METHODOLOGY FOR ESTIMATING MISSING INFORMATION AND ANOVA CORRECTION WITH THE CELL MEANS MODEL IN CONNECTED DESIGNS

Diana C. Franco SOTO[1]
Oscar O. Melo MARTÍNEZ[1]

▪ *ABSTRACT: In this paper we present a non-iterative methodology to estimate missing data in connected cell means designs. This methodology improves on that of Melo (2002), reducing the correlation between observed and estimated information. After imputing the missing information, we suggest a way to analyze the variance by accounting for the subvector's covariance structure of the information vector. This analysis guarantees that the test statistic follows a central F distribution under H_0. Finally, we work on a case study of a factorial design, in which the proposed methodology is applied.*

▪ *KEYWORDS: Cell means model; connected designs; missing information; estimation and imputation; distribution of quadratic forms.*

1 Introduction

In experimental design, after carrying out the relevant field proofs, it is possible that some of the observable or experimental units are lost. This might happen for unexpected reasons, even though the researcher carefully watches over the process. The existence of that missing information in some cases can invalidate the experimental process and make the use of test statistics intented to verify the hypotheses of interest, even more complex.

In multifactorial cell means designs, the presence of empty cells implies that not every means can be estimated and some hypotheses concerning the linear combinations of the parameters are out of place. In such situations the designed experiment is no longer balanced and loses its symmetry and orthogonality -though it is often assumed- and this brings about the need for a clear identification of the

[1]Department of Statistics, Universidad Nacional de Colombia, Bogotá, D. C., Colombia. E-mail: *dcfrancos@unal.edu.co / oomelom@unal.edu.co*

functions that can be estimated, whose nature depends on whether the model is connected or not. Regarding this latter aspect of concern, it is worth mentioning the works by Weeks and Williams (1964), Murray and Smith (1985) and Dodge (1985) in models with n-ways of classification.

The lack of information in experimental design is a subject to which particular attention has been paid over the past years, see for example, Allan and Wishart (1930), Yates (1933), Bartlett (1937), John and Prescott (1975), Jarrett (1978), Little (1988), Liu (1996), Little and Rubin (2002). One of the most recent works about this is presented in Melo (2002), in which three methodologies for estimation of empty cells with the cell means model in connected designs are proposed.

The theoretical results in the aforementioned study suggest and guide the work developed herein. That is because, in this study, as most studies regarding missing information, after carrying out the imputation, the common analysis of variance is performed. However, a detailed study of the distributional properties of the test statistic is lacking. Furthermore, the correlation and appropriate structure of the covariability that exists between the observed and estimated information is also omitted. Such omission might bring about serious consequences regarding the validity of the results obtained for the test of hypotheses, and that may lead to erroneous conclusions.

With the aim of overcoming the foregoing difficulties, in this work we propose a methodology for handling missing information in connected cell means designs. In the next section some basic ideas about linear models are presented. We present a non-iterative methodology to estimate missing information in connected designs with the cell means model in section 3. This methodology improves on that of Melo (2002), reducing the correlation between observed and estimated information. After imputing the missing information, we suggest a way to analyze the variance by accounting for the subvector's covariance structure of the information vector. This analysis guarantees that the test statistic follows a central F under H_0. In the fourth section, we show a numerical example on a case study of a factorial design, in which the proposed techniques are applied. Finally, we present some conclusions about the work in the last section.

2 Preliminary concepts

The methodology for the theoretical development of the problem is based on the cell means model and modified cell means model, which are presented in this section.

The cell means model presented in Searle (1987), Hocking (1985) and Hocking (1996), is given by

$$y = W\mu + e \tag{1}$$

subject to the constraint

$$G\mu = g$$

where $y_{(n\times 1)}$ is the vector of random variables, $W_{(n\times p)}$ is the incidence matrix that associates each observation with its average values, $\mu_{(p\times 1)}$ is the vector of population means, $G_{(s\times p)}$ is the contrast matrix representing the known linear relations from the cell means and in which the noninteraction constraint are usually specified such that $\text{rank}(G) = s$, $g_{(s\times 1)}$ is the vector of unknown constants, and $e_{(n\times 1)}$ is the vector of non observable random variables such that $e \sim N(0_{(n\times 1)}, \sigma^2 I_n)$. Adittionally, n is the number of total observations, m is the number of missing data, p is the number of parameters (population means) and s is the number of linearly independent restrictions over the cell means.

In cases where there are no empty cells, W has a full column rank. Otherwise W has a zero column for each empty cell, whereas μ mantains its original structure as if every cell had been observed.

The modified cell means model is a special case of model (1), with

$$G\mu = 0 \tag{2}$$

In particular, the modified cell means model is useful in cases with empty cells, because this model reduces the dimensionality of the original model by substituting the constraint set (2) in this model (Murray and Smith, 1985).

The constraint matrix $G_{(s\times p)}$ can be reorganized, reordered and partitioned into two submatrices, i.e. $G = (G_1 \vdots G_2)$, where $G_{1(s\times(p-s))}$ and $G_{2(s\times s)}$ is such that $\text{rank}(G_2) = s$. The partition of vector μ is independent of the data obtained, and, in particular, it is independent of the number of missing cells and its location.

The rows of $\mu_{(p\times 1)}$ can be reorganized in agreement with the partitioning of G, giving $\mu = (\mu_1 \vdots \mu_2)$, where $\mu_{1((p-s)\times 1)}$ and $\mu_{2(s\times 1)}$.

Thus, $G\mu = G_1\mu_1 + G_2\mu_2 = 0$, and since G_2 is a full column rank matrix, there is a unique solution for μ_2 in terms of μ_1 given by

$$\mu_2 = -G_2^{-1}G_1\mu_1 \tag{3}$$

Finally, this leads to a reordering of W, the incidence matrix, which should be in agreement with the partition of both G and μ. Hence we obtain

$$\begin{aligned} y &= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + e \\ &= \begin{pmatrix} w_1 & w_2 \end{pmatrix} \begin{pmatrix} I \\ -G_2^{-1}G_1 \end{pmatrix} \mu_1 + e \\ &= V\mu_1 + e \end{aligned} \tag{4}$$

where $V_{(n\times(p-s))} = w_1 - w_2 G_2^{-1} G_1$.

If the rank of V is $p-s$, then $V'V$ is non-singular and the following estimators for the components of the vector μ are obtained

$$\tilde{\mu}_1 = (V'V)^{-1}V'y \tag{5}$$

$$\tilde{\mu}_2 = -G_2^{-1}G_1\tilde{\mu}_1 \tag{6}$$

2.1 The constrained least squares non-iterative method of estimation

If there are m missing values, the cell means model can be written as

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \mu + e \tag{7}$$

where the vector of random variables $y_{(n\times 1)}$ is partitioned into two subvectors consisting of the observed information $y_{1((n-m)\times 1)}$ and another vector associated with the missing information $y_{2(m\times 1)}$ (which could be a zero vector or another vector with presupposed initial values). Besides, $W_{(n\times p)}$ is also partitioned into two submatrices associated with the observed and non-observed information, $W_{1((n-m)\times p)}$ and $W_{2(m\times p)}$ respectively.

If a subsequent partition is carried out on the submatrices W_1 and W_2, taking into account the restriction matrix G for the modified cell means model, then

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + e \tag{8}$$

with $W_{11((n-m)\times(p-s))}$, $W_{12((n-m)\times s)}$, $W_{21(m\times(p-s))}$ and $W_{22(m\times s)}$.

Substituting μ_2 by (3), we obtain

$$y_1 = W_{11}\mu_1 - W_{12}G_2^{-1}G_1\mu_1 + e_1 = V_1\mu_1 + e_1 \tag{9}$$

where $V_{1((n-m)\times(p-s))} = W_{11} - W_{12}G_2^{-1}G_1$ is the submatrix with the restriction of non-interaction associated with the observed data.

The least squares estimator for μ_1 is given by

$$\hat{\mu}_1 = (V_1'V_1)^{-1}V_1'y_1 \tag{10}$$

Similarly

$$y_2 = W_{21}\mu_1 - W_{22}G_2^{-1}G_1\mu_1 + e_2 = V_2\mu_1 + e_2 \tag{11}$$

where $V_{2(m\times(p-s))} = W_{21} - W_{22}G_2^{-1}G_1$ is the submatrix with the restriction of non-interaction associated with the missing data.

From the above we can find an alternative expression for the estimation of missing information given by

$$\hat{y}_2 = V_2\hat{\mu}_1 = V_2(V_1'V_1)^{-1}V_1'y_1 \tag{12}$$

whose expected value and variance are respectively

$$E(\hat{y}_2) = V_2\mu_1 \quad \text{and} \quad \text{Var}(\hat{y}_2) = V_2(V_1'V_1)^{-1}V_2'\sigma^2$$

In order to achieve the imputation of missing information, the estimations are obtained making use of expression (12) and the empty cells are replaced by the respective estimated values.

Then, in this case the model is given by

$$\begin{aligned} y_0 &= \begin{pmatrix} y_1 \\ \hat{y}_2 \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \mu_1 + \begin{pmatrix} e_{01} \\ e_{02} \end{pmatrix} \\ y_0 &= V\mu_1 + e_0 \end{aligned} \tag{13}$$

2.1.1 Decomposition of sums of squares after imputing estimated data

In this subsection, we present the decomposition of sums of squares after substituting the empty cells by the respective estimated values. The sum of squares is corrected by mean.

The total sum of squares with imputed data, corrected by mean (TSS_{WIDC}) is

$$\begin{aligned} TSS_{\text{WIDC}} &= \begin{pmatrix} y_1' & \hat{y}_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \hat{y}_2 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} y_1' & \hat{y}_2' \end{pmatrix} \begin{pmatrix} 1_1 \\ 1_2 \end{pmatrix} \begin{pmatrix} 1_1' & 1_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \hat{y}_2 \end{pmatrix} \\ &= y_1'y_1 + \hat{y}_2'\hat{y}_2 - \frac{1}{n}(y_1'1_1 1_1'y_1 + 2\hat{y}_2'1_2 1_1'y_1 + \hat{y}_2'1_2 1_2'\hat{y}_2) \end{aligned} \tag{14}$$

When data is not imputed, the total sum of squares corrected by mean (TSS_{NIDC}) is

$$TSS_{\text{NIDC}} = y_1'y_1 - \frac{1}{n_1}(y_1'1_1 1_1'y_1) \tag{15}$$

We observe clearly that

$$TSS_{\text{WIDC}} \geq TSS_{\text{NIDC}}$$

The model sum of squares with imputed data, corrected by mean (MSS_{WIDC}) is

$$MSS_{\text{WIDC}} = y_1'V_1(V_1'V_1)^{-1}V_1'y_1 + \hat{y}_2'\hat{y}_2 - \frac{1}{n}(y_1'1_1 1_1'y_1 + 2\hat{y}_2'1_2 1_1'y_1 + \hat{y}_2'1_2 1_2'\hat{y}_2) \tag{16}$$

When data is not imputed, the model sum of squares corrected by mean (MSS_{NIDC}) is

$$MSS_{\text{NIDC}} = y_1' V_1 (V_1' V_1)^{-1} V_1' y_1 - \frac{1}{n_1}(y_1' 1_1 1_1' y_1) \tag{17}$$

Similarly

$$MSS_{\text{WIDC}} \geq MSS_{\text{NIDC}}$$

Finally, after imputing estimated data, the residual sum of squares (RSS_{WID}) is

$$\begin{aligned} RSS_{\text{WID}} &= \begin{pmatrix} y_1' & \hat{y}_2' \end{pmatrix} \begin{pmatrix} y_1 \\ \hat{y}_2 \end{pmatrix} - y_1' V_1 (V_1' V_1)^{-1} V_1' y_1 - \hat{y}_2' \hat{y}_2 \\ &= y_1' (I - V_1 (V_1' V_1)^{-1} V_1') y_1 \end{aligned} \tag{18}$$

When data is not imputed, the residual sum of squares (RSS_{NID}) is

$$RSS_{\text{NID}} = y_1' (I - V_1 (V_1' V_1)^{-1} V_1') y_1 \tag{19}$$

In this case

$$RSS_{\text{WID}} = RSS_{\text{NID}}$$

From the above, the analysis of variance after imputing estimated data corrected by mean, is presented in Table 1.

Table 1 - *ANOVA obtained by means of the constrained least squares method*

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Model	$p - s - 1$	MSS_{WIDC}	MSM_{WIDC}	$\frac{MSM_{\text{WIDC}}}{RMS_{\text{WID}}}$
Residual	$n - p + s - m$	RSS_{WID}	RMS_{WID}	
Total	$n - m - 1$	TSS_{WIDC}		

3 Proposed methodology for handling missing information

In this section, we studied the analysis of variance construction after imputing estimated data through the non-iterative method given by Melo (2002). Next we found the structure of the variance and covariance matrix between observed and estimated information, which shows the problems relatives to the traditional

analysis of variance. Then, we present a non-iterative methodology to estimate missing information in connected designs with the cell means model which reduces the correlation between observed and estimated information. After imputing the missing information, we suggest a way to analyze the variance by accounting for the subvector's covariance structure of the information vector. This analysis guarantees that the test statistic follows a central F distribution under H_0.

In this case, the model of interest in terms of the observed cell means vector μ_1 for the analysis of variance as regards the techniques presented in Melo (2002) is given by (13). For that model, where $V_{(n\times(p-s))}$ is a full column rank matrix such that $rank(V) = p - s$, the unbiased minimum variance estimators of μ_1 and σ^2, obtained through the likelihood function, are given by (5) and

$$\hat{\sigma}^2 = S_0^2 = \frac{RSS_{\text{NID}}}{n-p+s} \tag{20}$$

respectively. These estimators are independently distributed with distributions

$$\hat{\mu}_1 \sim N(\mu_1, \sigma^2(V'V)^{-1}) \quad and \quad S_0^2 \sim \frac{\sigma^2}{n-p+s}\chi^2_{(n-p+s)}$$

The likelihood function to be maximizied is

$$L(\mu_1, \sigma^2) = f(y, \mu_1, \sigma^2) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\|y - V\mu_1\|^2\right]. \tag{21}$$

The general hypothesis relative to the parameters of the linear model of interest, is given by

$$H_0 : V\mu_1 = 0 \quad vs \quad H_a : V\mu_1 \neq 0, \;\; \sigma^2 > 0 \tag{22}$$

Under H_0 we have a family of distributions for each value of σ^2. In order to carry out the test we start out with the likelihood function expressed in (21). The test of hypotheses is carried out by means of the generalized likelihood ratio principle, given by

$$\lambda = \frac{\max\limits_{H_0} L(y, \mu_1, \sigma^2)}{\max\limits_{\Omega} L(y, \mu_1, \sigma^2)} \tag{23}$$

In such a case, we make use of maximum likelihood estimators of σ^2 in (20). Next we construct the relevant statistic which is commonly used in order to test H_0 given by

$$F = \frac{(n-p+s-m)(y'(V(V'V)^{-1}V')y)}{(p-s)(y'[I - V(V'V)^{-1}V']y)} \tag{24}$$

It is assumed that such statistic is distributed as $F'_{(a_1,a_2,\phi_0)}$, with $a_1 = p - s$, $a_2 = n - p + s - m$ and $\phi_0 = \frac{\mu_1' V' V \mu_1}{2\sigma^2}$. This statistic is central if and only if $H_0 : V\mu_1 = 0$ is true.

However, it is clear that the components of vector y are not independent, since the estimated information vector $\hat{y}_2$ is a linear combination of the observed information vector y_1, as in expression (12).

Furthermore, if we carry out the analysis of sum of squares' distribution expressed only in terms of the vector of observed values y_1, we obtain

$$\begin{aligned} MSS_{\text{WID}} &= y'V(V'V)^{-1}V'y = y_1'My_1 \\ RSS &= y'\left[I - V(V'V)^{-1}V'\right]y = y_1'Qy_1 \end{aligned}$$

where $M = V_1(V_1'V_1)^{-1}V_1' + V_1(V_1'V_1)^{-1}(V_2'V_2)(V_1'V_1)^{-1}V_1'$ and $Q = I_1 - V_1(V_1'V_1)^{-1}V_1'$.

However, the matrices associated with the quadratic forms are not all idempotents:

$$\begin{aligned} MM =& V_1(V_1'V_1)^{-1}V_1' + 2V_1(V_1'V_1)^{-1}(V_2'V_2)(V_1'V_1)^{-1}V_1' \\ &+ V_1(V_1'V_1)^{-1}(V_2'V_2)(V_1'V_1)^{-1}(V_2'V_2)(V_1'V_1)^{-1}V_1' \\ \neq& M \end{aligned}$$

and $QQ = Q$. Thus, the model sum of squares after imputing the estimated data (MSS_{WID}) is

$$y_1'My_1 \not\sim \chi^{2'}_{(p-s,\phi_1)}$$

and the residual sum of squares after imputing the estimated data is

$$y_1'Qy_1 \sim \chi^2_{(n-p+s-m)}$$

In such a case, the ratio found by using the likelihood function is not an F central under H_0. However, it is commonly used in order to test H_0.

On the other hand, the structure of the variance and covariance matrix between observed and estimated information, is given by

$$D' = \sigma^2 D'' = \sigma^2 \begin{pmatrix} [I]_{((n-m)\times(n-m))} & \vdots & \left[V_1(V_1'V_1)^{-1}V_2'\right]_{((n-m)\times m)} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \left[V_2(V_1'V_1)^{-1}V_1'\right]_{(m\times(n-m))} & \vdots & \left[V_2(V_1'V_1)^{-1}V_2'\right]_{(m\times m)} \end{pmatrix}$$

The previous matrix is symmetric and singular. This structure of the covariance matrix allows us to show another problem overlooked by the traditional analysis of variance, which is to assume that $e \sim N(0_{(n\times 1)}, \sigma^2 I_n)$. This is because the error is not truly homoscedastics, since $e \sim N(0_{(n\times 1)}, D')$.

3.1 Estimation and imputation of missing information

In the constrained least squares non-iterative method of estimation, the expression for the estimation of missing information is given by (12). In such expression we can see that the estimated information vector $\hat{y}_2$ is obtained as a linear combination of the observed information vector y_1 and then, $\mathrm{Cov}(y_1, \hat{y}_2) = \sigma^2 V_2 (V_1' V_1)^{-1} V_1'$.

The fact that the structure of covariability between the two vectors is different from zero leads us to think of the necessity of a methodology with which the correlations between the y's subvectors decreases, reducing the dependence between the estimated and observed information vectors.

In order to achieve such goal, the proposed non-iterative methodology consists in the adittion of a random component $e^*_{(m\times 1)}$ to the estimated observations vector $\hat{y}_{2(m\times 1)}$, such that

$$e^* \sim N(0_{(m\times 1)}, \sigma^2 I_m) \tag{25}$$

where σ^2 is the variance of errors obtained by fitting a model like (9), assuming

$$e_1 \sim N(0_{((n-m)\times(n-m))}, \sigma^2 I_{(n-m)})$$

where σ^2 is estimated with the residual mean square without imputed data (RMS_{NID}). Here it is important to note that the use of the residual mean square instead of a known variance, would be able to have an impact on the properties of the estimators obtained.

The randomly generated errors $e^*_{(m\times 1)}$ must satisfy the foregoing condition and satisfy the presupposition of uncorrelation with errors $e_{1((n-m)\times 1)}$, such that $\mathrm{Cov}(e_1, e^*) = 0_{((n-m)\times m)}$. Therefore, we propose the following expression in order to estimate the missing information

$$\hat{y}_2^* = V_2 \hat{\mu}_1 + e^* = V_2 (V_1' V_1)^{-1} V_1' y_1 + e^* \tag{26}$$

The addition of such random component, leads to a decrease of correlation between the y's subvectors.

The structure of the variance and covariance matrix between observed and estimated information using the proposed non-iterative technique of estimation, is given by

$$D^* = \sigma^2 D = \sigma^2 \begin{pmatrix} [I]_{((n-m)\times(n-m))} & \vdots & \left[V_1 (V_1' V_1)^{-1} V_2'\right]_{((n-m)\times m)} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \left[V_2 (V_1' V_1)^{-1} V_1'\right]_{(m\times(n-m))} & \vdots & \left[I + V_2 (V_1' V_1)^{-1} V_2'\right]_{(m\times m)} \end{pmatrix}$$

The previous matrix is symmetric and its rows and columns are linearly independent, and thus $D^* = \sigma^2 D$ is non-singular.

In order to achieve the imputation of missing information, the estimations are obtained by making use of expression (26) and the empty cells are replaced by the respective estimated values and conserving the initial places in the design.

In this case, the model is given by

$$\begin{aligned} y^* &= \begin{pmatrix} y_1 \\ \hat{y}_2^* \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix} \mu_1 + \begin{pmatrix} e_1^* \\ e_2^* \end{pmatrix} \\ y^* &= V\mu_1 + \varepsilon \end{aligned} \tag{27}$$

such that $\varepsilon \sim N(0_{(n\times n)}, D^*)$.

Due to the characteristics of the errors in this model it is appropriate to use the weighted least squares technique presented by Montgomery and Peck (1992).

In this case, since D is a non-singular symmetric matrix, it is possible to claim that there is another non-singular symmetric matrix K, such that $K'K = KK = D$. Such matrix K is known in the literature as the square root of D.

If in model (27), we premultiply by K^{-1}, then

$$\begin{aligned} K^{-1}y^* &= K^{-1}V\mu_1 + K^{-1}\varepsilon \\ z &= B\mu_1 + \varepsilon^* \end{aligned} \tag{28}$$

where $z = K^{-1}y^*$, $B = K^{-1}V$ and $\varepsilon^* = K^{-1}\varepsilon$.

For this new model, we have that $\varepsilon^* \sim N(0_{(n\times n)}, \sigma^2 I_n)$. Since all of the presuppositions of least squares are satisfied by model (28), the generalized unbiased and minimum variance estimator is given by

$$\hat{\mu}_1^* = (V'D^{-1}V)^{-1}(V'D^{-1}y^*) \tag{29}$$

such that

$$\hat{\mu}_1^* \sim N(\mu_1, \sigma^2(V'D^{-1}V)^{-1}) \tag{30}$$

3.2 Analysis after the imputation of missing information by means of the proposed non-iterative methodology

In this section, we present the relevant results relatives to the parameter estimation and test of hypotheses obtained through the likelihood function, after the imputation of missing information by means of the proposed non-iterative methodology.

3.2.1 Parameter estimation and test of hypotheses with the maximum likelihood function

The unbiased and minimum variance estimators are given by (29) and

$$\hat{\sigma}^2 = S^2 = \frac{RSS^*{}_{\text{WID}}}{n-p+s} \tag{31}$$

with $RSS^*{}_{\text{WID}}$ being the residual sum of squares after imputing estimated data, given by

$$RSS^*{}_{\text{WID}} = y^{*'}\left[D^{-1} - D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}\right]y^* = y^{*'}Q_1y^* \tag{32}$$

where $Q_1 = D^{-1} - D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}$.

Furthermore we consider the general test of hypotheses expressed in (22). In order to carry out the test we use the generalized likelihood ratio presented previously.

In this case we have

$$\lambda^{2/n} = \frac{1}{\left(1 + \frac{y^{*'}D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}y^*}{RSS^*{}_{\text{WID}}}\right)} \tag{33}$$

where $\lambda^{2/n}$ is small if $\frac{y^{*'}D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}y^*}{RSS^*{}_{\text{WID}}}$ is big and is a monotone function that allows us to use $\frac{y^{*'}D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}y^*}{RSS^*{}_{\text{WID}}}$ in order to carry out the general test of hypotheses.

To take into account from the above, below we carry out the analysis of sum squares' distribution.

3.2.2 Distributions of obtained quadratic forms

Following theorems of distribution of quadratic forms presented in Searle (1971) and Hocking (1996), we verified the idempotence of the matrices associated with the sum of squares found in the analysis of the maximum likelihood function, which are given by

$$MSS^*{}_{\text{WID}} = y^{*'}\left[D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}\right]y^* = y^{*'}Q_2y^* \tag{34}$$

$$TSS^*{}_{\text{WID}} = y^{*'}D^{-1}y^* \tag{35}$$

where $Q_2 = D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}$.

We have that the matrices associated with the quadratic forms of $MSS^*{}_{\text{WID}}$ and $RSS^*{}_{\text{WID}}$ are idempotents (Franco 2003).

Therefore

$$\frac{MSS^*{}_{\text{WID}}}{\sigma^2} = \frac{y^{*'}[D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}]y^*}{\sigma^2} \sim \chi^{2'}_{(p-s,\phi)} \tag{36}$$

with $\phi = \frac{\mu_1'V'D^{-1}V\mu_1}{2\sigma^2}$, and besides

$$\frac{RSS^*{}_{\text{WID}}}{\sigma^2} = \frac{y^{*'}[D^{-1} - D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}]y^*}{\sigma^2} \sim \chi^2_{(n-p+s-m)} \tag{37}$$

The independence of the matrices associated with the sums of squares found in the analysis of the maximum likelihood function is verified, then $MSS^*{}_{\text{WID}}$ and $RSS^*{}_{\text{WID}}$ are independent.

Such as is presented in Franco (2003), since by means of the proposed non-iterative methodology of estimation the subvectors of the observations vector y are not dependent and each one of the quadratic forms has distribution χ^2 and these are independent, we have

$$\frac{(n-p+s-m)(y^{*'}[D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}]y^*)}{(p-s)(y^{*'}[D^{-1} - D^{-1}V(V'D^{-1}V)^{-1}V'D^{-1}]y^*)} \sim F'_{(p-s,n-p+s-m,\phi)} \tag{38}$$

The previous ratio has distribution central F under $H_0 : V\mu_1 = 0$.

3.2.3 Analysis of variance for the model after imputing estimated data

For the proposed non-iterative methodology of estimation, the correction factor by mean for the model and total sums of squares ($CFSS^*$) is given by

$$CFSS^* = y^{*'}\left[D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}\right]y^* = y^{*'}Q_3y^* \tag{39}$$

where $1_{(n\times 1)}$ is a vector of ones of size n and $Q_3 = D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}$.

The matrix associated with the quadratic form of the correction factor by mean $CFSS^*$ is idempotent and therefore

$$\frac{CFSS^*}{\sigma^2} = \frac{y^{*'}[D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}]y^*}{\sigma^2} \sim \chi^{2'}_{(1,\phi_1)} \tag{40}$$

with $\phi_1 = \frac{\mu_1'V'[D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}]V\mu_1}{2\sigma^2}$.

From the above, the model sum of squares after imputing data corrected by mean ($MSS^*{}_{\text{WIDC}}$) is given by

$$\begin{aligned} MSS^*{}_{\text{WIDC}} &= y^{*'}D^{-1}\left[V(V'D^{-1}V)^{-1}V'D^{-1} - 1(1'D^{-1}1)^{-1}1'D^{-1}\right]y^* \\ &= y^{*'}D^{-1}[H_V - M]y^* \end{aligned} \tag{41}$$

and the total sum of squares after imputing data corrected by mean ($TSS^*{}_{\text{WIDC}}$) is

$$\begin{aligned} TSS^*{}_{\text{WIDC}} &= y^{*'}\left[D^{-1} - D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}\right]y^* \\ &= y^{*'}D^{-1}[I - M]y^* \end{aligned} \tag{42}$$

where $H_V = V(V'D^{-1}V)^{-1}V'D^{-1}$ and $M = 1(1'D^{-1}1)^{-1}1'D^{-1}$ are idempotent matrices.

In such case, we verified the idempotence of quadratic forms and then

$$\frac{b_2(y^{*'}D^{-1}[V(V'D^{-1}V)^{-1}V' - 1(1'D^{-1}1)^{-1}1']D^{-1}y^*)}{b_1(y^{*'}D^{-1}[I - V(V'D^{-1}V)^{-1}V'D^{-1}]y^*)} \sim F'_{(b_1,b_2,\phi_1^*)} \tag{43}$$

with $\phi_1^* = \frac{\mu_1'V'[D^{-1}]V\mu_1 - \mu_1'V'[D^{-1}1(1'D^{-1}1)^{-1}1'D^{-1}]V\mu_1}{2\sigma^2}$, $b_1 = p - s - 1$ and $b_2 = n - p + s - m$.

Under the null hyphoteses $H_0 : V\mu_1 = 0$, $\phi_1^* = 0$, and then ratio (43) is distributed as central F. Therefore, the analysis of variance obtained by means of the proposed non-iterative methodology, corrected by mean, is given in Table 2.

Table 2 - *ANOVA obtained by means of the proposed methodology*

Source of variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Model	$p - s - 1$	MSS^*_{WIDC}	MSM^*_{WIDC}	$\frac{MSM^*_{\text{WIDC}}}{RMS^*_{\text{WID}}}$
Residual	$n - p + s - m$	RSS^*_{WID}	RMS^*_{WID}	
Total	$n - m - 1$	TSS^*_{WIDC}		

4 Application

The data corresponds to a $2 \times 3 \times 4$ factorial design presented in Myers and Montgomery (1995). The experiment considers the effects of corrosion rate (inch/min)(A), cut depth (inch)(B) and kind of material (C) on the life of a set of cutting tools of a metal particle. Two replicates of a $2 \times 3 \times 4$ factorial design were run. The results are presented in table 3.

The model associated with the $2 \times 3 \times 4$ factorial design without interaction is given by

$$y_{ijkl} = \mu_{ijk} + e_{ijkl} \tag{44}$$

$i = 1, 2$, $j = 1, 2, 3$, $k = 1, 2, 3, 4$ and $l = 1, 2, \ldots, n_{ijk}$; and where y_{ijkl} is the l-th observation relative to the i-th material, j-th corrosion rate and k-th cut depth, μ_{ijk} is the ijk-th cell mean and e_{ijkl} is the random error component such that $e_{ijkl} \sim N(0, \sigma^2)$.

The model (44) is subject to the non-interaction among AB, AC, BC and ABC. Of these 50 constraints, there are only

Table 3 - $2 \times 3 \times 4$ *Factorial design with all observed data*

Corrosion Rate	Material I				Material II			
	Cut depth				Cut depth			
	0.15	0.20	0.30	0.40	0.15	0.20	0.30	0.40
0.20	74	79	89	102	63	73	77	101
	78	82	94	98	68	74	79	103
0.25	98	97	98	105	74	85	83	105
	91	93	105	102	77	81	87	104
0.30	114	115	122	133	100	105	111	118
	108	111	117	138	97	108	107	122

$$\begin{aligned} s =& (a-1)(b-1) + (a-1)(c-1) + (b-1)(c-1) + (a-1)(b-1)(c-1) \\ =& 2 + 3 + 6 + 6 = 17 \end{aligned}$$

linearly independent ones. One such set of linearly independent constraints is the set with $i' = a = 2, j' = b = 3$ and $k' = c = 4$ given by

$$\begin{gathered} \mu_{1j.} - \mu_{13.} - \mu_{2j.} + \mu_{23.} = 0 \\ \mu_{1.k} - \mu_{1.4} - \mu_{2.k} + \mu_{2.4} = 0 \\ \mu_{.jk} - \mu_{.j4} - \mu_{.3k} + \mu_{.34} = 0 \\ \{\mu_{1jk} - \mu_{13k} - \mu_{1j4} + \mu_{134}\} - \{\mu_{2jk} - \mu_{23k} - \mu_{2j4} + \mu_{234}\} = 0 \end{gathered}$$

with $j = 1, 2$ and $k = 1, 2, 3$.

In terms of the modified cell means model, (44) is given by (1) subject to (2), where the contrast matrix G is given by

$$G = \begin{pmatrix} D_2 \otimes D_3 \otimes J_4' \\ D_2 \otimes J_3' \otimes D_4 \\ J_2' \otimes D_3 \otimes D_4 \\ D_2 \otimes D_3 \otimes D_4 \end{pmatrix} \tag{45}$$

with $D_2 = \begin{pmatrix} 1 & -1 \end{pmatrix}$, $D_3 = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}$, $D_4 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}$, $J_4' = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}$, $J_3' = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$ and $J_2' = \begin{pmatrix} 1 & 1 \end{pmatrix}$.

The columns of G_2 are relative to the cells means

$$\begin{gathered} \mu_{111}, \mu_{112}, \mu_{113}, \mu_{114}, \mu_{121}, \mu_{122}, \mu_{123}, \mu_{124}, \\ \mu_{131}, \mu_{132}, \mu_{133}, \mu_{211}, \mu_{212}, \mu_{213}, \mu_{221}, \mu_{222}, \mu_{223} \end{gathered}$$

and the columns of G_1 are relative to the cells means

$$\mu_{134}, \mu_{214}, \mu_{224}, \mu_{231}, \mu_{232}, \mu_{233}, \mu_{234}$$

Below we consider two situations of randomized loss of information on this case study and apply the proposed non-iterative methodology.

In the next section, the estimations are obtained through our proposed non-iterative methodology in cases where there are $m = 6$ and $m = 30$ missing data in the design.

4.1 Missing information estimation

In Tables 4 and 6, the data of factorial design with all of the randomized loss of information of interest are showed. In Tables 5 and 7, the estimated missing information for cases where there are $m = 6$ and $m = 30$ missing data, respectively, is presented . In these tables, y_2 corresponds to the observed information vector, $\hat{y}_2$ is the estimated information vector using the non-iterative technique presented in Melo (2002), e^* is the random error vector and $\hat{y}_2^*$ is the estimated information vector using the non-iterative technique proposed in this work.

Table 4 - $2 \times 3 \times 4$ *Factorial design with* $m = 6$ *missing data*

Corrosion Rate	Material I				Material II			
	Cut depth				Cut depth			
	0.15	0.20	0.30	0.40	0.15	0.20	0.30	0.40
0.20	74	79	89	102	63	73	77	101
	78	82	x	98	68	x	79	103
0.25	98	97	98	105	74	85	83	105
	91	x	105	102	x	81	87	104
0.30	114	115	122	133	100	105	111	118
	108	111	x	138	97	108	x	122

Table 5 - *Estimated information of* $2 \times 3 \times 4$ *factorial design with* $m = 6$ *missing data*

y_2	$\hat{y}_2$	e^*	$\hat{y}_2^*$	y_2	$\hat{y}_2$	e^*	$\hat{y}_2^*$
94	88.58	1.37	89.95	74	73.11	-2.11	71.00
93	92.79	-0.60	92.19	77	78.10	-0.81	72.29
117	120.13	-1.67	118.46	107	110.34	2.80	113.14

In case $m = 6$ missing data, $\hat{\mu}_1' = (133.54, 92.21, 102.09, 99.76, 104.66, 110.34, 123.75)$ and $e^* \sim N(0_{(6\times1)}, 22.16)$.

Table 6 - $2 \times 3 \times 4$ *Factorial design with* $m = 30$ *missing data*

Corrosion Rate	Material I				Material II			
	Cut depth				Cut depth			
	0.15	0.20	0.30	0.40	0.15	0.20	0.30	0.40
0.20	x	79	89	102	x	73	77	101
	x	x	x	x	x	x	x	x
0.25	98	x	98	105	74	x	83	105
	x	x	x	x	x	x	x	x
0.30	x	115	122	133	x	105	111	118
	x	x	x	x	x	x	x	x

Table 7 - *Estimated information of* $2 \times 3 \times 4$ *factorial design with* $m = 30$ *missing data*

y_2	$\hat{y}_2$	e^*	$\hat{y}_2^*$	y_2	$\hat{y}_2$	e^*	$\hat{y}_2^*$
74	84.84	0.53	85.37	63	74.40	2.88	77.28
78	84.84	-2.27	82.57	68	74.40	0.63	75.03
82	82.97	-2.23	80.74	74	72.52	0.87	73.39
94	89.59	1.07	90.66	79	79.15	-2.09	77.06
98	103.59	-1.86	101.73	103	93.15	-0.29	92.86
91	91.22	-1.45	89.77	77	80.77	-0.20	80.57
97	89.34	-1.36	87.98	85	78.90	-0.31	78.59
93	89.34	-2.51	86.83	81	78.90	1.18	80.08
105	95.97	2.14	98.11	87	85.52	0.88	86.40
102	109.97	1.48	111.45	104	99.52	1.67	101.19
114	115.34	-2.16	113.18	100	104.90	-1.32	103.58
108	115.34	2.25	117.59	97	104.90	1.40	106.30
111	113.47	-2.03	111.44	108	103.02	-0.78	102.24
117	120.09	-2.56	117.53	107	109.65	2.36	112.01
138	134.09	-2.83	131.26	122	123.65	1.45	125.10

In case where there are $m = 30$ missing data, $\hat{\mu}_1' = (134.09, 93.15, 99.52, 104.90, 103.02, 109.65, 123.65)$ and $e^* \sim N(0_{(30\times1)}, 26.08)$.

From Tables 5 and 7, we observe that the estimations obtained by means of the non-iterative methodologies of estimation considered in this work are near the observed data, though a few data are underestimated and others are overestimated.

In the next section, we present the analysis of variance obtained through our proposed methodology in cases where there are $m = 6$ and $m = 30$ missing data in the design.

4.2 Analysis of variance

In Tables 9 and 10, the analysis of variance after imputing the missing information is showed. In Table 8, we explain the meaning of the columns in these tables, in which the analysis of variance are shown.

Table 8 - *Meaning of columns of tables in which the analysis of variance are shown*

Convention	Analysis of variance
$WAD(cm)$	*With all observed data, corrected by mean.*
$WID(cm)$	*With imputed data using the technique presented in Melo 2002, corrected by mean.*
$WID^*(cm)$	*With imputed data using the new technique, corrected by mean.*
$NID(cm)$	*Without imputed data, corrected by mean.*

In a case with $m = 6$ missing data, the degrees of freedom are given by $p - s = 24 - 17 = 7$, $n - p + s - m = 48 - 24 + 17 - 6 = 35$, $n - m = 48 - 6 = 42$.

Table 9 - *Analysis of variance of* $2 \times 3 \times 4$ *factorial design with* $(m = 6)$ *estimated data*

Methodology	MSS	RSS	TSS	MSM	RMS	F
$WAD(cm)$	13024.63	821.85	13846.48	2170.77	20.05	108.29
$WID(cm)$	13346.48	775.66	14122.15	2224.41	22.16	100.37
$WID^*(cm)$	11617.40	793.64	12411.04	1936.23	22.67	85.38
$NID(cm)$	11617.40	775.66	12393.07	1936.23	22.16	87.36

In a case with $m = 30$ missing data, the degrees of freedom are given by $p - s = 24 - 17 = 7$, $n - p + s - m = 48 - 24 + 17 - 30 = 11$, $n - m = 48 - 30 = 18$.

From Tables 9 and 10, we observe that in the analysis of variance obtained by means of the methodologies previously considered, the value of the F statistic tends to decrease as long as we increment the quantity of missing information in the design, leading to a progressively different value from that obtained when we have all the original observed data. Incrementing the rate of missing information from 12.5% up to 62.5%, leads to a bold reduction of the value of the F statistic. In all cases, this value is less than the value obtained when no such imputation of missing information is performed. Even so, the F test continues powerful enough and the decision regarding $H_0 : V\mu_1 = 0$ is not affected.

Table 10 - *Analysis of variance of* $2 \times 3 \times 4$ *factorial design with* $(m = 30)$ *estimated data*

Methodology	MSS	RSS	TSS	MSM	RMS	F
$WAD(cm)$	13024.63	821.85	13846.48	2170.77	20.05	108.29
$WID(cm)$	12719.96	286.98	13006.95	2119.99	26.08	81.25
$WID^*(cm)$	4821.01	378.40	5199.42	803.50	34.40	23.35
$NID(cm)$	4821.01	286.98	5108	803.50	26.08	30.79

Conclusions

We are conscious that nowadays there are many programs and theories to work with non orthogonal and unbalanced designs. Many of these methodologies are based on iterative procedures that, in some situations, are complex and can provide estimators that are not unique, unbiased, efficient and the minimum variance. For that, in this work, we proposed a simple and non-iterative methodology for estimating missing information with the cell means model in connected designs, which provides us with unbiased estimators and allows to estimate individual data in the cells and the value of the empty cells as well.

In our proposed non-iterative methodology, we added a random error to the estimated values vector, with the aim of diminishing the linear correlation between the subvectors of the information vector. Furthermore, such random error allows to guarantee the non-singularity of the variance and covariance matrix of such subvectors.

On the other hand, although the proposed non-iterative methodology of estimation is one of the main contributions carried out, the results obtained in the analysis of variance after imputing the missing information in the design are more important, because the foregoing considerations allow to find, in a very detailed fashion, the distribution of the test statistic, and such distribution turns out to be an F.

As we can see, the main purpose of our work, far from limited in order to propose a new non-iterative methodology for estimating missing information consisted of a revision and construction of the exact test to carry out the ANOVA, considering that the traditional test of analysis of variance when there are missing cells and these are substituted in the design. It is not exact and appropriate to verify the hypothesis of interest.

Finally, the analysis of variance allows to bring out the influence that the rate of missing information has over the value of the statistic. This is a compelling reason to be careful when dealing with situations of missing information, in which it is not appropriate to impute a large amount of information in the design.

Acknowledgments

The authors gratefully acknowledge the comments and suggestions of the associate editor and two anonymous referees that helped improve the article immensely.

SOTO, D. C. F.; MARTÍNEZ, O. O. M. Metodologia para estimar dados perdidos e correção na análise da variância em delineamentos conectados de médias de caselas. *Rev. Mat. Estat.,* São Paulo, v.25, n.1, p.45-64, 2007.

- RESUMO: Neste trabalho apresentamos uma metodologia não-iterativa para estimar dados perdidos em delineamentos conectados de médias de caselas. Esta metodologia melhora aquela de Melo (2002), reduzindo a correlação entre as informações observadas e estimadas. Depois de estimar as informações perdidas, nós sugerimos um caminho para a análise de variância considerando a estrutura de covariância dos sub-vetores do vetor de informação. Esta análise garante que a estatística do teste segue uma distribuição F central sob H_0. Finalmente, nós trabalhamos em um estudo de caso de um delineamento fatorial, em que são aplicados as técnicas sugeridas.

- PALAVRAS-CHAVE: Delineamento de média de célula; delineamentos conectados; informações perdidas; estimação e substituição; distribuição de formas quadráticas.

References

ALLAN, F.; WISHART, J. A Method of estimating the yield of a missing plot in field experiments. *J. Agric. Sci.*, Cambridge, v.20, p.399-406, 1930.

BARTLETT, M. S. Some examples of statistical methods of research in agriculture and applied biology. *J. R. Stat. Soc.*, London, v.4, n.2, p.137-183, 1937.

DODGE, Y. *Analysis of experiments with missing data.* New York: John Wiley & Sons, 1985. 498p.

FRANCO, D. C. Metodología para la estimación de información faltante, imputación y corrección en el análisis de varianza con el modelo de medias de celda en diseños conectados. 2003. 84f. Trabajo (Grado) – Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, 2003.

HOCKING, R. R. *The analysis of linear models.* Monterrey: Brooks/Cole, 1985. 385p.

HOCKING, R. R. *Methods and applications of linear models.* New York: John Wiley & Sons, 1996. 731p.

JARRETT, R. The analysis of designed experiments with missing observations. *J. R. Stat. Soc. Ser. C: Appl. Stat.*, London, v.27, n.1, p.38-46, 1978.

JOHN, J. ; PRESCOTT, P. Estimating missing values in experiments. *J. R. Stat. Soc. Ser. C: Appl. Stat.*, London, v.24, n.2, p.190-192, 1975.

LITTLE, R. ; RUBIN, D. *Statistical analysis with missing data*. New York: John Wiley & Sons, 2002. 408p.

LITTLE, R. Robust estimation of the mean and covariance matrix from data with missing values. *J. R. Stat. Soc. Ser. C: Appl. Stat.*, London, v.37, n.1, p.23-38, 1988.

LIU, C. Bayesian robust multivariate linear regression with incomplete data. *J. Am. Stat. Assoc.*, New York, v.91, n.435, p.1219-1227, 1996.

MELO, S. E. *Metodología para la estimación de celdas vacías con el modelo de medias de celdas en diseños conectados*. 2002. 91f. Tesis (Maestría) – Universidad Nacional de Colombia, Bogotá, 2002.

MONTGOMERY, D. ; PECK, E. *Introduction to linear legression analysis*. New York: John Wiley & Sons, 1992. 527p.

MURRAY, L. W.; SMITH, D. W. Estimability, testability and connectedness in the cell means model. *Commun. Stat.*, New York, v.14, n.8, p.1889-1917, 1985.

MYERS R. ; MONTGOMERY, D. *Response surface methodology, process and product optimization using designed experiments*. New York: John Wiley & Sons, 1992. 824p.

SEARLE, S. R. *Linear models*. New York: John Wiley & Sons, 1971. 532p.

SEARLE, S. R. *Linear models for unbalanced data*. New York: John Wiley & Sons, 1987. 533p.

WEEKS, D. L.; WILLIAMS, D. R. A note on the determination of connectedness in an n-way cross classification. *Technometrics*, Washington, v.6, n.3, p.319-324, 1964.

YATES, F. The Analysis of replicate experiments when the field results are incomplete. *Emp. J. Exper. Agric.*, Oxford, v.1, p.129-142, 1933.

Received in 12.02.2006.

Approved after revised in 25.05.2007.