Imputation of Missing Data in a Split Questionnaire Design

Matt Stuart

Iowa State University – CSSM

11/4/2016
1 Motivation

2 Available Case

3 Multiple Imputation

4 Simulation

5 Next Steps
AVMA Pet Demographic Survey

- The American Veterinary Medical Association (AVMA) has tasked the CSSM to complete a survey on Pet Demographics. This is a survey that is conducted every five years.
- The survey is broken down into fifteen different modules based on fifteen types of pets surveyed, with the number of questions per module based on the type of pet.
The current survey is too long, for example, 112 different variables amongst 22 questions in the dog module.

The response rate among the 222,244 households that received the survey was only 22.7 %

This low response rate along with response fatigue with the long questionnaire may have led to data that is inherently biased.
Split the questions in each module into smaller parts that takes at most 20 minutes to complete.

Impute the values for the missing questions based on the answers of the previous questions.

Further Design Issues:
- How to make up the splits?
- How to impute the missing values?
Multiple Impuation

- This discussion is regarding the imputation of the missing values using methods from (Raghunathan, et al 1995).
- Compare methods of available case vs. multiple imputation
1 Motivation

2 Available Case

3 Multiple Imputation

4 Simulation

5 Next Steps
Method

- Want to relate $Y$ to predictors $X_1$ and $X_2$ with each $X$ being the average of the values measured on the item.
- Mean and variance of $Y$ is from all sampling units
- Mean and variance of $X_1$ and covariance between $X_1$ and $Y$ is from those who received Part 1
- Mean and variance of $X_2$ and covariance between $X_2$ and $Y$ is from those who received Part 2
- These statistics are substituted for the corresponding complete data sufficient statistics to estimate regression coefficients
- Convenient algorithm uses SWEEP operator of matrix of means variances and covariances, which can also obtain an approximate covariance matrix
Though the covariance estimate is used in practice, it lacks any theoretical justification.

An alternate is available using asymptotic arguments from, though this method is computationally complex to implement.

Because of the different sample spaces for means, variances and covariances, the $\mathbf{X}^t\mathbf{X}$ may not be positively definite.

It also fails to utilize the correlation between variables in different parts, which can gain efficiency, especially if the splits were designed so variables with high correlation were assigned to different parts.
1. Motivation

2. Available Case

3. Multiple Imputation

4. Simulation

5. Next Steps
Model Assumptions

- Suppose we observe a $p$-dimensional continuous variable $Y$ and a $q$-dimensional categorical variable $Z$.
- The individual variable $Z_j$ has $C_j$ levels.
- $n_{i_1 i_2 \ldots i_q}$: the number of individuals where $Z_1 = i_1, \ldots, Z_q = i_q$
- $i_j = 1, 2, \ldots, C_j$: the $j$th level of variable $i$. 
The cells form a multinomial random variable: \( \pi = (\pi_{i_1\ldots i_q}) \).

Due to the high number of cells (C), a log-linear structure on the cell probabilities is used:

\[
\log \pi = \alpha + \sum_{j=1}^{q} \beta_{ij} + \sum_{j=1}^{q} \sum_{k=1, k \neq j}^{q} \gamma_{ij}i_{jk} + \\
\sum_{j=1}^{q} \sum_{k=1, k \neq j}^{q} \sum_{l=1, l \neq j, l \neq k}^{q} \delta_{ij}i_{jk}i_{lj} + \ldots \quad (1)
\]

Most times, second-order terms are sufficient, but can include higher terms if warranted.
Model Assumptions (cont.)

- $\mathbf{Y} \sim \text{Normal}(\mu = \mu_m, \sigma = \Sigma)$ for a cell $m = i_1...i_q$ (2)
- $\mu_m \sim \text{Normal}(\mu = V_m \tau, \sigma = \Omega)$ (3)
- $V_m$ is a $p \times r$ matrix of covariates defined by $\mathbf{Z}$.
- $\tau$ is a $r \times 1$ matrix of regression coefficients.
Model Assumptions (cont.)

- \( \omega = (\tau, \Omega, \Sigma, \alpha, \beta_i, \ldots) \) (4)
- A prior distribution for \( \omega \) is assumed where \( P(\omega) \propto \frac{1}{|\Omega||\Sigma|} \)
Gibbs Sampling

- Using these model specifications, imputations will be made using Gibbs sampling in a cyclic manner from the multivariate posterior distribution.
- Run an initializing phase to eliminate the effect of the initial draw as a prespecified number of cycles.
- $M$ is the number of values we will store as values for the missing data
- $P$ is the ordered value to be stored
Gibbs Sampling

- Draw a total of $MP$ values from the relevant conditional distributions in the cyclic manner.
- We store every $P$th draw as an approximately independent draw from the joint posterior distribution to obtain a total of $M$ imputed values.
Let $\mathbf{T}$ denote a $k$-dimensional population quantity of interest.
- Population mean of $k$ characteristics
- Regression coefficients of a model

Let $\hat{T}_{*,w}$ and $\hat{V}_{*,w}$ be the estimate and associated variance based on the $w$th completed data from the Gibbs sampling.

$w = 1, 2, \ldots, M$
Multiple Imputation

Estimation Using Multiply Imputed Data (cont.)

\[ \hat{T}_{MI} = \sum_{w=1}^{M} \frac{\hat{T}_{*w}}{M} \]  
\[ \hat{V}_{MI} = (1 + r_m) \sum_{w=1}^{M} \frac{\hat{V}_{*w}}{M} \]  
\[ r_m = (1 + \frac{1}{M}) \sum_{w=1}^{M} (\hat{T}_{*w} - \hat{T}_{MI}) \hat{V}_{*w}^{-1} (\hat{T}_{*w} - \hat{T}_{MI})^t \left( \frac{1}{k(M-1)} \right) \]

(Li, et. al 1991) suggests using a t-distribution with 
\[ df = 4 + (t - 4) * \left(1 + \frac{1-2}{r_m} \right)^2 \]  
where \( t = k(M - 1) \) to construct confidence intervals.
Suppose we have $R$ models to be considered for either $\pi_m$ or $\mu_m$.

Denote

- $X_{obs}$ as the observed values of $Y$ and $Z$
- $X_{mis}(t)$ as the drawn values of the missing $Y$ and $Z$ at the $t$th cycle
- $\omega(t)$ as the $t$th drawn value of $\omega$
- $L_r(t) = L_r(\omega(t)|X_{obs}, X_{mis}(t))$ (8) as the complete data posterior density under model $r$. 
Run $T = MP$ Gibbs Cycles
Marginal Likelihood of model $r$ denoted as:

$$L_r = \sum_{t=1}^{T} \frac{L_r(t)}{T} \quad (9)$$

If all $R$ models are equally preferable, then the posterior probability of model $r$ being correct is approximately:

$$P_r = \frac{L_r}{\sum_{r=1}^{R} L_r} \quad (10)$$
Smoking

- Survey conducted on smoking and alcohol consumption, along with prevalence of other health risks and prevention methods via the Cancer Risk Behavior study.
- Random-digit dialing with an oversampling of households from large cities or communities conducted across state of Washington.
- Obtained full data estimates on the 1989 respondents of a possible 2590 individuals.
- Goal is to reduce questionnaire response time by 60%.
Smoking (cont.)

- 127 variables were of interest for the survey
  - 14 corresponded to common items of a core component
  - 113 others were assigned to five different question parts
  - Items with higher correlation coefficients were placed in different parts (will discuss later).

- 12 important variables will be inspected for comparison
20 roughly equivalent splits were constructed
Each was analyzed to assess sensitivity of inference

Define:
- $G_i$ be the point estimate of a variable based on the $i$th split data
- $G$ be the point estimate of a variable based on the complete data
- $S_i$ be the variable’s standard error based on the $i$th split data
- $S$ be the variable’s standard error based on the complete data
The discrepancy measures $A$ and $B$ are defined as:

- $A = 100 \times \frac{1}{20} \sum_{i=1}^{20} \left( \frac{(G_i - G)^2}{G} \right)^{0.5}$
- $B = 100 \times \frac{1}{20} \sum_{i=1}^{20} \left( \frac{(S_i - S)^2}{S} \right)^{0.5}$
<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Whole questionnaire</th>
<th>Split questionnaire method</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Is diet related to cancer?</td>
<td>64  1.6</td>
<td>63  1.8</td>
<td>4.6</td>
<td>7.3</td>
<td>62  2.1</td>
</tr>
<tr>
<td>If yes, is it strong?</td>
<td>39  2.2</td>
<td>41  2.8</td>
<td>5.7</td>
<td>8.2</td>
<td>40  3.6</td>
</tr>
<tr>
<td>Is diet related to heart disease?</td>
<td>76  0.9</td>
<td>79  1.2</td>
<td>4.1</td>
<td>5.1</td>
<td>77  1.6</td>
</tr>
<tr>
<td>If yes, is it strong?</td>
<td>62  1.1</td>
<td>60  1.4</td>
<td>4.5</td>
<td>6.9</td>
<td>60  1.5</td>
</tr>
<tr>
<td>Do you eat vegetable other than salad in lunch?</td>
<td>37  2.1</td>
<td>35  2.3</td>
<td>8.5</td>
<td>8.2</td>
<td>33  3.4</td>
</tr>
<tr>
<td>Do you eat vegetable other than salad in dinner?</td>
<td>61  1.2</td>
<td>62  1.3</td>
<td>4.2</td>
<td>5.7</td>
<td>60  1.6</td>
</tr>
<tr>
<td>Have you decreased your fat intake?</td>
<td>73  1.0</td>
<td>72  1.3</td>
<td>6.2</td>
<td>8.7</td>
<td>70  1.9</td>
</tr>
<tr>
<td>To make your diet healthy would you Take a supplement?</td>
<td>15  1.3</td>
<td>19  1.4</td>
<td>9.1</td>
<td>9.4</td>
<td>18  3.2</td>
</tr>
<tr>
<td>Eat vegetables?</td>
<td>82  0.8</td>
<td>86  1.3</td>
<td>6.8</td>
<td>9.2</td>
<td>84  2.1</td>
</tr>
<tr>
<td>Have you ever had mammography?</td>
<td>70  2.0</td>
<td>70  2.2</td>
<td>2.7</td>
<td>3.2</td>
<td>68  3.7</td>
</tr>
<tr>
<td>Have you ever had Pap smear?</td>
<td>88  1.1</td>
<td>87  1.2</td>
<td>2.2</td>
<td>3.6</td>
<td>88  1.8</td>
</tr>
<tr>
<td>Have you ever had hemocult test?</td>
<td>60  2.2</td>
<td>60  2.4</td>
<td>5.1</td>
<td>5.9</td>
<td>60  4.1</td>
</tr>
<tr>
<td>Have you ever had digital rectal exam?</td>
<td>48  2.4</td>
<td>44  2.9</td>
<td>11.2</td>
<td>12.5</td>
<td>46  3.1</td>
</tr>
<tr>
<td>Does having two drinks a day during pregnancy cause problems?</td>
<td>92  1.1</td>
<td>90  1.1</td>
<td>1.3</td>
<td>2.1</td>
<td>90  1.2</td>
</tr>
<tr>
<td>Is it dangerous to drive after two drinks (moderate or strong)?</td>
<td>60  1.3</td>
<td>62  1.8</td>
<td>4.2</td>
<td>3.9</td>
<td>64  2.9</td>
</tr>
<tr>
<td>Have you seen warning label on alcoholic beverages?</td>
<td>28  2.2</td>
<td>25  3.1</td>
<td>13.2</td>
<td>15.2</td>
<td>29  5.2</td>
</tr>
</tbody>
</table>

NOTE: The discrepancy measures A and B defined in Section 4 are expressed as percentages.
Simulation # 2

- Objectives:
  - Explore Loss of Efficiency of Split Questionnaire as a function of correlation of variables within and between question components
  - Explore sensitivity to departures of assumed normality of $Y$ and $\mu_m$
Full questionnaire contains 15 continuous variables ($\mathbf{Y}$) and 5 binary variables ($\mathbf{Z}$)

Complete data were generated using general model from (1), (2), and (3) where (1) uses third-order interaction terms and lower

$V_m \tau = 0$

$\Omega, \Sigma$ chosen to ensure:
- Correlation coefficient between any two variables in same part is $\rho_w$
- Correlation coefficient between any two variables in different parts is $\rho_b$
Levels of $\rho_w$ and $\rho_b$ have been specified as:
- $\rho_w = (0.1, 0.3)$
- $\rho_b = (0.3, 0.5, 0.8)$

Distributions of $Y$ and $\mu_m$ will vary as follows:
- Normal distribution
- $t$-distribution (df=3)
- Shifted exponential (location=0)
For each condition 1000 complete data sets were generated.
Each simulated complete data set had a sample of 1000 units.
Assuming fourth-order or higher interaction terms were of no interest, the sample was randomly assigned one of 10 combinations of 3 components (out of 5) such that each combination went to 100 sampling units.
The model for imputations included a log-linear model for categorical variables with all main, two-factor, and three-factor interaction effects.

50 imputed sets of values were obtained from 5 parallel Gibbs sequences and storing every 250th draw for $P = 100$

1000 estimates based on split data and complete data were computed along with the following statistics:

- Bias
- Mean Square Error (MSE)
- Exact Coverage of 90% CI
- Length of CI
Table 2. Comparison of Split and Complete Data Estimates of the Regression Coefficient in the Logistic Model.

<table>
<thead>
<tr>
<th>Distribution of $\mu_m$</th>
<th>$Y$</th>
<th>Characteristic</th>
<th>$\rho_w = .3$</th>
<th>$\rho_w = .1$</th>
<th>$\rho_w = .8$</th>
<th>$\rho_w = .3$</th>
<th>$\rho_w = .1$</th>
<th>$\rho_w = .8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\rho_B = .3$</td>
<td>$\rho_B = .5$</td>
<td>$\rho_B = .8$ (1)</td>
<td>$\rho_B = .3$</td>
<td>$\rho_B = .5$</td>
<td>$\rho_B = .8$ (1)</td>
</tr>
<tr>
<td>$N$</td>
<td>$N$</td>
<td>Bias</td>
<td>.0001</td>
<td>.0002</td>
<td>-.0002</td>
<td>.0001</td>
<td>.0002</td>
<td>-.0002</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>MSE</td>
<td>122.7</td>
<td>110.2</td>
<td>101.2</td>
<td>131.9</td>
<td>118.3</td>
<td>108.3</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Coverage</td>
<td>91.3</td>
<td>92.3</td>
<td>92.5</td>
<td>91.7</td>
<td>91.3</td>
<td>9.17</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Length</td>
<td>112.1</td>
<td>106.7</td>
<td>106.6</td>
<td>114.8</td>
<td>111.7</td>
<td>105.9</td>
</tr>
<tr>
<td>$N$</td>
<td>$t_3$</td>
<td>Bias</td>
<td>.0013</td>
<td>.0022</td>
<td>-.0032</td>
<td>.0011</td>
<td>.0012</td>
<td>-.0012</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>MSE</td>
<td>123.1</td>
<td>114.5</td>
<td>106.9</td>
<td>124.3</td>
<td>119.2</td>
<td>107.5</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Coverage</td>
<td>93.5</td>
<td>92.3</td>
<td>93.5</td>
<td>92.5</td>
<td>91.3</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Length</td>
<td>105.8</td>
<td>102.4</td>
<td>101.6</td>
<td>115.5</td>
<td>111.2</td>
<td>103.1</td>
</tr>
<tr>
<td>$N$</td>
<td>$E$</td>
<td>Bias</td>
<td>.0322</td>
<td>.0113</td>
<td>.0132</td>
<td>.0113</td>
<td>.0122</td>
<td>.0032</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>MSE</td>
<td>185.6</td>
<td>158.2</td>
<td>136.9</td>
<td>192.6</td>
<td>132.7</td>
<td>124.5</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Coverage</td>
<td>86.5</td>
<td>87.3</td>
<td>87.5</td>
<td>90.5</td>
<td>89.3</td>
<td>89.5</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Length</td>
<td>168.2</td>
<td>146.2</td>
<td>119.5</td>
<td>150.8</td>
<td>131.6</td>
<td>119.5</td>
</tr>
<tr>
<td>$t_3$</td>
<td>$N$</td>
<td>Bias</td>
<td>.0001</td>
<td>.0002</td>
<td>-.0002</td>
<td>.0001</td>
<td>.0002</td>
<td>-.0002</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>MSE</td>
<td>121.7</td>
<td>109.8</td>
<td>100.5</td>
<td>117.5</td>
<td>109.9</td>
<td>102.0</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Coverage</td>
<td>91.9</td>
<td>92.9</td>
<td>91.5</td>
<td>91.9</td>
<td>92.9</td>
<td>91.5</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Length</td>
<td>111.6</td>
<td>107.8</td>
<td>101.7</td>
<td>112.5</td>
<td>109.5</td>
<td>103.8</td>
</tr>
<tr>
<td>$t_3$</td>
<td>$t_3$</td>
<td>Bias</td>
<td>.0015</td>
<td>.0012</td>
<td>-.0012</td>
<td>.0015</td>
<td>.0012</td>
<td>-.0012</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>MSE</td>
<td>127.6</td>
<td>112.0</td>
<td>101.5</td>
<td>129.8</td>
<td>104.3</td>
<td>112.2</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Coverage</td>
<td>92.5</td>
<td>92.3</td>
<td>92.5</td>
<td>92.5</td>
<td>92.3</td>
<td>92.5</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Length</td>
<td>114.7</td>
<td>109.4</td>
<td>103.1</td>
<td>115.3</td>
<td>106.2</td>
<td>102.2</td>
</tr>
<tr>
<td>$t_3$</td>
<td>$E$</td>
<td>Bias</td>
<td>.0462</td>
<td>.0399</td>
<td>.0213</td>
<td>.0063</td>
<td>.0062</td>
<td>.0029</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>MSE</td>
<td>196.9</td>
<td>161.4</td>
<td>125.5</td>
<td>191.9</td>
<td>158.8</td>
<td>110.8</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Coverage</td>
<td>86.5</td>
<td>87.3</td>
<td>86.5</td>
<td>91.5</td>
<td>91.3</td>
<td>90.5</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Length</td>
<td>160.0</td>
<td>140.2</td>
<td>111.2</td>
<td>149.9</td>
<td>138.7</td>
<td>114.1</td>
</tr>
<tr>
<td>$E$</td>
<td>$N$</td>
<td>Bias</td>
<td>.0011</td>
<td>.0012</td>
<td>-.0012</td>
<td>.0031</td>
<td>.0011</td>
<td>-.0011</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>MSE</td>
<td>124.6</td>
<td>109.5</td>
<td>102.9</td>
<td>132.5</td>
<td>110.1</td>
<td>106.7</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Coverage</td>
<td>89.3</td>
<td>91.3</td>
<td>91.5</td>
<td>89.9</td>
<td>91.3</td>
<td>91.8</td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>Length</td>
<td>119.4</td>
<td>111.1</td>
<td>101.3</td>
<td>118.4</td>
<td>115.5</td>
<td>105.4</td>
</tr>
<tr>
<td>$E$</td>
<td>$t_3$</td>
<td>Bias</td>
<td>.0013</td>
<td>.0022</td>
<td>-.0032</td>
<td>.0014</td>
<td>-.0012</td>
<td>-.0002</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>MSE</td>
<td>120.2</td>
<td>117.2</td>
<td>113.5</td>
<td>140.0</td>
<td>132.6</td>
<td>119.2</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Coverage</td>
<td>91.5</td>
<td>91.3</td>
<td>92.5</td>
<td>91.5</td>
<td>91.3</td>
<td>91.5</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>Length</td>
<td>115.5</td>
<td>108.4</td>
<td>103.7</td>
<td>126.8</td>
<td>109.8</td>
<td>102.9</td>
</tr>
<tr>
<td>$E$</td>
<td>$E$</td>
<td>Bias</td>
<td>.0413</td>
<td>.0224</td>
<td>.0132</td>
<td>.0103</td>
<td>.0149</td>
<td>.0102</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>MSE</td>
<td>196.9</td>
<td>168.2</td>
<td>123.5</td>
<td>203.6</td>
<td>177.6</td>
<td>121.9</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Coverage</td>
<td>84.5</td>
<td>86.3</td>
<td>87.5</td>
<td>82.5</td>
<td>86.3</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>$E$</td>
<td>Length</td>
<td>217.9</td>
<td>145.2</td>
<td>113.2</td>
<td>217.8</td>
<td>159.7</td>
<td>139.5</td>
</tr>
</tbody>
</table>

NOTE: This table shows the bias, the mean squared error of the split data estimate as a percent of the mean squared error of the complete data estimate, the exact coverage in percent of the nominal 90% confidence interval, and the expected length of the nominal 90% confidence interval based on the split data as a percent of the expected length of the complete data confidence interval across 54 conditions. The distributions of $Y$ and $\mu_m$ considered are $N$, Normal; $E$, shifted exponential; and $t_3$, $t$ on 3 degrees of freedom.
1 Motivation

2 Available Case

3 Multiple Imputation

4 Simulation

5 Next Steps
Next Steps

- Determine if this method will be implemented for the AVMA Pet Survey
- If this is the case:
  - Determine splits that minimize the amount of information lost from the survey
  - Run multiple imputation and determine difference between our results and that of the survey
References


Thank You

Questions?