SEM 1: Confirmatory Factor Analysis
Week 4 - Missing data

Sacha Epskamp

2020
Why are data missing? In a general $X$ predicts $Y$ case:

- **Missing completely at random (MCAR)**
  - Missingness is independent of $Y$ or $X$
  - Everything is fine!

- **Missing at random (MAR)**
  - Missingness is independent of $Y$, but not of $X$
  - Example: Men less willing to respond to mental health questionnaire
  - Not a big problem

- **Missing not at random (MNAR)**
  - Missingness depends on $Y$
  - Example: People with severe mental health problems fill in less questionnaires
  - This is bad :( 

Unfortunately, there is no way to know exactly how your data is missing.
Missing completely at random (MCAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Regression based on observed data

Regression based on all data
Missing at random (MAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Regression based on observed data
Regression based on all data
Missing not at random (MNAR)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
</tbody>
</table>

Regression based on observed data
Regression based on all data
“Older” methods of handling missing data:

- **Compute SSS using listwise deletion**
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- **Compute SSS using pairwise estimation**
  - Assumes MAR
  - Estimate each element of SSS using all available data
  - Downside: Each covariance is based on different $n$ and variance–covariance matrix might not be positive definite

- **(Multiple) inputations**
  - Assumes MAR
  - Inpute missingness using mean scores or regression models
  - Downside: complicated, can increase bias if MNAR
“Older” methods of handling missing data:

- Compute $S$ using listwise deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- (Multiple) inputations
  - Assumes MAR
  - Inpute missingness using mean scores or regression models
  - Downside: can increase bias if MNAR
“Older” methods of handling missing data:

- Compute $S$ using **listwise** deletion
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- Compute $S$ using **pairwise** estimation
  - Assumes MAR
  - Estimate each element of $S$ using all available data
  - Downside: Each covariance is based on different $n$ and variance–covariance matrix might not be positive definite
“Older” methods of handling missing data:

- **Compute $S$ using listwise deletion**
  - Assumes MCAR
  - Delete all rows with a missing value
  - Downside: deletes observed data

- **Compute $S$ using pairwise estimation**
  - Assumes MAR
  - Estimate each element of $S$ using all available data
  - Downside: Each covariance is based on different $n$ and variance–covariance matrix might not be positive definite

- **(Multiple) inputations**
  - Assumes MAR
  - Inpute missingness using mean scores or regression models
  - Downside: complicated, can increase bias if MNAR
Full-information maximum likelihood (FIML):

- Compute likelihood for each person or each data subset with the same missingness pattern
- Assumes MAR
- Uses the full data set and all observations
- Downside: full data needed (analysis can not be done using covariance matrix)
- Implemented in most software (e.g., lavaan, Mplus, psychonetrics)
  - Use missing = "FIML" in lavaan or estimator = "FIML" in psychonetrics
FIML estimator in *psychonetrics* for every subset of data \( i \) with same missingness pattern:

\[
F_{\text{FIML}} = \frac{1}{n} \sum_i n_i \left( \text{trace} \left( S_i \Sigma_i^{-1} \right) + (\bar{y}_i - \mu_i)^\top \Sigma_i^{-1} (\bar{y}_i - \mu_i) - \ln |\Sigma_i^{-1}| \right)
\]

- \( n_i \): sample size of subset \( i \)
- \( S_i \): sample covariances (ML) of subset \( i \) (note, \( S_i = O \) if \( n_i = 1 \))
- \( \bar{y}_i \): sample means of of subset \( i \) (note, same as the observed score if \( n_i = 1 \))
- \( \Sigma_i \): Subset of \( \Sigma \) containing only elements of observed data in subset \( i \)
- \( \mu_i \): Subset of \( \mu \) containing only elements of observed data in subset \( i \)

Downside: saturated model needs to be computed as well.
Assumptions of maximum likelihood estimation

1. Independence: Observations are a simple random sample from some population
   ▶ Consequence: underestimated standard errors, inflated Type-I error rates
   ▶ Solution 1: use SE correction for dependence structure
   ▶ Solution 2: multilevel SEM

2. Multivariate Normality: Variables are univariate normally distributed at levels of all other variables, residuals are normal and homoscedastic, latent variables are normal, bivariate relations are linear
   ▶ Consequence: standard errors are incorrect (probably too low), \( \chi^2 \) test value is not accurate (probably too high)
   ▶ Solution 1: use “robust” standard errors (estimator = 'MLM', with complete data; estimator = 'MLR' with incomplete data)
   ▶ Solution 2: use bootstrapped standard errors & test statistic