BIOSTAT 2065
Analysis of Incomplete Data

Gong Tang
Dept of Biostatistics
University of Pittsburgh

Fall, 2005
Textbook


Some References
Course Requirements

• Homework assignments.
• Mid-term exam.
  --- Cover part I and II of the textbook.
  --- Open book.
• One final project.
  --- Data analysis or literature review.
Incomplete (Missing) Data

Examples

• Random sample.
• Missing values of variables.
• Censored data.

Importance

• Standard statistical techniques are usually based on complete data (sets).
• Ignoring the nature of missing data often leads to biased estimates.
Univariate Missing Data (I)

- Missing values occur from a sample of $X \sim N(0,1)$, for example,

(i) Randomly missing:

each observation has prob=0.5 to be observed.

(ii) Censored:

observation that is over 0 is missing.

(iii) Stochastic censoring:

an observation $x$ has $\text{prob}=\Phi(x + 0.5)$
to be observed.
Univariate Missing Data (II)

- $X$ is fully observed.
- $Y$ is missing for some subjects.
- Cause unbalance in experiments.
**Complete Data**

![Complete Data Diagram]

**Missing for x > 1**

![Missing for x > 1 Diagram]

**Missing for y > 1**

![Missing for y > 1 Diagram]

**Fit Comparison**

![Fit Comparison Diagram]

\( Y = 1 + X + N(0, 0.25) \)
Unit and Item Nonresponse in Surveys

- \( X \) --- design variables
  (e.g., household location, age, sex etc.)

- \( Y \) --- survey variables
  (e.g., income, energy consumption)

- \( X \) --- design variables
  (e.g., household location, age, sex etc.)

- Unit nonresponse:
  all the values of \( Y \) are missing.

- Item nonresponse:
  the values of \( Y \) are partially observed.
Longitudinal Dropout Data

In longitudinal studies, some subjects may drop out prematurely due to various reasons: moving, treatment failure, death, consent withdrawal, etc.

- Missing data pattern: monotone.
Latent Variables

- Measurement error model
  \[ Z = X + \delta \]
  \[ Y = \beta_0 + \beta_1 Z + \varepsilon \]
- Random-effects models
  \[ Y_{ij} = \beta_0 + \beta_1 X_{ij} + Z_i + \varepsilon_{ij}, \]
  \[ Z_i \sim N(0, \nu^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2) \]
  \[ i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, J. \]
An Assumption (I)

- **Missing data indicators hide true values that are meaningful for analysis.**
- Causal effects in randomized clinical trials.
  (a) Patients are randomized into two intervention groups: $T=0, 1$.
  The outcome is death status: $D=0, 1$.
  Let $D(t)$ be the potential outcome given $T=t$.
  Individual causal effect = $D_i(1)-D_i(0)$.
  (b) Causal effect: $E\{D(1)-D(0)\}$.
  (c) For subject $i$, only one of $\{D_i(0), D_i(1)\}$ is observed; the other one is missing by design.
An Assumption (II)

• If the interest is “quality-of-life health” indicator $Y$, given the patients is alive.
• Split the population into groups:
  (a) $LL$: $D(0)=D(1)=0$;
  (b) $LD$: $D(0)=0$, $D(1)=1$;
  (c) $DL$: $D(0)=1$, $D(1)=0$;
  (d) $DD$: $D(0)=D(1)=1$.
• Causal inference of $T$ on $Y$ can only made through the subpopulation $LL$. 
The Joint Distribution

- Complete data $Y = (y_{ij})$,

Missing-data indicators $M = (M_{ij})$,

where, $M_{ij} = \begin{cases} 0, & \text{if } y_{ij} \text{ is observed;} \\ 1, & \text{if } y_{ij} \text{ is missing.} \end{cases}$

$i = 1, 2, ..., n; \ j = 1, ..., J$.

- Joint distribution

$$f(Y, M; \theta, \phi) = f(Y; \theta) f(M | Y; \phi)$$

$$= \prod_{i=1}^{n} f(Y_i; \theta) \prod_{i=1}^{n} f(M_i | Y_i; \phi) \quad \text{<independent>$$}
Missing-data Mechanism

- **Missing completely at random (MCAR)** if
  \[ f(M | Y; \phi) = f(M; \phi), \text{ for all } Y. \]

- **Missing at random (MAR)** if
  \[ f(M | Y; \phi) = f(M | Y_{obs}; \phi), \text{ for all } Y. \]
  where \( Y = (Y_{obs}, Y_{mis}). \)

- **Not missing at random.**

- **Examples**
  (a) Longitudinal dropout data.
  (b) General bivariate missing data.
General Approaches for Missing Data

- Complete-case analysis.
- Weighting procedures.
- Imputation-based procedures.
- Model-based procedures.
  (a) Maximum likelihood based.
  (b) Estimating equations based.
  (c) Gibbs sampler.
1.2. MISSING-DATA PATTERNS

(a) Univariate Nonresponse

\[ Y_1, Y_2, Y_3, Y_4, Y_5 \]

(b) Multivariate Two Patterns

\[ Y_1, Y_2, Y_3, Y_4, Y_5 \]

(c) Monotone

\[ Y_1, Y_2, Y_3, Y_4, Y_5 \]

(d) General

\[ Y_1, Y_2, Y_3, Y_4, Y_5 \]

(e) File Matching

\[ Y_1, Y_2, Y_3 \]

(f) Factor Analysis

\[ X, Y \]

Figure 1.1. Examples of missing-data patterns. Rows correspond to observations, columns to variables.

EXAMPLE 1.2. Unit and Item Nonresponse in Surveys. Another common pattern is obtained when the single incomplete variable \( Y_K \) in Figure 1.1a is replaced by a set of variables \( Y_{J+1}, \ldots, Y_K \), all observed or missing on the same set of cases (see Figure 1.1b, where \( K = 5 \) and \( J = 2 \)). An example of this pattern is unit nonresponse in sample surveys, where a questionnaire is administered and a subset of sampled individuals do not complete the questionnaire because of noncontact, refusal, or some other reason. In that case the survey items are the incomplete variables, and the fully observed variables consist of survey design variables measured for respondents and nonrespondents, such as household location or characteristics measured in a listing operation prior to the survey. Common techniques for addressing unit nonresponse in surveys are discussed in Chapter 3.
1.2. MISSING-DATA PATTERNS

categories and the obesity variables are binary, the data can be displayed as counts in a contingency table. Table 1.3 displays the data in this form, with missingness of obesity treated as a third category of the variable, where O = obese, N = not obese, and M = missing. Thus the pattern MON denotes missing at the first round, obese at the second round, and not obese at the third round, and other patterns are defined similarly.

Woolson and Clarke analyze these data by fitting multinomial distributions over the $3^3 - 1 = 26$ response categories for each column in Table 1.3. That is, missingness is regarded as defining strata of the population. We suspect that for these data it

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NNN</td>
<td>90</td>
<td>150</td>
<td>152</td>
<td>119</td>
<td>101</td>
<td>75</td>
<td>154</td>
<td>148</td>
<td>129</td>
<td>91</td>
</tr>
<tr>
<td>NNO</td>
<td>9</td>
<td>15</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>8</td>
<td>14</td>
<td>6</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>NON</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>13</td>
<td>10</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>NOO</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>19</td>
<td>8</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>ONN</td>
<td>0</td>
<td>8</td>
<td>7</td>
<td>13</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>ONO</td>
<td>1</td>
<td>9</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>OON</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>OOO</td>
<td>8</td>
<td>20</td>
<td>25</td>
<td>16</td>
<td>15</td>
<td>8</td>
<td>21</td>
<td>27</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>NNH</td>
<td>16</td>
<td>38</td>
<td>48</td>
<td>42</td>
<td>82</td>
<td>20</td>
<td>25</td>
<td>36</td>
<td>36</td>
<td>83</td>
</tr>
<tr>
<td>NOM</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>ONM</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>OOM</td>
<td>0</td>
<td>11</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>4</td>
<td>17</td>
<td>13</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>NMN</td>
<td>9</td>
<td>16</td>
<td>13</td>
<td>14</td>
<td>6</td>
<td>7</td>
<td>16</td>
<td>8</td>
<td>31</td>
<td>5</td>
</tr>
<tr>
<td>NMO</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>OMN</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>OMO</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>MNN</td>
<td>129</td>
<td>42</td>
<td>36</td>
<td>18</td>
<td>13</td>
<td>109</td>
<td>47</td>
<td>39</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>MNO</td>
<td>18</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>22</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MON</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>MOO</td>
<td>13</td>
<td>13</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>NMN</td>
<td>32</td>
<td>45</td>
<td>59</td>
<td>82</td>
<td>95</td>
<td>23</td>
<td>47</td>
<td>53</td>
<td>58</td>
<td>89</td>
</tr>
<tr>
<td>OMM</td>
<td>5</td>
<td>7</td>
<td>17</td>
<td>24</td>
<td>23</td>
<td>5</td>
<td>7</td>
<td>16</td>
<td>37</td>
<td>32</td>
</tr>
<tr>
<td>MNN</td>
<td>33</td>
<td>33</td>
<td>31</td>
<td>23</td>
<td>34</td>
<td>27</td>
<td>23</td>
<td>25</td>
<td>21</td>
<td>43</td>
</tr>
<tr>
<td>MOM</td>
<td>11</td>
<td>4</td>
<td>9</td>
<td>6</td>
<td>12</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>MMN</td>
<td>70</td>
<td>55</td>
<td>40</td>
<td>37</td>
<td>15</td>
<td>65</td>
<td>39</td>
<td>23</td>
<td>23</td>
<td>14</td>
</tr>
<tr>
<td>MMO</td>
<td>24</td>
<td>14</td>
<td>9</td>
<td>14</td>
<td>3</td>
<td>19</td>
<td>13</td>
<td>8</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>


$^a$NNN indicates not obese in 1977, 1979, and 1981; O indicates obese, and M indicates missing in a given year.