JOURNAL

OF THE

Indian Society of Agricultural Statistics

Vol. II] 1950 [No. 2

ERROR VARIANCE OF TREATMENT CONTRASTS IN AN EXPERIMENT WITH MISSING OBSERVATIONS (WITH SPECIAL REFERENCE TO INCOMPLETE LATIN SQUARES)

By H. Fairfield Smith

The Rubber Research Institute of Malaya

In the original paper, which put the method of analysing experiments with missing plots in its modern form, Yates (1933) indicated that the error variance of treatment comparisons could be obtained by considering the linear function of observations giving the estimate of any comparison in question. This method is however intolerably tedious, except in the simplest cases, and Yates (1933 and 1936) suggested some approximations. That for randomised blocks has since been improved by Taylor (1948), and for the simple treatment contrasts considered by him no greater accuracy will usually be required. It is nevertheless advisable to be able to derive the appropriate estimate of error for any case which may arise, and the procedure indicated below will give the exact formula—without undue labour if the number of missing observations is not abnormally large.

Bartlett (1937) pointed out that, when some observations of an experimental design are missing, estimates of the missing observations, which minimise the sum of squares of deviations from fitted constants, can be obtained, in the form of regression coefficients, from analysis of covariance on dummy variates. Wishart (1936) showed how to

compute the error variance of treatment comparisons adjusted by a regression on concomitant variables following analysis of covariance. It is implied by Bartlett's presentation that the standard errors of individual treatment contrasts involving estimates of missing observations will be given in the same way; but, while the procedure to obtain the exact z test of significance among a group of treatments has become well known (having been illustrated by Bartlett, *loc. cit.*), the procedure for computing the standard error of individual contrasts seems to have been generally overlooked by research workers.

Two instructive examples, (1) with a single missing plot, (2) with a complex of abnormal plots, are provided by an experiment on manuring rubber trees, which was composed of two incomplete (6×5) latin squares. On one of these the felled jungle timber was left to rot *in situ*, on the other it was burnt.

Example (1) *Unburnt area.*—Data for mean girth per tree per plot are given in Table I: one plot, marked *, was affected by root disease and heavily "supplied" (*i.e.*, partially replanted) and consequently has

TABLE I

Girths less 20 *inches, in units of* ·01 *inch*: *Unburnt area.*

	Column 1	2	3	4	5	6	Row Total	Treatment	Total
Row 1 ..	79*A*	67*C*	501*D*	94*E*	26*B*	193*F*	960	*A* (1)	609
2 ..	250*D*	277*F*	331*E*	−5*C*	22*A*	−290**B*	585	*B n*	37
3 ..	61*B*	208*A*	303*C*	270*F*	244*E*	367*D*	1453	*C nk*	386
4 ..	433*F*	282*D*	210*A*	87*B*	−3*C*	267*E*	1276	*D np*	1699
5 ..	290*E*	153*B*	404*F*	90*A*	299*D*	24*C*	1260	*E pk*	1226
Treatm. of "missing" row	*C*	*E*	*B*	*D*	*F*	*A*	..	*F npk*	1577
Column Total	1113	987	1749	536	588	561	5534		

many trees of abnormally small girth. Using as dummy variate $(x) - 1$ for the plot * and 0 for others, analysis of covariance, following the method of analysis described by Yates (1936), is obtained as in Table II. The sums of squares for x are simply the degrees of freedom for each term divided by the total number of plots (30) (Anderson, 1946). For sums of products one simply replaces squares

TABLE II

	d.f.	Sums of Squares and Products		
		y^2	x^2	xy
Rows ..	4	77629·5	·13	86·96
Columns ..	5	223101·5	·16	72·26
Treatments ..	5	385457·9	·16	166·06
Remainder ..	15	99564·6	·5	149·16
Regression ..	1	44501·4	$F = 11·315$	
Remainder ..	14	55063·2		

Rem. M.Sq. 3933·09

Regression coefficient (b) = 149·16/·5 = 298·3

"Estimated value" of abnormal observation = $-290+b = 8·3$

$V(b) = 3933·09/·5 = 7866·18$

in the formulæ of Yates (1936) by products, *e.g.*, sum of products ascribed to treatments

$$= \frac{\Sigma\{(p-1)\,T_{yi} + C_{yi}\}\{(p-1)\,T_{xi} + C_x\} - pG_yG_x}{p\,(p-1)\,(p-2)}, \qquad (1)$$

where p is here 6,

T_{yi} = total of observed girths (y) for treatment i,

T_{xi} = total x for treatment i; here -1 for treatment B and otherwise 0,

C_{yi} = total y for the column in which treatment i is missing,

C_{xi} = total x for the column in which treatment i is missing; here -1 for column A and otherwise 0,

G_y = grand total of y,

G_x = grand total of x; here -1.

Estimation of treatment contrasts and their standard errors when data are complete.—From the formulæ given by Yates, p. 303, the estimate of the mean of treatment i, with complete data, is

$$m + t_i = \frac{(p-1)\,T_i + C_i}{p\,(p-2)} - \frac{G}{p\,(p-1)\,(p-2)} = t_i' - \text{const.} \quad (2)$$

In considering treatment contrasts the last term cancels and so need not be considered. Expressing the first term as a linear function of plot yields its error variance is easily seen to be

$$V(t') = \frac{(p-1)\,(p^2-2p+2)}{p^2\,(p-2)^2}\,V_1, \quad (3)$$

where V_1 is the variance of a single plot. Treatment constants (t_i') are however correlated in that any pair have two plots in common with coefficients $(p-1)$ and 1, the coefficients being reversed in the two treatments. The covariance of any two constants is therefore

$$\text{Cov.}\,(t_i' t_j') = \frac{2\,(p-1)}{p^2\,(p-2)^2}\,V_1. \quad (4)$$

From these expressions can be obtained the variance of any linear function of the t's used to estimate a treatment contrast. For example, if the experiment of Table I be regarded as an incomplete factorial with treatments p and k missing from the standard $2\times2\times2$ form, and if all interactions can be assumed to be zero, the overall estimate with minimum variance for response to phosphate is (Smith, 1949)

$$\bar{P} = \tfrac{1}{4}\,[npk + pk + 2np - 2nk - n - (1)] \quad (5)$$

whence

$$V(\bar{P}) = \tfrac{1}{16}\,\{12V(t') - 12\ \text{Cov.}\,(t_i' t_j')\} = \frac{3\,(p-1)}{4p\,(p-2)}\,V_1. \quad (6)$$

Estimation of treatment contrasts and standard errors with a "missing" observation.—Following the usual method with an abnormal or missing observation, we can replace it by its "estimated value" (here 8·33) and estimate treatment contrasts as for complete data. To estimate the standard error of such a contrast (E), one notes that it can be made up of two terms

$$E = Y - bX, \quad (7)$$

where Y is a linear function of *observed* treatment means or totals (see comment 2 below) and X is the same linear function of the dummy (or concomitant) variate. The regression coefficient, b, being determined from the remainder line of the analysis of variance and

covariance, is orthogonal to treatment contrasts as evaluated from the observations, and therefore the error variance of E is given by

$$V(E) = V(Y) + X^2 V(b), \tag{8}$$

For example, consider the estimation of contrasts between any pair of treatments. In an incomplete latin square, from (2), using observed "yields" only

$$Y = t_i' - t_j' = \frac{(p-1)T_i + C_i - (p-1)T_j - C_j}{p(p-2)} \tag{9}$$

$$V(Y) = 2V(t') - 2\,\text{Cov.}(t_i' t_j') = \frac{2(p-1)}{p(p-2)} V_1. \tag{10}$$

If an observation is "missing", or adjusted, in treatment i and column j

$$X(t_i') = \frac{-(p-1)}{p(p-2)} \quad X(t_k') = 0 \tag{11}$$

$$X(t_j') = \frac{-1}{p(p-2)} \quad X(t_i' - t_j') = \frac{-1}{p}$$

$$V(b) = \frac{p(p-1)}{(p-1)(p-3)} V_1, \text{ etc.}$$

where t_k' is a treatment "mean" containing no estimated "yield". Consequently the error variances of the three possible types of contrast between pairs of treatments are given by

$$V(t_i' - t_k') = \left\{\frac{2(p-1)}{p(p-2)} + \frac{(p-1)^2 p}{p^2(p-2)^2(p-3)}\right\} V_1$$

$$= \frac{145}{288} V_1 \text{ for a } 6\times 5 \text{ "square"} \tag{12}$$

$$V(t_j' - t_k') = \left\{\frac{2(p-1)}{p(p-2)} + \frac{p}{p^2(p-2)^2(p-3)}\right\} V_1$$

$$= \frac{121}{288} V_1 \text{ for a } 6\times 5 \text{ "square"}$$

$$V(t_i' - t_j') = \left\{\frac{2(p-1)}{p(p-2)} + \frac{p}{p^2(p-3)}\right\} V_1$$

$$= \frac{136}{288} V_1 \text{ for a } 6\times 5 \text{ "square"}$$

as compared to $\frac{120}{288} V_1$ for a comparison with complete data.

For the estimate of $\bar{P}$, formula (5), in the example of Table I where an observation is "missing" in treatment n and in the column with treatment (1) omitted, we have,

$$X(\bar{P}) = \tfrac{1}{4}(0 + 0 + 2.0 - 2.0 + \frac{5}{24} + \frac{1}{24}) = \frac{1}{16} \qquad (13)$$

Therefore from (6) and (8)

$$V(\bar{P}) = (\frac{15}{96} + \frac{2}{256}) V_1 = \frac{21}{128} \times 3933.09 = 645.27 \qquad (14)$$

Comment (1).—Other procedures are possible to estimate the missing value (or the regression coefficient b). For example, writing y for the missing value, the analysis of variance of the "yields" of Talbe I could be evaluated as in Table III. Whence, differentiating

TABLE III

	d.f.	S.Sq.	if $y = 8\cdot3$ S.Sq.
Rows ..	4	$38402\cdot2 - 96\cdot6y + \cdot13y^2$	37 606·4
Columns ..	5	$195203\cdot5 - 47\cdot86y + \cdot16y^2$	194 816·2
Treatments ..	5	$303155\cdot9 - 235\cdot46y + \cdot16y^2$	301 205·2
Remainder ..	15	$55098\cdot0 - 8\cdot3y + \cdot5y^2$	55 063·2 (14 *d.f.*)
Total ..	29	$591859\cdot5 - 388\cdot26y + \cdot96y^2$	

the remainder sum of squares with respect to y, it is quickly seen that the "estimated value" is 8 3. Inserting this in each row gives the analysis of variance as it would be worked out using the estimate for the missing observation.

The error variance of the adjustment for y is $V(b)$ as above, and evaluation of errors of treatment contrasts proceeds as before. Since the construction of the columns for x^2 and xy in Table II is very easy, there is not much to choose with respect to computing labour in the two methods. To one who has mastered the technique of analysis of covariance the procedure of Table II leads with less thought directly to estimates of error variances and of reduced mean squares to test significance of treatment (or column, etc.) effects. (The latter being not here required since they are obviously significant.)

Comment (2).—The computation of Table II could have been made slightly easier by using $y = 0$ for plot * instead of the observed value, −290, leading to $b = 8{\cdot}3$ and subsequent work as before. When the observed yield is used the mean square ascribed to the 1 d.f. for regression (Table II) shows immediately the improvement in error sum of squares obtained by rejecting the abnormal observation. However if any other value of y had been used the same thing would be easily shown by comparing the difference between observed and estimated "yields" with $\sqrt{V(b)}$; for example, if we had here used $y = 0$, we would have $t^2 = \{8.33 - (-290)\}^2/7866.18 =$ the variance ratio for regression using the observed value $= 44501/3933$.

If analysis of covariance is carried out with $y = 0$ for the "missing" plot, it may seem at first sight that this is an arbitrary value free of error and that the total number of observations is, in this example, 29 instead of 30. But $V(Y)$ in equation (8) has to be written as if the data were complete. The reason is easily seen by regarding y as an observed yield, subject to the same degree of random error as other plot yields, but affected by some additional treatment whose effect is to be estimated by a fitted constant (b). Since there is only one observation for measurement of this effect, any difference is fully absorbed by the constant, and (except for its value) all other results will be the same irrespective of what initial value was used. But if an arbitrary value, say 0, were regarded, like the dummy variate x, as free of error, then b would not be independent of Y, and Wishart's formula for variance of an adjusted estimate would no longer hold.

Some experimentation with a simple case, such as randomised blocks, is instructive in showing how the covariance method gives the same results as the more tedious procedure of writing out the estimate of a treatment mean or contrast as a linear function of the observed yields.

Although description may appear complicated, the procedure is quite easy to apply in practice once one has mastered the technique of analysis of covariance and the estimation of error variances of means adjusted by regression. Despite the well-known difficulties of multiple regression with several independent variates, a moderate number of missing values can be handled without undue difficulty because the analyses of variance and covariance of the dummy variates are very simple, and their variance-covariance matrix in the normal equations is usually composed of fairly simple fractions.

Example (2) *Burnt area.*—Data for mean girth per tree per plot are given in Table IV. In this section there are two patches in which diseased trees have been replaced by supplies, one affecting a single

TABLE IV

Girths less 20 *inches, in units of* ·01 *inch*: *Burnt area*

	Column 1	2	3	4	5	6	R		
Row 1 ..	(−353)C*	46E	129D	−101A	(−227)B†	372F	446	A (1)	−277
2 ..	369F	−63C	82E	242D	(−222)A‡	193B	828	B n	282
3 ..	232D	22B	−142A	352F	74C	186E	724	C nk	61
4 ..	103B	−12A	343F	132E	199D	25C	790	D np	1075
5 ..	−27A	358F	25C	−36B	23E	273D	616	E pk	469
Treatment of 'missing' row	E	D	B	C	F	A	..	F npk	1794
C	677	351	437	589	301	1049	3404		

Marginal totals given for substitution of arbitrary values 0 in plots *, †, and 5 in ‡

plot marked *, and one affecting two adjacent plots † and ‡. It is reasonable to suppose that both plots † and ‡ may be approximately equally affected and so may be adjusted by fitting a single constant. We therefore use, as concomitant variates, $u = -1$ for plot *, and 0 elsewhere; $w = -1$ for each of plots † and ‡, and 0 elsewhere. As before, for computation, any arbitrary values can be assigned to the three abnormal plots, provided that the observed difference is ratained between plots † and ‡. Table V gives the analysis of variance and covariance with $y = 0$ in plots * and †, and 5 in ‡. The remainder

TABLE V

	d.f.	y^2	u^2	w^2	uw	uy	wy
Rows ..	4	15798·13	·13	·2	·1	39·13	14·6
Coloumns ..	5	75843·86	·16	·6	−·06	−21·93	106·53
Treatments ..	5	522260·38	·16	·25	−·1	107·76	122·925
Remainder ..	15	40263·08	·5	·75	0	−11·5	−22·125
Total ..	29	654165·46	·96	1·86	−·06	113·46	−221·93

$b_u = -11{\cdot}5/{\cdot}5 = -23$ = estimated value for plot *
$b_w = -22{\cdot}125/{\cdot}75 = -29{\cdot}5$ = ,, ,, †
Regr. S.Sq. 2 *d.f.* 917·1875
Rem. S.Sq. 13 *d.f.* 39 345·896
M.Sq. (V_1) 3 026·61

line gives two normal equations to solve for b_u and b_w; but in this case, since the covariance of u and w is zero, the coefficients are seen to be independent and it is not necessary explicitly to set up the equations.

If the observed values for the abnormal plots had been used we would have obtained $b_u' = -23 - (-353) = 330$, $b_w' = -29.5 - (-227) = 197.5$, and the reduction in the remainder sum of squares due to the adjustments, or "regression" sum of squares (2 d.f.), is:

$$\frac{b_u'^2\, c_{ww} + b_w'^2\, c_{uu} - 2b_u' b_w' c_{uw}}{c_{uu}\, c_{ww} - c_{uw}^2}; \qquad (15)$$

where $[c]$ is the reciprocal of the matrix of the variances and covariance of u and w used to get the regression coefficients. Since in this case $c_{uw} = 0$, it is here simply the sum of squares attributable to each adjustment separately, that is

$$\cdot 5 \times 330^2 + \cdot 75 \times 197{\cdot}5^2 = 83\,704{\cdot}69;$$

and it is obvious that the effect of disease and supplying has been very highly significant.

Estimates of contrasts using the estimated value can, as before, be represented by a function Y of observed yields adjusted by the regression coefficients multiplied by corresponding functions, U and W, of the dummy variates.

$$E = Y - b_u U - b_w W. \tag{16}$$

With such multiple adjustments the b's will not in general be independent; accordingly

$$V(E) = V(Y) + U^2 V(b_u) + W^2 V(b_w) + 2UW \text{ Cov. } (b_u b_w) \tag{17}$$

which may, with some advantage for ease of computation, be expressed in terms of the elements of the reciprocal matrix $[c]$

$$V(E) = (a + U^2 c_{uu} + W^2 c_{ww} + 2UW c_{uw})\, V_1 \tag{18}$$

where a is as usual the sum of squares of the coefficients of each observation in the linear function Y.

For example, to estimate the error variance of $\bar{P}$ as defined by (5) above,

$$a = \frac{15}{96} \text{ as in (6) or (14)}$$

$$U = \frac{1}{4.24}\,[0 - 1 + 2.0 - 2\,(-5) - 0 - 0] = \frac{3}{32}$$

$$W = \frac{1}{4.24}\,[-2 + 0 + 2.0 - 2.0 - (-5)\,(-5)] = \frac{3}{16}$$

$$c_{uu} = \frac{1}{.5},\ c_{ww} = \frac{1}{.75},\ c_{uw} = 0$$

$$V(\bar{P}) = \left(\frac{15}{96} + \frac{9}{512} + \frac{3}{64}\right) 3026.61 = 667.98$$

The estimated values (or b's) could of course again have been obtained by a similar procedure to that illustrated in Table III; but with more than one value to be estimated the covariance approach seems definitely easier, as well as giving more readily the variance-covariance matrix of the dummy variates from which to evaluate the variance of adjustments.

Further comments.—Some other features of this experiment are of passing interest. It seems that the original intention when the experiment was designed (in 1935) was to lay down two latin squares, but when there was found to be insufficient space it was supposedly converted to two groups of five randomised blocks by omitting one of the rows. It is accordingly of some interest to compare the analyses which would be given by accepting the design as intended, and as worked out above. This is done in Table VI. Here the mean squares for columns are adjusted for treatments; the crude sums of squares

TABLE VI

	d.f.	Mean Squares	
		Unburnt	Burnt
Observations treated as 5 randomised blocks			
Rows ..	4	19,407	23,250
Treatments ..	5	91,746	152,490
Remainder ..	20	12,470	9.277
Adjusted for columns and diseased patches			
Rows ..	4	9,402	4,710
Col. adj. for treatm.	5	25,122	6,712
Treatm. ,, col. ..	5	60,241	106,928
Adjustments for disease patches	1 or 2	44,501	41,852
Remainder ..	14 or 13	3,933	3,027

as entered in Tables III and V, to estimate remainder variances, being inflated by treatment differences, exaggerate the real column effects. The mean squares for rows, columns and treatments in the second part are those given by using adjusted values for the diseased plots. For tests of significance they would require to be further reduced in the usual way to allow for errors of adjustment; but this is of no interest here (the overall treatment effects are obviously significant and interest centres only in testing responses to individual fertiliser ingredients), and these values indicate the sources of heterogeneity responsible for inflating error in the simple analyses. By paying attention to columns and disease patches accuracy has been increased

threefold. The restriction on columns represents a balancing of extraneous variation which affects all treatments. The adjustments for diseased patches are somewhat different. All information, and more, from these is absorbed in estimating the adjustments [*cf.* variances indicated by equations (12)]. If they were not applied errors of treatments not affected would be seriously overestimated, of affected treatments would be underestimated.

In the type of land used for forest crops atypical patches of soil are often met with and can sometimes be detected in advance of beginning an experiment. Common examples are areas of swampy or clay soil or rock outcrops. Usually their effect is smaller than in the example here given. While in some ways their effect may be thus less serious, in other ways it is worse by adding to problems of interpretation some difficulty to decide whether or not it may be advisable or worth while to evaluate adjustments for them. Where possible such areas should be excluded from an experiment, even if it means abandoning the two way control given by a latin square and reverting to more flexible randomised blocks. At the same time it has to be admitted that, when an experiment is planned for an area still under jungle 200 miles from the research centre, detection and exclusion of abnormal soil patches raises problems of administration which are more than trivial. An administrator could with some reason take the view that the chance of improvement with respect both to accuracy and computing labour, was not good enough to justify the cost of interrupting smooth execution of plans. In this particular area the defects could have been spotted and eliminated only if the commencement of experimental treatments had been delayed for about two years after planting, and therefore was not possible if it were essential for experimental treatments to begin immediately after opening from jungle. Nevertheless wherever it may be possible to detect and avoid such patches, with their (seemingly) never-ending complications in a longterm experiment, the value of doing so can hardly be overemphasised.

The adjustments for abnormal patches, as derived above for each separately, differ only to the extent of the standard errors of the estimates; it is therefore reasonable to suppose that all four affected plots have been retarded by approximately equal amounts. Consequently, the error variances of both sections being similar, future records may be satisfactorily analysed by calculating from the pooled remainder sum of squares only a single adjustment to be applied to all four plots.

Adjustments for columns and diseased patches have reduced mean squares between treatments to two-thirds of the values given by simple randomised block analyses. Apart from a small reduction due to weights of treatment means being reduced from 5 to 24/5 after adjustment for columns, this represents a smaller spread of treatment means after adjustments. That it really does represent a reduction in errors can be well demonstrated in a way which may appeal to the " practical " worker by comparing the crude and adjusted means with the treatments, when it is evident that the adjusted values are more " sensible " in approximately the same proportion as the standard errors have been reduced.

Combination of both sections shows that the difference of the two treatment mean squares is significant, being due to a very highly significant interaction of nitrogen × burning. The complete set of treatment responses will be given elsewhere (Smith, 1949).

Summary

It is well known that analysis of covariance on dummy variates can be used to estimate values of missing observations as a simplified method of fitting constants when experimental data, in which effects were intended to be orthogonal, has been slightly deranged. It is pointed out that the method of evaluating standard errors of treatment contrasts after adjustment by regression, as described by Wishart (1936), can be used to evaluate the error variances of contrasts involving estimates of missing observations. The procedure is illustrated for incomplete latin squares where a few plots, which may be considered either individually or in groups, have been affected by abnormal soil conditions. The exact formulæ are derived for standard errors of treatment means and simple contrasts in an incomplete latin square with one missing plot. The derivation for more complex contrasts is illustrated.

Acknowledgment

The basis of this paper was outlined when the author was guest research worker at the Institute of Statistics of the University of North Carolina on a project under contract with the Office of Naval Research, U.S.A.; and owes not a little to the benefit of discussion with Professor W. G. Cochran.

Postscript

Since the above note was written I have received the December 1948 issue of *Biometrics* containing a paper by Quenouille which

develops essentially the same ideas as used in example (2) above. That paper however is concerned with data which are fundamentally non-orthogonal and notes application of the method to missing observations only in passing. This aspect seems to deserve further emphasis.

LITERATURE CITED

Anderson, R. L. .. *Biometrics Bull.*, 1946, **2**, 41–47.

Bartlett, M. S. .. *J. Roy. Stat. Soc. Suppl.*, 1937, **4**, 151.

Quenouille, M. H. .. *Biometrics*, 1948, **4**, 240–46.

Smith, H. Fairfield .. "Effect of fertilisers on growth of Hevea" (not yet. published), 1949.

Taylor, J. .. *Nature*, 1948, **162**, 262.

Wishart, J. .. *J. Roy. Stat. Soc. Suppl.*, 1936, **3**, 79–82.

Yates, F. .. *Emp. Journ. Expt. Agric.*, 1933, **1**, 129–42.

——— .. *J. Agric. Sci.*, 1936, **26**, 301–15.