Missing Data
Module 1: Introduction, overview

Roderick Little and Trivellore Raghunathan
Course Outline

9:00-10:00  Module 1 Introduction and overview
10:00-10:30 Module 2 Complete-case analysis, including weighting methods
10:30-10:45 Break
10:45-12:00 Module 3 Imputation, multiple imputation
1:30-2:00  Module 4 A little likelihood theory
2:00-3:00  Module 5 Computational methods/software
3:00-3:30  Module 6: Nonignorable models
Module 1: Introduction and Overview

• Missing data defined, patterns and mechanisms

• Analysis strategies
  – Properties of a good method
  – complete-case analysis
  – imputation, and multiple imputation
  – analysis of the incomplete data matrix
  – historical overview

• Examples of missing data problems
Module 1: Introduction and Overview

• Missing data defined, patterns and mechanisms

• Analysis strategies
  – Properties of a good method
  – complete-case analysis
  – imputation, and multiple imputation
  – analysis of the incomplete data matrix
  – historical overview

• Examples of missing data problems
Missing data defined

- Always assume missingness hides a meaningful value for analysis

- Examples:
  - Missing data from missed clinical visit (√)
  - No-show for a preventive intervention (?)
  - In a longitudinal study of blood pressure medications:
    - losses to follow-up (√)
    - deaths (×)
Patterns of Missing Data

- Some methods work for a general pattern

- Other methods apply only to special patterns
Pattern versus mechanism

• Pattern: Which values are missing?
• Mechanism: Why? Reasons related to the study variables?

\[ Y = \text{data matrix, if no data were missing} \]
\[ M = \text{missing-data indicator matrix} \]

\((i,j)\) th element indicates whether \((i,j)\) th element of \(Y\) is missing (1) or observed (0)

– Pattern concerns distribution of \(M\)
– Mechanism concerns distribution of \(M\) given \(Y\)
More on mechanisms

• Data are:
  – missing completely at random (MCAR) if missingness independent of $Y$:
    \[ p(M \mid Y) = p(M) \text{ for all } Y \]
  – missing at random (MAR) if missingness only depends on observed components $Y_{\text{obs}}$ of $Y$:
    \[ p(M \mid Y) = p(M \mid Y_{\text{obs}}) \text{ for all } Y \]
  – missing not at random (MNAR) if missingness depends on missing (as well as perhaps on observed) components of $Y$
MAR for univariate nonresponse

\( X_j = \) complete covariates
\( Y = \) incomplete variable
\( M = 1, Y \) missing
\( 0, Y \) observed

MAR: missingness independent of \( Y \) given \( X_1...X_k \)
That is, \( M \) can depend on \( X’s \) …
but not on \( Y \) given \( X’s \)

Module 1: Introduction and Overview
MAR for monotone missing data
MAR if dropout depends on values recorded prior to drop-out
MNAR if dropout depends on values that are missing (that is, after drop-out)
Censoring by end of study: plausibly MCAR
Drop-out from side effect: MAR if side effect is measured and included in analysis
A non-monotone example

Mechanism is MAR if

\[ \Pr(Y_2 \text{ missing}) = g(Y_1) \]

\[ \Pr(Y_1 \text{ missing}) = f(Y_2) \]

\[ \Pr(\text{complete}) = 1 - f(Y_2) - g(Y_1) \]
Module 1: Introduction and Overview

• Missing data defined
• Patterns and mechanisms
• Analysis strategies
  – Properties of a good method
  – complete-case analysis
  – imputation, and multiple imputation
  – analysis of the incomplete data matrix
  – historical overview
• Examples of missing data problems

Module 1: Introduction and Overview
Properties of a good missing-data method

• Makes use of partial information on incomplete cases, for reduced bias, increased efficiency

• Frequency valid ("calibrated") inferences under plausible model for missing data (e.g. confidence intervals have nominal coverage)

• Propagates missing-data uncertainty, both within and between imputation models

• Favor likelihood based approaches
  – Maximum Likelihood (ML) for large samples
  – Multiple Imputation/Bayes for small samples
Module 1: Introduction and Overview

General Strategies

Imputation
- Complete cases
- Imputations

Complete-Case Analysis
- Complete cases
- Weights
- Imputations
- Discard

Analyze Incomplete
- Complete cases
- e.g. maximum likelihood

Module 1: Introduction and Overview
Module 1: Introduction and Overview

• Missing data defined
• Patterns and mechanisms
• Analysis strategies
  – Properties of a good method
  – complete-case analysis
  – imputation, and multiple imputation
  – analysis of the incomplete data matrix
  – historical overview

• Examples of missing data problems
Ex 1. Missing data in surveys

- National Health and Nutrition Examination Survey (NHANES) III
- Public Use Files subject to:
  - Unit nonresponse
    - noncontact
    - refusal
  - Partial response
    - questionnaire interview complete
    - health examination missing
  - Item nonresponse
Ex. 1 contd. Survey nonresponse

• Issue: different users of public use files get different answers to the same question because missing data are handled in different, sometimes naïve, ways

• Methods
  – weighting for unit nonresponse
  – multiple imputation for partial and item nonresponse
Unit nonrespondents in surveys

• Unit nonrespondents may differ from respondents, leading to data that are missing not at random, biased estimates.
  – A simple formula for means:

\[
\bar{Y}_R - \bar{Y} = \pi_{NR} \times (\bar{Y}_R - \bar{Y}_{NR})
\]

Bias = NR rate * difference in R and NR means
NR = nonrespondent, R = respondent
Sampling unit nonrespondents

• One approach is to follow up a subsample of nonrespondents with special efforts:
  – abbreviated interview
  – monetary incentives

• Data collected can be
  – weighted to represent all nonrespondents
  – used to (multiply) impute other nonrespondents
Ex. 2. Administrative Censoring in Longitudinal Studies

- Censoring by termination of study

Start \(\quad\) end

\[
\begin{array}{c}
\text{Observed} \\
\text{missing}
\end{array}
\]
Ex. 3. Attrition in Longitudinal Studies

• Longitudinal studies often have drop-outs
  – Move out of study catchment area
  – Participation becomes too onerous

• Common analyses have problems:
  – complete case analysis is biased if drop-outs differ
  – Naïve imputation (e.g. last observation carried forward) involves unrealistic assumptions

• We discuss better alternatives
Ex 4. Missing Data in Clinical Trials

• “A long standing issue in clinical trials, and especially in regulatory submissions that contain clinical trials intended to support efficacy and safety and marketing approval
• ICH E9 addresses it briefly but no analysis advice
• FDA’s critical path initiative identified this as a topic in the streamlining of clinical trials and PhRMA, in negotiating the PDUFA 4 agreement, wanted FDA to bring consensus to this topic (FDAAA)”
• (From presentation by Robert T. O’Neill Ph.D., Director, Office of Biostatistics Center for Drug Evaluation and Research, FDA ,at the Annual Conference of the International Society for Clinical Biostatistics, Prague, Aug, 2009)
Panel on Handling Missing Data in Clinical Trials

- RODERICK J. A. LITTLE (*Chair*), Department of Biostatistics, University of Michigan
- KAY DICKERSIN, Department of Epidemiology, Johns Hopkins University
- RALPH D’AGOSTINO, Department of Mathematics and Statistics, Boston University
- SCOTT S. EMERSON, Department of Biostatistics, University of Washington, Seattle,
- JOHN T. FARRAR, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine
- CONSTANTINE FRANGAKIS, Department of Biostatistics, John Hopkins University
- JOSEPH W. HOGAN, Center for Statistical Sciences, Program in Public Health, Brown University
- GEERT MOLENBERGHS, International Institute for Biostatistics and Statistical Bioinformatics, Universiteit Hasselt, The Netherlands
- SUSAN A. MURPHY, Department of Statistics, University of Michigan, Ann Arbor
- JAMES D. NEATON, School of Public Health, University of Minnesota
- ANDREA ROTNITZKY, Departamento de Economia, Universidad Torcuato Di Tella, Buenos Aires, Argentina
- DANIEL SCHARFSTEIN, Department of Biostatistics, Johns Hopkins University,
- WEICHUNG (JOE) SHIH, Department of Biostatistics, University of Medicine and Dentistry of New Jersey School of Public Health
- JAY P. SIEGEL, Johnson & Johnson, Radnor, Pennsylvania
- HAL STERN, Department of Statistics, University of California, Irvine

- MICHAEL COHEN (NAS) Study Director

Module 1: Introduction and Overview
Key Take-Home Messages

- Missing data undermines randomization, the lynchpin of inferences in confirmatory trials
- Limiting missing data should be a major consideration when weighing alternative study designs
  - Analysis methods come with unverifiable assumptions, and limiting these assumptions is crucial
- Careful attention to avoiding missing data in trial conduct can greatly limit the scope of the problem
- Analysis methods need to be driven by plausible scientific assumptions
- Sensitivity analyses to assess robustness to alternative analysis models are needed
  - Lack of robust treatment effect from these analyses reinforces the need to limit missing data in trial design and conduct
Missing Data in Clinical Trials

• Clinical trials record and compare treatment-specific outcomes. So it is important to distinguish between
  – Treatment discontinuation: Treatment-specific outcomes that are not recorded when participants go off-protocol, i.e. discontinue their assigned treatments (lack of efficacy, lack of tolerability)
  – Analysis dropouts: Missing data arising from inability to record outcomes, e.g. from missed clinic visits, attrition. Participants may or may not be off protocol
  – Indicate below that these are somewhat different examples of missing data problems.
Ex 5 -- Dose-Titration Study of Tacrine for Alzheimer’s Disease

- Randomized, double-blind dose-escalation study (Knapp et al. 1994). Outcome - ADAS-COG

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Time (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3  4  5</td>
</tr>
<tr>
<td>Placebo</td>
<td>0  0  0  0  0</td>
</tr>
<tr>
<td>80mg</td>
<td>40  80 80 80 80</td>
</tr>
<tr>
<td>120mg</td>
<td>40  80 120 120 120</td>
</tr>
</tbody>
</table>
The Drop-Out Problem

• Titration to higher dosages to avoid side-effects on liver function

• Patients with side effects removed from double-blind study

• Other drop-outs from lack of compliance, dose-related adverse events

• Substantial differential drop-out rate at t=5:
  – Placebo 44/184 (24%)
  – 80mg 31/61 (51%)
  – 120mg 244/418 (57%)
Ex. 6. Complete-data problems formulated as missing data

Some complete-data problems can be formulated with underlying unobserved data, allowing incomplete-data methods of analysis, e.g. EM algorithm

– Factor analysis -- multivariate regression with unobserved regressors
– Mixed-effects models -- random effects are unobserved “missing” data
– Genetics -- genotypes are “missing”
Conclusions

• Some methods apply to particular patterns, others to any pattern
• Properties of methods vary depending on mechanism
• All methods have limitations -- better to avoid missing values, or try to minimize the problem
• Course will examine methods in more detail
Missing Data
Roderick Little and Trivellore Raghunathan

Module 2
Complete-case analysis, including weighting methods
Complete-Case Analysis

- Default Analysis in Statistical Packages
- Simple but has limitations
Unweighted CC Analysis

• Easy (but missing values must be flagged!)
• Does not invent data
• Simple and may be good enough with small amounts of missing data
  – but defining “small” is problematic; depends on
    • fraction of incomplete cases
    • recorded information in these cases
    • parameter being estimated
Limitations of CC Analysis

• Loss of information in incomplete cases has two aspects:
  – Increased variance of estimates
  – Bias when complete cases differ systematically from incomplete cases

• restriction to complete cases requires that the complete cases are representative of all the cases for the analysis in question – for some (not all) analyses this implies MCAR

• this assumption is often questionable!
Increased variance from CC analysis

Inefficiency of CC Analysis depends on how much information is contained in the discarded incomplete cases.

In a likelihood analysis, this is measured by the information matrices for the complete cases and all the cases, as discussed below.
Information measures

\[ I_{\text{full}} = I_{\text{obs}} + I_{\text{mis}} \]

fraction of missing information \( \lambda_{\text{mis}} = I_{\text{mis}} I_{\text{full}}^{-1} \)

\[ I_{\text{obs}} = I_{\text{CC}} + I_{\text{IC}} \] (assuming MCAR)

fraction of info in incomplete cases \( \lambda_{\text{IC}} = I_{\text{IC}} I_{\text{obs}}^{-1} \)

\( \lambda_{\text{mis}} \) is not recoverable by analysis, \( \lambda_{\text{IC}} \) is ...

- Loss of information depends on the parameter of interest ...

2. Complete-case analysis, weighting
Example: univariate nonresponse

\[ X_1 \ X_2 \ldots X_p \quad Y \]

Suppose \( X_1, \ldots, X_p \) are strong predictors of \( Y \)

\( I_{IC} \) is substantial for unconditional mean of \( Y \)

\( I_{IC} = 0 \) for conditional mean of \( Y \) given \( X_1, \ldots, X_p \)!
Bias of CC analysis: Simulation Study

• True model:

\[
X \sim N(0,1)
\]
\[
\text{Logit}[\Pr(E=1|X)]=0.5+X
\]
\[
\text{logit}[\Pr(D=1|E,X)]=0.25+0.5X+1.1E
\]

• Sample size: 500
• Number of Replicates: 5000
• Before Deletion Data Sets
Missing-Data Mechanism

- $D$ and $E$: completely observed
- $X$: sometimes missing
- Values of $X$ in each cell are set to missing with the following underlying probabilities:

  $D=0, E=0$: $p_{00} = 0.19$
  $D=0, E=1$: $p_{01} = 0.09$
  $D=1, E=0$: $p_{10} = 0.015$
  $D=1, E=1$: $p_{11} = 0.055$

MAR mechanism
Before Deletion Estimates

- Histogram of 5000 estimates before deleting values of $X$
- logistic model
  
  $\text{logit } Pr(D=1|E,X) = \beta_0 + \beta_1 E + \beta_2 X$

2. Complete-case analysis, weighting
Complete-Case Estimates

Histogram of complete-case analysis estimates

Delete subjects with missing X values

True value = 1.1, serious negative bias
Weighted CC Analysis

- **Weight** respondents differentially to reduce nonresponse bias – e.g. mean becomes weighted mean.
- Common for unit nonresponse in surveys.
Sampling weights

• In a probability survey, each sampled unit $i$ represents $w_i$ units of the population, where

$$w_i = \frac{1}{\Pr(\text{unit } i \text{ sampled})}$$

$w_i$ is determined by the sample design and hence *known*

• Extend this idea to unit nonresponse ...
Unit Nonresponse Weights

- If probability of response was known, could obtain weight for units that are sampled and respond:

\[ w_i = \frac{1}{pr(\text{unit } i \text{ is sampled and responds})} \]

\[ = \frac{1}{pr(\text{i sampled})} \times \frac{1}{pr(\text{i responds|sampled})} \]

\[ =(\text{sampling weight}) \times (\text{response weight}) \]

Since prob of response is not known, we need to estimate it.
Adjustment Cell method

- Group respondents and nonrespondents into adjustment cells with similar values on variables recorded for both:
- e.g. white females aged 25-35 living in SW

\[ \begin{array}{c}
\text{100 in sample} < \\
\text{80 respondents} \\
\text{20 nonrespondents} \\
\text{pr(response in cell) = 0.8} \\
\text{response weight = 1.25}
\end{array} \]
Simulation Study with weights

- $r_{ij}$= response rate in cell $D=i$ and $E=j$

- weight respondents in cell $D=i$ and $E=j$ by

$$w_{ij} = \frac{1}{r_{ij}}$$
Weighted Estimates

- Weighted Logistic Regression Model

Weighted by inverse of cell-specific response rates

True Value: 1.1

Bias is removed
Choice of Adjustment Cells

• With extensive covariate information, can’t cross-classify on all of them
• How do we choose which variables to use?
• Consider problem of estimating the mean of an outcome \( Y \)

\[
\begin{array}{cccccc}
X_1 & X_2 & \ldots & X_p & Y & M \\
0 & 0 & 0 & ? & ? & ? \\
\end{array}
\]

Complete cases
## Effect of NR weighting on bias and variance

<table>
<thead>
<tr>
<th>Assoc of Adj Cell $X$ with Nonresponse $M$</th>
<th>Assoc of Adj Cell $X$ with Outcome $Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Bias: ---</td>
<td>Bias: ---</td>
</tr>
<tr>
<td>Var: ---</td>
<td>Var: ↓</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Bias: ---</td>
<td>Bias: ↓</td>
</tr>
<tr>
<td>Var: ↑</td>
<td>Var: ↓</td>
</tr>
</tbody>
</table>

See Little and Vartivarian (2005 Survey Methodology)
Choice of Adjustment Cells

• To reduce bias, adjustment cell variable needs to be related to nonresponse $M$ and outcome $Y$
• To reduce variance, adjustment cell variables need to be related to outcome (otherwise weighting increases variance)
• Two methods for creating adjustment cells with multiple $X$’s are
  – Response propensity stratification, based on regression of $M$ on $X$
  – Predictive mean stratification, based on regression of $Y$ on $X$
Response propensity stratification

- $X = \text{covariates observed for respondents and nonrespondents, } Y \text{ missing}$
- $M = \text{missing-data indicator}$
  - nonrespondent $= 1$, respondent $= 0$
- (A) Regress $M$ on $X$ (probit or logistic), using respondent and nonrespondent data $\hat{p}(M = 0 | X) = \text{propensity score}$
- (B1) Weight respondents by inverse of propensity score from (A), $1/ \hat{p}(M = 0 | X)$, or:
- (B2) form adjustments cells by categorizing $1/ \hat{p}(M = 0 | X)$
- Note that this method is only effective if propensity is also related to the outcome
Predictive mean stratification

- $X = \text{covariates observed for respondents and nonrespondents}
- Y = \text{outcome with missing data}
- (A) Regress $Y$ on $X$ (linear, other as appropriate) using respondent data only
- (B) Form adjustments cells with similar values of predictions from the regression $\hat{Y}(X)$

- This method has potential to reduce both bias and variance
- Note that the adjustment cells depend on outcome, so this method yields different cells for each outcome, hence is more complex with many outcomes with missing values
Inference from Weighted Data

• Role of weights in analytical inference (regression, factor analysis, …) is controversial

• Can use packages for computing standard errors for complex sample designs -- but often these do not take into account sampling uncertainty in weights

• Bootstrap/Jackknife of weighting procedure propagates uncertainty in weights – but weights need to be recalculated on each BS/JK sample
Post-stratification

- Weight respondents to match distribution of a categorical variable with distribution known from external data (for example the census)

<table>
<thead>
<tr>
<th>Age post-stratum</th>
<th>20-29</th>
<th>30-39</th>
<th>40-49</th>
<th>&gt;50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondents</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td>$m_4$</td>
</tr>
<tr>
<td>Population</td>
<td>$N_1$</td>
<td>$N_2$</td>
<td>$N_3$</td>
<td>$N_4$</td>
</tr>
</tbody>
</table>

Assign respondents in Age category $j$ the weight:

$$w_j = C \times \frac{N_j}{m_j}$$

$C$ chosen so that weights sum to number of respondents
Post-stratification (continued)

- Often post-stratification is a final step after other weighting adjustments
- Corrects for bias from differential nonresponse across post-strata
- Useful post-stratification variables are predictive of both nonresponse and survey outcomes
- Statistician does not control the counts $m_j$
  - If these are too noisy the weights may need smoothing
- *Raking* extends method to cases where more than one margin is available
Summary of Weighting Methods

- Weighting is a relatively simple device for reducing bias from complete-case analysis.
- Same weight for all variables -- simple, but better methods tune adjustment according to outcome.
- No built in control of variance
  - Ad-hoc trimming is common in surveys.
- Less useful when:
  - Covariate information is extensive.
  - Pattern of missing-data is non-monotone.
Missing Data
Roderick Little and Trivellore Raghunathan

Module 3
Imputation, multiple imputation
Features of Imputation

**Good**
- Rectangular File
- Retains observed data
- Handles missing data once
- Exploits incomplete cases

**Bad**
- Naïve methods can be bad
- Invents data –
- Understates uncertainty

3. Imputation, Multiple Imputation
Imputing Means

Unconditional

\[ \hat{E}(Y_2) = \bar{y}_2 \]

Conditional on observed variables

\[ \hat{y}_{i2} = \hat{E}(y_{i2} \mid y_{i1}) = \hat{\beta}_{20} + \hat{\beta}_{21} y_{i1} \]
Properties of Mean Imputation

• Marginal distributions, associations estimated from filled-in data are distorted

• Standard errors of estimates from filled-in data are too small, since
  – Standard deviations are underestimated
  – “Sample size” is overstated

• Conditional better than unconditional mean, which can be worse than complete cases
Imputing Draws

- Imputations can be random draws from a predictive distribution for the missing values.

\[
\hat{y}_{i2} = \hat{E}(y_{i2} | y_{i1}) + r_i
\]

\[
r_i \sim N(0, s_{221}), s_{221} = \text{resid variance, or}
\]

\[
r_i = \text{residual from randomly selected complete case}
\]
Imputing draws for binary data

- For binary (0-1) data, impute 1 with probability = predicted prob of a one given observed covariates

\[ \hat{p}_{i2} = \Pr(y_{i2} = 1 \mid y_{i1}) \] (e.g. logistic regression)

\[ y_{i2} = \begin{cases} 1, \text{prob } \hat{p}_{i2} \\ 0, \text{prob } 1 - \hat{p}_{i2} \end{cases} \]
Properties of Imputed Draws

- Adds noise, less efficient than imputing means, but:
- No (or reduced) bias for estimating distributions
- More robust to nonlinear data transformations
- Conditional draws better than unconditional:
  - Improved efficiency
  - Preserves associations with conditioned variables
- Standard errors from filled-in data are improved, but still wrong:
  - Standard deviation is ok
  - Information is overstated; multiple imputation fixes this
What should imputes condition on?

• In principle, all observed variables
  – Whether predictors or outcomes of final analysis model
  – May be impractical with a lot of variables

• Variable selection
  – Similar ideas to weighting adjustments apply
  – Priority to variables predictive of missing variable (and nonresponse)
  – Favor inclusion over exclusion (more later)
Creating the predictive distribution

All imputation methods assume a model for the predictive distribution of the missing values

- *Explicit*: predictive distribution based on a formal statistical model (e.g. multivariate normal); assumptions are explicit
- *Implicit*: focus is on an algorithm, but the algorithm implies an underlying model; assumptions are implicit
Two implicit modeling procedures

• Last observation carried forward (LOCF) imputation for repeated measures with drop-outs:
  – impute last recorded value
  – implicit model: values are unchanged after drop-out

• Hot deck imputation (see Andridge and Little 10)
  – classify respondents, nonrespondents into adjustment cells with similar observed values
  – impute values from random respondent in same cell
  – implicit model: regression of missing variables on variables forming cells, including all interactions
Current Population Survey Hot Deck

• Missing ($Y$): Earnings Variables

• Observed ($X$):
  – Age, Race, Sex, Family Relationship, Children, Marital Status, Occupation, Schooling, Full/Part time, Type of Residence, Income Recipiency Pattern

• Flexible matching:
  – Joint Classification by $X$ yields giant matrix. If a match is not found, table is coarsened or collapsed in stages until a match is found
CPS Hot Deck (continued)

Good Features

- Imputes real values
- multivariate: associations preserved
- Conditions on $X$’s
- Assessments suggest method works quite well with large data sets

Bad Features

- Does not exploit previous earnings models
- Includes high order interactions at expense of main effects of omitted $X$’s
- Imputation uncertainty not included in standard errors

For comparison of CPS Hot Deck with stochastic regression imputation see David et al. (1986)
Other matching methods

• More generally, nonrespondents $j$ can be matched to respondents $i$ based on a closeness metric $D(i, j)$

  - Adjustment cell: $D(i, j) = \begin{cases} 0, & \text{if } i, j \text{ belong to same cell} \\ 1, & \text{if } i, j \text{ belong to different cells} \end{cases}$

  - Mahalanobis: $D(i, j) = (x_i - x_j)^T S_X^{-1} (x_i - x_j)$

  - Predictive Mean: $D(i, j) = (\hat{y}_i - \hat{y}_j)^T S_{Y,X}^{-1} (\hat{y}_i - \hat{y}_j)$

    $\hat{y}_i = \text{regression prediction of } Y \text{ given } X$

    $S_{Y,X} = \text{resid covariance matrix}$
Properties of matching methods

- Imputation error not propagated in standard errors from filled-in data
- One metric irrespective of outcome -- in contrast, models tailor adjustment to individual $Y$’s
- Predictive mean metric better than Mahalanobis metric, since more targeted to $Y$’s.
- Robust to model misspecification, but needs large samples: poor matches when sample is thin

See Little (1988 JBES) for more discussion
Longitudinal Data: a simple, better alternative to LOCF

- For repeated measures, the following Row +/- Col methods includes individual (row) and time (column) effects

\[ y_{it} = \text{value for subject } i, \text{ time } t \]

\[ \hat{y}_{it} = m + a_i + b_t + r_{kt} \quad (\text{Row + Col}) \]

\[ \hat{y}_{it} = m \times a_i \times b_t \times r_{kt} \quad (\text{Row } \times \text{ Col}) \]

\[ m = \text{grand mean} \]

\[ a_i = \text{row effect, from deviations of row } i \]

\[ b_t = \text{column effect, from deviations of col } t \]

\[ r_{kt} = \text{residual from matched respondent } k \]
Example 2: Logistic simulation example revisited

- Simple to create hot-deck imputations
- $n_{ij} =$ Observed Sample size in cell $D=i, E=j$
- $m_{ij} =$ Number of missing values
- Randomly draw $m_{ij}$ values from $n_{ij}$ observed values with replacement
Hot-deck Single Imputation Estimates

- Single Imputation
- Imputed Data Sets Analyzed as if Complete Data
- **TRUE VALUE 1.1:** estimates are unbiased
Summary of imputation methods

• Imputations should:
  – condition on observed variables
  – be multivariate to preserve associations between missing variables
  – generally be draws rather than means

• Key problem: single imputations do not account for imputation uncertainty in se’s. Consider next two approaches to this problem
  – bootstrapping the imputation method
  – multiple imputation
Accounting for Imputation Uncertainty

• Imputation “makes up” the missing data
  – treats imputed values as the truth
• For statistical inference (standard errors, P-Values, confidence intervals) need methods that account for imputation error
  – (A) redo imputations using sample reuse methods – bootstrap, jackknife
  – (B) Multiple imputation (Rubin 1987)
Bootstrapping: with complete data

• A bootstrap sample of a complete data set $S$ with $n$ observations is a sample of size $n$ drawn with replacement from $S$
  – Operationally, assign weight $w_i$ to unit $i$ equal to number of times it is included in the bootstrap sample

$$w_1, \ldots, w_n \sim \text{MNOM}(n; \frac{1}{n}, \ldots, \frac{1}{n})$$
Bootstrap distribution

• Let $\hat{\theta}^{(b)}$ be a consistent parameter estimate from the $b$th bootstrap data set.

• Inference can be based on the bootstrap distribution generated by values of $\hat{\theta}^{(b)}$.

• In particular the bootstrap estimate is

$$\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$$

with variance

$$\hat{V}_{\text{boot}} = \frac{1}{B - 1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2$$
Bootstrapping incomplete data

- For incomplete data:
  - bootstrap the incomplete cases
  - impute bootstrapped data set
  - $\hat{\theta}^{(b)} =$ consistent estimate from $b$th data set, with values imputed; then as before:

$$
\hat{\theta}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)} \\
\hat{V}_{\text{boot}} = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{(b)} - \hat{\theta}_{\text{boot}})^2
$$

* Bootstrap then impute, not
* Impute then bootstrap
Imputing the bootstrap sample

- Impute so that the estimate $\hat{\theta}_b$ from imputed data is consistent. In particular:
  - conditional mean ok for linear statistics
  - conditional draw ok for linear or nonlinear statistics; more general, but loss of efficiency
- Computationally intensive: imputations created for each bootstrap data set
  $B=200, 1000$ are typical numbers
Multiple Imputation

- Create $D$ sets of imputations, each set a draw from the predictive distribution of the missing values
  - e.g. $D=5$

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.1</td>
<td>2.7</td>
<td>1.9</td>
<td>2.5</td>
<td>2.3</td>
</tr>
<tr>
<td>4.5</td>
<td>5.1</td>
<td>5.8</td>
<td>3.9</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>31</td>
<td>32</td>
<td>18</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
Multiple Imputation Inference

• $D$ completed data sets (e.g. $D = 5$)
• Analyze each completed data set
• Combine results in easy way to produce multiple imputation inference
• Particularly useful for public use datasets
  – data provider creates imputes for multiple users, who can analyze data with complete-data methods
MI Inference for a Scalar Estimand

\( \theta = \) estimand of interest

\( \hat{\theta}_d = \) estimate from \( d \)th dataset \( (d = 1, ..., D) \)

The MI estimate of \( \theta \) is

\[
\bar{\theta}_D = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d
\]

\( W_d = \) estimate of variance of \( \hat{\theta}_d \) from \( d \)th dataset

The MI estimate of variance is

\[
T_D = \bar{W}_D + (1 + 1/D)B_D
\]

\[
\bar{W}_D = \frac{1}{D} \sum_{d=1}^{D} W_d = \text{Within-Imputation Variance}
\]

\[
B_D = \frac{1}{D-1} \sum_{d=1}^{D} (\hat{\theta}_d - \bar{\theta}_D)^2 = \text{Between-Imputation Variance}
\]
Example of Multiple Imputation

• First imputed dataset

<table>
<thead>
<tr>
<th>Dataset ($d$)</th>
<th>$\mu_1$</th>
<th>$\beta_{531234}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>12.6 ($3.6^2$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4.32 ($1.95^2$)</td>
</tr>
</tbody>
</table>

2.1
4.5
24 1
• Second imputed dataset

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_1$</td>
<td>$Y_2$</td>
<td>$Y_3$</td>
<td>$Y_4$</td>
<td>$Y_5$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.7</td>
<td>5.1</td>
<td>31</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset ($d$)</th>
<th>$\mu_1$</th>
<th>$\beta_{53,1234}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6 (3.6$^2$)</td>
<td>4.32 (1.95$^2$)</td>
</tr>
<tr>
<td>2</td>
<td>12.6 (3.6$^2$)</td>
<td>4.15 (2.64$^2$)</td>
</tr>
</tbody>
</table>
3. Imputation, Multiple Imputation

- Third imputed dataset

<table>
<thead>
<tr>
<th>Dataset (d)</th>
<th>$\mu_1$</th>
<th>$\beta_{531234}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6 (3.6$^2$)</td>
<td>4.32 (1.95$^2$)</td>
</tr>
<tr>
<td>2</td>
<td>12.6 (3.6$^2$)</td>
<td>4.15 (2.64$^2$)</td>
</tr>
<tr>
<td>3</td>
<td>12.6 (3.6$^2$)</td>
<td>4.86 (2.09$^2$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
<th>$Y_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.9</td>
<td>5.8</td>
<td>32</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

3. Imputation, Multiple Imputation
- **Fourth imputed dataset**

<table>
<thead>
<tr>
<th>Dataset ($d$)</th>
<th>$\mu_1$</th>
<th>$\beta_{53\cdot1234}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6 (3.6$^2$)</td>
<td>4.32 (1.95$^2$)</td>
</tr>
<tr>
<td>2</td>
<td>12.6 (3.6$^2$)</td>
<td>4.15 (2.64$^2$)</td>
</tr>
<tr>
<td>3</td>
<td>12.6 (3.6$^2$)</td>
<td>4.86 (2.09$^2$)</td>
</tr>
<tr>
<td>4</td>
<td>12.6 (3.6$^2$)</td>
<td>3.98 (2.14$^2$)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$Y_4$</th>
<th>$Y_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• Fifth imputed dataset

<table>
<thead>
<tr>
<th>Dataset (d)</th>
<th>$\mu_1$</th>
<th>$\beta_{53\cdot1234}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6 ($3.6^2$)</td>
<td>4.32 ($1.95^2$)</td>
</tr>
<tr>
<td>2</td>
<td>12.6 ($3.6^2$)</td>
<td>4.15 ($2.64^2$)</td>
</tr>
<tr>
<td>3</td>
<td>12.6 ($3.6^2$)</td>
<td>4.86 ($2.09^2$)</td>
</tr>
<tr>
<td>4</td>
<td>12.6 ($3.6^2$)</td>
<td>3.98 ($2.14^2$)</td>
</tr>
<tr>
<td>5</td>
<td>12.6 ($3.6^2$)</td>
<td>4.50 ($2.47^2$)</td>
</tr>
</tbody>
</table>

Mean: 12.6 ($3.6^2$) 4.36 ($2.27^2$)  
Var: 0 0.339
### Summary of MI Inferences

<table>
<thead>
<tr>
<th></th>
<th>$\bar{\theta}_D$</th>
<th>$\bar{W}_D$</th>
<th>$B_D$</th>
<th>$\sqrt{T_D} = \sqrt{\bar{W}_D + \frac{6}{5} B_D}$</th>
<th>$\hat{\gamma}_D = \frac{1.2 B_D}{(1.2 B_D + \bar{W}_D)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>12.6</td>
<td>3.6</td>
<td>0</td>
<td>3.6</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_{531234}$</td>
<td>4.36</td>
<td>2.27</td>
<td>0.339</td>
<td>2.36</td>
<td>0.073</td>
</tr>
</tbody>
</table>

\[
\hat{\gamma}_D = \frac{(1+1/D)B_D}{(1+1/D)B_D + \bar{W}_D}
\]

= estimated fraction of missing information

---

3. Imputation, Multiple Imputation

32
Creating Multiple Imputations

- Multiple Imputations created within a single model take into account within-model uncertainty
- Multiple Imputations can also be created under alternative models, to account for imputation model uncertainty
- Imputations can be based on implicit or explicit models, as for single imputation
Examples of draws for $d$th set of MI’s

- Hot Deck: create $D$ candidate donors that are close to incomplete case, and draw $d$th value from this set with replacement

- Regression: add normal draws $r^{(d)}_i$ to regression predictions

$$y^{(d)}_i = \hat{E}(y_{i2} \mid y_{i1}) + r^{(d)}_i$$

$r^{(d)}_i \sim N(0, \hat{\sigma}^2)$

These methods are simple but *improper* – do not account for parameter uncertainty
Later consider *proper* methods that take into account uncertainty in regression coefficients
Imperfect MI

1. Estimate parameters (e.g. using complete cases)
2. Impute missing values given estimated parameters
3. Repeat (2) for MI data sets
4. Use MI formula for variance

Note: only works for small amounts of missing data
Example: 2x2 Table

Estimands:
- Cell (1,1) proportion
- Odds ratio

Multiple Imputation (D=5): Draw 5 sets of independent Binomial random variables
- A ~ Bin(30, 100/150)
- B ~ Bin(60, 75/150)
- C ~ Bin(28, 100/175)
- D ~ Bin(60, 50/125)

Improper!

\[
\begin{array}{c|c}
Y_1 & Y_2 \\
1 & 100 \, 50 \\
2 & 75 \, 75 \\
\end{array}
\]

\[
\begin{array}{c|c}
Y_1 & Y_2 \\
1 & ? \, ? \\
2 & ? \, ? \\
\end{array}
\]

\[
\begin{array}{c|c}
Y_1 & Y_2 \\
1 & 28 \, 60 \\
2 & ? \, ? \\
\end{array}
\]

\[
\begin{array}{c|c}
Y_1 & Y_2 \\
1 & ? \, ? \\
2 & ? \, ? \\
\end{array}
\]

\[
\begin{array}{c|c}
Y_1 & Y_2 \\
1 & 100 + A + C \, 50 + (30 - A) + D \\
2 & 75 + (28 - C) + B \, 75 + (60 - D) + (60 - B) \\
\end{array}
\]
### Imputation, Multiple Imputation

Below is a table showing the data structure and calculations involved in multiple imputation.

<table>
<thead>
<tr>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_2$</th>
<th>$Y_2$</th>
<th>$Y_2$</th>
<th>$Y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>138</td>
<td>137</td>
<td>136</td>
<td>133</td>
<td>135</td>
</tr>
<tr>
<td>2</td>
<td>83</td>
<td>84</td>
<td>81</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>107</td>
<td>150</td>
<td>112</td>
<td>114</td>
<td>115</td>
<td>128</td>
</tr>
<tr>
<td>114</td>
<td>145</td>
<td>147</td>
<td>144</td>
<td>133</td>
<td>133</td>
</tr>
</tbody>
</table>

The table includes the following calculations:

- $\hat{\theta}_{11}$
- $\hat{\theta}_{11}(1 - \hat{\theta}_{11}) / 478$
- $\ln(OR)$
- $\text{Var}(\ln(OR)) = (1/\hat{\theta}_{11} + 1/\hat{\theta}_{12} + 1/\hat{\theta}_{21} + 1/\hat{\theta}_{22}) / 478$

The MI estimates are as follows:

<table>
<thead>
<tr>
<th>$\hat{\theta}_{11}$</th>
<th>$\hat{\theta}<em>{11}(1 - \hat{\theta}</em>{11}) / 478$</th>
<th>$\ln(OR)$</th>
<th>$\text{Var}(\ln(OR))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.29</td>
<td>0.00043</td>
<td>0.85</td>
<td>0.0353</td>
</tr>
<tr>
<td>0.29</td>
<td>0.000428</td>
<td>0.75</td>
<td>0.0350</td>
</tr>
<tr>
<td>0.28</td>
<td>0.000426</td>
<td>0.77</td>
<td>0.0353</td>
</tr>
<tr>
<td>0.28</td>
<td>0.00042</td>
<td>0.66</td>
<td>0.0348</td>
</tr>
<tr>
<td>0.28</td>
<td>0.000424</td>
<td>0.54</td>
<td>0.0349</td>
</tr>
<tr>
<td>0.28</td>
<td>0.000425</td>
<td>0.71</td>
<td>0.0351</td>
</tr>
<tr>
<td>1.62E-05</td>
<td></td>
<td>0.014</td>
<td></td>
</tr>
</tbody>
</table>

The table also includes annotations for Between-variance and Within-variance, indicating the variance components.
$$\text{var}_{MI}(\hat{\theta}_{11}) = 0.000425 + \frac{5+1}{5} \times 1.62 \times 10^{-5} = 0.000445$$

$$r_m = \frac{\frac{5+1}{5} \times 1.62 \times 10^{-5}}{0.000425 + \frac{5+1}{5} \times 1.62 \times 10^{-5}} = 0.044$$

$$df = (5-1)/(0.044)^2 \approx 2066$$

95% confidence interval:

$$0.28 \pm 1.96 \times \sqrt{0.000445} = (0.24, 0.32)$$

Complete-case:

$$0.33 \pm 1.96 \times \sqrt{\frac{0.33 \times 0.67}{300}} = (0.28, 0.38)$$
\[
\text{var}_{MI} (\log(OR)) = 0.0351 + \frac{5+1}{5} 0.014 = 0.0519
\]

\[
r_M = \frac{\frac{5+1}{5} 0.014}{0.0351 + \frac{5+1}{5} 0.014} = 0.325
\]

\[
df = (5 - 1)/(0.325)^2 \approx 38
\]

95\% Confidence interval:

\[
0.71 \pm 2.024 \times \sqrt{0.0519} = (0.25, 1.17)
\]

Complete–case:

\[
0.69 \pm 1.96 \times \sqrt{\frac{1}{100} + \frac{1}{75} + \frac{1}{50} + \frac{1}{75}} = (0.22, 1.16)
\]
• The proper imputation approach should reflect uncertainty in the estimated proportions used in the binomial distribution.

• Using software that creates proper multiple imputation (CAT [discussed later]) on the same data set, we get

\[ \hat{\theta}_{11} = 0.2812, \ SE = 0.02088 \]

\[ \log(\text{OR}) = 0.7364, \ SE = 0.2276 \]
Creating proper MI’s via bootstrap

(1) Take Bootstrap sample
(2) Estimate parameters (e.g. using complete cases) on BS sample
(3) Impute missing values given estimated parameters
(4) Repeat (1)-(3) for MI data sets
(5) Use MI formula for variance

Note: estimating parameters on BS sample propagates imputation uncertainty
Example -- Dose-Titration Study of Tacrine for Alzheimer’s Disease

- Randomized, double-blind dose-escalation study (Knapp et al. 1994). Outcome - ADAS-COG

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Time (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3  4  5  [6]  [7]</td>
</tr>
<tr>
<td>Placebo</td>
<td>0  0  0  0  0  0  0</td>
</tr>
<tr>
<td>80mg</td>
<td>40 80 80 80 80 120 120</td>
</tr>
<tr>
<td>120mg</td>
<td>40 80 120 120 120 160 160</td>
</tr>
</tbody>
</table>
The Drop-Out Problem

- Titration to higher dosages to avoid side-effects on liver function
- Patients with side effects removed from double-blind study
- Other drop-outs from lack of compliance, dose-related adverse events
- Substantial differential drop-out rate at t=5:
  - Placebo 44/184 (24%)
  - 80mg 31/61 (51%)
  - 120mg 244/418 (57%)
MI model

• Missing values of ADAS-COG multiply imputed using a regression on dose, previous ADAS-COG values and baseline covariates. Two models:
  – Continuing dose model: assumes same dose after dropout as last dose before dropout
  – Zero-dose model: dose goes to zero after drop-out

• Contrast Intent-to-treat, where dose is based on original randomization
Ex. 1 contd. Tacrine Dataset

IT Analysis, Continuing Dose  MI Model: 80mg vs Placebo

<table>
<thead>
<tr>
<th>MI number</th>
<th>Treat.diff (s.e.)</th>
<th>p-value</th>
<th>95 % C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-3.486 (0.951)</td>
<td>0.0003</td>
<td>(-5.35,-1.62)</td>
</tr>
<tr>
<td>2</td>
<td>-3.682 (0.876)</td>
<td>0.0000</td>
<td>(-5.40,-1.97)</td>
</tr>
<tr>
<td>3</td>
<td>-3.142 (0.944)</td>
<td>0.0009</td>
<td>(-4.99,-1.29)</td>
</tr>
<tr>
<td>4</td>
<td>-4.889 (0.908)</td>
<td>0.0000</td>
<td>(-6.67,-3.11)</td>
</tr>
<tr>
<td>5</td>
<td>-4.633 (0.910)</td>
<td>0.0000</td>
<td>(-6.42,-2.85)</td>
</tr>
<tr>
<td>6</td>
<td>-4.146 (0.920)</td>
<td>0.0000</td>
<td>(-5.95,-2.34)</td>
</tr>
<tr>
<td>7</td>
<td>-5.239 (0.925)</td>
<td>0.0000</td>
<td>(-7.05,-3.43)</td>
</tr>
<tr>
<td>8</td>
<td>-4.463 (0.933)</td>
<td>0.0000</td>
<td>(-6.29,-2.63)</td>
</tr>
<tr>
<td>9</td>
<td>-4.511 (0.953)</td>
<td>0.0000</td>
<td>(-6.38,-2.64)</td>
</tr>
<tr>
<td>10</td>
<td>-3.497 (0.899)</td>
<td>0.0001</td>
<td>(-5.26,-1.73)</td>
</tr>
</tbody>
</table>

MI Inference: **-4.169 (1.173)**  **0.0039**  **(-6.72,-1.62)**
Summary of Multiple Imputation

• Retains advantages of single imputation
  – Consistent analyses
  – Data collectors knowledge
  – Rectangular data sets

• Corrects disadvantages of single imputation
  – Reflects uncertainty in imputed values
  – Corrects inefficiency from imputing draws
    • estimates have high efficiency for modest $M$, e.g. 10
Missing Data
Roderick Little and Trivellore Raghunathan

Module 4
A little likelihood theory for incomplete data
Likelihood methods

• Statistical model + data ⇒ Likelihood

• Two general approaches based on likelihood
  – maximum likelihood inference for large samples
  – Bayesian inference for small samples:
    \[ \log(\text{likelihood}) + \log(\text{prior}) = \log(\text{posterior}) \]

• Methods can be applied to incomplete data
  – do not require rectangular data sets

• First review main ideas for complete data
  – Little and Rubin (2002, chapter 6)
Definition of Likelihood

• Data $Y$
• Statistical model yields probability density $f(Y|\theta)$ for $Y$ with unknown parameters $\theta$

  $L(\theta|Y) = \text{const} \times f(Y|\theta)$

• Likelihood function is then a function of $\theta$

• Loglikelihood is often easier to work with:

  $\ell(\theta|Y) = \log L(\theta|Y) = \text{const} + \log \{ f(Y|\theta) \}$

  Constants can depend on data but not on parameter $\theta$
Example: Normal sample

- $Y = (y_1, ..., y_n)$ univariate iid normal sample

$$\theta = (\mu, \sigma^2)$$

$$f(Y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right)$$

$$\ell(\mu, \sigma^2 \mid Y) = -\frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$
Maximum Likelihood Estimate

• The maximum likelihood (ML) estimate $\hat{\theta}$ of $\theta$ maximizes the likelihood, or equivalently the log-likelihood
  \[
  L(\hat{\theta} | Y) \geq L(\theta | Y) \quad \text{for all } \theta \\
  \log L(\hat{\theta} | Y) \geq \log L(\theta | Y) \quad \text{for all } \theta
  \]

• The ML estimate is the
  “value of the parameter that makes the data most likely”

• The ML estimate is not necessarily unique, but is for many regular problems given enough data
Computing the ML estimate

• In regular problems, the ML estimate can be found by solving the likelihood equation

\[ S(\theta|Y) = 0 \]

where \( S \) is the score function, defined as the first derivative of the loglikelihood:

\[ S(\theta|Y) \equiv \frac{\partial \log L(\theta|Y)}{\partial \theta} \]
Normal Examples

• Univariate Normal sample  \( Y = (y_1, \ldots, y_n) \)  \( \theta = (\mu, \sigma^2) \)

\[
\hat{\mu} = \bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

(Note the lack of a correction for degrees of freedom)

• Multivariate Normal sample

\[
\hat{\mu} = \bar{y}, \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^T
\]

• Normal Linear Regression (possibly weighted)

\[
(y_i \mid x_{i1}, \ldots, x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2 / w_i)
\]

\[
\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \text{weighted least squares estimates}
\]

\[
\hat{\sigma}^2 = (\text{weighted residual sum of squares})/n
\]
Properties of ML estimates

• Under assumed model, ML estimate is:
  – Consistent (not necessarily unbiased)
  – Efficient for large samples
  – Not necessarily the best for small samples

• ML estimate is transformation invariant
  – If $\hat{\theta}$ is the ML estimate of $\theta$
    Then $\phi(\hat{\theta})$ is the ML estimate of $\phi(\theta)$
Large-sample ML Inference

• Basic large-sample approximation:
  
  for regular problems, 
  \[ \theta - \hat{\theta} \sim N(0, C) \]

  where \( C \) is a covariance matrix estimated from the sample
  – Frequentist treats \( \hat{\theta} \) as random, \( \theta \) as fixed; equation defines the sampling distribution of \( \hat{\theta} \)
  – Bayesian treats \( \theta \) as random, \( \hat{\theta} \) as fixed; equation defines posterior distribution of \( \theta \)
Forms of precision matrix

• The precision of the ML estimate is measured by $C^{-1}$

Some forms for this are:

– Observed information (recommended)

$$C^{-1} = I(\hat{\theta} | Y) = -\frac{\partial^2 \log L(\theta | Y)}{\partial \theta \partial \theta} \bigg|_{\theta=\hat{\theta}}$$

– Expected information (not as good, may be simpler)

$$C^{-1} = J(\hat{\theta}) = E[I(\theta | Y, \theta)] \bigg|_{\theta=\hat{\theta}}$$

– Some other approximation to curvature of loglikelihood in the neighborhood of the ML estimate
Interval estimation

- 95% (confidence, probability) interval for scalar \( \theta \) is:
  \[
  \hat{\theta} \pm 1.96 \, \sigma^{1/2},
  \]
  where 1.96 is 97.5 percentile of normal distribution.

- Example: univariate normal sample

\[
I = J = \begin{bmatrix}
  n / \hat{\sigma}^2 & 0 \\
  0 & n / (2\hat{\sigma}^4)
\end{bmatrix}
\quad \Rightarrow
C = \begin{bmatrix}
  \hat{\sigma}^2 / n & 0 \\
  0 & 2\hat{\sigma}^4 / n
\end{bmatrix}
\]

Hence some 95% intervals are:

\[
\bar{y} \pm 1.96 \, s / \sqrt{n} \quad \text{for} \ \mu
\]

\[
s^2 \pm 1.96 \, s^2 / \sqrt{n/2} \quad \text{for} \ \sigma^2
\]

\[
\ln(s) \pm 1.96 \, \sqrt{2 / n} \quad \text{for} \ \ln(\sigma)
\]
Significance Tests

Tests based on likelihood ratio (LR) or Wald (W) statistics:

\[ \theta = (\theta_{(1)}, \theta_{(2)}); \theta_{(1)0} = \text{null value of } \theta_{(1)}; \theta_2 = \text{other parameters} \]

\[ \hat{\theta} = \text{unrestricted ML estimate} \]

\[ \tilde{\theta} = (\theta_{(1)0}, \tilde{\theta}_{(2)}); \tilde{\theta}_{(2)} = \text{ML estimate of } \theta_{(2)} \text{ given } \theta_{(1)} = \theta_{(1)0} \]

**LR statistic:**

\[ \text{LR}(\hat{\theta}, \tilde{\theta}) = 2 \left[ \ell(\hat{\theta} | Y) - \ell(\tilde{\theta} | Y) \right] \]

**Wald statistic:**

\[ W(\hat{\theta}, \tilde{\theta}) = (\theta_{(1)0} - \hat{\theta}_{(1)})^T C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)}) \]

\[ C_{(11)} = \text{covariance matrix of } (\theta_{(1)} - \hat{\theta}_{(1)}) \]

yield P-values

\[ P = pr \left( \chi_q^2 > D(\hat{\theta}, \tilde{\theta}) \right) \]

\[ D = \text{LR or Wald statistic}; \ q = \text{dimension of } \theta_0 \]

\[ \chi_q^2 = \text{Chi-squared distribution with } q \text{ degrees of freedom} \]
Bayes inference

- Given a prior distribution $p(\theta)$ for the parameters, inference can be based on the posterior distribution using Bayes’ theorem:

$$p(\theta|Y) = \text{const.} \times p(\theta) \times L(\theta|Y)$$

- For small samples, we prefer Bayes’ inference based on the posterior to the large sample ML approximation.
  - In important standard problems with non-informative priors, Bayes yields inference comparable to small-sample frequentist inference
  - In many non-standard problems, Bayes yields answers where no exact frequentist answer exists
Example: linear regression

The normal linear regression model:

\[ (y_i | x_{i1}, \ldots, x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2) \]

with non-informative “Jeffreys” prior:

\[ p(\beta_0, \ldots, \beta_p, \log \sigma^2) = \text{const.} \]

yields the posterior distribution of \((\beta_0, \ldots, \beta_p)\) as multivariate

\( T \) with mean given by the least squares estimates \((\widehat{\beta}_0, \ldots, \widehat{\beta}_p)\)

covariance matrix \((X^T X)^{-1} s^2\), where \(X\) is the design matrix,

and degrees of freedom \(n - p - 1\).

Resulting posterior probability intervals are equivalent to

standard \( t \) confidence intervals.
Simulating Draws from Posterior Distribution

• With problems with high-dimensional $\theta$, it is often easier to draw values from the posterior distribution, and base inferences on these draws.

• For example, if

$$(\theta^{(d)}_1 : d = 1, \ldots, D)$$

is a set of draws from the posterior distribution for a scalar parameter $\theta_1$, then

$$\bar{\theta}_1 = D^{-1} \sum_{d=1}^{D} \theta^{(d)}_1$$ approximates posterior mean

$$s^2_{\theta} = (D - 1)^{-1} \sum_{d=1}^{D} (\theta^{(d)}_1 - \bar{\theta}_1)^2$$ approximates posterior variance

$$(\bar{\theta}_1 \pm 1.96 s_{\theta})$$ or 2.5th to 97.5th percentiles of draws approximates 95% posterior credibility interval for $\theta$
Example: Posterior Draws for Normal Linear Regression

\((\hat{\beta}, s^2) = \text{ls estimates of slopes and resid variance}\)

\(\sigma^{(d)2} = (n - p - 1)s^2 / \chi^2_{n-p-1}\)

\(\beta^{(d)} = \hat{\beta} + A^T z \sigma^{(d)}\)

\(\chi^2_{n-p-1} = \text{chi-squared deviate with } n - p - 1 \text{ df}\)

\(z = (z_1, ..., z_{p+1})^T, \ z_i \sim N(0,1)\)

\(A = \text{upper triangular Cholesky factor of } (X^T X)^{-1}\):

\(A^T A = (X^T X)^{-1}\)

- Easily extends to weighted regression: see Example 6.19
Likelihood methods with incomplete data

• Statistical model + incomplete data $\Rightarrow$ Likelihood

• Statistical models needed for:
  – data without missing values
  – missing-data mechanism

• Model for mechanism not needed if it is ignorable (to be defined later)

• With likelihood, proceed as before:
  – ML estimates, large sample standard errors
  – Bayes posterior distribution
  – Little and Rubin (2002, chapter 6)
The Observed Data

\[ Y = (y_{ij})_{n \times K} = (Y_{\text{obs}}, Y_{\text{mis}}) \]

\[ M = (m_{ij})_{n \times K} \]

\[ m_{ij} = \begin{cases} 0, & y_{ij} \text{ observed} \\ 1, & y_{ij} \text{ missing} \end{cases} \]
Model for $Y$ and $M$

\[ f(Y, M \mid \theta, \psi) = f(Y \mid \theta) \times f(M \mid Y, \psi) \]

Complete-data model  model for mechanism

Example: bivariate normal monotone data

**complete-data model:**

\[(y_{i1}, y_{i2}) \sim_{iid} N_2(\mu, \Sigma)\]

**model for mechanism:**

\[(m_{i2} \mid y_{i1}, y_{i2}) \sim_{ind} Bern[\Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2})]\]

\[\Phi = \text{Normal cumulative distribution function}\]
Two likelihoods

• **Full likelihood** - involves model for \( M \)

\[
 f (Y_{\text{obs}}, M \mid \theta, \psi) = \int f (Y_{\text{obs}} \mid Y_{\text{mis}}, \theta) f (M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) dY_{\text{mis}} 
\]

\[
 \Rightarrow L_{\text{full}} (\theta, \psi \mid Y_{\text{obs}}, M) = \text{const} \times f (Y_{\text{obs}}, M \mid \theta, \psi)
\]

• **Likelihood ignoring the missing-data mechanism** \( M \)
  • simpler since it does not involve model for \( M \)

\[
 f (Y_{\text{obs}} \mid \theta) = \int f (Y_{\text{obs}} \mid Y_{\text{mis}}, \theta) dY_{\text{mis}} 
\]

\[
 \Rightarrow L_{\text{ign}} (\theta \mid Y_{\text{obs}}) = \text{const} \times f (Y_{\text{obs}} \mid \theta)
\]
Ignoring the missing-data mechanism

• Note that if:

\[ L_{\text{full}}(\theta, \psi | Y_{\text{obs}}, M) = L(\psi | M, Y_{\text{obs}}) \times L_{\text{ign}}(\theta | Y_{\text{obs}}) \]

where \( L(\psi | M, Y_{\text{obs}}) \) does not depend on \( \theta \)
then inference about \( \theta \) can be based on \( L_{\text{ign}}(\theta | Y_{\text{obs}}) \)

• The missing-data mechanism is then called *ignorable* for likelihood inference
Ignoring the md mechanism continued

• It is easy to show that sufficient conditions for ignoring the missing-data mechanism are:
  (A) Missing at Random (MAR):
  
  \[
  f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) = f(M \mid Y_{\text{obs}}, \psi) \text{ for all } Y_{\text{mis}}
  \]

  (B) Distinctness:
  \( \theta \) and \( \psi \) have distinct parameter spaces
  (Bayes: priors distributions are independent)

Proof: 

\[
\begin{align*}
  f(Y_{\text{obs}}, M \mid \theta, \psi) &= \int f(Y_{\text{obs}}, \ Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi) \, dY_{\text{mis}} \\
  &= \overset{(\text{MAR})}{\int} f(Y_{\text{obs}}, \ Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, \psi) \, dY_{\text{mis}} \\
  &= f(M \mid Y_{\text{obs}}, \psi) \times \int f(Y_{\text{obs}}, \ Y_{\text{mis}} \mid \theta) \, dY_{\text{mis}} \\
  &= f(M \mid Y_{\text{obs}}, \psi) \times f(Y_{\text{obs}} \mid \theta)
\end{align*}
\]
Ignoring the md mechanism continued

- If MAR holds but not distinctness, ML based on ignorable likelihood is valid but not fully efficient,

- So MAR is the key condition
Bivariate Normal Monotone Data

\[ L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = \prod_{i=1}^{r} N_2(y_{i1}, y_{i2} \mid \mu, \Sigma) \times (1 - \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2})) \]

\[ \times \prod_{i=r+1}^{n} \int N_2(y_{i1}, y_{i2} \mid \mu, \Sigma) \times \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}) \, dy_{i2} \]

Under MAR: \( \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}) = \Phi(\psi_0 + \psi_1 y_{i1}) \)

\[ L_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) = \prod_{i=1}^{r} N_2(y_{i1}, y_{i2} \mid \mu, \Sigma) \times (1 - \Phi(\psi_0 + \psi_1 y_{i1})) \]

\[ \times \prod_{i=r+1}^{n} N_2(y_{i1} \mid \mu, \Sigma) \times \Phi(\psi_0 + \psi_1 y_{i1}) \]

\[ = L_{\text{ign}}(\theta \mid Y_{\text{obs}}, M) \times \prod_{i=1}^{r} (1 - \Phi(\psi_0 + \psi_1 y_{i1})) \times \prod_{i=r+1}^{n} \Phi(\psi_0 + \psi_1 y_{i1}) \]

4. Likelihood theory
Bayes: add prior distributions

\[
p_{\text{complete}}(\theta, \psi \mid Y, M) = \pi(\theta, \psi) \times f(Y \mid \theta) \times f(M \mid Y, \psi)
\]

Prior dn \text{ Complete-data model} model for mechanism

\[
p_{\text{full}}(\theta, \psi \mid Y_{\text{obs}}, M) \propto \pi(\theta, \psi) \times f(Y_{\text{obs}}, M \mid \theta, \psi)
\]

- **Full** posterior dn - involves model for \( M \)

\[
f(Y_{\text{obs}}, M \mid \theta, \psi) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta) f(M \mid Y_{\text{obs}}, Y_{\text{mis}}, \psi)dY_{\text{mis}}
\]

- Posterior dn **ignoring the missing-data mechanism** \( M \) (simpler since it does not involve model for \( M \))

\[
p_{\text{ign}}(\theta \mid Y_{\text{obs}}) \propto \pi(\theta) \times f(Y_{\text{obs}} \mid \theta)
\]

\[
f(Y_{\text{obs}} \mid \theta) = \int f(Y_{\text{obs}}, Y_{\text{mis}} \mid \theta)dY_{\text{mis}}
\]
Summary of Theory

• Likelihood Inference Ignoring the Missing Data Mechanism is valid if
  – Model for $Y$ is correctly specified
  – Data are MAR
  – Fully efficient if distinctness condition holds

• In contrast, many ad-hoc methods require the stronger MCAR assumption
Missing Data
Roderick Little and Trivellore Raghunathan

Module 5
Maximum Likelihood/Bayes computational tools
Computational tools

• Tools for monotone patterns
  – Maximum likelihood based on factored likelihood
  – Draws from Bayesian posterior distribution based on factored posterior distributions

• Tools for general patterns
  – EM/ECM
  – Bayes based on Gibbs’ sampler (also provides multiple imputations of missing values)

• Illustrate with bivariate normal data
Bivariate Monotone Data

- Maximum Likelihood by factoring the likelihood (Anderson 1957: Little and Rubin 2002, chapter 7):

\[ f(y_1, y_2 | \theta) = f(y_1 | \phi_1(\theta)) \times f((y_2 | y_1, \phi_2(\theta)) \]

\[ \Rightarrow L(\theta | Y_{obs}) = L_1(\phi_1 | \{y_{i_1} : i = 1, ..., n\}) \times L_2(\phi_2 | \{(y_{i_1}, y_{i_2}), i = 1, ... r\}) \]

- Method only works when parameters \( \phi_1 \) and \( \phi_2 \) are distinct

5. Likelihood based methods
Bivariate Normal Data

• For Bivariate Normal data:

\[ \phi = (\phi_1, \phi_2) \]
\[ \phi_1 = (\mu_1, \sigma_{11}) \]
\[ \phi_2 = (\beta_{20.1}, \beta_{21.1}, \sigma_{22.1}) \]

• ML estimates are:

\[ \hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^{n} y_{i1}, \quad \hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^{n} (y_{i1} - \hat{\mu}_1)^2 \]

\[ \hat{\beta}_{21.1} = s_{12} / s_{11} \]
\[ \hat{\beta}_{20.1} = \bar{y}_2 - \hat{\beta}_{21.1} \bar{y}_1 \]
\[ \hat{\sigma}_{22.1} = s_{22.1} = s_{22} - s_{12}^2 / s_{11} \]

\[ \{ \text{Least squares on } r \text{ complete cases} \} \]
Estimates of other parameters

- ML for other parameters by expressing them as functions of \( \phi \) and applying transformation property of ML:

\[
\mu_2 = \beta_{20} + \beta_{21} \mu_1 \\
\Rightarrow \hat{\mu}_2 = \hat{\beta}_{20} + \hat{\beta}_{21} \hat{\mu}_1 = \bar{y}_2 + \hat{\beta}_{21} (\hat{\mu}_1 - \bar{y}_1)
\]

which is known as the **regression estimate** of the mean

\[
\hat{\sigma}_{22} = \hat{\sigma}_{22} + \hat{\beta}^2_{22} \hat{\sigma}_{11} = s_{22} + \hat{\beta}^2_{21} (\hat{\sigma}_{11} - s_{11}) \\
\hat{\sigma}_{12} = \hat{\beta}_{21} \hat{\sigma}_{11} = s_{12} (\hat{\sigma}_{11} / s_{11})
\]
Bayes Inference

• An attractive alternative to ML inference is to draw parameters from Bayes posterior distribution
  – Estimate = posterior mean
  – Standard error = posterior standard deviation
  – Standard errors are often easier than ML!
Normal problem posteriors

- In normal example with reference priors:
  \[ \pi_1(\mu_1, \log \sigma_{11}) \pi_2(\beta_{20,1}, \beta_{21,1}, \log \sigma_{22,1}) \propto \text{const} \]

standard complete-data calculations yield:

\[ \sigma_{11} | data \sim n\hat{\sigma}_{11} / \chi^2_{n-1} \quad \text{scaled inverse chi-squared dn} \]

\[ \mu_1 | \sigma_{11}, data \sim N(\hat{\mu}_1, \sigma_{11} / n) \quad \text{normal (unconditionally } \mu_1 \text{ is } T) \]

\[ \sigma_{22,1} | data \sim rs_{22,1} / \chi^2_{r-2} \]

\[ \beta_{21,1} | \sigma_{22,1}, data \sim N(\hat{\beta}_{21,1}, \sigma_{22,1} / (rs_{11})) \]

\[ \beta_{20,1} | \beta_{21,1}, \sigma_{22,1}, data \sim N(\bar{y}_2 - \beta_{21,1}\bar{y}_1, \sigma_{22,1} / r) \]
Postiors of other parameters

- Postiors of other parameters can be simulated by expressing them as functions of $\phi$ and substituting draws $\phi^{(d)}$ from the posterior distribution of $\phi$:

\[
\begin{align*}
\mu_2 &= \beta_{20:1} + \beta_{21:1} \mu_1 \\
\mu_2^{(d)} &= \beta_{20:1}^{(d)} + \beta_{21:1}^{(d)} \mu_1^{(d)} \\
\sigma_{22}^{(d)} &= \sigma_{22:1}^{2(d)} + \beta_{22:1}^{2(d)} \sigma_{11}^{(d)} \\
\sigma_{12}^{(d)} &= \beta_{21:1}^{(d)} \sigma_{11}^{(d)}
\end{align*}
\]

- Idea is analogous to ML, with draws replacing ML estimates

- S.e.’s are estimated as sample s.d. of the drawn parameters, without the need to invert an information matrix
Monotone Multivariate Normal Data

Regress current on more observed variables using available cases; e.g. 3 variables:

Marginal of $Y_1$

Regression $Y_2$ on $Y_1$

Regression $Y_3$ on $Y_1$ $Y_2$
Summary

• Factoring into conditional distributions is useful for monotone missing data

• Requires distinctness of parameters of distribution. E.g. does not work for bivariate normal with covariance matrix

\[ \Sigma = \begin{pmatrix} \sigma^2 & \rho \sigma^2 \\ \rho \sigma^2 & \sigma^2 \end{pmatrix} \]

Since parameters \( \phi_1, \phi_2 \) are no longer distinct

• In such cases, and when the pattern is non-monotone, iterative algorithms such as EM are needed for ML estimation. Consider these approaches next.
ML for General Bivariate Pattern

Model \((y_{i1}, y_{i2}) \sim N_2(\mu, \Sigma)\)

Mechanism: MAR

\[
L(\theta | Y_{obs}) \propto \prod_{i=1}^{r} N_2(y_{i1}, y_{i2} | \mu, \Sigma) \times \prod_{i=r+1}^{r+m_1} N_1(y_{i1} | \mu_1, \sigma_{11}) \times \prod_{i=r+m_1+1}^{r+m_1+m_2} N_1(y_{i2} | \mu_2, \sigma_{22})
\]

- Likelihood no longer factors!

5. Likelihood based methods
EM algorithm

Choose starting values $\theta^{(0)}$

Current estimate at iteration $t$: $\theta^{(t)}$

Iteration $t+1$

E(xpectation) Step:
Compute $Q(\theta, \theta^{(t)}) \equiv E\left[ \log p(Y_{\text{obs}}, Y_{\text{mis}} | \theta) | Y_{\text{obs}}, \theta^{(t)} \right]$

M(aximization) Step:
Choose $\theta^{(t+1)}$ to maximize $Q(\theta, \theta^{(t)})$ with respect to $\theta$
Beyond EM

- **Generalized** EM algorithms replace maximization in M Step by a step that increases the Q function; examples are:
  - ECM (LR Section 8.5.1) Conditional maximization replaces full maximization when the latter is iterative
  - ECME (LR Section 8.5.2) Some conditional M steps max full likelihood rather than Q function
- PXEM (LR Section 8.5.3) Imbed in larger model to speed up the algorithm
- Hybrid methods (LR Section 8.6) combine EM and Newton or other steps
Standard errors and EM

• Some approaches to computing standard errors are:
  – Compute and invert information matrix at end
  – Supplemented EM algorithm: computes information matrix using EM iterates
  – Apply to Bootstrap samples and build bootstrap distribution of estimates
  – Bayes’ simulation methods (more below)
Bayes/MI for general patterns

- Superior to ML – readily provides estimates of uncertainty
- MI combining rules are based on Bayesian theory of multiple imputation under explicit models
MI approximation to posterior mean

\[
E(\theta \mid Y_{\text{obs}}) = \int E(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}) p(Y_{\text{mis}} \mid Y_{\text{obs}}) dY_{\text{mis}}
\]

\[
\approx \frac{1}{D} \sum_{d=1}^{D} E(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}) = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d,
\]

where \( \hat{\theta}_d \) = is posterior mean from \( d \)th imputed dataset
MI approximation to posterior variance

\[
\text{Var}(\theta \mid Y_{\text{obs}}) = E(\theta^2 \mid Y_{\text{obs}}) - \left( E(\theta \mid Y_{\text{obs}}) \right)^2 \approx \bar{V} + B
\]

\[
\bar{V} = \frac{1}{D} \sum_{d=1}^{D} V_d = \text{within-imputation variance},
\]

\[
V_d = \text{Var}(\theta \mid Y_{\text{obs}}, Y_{\text{mis}}^{(d)}) \text{ is posterior variance from } d\text{th dataset}
\]

\[
B = \frac{1}{D-1} \sum_{d=1}^{D} \left( \hat{\theta}_d - \bar{\theta}_D \right)^2 = \text{between-imputation variance}
\]
Refinements for small $D$

(A): $\text{Var}(\theta \mid Y_{\text{obs}}) \approx \bar{V} + (1 + 1/D)B$

(B) Replace normal reference distribution by t distribution with df

$$\nu = (D - 1) \left(1 + \frac{D\bar{V}}{D + 1B}\right)^2$$

(C) For normal sample with variance based on $\nu_{\text{com}}$ df, replace $\nu$ by

$$\nu^* = \left(\nu^{-1} + \hat{\nu}_{\text{obs}}^{-1}\right)^{-1}, \hat{\nu}_{\text{obs}} = (1 - \hat{\gamma}_D) \left(\frac{\nu_{\text{com}} + 1}{\nu_{\text{com}} + 3}\right)\nu_{\text{com}}$$

$$\hat{\gamma}_D = \frac{(1 + D^{-1})B}{\bar{V} + (1 + D^{-1})B} = \text{estimated fraction of missing information}$$
Improper Multiple Imputation

- The predictive distribution $p(Y_{\text{mis}} | Y_{\text{obs}})$ is usually not easy to draw from directly, even for simple problems.

A simple approach is to draw $Y_{\text{mis}}$ from

$p(Y_{\text{mis}} | Y_{\text{obs}}, \tilde{\theta})$

where $\tilde{\theta}$ is an estimate of $\theta$

This method is called improper, since it does not allow for uncertainty in estimating $\theta$

OK with small amounts of missing data
Example: monotone bivariate normal

For bivariate normal data:

\[
(y_{i2} \mid Y_{\text{obs}}, \hat{\theta}) \sim \text{ind } N(\hat{\beta}_{20} + \hat{\beta}_{21} y_{i1}, \hat{\sigma}_{22}), \text{ so }
\]

\[
\hat{\gamma}^{(d)}_{i2} = \hat{\beta}_{20} + \hat{\beta}_{21} y_{i1} + z_{il} \sqrt{\hat{\sigma}_{22}}
\]

where \( z_{il} \sim N(0, 1), \)

\( \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\sigma}_{22} \) LS estimates from complete cases

This is simply regression imputation with random normal draws for the missing values.

- Generalization to multivariate missing data is easy
Proper MI methods

• Proper (or approximately proper) methods generate a draw of $\theta$ for the $d$th dataset, and then impute missing values conditional on the drawn value of $\theta$

• These methods are better since they propagate the error in estimating $\theta$ in the imputations

For bivariate normal montone data, proper draws for data set $d$ are:

$$\hat{y}_{i2}^{(d)} = \hat{\beta}_{20}^{(d)} + \hat{\beta}_{21}^{(d)} y_{i1} + z_{il} \sqrt{\hat{\sigma}_{22}^{(d)}}$$

where $z_{il} \sim N(0, 1)$, and $$\hat{\beta}_{20}^{(d)}, \hat{\beta}_{21}^{(d)}, \hat{\sigma}_{22}^{(d)}$$ draws from their posterior distribution based on complete cases
Proper method 1: bootstrap ML

Let \( Y_{\text{obs}}^{(\text{boot},d)} \) denote a bootstrap sample of the complete and incomplete cases.

One proper approach is to compute as the \( d \)th draw: \( \hat{\theta}^{(d)} = \hat{\theta}(Y_{\text{obs}}^{(\text{boot},d)}) \)

the ML estimate computed on the bootstrap sample

This method is valid for large samples

Used in the Amelia MI program
Method 2: Gibbs’ Sampler

The best method is to iterate between draws of the missing values and draws of the parameters:

For data set $d$, given current draws $Y_{\text{mis}}^{(dt)}$, $\theta^{(dt)}$

Iteration $t + 1$ consists of:

I (imputation) step: $(Y_{\text{mis}}^{(d,t+1)} | Y_{\text{obs}}, \theta^{(dt)}) \sim p(Y_{\text{mis}} | Y_{\text{obs}}, \theta^{(dt)})$

P (posterior draw) step: $(\theta^{(d,t+1)} | Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)}) \sim p(\theta | Y_{\text{obs}}, Y_{\text{mis}}^{(d,t+1)})$

As $t$ tends to infinity, this sequence converges to a draw from the joint posterior distribution of $(Y_{\text{mis}}, \theta | Y_{\text{obs}})$, as required.

This is one of the first applications of the Gibbs’ sampler (Tanner 1991; Gelman et. al 1996; Schafer 1996).
Logistic regression example revisited

- Imputation Model
  \[ X_{edi} \sim \text{iid } N(\mu_{ed}, \sigma^2); \]
  \[ e=0,1, \quad d=0,1, \quad \text{subject } i \]
- Imputations are draws from the posterior predictive distribution
- Draw \( \sigma^2 \), then \( \mu_{ed} \) and then missing \( X_{edi} \)
Predictive Distributions

- Draw $\sigma^2$
  
  $$\sigma^2 \sim \frac{WSS}{\chi^2_{r-4}}$$

- Draw $\mu_{ed}$

- $WSS =$ Residual sum of squares

- $r_{ed} =$ Number of respondents in cell $ed$

- $\bar{x}_{ed} =$ Mean for cell $ed$

- $\mu_{ed} | X_{obs}, D, E, \sigma^2 \sim N(\bar{x}_{ed}, \sigma^2 / r_{ed})$

- Draw $X_{edi} \sim N(\mu_{ed}, \sigma^2)$

5. Likelihood based methods
5. Likelihood based methods

Histogram of Multiple Imputation Estimates

- 5 Imputations per missing value
- 5 completed Datasets
- Analyze each separately
- Combine using the formulae given earlier
## Coverage and MSE of Various Methods

<table>
<thead>
<tr>
<th>METHOD</th>
<th>COVERAGE (95% Nominal)</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Deletion</td>
<td>94.68</td>
<td>0.0494</td>
</tr>
<tr>
<td>Complete-case</td>
<td>37.86</td>
<td>0.4456</td>
</tr>
<tr>
<td>Weighted Complete-case</td>
<td>97.42</td>
<td>0.0538</td>
</tr>
<tr>
<td>Hot-Deck Single Imputation</td>
<td>90.28</td>
<td>0.0566</td>
</tr>
<tr>
<td>Multiple Imputation</td>
<td>94.56</td>
<td>0.0547</td>
</tr>
</tbody>
</table>
Use of Auxiliary Information in Imputations

• Imputation may involve many more variables though a particular substantive analysis may only use a subset of variables

• Example: Public use data sets or a data set to be used by multiple researchers from different perspectives

• Improve efficiency, reduce bias
### Expanded Simulation Study

- Add auxiliary variable: \( Z \sim N(0,1) \), \( \text{Corr}(Z, X) = \rho \)

<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Efficiency of MI Using Z compared to Ignoring Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.89</td>
<td>1.42</td>
</tr>
<tr>
<td>0.71</td>
<td>1.31</td>
</tr>
<tr>
<td>0.55</td>
<td>1.21</td>
</tr>
<tr>
<td>0.35</td>
<td>1.12</td>
</tr>
<tr>
<td>0</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Chained Equations or Sequential Regression Approach

- Types of variables
  - Continuous
  - Categorical
  - Count
  - Mixed or semi-continuous
  - Ordinal
  - Censored skewed

Completeness Observed variables: $U$

$Y_1, Y_2, Y_3, ..., Y_p$ : Variables with $Y_1$ having the least number of missing values and $Y_p$ the highest.

- Regress $Y_1$ on $U$; Impute missing $Y_1$
- Regress $Y_2$ on $U, Y_1$; Impute missing $Y_2$
- Using the fully Bayesian regression model (draws)
- and so on

$Y_p$ on $U, Y_1, ..., Y_{p-1}$

Iterations

1, 2, 3

www.isr.umich.edu/src/s...
Sequential Regressions

\[ Y_1 | Y_2, Y_3, \ldots, Y_p \]
\[ Y_2 | Y_1, Y_3, \ldots, Y_p \]
\[ \ldots \]
\[ Y_p | Y_1, Y_2, \ldots, Y_p \]

- Linear regression
- Logistic regression
- Poisson regression
- Multinomial Logit regression
- Two stage binary/linear regression model for semi-continuous
- Ordinal

- Add interactions
- Stepwise to reduce computation
- Bounding of imputation
- Skip patterns
- Regression Diagnostics to fine tune each model

Technical Issue:
- There may not be a joint distribution with these conditional distributions
- Does it matter?
• Restrictions
  – Regression model is fitted only to the relevant subset
• Bounds
  – Draws from a truncated distribution from the corresponding regression model
• Stepwise selection possible at each step to save computation time
• Ability to specify individual regression model: Parametric or semi-parametric models
• Models each conditional distribution. There is no guarantee that a joint distribution exists with these conditional distributions
• How many iterations?
  – Empirical studies show that nothing much changes after 5 or 6 iterations
• Several completed data statistics seem to converge to the same value regardless of seeds
Software for Multiple Imputation Analysis

For Creating Imputations

- **SAS**
  - PROC MI
  - IVEware
- **Standalone**
  - SRCware
- **Other options**
  - STATA
    - MI IMPUTE
  - R
    - MICE
  - SOLAS
  - SPSS (limited)

For Analysis of multiply imputed data

- **SAS**
  - PROC MIANALYZE
  - IVEware
- **Standalone**
  - SRCware
- **Other options**
  - STATA
    - MI ESTIMATE
  - SUDAAN
  - R

6. Chained Equation Multiple Imputation
PROC MI

- Multivariate normal for continuous variable
- Flexible continuous specification (FCS) or Sequential regression
  - Linear regression or predictive mean matching for continuous variables
  - Logistic for binary and ordinal logistic for polytomous; discriminant analysis for classification variables
- Several nice options and print out missing data pattern and other descriptive information
- Stores all the completed data sets in a single data set (use the system variable _IMPUTE_ to select one imputed data set)
IVEware (www.iveware.org)

- A routine called within SAS to create multiple imputations
- Sequential Regression Multivariate Imputation or Chained Equations or Fully Conditional Specification
- Variables
  - Continuous (linear)
  - Binary (logistic)
  - Count (Poisson)
  - Polytomous (Multinomial logit)
  - Semi-Continuous (Two stage logistic, linear regression)
- Model building: Minimum additional R-square, Best p predictors, inclusion of interactions
- Bounds: Ensure that the imputed continuous variable falls within a specified range
- Restriction: Subset cases to be imputed
- Data preparation allows creation new variables
- The output data can be stored
IVEware

- **COMBINE**: Designed to pool data from multiple studies.
  - Suppose Study A measures \((X,Y)\), Study B measures \((X,Z)\) and Study C measures \((Y,Z)\). Combining A, B and C creates a concatenated data set with \((X,Y,Z)\)
  - Next into IMPUTE for multiply imputing the missing \(Z\) (in A), \(Y\) (in B) and \(X\) (in C)
  - Analyze completed \((X,Y,Z)\)

- **SYNTESIZE**: Multiple Imputation for Disclosure Limitation

6. Chained Equation Multiple Imputation
SRCware

• IMPUTE is identical to IVEware
• Reads in data from a text file with the variable name in the first row.
  – Specify delimiter between columns
  – Missing value code is “.”
• Output stored as text file with variable names as the first row
Analysis of Multiply Imputed Data

• PROC MIANALYZE
• Analyze each imputed data using various PROCs
  – Output Parameter estimates and Covariance matrix (syntax for doing this varies by PROCs)
• Input the parameter estimates and the covariance matrices
MI Analysis (Contd.)

- Iveware
  - Describe (mean, proportions, contrasts) (Also available in SRCware)
  - Regress (linear, logistic, Poisson, Tobit and Proportional Hazard) (Also available in SRCware)
  - SASMOD (Structural Equation Model, Nonlinear regression, Hierarchical Models, Generalized regression model, contingency table analysis etc)
  - Complex Survey Design (uses Jackknife Repeated Replication Techniques)
Conclusions

- Multiple imputation based on Bayesian principles a powerful tool for missing data problems
- link between multiple imputation and modern tools of stochastic simulation
- other methods (SIR, Metropolis-Hastings) can handle more difficult models
Missing Data
Roderick Little and Trivellore Raghunathan

Module 6
Nonignorable Models
Likelihood methods for non-ignorable models

- Statistical model + incomplete data $\Rightarrow$ Likelihood
- For non-ignorable missing data, need a model for the missing-data mechanism as well as a model for the data
- ML estimates, large sample standard errors
  - tools like factored likelihood, EM apply to these models too
- Bayes posterior distribution
  - stochastic simulation methods apply
- But beware of unidentified or poorly-identified models, leading to numerical problems (e.g. lack of convergence). (Little and Rubin 2002, chapt. 15)
Models for $Y$ and $M$

- Let $(y_i, m_i)$ denote the complete-data vector and missing-data indicator for the $i$th unit, and assume independence across units. Two generic modeling approaches are:

  **Selection models, which factor:**

  $$f(y_i, m_i | \theta, \psi) = f(y_i | \theta) \times f(m_i | y_i, \psi)$$

  - Complete-data model
  - Model for md mechanism

  **Pattern-mixture models, which factor:**

  $$f(y_i, m_i | \phi, \gamma) = f(y_i | m_i, \phi) \times f(m_i | \gamma)$$

  - Model for $y$’s within pattern $m_i$
  - Probability of pattern $m_i$
Selection or Pattern-Mixture Models?

• Selection models are:
  – more natural substantive formulation of model, if inference concerns the entire population
  – more common approach in literature
  – sensitive to specification of the form of the missing-data mechanism, which is often not well understood

• Pattern-mixture models are:
  – More natural when interest is in population strata defined by missing-data pattern
  – closer to the form of the data, sometimes simpler to fit
  – can avoid specifying the form of the md mechanism, which is incorporated indirectly via parameter restrictions.
Example 1: Bivariate Normal

• **Probit selection model** for monotone bivariate normal data:

\[
(y_{i1}, y_{i2}) \sim_{iid} N_2(\mu, \Sigma)
\]

\[
(m_{i2}|y_{i1}, y_{i2}) \sim_{ind} Bern[\Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2})]
\]

\[\Phi = \text{Normal cumulative distribution function}\]

• **Pattern-mixture model**:

\[
(y_{i1}, y_{i2}|m_{i2} = r) \sim_{iid} N_2(\mu^{(r)}, \Sigma^{(r)}) \quad \rightarrow \quad (y_{i1}, y_{i2}) \sim \text{normal mixture}
\]

\[
(m_{i2}) \sim_{ind} Bern[\gamma]
\]

• Models are equivalent assuming MCAR with

\[\psi_0 = \gamma, \psi_1 = \psi_2 = 0, \mu^{(r)} = \mu, \Sigma^{(r)} = \Sigma\]
Key problem: lack of information

Consider the probit selection model:

\[ pr(m_{i2} = 1 \mid y_{i1}, y_{i2}) = \Phi(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}) \]

There is essentially no information in the data about the coefficient \( \psi_2 \) since:
- only cases with \( m_{i2} = 0 \) have \( y_{i2} \) observed

In MAR analyses this problem is avoided by setting \( \psi_2 = 0 \)

For non-MAR models there are two approaches:
(a) make other structural assumptions to identify \( \psi_2 \), e.g. \( \psi_1 = 0 \).
(b) assess sensitivity of inferences to difference choices of \( \psi_2 \)
Pattern-Mixture Models

• In pattern-mixture models lack of information about the mechanism is reflected in unidentified parameters for the incomplete patterns:
e.g. in the bivariate normal model only two of the five parameters \( (\mu^{(1)}, \Sigma^{(1)}) \) are identified, namely \( \mu_{11}^{(1)} \) and \( \sigma_{11}^{(1)} \).

• Assumptions about the mechanism yield restrictions on the parameters that may identify the model
E.g. for the normal model if:
\[
\Pr(m_{i2} = 1 | y_{i1}, y_{i2}) = g(y_{i2}) \quad \text{for any function } g, \text{ then}
\]
\[
y_{i1} \text{ is independent of } m_{i2} \text{ given } y_{i2}
\]
\[
\Rightarrow \beta_{10}^{(0)} = \beta_{10}^{(1)}, \beta_{12}^{(0)} = \beta_{12}^{(1)}, \sigma_{11}^{(0)} = \sigma_{11}^{(1)}
\]
These \( 3 = 5 - 2 \) restrictions yield an identified model.

Nonignorable models
ML for Normal Pattern-Mixture Model

• More generally, if

$$\Pr(m_{i2} = 1 | y_{i1}, y_{i2}) = g(y_{i1} + \lambda y_{i2})$$

for known $\lambda$, the ML estimate of the marginal mean of $Y_2$ is:

$$\hat{\mu}_2 = \bar{y}_2 + \hat{\beta}^{(\lambda)}_{21} (\hat{\mu}_1 - \bar{y}_1), \quad \hat{\beta}^{(\lambda)}_{21} = \frac{s_{12} + \lambda s_{22}}{s_{11} + \lambda s_{12}}$$

where $(s_{jk})$ is the sample cov. matrix from complete cases. Other parameters have similarly simple expressions.

• Extends the regression estimator to non-MAR mechanisms
• Form of mechanism unspecified, yielding robustness
• Again, no information in the data about $\lambda$. We can assess sensitivity to different values of this parameter

Nonignorable models
Ex. 2. Haloperidol Study

- Clinical trial to compare three alternative dose regimens of haloperidol for schizophrenia (Little and Wang 96).
  - 65 patients assigned to receive 5, 10 or 20 mg./day of haloperidol for 4 weeks. $X = 3$ indicators for drug dose (or equivalently two indicators and a constant term).
  - $Y =$ Brief Psychiatric Rating Scale Schizophrenia (BPRSS) factor, measured at baseline, week 1 and week 4. Main interest in mean change in BPRSS between baseline and week 4 for each dose group.
  - 29 patients dropped out of the study before the 4th week for a variety of reasons, including side effects of the drug. Proportions dropping out varied across dose groups, suggesting that missingness was related to dose.
Pattern-mixture model for this example

• Little and Wang (96) assume

\[(m \mid x) \sim \text{Bernoulli}\left(p(x \mid \pi)\right);\]
\[
\log\left(\frac{p(x \mid \pi)}{1 - p(x \mid \pi)}\right) = \pi^T x
\]

• This logistic regression of \(m\) on \(x\) is assumed saturated, that is, the fraction of incomplete cases is estimated independently for each treatment group.

\[(y \mid x, m = k) \sim \text{ind } N_3(B^{(k)} x, \Sigma^{(k)})\]

• Parameters \(B^{(k)}, \Sigma^{(k)}\) are identified by alternative assumptions about the missing-data mechanism...
Four alternative md mechanisms

- The effect of nonignorable nonresponse is assessed by computing estimates under a range of assumptions about the missing-data mechanism. Specifically, missingness is assumed to be an arbitrary function of treatment group and one of the following four linear combinations of $Y$:

\[
\begin{align*}
A. \quad y_A^* &= 0.4Y_1 + 0.4Y_2 + 0.2Y_3 \\
B. \quad y_B^* &= 0.3Y_1 + 0.3Y_2 + 0.4Y_3 \\
C. \quad y_C^* &= 0.1Y_1 + 0.1Y_2 + 0.8Y_3 \\
D. \quad y_D^* &= Y_3
\end{align*}
\]

A is closest to ignorable, where coefficient of $Y_3 = 0$; D is the most extreme departure from ignorable nonresponse.
Comparison of Estimates

- Following table shows estimated difference in mean BPRSS between baseline and week 4 for the 3 dosage groups, for:
  - Complete-Case (CC) analysis.
  - Ignorable ML, where missingness is assumed to depend on the BPRSS scores at baseline and week 1. Standard errors are the sds of estimates from 1000 bootstrap samples.
  - (a) ML for the four pattern-mixture models. Asymptotic standard errors were computed using the SEM algorithm.
  - (b) Bayes for the four pattern-mixture models. Mean and variance of the posterior distributions were simulated via Gibbs sampling. The prior was

\[
p(\pi, \phi) \propto |\Sigma_{11,2}|^{-(p_1+1)/2} |\Sigma_2^{(0)}|^{-1/2} |\Sigma_2^{(1)}|^{-1/2}
\]

- ML estimates for the probit selection model.
## Estimates for Haloperidol Data

<table>
<thead>
<tr>
<th></th>
<th>Treatment Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dose 5</td>
</tr>
<tr>
<td>Sample Size</td>
<td>15</td>
</tr>
<tr>
<td>Fraction missing</td>
<td>.33</td>
</tr>
<tr>
<td>CC Analysis</td>
<td>3.700 (1.027)</td>
</tr>
<tr>
<td>Ignorable ML</td>
<td>3.291 (0.897)</td>
</tr>
<tr>
<td>Pattern-Mixture ML</td>
<td></td>
</tr>
<tr>
<td>Mechanism A</td>
<td>3.276 (0.898)</td>
</tr>
<tr>
<td>Mechanism B</td>
<td>3.251 (0.909)</td>
</tr>
<tr>
<td>Mechanism C</td>
<td>3.181 (0.945)</td>
</tr>
<tr>
<td>Mechanism D</td>
<td>3.140 (0.968)</td>
</tr>
<tr>
<td>Probit Selection ML</td>
<td>3.345 (1.027)</td>
</tr>
</tbody>
</table>

Nonignorable models
## Estimates for Haloperidol Data

### (contd)

#### Treatment Group

<table>
<thead>
<tr>
<th></th>
<th>Dose 5</th>
<th>Dose 10</th>
<th>Dose 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size</td>
<td>15</td>
<td>34</td>
<td>16</td>
</tr>
<tr>
<td>Fraction missing</td>
<td>.33</td>
<td>.41</td>
<td>.63</td>
</tr>
<tr>
<td>CC Analysis</td>
<td>3.700 (1.027)</td>
<td>4.350 (0.726)</td>
<td>5.667 (1.326)</td>
</tr>
<tr>
<td>Ignorable ML</td>
<td>3.291 (0.897)</td>
<td>4.087 (0.618)</td>
<td>6.463 (1.044)</td>
</tr>
<tr>
<td><strong>Pattern-Mixture Bayes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanism A</td>
<td>3.229 (0.985)</td>
<td>4.070 (0.710)</td>
<td>6.464 (1.185)</td>
</tr>
<tr>
<td>Mechanism B</td>
<td>3.212 (1.016)</td>
<td>4.133 (0.717)</td>
<td>6.559 (1.221)</td>
</tr>
<tr>
<td>Mechanism C</td>
<td>3.133 (1.117)</td>
<td>4.226 (0.772)</td>
<td>6.812 (1.393)</td>
</tr>
<tr>
<td>Mechanism D</td>
<td>3.075 (1.187)</td>
<td>4.258 (0.820)</td>
<td>6.964 (1.526)</td>
</tr>
<tr>
<td>Probit Selection ML</td>
<td>3.345 (1.027)</td>
<td>4.155 (0.560)</td>
<td>6.586 (1.194)</td>
</tr>
</tbody>
</table>

Nonignorable models
Haloperidol Data Findings

• Complete-case estimates deviate noticeably from estimates from the other methods

• Estimates for the probit selection model are similar to those for the ignorable selection model and pattern mixture model with mechanism (A).
  – Bootstrap distributions of the estimated coefficients for selection were very dispersed, confirming that the ability to estimate these parameters simultaneously from the data is very limited.
  – This approach is not recommended
Haloperidol Findings continued

- ML and Bayes estimates for pattern-mixture models are similar.
- Size of treatment effects is only moderately sensitive to the choice of pattern-mixture model (ignorable, A-D):
  - As missingness becomes increasingly dependent on the missing Week 4 BPRSS value, differentials in treatment effects by size of dose increase slightly
- Asymptotic se’s are a bit smaller than the posterior standard errors; the latter are superior in that they do not rely on quadratic approximations to the log likelihood
- Se’s of the pattern-mixture model estimates increase from models A though D, reflecting a loss of information with increasing degree of non-ignorable nonresponse
Ex 3: A simple sensitivity analysis for Clinical Trials

- Recent National Academy report on missing data in clinical trials recommends sensitivity analysis
- Parameters of MNAR models cannot be reliably estimated – identifiability requires structural assumptions that are often questionable
- Varying certain parameters in a sensitivity analysis is the preferred approach
- In many (not all) situations, it would be reasonable to choose an MAR primary model, and look at MNAR models via a sensitivity analysis to assess plausible deviations from MAR

Nonignorable models
Simple example

Consider first two treatments $T = 1,2$, single outcome $Y$, no auxiliary data
Let $M = 0$ if $Y$ is observed, $M = 0$ otherwise
Problem is to estimated mean in each treatment arm based on data from subjects in each arm, where some of the $Y$'s are missing.
MAR would assume that within each treatment arm, the distribution of $Y$ for respondents was the same as for nonrespondents.
Simple example

For treatment $t$, let

$$\mu_{0t} = E(Y \mid T = t, M = 0), \mu_{1t} = E(Y \mid T = t, M = 1)$$

Goal is to estimate $\mu_t = \pi_t \mu_{0t} + (1 - \pi_t) \mu_{1t}$

There is information about respondent means

$\{\mu_{0t}\}$ and response rates $\{\pi_t\}$

No information about nonrespondent means $\{\mu_{1t}\}$
Simple pattern-mixture model for sensitivity analysis

\[ Y \mid M = 0, T = t \sim N(\mu_{0t}, \sigma_t^2) \]

\[ Y \mid M = 1, T = t \sim N(\mu_{0t} + \Delta_t, \sigma_t^2) \]

\[ \Pr(M = 0 \mid T = t) = \pi_t \]

MAR: \( \Delta_1 = \Delta_2 = 0 \)

MNAR: For prespecified plausible values of \( \Delta_1, \Delta_2 \):

generate inferences about \( \mu_1, \mu_2 \) and \( \mu_2 - \mu_1 \)

by replacing \( \mu_{0t}, \sigma_t^2 \) and \( \pi_t \) by sample estimates, with associated standard errors.

Examples of choices of \( \Delta_1, \Delta_2 \):

\( \Delta_j = k\sigma_j, \; k = 0.2, 0.5 \)

Nonignorable models
Variants and extensions

(A) Assume $\mu_{1t} = g^{-1}\{g(\mu_{0t}) + \Delta_t\}$,
where $g$ is a link function suitable for the outcome, e.g. logit for binary outcome

(B) Include covariates and auxiliary variables, $Z$:

$[Y_i \mid M_i = 0, T_i = t, z_i] \sim N(\alpha_{0t} + \beta_t z_i, \sigma^2_j)$

$[Y_i \mid M_i = 1, T_i = t, z_i] \sim N(\alpha_{0t} + \Delta_t + \beta_t z_i, \sigma^2_j)$

Multiple imputation is a convenient way of marginalizing over auxiliary $z_i$.
Setting $\Delta_t = k\sigma_t$, good covariates are "rewarded" by a reduction in the size of $\sigma_j$, and hence $\Delta_j$
Possible choices of $k$ are 0.2, 0.5,...
Sensitivity analysis for selection models

- For selection models, a sensitivity analysis can be conducted by making an assumption about how the odds of nonresponse change with the values of the outcome $Y$.

- For example, one can assume that the log odds of nonresponse differs by $\alpha$ for those who differ by one unit on $Y$, that is

$$\text{logit} \{P[M = 1 | V, Y = y]\} = h(V) + \alpha y.$$ 

This is a selection model since it models the probability of nonresponse as a function of the outcome and auxiliary variables $V$. 

Nonignorable models
Sensitivity analysis for selection models

Adopting a value of $\alpha$ corresponds to adopting a known link between the distribution of the respondents and the nonrespondents.

A sensitivity analysis consists of repeating the inference for $\mu$ at different plausible values of $\alpha$ so as to assess the sensitivity of inferences about $\mu$ to deviations from MAR.

I prefer the pattern-mixture approach since it is easier to implement and explain to clinicians, and $\alpha$ has a tricky interpretation.
Summary

- Sensitivity analysis is a scientific way of attempting to reflect uncertainty arising from potentially MNAR missing data.
- Deciding on how to implement and interpret a sensitivity analysis in the regulatory setting is challenging.
- The need and importance of sensitivity analysis increases with the amount of potentially MNAR missing data.
- This reinforces the need to limit missing data in the design and implementation stage.
  - Avoiding substantial amounts of missing data is key!
Conclusions

• Nonignorable mechanisms can be included in a missing-data analysis, but this is a difficult modeling problem
• Often little is known about the missing-data mechanism, and results may be sensitive to formulation
• Parameters of missing-data are often unidentified or weakly-identified from the data ...
• As a result, it may be more appropriate to do a sensitivity analysis, fixing weakly identified parameters at different values.
• Software for fitting non-ignorable models is not widely available – pattern-mixture approach is
• Design to avoid nonignorable missing data is preferable if possible