Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 6 (January 23, 2007): Introduction and Naïve Methods on anaysis of missing data

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Problem of missing data

Standard statistical methods have been developed to analyze rectangular data sets.

- The rows are observation units
- The columns are va variables.
- The entries are values (real numbers).

The concern is what happen when some of these values are not observed.

- Most statistical software creates special codes for the missing values
- Some statistical software exclude subjects with missing values - "complete-case analysis", which will be valid in a very limited cases.

Missing-data patterns

- Notations: For subject i, K variables, $Y_1, \ldots, Y_K$, are measured, and y_{ij} is the value of variable j, Y_j. Here Y_j should be either an independent or dependent variable.
- $Y = (y_{ij})$: an (nxK) rectangular data set without missing values - complete data.
- $M = (m_{ij})$: the missing-data indicator matrix such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present.

Some common missing-data patterns

- Univariate nonresponse where missingness is confined to a single variable, say Y_K.
- Unit and Item nonresponse in survey.
  - Unit nonresponse: a questionnaire is administered and a subset of subjects do not complete the questionnaire because of noncontact, refusal, or some other reason.
  - Item nonresponse: some subjects complete the questionnaire but do not answer all items in the questionnaire - a haphazard pattern.

Some common missing-data patterns, cont

- Attrition in longitudinal studies: some subjects drop out prior to the end of the study and do not return. The patter of attribution is an example of monotone missing data.

Assumption on missing data

Assumption 1: Missingness indicators hide true values that are meaningful for analysis.

Some Examples

- Example 1: No response in three binary outcomes measured for the same patient.
- Example 2: Effects of treatments on quality of life outcomes.
  - Consider a randomized trial with two treatments, T=0 or 1, and suppose that a primary outcome of the study is a "quality-of-life health score" Y measured one year after randomization to treatment.
  - For participants who die within a year of randomization to treatment, Y is undefined in some sense or "censored" due to death.
  - It does not make sense to treat those outcomes as missing as missing values as in Assumption above, given that quality of life is a meaningless concept for people who are not alive.

Some examples, cont

- Example 3. Nonresponse in Opinion Polls.
  - We are interested in polling individuals about how they will vote in a future referendum, where the available responses are "yes", "no", or "missing".
  - Individuals who fail to respond to the question may be refusing to reveal real answers or may have no interest in voting.
  - Assumption 1 would not apply to individuals who would not vote, and these individuals define a stratum of the population that is not relevant to the outcome of the referendum.
  - Assumption 1 would apply to individuals who do not respond to the initial poll but would vote in the referendum. For these individuals it would make sense to treat their responses on the referendum as missing.

Mental health study in 2005 Ph.D applied exam

- A multi-clinic observational study on a prospective cohort of primary care patients with clinical depression.
- The study evaluated depressive symptoms, mental and physical health for 966 clinically depressed persons from 6 large U.S. clinics.
- These persons responded to several questionnaires that measured their physical and mental health at baseline, 6 weeks, 3 months, and nine months after baseline.
- At 9 month after baseline each patient was interviewed by a psychiatrist who determined whether or the patient still suffered from clinical depression after nine months.

Missing data pattern on MH measure

- Unit non-response over time (X: observed, M=missing)

Baseline	6 weeks	3 months	nine months	Freq
X	X	X	X	759
X	X	X	M	92
X	X	M	X	27
X	X	M	M	27
X	M	X	X	22
X	M	X	M	9
X	M	M	X	3
X	M	M	M	2
M	X	X	X	14
M	X	X	M	7
M	X	M	M	2
M	M	X	X	2

Clinical Diagnosis of Alzheimer's Disease (AD)

- The National Alzheimer Coordinating Center (NACC) maintains a database of subjects from 30 Alzheimer disease (AD) centers.
- The data contain clinical assessments on the AD status of all subjects, and only some of them have the neuropathologic examination results when they died - gold standard.
- Some subjects with the clinical assessment results are missing their true disease status.
- If we only use subjects with the neuropathologic examination results, estimated sensitivity and specificity of the clinical diagnosis of AD would be at risk for verification bias.

Neuropathologic Diagnosis of Alzheimer's Disease (AD)

The three sets of criteria for neuropathologic diagnosis of Alzheimer's Disease (AD):

- Khatchaturian /ADRDA (Alzheimer's Disease and Related Diseases Association) - 1984 - based on number of senile plaques per microscopic field in certain areas of the brain (neocortex), requiring different levels of plaque density depending on age and presence of clinical symptoms.
- CERAD (Consortium to Establish a Residency for Alzheimer's Disease) - 1991 - based on semiquantitative plaque counts in 3 specific areas of the neocortex; standardized staining techniques; emphasized "neuritic" over "diffuse" plaques; plaque density required still varied with age and presence of clinical dementia symptoms.
- NIA/Reagan - 1997 - built on CERAD, but explicitly included frequency of neurofibrillary tangles in addition to neuritic plaques, both to be assessed semiquantitatively in samples from specified parts of the brain using standardized staining techniques.

Neuropathologic Diagnosis of Alzheimer's Disease (AD), Cont

- We are interested in assessing the agreement among three neuropathologic diagnoses of AD.
- We are interested in the comparison of the diagnostic accuracy of clinical assessments using the three different neuropathologic diagnoses of AD.

Reasons for missing Neuropathologic Diagnosis of AD

- Many patients died
  1. in years before any of the criteria had been developed,
  2. in the era when only Khatchaturian/ADRDA had been developed, or
  3. in the era when CERAD had replaced Khatchaturian/ADRDA but NIA/Reagan criteria had not yet been developed.
- Once newer criteria came into vogue, older ones tended to fall out of favor, so the older criteria are often missing on patients with later death years.
- In some instances, later criteria were more specific about areas of the brain from tissue had to be available, or about staining techniques, which may not have been possible to satisfy if the original brain had been destroyed after tissue samples were made. Thus it may not have been possible to apply newer criteria if the required samples or stains were unavailable.

An example study on Alzheimer’s Disease

- Education has been reported to influence the risk and clinical course of Alzheimer’s disease (AD), but the mechanisms remain unclear.
- More-educated persons may have better test-taking skills, "cognitive reserve" that delays appearance of cognitive decline, or slower rates of cognitive decline with advancing neuropathology.
- We are interested in associations among cognitive function, severity of AD neuropathology, and education in a large national sample.

Example, cont

- We studied 2,792 autopsied patients age 65+ years enrolled before July, 2002 at AD Centers throughout the U.S. who had taken the Mini-Mental State Examination (MMSE).
- Braak & Braak stage, neuritic plaque density, and CERAD and NIA/Reagan diagnostic classifications were used to measure AD neuropathologic severity.
- Outcome: MMSE
  Independent variables: education and neuropathologic severity, and age.
- Only 1,157 patients had the complete information on all variables.

Missing data mechanism

- Relationship of the missingness to the variables
- $Y = (y_{ij})$: an (nxK) rectangular data set without missing value.
- With ith row, $y_i = (y_{i1}, \ldots, y_{iK})$: y_{ij} is the value of the jth variable Y_j for subject i
- The missing-data indicator matrix, $M = (m_{ij})$, where $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present.
- $Y = (Y_{obs}, Y_{mis})$.

Missing data mechanism, cont

- The missing data mechanism is characterized by the conditional distribution of M given Y, $f(M \mid Y, \phi)$.
- MCAR: $f(M \mid Y, \phi) = f(M \mid \phi)$, the missingness does not depend on the data.
- MAR: $f(M \mid Y, \phi) = f(M \mid Y_{obs}, \phi)$, the missingness depends only on the observed values
- NMAR: $f(M \mid Y, \phi) = f(M \mid Y_{obs}, Y_{mis}, \phi)$, the missingness depends on missing values
- Under NMAR, one needs to model the missingness

Example 1: univariate nonresponse

- Suppose there are missing data on a particular variable Y and complete data on other variables X.
- The missing data on Y is MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the study.
- The MCAR assumption would be violated if people who are missing Y were younger, on average, than people who have values of Y.

Missing at Random (MAR)

- Missing data on Y are said to be MAR if the probability of missing data on Y does not depend on Y after controlling for other observed variables X.
- In general, data are not MAR if those individuals with missing values on Y tend to have lower (or higher) values on Y than those with data on Y, controlling for other observed variables.
- It is impossible to test whether the MAR condition is satisfied.

Ignorable and Nonignorable

- The missing data mechanism is said to be be ignorable if (a) the missing data are MAR and (b) the parameters that govern the missing data process are unrelated to the parameters to be estimated.
- If the missing data are not MAR, the missing data mechanism is said to be nonignorable.

Conventional methods for missing data

- Complete-case (listwise deletion) method: delete from sample any observations that have missing data on any variables and then apply conventional methods of analysis for complete data.
- Available case (pairwise deletion) method: It is well known, for example, that a linear regression can be estimated using only the sample means and covariance matrix. The idea of pairwise deletion is to compute each of these summary statistics using all the cases that are available.
- Dummy variable adjustment: Suppose that some data are missing on an independent variable X in a regression model. We create a dummy variable D that is equal to 1 if data are missing on X and equal to zero otherwise. We also create one additional variable X^* such that

$$X^* = \begin{cases} X \\ c \end{cases},$$

where c is any constant. We then fit a regression model on Y using X^*, D, and other independent variables.

Validity of complete-data method

- If missing data are MCAR, the complete-data method will produce valid results.
- If missing data are not MCAR, but only MAR, the complete-data method can produce biased estimates.
- The complete-data method is most robust to violations of MAR among independent variables in a regression analysis.
- If the probability of missing data on any of the independent variables does not depend on the values of the dependent variable, then regression estimates using the complete-data method will be valid.
  Proof: let R be an indicator for observed data ($R = 1$ if all variables are observed; otherwise, $R = 0$). We are interested in estimating $f(Y \mid X)$, the conditional probability of Y given X. Note that

$$f(Y \mid X, R = 1) = \frac{f(Y, X, R = 1)}{f(X, R = 1)} = \frac{f(R = 1 \mid Y, X) f(Y \mid X) f(X)}{f(R = 1 \mid Y) f(X)}.$$

  If $f(R = 1 \mid Y, X) = f(R = 1 \mid X)$, $f(Y \mid X, R = 1) = f(Y \mid X)$. Therefore,

$$f(Y \mid X, R = 1) = f(Y \mid X).$$

Validity of complete-data method, cont

- This conclusion holds not only for linear regression models but also for logistic regression and Cox regression models.
- For logistic regression models, if the probability of missing data on any variables depends on the value of the dependent variable but does not depend on independent variables, then logistic regression with complete-data yields consistent estimates of the slope coefficients and their standard errors, but not on the intercept estimates (Vach, 1994, Logistic Regression with Missing Values in the Covariates. Springer-Verlag, New York).
- In summary, the complete-data analysis is not a bad alternative approach, particularly when the probability of missing data on one independent variable depends on the value of that variable but not on the dependent variable.

Single imputation

- Single imputation substitutes some reasonable guess (imputation) for each missing value and then perform the analysis as if there were no missing data.
- Imputation-based procedures - explicit modeling methods
  - Mean imputation.
  - Regression imputation: replace missing values by predicted values from a regression of the missing item on items observed for the unit, usually calculated from units with both observed and missing variables present. Mean imputation is a special case.
  - Stochastic regression imputation: replace missing values by a value predicted by regression imputation plus a residual, drawn to reflect uncertainty in a the predicted value.

Taxonomy of missing-data methods, cont

- Imputation-based procedures - implicit modeling methods
  - Hot deck imputation: replacing missing values by values from "similar" responding units in the sample.
  - Cold deck imputation: replace a missing value of an item by a constant value from an external source.
- Model-based procedures: define a model for the observed data and base inferences on the likelihood or posterior distribution under that model.

Nearest neighbor hot deck

- Define a metric to measure distance between units, based on the values of covariates
- Choose imputed values that come from responding units close to the units with missing value.

Nearest neighbor hot deck, Cont

- Let $X = (X_1, \ldots, X_K)$ be K covariates that are observable for all units, and let $x_i = (x_{i1}, \ldots, x_{iK})^T$ be the values of K covariates for a unit i for which y_i is missing.
- Let $d(i, j)$ be the metric between two covariates x_i and x_j of unit i and unit j.

Nearest neighbor hot deck, Cont

- For the unit i, we choose an imputed value for y_i from those units j that are such that (1) $y_j, x_{j1}, \ldots, x_{jK}$ are observed, and (2) $d(i,j)$ is less than some value d_0.

Some special cases of nearest neighbor hot deck

- Maximum deviation: $d(i,j) = max_k \mid x_{ik} - x_{jk} \mid$.
- Mahalanobis: $d(i,j) = (x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)$, where S_{xx} is an estimate of the covariance matrix of x_i.
- Predictive mean: $d(i,j) = [\widehat{y}(x_i) - \widehat{y}(x_j)]^2$, where $\widehat{y}(x_i)$ is the predicted value of Y from the regression of Y on the $x's$ computed using the complete cases.

Comments on imputation

Imputation should generally be

- Conditional on observed variable, to reduce bias due to response, improve precision, and preserve association between missing and observed variables.
- Multivariate, to preserve associations between missing variables.
- Draws from the predictive distribution rather than means.

Account for imputation uncertainty $\longrightarrow$ multiple imputation.

Multiple imputation, cont

- Imputation: Create D imputations of the missing data, $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(D)}$, under a suitable model.
- Analysis: Analyze each of the D completed datasets in the same way.
- Combination: Combine the m sets of estimates and SE's using Rubin's (1987) rules.