Imputing Missing Data for Scale Construction: An Example

Ed Stanek

Introduction

We describe a simple example of data imputation in the context of scale construction. A typical setting consists of a set of questions (items) completed by a set of subjects. We assume that some items are left missing by some respondents. We also assume that a ‘scale’ is desired that is defined as the average response of the full set of items for each subject. For subjects missing a response where the average ‘item’ score was high, their ‘scales’, if computed based on the available items will be too low. Similarly, for subjects missing a response where the average ‘item’ score is low will have too high ‘scales’. As a result, some accounting is needed for the missing items.

We describe a strategy that can be used to impute the ‘missing data’ and still preserve the main effects for subjects and items. The strategy is a simple application of the Healy and Westmacott (1956) strategy (see discussion by Little and Rubin, p26 (1987)). As noted by Little and Rubin, this strategy is an example of an EM algorithm.

The imputation strategy can be described briefly as follows. First, fit a two way factorial model to subject and item responses, with main effects for subjects and for items. Obtain the predicted values for the subject-item combinations where data were missing. Then, substitute the predicted values for the missing items into the appropriate subject-item response. Use the original response for items where response was given.

Illustration

We illustrate this strategy with an example, using data from Little and Rubin (1987), p31. The example consists of a randomized block study. We substitute ‘subjects’ for “blocks” and ‘items’ for ‘factor levels’. With these changes, the resulting data is given in Table 1.

Table 1a. Two Factor Study with Missing Data (Little and Rubin, 1987, p31).

<table>
<thead>
<tr>
<th></th>
<th>Subject 1</th>
<th>Subject 2</th>
<th>Subject 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>Missing</td>
<td>8.0</td>
<td>7.93</td>
</tr>
<tr>
<td>Item 2</td>
<td>8.14</td>
<td>8.15</td>
<td>7.87</td>
</tr>
<tr>
<td>Item 3</td>
<td>7.76</td>
<td>Missing</td>
<td>7.74</td>
</tr>
<tr>
<td>Item 4</td>
<td>7.17</td>
<td>7.57</td>
<td>7.80</td>
</tr>
<tr>
<td>Item 5</td>
<td>7.46</td>
<td>7.68</td>
<td>7.21</td>
</tr>
</tbody>
</table>

Note that two items are missing for two of the subjects.

Imputing the Missing Data

The imputation strategy we use is based on a model. The model we assume is a main effects model. This means that subjects are assumed to have different average item responses, and the
average responses for any two items is allowed to differ. It is noteworthy here that the average response for the subjects is assumed to be the ‘scale’ of interest. In some settings, this definition of the ‘scale’ may not be correct.

With the two factor main effects model, standard analysis of variance methods are used to read in the data (Table 1). Using predicted values from this model fit, a new variable is constructed that replaces the missing values by their predictors (Table 2). A comparison of estimates of the main effects from two way models using the original data, and data including the predicted values reveals that the main effects are identical, but the standard error of the estimates including the predicted values is reduced (Table 3). This reduction in the magnitude of the standard errors is artificial due to the data imputation.

Source:ptes01p08.sas 8/28/01 By EJS
Table 1. List of data from Little and Rubin (1987, p31)

<table>
<thead>
<tr>
<th>Item</th>
<th>subject</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>8.00</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>7.93</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.14</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8.15</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7.87</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7.76</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7.74</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7.17</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7.57</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7.80</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.46</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>7.68</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7.21</td>
</tr>
</tbody>
</table>

Source:ptes01p08.sas 8/28/01 By EJS
Table 2. List of Data with Predicted Values replacing Missing Data

<table>
<thead>
<tr>
<th>Item</th>
<th>subject</th>
<th>response</th>
<th>Pred</th>
<th>resp2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>.</td>
<td>7.85492</td>
<td>7.85492</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>8.00</td>
<td>8.04206</td>
<td>8.00000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>7.93</td>
<td>7.88794</td>
<td>7.93000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>8.14</td>
<td>7.97995</td>
<td>8.14000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8.15</td>
<td>8.16709</td>
<td>8.15000</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>7.87</td>
<td>8.01296</td>
<td>7.87000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>7.76</td>
<td>7.73349</td>
<td>7.76000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>.</td>
<td>7.92063</td>
<td>7.92063</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>7.74</td>
<td>7.76651</td>
<td>7.74000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>7.17</td>
<td>7.43995</td>
<td>7.17000</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>7.57</td>
<td>7.62709</td>
<td>7.57000</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7.80</td>
<td>7.47296</td>
<td>7.80000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.46</td>
<td>7.37661</td>
<td>7.46000</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>7.68</td>
<td>7.56376</td>
<td>7.68000</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>7.21</td>
<td>7.40963</td>
<td>7.21000</td>
</tr>
</tbody>
</table>
Table 3.  Comparison of estimates of main effects using on Imputed Data

<table>
<thead>
<tr>
<th>Effect</th>
<th>Item</th>
<th>subject</th>
<th>Estimates based on</th>
<th>Estimates based on</th>
<th>SE Estimates based on</th>
<th>SE Estimates based on</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>_</td>
<td>7.928307</td>
<td>7.928307</td>
<td>0.163605</td>
<td>0.110810</td>
</tr>
<tr>
<td>Item</td>
<td>2</td>
<td>_</td>
<td>8.053333</td>
<td>8.053333</td>
<td>0.127952</td>
<td>0.110810</td>
</tr>
<tr>
<td>Item</td>
<td>3</td>
<td>_</td>
<td>7.806878</td>
<td>7.806878</td>
<td>0.163605</td>
<td>0.110810</td>
</tr>
<tr>
<td>Item</td>
<td>4</td>
<td>_</td>
<td>7.513333</td>
<td>7.513333</td>
<td>0.127952</td>
<td>0.110810</td>
</tr>
<tr>
<td>Item</td>
<td>5</td>
<td>_</td>
<td>7.450000</td>
<td>7.450000</td>
<td>0.127952</td>
<td>0.110810</td>
</tr>
<tr>
<td>subject</td>
<td>_</td>
<td>1</td>
<td>7.676984</td>
<td>7.676984</td>
<td>0.116470</td>
<td>0.085833</td>
</tr>
<tr>
<td>subject</td>
<td>_</td>
<td>2</td>
<td>7.864127</td>
<td>7.864127</td>
<td>0.116470</td>
<td>0.085833</td>
</tr>
<tr>
<td>subject</td>
<td>_</td>
<td>3</td>
<td>7.710000</td>
<td>7.710000</td>
<td>0.099112</td>
<td>0.085833</td>
</tr>
</tbody>
</table>

A SAS MACRO to Impute Missing Data based on a 2-Factor Main Effect Model, and create a Score

We developed a SAS macro to impute missing values based on a 2-factor main effect model, and then compute a score for each subject based on the observed and imputed values. The development program for the macro is ptrz03p01.sas. The macro is contained in ptrz03p03.sas. A test of the macro using variables for a prevention scale in the Pathways study is contained in the program ptrz03p05.sas. A test of the macro using the data (contained above) from Little and Rubin, 1987, p31 is contained in the program ptrz03p07.sas.

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ptes01p08.sas</td>
<td>Comparison of ANOVA and predicted values using imputation</td>
</tr>
<tr>
<td>ptrz03p01.sas</td>
<td>Develop imputation process</td>
</tr>
<tr>
<td>ptrz03p03.sas</td>
<td>Macro to impute missing value and construct subject scores</td>
</tr>
<tr>
<td>ptrz03p05.sas</td>
<td>Test of macro using Pathways Prevention scores</td>
</tr>
<tr>
<td>ptrz03p07.sas</td>
<td>Test of macro using Little and Rubin (1987) p31 data.</td>
</tr>
</tbody>
</table>

References:
