Comparison of Imputation Accuracy Based on Imputation Class Definitions for IPEDS Finance Survey

Marcus Berzofsky

Joint Statistical Meetings

Minneapolis, MN  August 7-11, 2005
Contents

- General background
- Overview of problem
- Methodology
  - Test if response is missing completely at random
  - Comparison of imputation class designs
  - Analysis of precision and accuracy of imputation classes
- Results & Conclusions
IPEDS, Integrated Postsecondary Education Data System, is a set of annual surveys that institutions that participated in title IV student aid programs are required to complete.

The Finance survey is a required component that obtains information on an institution’s assets as well as their revenue and expenditures.

Different survey form for Public, Private not-for-Profit and For-Profit institutions.
Nonresponding institutions’ data are imputed

Survey contains over 200 variables that are simultaneously imputed (i.e. one donor institution) to maintain consistency across variables

Missing data imputed by 3 different methods

Will focus on nearest neighbor imputation method
  - Procedure used when nonrespondent institution does not have prior year data
Two ways to form imputation classes:

- A non-statistical imputation class using subject matter expertise
- A statistical imputation class using information from principle component analysis (PCA) and Chi-square automatic detector indicator (CHAID)

Need to determine which way results in more accurate imputation
Methodology

- Test if nonresponse is missing completely at random
  - Define universe of eligible institutions as those that responded to 2005 survey (N=5,084)
  - Define nonrespondent universe as those that responded to 2005 survey, but were nonrespondents in either 2002, 2003 or 2004 surveys (N=388)
Select 500 sets of samples (n=49) from both sets of institutions

- Random nonresponse samples selected via stratified SRS proportionally to the distribution of institutions by institutional level and control from among all eligible institutions
- Nonrandom nonresponse samples selected by SRS among nonrespondent institutions
- Sample size based on number of total nonrespondents to 2005 survey
For each sample, removed respondent data and imputed using both the PCA/CHAID and the non-statistical imputation classes.

The nearest neighbor procedure was used to compute imputations:
- Uses distance formula to select donor institution from imputation class.

Regress the imputed value onto the respondent value for each variable using the no intercept model:
- \[ \text{IMPUTED} = B \times \text{RESPONSE} + e \]
Methodology (cont.)

- Compare distribution of beta estimates for each variable across all 500 samples and across all finance variables
  - Calculate MSE of beta estimates for each variable
    - \( \text{MSE}(B_i) = (B_i - 1)^2 + V(B_i) \)
  - Compare distribution of MSEs across all variables
## Distribution of Eligible Population and Samples by Institutional Level and Control

<table>
<thead>
<tr>
<th>Institutional Level and Control</th>
<th>% Dist. of Eligible Population (N=5,084)</th>
<th>% Dist. of Nonrespondent Universe (N=388)</th>
<th>Avg. % Random Sample Dist. (n=49)</th>
<th>Avg. % Nonrandom Sample Dist. (n=49)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administrative unit only</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Public, 4 year and above</td>
<td>11.15</td>
<td>1.29</td>
<td>11.14</td>
<td>1.31</td>
</tr>
<tr>
<td>Private not-for-profit, 4 year and above</td>
<td>28.40</td>
<td>15.98</td>
<td>28.41</td>
<td>16.04</td>
</tr>
<tr>
<td>Private for-profit, 4 year and above</td>
<td>4.45</td>
<td>2.58</td>
<td>4.44</td>
<td>2.57</td>
</tr>
<tr>
<td>Public, 2 year</td>
<td>18.17</td>
<td>10.05</td>
<td>18.17</td>
<td>10.10</td>
</tr>
<tr>
<td>Private not-for-profit, 2 year</td>
<td>2.68</td>
<td>4.38</td>
<td>2.69</td>
<td>4.08</td>
</tr>
<tr>
<td>Private for-profit, 2-year</td>
<td>11.53</td>
<td>16.24</td>
<td>11.50</td>
<td>16.05</td>
</tr>
<tr>
<td>Public, less than 2-year</td>
<td>3.11</td>
<td>7.73</td>
<td>3.20</td>
<td>7.76</td>
</tr>
<tr>
<td>Private not-for-profit, less than 2-year</td>
<td>1.24</td>
<td>1.55</td>
<td>1.19</td>
<td>1.54</td>
</tr>
<tr>
<td>Private for-profit, less than 2-year</td>
<td>19.28</td>
<td>40.21</td>
<td>19.27</td>
<td>40.55</td>
</tr>
</tbody>
</table>
Using a chi-square test of homogeneity, the distributions of the entire population and the nonrespondent population were not equal

- Test Statistic = 215.36
- P-value less than 0.0001

Concluded that response is not missing completely at random
Non-statistical Imputation Classes

- Based on subject matter expertise of the data set
- Identified following variables for classes
  - Form used (3 different forms; Public, Private not-for-profit, Private for-profit)
  - Institutional level and Control
  - Medical school
  - Level of offering (first-professional, graduate, undergraduate)
  - Census division (for public institutions only)
- Minimum allowable size of class was 9 institutions
  - If class was less than 9 institutions, it was collapsed with closely related class
- 31 imputation classes created
PCA/CHAID Imputation Classes

- Selected summary variables from each form (about 20 variables across all three forms)
- PCA provides means to summarize all 200 variables into one single value
  - This allows pooling of institutions with similar responses across all variables
- PCA was applied to generate weight for each variable
- Index for each institution created by summing the product of weight*variable across all summary variables
PCA/CHAID Imputation Classes (cont.)

- Answer Tree 2.0 software was used in generating imputation classes
- CHAID uses Chi-square or F statistics to identify optimal splits
- Index was used as outcome variable (i.e., institutions with similar index values would be grouped together)
- Predictor variables were determined by prior knowledge about the institutional characteristics. The significance of the variables on the finance data set were determined through regression analysis using the index variable as the dependent outcome
PCA/CHAID analysis was computed for the past three years of finance data. Variables consistently significant across all three years of data were used to create the final imputation classes.

Imputation classes were defined by common branches generated from three separate Decision Trees for each survey form.

Variables that were used in defining imputation classes include: degree granting status, institutional level and control, Offer graduate/first-professional, medical school, FTE groupings, student services and sports variables.

State (FIPS) was never part of the main branches and, therefore, not included in defining imputation groups.

Total of 45 imputation classes were created from Decision Tree.
CHAID Tree Example

Index
Mean: 0.1150
SD: 1.96
n: 1675
Predict: 0.1150

DEGGRANT
P-value 0.000; F=1022.19; df=1,1673

1
Mean: 0.5615
SD: 1.5073
n: 1474 (88.0%)
Predicted: 0.5615

2
Mean: -3.1595
SD: 1.8188
n: 201 (12.0%)
Predicted: -3.1595
Analysis

Following tables based on results from nonrespondent set of samples
PCA/CHAID Method vs. Non-statistical Method

Public Institutions Net Assets Variables
Nonrandom Sample

<table>
<thead>
<tr>
<th>Variable</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1=Current Assets</td>
<td>1</td>
</tr>
<tr>
<td>2=Capital Assets</td>
<td>2</td>
</tr>
<tr>
<td>3=Accumulated Depreciation</td>
<td>3</td>
</tr>
<tr>
<td>5=Total Noncurrent Assets</td>
<td>5</td>
</tr>
<tr>
<td>7=Long Term Debt, current portion</td>
<td>7</td>
</tr>
<tr>
<td>9=Total Current Liabilities</td>
<td>9</td>
</tr>
<tr>
<td>10=Long Term Debt</td>
<td>10</td>
</tr>
<tr>
<td>12=Total Noncurrent Liabilities</td>
<td>12</td>
</tr>
<tr>
<td>14=Invested in Capital Assets</td>
<td>14</td>
</tr>
<tr>
<td>15=Restricted-expendable</td>
<td>15</td>
</tr>
<tr>
<td>16=Restricted-nonexpendable</td>
<td>16</td>
</tr>
</tbody>
</table>
PCA/CHAID Method vs. Non-statistical Method
Private Not-for-Profit Institutions Asset Variables
Nonrandom Sample

Variable:

1 = Long-term investment
2 = Total assets
3 = Total liabilities
4 = Total unrestricted net assets
5 = Total restricted net assets
6 = Total net assets

Beta

PCA/CHAID
Non-Statistical
PCA/CHAID Method vs. Non-statistical Method

Private For-Profit Institutions Asset Variables

Nonrandom Sample

1=Total assets        2=Total liabilities        5=Total revenues        6=Total expenses

Beta

1a 1b 2a 2b 5a 5b 6a 6b

PCA/CHAID

Variable

Non-Statistical
Comparison of MSE Across All Samples
All Institutions by Control
Nonrandom Sample

Control by Method

PCA/CHAID Non-Statistical
Comparison of MSE Across All Samples

All Institutions
Nonrandom Sample

MSE

CHAID  Non-statistical

Method

PCA/CHAID  Non-Statistical
## Distribution of MSE Across All Variables by Control and Imputation Class Method

<table>
<thead>
<tr>
<th></th>
<th>Public</th>
<th>Private not-for-Profit</th>
<th>Private for-Profit</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCA/CHAID</td>
<td>Non-Stats</td>
<td>PCA/CHAID</td>
<td>Non-Stats</td>
</tr>
<tr>
<td>Mean</td>
<td>46,201,054</td>
<td>10,108.86</td>
<td>37.12</td>
<td>1,571.54</td>
</tr>
<tr>
<td>Min</td>
<td>0.10408</td>
<td>0.10312</td>
<td>0.15912</td>
<td>0.09906</td>
</tr>
<tr>
<td>25&lt;sup&gt;th&lt;/sup&gt; %</td>
<td>0.30496</td>
<td>0.33615</td>
<td>0.96128</td>
<td>0.92724</td>
</tr>
<tr>
<td>Median</td>
<td>0.99953</td>
<td>0.92544</td>
<td>1.44954</td>
<td>1.32948</td>
</tr>
<tr>
<td>75&lt;sup&gt;th&lt;/sup&gt; %</td>
<td>23.2158</td>
<td>14.3219</td>
<td>10.5242</td>
<td>5.50132</td>
</tr>
<tr>
<td>Max</td>
<td>2.5 Billion</td>
<td>296,119.62</td>
<td>1,884.01</td>
<td>110,965</td>
</tr>
</tbody>
</table>

Mean: 22537115, 5,570.8

Min: 0.10408, 0.46153

25<sup>th</sup> %: 0.10408, 0.46153

Median: 0.99953, 1.19236

75<sup>th</sup> %: 23.2158, 10.9853

Max: 2.5 Billion, 296,119
Conclusions

- The PCA/CHAID method was hampered by the following limitations and constraints:
  - Require one donor institution per imputee for all variables
  - About 20 key summery finance variables were used in PCA to represent over 200 finance variables. These variables only explained 53% to 70% of the variation
  - Only kept common major branches in the decision tree due to year to year consistency issues
Conclusions (cont.)

- When specific knowledge about the data is known prior to imputation, it is possible to create imputation classes that generate as accurate imputations as the PCA/CHAID produce.

- We recommend the non-statistical method because it is easier to implement and provides homogeneous imputation classes across years.
Questions?