MISSING CELLS DATA ANALYSIS UNDER A CONSTRAINED "U" MODEL

Dorinda Love Casanova, Department of Experimental Statistics, LSU

Abstract

Analysis of data involving missing cells is of major concern to many statisticians. This paper proposes an alternate method of dealing with this problem, whereby the "u" model is applied under a constraint condition, to data with missing cells. Instead of the traditional generalized inverse, a partitioned inverse is utilized. The end product is a cell means solution vector which provides an estimate for all cells including the missing ones when a constraint condition is imposed upon the model.

Introduction

Whether the incomplete data set is due to an unsampled population or unobserved treatment combination, the question arises of what should be done in the event of missing cells? Some examples of missing cells resulting from unobserved treatment combinations due to problems encountered with the sampling unit including: (1) in agriculture, a particular crop or field in a study that was damaged or destroyed by a flooding rain,

(2) in dairy science, an animal becomes very ill or possibly dies before the completion of the experiment,

(3) in social sciences, human subjects being absent when scheduled to participate in an experiment,

(4) in a survey type situation, the subjects do not follow the directions properly resulting in portions of the questionnaire filled out incorrectly,

(5) equipment malfunctions such as

(a) in fisheries, a net breaks on a cast at a particular location, or

(b) in physical education, the spring on the caliper may slip when conducting a skinfold test.

Missing cells may be caused by an unsampled population due to problems involving the experimental unit, that is, when it is not possible, humanly or otherwise, to apply the treatment combination to the experimental unit. This results in no measurement being taken. Finally, errors may be made in the application of the treatment, in the recording of the data, or losing the data.

Over the years, statisticians have been perplexed by the problem of missing cells, still with no generally applicable procedure available. However, there are several options a researcher has when analyzing a data set of this nature. First, it can be left in its original form and analyzed as is. With a completely randomized design, there are no problems in doing this because only the sample size is reduced. If it is a more advanced design, it may be possible to delete the incomplete observations and then proceed with the analysis. Finally, there are techniques for estimating the missing cells. This can be done only if the cause is not related to the nature of the treatment used in the experiment. The formula for estimation was developed by Yates in 1933. Keep in mind that this does not imply additional information to the experiment, it only facilitates the analysis of the remaining data. One degree of freedom is removed from error for each missing cell. If there are more than one missing cell, then an iterative process is required until all estimates are stabilized, which usually takes two or three cycles. In order to estimate these cells, the assumption of additivity must be made which places a restriction on the model.

A brief discussion of the difference between a constraint and a restriction is needed for clarification purposes. A constraint on the solution is a relationship imposed on the set of normal equations in order to obtain a solution. Constraint conditions need only apply to the elements of the solution vector. They do not apply to the parameters in the model and are imposed solely for deriving a solution. Any constraint that leads to a solution is sufficient. These constraint conditions do not affect the estimable functions or testable hypothesis. A restriction on the parameters of the model is a relationship among the parameters which is considered to be an integral part of the model that must be accounted for in the estimation and hypothesis testing. A restriction is not a hypothesis to be tested, it is a known fact and serves as an assumption. Certain kinds of restrictions can be applied as constraint conditions, but constraints do not imply restrictions. Restrictions may be based on prior knowledge of past research to restrict the parameters of the model.

The cell means or "u" model is an approach to the analysis of linear models in which it is assumed that each observation is drawn from a separate population having its own mean and variance. By the utilization of this model, the population means are the basis for estimation of the parameters and testing of all desired hypotheses. Some of the advantages of the "u" model include: (1) balanced cases need no separate theoretical consideration,

(2) no additional computation is needed when dealing with missing cells,

(3) only hypotheses specified directly by the researcher are tested, and

(4) the interpretation of the main effects and interactions are more meaningful since they are from population means.

Because it is expressed in the traditional matrix notation, the "u" model provides a theory of linear models that is not only mathematically brief, but also conceptually clear from the statistician's point of view(Speed,1969).

The objective of this paper is to offer an alternative method of analysis for data hindered by missing cells. The "u" model is utilized subject to a constraint condition. An estimate for each missing cell may be obtained from the iterative method of calculation developed by.
Yates. However, since the population means are the basis for the analysis, the solution vector resulting from the use of the cell means model approach provides the estimate which bypasses the Yates' technique. A partitioned inverse rather than a generalized inverse is employed because it requires less core storage rendering a more efficient procedure. Although the main thrust of this paper deals with missing cells, the procedure lends itself to both balanced and unbalanced data sets with restrictions and also balanced data sets without restrictions.

The following is an example illustrating a case in which this analytical method would be applicable:

Four primates were administered a special diet for a certain time period. Afterwards, a blood chemistry test to measure cortisone level was conducted at five unequally spaced time intervals. Measurements were not obtained for all of the animals under all five time intervals.

General questions about main effects and interactions are usually of interest. This example is also concerned with answering questions involving time trend analysis. When testing hypotheses with the "u" model, each hypothesis must be stated explicitly before it can be tested.

Preliminary Notes

Speed (1969) introduced the following linear model

\[ Y = W \mu + e \]

subject to \( Q\mu = 0 \)

where \( Y \) is a nx1 vector of observations,
\( \mu \) is a px1 vector of cell means,
\( W \) is a pxp design matrix,
\( e \) is a nx1 vector of random variables,
and \( G \) is a pxp matrix of rank \( r \) that imposes the constraint condition known about the cell means.

Since \( Y = N(W\mu, \sigma^2I) \), it can be assumed that \( p \) populations have been sampled, each having a mean and common variance. These populations may be sampled any number of times. Previously, only those populations with a minimum sample size of one were included in the analysis. Now, however, each treatment combination will be represented in the model as a cell mean with no minimum sample size requirement. Although only the observed values will be used in the actual analysis, the final product will result in an estimation for all of the parameters; missing and observed. The estimates for the missing cells are exactly the same numerical values which would result if the iterative process of Yates' formula were used:

The principle of least squares is applied to obtain a minimum of the function

\[ Q(\mu) = (Y - W\mu)'(Y - W\mu). \]

Lagrangian multipliers, which incorporate the constraint condition, are then used to develop the appropriate set of normal equations. At this point, it is possible to solve for the cell means solution vector, \( \mu \).

Before continuing, it is necessary to establish the following matrix notation which will be needed to understand the development of the solution vector:

1. The "u" model dealing only with the observed values is denoted:

\[ Y = W\mu + e \]

where \( n \) is the total number of observed cells,
\( Y_0 \) is a \( n \times 1 \) vector of observed cells,
\( W_0 \) is a \( n \times p \) design matrix only for observed sample sizes,
\( e \) is a \( n \times 1 \) vector of random variables associated with the observed values only, and

\[ Y = N(W\mu, \sigma^2I). \]

2. The "u" model representing the missing values (NOTE: This population has not been sampled or is unobservable.)

\[ Y = W\mu + e \]

where \( n \) is the total number of missing cells,
\( Y_m \) is a \( n \times 1 \) vector of missing observations,
\( W_m \) is a \( n \times p \) design matrix for each missing cell,
\( e \) is a \( n \times 1 \) vector of random variables associated with the missing values only.

3. The constraint matrix, \( G \), partitioned according to the observed cells and the missing cells represented by

\[ G = \begin{bmatrix} G_0 & G_m \end{bmatrix} \]

where \( C_0 \) is a \( r \times p \) matrix defining the portion of the constraint matrix associated with the observed cells,
\( C_m \) is a \( r \times q \) matrix defining the portion of the constraint matrix associated with the missing cells.

4. The following matrices represent a shorthand notation for the elements composing the partitioned inverse matrix:

\[ W = W_0 \cdot W_m \]

\[ W_0 = W_0 \cdot W_m \]

\[ W_m = W_0 \cdot W_0 \cdot W_0 \]

5. The partitioned inverse matrix, \( WPWI \), is represented by

\[ WPWI = \begin{bmatrix} \begin{bmatrix} E_0^{-1} & 0 \\ 0 & G_0^{-1} \end{bmatrix} \begin{bmatrix} N_0^{-1} & 0 \\ 0 & G_m^{-1} \end{bmatrix} \\ \begin{bmatrix} N_0^{-1} & 0 \\ 0 & G_m^{-1} \end{bmatrix} \end{bmatrix} \]

Development of Solution

In order to obtain an estimate for each missing cell, in addition to the observed cells, the parameters for the missing cells must be included in the \( \mu \) vector. This results in a singular design matrix which does not have an inverse. The original "u" model, along with the set of normal equations, is partitioned into an observed segment followed by a missing segment in the following manner:

\[ Y = W\mu + e \]

where

\[ Y = N(Wg, \sigma^2I) \]

and

\[ Y = W\mu + e \]

where

\[ Y = N(Wg, \sigma^2I) \]

192
\[(W'W)u + G' \cdot N'Y \quad \begin{bmatrix} W'Y & W'X & Y & X \\ \end{bmatrix} \begin{bmatrix} \mu \\ \mu + a \\ \lambda + a \\ \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \end{bmatrix} + \begin{bmatrix} 0 \\ a \cdot m \\ \lambda \cdot m \\ \end{bmatrix}
G = 0 \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ \end{bmatrix}
\]

The normal equations may be written in matrix notation:
\[
\begin{bmatrix} \mu \\ \mu + a \\ \lambda + a \\ \end{bmatrix} = \begin{bmatrix} W'Y \\ W'X \\ \end{bmatrix}
\]

To solve for the cell means solution vector, the inverse of the above matrix is determined and the resulting matrix is
\[
\begin{bmatrix} BL^{-1} & -L^{-1} \cdot G \cdot H^{-1} \\ -N^{-1} \cdot G \cdot H^{-1} & L^{-1} \cdot G \cdot AM^{-1} \\ \end{bmatrix}
\]

However, to evaluate the \( \beta \) vector, only the first two rows of the first two columns of this matrix are needed. The cell means solution vector is estimated by
\[
\begin{bmatrix} \beta_0 \\ \beta_m \\ \end{bmatrix} = \begin{bmatrix} BL^{-1} \cdot W'Y \\ -N^{-1} \cdot G \cdot H^{-1} \cdot W'Y \\ \end{bmatrix}
\]

In order to calculate a sum of squares for an F test, the question of interest must be stated explicitly. The matrix notation for the null hypothesis is
\[
H_0: H = 0
\]

where \( H \) is a hypothesis matrix and \( W=WH' \). It has been proven that
\[
F = \frac{\sum (s_{0,0} - \mu) \cdot (s_{0,0} - \mu)}{\sigma^2}
\]

Application to Example

As mentioned previously, the following case serves as a numerical example. Four primates were fed a special diet for a predetermined time period. Blood samples were drawn at five different time intervals. Subsequently, blood chemistry tests for cortisone level were conducted on each sample and the results recorded. However, due to extenuating circumstances, measurements were not reported for all animals under all time intervals. The intervals were irregularly spaced at 500, 600, 1200, 1800, and 2200 units of time from some starting point. It was of interest to determine if a time trend was established by the data.

This experiment was a randomized block design with a random arrangement of treatments. There were two factors, animal and time, with four levels of animal and five levels of time. Each of the four animals served as a complete replication of the experiment. Since all animals were to be tested under each time interval, there would have been one observation per cell. Therefore, the additivity assumption was made which restricted the parameters of the model to no animal x time interaction. The resultant data set consisted of a total of twenty cells including seventeen observed cells and three missing cells. The data which indicates the location of the missing cells are summarized in Table 1. Animal 32 was not tested under time intervals 300 and 1200 and Animal 52 was not tested under interval 1200.

Because the time intervals were unequally spaced, the coefficients were derived by using the orthogonal polynomial command in the MATRIX procedure. Table 2 displays the analyses of variance table. There were no significant animal or time main effects found. A cubic time trend was indicated by the data, but the linear, quadratic, and quartic trends were not significant. The estimates of the cell means are presented in Table 3 with the missing cell estimates included.

Discussion

By using the Yates method, an estimate for each missing cell can be calculated. However, this lengthy process is time consuming when there is more than one missing cell in the data. The estimates produced in the solution vector are identical to those calculated by the iterative process involving the Yates method.

The data were also analyzed by the General Linear Model procedure (GLM) in SAS79. The results were identical. All sums of squares (type IV), degrees of freedom, F values, and least squares means were the same. However, GLM requires more core storage than the "u" model approach. In manually deriving the coefficients for the time trends, the multipliers were essen-
Some of the advantages of the "u" model approach have been mentioned; others deserve consideration. For the actual analysis, the core storage space needed for the cell means model was smaller than for the General Linear Model. An estimate for each missing cell is obtained with no extra calculations or time required. Any constraint condition can be imposed. It is not limited to the no interaction restriction only. Furthermore, any desired hypothesis can be tested. In fact, only the hypotheses specified are tested. With the "u" model, interaction is assumed until it is consciously restricted.

Why would one want to use this method when another method is so readily available and widely used? First, there is no question about the hypotheses being tested and what restrictions, if any, are being imposed. Second, the interpretation of the analysis is more straightforward when population means are compared. Third, many times it is not clear what the General Linear Model procedure uses for the tests, especially when weighted hypotheses or individual comparisons are warranted. Fourth, with the "u" model approach, the researcher need not wonder about the test hypothesis. The GLM procedure assumes nothing until specified, whereas the "u" model assumes all until restricted.

In order to set up the necessary hypotheses and constraint conditions, this approach requires more reasoning and thought on the part of the researcher. However, the researcher will be more confident in the findings and not perplexed by the NON-EST on the SAS printout.

With a few additions, this method is also applicable to all types of designs with various arrangement of treatments for balanced and unbalanced data sets with constraints and balanced data sets without constraints. Several slight modifications are required in the theoretical background and programming procedure.

Table 1.-Location of the observed and missing cells in the primate example.

<table>
<thead>
<tr>
<th>TIME INTERVALS</th>
<th>300</th>
<th>600</th>
<th>1200</th>
<th>1800</th>
<th>2200</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>26.0</td>
<td>67.0</td>
<td>46.0</td>
<td>18.0</td>
<td>27.5</td>
</tr>
<tr>
<td>ANIMALS</td>
<td>32</td>
<td>30.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>25.5</td>
<td>43.0</td>
<td>36.0</td>
<td>33.0</td>
<td>20.5</td>
</tr>
<tr>
<td>52</td>
<td>37.0</td>
<td>34.5</td>
<td></td>
<td>26.0</td>
<td>38.5</td>
</tr>
</tbody>
</table>

Table 2.-Analysis of Variance Table for primate example.

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>3</td>
<td>326.4299</td>
<td>108.8100</td>
<td>1.0103</td>
</tr>
<tr>
<td>Time</td>
<td>4</td>
<td>1060.2465</td>
<td>265.0620</td>
<td>2.4611</td>
</tr>
<tr>
<td>Linear</td>
<td>1</td>
<td>212.6047</td>
<td>212.6047</td>
<td>1.9740</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1</td>
<td>219.8655</td>
<td>219.8655</td>
<td>2.0414</td>
</tr>
<tr>
<td>Cubic</td>
<td>1</td>
<td>551.4873</td>
<td>551.4873</td>
<td>5.1205</td>
</tr>
<tr>
<td>Quartic</td>
<td>1</td>
<td>4.5899</td>
<td>4.5899</td>
<td>0.0426</td>
</tr>
<tr>
<td>Error</td>
<td>9</td>
<td>969.3201</td>
<td>107.7020</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>2372.9706</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.-Estimate of the cells means(1) for the primate example.

<table>
<thead>
<tr>
<th>TIME INTERVALS</th>
<th>300</th>
<th>600</th>
<th>1200</th>
<th>1800</th>
<th>2200</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>31.671</td>
<td>48.560</td>
<td>43.650</td>
<td>29.310</td>
<td>31.310</td>
</tr>
<tr>
<td>ANIMALS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>18.944</td>
<td>35.833</td>
<td>30.926</td>
<td>16.593</td>
<td>18.583</td>
</tr>
<tr>
<td>41</td>
<td>26.371</td>
<td>45.260</td>
<td>38.350</td>
<td>24.010</td>
<td>26.010</td>
</tr>
<tr>
<td>52</td>
<td>30.458</td>
<td>47.347</td>
<td>42.438</td>
<td>28.097</td>
<td>30.097</td>
</tr>
</tbody>
</table>

Bibliography


