Strategies and Guidelines for Handling Missing Data in Social Work Research

Annual Conference of the Society for Social Work and Research
January 2020
Christopher J. Wretman, PhD ¹
wretman@email.unc.edu

Roderick A. Rose, PhD ²
rrose@ssw.umaryland.edu

Jeongsuk Kim, PhD ³
jeongsuk@email.unc.edu

¹ The Cecil G. Sheps Center for Health Services Research & School of Social Work, The University of North Carolina at Chapel Hill

² School of Social Work, The University of Maryland–Baltimore

³ School of Social Work, The University of North Carolina at Chapel Hill
I. What are missing data and why should I care?
Missing data = **blank values** on variables among observations.
<table>
<thead>
<tr>
<th>a2_2</th>
<th>a2_3</th>
<th>a2_4</th>
<th>a2_5</th>
<th>a3_1</th>
<th>a3_2</th>
<th>a3_3</th>
<th>a3_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>.</td>
<td>.</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>.</td>
<td>1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.</td>
<td>1</td>
<td>1</td>
<td>.</td>
</tr>
</tbody>
</table>
Sources & Reasons

Non-response
Refusal
Skips
Data collection error
Data collection change
Data entry error
Late entry/participation
Attrition
Useful Questions

Observation-level v. variable-level?
Time invariant v. time varying?
Reasonable v. problematic?
Design-based v. other?
Sample v. sub-sample?
Patterns v. mechanisms?
Technical analysis of MD begin by examining the mechanism.
Substantive question: Why are certain data missing?

Let ...

\[ X = \text{observed data matrix} \]

\[ Y = \text{hypothetical non-missing data matrix} \]

\[ M = \text{MD matrix where the \((i,j)\) th element indicates whether \((i,j)\) th element of \(Y\) is either missing (1) or observed (0)} \]

Statistical question: What is the distribution of \(M\) given \(Y\)?
Mechanism Classifications

**Missing Completely at Random (MCAR):** \( p(Y = M | X, Y) = p(Y = M) \)

- Formal: The probability that \( Y \) is missing does not depend on \( X \) or \( Y \)
- Informal: “Unconditional/Independent” = no patterns = random sub-sample/mechanism
- Unlikely; Testable but not provable with Little’s test

**Missing at Random (MAR):** \( p(Y = M | X, Y) = p(Y = M | X) \)

- Formal: The probability that \( Y \) is missing depends on \( X \) but not \( Y \); The values of \( Y \) depend on \( X \) but not \( Y \)
- Informal: “Conditional/Systematic” = patterns based on observed data
- Commonly assumed; “testable” in a sense with bivariate tests ... but not really

**Missing Not at Random (NMAR):** \( p(Y = M | X, Y) = p(Y = M | X, Y) \)

- Formal: The probability that \( Y \) is missing depends on \( Y \) (and \( X \)); The values of \( Y \) depend on \( Y \) (and \( X \))
- Informal: “Conditional/Systematic” = patterns base on unobserved data
- Non-ignorable; Worst case scenario ...
What might be a scenario in which we have missing data in our sample but we can reasonably assume the mechanism for missingness is MAR?

What about when we might be forced to assume mechanism for missingness is NMAR?
No matter what ... missing data weaken evidence.
Missing Data Consequences

Decreases to ...

Statistical power: reduced sample size

Internal validity: increased variance + representativeness questions vis-a-vis analytic sample v. sample

External validity: representativeness questions vis-a-vis sample v. population

Increases to ...

Bias & doubt

Heart palpitations, sweating, nausea, anxiety, sleeplessness, shakiness, loss of purpose in one’s personal & professional life
Decreased evidence quality hurts our ability to affect change.
Statistically, the presence of MD threatens assumptions about the completeness of data, potentially jeopardizing statistical power and validity. Our sample is both smaller and perhaps not representative. Substantively, the presence of MD is particularly important given that (a) missing observations are frequently participants in a study who are of interest and (b) analyses that do not consider the implications of missing data can produce misleading results.
One way to understand MD is through data simulations ...
Data: School Success Profile

Sample:

2011–2015; 8 states: NC, CA, SC, IL, etc.

\(N = 9,536\) middle & high school school students

Mean age = 13.0 years; 47% White, 27% African American, 20% Hispanic

Analytic sample = random 1,000 observations

Dependent variables:

“h_total”: 5 health/well-being items (e.g., “I feel good about myself“); 5–15

“f_total”: 6 family items (e.g., “Encourage you to do well in school“); 6–18

“s_total”: 7 school items (e.g., “This is a very good school to attend“); 7–28
Simulations

Software = Stata

#1: Drop n random observations

#2: Drop n random observations k times

#3: Drop n random observations k times & generate p values based on independent t-tests

#4: Drop n random observations k times & generate p values based on independent t-tests & graph
Simulation #1 Syntax

program md1
quietly {
    set seed `1'
    summarize `2', detail
    local count_1 = r(N)
    local count_2 = round((r(N))*`3')
    local mean_1 = round(r(mean), .001)
    local sd_1 = round(r(sd), .001)
    randomtag, count(`count_2') generate(temp)
    summarize `2' if temp==1, detail
    local mean_2 = round(r(mean), .001)
    local sd_2 = round(r(sd), .001)
    local diff = round(abs(`mean_1'-'mean_2'), .001)
    drop temp
}
display "`2' Full: n=`count_1'; m=`mean_1'; sd=`sd_1''
display "`2' Reduced: n=`count_2'; m=`mean_2'; sd=`sd_2''
display "`2' Difference: `diff''
end
Simulation #1 Results: \( n = 90\%, 75\%, 50\%; \ k = 1 \)

```plaintext
foreach a in 90 75 50  {
    md1 11979 h_total .`a'
    md1 11979 f_total .`a'
    md1 11979 s_total .`a'
}
```

<table>
<thead>
<tr>
<th>( N )</th>
<th>( h_{\text{total}} ) ( M ) (SD)</th>
<th>( f_{\text{total}} ) ( M ) (SD)</th>
<th>( s_{\text{total}} ) ( M ) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>13.02 (2.53)</td>
<td>13.99 (3.18)</td>
<td>22.15 (3.86)</td>
</tr>
<tr>
<td>900</td>
<td>13.07 (2.50) +.05</td>
<td>14.05 (3.15) +.07</td>
<td>22.24 (3.86) +.09</td>
</tr>
<tr>
<td>750</td>
<td>13.09 (2.47) +.07</td>
<td>14.15 (3.14) +.16</td>
<td>22.34 (3.88) +.19</td>
</tr>
<tr>
<td>500</td>
<td>13.04 (2.47) +.01</td>
<td>14.16 (3.14) +.18</td>
<td>22.35 (3.96) +.20</td>
</tr>
</tbody>
</table>
Simulation #2 Syntax

program md2
quietly {
    summarize `2', detail
    local count_1 = r(N)
    local count_2 = round((r(N))*`3')
    local mean_1 = round(r(mean), .001)
    local sd_1 = round(r(sd), .001)
    forvalues i = 1/`4' {
        set seed `1'`i'
        randomtag, count(`count_2') generate(temp)
        summarize `2' if temp==1, detail
        generate mean_`1'`i' = round(r(mean), .001)
        generate sd_`1'`i' = round(r(sd), .001)
        drop temp
    }
    egen mean_2 = rowmin(mean_*)
    egen mean_3 = rowmax(mean_*)
    egen sd_2 = rowmin(sd_*)
    egen sd_3 = rowmax(sd_*)
    foreach v of varlist mean_2 mean_3 sd_2 sd_3 {
        summarize `v', detail
        local `v' = round(r(mean), .001)
    }
    local diff_1 = round((`mean_1' -`mean_2'), .001)
    local diff_2 = round((`mean_1' -`mean_3'), .001)
    drop mean_* sd_*
}
display "`2' Full: n=`count_1'; m=`mean_1'; sd=`sd_1''
display "`2' Reduced: n=`count_2'; m_min=`mean_2', m_max=`mean_3'; sd_min=`sd_2', sd_max=`sd_3''
display "`2' Difference: diff_min=`diff_1''; " "diff_max=`diff_2''
end
Simulation #2 Results: $n = 90\%, 75\%, 50\%; k = 1000$

```plaintext
foreach a in 90 75 50 {
    md2 11979 h_total .`a' 1000
    md2 11979 f_total .`a' 1000
    md2 11979 s_total .`a' 1000
}
```

<table>
<thead>
<tr>
<th>$N$</th>
<th>$h_{total}$</th>
<th>$f_{total}$</th>
<th>$s_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$ (SD)</td>
<td>$M$ (SD)</td>
<td>$M$ (SD)</td>
</tr>
<tr>
<td>1000</td>
<td>13.02 (2.53)</td>
<td>13.99 (3.18)</td>
<td>22.15 (3.86)</td>
</tr>
<tr>
<td></td>
<td>–.09, +.10</td>
<td>–.09, +.14</td>
<td>–.12, +.18</td>
</tr>
<tr>
<td></td>
<td>–.18, +.19</td>
<td>–.17, +.16</td>
<td>–.20, +.25</td>
</tr>
<tr>
<td></td>
<td>–.29, +.24</td>
<td>–.31, +.31</td>
<td>–.43, .46</td>
</tr>
</tbody>
</table>
program md3
quietly {
    ttest `2', by(`3')
    local p_1 = round(`r(p)', .00001)
    summarize `2', detail
    local count = round((r(N))*`4')
    forvalues i = 1/`5' {
        set seed `1'`i'
        randomtag, count(`count') generate(temp)
        ttest `2' if temp==1, by(`3')
        generate p_`1'`i' = round(`r(p)', .00001)
        drop temp
    }
    egen p_2 = rowmin(p_*)
    egen p_3 = rowmax(p_*)
    foreach v of varlist p_2 p_3 {
        summarize `v', detail
        local `v' = round(r(mean), .00001)
    }
    drop p_*
}
display "`2' Full: p=`p_1'"
display "`2' Reduced: p_min=`p_2'"; " "p_max=`p_3'
end
Simulation #3 Results: \( n = 90\%, 75\%, 50\%; \ k = 1000; \) \( t \)-test by gender

```plaintext
foreach a in 90 75 50  {
    md2 11979 h_total gender .`a' 1000
    md2 11979 f_total gender .`a' 1000
    md2 11979 s_total gender .`a' 1000
}
```

<table>
<thead>
<tr>
<th>( N )</th>
<th>( h_{\text{total}} \times \text{gender} ) ( p )</th>
<th>( f_{\text{total}} \times \text{gender} ) ( p )</th>
<th>( s_{\text{total}} \times \text{gender} ) ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.015</td>
<td>0.092</td>
<td>0.183</td>
</tr>
<tr>
<td>900</td>
<td>0.0003–.203</td>
<td>0.009–.586</td>
<td>0.022–.823</td>
</tr>
<tr>
<td>750</td>
<td>0.0002–.703</td>
<td>0.003–.972</td>
<td>0.006–.995</td>
</tr>
<tr>
<td>500</td>
<td>0.0001–.999</td>
<td>0.0004–.997</td>
<td>0.002–.999</td>
</tr>
</tbody>
</table>
Simulation #4 Syntax

program md4
quietly {
    use "2015 SSP (Clean).dta", clear
    ttest `2', by(`3')
    generate p = round(`r(p)', .00001)
    summarize `2', detail
    local count = round((r(N))*`4')
    forvalues i = 1/`5' {
        set seed `1'`i'
        randomtag, count(`count') generate(temp)
        ttest `2' if temp==1, by(`3')
        generate p_`1'`i' = round(`r(p)', .00001)
        drop temp
    }
    keep in 1
    reshape long p_, i(id_part) j(temp)
    summarize p_
    local m = r(max)
    histogram p_, freq bcolor(navy) width(.0005) xlabel(.000(.01)`m')
}
end
Simulation #4 Results: $n = 90\%; \ k = 1000; \ t$-test by gender

$md4 \ 11979 \ h_{total} \ gender \ .90 \ 1000$
Ultimately ... you don’t 100% know the mechanism/probability.
<table>
<thead>
<tr>
<th></th>
<th>a2_2</th>
<th>a2_3</th>
<th>a2_4</th>
<th>a2_5</th>
<th>a3_1</th>
<th>a3_2</th>
<th>a3_3</th>
<th>a3_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What are some key takeaways from these simple though illustrative simulations regarding the potential effects of missing data?
II. Coping Strategies
Your first priority should always be initial minimization.
Starting Off Well ...

Articulate need & impact = “want to do it”

Incentivize = “start to do it”

Design = “will complete it” ...

Parsimonious

Relevant

Appropriate

Fair
Describe the MD

Summarize missingness

Explore patterns

Test for MCAR

“Test” for MAR
### Stata: `misschk`

```
. misschk
Variables examined for missing values

<table>
<thead>
<tr>
<th>#</th>
<th>Variable</th>
<th># Missing</th>
<th>% Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>groupid</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>a1</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>a2</td>
<td>22</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>a3</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>a4</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>a5</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>a6</td>
<td>4655</td>
<td>48.8</td>
</tr>
<tr>
<td>8</td>
<td>a7</td>
<td>31</td>
<td>0.3</td>
</tr>
<tr>
<td>9</td>
<td>a8</td>
<td>54</td>
<td>0.6</td>
</tr>
<tr>
<td>10</td>
<td>a9</td>
<td>227</td>
<td>2.4</td>
</tr>
</tbody>
</table>
```

### Stata: `mvpatterns`

```
. mvpatterns
variables with no mv's: groupid a1 a3 a4 a5

<table>
<thead>
<tr>
<th>Variable</th>
<th>type</th>
<th>obs</th>
<th>mv</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>groupstate</td>
<td>str2</td>
<td>9307</td>
<td>229</td>
<td>AGE IN YEARS</td>
</tr>
<tr>
<td>a2</td>
<td>byte</td>
<td>9514</td>
<td>22</td>
<td>FREE LUNCH</td>
</tr>
<tr>
<td>a6</td>
<td>byte</td>
<td>4881</td>
<td>4655</td>
<td>FAMILY CONSTELLATION</td>
</tr>
<tr>
<td>a7</td>
<td>byte</td>
<td>9505</td>
<td>31</td>
<td>LANGUAGE SPOKEN AT HOME</td>
</tr>
<tr>
<td>a8</td>
<td>byte</td>
<td>9482</td>
<td>54</td>
<td>ADULTS IN ARMED FORCES</td>
</tr>
</tbody>
</table>

Patterns of missing values

```

```
<table>
<thead>
<tr>
<th>_pattern _mv _freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>++.++. 1 4561</td>
</tr>
<tr>
<td>+++++. 0 4455</td>
</tr>
<tr>
<td>++++. 1 218</td>
</tr>
<tr>
<td>+++. 1 126</td>
</tr>
<tr>
<td>++. 1 68</td>
</tr>
</tbody>
</table>

| +++. 1 20         |
| ++++ 2 20         |
| +++. 1 15         |
| ++. 1 15          |
| ++.+ 2 11         |

| .++. 2 5          |
| +++. 2 4         |
| .+.++ 2 4        |
| +.++. 3 3        |
| ++++. 2 2        |

| .++.++ 2 2       |
| .++. 2 2         |
| .+.++ 2 4       |
| +.++. 2 1        |
| ++++. 2 1       |
| +++.++ 2 1      |
```

Total | 9,536 | 100.00

```
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,673</td>
<td>49.00</td>
<td>49.00</td>
</tr>
<tr>
<td>1</td>
<td>4,746</td>
<td>49.77</td>
<td>98.77</td>
</tr>
<tr>
<td>2</td>
<td>110</td>
<td>1.15</td>
<td>99.93</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.05</td>
<td>99.98</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.02</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 9,536 | 100.00
```
Stata: “misstable sum, generate()”

```
. misstable sum, generate(x_)
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs=</th>
<th>Obs&gt;</th>
<th>Obs&lt;</th>
<th>Unique values</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>a2</td>
<td>22</td>
<td>9,514</td>
<td></td>
<td>12</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>a6</td>
<td>4,655</td>
<td>4,881</td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>a7</td>
<td>31</td>
<td>9,505</td>
<td></td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>a8</td>
<td>54</td>
<td>9,482</td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>a9</td>
<td>227</td>
<td>9,309</td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Stata: “logit …”

```
. logit x_a6 a1 a3 a4 a5 a7 a8 a9, nolog
```

Logistic regression

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>LR chi2(7)</th>
<th>Prob &gt; chi2</th>
<th>Pseudo R2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9255</td>
<td>872.95</td>
<td>0.0000</td>
<td>0.0680</td>
</tr>
</tbody>
</table>

Log likelihood = -5977.8107

| x_a6 | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|------|--------|-----------|-------|-----|---------------------|
| a1   | -.0377786 | .0436219 | -0.87 | 0.386 | -0.1232759 to -0.0477188 |
| a3   | -.1886021 | .0149654 | -12.60 | 0.000 | -.2179338 to -.1592704 |
| a4   | -1.124611 | .0920394 | -12.22 | 0.000 | -1.305005 to -.9442173 |
| a5   | .0792615 | .01601 | 4.95 | 0.000 | .0478824 to .1106406 |
| a7   | .4301455 | .041588 | 10.34 | 0.000 | .3486345 to .5116564 |
| a8   | -.6786111 | .0896586 | -7.57 | 0.000 | -0.8543387 to -.5028835 |
| a9   | -.9933882 | .091859 | -10.81 | 0.000 | -1.173428 to -.8133479 |
| _cons | 2.631229 | .1784085 | 14.75 | 0.000 | 2.281555 to 2.980903 |

Stata: “mcartest a1-a9”

Little's MCAR test

Number of obs = 9536
Chi-square distance = 1202.5550
Degrees of freedom = 113
Prob > chi-square = 0.0000

p > .05 = MCAR? 

p < .05 = Significant predictor of missingness = MAR?
Approaches

Missing Completely at Random (MCAR) & Missing at Random (MAR):

- Deletion
- Imputation
- Full information maximum likelihood

Missing Not at Random (NMAR):

- Pattern mixture models
- Heckman selection model
The first type of approach is simple but often unsatisfactory.
Deletion

Default setting in most software

Listwise deletion:

“complete case” analysis
Exclude observations missing any variable
Good: simple, easy, samples/results are comparable, unbiased if MCAR
Bad: reduces power, loses sample

Pairwise deletion:

“available case” analysis
Exclude observations missing any variable to the particular analysis
Good: Preserves sample, uses more information
Bad: samples/results not comparable
**Listwise:** User drops observations manually

**Pairwise:** Software drops observations automatically
Non-Response Weighting

Also discards/deletes data … but weights observed data “up” to the observed + non-observed (i.e., non-missing + missing)

Can get complicated … multiple variables, multiple imputations

Steps:

1. Build a model to predict non-response (i.e., 0,1; logistic regression)
2. Get predicted probabilities
3. Take the inverse = “inverse probability weights” = 1/1-p for non-missing observations
4. Apply weights in subsequent analyses to non-missing observations only to make the analytic sample (i.e., complete case) more representative of the entire sample
Approaches become more active when they use imputation.
Naive Imputation

Mean substitution: use the sample mean for each variable

Regression: model includes variables that predict missingness

Hot deck: matches MD observations with “similar” non-MD observations

Propensity score: multivariate hot deck

Notes:

Ad hoc

General strategy: Imputes 1 time in a way that equates to a random guess, but then does not allow for the uncertainty of that guess

Not typically recommended …

Still used sometimes & seen in older research
Less Naive Imputation

Ipsative mean imputation:

- Uses an observation/participant’s mean across a set of variables
- Less biased with more variables providing information
- Particularly useful with a “set” of variables that go together (e.g., a scale)
- Often involves setting a minimum # or % a priori

Last value carried forward:

- Also: “Next value carried backwards”
- Particularly useful/tempting for longitudinal data
Simple mean imputation:
Enter variable mean = “3.52”

Ipsative mean imputation:
Enter (2+3+2+2)/4 = “2.25”

Last value carried forward:
Enter “2”
Approaches become very active with multiple imputation.
Multiple Imputation: General Thoughts

Basic concept = Use a complicated Monte Carlo simulation to estimate multiple data sets that get combined into one (after lots of thinking & processing)

Inherently a model-based approach

Many, many options & details ... not for the faint of heart

Very software specific

Different approaches produce different results

Different models for different analyses = lots of care & time
Multiple Imputation: Guidelines

1. Model for data analysis must be included in the imputation model & include all relationships among variables

2. Variables must satisfy certain distributional assumptions

3. Data must support the model (i.e., the robustness or collinearity condition)

4. Results must be interpreted understanding that SEs are inflated
Multiple Imputation: Practicalities

General steps:

1. Do data cleaning, descriptive statistics, specify RQs, conduct diagnostic checks (e.g., VIF)
2. Tell the software you’re going to do MI & let it know what variables will be involved and what their relationships will be
3. Generate $D$ imputed datasets based off the specified parameters (typically 5-20)
4. Pool all of the estimates into a final single set of estimates
5. Pray you got it all correct …

Stata:

1. "mi set"
2. "mi register" + "mi describe"
3. "mi impute"
4. "mi estimate"
1. Impute
   Stata: “mi set” + “mi register” + “mi describe”

2. Analyze
   Stata: “mi impute”

3. Pool
   Stata: “mi estimate”
1. Fit the linear regression of $Y$ on $X$ using the complete cases:

$$Y = \alpha + \beta X + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

2. This gives estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$.

3. To create the $m$th imputed dataset:
   
   3.1 Draw new values $\alpha_m$, $\beta_m$ and $\sigma_m^2$ based on $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$.
   
   3.2 For each subject with observed $X_i$ but missing $Y_i$, create imputation $Y_{i(m)}$ by:

$$Y_{i(m)} = \alpha_m + \beta_m X_i + \epsilon_{i(m)}$$

where $\epsilon_{i(m)}$ is a random draw from $N(0, \sigma_m^2)$. 

Bartlett, 2012
<table>
<thead>
<tr>
<th>Subject</th>
<th>Data</th>
<th>Imputation 1</th>
<th>Imputation 2</th>
<th>Imputation 3</th>
<th>Imputation 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>1.1</td>
<td>3.4</td>
<td>1.1</td>
<td>3.4</td>
<td>1.1</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>3.9</td>
<td>1.5</td>
<td>3.9</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>2.3</td>
<td>2.6</td>
<td>2.3</td>
<td>2.6</td>
<td>2.3</td>
</tr>
<tr>
<td>4</td>
<td>3.6</td>
<td>1.9</td>
<td>3.6</td>
<td>1.9</td>
<td>3.6</td>
</tr>
<tr>
<td>5</td>
<td>0.8</td>
<td>2.2</td>
<td>0.8</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>6</td>
<td>3.6</td>
<td>3.3</td>
<td>3.6</td>
<td>3.3</td>
<td>3.6</td>
</tr>
<tr>
<td>7</td>
<td>3.8</td>
<td>1.7</td>
<td>3.8</td>
<td>1.7</td>
<td>3.8</td>
</tr>
<tr>
<td>8</td>
<td>?</td>
<td>0.8</td>
<td><strong>0.2</strong></td>
<td>0.8</td>
<td><strong>0.8</strong></td>
</tr>
<tr>
<td>9</td>
<td>?</td>
<td>2.0</td>
<td><strong>1.7</strong></td>
<td>2.0</td>
<td><strong>2.4</strong></td>
</tr>
<tr>
<td>10</td>
<td>?</td>
<td>3.2</td>
<td><strong>2.7</strong></td>
<td>3.2</td>
<td><strong>2.5</strong></td>
</tr>
</tbody>
</table>

Bartlett, 2012
Imputation is **user-determined estimation of values that are not there**. It inherently involves some level of subjective assumption. At its heart, imputation uses **existing and knowable information** among observations without missingness to provide estimates/ideas/guesses to edit observations with missingness. Otherwise known as **“filling in”** those blanks. Imputation has a long history, and has been thoroughly researched in a multitude of ways ... with somewhat **mixed results**.
An alternative model approach relies on maximum likelihood.
**Full Information Maximum Likelihood**

Identifies the set of parameter values that produces the highest log-likelihood (LL)

Estimate = value that is most likely to have resulted in the observed data

Directly estimates the parameters in the presence of MD

Enhances the analysis model $B = f(X, Y)$ by including the missing data model $R = g(X, Y, Z)$ such that … $B = f(X, Y; R) =$ adjusts likelihood function so each observation contributes what it can

**Thoughts:**

Consistently found to work quite well with large sample sizes

Good: Uses complete cases + incomplete cases to calculate LL; Unbiased estimates with MCAR/MAR

Bad: Standard errors biased downward
THE MISSING DATA MODEL

We can define a **missing data model**:

\[ R = g(X, Y, Z) \]

where these terms are as defined on the previous slide.

This is different from the **analysis model**:

\[ B = f(X, Y) \]
The total area of the circles is $g(X, Y, Z)$, the missing data model.
FIML will be better if the missing data model looks like this

In other words, the fewer “extra” variables, the better. There are risks to adding a bunch of variables to your model.
<table>
<thead>
<tr>
<th>a2_2</th>
<th>a2_3</th>
<th>a2_4</th>
<th>a2_5</th>
<th>a3_1</th>
<th>a3_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
FIML Practicalities

Simpler = relatively few “researcher degrees of freedom”

Most software defaults to FIML when only the DV has MD (e.g., SAS “proc mixed,” Stata “xtmixed”)

If covariates are missing ... special software/procedures are necessary

Stata: “sem” commands with “method(mlmv)” option

MPlus: default

Paul Allison: “better” than MI ... http://statisticalhorizons.com/ml-better-than-mi
III. Final Thoughts
Missing data matters both methodologically and substantively. And, for better or worse, its deleterious effect on your research matters differentially depending on the nature of your research questions, sample, analyses, implications, etc. It is all context dependent and typically not easily defined and addressed. Overall, dealing with MD is complex and frequently one of the most challenging components of an analysis. It’s hard.
The first and best approach is to be clear & transparent.

Michael S. Kelly, Andy Frey, Aaron Thompson, Heather Klemp, Michelle Alvarez, and Stephanie Cosner Berzin

The Second National School Social Work Survey in 2014 aimed to update knowledge of school social work practice by examining how practitioner characteristics, practice context, and practice choices have evolved since the last national survey in 2008. This second survey was also developed to assess how the new national school social work practice model created by the School Social Work Association of America aligns with early 21st century school social work practice realities. The second survey was conducted from February through April 2014 (3,769 total responses were collected) and represents the largest sample of American school social workers surveyed in two decades. Data from the Second National School Social Work Survey showed a field that still has not fully responded to calls to implement evidence-informed and data-driven practices. This article notes the need to better integrate pre- and postservice training in data-driven practices and provides recommendations for ways to overcome barriers that school social workers report facing.

KEY WORDS: evidence-based practice; national model; school social work; survey research

The appropriate role for school social workers has long been the subject of debate (Costin, 1969; Kelly, Berzin, et al., 2010; Meares, 1977). To clarify the roles and responsibilities of school social workers and address the need for a national model that guides practice, the School Social Work Association of America (SSWA) commissioned a survey in 2008. The first national survey of school social workers was conducted in 1997 by the National Association of School Social Workers (NASSW) and the National Association of Social Workers (NASW). The survey was conducted again in 2014 by the School Social Work Association of America (SSWA) to assess the evolution of school social work practice and the alignment of the national model with current realities. The results of this survey provide important insights into the current state of school social work practice and suggest areas for future research and development.
Analysis
Before describing the national sample of school social workers and conducting exploratory statistical tests to examine relationships between practitioner characteristics and practice behaviors reflected in the national model, we examined missing data patterns.

**Missing Data Analysis.** A total of 3,769 school social workers initiated the survey and 2,521 (67 percent) of those respondents provided responses to all questions, leaving 905 (23 percent) respondents with some missing data. Analysis of missing data patterns revealed no significant associations between response patterns and observed respondent characteristics (that is, gender, race, degree, years of practice, certification, and licensure). Furthermore, those with missing data did not differ from those with full data on demographics. Under these conditions, we assume the missingness in the data to meet the basic criteria for the designation of missing at random (Little, 1988; Little & Rubin, 1989) and all analyses and estimates reported here were generated using a traditional listwise deletion approach. Under this approach, each estimate generated from the data uses all responses available (Pigott, 2001). Although the listwise deletion approach does result in a decreased sample size, it has important advantages. First, it abides by the rule of parsimony and offers the simplest analyses facilitating clear understanding and communication of the findings (Kline, 2011); second, listwise deletion produces unbiased estimates under the “missing at random” classification (Baraldi & Enders, 2010).
The General MD Process

1. Understand what’s going on
2. Use what information you can/know
3. Do the best you can
4. Be transparent & clear
5. Beg forgiveness
6. Promise/Hope for better next time
MD Analysis Steps

1. Do the missing data matter?
2. What data are missing?
3. What mechanism/structure describes the missing data?
4. What method(s) should be used to handle the missing data?
5. How do we implement the method(s)?
6. Did it work?
Stata Resources


Comprehensive overview: https://www.stata.com/meeting/italy12/abstracts/materials/it12_bartlett.pdf

Course at Centre for Multilevel Modeling, University of Bristol: http://www.bristol.ac.uk/cmm/learning/online-course/course-topics.html#m14

Multiple imputation: https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/

Multiple imputation: https://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm
R Resources


Comprehensive overview: https://cran.r-project.org/web/packages/finalfit/vignettes/missing.html

Brief overview: http://r-statistics.co/Missing-Value-Treatment-With-R.html

Focus on graphing: https://rstudio-pubs-static.s3.amazonaws.com/4625_fa990d611f024ea69e7e2b10dd228fe7.html
Classic MD Literature


Social Work MD Literature


Thank you! Email me for resources: wretman@email.unc.edu
Thoughts? Questions?