How to Treat Missing Data

Tuesday, September 6, 2011
Speaker: Dr. Miyong Kim
Problems - Missing data

- Biased results
- Decreased statistical power
- Complicated analyses
Prevention!!
Instrument phase:
- Avoid ambiguous questions
- Clear wording
- Special attention to sensitive topics
- Instrument should be no longer than absolutely necessary
Collecting data Phase

- Unit non response rates - call backs or follow up
- Population study
  1) Subsample substitution
  2) A randomized response technique
- Use of proxy respondent
- Incentive
- Oversampling
- Use of the Auspices
- Training and supervision of research staff
Patterns of Missingness

- Missing at random
- Missing with bias
- Missing on specific indicators
- Missing on common factors (item vs. factor)
How do you know it is random?

- This is not an absolute science!
- Plotting the missing data
- Assess the magnitude (too many missing from one variable)
- Dummy coding
Data Deletion:
* Pairwise Vs. Listwise

Data Imputation:
* Between-Subjects “Univariate”
* Within-Subjects “Multivariate”
Pairwise Deletion

* Each bivariate correlation estimated on all data available for each successive pair of study variables

* Random missingness and large samples may produce good estimates of population correlation matrix

* However, biased missingness and small samples may produce ill-conditioned sample correlation matrix
Listwise Deletion:
( when the magnitude of missing data is not large )

* Random missingness and large samples may produce good estimates of population correlation matrix
* Systematic biases in data deleted may produce systematic biases in data retained
Detection of Biases following Listwise Deletion

- **Univariate Approach**
  - Construct dummy variable to code for missingness of each variable with missing data
  - Using pairwise deletion, test correlation of each dummy variable with every other study variable (Caution: This may grossly inflate alpha!)
  - Test correlation of each dummy variable with all other dummy variables coding for missing data
  - Factor analysis of dummy variables may reveal latent structure of missing data (Richard Gorsuch, personal communication)
  - However, biases on different variables with missing data may be quite heterogeneous
Detection of Biases following Listwise Deletion

- **Multivariate Approach**
  * Construct dummy variable to code for two subsamples, "listwise deleted" and "listwise retained"
  * Using pairwise deletion, test correlation of single dummy variable with every other study variable
  * Evaluate effect sizes of single dummy variable on study variables as measures of bottom-line listwise deletion bias
  * Homogeneous biases across study variables may tend to compound
  * Heterogeneous biases of different variables may tend to cancel out ("heterogeneous irrelevancies")
When you need to save N for your analyses!

*“Cold deck imputation”
*“Hot deck imputation”
BETWEEN-SUBJECTS "UNIVARIATE" DATA IMPUTATION

* Missing values on each variable are imputed for each subject based on observed values for other subjects on same variable

* Substantiate that subjects missing and not missing data are somehow similar (e.g., Matched Subjects)

* Assumption may have little justification if missing data is systematically biased, but can be evaluated by dummy variable method
WITHIN-SUBJECTS DATA "MULTIVARIATE" IMPUTATION

* Missing values on each variable are imputed for each subject based on observed values for same subject on other variables
* Substantiate that variables missing and not missing data are somehow similar (e.g., Parallel Measures)
* Assumption can be supported by results of multivariate analysis
The Estimation of Factor Scores

DIFFERENTIALLY WEIGHTED" FACTOR SCORES:

* Standard errors for different factor loadings typically large
* Estimates are "unstable" or sample-specific
* Seldom possible to discriminate between similar loadings
* Worsens when the sample sizes are smaller
UNIT WEIGHTED" FACTOR SCORES:

* All significant indicators weighted equally (i.e., 1.0)
* Correlated 0.95 to "differentially weighted" factor scores
* More generalizable across independent samples
* Much easier to calculate (summation of Z-scores)
AN APPROXIMATION TO CONFIRMATORY FACTOR ANALYSIS FOR SMALL SAMPLES

* An Approximate solution:
  * Factor model must be theoretically specified \textit{a priori} (\textit{i.e.}, confirmatory)
  * Indicators must be expressed in a common metric (\textit{e.g.}, Z-scores)
  * Factor scores can be estimated as arithmetic means of indicator scores
  * Means represent linear transformations of summed "unit weighted" factor scores
  * Standardizing eliminates any superficial difference
  * Bivariate correlations with indicator scores represent "factor structure" (\textit{lambda})
  * Bivariate correlations with each other represent "factor intercorrelation matrix" (\textit{phi})
  * Bivariate correlations with their own arithmetic mean represent "higher-order factor structure"
MULTIVARIATE IMPUTATION AND DETECTION OF BIASES

MULTIVARIATE IMPUTATION OF MISSING DATA AND DETECTION OF BIASES FOLLOWING MULTIVARIATE IMPUTATION

GENERALIZATION OF FACTOR STRUCTURES OBTAINED ACROSS LISTWISE-RETAINED AND LISTWISE-DELETED SUBSAMPLES:

* Using listwise deletion, construct and validate factor models (e.g., by CFA) for listwise-retained data
* Estimate unit-weighted factor scores as means of standardized (z) scores of significant indicators
* Compute unit-weighted factor scores for listwise-retained data using complete data
* Impute unit-weighted factor scores for listwise-deleted data using all nonmissing indicators
* Using pairwise deletion, construct factor structures for both subsamples by correlating unit-weighted factor scores to indicator variables
* Compare factor structure of listwise-deleted subsample to that of listwise-retained subsample
* Where factor structures cross-validate across subsamples, compute unit-weighted factor scores using multivariate imputation for combined sample
* Multivariate imputation will automatically "compute" factor scores for complete data and "impute" factor scores for incomplete data
MULTIVARIATE IMPUTATION WORKS WITH MISSING INDICATORS, BUT DOES NOT WORK WITH MISSING COMMON FACTORS

* All indicators of a common factor might be missing, a silence which truly betokens
* Univariate imputation of factor scores may be combined with dummy variable method in multivariate analysis
* Specify structural pathways from dummy variable coding for missing factor to all other study variables in the model
* Dummy variable pathways to other study variables may be indirect through non-missing common factors
* Alternatively, multi-sample analysis may be employed if number of missing factor subsamples is small (e.g., with "Structured" Missing Data)
Comparison of Missing Data Deletion and Missing Data Imputation Techniques:

* Pairwise deletion produces biased sample upon certain missing data conditions (e.g., missing factors)
* Listwise deletion performs surprisingly well under a wide array of missing data conditions
* All else being equal (*ceteris paribus*), listwise deletion appears to outperform multivariate imputation in most missing data conditions
* In reality, all else is seldom equal because listwise deletion typically produces much more missing data than multivariate imputation (e.g., with "Messy" Missing Data)
* Multivariate imputation produces conservative but robust estimates under most missing data conditions
* Multivariate imputation holds up well under varying degrees of missing data bias, but breaks down under conditions of very high measurement error (e.g., 70%)
* When both missing data bias and measurement error are very high, none of these techniques work very well
THE DUMMY VARIABLE METHOD (WHETHER UNIVARIATE OR MULTIVARIATE) IS NOT GENERALLY RECOMMENDED WHEN MISSING DATA EXCEEDS 20% OF THE TOTAL SAMPLE

* A large proportion of missing data is generally deemed to be a threat to the validity of most methods of data imputation
A LARGE PROPORTION OF DATA THAT IS MISSING AT RANDOM IS FUNCTIONALLY EQUIVALENT TO A "SMALL" (OR AT LEAST "SMALLER") SAMPLE SIZE

Where a large proportion of the data is missing at random, the only problem with data deletion methods is that the remaining sample may be of inadequate size.

Data that is missing at random does not compromise the representativeness of the original sample.
A LARGE PROPORTION OF DATA THAT IS MISSING AT RANDOM IS FUNCTIONALLY EQUIVALENT TO A "SMALL" (OR AT LEAST "SMALLER") SAMPLE SIZE

* Where a large proportion of the data is missing at random, the only problem with data deletion methods is that the remaining sample may be of inadequate size

* Data that is missing at random does not compromise the representativeness of the original sample

* What defines a "small" sample size?
* Please feel free to take this time to ask our speaker questions.
* Please don’t forget to fill out our short survey to better serve you for future P30 Methodological Lunch & Learns
* Our next Lunch & Learn will be:
  * **Tuesday, October 4, 2011**
  * **Overview of IRT and Innovative Measurements**
  Speaker: Dr. Kitty Chan (JHMI)

*Questions?