lambda_2$

9) $\sum_i\sum_k y_{ik}x_{ik} = m\sum_i\sum_k x_{ik} + \sum b_i\sum_k x_{ik} + \sum t_k\sum_i x_{ik} + r\sum_i\sum_k (x_{ik})^2$

or

10) $G_y = (ns-1)m + s\sum b_i + n\sum t_k + rG_x$

11) $G_y - B_{ij} = s(n-1)m + s\sum b_i + (n-1)\sum t_k + r[G_x - B_{ix}] + \lambda_1$

12) $G_y - T_{ky} = (s-1)nm + (s-1)\sum b_i + n\sum t_k + r[G_x - T_{kx}] + \lambda_2$

13) $H_{xy} = mG_x + B_{ix}\sum b_i + T_{kx}\sum t_k + rH_{xx}$

where G_y is the grand total,

B_{ih} is the total of the block (i) containing the missing value,

$T_{\alpha h}$ is the total of the treatment (α) containing the missing value,

H_{xy} is the sum of the products $x_{\alpha\nu}y_{\alpha\nu}$

H_{x2} is the sum of the squares $x_{\alpha\nu}^2$,

and h is x and y as before.

From these last equations, it is easily proved that $\lambda_1 = \lambda_2 = 0$. By subtracting

$$B_{iy} = (s-1)m + \Sigma t_\alpha + rB_{ix}$$

from equation (10) and equating the results to equation (11), it is found that

$$\lambda_1 = 0.$$

Then by subtracting

$$T_{\alpha y} = (n-1)m + \Sigma b_\alpha + rT_{\alpha x}$$

from equation (10) and equating that to equation (12), it is found that

$$\lambda_2 = 0.$$

In order to solve these normal equations for m, the b's, t's, and r, it is necessary to expand them into the following form with all of the symbols as defined above.

14)

$$G_y = (ns-1)m + sb_1 + \cdots + sb_{\alpha-1} + sb_{\alpha+1} + \cdots + sb_n + nt_1 + \cdots + nt_{\alpha-1} + nt_{\alpha+1} + \cdots + nt_s + rG_x$$

$$B_{1y} = m + sb_1 + t_1 + \cdots + t_{\alpha-1} + t_{\alpha+1} + \cdots + t_s + rB_{1x}$$

$$\vdots$$

$$B_{\alpha-1\,y} = m + sb_{\alpha-1} + t_1 + \cdots + t_{\alpha-1} + t_{\alpha+1} + \cdots + t_s + rB_{\alpha-1\,x}$$

$$B_{\alpha+1\,y} = m + sb_{\alpha+1} + t_1 + \cdots + t_{\alpha-1} + t_{\alpha+1} + \cdots + t_s + rB_{\alpha+1\,x}$$

$$\vdots$$

$$B_{ny} = m + sb_n + t_1 + \cdots + t_{\alpha-1} + t_{\alpha+1} + \cdots + t_s + rB_{nx}$$

$$T_{1y} = m + b_1 + \cdots + b_{\alpha-1} + b_{\alpha+1} + \cdots + b_n + nt_1 + rT_{1x}$$

$$\vdots$$

$$T_{\alpha-1\,y} = m + b_1 + \cdots + b_{\alpha-1} + b_{\alpha+1} + \cdots + b_n + nt_{\alpha-1} + rT_{\alpha-1\,x}$$

$$T_{\alpha+1\,y} = m + b_1 + \cdots + b_{\alpha-1} + b_{\alpha+1} + \cdots + b_n + nt_{\alpha+1} + rT_{\alpha+1\,x}$$

$$\vdots$$

$$T_{sy} = m + b_1 + \cdots + b_{\alpha-1} + b_{\alpha+1} + \cdots + b_n + nt_s + rT_{sx}$$

$$H_{xy} = G_x m + B_{1x} b_1 + \cdots + B_{\alpha-1\,x} b_{\alpha-1} + B_{\alpha+1\,x} b_{\alpha+1} + \cdots + B_{nx} b_n + T_{1x} t_1 + \cdots + T_{\alpha-1\,x} t_{\alpha-1} + T_{\alpha+1\,x} t_{\alpha+1} + \cdots + T_{sx} t_s + rH_x$$

These provide us with (n+s) equations with which to solve our (n+s) unknowns. With the identities for m, b_α, t_α, and r, it is possible to solve for the missing value $y_{\alpha\alpha}$, since it is known that the equation for $y_{\alpha\alpha}$ is

$$y_{\alpha\alpha} = m + b_\alpha + t_\alpha + r x_{\alpha\alpha}.$$

The formula derived in this manner is

15) $$y_{\alpha\alpha} = \left[\frac{nB_{\alpha x} + sT_{\alpha x} - G_x}{(n-1)(s-1)}\right] - r\left[\frac{nB_{\alpha x} + sT_{\alpha x} - G_x}{(n-1)(s-1)} - x_{\alpha\alpha}\right]$$

where n is the number of blocks,

s is the number of treatments,

$B_{\alpha h}$ is the sum of "h" in the ith block,

$T_{\alpha h}$ is the sum of "h" in the αth treatment,

G_h is the grand sum of "h",

and $$r = \frac{H_{xy} - \sum_\alpha B_{\alpha x}B_{\alpha y}/s - \sum_\alpha T_{\alpha x}T_{\alpha y}/n + [(ns-n-s)G_xG_y - n^2B_{\alpha x}B_{\alpha y} - s^2T_{\alpha x}T_{\alpha y} + n(B_{\alpha x}G_y + B_{\alpha y}G_x) + s(T_{\alpha x}G_y + T_{\alpha y}G_x) - ns(B_{\alpha x}T_{\alpha y} + T_{\alpha x}B_{\alpha y})]/ns(n-1)(s-1)}{H_{x^2} - \sum_\alpha B_{\alpha x}^2/s - \sum_\alpha T_{\alpha x}^2/n + [(ns-n-s)G_x^2 - n^2B_{\alpha x}^2 - s^2T_{\alpha x}^2 + 2nB_{\alpha x}G_x + 2sT_{\alpha x}G_x - 2ns\,B_{\alpha x}T_{\alpha x}]/ns(n-1)(s-1).}$$

It may be noticed at once that should x be absent, that is, if there is only one variable, this formula reduces to that for the simple randomized block as derived by Yates. From this it appears that the above formula may be extended to include as many variables as one may wish to include in the experiment.

A Special Case

Let us see how this process works on a special case; for example, a randomized block having two blocks and three treatments. Assume, as before, that there exists both an $x_{\alpha\alpha}$ value and a $y_{\alpha\alpha}$ value in every cell of our table except one. Suppose that the cell in the second block and in the second treatment contains only the x_{22} value, and it is desired to calculate a satisfactory estimate for y_{22}.

Table 2.

2 by 3 Randomized Block

Blocks	Treatments			
	1	2	3	Totals
1	h_{11}	h_{12}	h_{13}	B_{1h}
2	h_{21}	x_{22}	h_{23}	B_{2h}
Totals	T_{1h}	T_{2h}	T_{3h}	G_h

Let this be the data table where h stands for x and y, as in the general case, and $B_{\alpha h}$, $T_{\alpha h}$, and G_h stand for the block, treatment, and grand totals respectively.

As in the general case, the error sum of squares

16) $$F = \sum_{\alpha\alpha}(y_{\alpha\alpha} - m - b_\alpha - t_\alpha - rx_{\alpha\alpha})^2$$

is to minimized subject to the restrictions

$$\sum_{\alpha} b_{\alpha} = \sum_{\alpha} t_{\alpha} = 0.$$

Or, introducing the Lagrange multipliers, find the unrestricted minimum of

17) $$F = \sum_{\alpha\alpha}(y_{\alpha\alpha} - m - b_{\alpha} - t_{\alpha} - rx_{\alpha\alpha})^2 + 2\lambda_1 \sum_{\alpha} b_{\alpha} + 2\lambda_2 \sum_{\alpha} t_{\alpha}$$

where m, the b's, t's, and r are the statistical estimates of the population parameters. To convert equation (17) into the conventional style, it may be written

18) $$F = \sum_{\alpha\alpha}(y_{\alpha\alpha} - m - b_{\alpha}\delta_{\alpha\beta} - t_{\alpha}\delta_{\alpha\beta} - rx_{\alpha\alpha})^2 + 2\lambda_1 \sum_{\alpha} b_{\alpha} + 2\lambda_2 \sum_{\alpha} t_{\alpha}$$

When the derivative of equation (18) is taken with respect to m, b_1, b_2, t_1, t_2, t_3, and r, equated to zero, and summed, a set of normal equations are obtained.

19)
$$G_y = 5m + 3b_1 + 2b_2 + 2t_1 + t_2 + 2t_3 + rG_x$$
$$B_{1y} = 3m + 3b_1 + t_1 + t_2 + t_3 + rB_{1x}$$
$$B_{2y} = 2m + 2b_2 + t_1 + t_3 + rB_{2x}$$
$$T_{1y} = 2m + b_1 + b_2 + 2t_1 + rT_{1x}$$
$$T_{2y} = m + b_1 + t_2 + rT_{2x}$$
$$T_{3y} = 2m + b_1 + b_2 + 2t_3 + rT_{3x}$$
$$H_{xy} = G_x m + B_{1x}b_1 + B_{2x}b_2 + T_{1x}t_1 + T_{2x}t_2 + T_{3x}t_3 + rH_x 2$$

Since it is known that

$$\sum_{\alpha} b_{\alpha} = \sum_{\alpha} t_{\alpha} = 0,$$

these equations may be written

20)
$$G_y = 5m - b_2 - t_2 + rG_x$$
$$B_{1y} = 3m + 3b_1 + rB_{1x}$$
$$B_{2y} = 2m + 2b_2 - t_2 + rB_{2x}$$
$$T_{1y} = 2m + 2t_1 + rT_{1x}$$
$$T_{2y} = m - b_2 + t_2 + rT_{2x}$$
$$T_{3y} = 2m + 2t_3 + rT_{3x}$$
$$H_{xy} = G_x m + B_{1x}b_1 + B_{2x}b_2 + T_{1x}t_1 + T_{2x}t_2 + T_{3x}t_3 + rH_x 2$$

On solving these normal equations by any of the possible methods, the following identities are obtained

21)

$$m = \left[\frac{G_y}{12} + \frac{B_{2y}}{6} + \frac{T_{2y}}{4}\right] - r\left[\frac{G_x}{12} + \frac{B_{2x}}{6} + \frac{T_{2x}}{4}\right]$$

$$b_1 = \frac{B_{1y} - rB_{1x}}{3} - m$$

$$b_2 = \left[\frac{B_{2y}}{2} + \frac{T_{2y}}{4} - \frac{G_y}{4}\right] - r\left[\frac{B_{2x}}{2} + \frac{T_{2x}}{4} - \frac{G_x}{4}\right]$$

$$t_1 = \frac{T_{1y} - rT_{1x}}{2} - m$$

$$t_2 = \left[T_{2y} + \frac{B_{2y}}{3} - \frac{G_y}{3}\right] - r\left[T_{2x} + \frac{B_{2x}}{3} - \frac{G_x}{3}\right]$$

$$t_3 = \frac{T_{3y} - rT_{3x}}{2} - m$$

$$r = \frac{H_{xy} - (B_{1x}B_{1y} + B_{2x}B_{2y})/3 - (T_{1x}T_{1y} + T_{2x}T_{2y} + T_{3x}T_{3y})/2 + (G_xG_y - 4B_{2x}B_{2y} - 9T_{2x}T_{2y} + 2(B_{2x}G_y + B_{2y}G_x) + 3(T_{2x}G_y + T_{2y}G_x) - 6(B_{2x}T_{2y} + T_{2x}B_{2y}))/12}{H_{x2} - (B_{1x}^2 + B_{2x}^2)/3 - (T_{1x}^2 + T_{2x}^2 + T_{3x}^2)/2 + (G_x^2 - 4B_{2x}^2 - 9T_{2x}^2 + 4B_{2x}G_x + 6T_{2x}G_x - 12B_{2x}T_{2x})/12.}$$

If these values are tested, it will be found that

$$\sum b_\alpha = \sum t_\alpha = 0$$

according to the restrictions set in the beginning. It will also be found, when these values are inserted into

$$y_{22} = m + b_2 + t_2 + rx_{22},$$

that

22)

$$y_{22} = \left[\frac{2B_{2y} + 3T_{2y} - G_y}{2}\right] - r\left[\frac{2B_{2x} + 3T_{2x} - G_x}{2} - x_{22}\right]$$

where r is equal to the above. This agrees with the general case as discussed previously.

Conclusion

The formula and theory for estimating a missing value in the case of covariance in a randomized block has been presented in this paper. It has also been found that the formula given corresponds to Yates' formula for a missing value in a randomized block when there is only one variable present in the experiment. The formula derived is

$$23) \qquad y_{\alpha\alpha} = \left[\frac{nB_{\alpha y} + sT_{\alpha y} - G_y}{(n-1)(s-1)}\right] - r\left[\frac{nB_{\alpha x} + sT_{\alpha x} - G_x}{(n-1)(s-1)} - x_{\alpha\alpha}\right]$$

where

$$24) \qquad r = \frac{H_{xy} - \sum B_{\alpha x}B_{\alpha y}/s - \sum T_{\alpha x}T_{\alpha y}/n + \left[(ns-n-s)G_xG_y - n^2B_{\alpha x}B_{\alpha y} - s^2T_{\alpha x}T_{\alpha y} + n(B_{\alpha x}G_y + B_{\alpha y}G_x) + s(T_{\alpha x}G_y + T_{\alpha y}G_x) - ns(B_{\alpha x}T_{\alpha y} + T_{\alpha x}B_{\alpha y})\right]/ns(n-1)(s-1)}{H_{x^2} - \sum B_{\alpha x}^2/s - \sum T_{\alpha x}^2/n + \left[(ns-n-s)G_x^2 - n^2B_{\alpha x}^2 - s^2T_{\alpha x}^2 + 2nB_{\alpha x}G_x + 2sT_{\alpha x}G_x - 2nsB_{\alpha x}T_{\alpha x}\right]/ns(n-1)(s-1).}$$

Bibliography

Allan, F. E. & Wishart, J., "A Method of Estimating the Yield of a Missing Plot in Field Experimental Work", Journal of Agricultural Science, 1930, V. 20, pp. 399-406.

Anderson, R. L., "Missing-Plot Techniques", Reprint from Biometrics Bulletin, June 1946, V. 2, No. 3, pp. 41-47.

DeLury, D. B., "The Analysis of Latin Squares When Some Observations are Missing", Reprint from the Journal of the American Statistical Association, Sept. 1946, V. 41, pp. 370-389.

Goulden, C. H., Methods of Statistical Analysis, John Wiley and Sons, New York, 1939, pp. 261-265.

Paterson, D. D., Statistical Technique in Agricultural Research, McGraw-Hill Book Co, Inc., New York & London, 1939, pp. 180-188.

Snedecor, G. W., Statistical Methods, Iowa State College Press, Ames, Iowa, 1946, pp. 268-269.

Wilks, S. S., "The Analysis of Variance and Covariance of Non-Orthogonal Data", Metron, V. 13, N. 2, 1938.