# Missing Plot Technique

*by*

*Tanmay Kumar Maity*

Due to accident, mishandling, attack of pests (for agricultural experiments) or some other reason one or more observations from an experiment may be lost or the value may be suspicious so that it is wise to treat it as absent.

The correct procedure is to write down the observational equations for the available observations and to perform a least square analysis. But the resulting normal equations are very difficult to solve owing to the missing observations. Yates (1937) considered a method of estimating the missing values, inserting the estimates and analysing the data. This technique gives results identical with those obtained by the correct procedure. The theory has been developed under the assumption that all the treatment contrasts are estimable under missingness.

Suppose k values denoted by $x_1,...,x_k$ are missing from an experiment. The missing plot technique is consisting of the following steps:

**Step 1:** Write down the error sum of squares (SSE) using the available observations and $x_i$, $i=1(1)k$ for the missing observations. This will be a quadratic form E(x), (say), in $x=(x_1,...,x_k)$.

**Step 2:** Minimize E(x) w.r.t. $x_1,...,x_k$ by solving the k normal equations $\frac{\partial E(x)}{\partial x_i}=0, i=1(1)k.$

Let $\hat{x}_i$ be the least square estimate of $x_i$, i=1(1)k

**Step 3:** An analysis of variance is carried out by using $\hat{x}_i$ in place of $x_i$ .From the analysis we get,

i) $SS_{TR}(\hat{x})$ = Sum of squares (SS) due to treatment using $\hat{x}$ for $x$ with d.f. $\nu_{TR}$ and

ii) $SSE(\hat{x})$ = SS due to error using $\hat{x}$ for $x$ with d.f. $\nu_e$-k

Where $\nu_e$ and $\nu_{TR}$ are the error d.f. and d.f. due to treatment in a table with full observations (having no missing value).

SS due to other sources are computed with the modifications that total sum of squares (TSS) carries d.f. $\nu_T$ - k.

**Step 4:** An approximate test of the equality of treatment effects is performed using the statistic: $F_{01}=\frac{SS_{tr}(\hat{x})/\nu_{TR}}{SSE(\hat{x})/\nu_e-k} \sim F_{\nu_{TR},\nu_e-k}$ under null hypothesis $H_0$

This test can be shown to be biased, since

$$E[MS_{TR}(\hat{x})] > E[MSE(\hat{x})] \text{ under } H_0$$

If $H_0$ is accepted by this biased test, there is no need to perform a more accurate test. On the other hand, if this test comes out to be significant it is not commented immediately in favour of rejection of $H_0$, but proceeds to the next step.

**Step 5:** Calculate restricted estimator of *x* under $H_0$, by minimizing the error SS under $H_0$. Let the estimate of *x* be $\tilde{x} = (\tilde{x}_1, ..., \tilde{x}_k)$.

**Step 6:** Carry out ANOVA using $\tilde{x}$ for the missing values. For this, find SS due to all sources (except treatment), $TSS(\tilde{x})$ and $SSE(\tilde{x})$. Obtain adjusted treatment SS due to $H_0$,

$SS_{H_0} = SSE(\tilde{x}) - SSE(\hat{x})$ with d.f. $\nu_{TR}$.

**Step 7:** Carry out the test of significance for $H_0$ using the statistic

$$F_{02} = \frac{SS_{H_0} / \nu_{TR}}{SSE(\hat{x}) / (\nu_e - k)} \sim F_{\nu_{TR}, \nu_e - k} \text{ under } H_0$$

If this $F_{02}$ comes out to be significant we can reject $H_0$ at chosen level. Otherwise we have to accept $H_0$ and make our comment accordingly.

## Analysis of RBD with missing data:

Suppose without loss of generality that observation for block 1 and treatment 1, i.e., $y_{11}$ is missing in a RBD with t treatments and r blocks. Let $y_{11}$ = x. We calculate

$B_1^{'} = \sum_{j(\neq)1} y_{1j}$ = total of all available observations for block 1

$T_1^{'} = \sum_{i(\neq)1} y_{i1}$ = total of all available observations for treatment 1

$G^{'} = \sum_{i} \sum_{j}{}_{(i,j)\neq(1,1)} y_{ij}$ = total of all available (rt-1) observations

We calculate different SS's as follows:

$$SS_{BL} = \frac{(B_1^{'}+x)^2 + \sum_{2}^{r} B_j^2}{t} - \frac{(G^{'}+x)^2}{rt}$$

$$SS_{TR} = \frac{(T_1^{'}+x)^2 + \sum_{2}^{t} T_i^2}{r} - \frac{(G^{'}+x)^2}{rt}$$

$$TSS = \sum_{i}\sum_{\substack{j \\ (i,j)\neq(1,1)}} y_{ij}^2 + x^2 - \frac{(G^{'}+x)^2}{rt}$$

$$SSE = TSS - SS_{BL} - SS_{TR}$$

$$= x^2 + \frac{(G^{'}+x)^2}{rt} - \frac{(B_1^{'}+x)^2}{t} - \frac{(T_1^{'}+x)^2}{r} + \text{terms independent of } x$$

We minimize this SSE w.r.t. *x*. This is obtained by solving the equation

$$\frac{d(SSE)}{dx} = 2x + \frac{2(G^{'}+x)}{rt} - \frac{2(B_1^{'}+x)}{t} - \frac{2(T_1^{'}+x)}{t} = 0$$

which gives, $\hat{x} = \dfrac{rB_1^{'} + tT_1^{'} - G^{'}}{(r-1)(t-1)}$

$\hat{x}$ is the least square estimate of the missing observation.

An approximate test for the null hypothesis H$_0$ ( $\tau_1 = \tau_2 = ... = \tau_t$ ) is obtained using the test statistic $F_{01} = MS_{TR}(\hat{x}) / MSE(\hat{x}) \sim F_{t-1,(r-1)(t-1)-1}$ under H$_0$, one error d.f. being lost due to estimation of $y_{11}$. This test is, however, biased in the sense that expectation of treatment MS is greater than the expectation of error MS under null hypothesis.

If the approximate test does not reject H$_0$, there is no need to perform the accurate test of significance which is given below:

We find least square estimate of $y_{11}$ under H$_0$ ( $\tau_1 = \tau_2 = ... = \tau_t$ ). The expression for SSE is given by, $SSE = TSS - SS_{BL} = x^2 - \dfrac{(B_1^{'}+x)^2}{t}$ + terms not involving *x*

The minimization of SSE w.r.t. *x* gives the estimate of *x* as $\tilde{x} = \dfrac{B_1^{'}}{t-1}$.

Using this estimated value $\tilde{x}$, find $SS_{BL}(\tilde{x}), TSS(\tilde{x})\,\&\, SSE(\tilde{x})$ (with d.f. rt-r-1). The SS due to H$_0$, $SS_{H_0} = SSE(\tilde{x}) - SSE(\hat{x}), d.f. = t-1$.

The statistic for testing H$_0$ is given by, $F_{02}=\dfrac{SS_{H_0}/(t-1)}{SSE(\hat{x})/(rt-r-t)}\sim F_{t-1,rt-r-t}$ under H$_0$.

In the general case, when the observation $y_{kl}$, corresponding to kth block and lth treatment is missing, the least square estimate for the plot is given by,

$\hat{x}=\dfrac{rB_k^{'}+tT_l^{'}-G^{'}}{(r-1)(t-1)}$ where $B_k^{'}\ \&T_l^{'}$ are sum of all available observations for the kth block and the lth treatment respectively and $G^{'}$ is the total of all available observations.

**Two missing observations:**

For two missing values, say *x* and *y*, let B$_1$ and B$_2$ be the total of known observations in the blocks containing x and y respectively and T$_1$ and T$_2$ be the totals of known observations in the treatments containing x and y respectively. And let G be the total of all the known observations. Therefore error SS becomes

$E=x^2+y^2-\frac{1}{t}[(B_1+x)^2+(B_2+y)^2]-\frac{1}{r}[(T_1+x)^2+(T_2+y)^2]+\frac{1}{rt}(G+x+y)^2$ + terms independent of x.

For a minimum of SSE w.r.t. *x* and y, we must have

$$\left.\begin{aligned}\frac{\partial E}{\partial x}&=0=x-\tfrac{1}{t}(B_1+x)-\tfrac{1}{r}(T_1+x)+\tfrac{1}{rt}(G+x+y)\\ \frac{\partial E}{\partial y}&=0=y-\tfrac{1}{t}(B_2+x)-\tfrac{1}{r}(T_2+x)+\tfrac{1}{rt}(G+x+y)\end{aligned}\right\}$$

$$\Rightarrow\begin{cases}(r-1)(t-1)x=rB_1+tT_1-G-y\\(r-1)(t-1)y=rB_2+tT_2-G-x\end{cases}$$

Solving these equations simultaneously, we get the estimates of *x* and *y*.

Similarly we can obtain the normal equations for the estimation of more than two missing observations and carry out the analysis accordingly.

## Analysis of LSD with missing data:

Suppose that the yield of the plot in row 1, column 1and receiving treatment 1, say $y_{111}$ is missing. Let $R_1^{'},C_1^{'},T_1^{'},G^{'}$ denote, respectively, the total of available yield figures for row 1, column 1, treatment 1, and for the whole table respectively and $R_i(i\neq 1),C_j(j\neq 1),T_k(k\neq 1)$ denote total for the ith row, jth column and kth treatment respectively. The least square estimate of $y_{111}$ = *x* (say) is obtained by minimizing the error SS w.r.t *x*. Now, Row SS (SSR) = $\frac{1}{m}(R_1^{'}+x)^2+\frac{1}{m}\sum_{2}^{m}R_i^2-\frac{(G^{'}+x)^2}{m^2}$

Column SS (SSC) = $\frac{1}{m}(C_1^{'}+x)^2+\frac{1}{m}\sum_{2}^{m}C_j^2-\frac{(G^{'}+x)^2}{m^2}$

Treatment SS ($SS_{TR}$) = $\frac{1}{m}(T_1^{'}+x)^2+\frac{1}{m}\sum_{2}^{m}T_k^2-\frac{(G^{'}+x)^2}{m^2}$

Total SS (TSS) = $x^2+\sum_{(i,j,k)[\neq(1,1,1)]\in D} y_{ijk}^2+\frac{(G^{'}+x)^2}{m^2}$

Therefore error SS is obtained by,

SSE($x$) = TSS – SSR – SSC - $SS_{TR}$

$$= x^2+\frac{2(G^{'}+x)^2}{m^2}-\frac{1}{m}[(R_1^{'}+x)^2+(C_1^{'}+x)^2+(T_1^{'}+x)^2] \text{ + terms independent of } x.$$

The minimizing equation $\frac{dSSE(x)}{dx}=0$ gives, $\hat{x}=\frac{mR_1^{'}+mC_1^{'}+mT_1^{'}-2G^{'}}{(m-1)(m-2)}$.

In general if the observation $y_{uvw}$ is missing, $\hat{y}_{uvw}=\frac{m(R_u^{'}+C_v^{'}+T_w^{'})-2G^{'}}{(m-1)(m-2)}$

We can substitute this $\hat{x}$ for the missing value and perform analysis of variance with the modification of SSE and TSS carries $(m-1)(m-2)-1 = m^2-3m+1$ d.f. and $m^2-2$ d.f. respectively. The usual test for the null hypothesis $H_0$ ($\tau_1=\tau_2=...=\tau_m$) is obtained using the test statistic $F_{01}=MS_{TR}(\hat{x})/MSE(\hat{x})\sim F_{m-1,m^2-3m+1}$ under $H_0$. But the test will have an upward bias. If $F_{01}$ is not significant we accept $H_0$. If $F_{01}$ is significant it is not sure whether this is due to bias or it is due to differential effects of the treatments. In this case one may perform an exact test as follows:

The analysis is first performed with the original data (excluding the missing data) by breaking up the total SS into SS's due to rows, columns and (treatment + error) as follows:

| **Sources of variation** | **d.f.** | **SS** |
|---|---|---|
| Rows | $m-1$ | $\frac{R_1^{'2}}{m-1}+\sum_{2}^{m}\frac{R_i^2}{m}-\frac{G^{'2}}{m^2-1}$ |
| Columns | $m-1$ | $\frac{C_1^{'2}}{m-1}+\sum_{2}^{m}\frac{C_i^2}{m}-\frac{G^{'2}}{m^2-1}$ |
| Treatment and Errors | $m^2-2m$ | W ( By subtraction) |
| Total | $m^2-2$ | $\sum_{(i,j,k)[\neq(1,1,1)]\in D} y_{ijk}^2-\frac{G^{'2}}{m^2-1}$ |

The quantity $\frac{W-SSE(\hat{x})}{m-1}$ is the mean square due to treatment, where $SSE(\hat{x})$ is the error SS in the augmented table. The statistic

$$\left[\frac{W-SSE(\hat{x})}{m-1}\right] / \left[\frac{SSE(\hat{x})}{m^2-3m+1}\right] \sim F_{(m-1),(m^2-3m+1)} \text{ under } H_0$$

Alternatively, if the preliminary test with the augmented table shows $F_{01}$ is significant one may perform the more accurate test as follows. We find least square estimate of $y_{111}$ under $H_0$ ($\tau_1 = \tau_2 = ... = \tau_m$). This is obtained by minimizing the restricted SSE which is given by,

$SSE^*(x)$ = TSS – SSR – SSC = $x^2 + \frac{(G^{'}+x)^2}{m^2} - \frac{1}{m}[(R_1^{'}+x)^2 + (C_1^{'}+x)^2] +$ terms independent of *x*.

Now $\frac{dSSE^*(x)}{dx} = 0$ gives $\tilde{x} = \frac{m(R_1^{'}+C_1^{'})-G^{'}}{(m-1)^2}$ .

Using this estimated value, $\tilde{x}$, find SSR, SSC, TSS and SS due to error = SSE($\tilde{x}$) with d.f. *m(m-2)*. The SS due to $H_0$ ($SS_{H_0}$) is $SSE(\tilde{x}) - SSE(\hat{x})$, d.f. = *m-1*. Therefore, the testing the equally of treatment mean is done by the test statistic

$$\frac{SS_{H_0}/(m-1)}{SSE(\hat{x})/(m^2-3m+1)} \sim F_{(m-1),(m^2-3m+1)} \text{ under } H_0.$$

The same procedure may be followed for estimating more than one, say k missing values and where the missing values are obtained by solving k equations simultaneously.

----------***----------