A Brief Introduction to Modern Missing Data Analysis

Kyle M. Lang

Quantitative Psychology Training Program
University of Kansas
Topics to be Covered

- An introduction to the missing data problem
- Missing Data Mechanisms
- Ad hoc missing data techniques
- Modern missing data analysis
  - EM
  - MI
  - FIML
What is Missing Data?

- Unobserved data with true population values
  - Nonresponse on sample surveys
  - Attrition in longitudinal studies
  - Un-administered materials in planned missingness designs
  - Data lost unintentionally

- Note: Not all unobserved data are truly missing
The Causes of Missingness

- Missing data are missing for a reason
  - Missing Data Mechanism

- Missing Completely at Random (MCAR)

- Missing at Random (MAR)

- Missing Not at Random (MNAR)
Missing Completely at Random

- MCAR missing can be viewed a random sample of the complete data.
- There are no correlates of the missing data, either measured or unmeasured
- A survey is administered to an undergraduate biology class
  - At time 2 some respondents from time 1 do not complete the survey because they are home sick
- Almost entirely recoverable
Missing at Random

- Missingness is correlated with a measured variable other than that which contains the missingness.
- The missingness can have a systematic cause, but the cause has been captured in the dataset.
- Highly religious subjects fail to respond to items assessing attitudes towards sex. If you have a measure of religiosity in the data, this is MAR missing.
- Very well recovered with appropriate techniques.
Missing Not at Random

- The missingness is correlated with the variable containing the missingness or an unmeasured variable.

- Subjects with very high or very low household income fail to respond to an SES item
  - Participant level on an item is designating their response rates

- If there is no measure of religiosity in the previous example, that missingness is MNAR

- Very poorly recovered even with appropriate techniques
Ignorability

- Almost all missing data techniques require the missingness mechanism be ignorable
  - Ignorability

- To safely ignore the missing data mechanism, Rubin (1976) showed that
  - 1) A MAR mechanism must give rise to the loss of information
  - 2) The parameters of substantive interest ($\theta$) and those which contribute to the missingness ($\xi$) must be distinct
Ignorability II

- Specifically, for the first tenet to hold:

\[ P(R | Y_{obs}, Y_{miss}, \xi) = P(R | Y_{obs}, \xi), \]

- The probability of missingness conditioned on both \( Y_{obs} \) and \( Y_{miss} \) must be no more informative than that conditioned on only \( Y_{obs} \)
Ignorability III

- For the second tenet to hold:
  - **Frequentist:** The joint parameter space \((\theta,\xi)\) must be the Cartesian cross-product of the individual parameter spaces \(\theta\) and \(\xi\).
  
  - **Bayesian:** Any possible joint prior applied to \((\theta,\xi)\) must factor into unique marginal distributions of \(\theta\) and \(\xi\).
Ad Hoc Techniques

- Listwise & Casewise Deletion
- Mean Substitution
- Regression Imputation / Conditional Mean Substitution
Listwise & Casewise Deletion

- Both entail deleting unobserved information
  - Listwise deletion removes all cases with at least one unobserved value
  - Casewise deletion removes pairs of values from the calculation of sufficient statistics

- Both can entail a substantial loss of sample size
  - 100 X 100 Matrix, 20 missing values
    - .2% of data missing
    - 20% of data deleted

- In the case of a nonrandom missingness mechanism, the truncated dataset will cease to be generalizable to the larger population.
Mean Substitution

- Every missing value is replaced with the mean of its scale
- This can lead to a substantial loss of variability
- Causes regression towards the mean
  - Can attenuate the true population relationships between variables
Regression Imputation

- The variable with missingness is regressed on all other variables, and the predicted values are used as replacements for the missing values
  - Leads to artificially precise standard errors
  - Amplifies the conformity of the missing value with the linear trend
    - This can artificially inflate the correlations between variables
General Problem with Single Imputation Techniques

- Single imputation techniques assume that the imputed value is an unbiased estimate of the true population value

  - Leads to artificially precise standard errors

  - Biases Wald tests and Confidence Intervals
    - Coverage under 95% for an alpha of .05
    - Increased Type I error rate
Modern Missing Data Analysis

- Expectation Maximization
  - Dempster, Laird, & Rubin (1977)

- Multiple Imputation
  - Rubin (1976, 1987)

- Full Information Maximum Likelihood
  - A chance for audience participation!
Expectation Maximization (EM)

- Maximum likelihood approach to estimating a complete covariance matrix in the presence of missing data
- Calculates the expected values of sufficient statistics (E-Step)
- Maximizes the conditional loglikelihood derived from the above sufficient statistics (M-Step)
- Converges to a stationary covariance matrix when the change in the loglikelihood between two subsequent M-Steps falls below a specified convergence criterion
Factored complete data PDF

- Assuming the ignorability of the missing data mechanism:

\[ P(Y \mid \theta) = P(Y_{obs} \mid \theta)P(Y_{miss} \mid Y_{obs}, \theta) \]

- EM replaces \( Y_{miss} \) with the expected values of \( Y_{miss} \mid Y_{obs}, \theta \)
Expectation (E) Step

- Separate regression equations are calculated for every pattern of missingness.

- The predicted values from these equations are used to fill in the missing values.

- This “complete” data frame is used to calculate updated estimates of the sufficient statistics (i.e., sums of squares and sums of cross products).
Maximization (M) Step

- The sufficient statistics produced by the E-Step are used to compute the complete-data loglikelihood.

- This loglikelihood is maximized as a complete data problem to reach new estimates of $\mu$ and $\Sigma$. 
EM Continued

○ The E-Step is calculating:

\[ Q(\theta \mid \theta^{(t)}) = \int l(\theta \mid Y) P(Y_{\text{miss}} \mid Y_{\text{obs}}, \theta^{(t)}) dY_{\text{miss}} \]

○ The M-Step maximizes the above such that:

\[ Q(\theta^{(t+1)} \mid \theta^{(t)}) \geq Q(\theta \mid \theta^{(t)}), \text{ for all } \theta \]
Multiple Imputation

- Bayesian simulation approach to imputation
- Developed to address many of the shortcomings of single imputation and early ML techniques
- Accounts for our uncertainty in the true population values of $Y_{\text{miss}}$
  - Does not lead to downwardly biased standard errors
MI Continued

- Through an iterative simulation technique, \( m > 1 \) plausible values of \( Y_{miss} \) are estimated:
  - Creates \( m \) complete datasets
  - The variability in the values of \( Y_{miss} \) between every imputation (i.e., between imputation variance) quantifies our uncertainty in the true population values of \( Y_{miss} \)

- \( m \) analysis models are run, and the resulting estimates are combined for statistical inference
Data Augmentation

- Data Augmentation (Tanner & Wong, 1987) is a two step iterative procedure to create multiple imputations.

- I-Step uses a series of stochastic regression equations to fill in the missing values.

- P-Step pulls random draws from probability distributions derived from the complete data from the I-Step to introduce variability into the estimates of the parameters.
Priors for MI

- Prior distributions must be specified to inform initial estimates of $\theta$

- Informative Priors
  - Stationary EM covariance matrix
  - Sufficient statistics from a single regression imputation

- Uninformative Priors
  - $\mu$ is drawn from a standard normal distribution
  - $\Sigma$ is drawn from a Wishart distribution
Imputation (I) Step

- Initial estimates of $\mu$ and $\Sigma$ are used to calculate separate stochastic regression equations for each pattern of missingness.

- The predicted values from the above equations are used to replace $Y_{\text{miss}}$. 
Posterior (P) Step

- The complete data frame from the I-Step is used to parameterize a normal and $\chi^2$ distribution.

- New estimates of $\mu$ and $\Sigma$ are estimated based on random draws from the above distributions.
P-Step Continued

- New estimates of $\sigma^2$ are drawn based on the following equality:

$$s_Y^{2*} \sim \frac{\nu s_Y^2}{\chi^2_v}$$

- New estimates of $\mu$ are drawn as follows:

$$\bar{Y}_Y^* \sim N(\bar{Y}_Y, s_Y^{2*})$$
Markov Chains

- The estimates from each subsequent P-Step create a Markov Chain
  - $\theta_{t+1}$ is dependent on $\theta_t$

- After a sufficient number of iterations the elements in the chain become independent
  - $\theta_{t+p}$ is independent of $\theta_t$
Independent Draws

- To achieve valid imputations, a number of P-Steps must be discarded (burn-in iterations) before the first draw
  - Insures that initial imputed estimates are independent of the prior

- Between every imputation after the first, many more P-Steps must be discarded
  - Insures subsequent imputed values are independent
    - $m_t$ is independent of $m_{t+1}$

- A common recommendation for the number of P-Steps to discard is twice the number of iterations it takes the EM algorithm to converge
Combining Parameters in MI

- All $m$ sets of parameters must be combined into a single vector of statistics in order to reach an inferential conclusion.

- Rubin (1987) developed simple formulae to combine all $m$ point estimates and variances into single statistics which can be used for Wald tests or CIs.
  - Rubin’s Rules
Rubin's Rules for Point Estimates

- The aggregate point estimate is simple the mean of all $m$ statistics:

$$\overline{Q} = m^{-1} \sum_{j=1}^{m} \hat{Q}^{(j)}$$
Rubin’s Rules for Variance Components

- The variability around $\bar{Q}$ has two components:
  - Within Imputation Variance:
    \[
    \bar{U} = m^{-1} \sum_{j=1}^{m} U^{(j)}
    \]
  - Between Imputation Variance
    \[
    B = (m - 1)^{-1} \sum_{j=1}^{m} [\hat{Q}^{(j)} - \bar{Q}]^2
    \]
The total variance of the MI estimates is a weighted sum of the between and within imputation variances:

\[ T = \bar{U} + (1 + m^{-1})B \]

- As \( m \) becomes larger the influence of \( B \) is weighted downward and \( T \) becomes smaller.
Assumptions of MI

- Three assumptions of MI
  - The probability distribution giving rise to the data is of a specified form
    - Usually the normal distribution
  - The missingness mechanism is ignorable
  - The prior distribution used to inform the initial estimates of $\theta$ is appropriate

*Note:* Assumptions 1 & 2 must also be made for ML techniques
Full Information Maximum Likelihood (FIML)

- FIML is a maximum likelihood estimator which is robust to the presence of missing data.

- Calculates an individual loglikelihood for each case in the data frame:
  - Allows separate estimation of parameters for every pattern of missingness without the need to address the true population values of $Y_{miss}$.
Multivariate Normal PDF

- The multivariate normal PDF from which FIML derives the joint loglikelihood it maximizes is characterized by the following equation:

\[ f(y) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-\frac{(y-\mu)'\Sigma^{-1}(y-\mu)}{2}} \]
Casewise Loglikelihood

- For each case in the data, the following loglikelihood is computed:

\[ \ell_i = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (y_i - \mu)' \Sigma^{-1} (y_i - \mu) \]

- Which becomes the following in the presence of missing data:

\[ \ell_i = -\frac{p_i}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_i| - \frac{1}{2} (y_i - \mu_i)' \Sigma_i^{-1} (y_i - \mu_i). \]
Joint Loglikelihood

- Each of the previous casewise loglikelihoods is summed to get:

\[
\ell = \sum \left[ -\frac{p_i}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_i| - \frac{1}{2} (y_i - \mu_i)'\Sigma_i^{-1}(y_i - \mu_i) \right].
\]

- FIML then maximizes the above equation to reach estimates of \( \mu \) and \( \Sigma \).
FIML Continued

- Because an individual loglikelihood is calculated for every case, the final joint loglikelihood is based only on $Y_{obs}$

  - Allows the estimation of parameters based on only observed information

  - No need to address the true population values of $Y_{miss}$
Take-Home Message

- Missing data cannot simply be ignored
- Ad hoc techniques lead to substantial bias in all but the most restrictive of cases
- Modern missing data techniques recover information very well provided conformity to a few loose assumptions
- All of the techniques described today are very easily implemented in nearly all common statistical packages
- More precisely and efficiently recovered information leads to more valid estimates of parameters, higher statistical power, more valid statistical inferences and better science
Special Thanks

- Wei Wu
- Todd Little
- Kris Preacher
- Rawni Anderson
- Waylon Howard
- John Geldhof