HONAM MATHEMATICAL JOURNAL
Volume 5, Number 1, September, 1983

이 論文은 1982年度 文敎部 學術硏究助成費에 의하여 硏究되었음.

Estimation and Analysis of Designed Experiments with Missing Observations

By Jung-Wook Sim*

1. Introduction

Allen and Wishart (1930) provide formulae appropriate in the case of a single plot of a randomized block or Latin spuare. Yates (1933) gave a general formula for estimating a single missing value in a wide range of factorial analyses. This formula is got by minimizing the error variance obtained when an unknown **x** is substituted for the missing value. Also Yates (1933) suggested an iterative procedure to deal with more than one missing value. Both of these methods for qualitative designs minimize the error term in the analysis of variance.

Wilkinson (1958a, b) and Tocher (1952) provide examples of methods which can require considerable computation owing to matrix manipulation and inversion. An analisis of covariance would also involve a great deal of computation, especially with more than one missing value. None of these methods is much easier than performing a regression on the completed experiments.

A simpler technique for factorial designs which does not minimize the residual sum of squares, was suggested by Draper and Stoneman (1963). The method is to equate a certain number of high-order interactions to zero, and check on lack of bias in the lower effects by means of a half normal plot. In the case of one missing value estimation proceeds by equating the highest order interaction to zero, this is equivalent to minimizing the residual sum of squares where the residual is estimated by the highest order interaction. More than one missing value is dealt with by zeroing an equal number of interactions. The choice of these interactions involves writing out each estimated effect as a function of the missing values and choosing interactions to satisfy a nonsingularity constraint. After zeroing these interactions, their suitability is tested by means of a half normal plot of the main effects; if this is unsatisfactory another set of interactions must be chosen and the procedure re-

* This research is supported by Research grant of Ministry of Education in 1982.

peated.

2. Simple Methods for Full Factorial Designs

A formula for a missing value in an m-way complete randomized block design may be derived from Yates' paper:

$$x=\frac{p_1P_1+p_2P_2+\cdots+p_mP_m-(m-1)T}{n+(m-1)-p_1-p_2-\cdots-p_m}$$

where

p_i=the number of classes in the ith group;

P_i=the sum of the values in the class of the ith set of classes containing the missing value;

n=the total number of observations in the complete design;

T=the total of the known observations.

This formula may be used for a factorial design. For example, a missing value in a 2^6 design may be estimated by

$$x=\frac{2P_1+2P_2+2P_3+2P_4+2P_5+2P_6-5T}{64+5-2-2-2-2-2-2}$$
$$=(2P_1+2P_2+2P_3+2P_4+2P_5+2P_6-5T)/57$$

If the missing value was abd, for instance,

P_1 Would be the sum of all results involving factor A at a high level,
P_2 Would be the sum of all results involving factor B at a high level,
P_3 Would be the sum of all results involving factor C at a low level,
P_4 Would be the sum of all results involving factor D at a high level,
P_5 Would be the sum of all results involving factor E at a low level,
P_6 Would be the sum of all results involving factor F at a low level.

In the Yates' formula, the denominator is

$$E=(n-1)-\{(p_1-1)+(p_2-1)+\cdots\cdots+(p_m-1)\}$$
=number of error degrees of freedom in the absence of any missing value.

Let M=general mean of known values
$=T/(n-1)$

then,

$$x = M - (n/E)(M - G)$$

where

$$G = M + (p_1(P_1 + M)/n - M) + (p_2(P_2 + M)/n - M) + \cdots + (p_m(P_m + M)/n - M)$$

and M − G is residual for missing point when M is inserted in place of a single missing value. In 2 − way complete design, if M is inserted in place of a single missing value, the residual for the missing unit is

$$= M - (P_1 + M)/p_2 - (P_2 + M)/p_1 + (T + M)/n$$
$$= M - p_1(P_1 + M)/n - p_2(P_2 + M)/n + \{M(n-1) + M\}/n$$
$$= (M - p_1(P_1 + M)/n) + \{M - p_2(P_2 + M)/n\}$$
$$= M - G$$

Healy and Westmacott (1956) gave a procedure for dealing with missing values in experiments analysed by computer. This procedure involves repeated use of the analysis routines needed when there are no missing values. Healy − Westmacott's iterative missing value procedure is following:

1. Inserts guessed values for the missing units.
2. Calculates residuals.
3. Form a new value for each missing unit by (guessed value − residual).
4. Residuals are then calculated afresh, and so on, until the sum of squares of residuals for missing units is small enough.

The equation

$$x = M - (n/E)(M - G)$$

show that the Healy− Westmacott process can be modified by subtracting at each stage not the residual but n/E times the residual. Yates and Anderson (1966) recommended for any new general program. Modification of Healy − Westmacott procedure proposed by Preece (1971) is following:

1. Set each of the n' missing values to M, multiplier m to n/E.
2. After each iteration, replace each estimated missing value by

   (current value) − m(current residual)

3. Calculate the ratio;

$$X' = \frac{(\text{S. S. of current residuals for missing plots})/n'}{(\text{S. S. of all residuals after first iteration})/E}$$

4 . If a value of X′ is small enough, stop.

Haseman and Gaylor (1973) showed an example that have 3 missing observations for which Preece (1971) achieved convergence in 3 iterations utilizing a computer. John and Prescott (1975) showed an example that have 2 missing observations.

3 . Estimation Missing Points in 2^n Factorial Designs

In this section we shall discuss some of the consequences of one or two missing points in 2^n factorial designs. It will be assumed throughout that the several facto-rs have coordinates $x_i = \pm 1$; the term "effect" will be used to denote either a main effect or an interaction, and by an estimate of an effect we shall mean an estimate of the corresponding regression coefficient, β.

If P and Q are effects in the same alias set. then PQ is a defining contrast, and, if **Q** is suppressed, **P** is estimable from the half replicate defined either by I = +PQ, or I = −PQ, in which it is aliased with **Q**. If there are p suppressed eff-ects in the same alias set as **P** we obtain p estimates of β_p from different half replicates.

Suppose that one point in a 2^n factorial design is missing and that we desire to analyze the incomplete design. One approach is to compute a missing plot value x, for the missing point. A reasonable procedure for doing this is to suppress one effect, usually the n factor interaction, in the model, and to choose x so as to make the contrast for this effect zero. This procedure is equivalent to the method used in randomized complete block designs of choosing x so as to minimize the sum of squares for error, and it gives the least squares estimates of the remaining e-ffects from the $2^n - 1$ actual points. In practice the simplest way to do this is to carry out Yates' algorithm with a zero entry for the missing point and then add $\pm\ x$ as appropriate in the last column. We illustrate this in the following exam-ple. John (1979) also showed similar example for 2^3 experimental design.

Table 1

Trt. comb	Data	(1)	(2)	(3)	(4)	Contrast Identification
(1)	15	41	80	$138 + x$	$314 + x$	M
d	26	39	$58 + x$	176	$-10 + x$	D
c	18	$28 + x$	86	$-6 + x$	$-10 - x$	C

cd	21	30	90	−4	$54-x$	CD
b	28	42	14	$-x$	$-18+x$	B
bd	x	44	$-20+x$	−10	$-30+x$	BD
bc	11	51	−4	$28-x$	$-10-x$	BC
bcd	19	39	0	26	$46-x$	BCD
a	25	11	−2	$-22+x$	$38-x$	A
ad	17	3	$2-x$	4	$2-x$	AD
ac	20	$-28+x$	2	$-34+x$	$-10+x$	AC
acd	24	8	−12	4	$-2+x$	ACD
ab	29	−8	−8	$4-x$	$26-x$	AB
abd	22	4	$36-x$	−14	$38-x$	ABD
abc	16	−7	12	$44-x$	$-18+x$	ABC
abcd	23	7	14	2	$-42+x$	ABCD

In Example, bd is missing. The estimated missing value is $x=30$. Hence, estimated effect of **A** is

$$\hat{\beta}_1=(38-30)/8=1 \tag{1}$$

From the single half replicate defined by **I=+ABD**,

$$\hat{\beta}_1=(1/4)\ (abcd+abd+ac+a-bc-b-cd-d)=1 \tag{2}$$

This is the same as the (1). In our example we should now suppress **BCD**, **ACD**, **ABD**, **ABC** and **ABCD**, and minimize

$$S=(BCD)^2+(ACD)^2+(ABD)^2+(ABC)^2+(ABCD)^2$$
$$=(46-x)^2+(-2+x)^2+(38-x)^2+(-18+x)^2+(-42+x)^2$$

Let $\frac{\partial S}{\partial x}=0$, then, $x=29.2$

The least squares estimate of β_1 becomes

$$\hat{\beta}_1^*=(38-29.2)/8=1.1 \tag{3}$$

The estimate of β_1 from the half replicate defined by

I=+BC(A~ABC)

is

$$(abcd+abc+ad+a-bcd-bc-d-(1))/4=2.5 \tag{4}$$

The estimate of β_1 from the half replicate defined by

I=−BD (A~ABD)

is

$$(abc+ab+acd+ad-bc-b-cd-d)/4=0 \tag{5}$$

The estimate of β_1 from the half replicate defined by

$I = +CD\ (A \sim ACD)$

is

$$(abcd + ab + acd + a - bcd - b - cd - (1))/4 = 4.5 \tag{6}$$

The estimate of β_1 from the half replicate defined by

$I = -ABCD\ (A \sim BCD)$

is

$$(abc + abd + acd + a - bcd - b - c - d)/4 = -1 \tag{7}$$

The estimate of β_1 from the half replicate defined by

$I = +BCD\ (A \sim ABCD)$

is

$$(abcd + ab + ac + ad - bcd - b - c - d)/4 = -0.5 \tag{8}$$

From (4), (5), (6), (7) and (8), average $\hat{\beta}_1^*$ is

$$\hat{\beta}_1^* = (2.5 + 0 + 4.5 + (-1) + (-0.5))/5 = 1.1 \tag{9}$$

That is the same as the (3).

Now we assumed that two points a and cd are missing. Two missing values are calculated, x for a and y for cd. Yates' algorithm is performed with zero for a and cd, and ± x, ± y are added as appropriate.

Table 2

Trt. comb	Data	(1)	(2)	(3)	(4)	Contrast Identification
(1)	15	41	59+y	139+y	290+x+y	M
d	26	18+y	80	151+x	16−x+y	D
c	18	50	61+x	−5+y	−28−x+y	C
cd	y	30	90	21−x	−14+x+y	CD
b	28	17+x	−7+y	−43+y	50−x−y	B
bd	22	44	2	15−x	−12+x−y	BD
bc	11	51	21−x	−15+y	−36+x−y	BC
bcd	19	39	0	1+x	70−x−y	BCD
a	x	11	−23+y	21−y	12+x−y	A
ad	17	−18+y	−20	29−x	26−x−y	AD
ac	20	−6	27−x	9−y	58−x−y	AC
acd	24	8	−12	−21+x	16+x−y	ACD
ab	29	17−x	−29+y	3−y	8−x+y	AB
abd	22	4	14	−39+x	−30+x+y	ABD
abc	16	−7	−13+x	43−y	−42+x+y	ABC
abcd	23	7	14	27−x	−16−x+y	ABCD

Suppressing all five interactions **BCD**, **ACD**, **ABD**, **ABC** and **ABCD**, and we choose $\mathbf{x}$ and $\mathbf{y}$ so as to minimize.

$$S=(70-x-y)^2+(16+x-y)^2+(-30+x+y)^2+(-42+x+y)^2+(-16-x+y)^2$$

Let $\frac{\partial S}{\partial x}=0$, and $\frac{\partial S}{\partial y}=0$, then we find

$$x=47/3,\ \text{and}\ y=95/3.$$

Hence, $\hat{\beta}_1=(12+x-y)/8=-0.5$

$$\hat{\beta}_2=(50-x-y)/8=1/3$$

$$\hat{\beta}_3=(-28-x+y)/8=-1.5$$

$$\hat{\beta}_4=(16-x+y)/8=4$$

Also $\hat{\beta}_1$ is estimable from the half replicate in which **A** is aliased with **ABCD** (I= + BCD) ; i.e.,

$$\hat{\beta}_1=(abcd+ab+ac+ad-bcd-b-c-d)/4=-0.5$$

4. Estimation Missing Points and Effects in Fractional Factorial Designs

Shearer (1973) proposed the method of estimating missing points and effects in the fractional factorial design using the iterative procedure. Now we use Shearer's notation below.

Suppose we have a two-level design with independent variables $x_{ij}\,(\pm 1)$, dependent variable y_i and a single missing value. :

$$x_{11}\quad x_{12}\cdots\cdots x_{1m}\quad y_1$$

$$x_{21}\quad x_{22}\cdots\cdots x_{2m}\quad y_2$$

....

$$x_{miss1}\quad x_{miss2}\quad\cdots x_{missm}\quad y_{miss}$$

....

$$x_{n1}\quad x_{n2}\ \cdots\cdots x_{nm}\quad y_n$$

The m variables have main effects $b_j\,(j=1,\cdots\cdots,m)$. The procedure is a two stage iterative one:

$$y_{miss}^{(i+1)}=\sum_{j=1}^{m} b_j^{(i)} x_{missj}+\sum_{k=1}^{n} y_k^{(i)}/n \qquad (1)$$

$$b_j^{(i)}=\sum_{k=1}^{n} x_{kj} y_k^{(i)}/n,\quad (j=1,\cdots\cdots,m) \qquad (2)$$

Equation (2) is just the usual main effect estimation procedure for a 2^m design, and equation (1) re-estimates the missing value using the new main effect estimates. This latest estimate of y_{miss} is used to re-estimate the main effects and the cycle is repeated. Any starting value $y_{miss}^{(0)}$ is used to initiate the process. After the first iteration the calculations are very simple, as $y_{miss}^{(l)}$ and $b_j^{(l)}$ are the only values to change on the right-hand side of equations (1) and (2).

The procedure will minimize the residual sum of squares as required provided that it converges as this implies that the estimated value lies on the regression line constructed through the remaining $n-1$ points.

Define $\underline{b}'^{(l)} = (b_1^{(l)}, b_2^{(l)}, \cdots\cdots, b_m^{(l)})$ and

$$\underline{X}'_{miss} = (x_{miss1}, x_{miss2}, \cdots\cdots, x_{missm})$$

Then

$$\sum_{k=1}^{n} y_k^{(l-1)} - y_{miss}^{(l-1)} = \sum_{k=1}^{n} y_k^{(l-2)} - y_{miss}^{(l-2)}$$

$$\sum_{k=1}^{n} y_k^{(l-1)} = \sum_{k=1}^{n} y_k^{(l-2)} + y_{miss}^{(l-1)} - y_{miss}^{(l-2)} \tag{3}$$

$$b_j^{(l-1)} = \sum_{k=1}^{n} x_{kj} y_k^{(l-1)}/n, \quad b_j^{(l-2)} = \sum_{k=1}^{n} x_{kj} y_k^{(l-2)}/n, \quad \text{and}$$

$$\sum_{k=1}^{n} x_{kj} y_k^{(l-1)} - x_{missj} y_{miss}^{(l-1)} = \sum_{k=1}^{n} x_{kj} y_k^{(l-2)} - x_{missj} y_{miss}^{(l-2)}$$

$$\therefore b_j^{(l-1)} - b_j^{(l-2)} = (x_{missj} y_{miss}^{(l-1)} - x_{missj} y_{miss}^{(l-2)})/n$$

$$\underline{b}'^{(l-1)} - \underline{b}'^{(l-2)} = \underline{X}'_{miss} y_{miss}^{(l-1)}/n - \underline{X}'_{miss} y_{miss}^{(l-2)}/n$$

$$\therefore \underline{b}'^{(l-1)} = \underline{b}'^{(l-2)} + (y_{miss}^{(l-1)} - y_{miss}^{(l-2)}) \underline{X}'_{miss}/n \tag{4}$$

From (3) and (4), the first equation (1) is

$$y_{miss}^{(l)} = \underline{b}'^{(l-1)} \underline{X}_{miss} + \sum_{k=1}^{n} y_k^{(l-1)}/n$$

$$= (\underline{b}'^{(l-2)} + (y_{miss}^{(l-1)} - y_{miss}^{(l-2)}) \underline{X}'_{miss}/n) \underline{X}_{miss} + \sum_{k=1}^{n} y_k^{(l-2)}/n + (y_{miss}^{(l-1)} - y_{miss}^{(l-2)})/n$$

$$= \underline{b}'^{(l-2)} \underline{X}_{miss} + \sum_{k=1}^{n} y_k^{(l-2)}/n + (y_{miss}^{(l-1)} - y_{miss}^{(l-2)})\ (1 + \underline{X}'_{miss} \underline{X}_{miss})/n$$

$$= y_{miss}^{(l-1)} + (y_{miss}^{(l-1)} - y_{miss}^{(l-2)})\ (1 + \underline{X}'_{miss} \underline{X}_{miss})/n$$

$$\therefore y_{miss}^{(l)} - y_{miss}^{(l-1)} = (y_{miss}^{(l-1)} - y_{miss}^{(l-2)})\ (1 + \underline{X}'_{miss} \underline{X}_{miss})/n$$

$$\frac{y_{miss}^{(l)} - y_{miss}^{(l-1)}}{y_{miss}^{(l-1)} - y_{miss}^{(l-2)}} = \frac{1 + \underline{X}'_{miss} \underline{X}_{miss}}{n} = \frac{1+m}{n} \tag{5}$$

As $m < n-1$ for any estimation to be possible.

$$0 < (y_{miss}^{(l)} - y_{miss}^{(l-1)}) / (y_{miss}^{(l-1)} - y_{miss}^{(l-2)}) < 1 \tag{6}$$

convergence is thus assured.

Suppose that Y is the limit of the sequence $y_{miss}^{(l)}$ and that $p = (n-m-1)/n$, equation (5) becomes

$$(y^{(i)}_{miss}-y^{(i-1)}_{miss})\,/\,(y^{(i-1)}_{miss}-y^{(i-2)}_{miss})=1-p$$

and thus

$$(y^{(i)}_{miss}-y^{(i-1)}_{miss})\,/\,(Y-y^{(0)}_{miss})=p\,(1-p)^{i-1} \tag{7}$$

as

$$\sum_{i=1}^{\infty}(y^{(i)}_{miss}-y^{(i-1)}_{miss})\,/\,(Y-y^{(0)}_{miss})=1$$

From (7),

$$Y=(y^{(1)}_{miss}-y^{(0)}_{miss})\,/\,p+y^{(0)}_{miss} \tag{8}$$

If $y^{(0)}_{miss}=0$, (8) becomes

$$Y=y^{(1)}_{miss}/p \tag{9}$$

Example. Suppose we have a half replicate of a 2^4 design $I=+ABCD$ with one missing value (bd).

Experiment	(1)	ab	ac	ad	bc	bd	cd	abcd
Response	3	5	5	6	7	*	10	8

In this case $n=8$, $m=4$, $p=(n-m-1)/n=3/8$

If $y^{(0)}_{miss}=0$, we can use (9). The main effects are

$$A=(5+5+6+8-3-7-0-10)/8=0.5$$
$$B=(5+7+0+8-3-5-6-10)/8=-0.5$$
$$C=(5+7+10+8-3-5-6-0)/8=2$$
$$D=(6+0+10+8-3-5-5-7)/8=0.5$$
$$\bar{y}=44/8=5.5$$

From (1),

$$y^{(1)}_{miss}=\bar{y}-A+B-C+D=3$$

Using (9),

$$Y=y^{(1)}_{miss}/p=3\,(8/3)=8$$

Thus, main effects are

$$\begin{aligned}A&=(5+5+6+8-3-7-8-10)/8=-0.5\\B&=(5+7+8+8-3-5-6-10)/8=0.5\\C&=(5+7+10+8-3-5-6-8)/8=1\\D&=(6+8+10+8-3-5-5-7)/8=1.5\end{aligned} \tag{10}$$

If $y^{(0)}_{miss}=6$, the main effects are

$$A=(5+5+6+8-3-7-6-10)/8=-0.25$$
$$B=(5+7+6+8-3-5-6-10)/8=0.25$$

$C=(5+7+10+8-3-5-6-6)/8=1.25$

$D=(6+6+10+8-3-5-5-7)/8=0.8$

$\overline{y}=50/8$

From (1),

$y_{miss}^{(1)}=\overline{y}-A+B-C+D=27/4$

Using (8),

$Y=(y_{miss}^{(1)}-y_{miss}^{(0)})/p+y_{miss}^{(0)}$

$=(27/4-6)(8/3)+6=8$

Thus, main effects are

$$A=-0.5,\ B=0.5,\ C=1,\ D=1.5 \qquad (11)$$

That is the same as the (10).

Now we prove the value of Y is constant whatever $y_{miss}^{(0)}$ may be. In this example, let $y_{miss}^{(0)}=k$. Then

$A=(4-k)/8,\ B=(-4+k)/8$

$C=(16-k)/8,\ D=(4+k)/8$

$\overline{y}=(44+k)/8$

and $y_{miss}^{(1)}=\overline{y}-A+B-C+D=(24+5k)/8$

Using (8),

$Y=\{(24+5k)/8-k\}(8/3)+k=8.$

In the case of more than one missing value, equations (1) and (2) must be used as they stand. All missing values should be estimated followed by re-estimation of all regression coefficients. These steps are repeated until convergence is achieved. Clearly this will again minimize the residual sum of squares. The fraction p can be thought of as a rate of convergence constant for a given design, and, in fact, the rates of convergence of several missing values approximate very closely to the value p in a reasonable design even though p is strictly valid only for a single missing observation.

References

1. Allen, F. E. and Wishart, J. (1930), ***A method of estimating the yield of a missing plot in field experimental work.*** J. Agric. Sci., 20, 399-406.
2. Draper, N. R. and Stoneman, D. M. (1963), ***Estimating missing values in unre-***

plicated two-level factorial and fractional factorial designs. University of Wisconsin Technical Report No. 20.

3. Haseman, J. K. and Gaylor, D. W. (1973), *An algorithm for non-iterative estimation of multiple missing values for crossed classifcations.* Technometrics, vol. 15, No. 3. 631-636.
4. Healy, M. J. R. and Westmacott, M. H. (1956), *Missing values in experiments analysed on automatic computers.* Appl. Statist. 5, 203-206.
5. John, P. W. M. (1969), *Some non-orthogonal fractions of 2^n designs.* J. Roy. Statist. Soc., B, 31, 270-275.
6. John, P. W. M. (1971), *Statistical design and analysis of experiments.* New York, The Macmillan Company.
7. John, P. W. M. (1979), *Missing points in 2^n and 2^{n-k} factorial designs.* Technometrics, vol. 21, No. 2. 225-228.
8. John, J. A. and Prescott, P. (1975), *Estimating missing values in experiments.* Appl. Statist. 24, No. 2, 190-192.
9. Preece, D. A. (1971), *Iterative procedures for missing values in experiments.* Technometrics, vol. 13, No. 4, 743-753.
10. Shearer, P. R. (1973), *Missing data in quantitative designs.* Appl. Statist. vol. 22, 135-140.
11. Tocher, K. D. (1952), *The design and analysis of block experiments.* J. R. Statist, Soc. B, 14, 49-51.
12. Wilkinson, G. N. (1958 a), *Estimation of missing values for the analysis of incomplete data.* Biometrics, 14, 257-286.
13. Wilkinson, G. N. (1958 b), *The analysis of variance and deviation of standard errors for incomplete data.* Biometrics, 14, 360-384.
14. Yates, F. (1933), *The analysis of replicated experiments when the field results are incomplete.* Emp. J. Exp. Agric. 1, 129-142.
15. Yates, F. and Anderson, A. J. B. (1966), *A general computer program for the analysis of factorial experiments.* Biometrics, 22, 503-524.

Chonnam National University