Statistical Methods for Correlated Data

Handling Missing Data

Michele Guindani

Department of Biostatistics
UCLA

Learning goals

- Taxonomy of the mechanisms by which there are missing data in the response
- Consequences of Missing Data Mechanisms
- Dropout - effect of dropout

Missing data (MD)

- MD are to be expected in LDA where non-responses can occur at any occation
- The term non-response here denotes that an *intended* measurement could not be obtained:
  - In LDA, an individual's response can be missing at one follow-up time but be measured at a later follow-up ⇛ various missingness patterns
  - LD studies also suffer from "attrition" or "drop-out": some individuals withdraw from the study before its intended completion
- When LD data are incomplete, there are important implications for their analysis

Two major consequences of Missingness patterns

- On the one hand, with missing data, there is necessary a loss of information and a reduction in precision of the methods of analysis
  - The reduction in precision is directly related to the amount of missing data and depends on the method of analysis
- More importantly, there is a concern for bias. Missing data can introduce bias and thereby lead to misleading inferences about changes in the response over time

Two major consequences of Missingness patterns

- On the one hand, with missing data, there is necessary a loss of information and a reduction in precision of the methods of analysis
  - The reduction in precision is directly related to the amount of missing data and depends on the method of analysis
- More importantly, there is a concern for bias. Missing data can introduce bias and thereby lead to misleading inferences about changes in the response over time
- What methods of analysis are most robust to the different patterns of missingness in the data?

Missing-data mechanisms

- The answer is complicated and depends on the missing-data mechanism
- "Missing data mechanism" $\rightarrow$ probability model for the distribution of the set of response indicator variables (1 if response was measured; 0 if response is missing)
- In short, the validity of any method of analysis will require that certain assumptions about the missing-data mechanism (MDM) (i.e., the reasons for missingness)

Missing-data mechanism

- In general, the key issue is whether the reasons for missingness are related to the outcome of interest.
  - When missingness is unrelated to the outcome, the impact of missing data is not much consequential
  - When the missingness is related to the outcome, care is needed since there is potential for bias if the individual with missing data differ in important ways from those with complete data
- The taxonomy of missing-data mechanisms has been introduced originally by D. Rubin (1976). These missing-data mechanisms differ in terms of assumptions about whether missingness is related to observed and unobserved responses.

Notation

- We intend to take n repeated measures of the response variable on the same individual.
- $n \times 1$ response vector for individual i :

$$Y_i = (Y_{i1} \quad Y_{i2} \quad \cdots \quad Y_{in})'$$

- $\cdot n \times p$ matrix of covariates for individual i (each row corresponds to covariates at one occasion):

$$X_i = \begin{pmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{in_n} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in,1} & X_{in,2} & \cdots & X_{inp} \end{pmatrix}$$

Notation

- A few remarks:
  - In the following, we focus exclusively on missingness in the response; however, missingness can also arise in covariates, and raises similar considerations.
  - The so-called "complete data" are the $n \times 1$ vector of *intended* responses (and the covariates)
  - Since we do not consider missingness in the covariates, we assume that any time-varying covariates are fixed by design

Notation

- $n \times 1$ vector of response indicators for individual i

$$R_i = \begin{pmatrix} R_{i1} & R_{i2} & \cdots & R_{in} \end{pmatrix}$$

where

$$R_{ij} = \begin{cases} 1 \text{ if } Y_{ij} \text{ is observed} \\ 0 \text{ if } Y_{ij} \text{ is missing} \end{cases}$$

- We can then partition the complete set of responses, Y_i into two components:

$Y_i^o =$ vector of observed responses on the $i-$ th subject
$Y_i^M =$ vector of responses on the i^{th} subject that would have been observed, but are missing

Hierarchy of missing-data mechanisms

- A hierarchy of three different types of missing-data mechanisms can be distinguished by considering how $\boldsymbol{R}_i$ is related to $\boldsymbol{Y}_i$:
  - a) missing completely at random (MCAR)
  - b) missing at random (MAR)
  - c) not missing at random (NMAR)
- The type of MDM determines the appropriateness of different methods of LDA

Missing completely at random (MCAR)

- Missingness mechanism for a response is independent of all values of the response, including the unobserved values:

$$\Pr\left(\boldsymbol{R}_i | \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) = \Pr\left(\boldsymbol{R}_i\right)$$

  i.e, LD are MCAR when $\boldsymbol{R}_i$ is independent of both $\boldsymbol{Y}_i^o$ and $\boldsymbol{Y}_i^m$, the observed and unobserved components of $\boldsymbol{Y}_i$

- As such, missingness in $\boldsymbol{Y}_i$ is simply the result of a chance mechanism that does not depend on observed or unobserved components of $\boldsymbol{Y}_i$. The observed values can be viewed as a random sample of the complete data

A Remark

- In the statistical literature there does not appear to be universal agreement on whether the definition of MCAR also assumes no dependence of missingness on the covariates, X_i
- Little (1995) considers *covariate-dependent missingness* as:

$$\Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) = \Pr\left(\boldsymbol{R}_i|X_i\right)$$

  but this terminology is not universal, so pay attention to the definition in each study
- Note that the probability model for the missing data

$$\begin{aligned}\Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) &= \frac{\Pr\left(\boldsymbol{R}_i, \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)} \\ &= \frac{\Pr\left(\boldsymbol{R}_i|X_i\right)\cancel{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}}{\cancel{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}}\end{aligned}$$

  think of side effects in clinical trials: they usually affect $\Pr\left(\boldsymbol{R}_i|X_i\right)$ but not $\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)$.

A Remark

- In the statistical literature there does not appear to be universal agreement on whether the definition of MCAR also assumes no dependence of missingness on the covariates, X_i
- Little (1995) considers *covariate-dependent missingness* as:

$$\Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) = \Pr\left(\boldsymbol{R}_i|X_i\right)$$

  but this terminology is not universal, so pay attention to the definition in each study
- Note that the probability model for the missing data

$$\begin{aligned}\Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) &= \frac{\Pr\left(\boldsymbol{R}_i, \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)} \\ &= \frac{\Pr\left(\boldsymbol{R}_i|X_i\right)\cancel{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}}{\cancel{\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)}}\end{aligned}$$

  think of side effects in clinical trials: they usually affect $\Pr\left(\boldsymbol{R}_i|X_i\right)$ but not $\Pr\left(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m|X_i\right)$.
- A subtle, but important, issue arises. The conditional independence of $\boldsymbol{Y}_i$ and $\boldsymbol{R}_i$, given X_i, may not hold when conditioning on only a subset of the covariates. Consequently, when an analysis is based on a subset of X_i that excludes a covariate predictive of $\boldsymbol{R}_i, \boldsymbol{Y}_i$ is no longer unrelated to $\mathbf{R}_i$.

MCAR Examples

- Samples are lost or destroyed.
- Missing laboratory values because of a batch of lab samples was improperly processed
- "Rotating panel" study design
  - Individuals rotate in and out of the study after providing a pre-determined number of repeated measurements
  - Commonly used in health surveys to reduce response burden
  - Missingness mechanism is under control of the investigator and determined a priori
- In a voter choice survey, independents may be less likely to answer a vote choice question (and party affiliation is measured).
  - Example of covariate-dependent missingness (some consider this MAR)
  - Note that if party affliation had not been measured and included in the model, then this missingness is no longer MCAR!

MCAR analysis

- The essential feature of MCAR is that the observed data can be thought of as a random sample of the complete data

☞ All moments, and even the joint distribution, of the observed data do not differ from the corresponding moments or joint distribution of the complete data.

  ☞ The "completers" (subjects with no-missing data) can be regarded as a random sample from the target population
  ☞ any method of analysis that yields valid inferences w/o missing data will also yield valid (albeit inefficient and wasteful in terms of sample-size) inferences when the analysis is restricted to the "completers" only (complete-case analysis)

- The distributions of $\boldsymbol{Y}_i^m$ and $\boldsymbol{Y}_i^o$ coincide with the distribution of $\boldsymbol{Y}_i$
  - Hence, all available data can be used to obtain valid estimates of moments such as means, variances, and covariances

Missing at Random (MAR)

- The Missingness mechanism for a response is only a function of the observed responses, but independent of specific missing values that would have been obtained (given covariates):

$$\Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i\right) = \Pr\left(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, X_i\right)$$

  i.e. longitudinal data are MAR when $\boldsymbol{R}_i$ is conditionally independent of $\boldsymbol{Y}_i^m$, given $\boldsymbol{Y}_i^o$

- Since the missing-data mechanism depends upon $\boldsymbol{Y}_i^o$, the distribution of $\boldsymbol{Y}_i$ in each of the distinct strata defined by the patterns of missingness is not the same as the distribution of $\boldsymbol{Y}_i$ in the target population

- Missing at random means there might be systematic differences between the missing and observed values, but these can be entirely explained by other observed variables.

MAR examples

- Study protocol requires that a subject be removed from the study whenever the value of the response falls outside a certain clinical range of values
  - Missingness in $\boldsymbol{Y}_i$ is under the control of the investigator and is related to observed components of $\boldsymbol{Y}_i$ only.
- If blood pressure data are missing at random, conditional on age and sex, then the distributions of missing and observed blood pressures will be similar among people of the same age and sex (e.g. within age/sex strata).
- Another example where missing data are MAR is in the Six Cities Study of Air Pollution and Health, when children moved out of the school district because they developed respiratory problems. If the decision to relocate could be predicted based only on the recorded history of pulmonary function measurements (i.e., the observed components of Y_i only), then the missing data are MAR. However, the MAR assumption would not hold if the decision to relocate was based on some extraneous variable, unavailable to the investigators, that was predictive of the future but unobserved, pulmonary function measurements.

Implications of MAR mechanism

- Because the missing data mechanism now depends on Y_i^O, the distribution of Y_i in each of the distinct sub-populations defined by the missing data patterns is not the same as the distribution of Y_i in the target population.

☞ the "completers" are a biased sample from the target population; consequently, an analysis restricted to the "completers" is not valid (no valid complete-case analysis).

☞ the sample means, variances, and covariances based on either the "completers" or the available data are biased estimates of the corresponding moments in the target population.

☞ With MAR, the observed data cannot be viewed as a random sample of the complete data.

Ignorability

- If the MDM is MAR, missing values can be validly predicted using the observed data and a valid model for the joint distribution of Y_i:

$$\begin{aligned} P(Y_i, R_i|X_i) &= P(Y_i^o, Y_i^m, R_i|X_i) \\ &= P(R_i|X_i, Y_i^o, Y_i^m) \times P(Y_i^o, Y_i^m|X_i) \\ &= P(R_i|X_i, Y_i^o) \times P(Y_i^o, Y_i^m|X_i) \end{aligned}$$

☞ no need to use the model for $Pr(R_i|Y_i, X_i)$ to obtain valid likelihood-based inferences, as long as we have a model for Y_i given X_i.

- Notice that not using $\Pr(R_i|Y_i, X_i)$ in the analysis has the important implication that we do not need to even posit a specific model for $\text{Prx}(R_i|Y_i, X_i)$ other than to say it does not depend on the missing observations.
- Since MCAR is a special case of MAR, the same is also true of MCAR.

$$P(Y_i^o, R_i \mid X_i) = P(R_i \mid X_i) \cdot P(Y_i^o \mid X_i)$$

Ignorability

- Specifically, Rubin (1976) showed that likelihood-based inferences can be based on the likelihood ignoring the missing-data mechanism, obtained by integrating the missing responses from the joint distribution, $f\left(\boldsymbol{Y}_i|X_i,\boldsymbol{\gamma}\right)$:

$$L\left(\gamma;Y_i^o,X_i\right) = \text{ constant } \times \prod_{i=1}^{N} \int f\left(\boldsymbol{Y}_i^o,\boldsymbol{Y}_i^m|X_i,\boldsymbol{\gamma}\right) d\boldsymbol{Y}_i^m$$

  Thus, when data are MAR, the missing values can be "predicted" from $P(Y^m|Y^o,X_i)$ using the observed data and a model for the joint distribution of Y_i.

- However, the validity of the predictions of the missing values rests upon correct specification of the entire joint distribution.

- Since it is common to use a model for $f\left(Y_i|X_i\right)$, valid likelihood-based analyses can be obtained with MAR or MCAR data with no extra assumptions, other than the general statement of MCAR or MAR.

Ignorability

- For this reason MCAR and MAR are often referred to as ignorable mechanisms.

☞ it does not mean we can ignore the missing data problem and use complete-case/ available-data analysis.

- Instead, once we establish that $\Pr(R_i|Y_i, X_i)$ does not depend on missing observations, we can ignore $\Pr(R_i|Y_i, X_i)$ and obtain a valid likelihood-based analysis (under correct model for $f(Y_i|X_i)$).
- In contrast, the GEE methods require a model for the mean response but do not specify the multivariate joint distribution for the response vector. As a result standard GEE methods do not provide valid estimates of the regression parameters when data are MAR but not MCAR.
- However, GEE estimators of can be adapted to provide a valid analysis by explicitly modeling $Pr(R_i|Y_i, X_i)$ and weighting the analysis accordingly (see later).

Not missing at random

- Missing data are said to be not missing at random when the probability that responses are missing is related to the specific values that should have been obtained, in addition to the ones actually obtained, i.e. $\Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, X_i)$ depends on at least some components of $\boldsymbol{Y}_i^m$
- The NMAR mechanism is often referred to as non-ignorable missingness because the missing-data mechanism cannot be ignored when the goal is to make inferences about the distribution of the complete longitudinal responses.

☞ The likelihood-based methods can no longer factorize the individual likelihood contributions

☞ All standard methods for LDA are not valid

☞ Any valid inferential method under NMAR requires specification of a model for the missing-data mechanism.

NMAR examples

- Refusal to answer because of unusual value of variable (e.g., high/low socioeconomic status)
- Censored data, such as all autism counts < 5 reported as NA or unobserved age at death for living subjects at study's end
- Quality of life surveys: if the QoL is bad, people may not even be able to complete a survey

Testing MCAR

- Little's MCAR test (1988) has been used to test the assumption of MCAR *vs* MAR for multivariate, partially observed quantitative data.
- Some notation (see Little, 1988, JASA):
- Let $\mathbf{y}_i = (1 \times p)$ vector of values for case i, in the absence of missing data.
- $\mathbf{r}_i = (1 \times p)$ vector of missing-data indicators for case i.
- J = number of distinct missing-data patterns $\mathbf{r}_i$ in the data set. Fully observed cases, if present, count as a pattern.
- S_j = set of cases with missing-data pattern j $(j = 1, \ldots, J)$.
- m_j = number of cases in S_j; $\Sigma m_j = n$.
- p_j = number of observed variables for cases in S_j.
- $\mathbf{D}_j = (p \times p_j)$ matrix indicating which variables are observed for pattern j. The matrix has one column for each variable present, consisting of $(p-1)$zero's and one 1 corresponding to the variable identified.

Little's MCAR test

- Let: $\mathbf{y}_{\text{obs},i} = (1 \times p_j)$ vector of values of observed variables in case i.
- $\overline{\mathbf{y}}_{\text{obs,j}} \equiv m_j^{-1} \sum_{i \in S_j} \mathbf{y}_{\text{obs},i} = (1 \times p_j)$ vector of means of observed variables for pattern j
- $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ $1 \times p$ population mean vector and $p \times p$ covariance matrix of $\mathbf{y}_i$.
- $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$ = ML estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, assuming the $\mathbf{y}_i$ are iid normal and the missing-data mechanism is ignorable.
- $\tilde{\boldsymbol{\Sigma}} = n\hat{\boldsymbol{\Sigma}}/(n-1)$, the ML estimate of Σ with a correction for degrees of freedom
- $\boldsymbol{\mu}_{\text{obs},j} \equiv \boldsymbol{\mu}\mathbf{D}_j$ $(1 \times p_j)$ vector of means of observed variables in pattern j
- $\sum_{\text{obs},j} \equiv \mathbf{D}_j^T \boldsymbol{\Sigma} \mathbf{D}_j$ $p_i \times p_i$ covariance matrix of observed variables in pattern j.

- Assume for simplicity Σ is known. Let $\boldsymbol{\mu}^*$ denote the ML estimate of μ; assuming the missing data are MAR and known Σ, and let $\mu^*_{\text{obs },j} = \mu^* \mathbf{D}_j$
- For the MAR assumption consider the statistics:

$$d_0^2 = \sum_{j=1}^{J} m_j \left(\overline{\mathbf{y}}_{\text{obs},j} - \boldsymbol{\mu}^*_{\text{obs},j}\right) \boldsymbol{\Sigma}^{-1}_{\text{obs},j} \left(\overline{\mathbf{y}}_{\text{obs},j} - \boldsymbol{\mu}^*_{\text{obs},j}\right)^T$$

- Suppose that $\mathbf{y}_i$ is multivariate normal distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. If the data are **MCAR**, then conditional on $\mathbf{r}_i$,

$$(\mathbf{y}_{\text{obs},i} | \mathbf{r}_i) \sim_{\text{ind}} N\left(\boldsymbol{\mu}_{\text{obs},j}, \boldsymbol{\Sigma}_{\text{obs},j}\right), \quad i \in S_j, 1 \leq j \leq J$$

- If the data are not MCAR then the means of the observed variables can vary across the patterns, suggesting the alternative model:

$$(\mathbf{y}_{\text{obs},i} | \mathbf{r}_i) \sim_{\text{ind}} N\left(\boldsymbol{\upsilon}_{\text{obs},j}, \boldsymbol{\Sigma}_{\text{obs},j}\right), \quad i \in S_j, 1 \leq j \leq J$$

where $\{\boldsymbol{\upsilon}_{\text{obs },j}, j = 1, \ldots, J\}$ are $(1 \times p_i)$ vectors of mean parameters for observed variables that (unlike $\boldsymbol{\mu}_{\text{obs.j}}$) are distinct for each pattern j.

- Roderick Little (1988) shows that d_0^2 is the likelihood ratio statistic for testing the first model against the second one
- Under the null hypothesis of MCAR d_0^2 has a chi-squared distribution with $f = \Sigma p_j - p$ df.
- If the data are MCAR and y_i has any distribution with mean μ and covariance matrix Σ, d_0^2 is asymptotically chi- squared with f df. That is, for large samples the assumption of normality can be relaxed.

Testing MCAR

- With regard to MCAR, it is also possible to evaluate whether the various patterns of missing data are consistent with sampling from a single normal population.
- This can be done by testing homogeneity of means, covariances, or homogeneity of both means and covariances (Kim and Bentler, 2002).
- What is often done: run t-tests and chi-square tests between R_i and other variables in the data set to see if the missingness on this variable is related to the values of other variables (e.g. t-test to check if there is a higher percentage of missingness in one group vs another group, males vs females, etc).

Dropout

- Most longitudinal studies are designed to collect data on every individual in the sample at a planned sequence of occasions.
- However, longitudinal studies habitually suffer from the problem of attrition; that is, some individuals "dropout" of the study prematurely
- The term dropout refers to the special case where if Y_{ik} is missing, then $Y_{ik+1}, \ldots, Y_{in}$ are also missing.
- Alternatively, when expressed in terms of the response indicators, dropout refers to the case where if $R_{ik} = 0$ then $R_{ik+1} = \cdots = R_{in} = 0$ ("monotone" missing data pattern)

Dropout

- When there is dropout in a longitudinal study, the key issue is whether those who "drop out" and those who remain in the study differ in any further relevant way.
- If they do not, then analyses restricted to those remaining in the study yield valid, albeit inefficient, inferences.
- If they do differ, then such "complete-case" analyses are potentially biased.
- Dropout can be completely at random, at random, or not at random.

Dropout

- When dropout is completely at random the probability of dropout at each occasion is independent of all past, current, and future outcomes (given the covariates). With completely random dropout, an individual leaves the study by a process unrelated to that individual's outcomes.
- In contrast, when dropout is at random, the probability of dropout at each occasion can depend on the previously observed outcomes up to, but not including, the current occasion. However, given the observed outcomes, dropout is assumed to be independent of the current and future outcomes.
- That is, with random dropout the process can depend on the outcomes that have been observed in the past, but given this information, it is unrelated to all future (unobserved) values of the outcome variable following dropout.

Dropout

- When dropout is not at random, the probability of dropping out at each occasion can depend on current and future unrecorded values of the outcome variable that would have been observed had the individual remained in the study
- In the context of dropout in a longitudinal study, the term "informative" dropout often is used to refer to dropout that is NMAR: informative about the distribution of future observations.
- For example, consider two subjects with the same past history of responses (and covariates) up to time t. One drops out and the other does not. With MAR, their future observations have the same distribution. In contrast, dropout that is NMAR informs us that the distributions of the future observations will differ.
- In the NMAR case, nothing in the data can be used to determine the distribution of the future observations of the dropouts; hence the analysis depends strongly on the specification of $\Pr(R_i|Y_i, X_i)$

Nonignorable dropout examples

- If the repeated measures are of pain and drop-out depends on the value of the pain variable at the time of drop-out, that would be outcome-dependent, because missingness then would depend on the (unobserved) value of Y at the time of drop-out.
- People who have more rapid decline in a health outcome tend to drop out more frequently than those with less rapid decline, dropping out depends on this underlying, unobserved slope.

Illustration of the effect of dropout

Suppose data are generated from a multivariate normal with mean

$$E\left(Y_{it}\right) = \mu_{it} = \beta_1 + \beta_2 t$$

and covariance

$$\text{Cov}\left(Y_{is}, Y_{it}\right) = \rho^{|s-t|}, \text{ for } \rho \geq 0$$

with $\beta_1 = 5$, $\beta_2 = 0.25$ and $\rho = 0.7$.

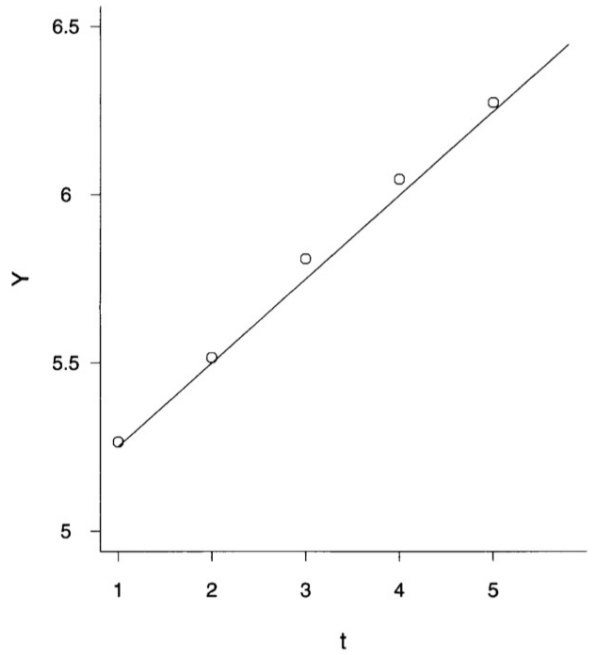

Fig. 17.2 Population regression line and empirical means at each occasion for simulated complete data.

Suppose that there is dropout. We replace the vector of response indicators, $R_{it}(t = 1, \ldots, 5)$, with a simple dropout indicator variable, D_i, for each individual. The random variable D_i is recorded for all individuals and $D_i = k$ if an individual drops out between the $(k-1)^{th}$ and k^{th} occasion; that is, only the first $D_i - 1$ responses are observed. Assume:

$$\log\left\{\frac{\Pr\left(D_i = k | D_i \geq k, Y_{i1}, \ldots, Y_{ik}\right)}{\Pr\left(D_i > k | D_i \geq k, Y_{i1}, \ldots, Y_{ik}\right)}\right\} = \theta_1 + \theta_2\left(Y_{ik-1} - \mu_{ik-1}\right) + \theta_3\left(Y_{ik} - \mu_{ik}\right)$$

- The previous model specifies that the probability of dropout at any occasion, given dropout has not previously occurred, can depend on the current value and the prior value of the response variable (relative to its mean)
- If the MD generating mechanism for the dropout is
  (a) MCAR: $\theta_2 = \theta_3 = 0$
  (b) MAR: $\theta_3 = 0$.
  (c) NMAR: $\theta_3 \neq 0$.

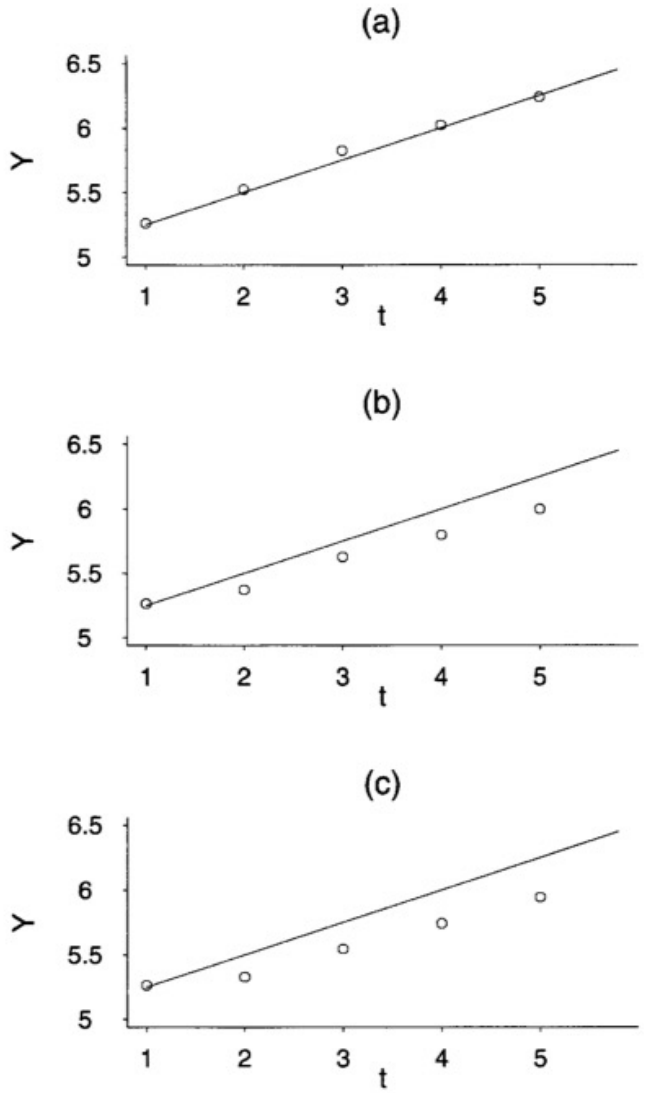

Fig. 17.3 Population regression line and observed data means at each occasion for simulated data when dropout is (a) completely at random (MCAR), (b) at random (MAR), and (c) not at random (NMAR).

Table 17.2 Parameter estimates and standard errors for correctly specified likelihood analysis (ML) and "working independence" analysis (OLS/GEE) based on simulated data when dropout is (a) completely at random, (b) at random, and (c) not at random. The true regression parameters are $\beta_1 = 5.0$ and $\beta_2 = 0.25$.

		ML		OLS/GEE	
Dropout	Parameter	Estimate	SE	Estimate	SE[a]
MCAR	Intercept	5.015	0.031	5.022	0.032
	t	0.257	0.016	0.253	0.018
MAR	Intercept	5.003	0.041	5.062	0.043
	t	0.261	0.016	0.182	0.018
NMAR	Intercept	5.058	0.040	5.071	0.043
	t	0.201	0.016	0.162	0.018

[a] Standard errors for OLS/GEE are based on sandwich variance estimator.

Statistical Methods for Correlated Data

Methods for handling data missing at random

Michele Guindani

Department of Biostatistics
UCLA

Learning goals

- How to handle MAR
  - Inverse Probability Weighted Augmented Estimators
  - Likelihood-based methods
- MNAR: ideas

Multiple Imputations (MI)

- MI assumes the data to come from a continuous multivariate distribution and contain missing values that can occur for any of the variables.
- Methodologically, its development is motivated by Bayes' theory (Rubin, 1987).
- In this approach, a normal predictive distribution of the missing outcome variable is introduced in the analysis of a complete dataset.
- Let $\hat{\theta}$ be a complete-data estimate of a population quantity θ. In standard asymptotic, $\hat{\theta} \sim N(\theta, V)$.
- With a well-assumed prior distribution of missing data, the posterior of θ is

$$P\left(\theta|Y_{obs}\right) = \int P\left(\theta|Y_{obs}, Y_{mis}\right) P\left(Y_{mis}|Y_{obs}\right) dY_{mis}$$

- Given the Bayes formulation, the MI approach is basically a statistical procedure for creating multiple sets of plausible values of missing data for reflecting uncertainty under a MAR assumption.
- The fundamental MI approach is repeated imputations, which, operationally, are drawn from the posterior predictive distribution of missing values under a particular, correctly assumed Bayes model on both the data and the missing-data mechanism.

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|\theta)\, P(\theta|Y_{obs})\, d\theta$$

Each set of imputations is used to create a completed dataset for deriving the completed-data estimators $\hat{\theta}$ and V with certain statistical procedures.

Multiple Imputations

- Suppose m imputations resulting in m repeated completed-data statistics: $\hat{\boldsymbol{\theta}}_{*1}, \ldots, \hat{\boldsymbol{\theta}}_{*_m}$ and $V_{*1}, \ldots, V_{*m}$ over the m completed datasets.
- These statistics are then combined to form one repeated-imputation inference for each assumed DGM.
- The repeated-imputation estimate of θ is

$$\bar{\theta}_m = \frac{\sum_{l=1}^{m} \hat{\theta}_{*l}}{m}$$

The associated variance-covariance of $\bar{\theta}_m$, denoted by V_m, is

$$V_m = \bar{V}_m + \frac{m+1}{m} B_m$$

where $\bar{V}_m$ and B_m represent the within-imputation and between-imputations variability, respectively, mathematically defined by

$$\bar{V}_m = \frac{\sum_{i=1}^{m} V_{*1}}{m}$$

and

$$B_m = \frac{\sum_{l=1}^{m} \left(\hat{\theta}_{*l} - \hat{\theta}_m\right) \left(\hat{\theta}_{*l} - \hat{\theta}_m\right)'}{(m-1)}$$

Then, asymptotically, $\sqrt{m}(\bar{\theta} - \theta) \overset{a}{\sim} N(0, \boldsymbol{V}_m)$

- Typically, large samples are not needed.
- In R the package MICE is one of the most commonly used for MI

`https://www.jstatsoft.org/article/view/v045i03/v45i03.pdf`
`https://stefvanbuuren.name/mice/`
`https://statistics.ohlsen-web.de/multiple-imputation-with-mice/`

Likelihood-based methods

- Likelihood-based methods, regardless of whether the missing-data mechanism is ignored or modeled, can also be thought of as imputation methods.
- When missingness is ignorable, likelihood-based methods can be used based solely on the marginal distribution of the observed data.
- Maximum likelihood estimates stimates can be obtained by maximizing $f\left(\boldsymbol{Y}_i^o|X_i,\gamma\right)$, the ordinary marginal distribution of the particular subset of $\boldsymbol{Y}_i$ determined by $\boldsymbol{Y}_i^o$, and the missing values are validly predicted by the observed data via the model for the conditional mean $E\left(\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o,X_i,\boldsymbol{\gamma}\right)$.
- This is more clear if we consider the EM algorithm for estimation (Dempster, Laird, and Rubin, 1977).

EM Algorithm

- In the EM algorithm, a two-step iterative algorithm alternates between filling in missing values with their conditional means, given the observed responses and parameter estimates from the previous iteration (the expectation or E-step), and maximizing the likelihood for the resulting "complete data" (the maximization or M-step).

EM Algorithm

- In the EM algorithm, a two-step iterative algorithm alternates between filling in missing values with their conditional means, given the observed responses and parameter estimates from the previous iteration (the expectation or E-step), and maximizing the likelihood for the resulting "complete data" (the maximization or M-step).
- For example, for multivariate normal responses, in the E-step of the EM algorithm one computes the "predictions" of the missing values based on

$$E\left(\boldsymbol{Y}_i^m|\boldsymbol{Y}_i^o,\boldsymbol{\gamma}\right)=\boldsymbol{\mu}_i^m+\Sigma_i^{mo}\Sigma_i^{o-1}\left(\boldsymbol{Y}_i^o-\boldsymbol{\mu}_i^o\right)$$

  Thus, when missingness is ignorable, likelihood-based inference does not require specification of the missing-data mechanism, but does require full distributional assumptions about $\boldsymbol{Y}_i$. Furthermore, the model for $f\left(\boldsymbol{Y}_i|X_i,\gamma\right)$ must be correctly specified. Any misspecification of the model for the covariance will, in general, yield biased estimates of the mean response trend.
- When missingness is non-ignorable, a joint model (e.g., selection or pattern-mixture model, soon) is required and inferences are sensitive to model assumptions.

Weighting methods

- An alternative approach for handling missing data is to weight the observed data appropriately
- Inverse probability weighted methods and propensity weights have a long history in statistics, from the sample survey literature (Horvitz and Thompson, 1952).
- In weighting methods, the underlying idea is to base estimation on the observed responses but weight them to account for the probability of non-response.
- Under MAR, the propensity for non response can be estimated as a function of the observed responses and any covariate that could predict non-response.
- For example, the GEE approach can be adapted by making adjustments for the propensities for dropout:

  Given estimated probabilities that subject i is still in the study at occasion j, $\hat{w}_{ij}$, a weighted analysis can be performed where the data available at the j-th occasion are weighted by $\hat{w}_{ij}^{-1}$.

Inverse Probability Weighted Augmented (IPWA) Estimators

- IPW and IPWA estimators for parameters of longitudinal data models were originally introduced by Robins and Rotnitzky (1992) as part of a general estimating function methods for incomplete data.
- From a practical point of view, this theory is the basis for the class of estimators for full data model parameters using what are often called in the context of missing data problems weighted estimating equations, or WGEEs.
- In order to discuss the IPWA, let's first consider a simple IPW estimator
- Let $C_i = 1$ if Y_i is observed and 0 otherwise.
- If we are willing to assume that missingness of Y depends only on covariates V and not on Y

$$\text{pr}(C = 1|Y, V) = \text{pr}(C = 1|V) = \pi(V)$$

- The complete case estimator, the sample mean of the Y_i for the individuals on whom Y is observed,

$$\widehat{\mu}^{cc} = \frac{\sum_{i=1}^{N} C_i Y_i}{\sum_{i=1}^{N} C_i}$$

  is in general not a consistent estimator for μ.

- Equivalently, $\widehat{\mu}^{cc}$ solves the estimating equation

$$\sum_{i=1}^{N} C_i \left(Y_i - \mu\right) = 0$$

- An inverse probability weighted estimator that is consistent can be derived by weighting the complete case equation:

$$\sum_{i=1}^{N} \frac{C_i}{\pi\left(V_i\right)} \left(Y_i - \mu\right) = 0$$

  which weighs the contribution of each complete case i by the inverse (reciprocal) of $\pi\left(V_i\right)$

- In order to show that is to be the case it is enough to show that the IPW estimating equation is unbiased, i.e. the estimating function

$$\frac{C_i}{\pi(V_i)}(Y_i - \mu)$$

satisfies

$$E_\mu \left\{ \frac{C_i}{\pi(V_i)}(Y_i - \mu) \right\} = 0$$

- This follows from

$$\begin{aligned} E_\mu \left\{ \frac{C_i}{\pi(V_i)}(Y_i - \mu) \right\} &= E_\mu \left[E \left\{ \frac{C_i}{\pi(V_i)}(Y_i - \mu) | Y_{-i}, V_i \right\} \right] \\ &= E_\mu \left\{ \frac{E(C_i | Y_{-i}, V_i)}{\pi(V_i)}(Y_i - \mu) \right\} \\ &= E_\mu \left\{ \frac{\pi(V_i)}{\pi(V_i)}(Y_i - \mu) \right\} \\ &= E_\mu (Y_i - \mu) = 0 \end{aligned}$$

since $E(C_i|Y_{-i}, V_i) = \text{pr}(C_i = 1|Y_{-i}, V_i) = \text{pr}(C_i = 1|V_i) = \pi(V_i)$ under MAR

- The estimator solving the previous EE is

$$\widehat{\mu}^{ipw2} = \left\{ \sum_{i=1}^{N} \frac{C_i}{\pi\left(V_i\right)} \right\}^{-1} \sum_{i=1}^{N} \frac{C_i Y_i}{\pi\left(V_i\right)}$$

  which is a weighted average of the observed Y_i's

- An alternative class of estimators involves augmenting the simple inverse probability weighted complete case estimating equation for μ. Estimators in this class can yield improved efficiency.

- The optimal estimator for μ within this class is the solution to the estimating equation

$$\sum_{i=1}^{N} \left[\frac{C_i}{\pi\left(V_i; \widehat{\psi}\right)} \left(Y_i - \mu\right) - \frac{C_i - \pi\left(V_i; \widehat{\psi}\right)}{\pi\left(V_i; \widehat{\psi}\right)} E\left\{ \left(Y_i - \mu\right) | V_i \right\} \right] = 0$$

- After some algebra, the previous expression can be written as

$$\sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi\left(V_i; \hat{\psi}\right)} - \frac{C_i - \pi\left(V_i; \widehat{\psi}\right)}{\pi\left(V_i; \widehat{\psi}\right)} E\left(Y_i | V_i\right) - \mu \right\} = 0$$

and leads to the estimator

$$\widehat{\mu} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi\left(V_i; \widehat{\psi}\right)} - \frac{C_i - \pi\left(V_i; \widehat{\psi}\right)}{\pi\left(V_i; \widehat{\psi}\right)} E\left(Y_i | V_i\right) \right\}$$

However, the conditional expectation $E(Y|V)$ is not known, and needs to be estimated from the data.
If the mechanism is MAR, $E(Y|V) = E(Y|V, C = 1)$, so we can estimate it from the data.

- More specifically, we can posit and fit a model for $E(Y|V = v)$,

$$m(v;\xi)$$

say, involving a finite-dimensional parameter ξ, on the complete cases $\{i : C_i = 1\}$. Specifically, if Y is continuous, for example, we might derive an estimator $\widehat{\xi}$ for ξ by using GEE under independence (OLS), solving in ξ

$$\sum_{i=1}^{N} C_i \frac{\partial}{\partial \xi} \left\{ m\left(V_i;\xi\right) \right\} \left\{ Y_i - m\left(V_i;\xi\right) \right\} = 0$$

- Substituting in the previous estimator $\hat{\mu}$, the resulting AIPW estimator for μ is

$$\widehat{\mu}^{\text{aipw}} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{C_i Y_i}{\pi\left(V_i;\hat{\psi}\right)} - \frac{C_i - \pi\left(V_i;\widehat{\psi}\right)}{\pi\left(V_i;\hat{\psi}\right)} m\left(V_i;\widehat{\xi}\right) \right\}$$

It can be shown that $\widehat{\mu}^{\text{aipw}}$ relatively more efficient than the simple inverse probability weighted estimator $\widehat{\mu}^{\text{ipw}}$. Moreover, it also has the property of double robustness.

- DOUBLE ROBUSTNESS: It can be shown that the estimator $\widehat{\mu}^{\text{aipw}}$ is a consistent estimator for μ if $EITHER$
- the model $\pi(v;\psi)$ for $\text{pr}(C=1|V=v)$ is correctly specified, OR
- The model $m(v;\xi)$ for $E(Y|V=v)$ is correctly specified

(or both). This property is referred to as double robustness, and the estimator $\widehat{\mu}^{\text{aipw}}$ is said to be doubly robust because its consistency is robust to mis-specification of either of these models.

Double Robustness

- A heuristic demonstration of this double robustness property is as follows. Under regularity conditions, $\widehat{\mu}^{\text{aipw}}$ converges in probability to

$$E\left\{\frac{CY}{\pi\left(V;\psi^*\right)}-\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}m\left(V;\xi^*\right)\right\}$$

where ψ^* and ξ^* are the limits in probability of $\widehat{\psi}$ and $\widehat{\xi}$. Adding and subtracting common terms

$$E\left[Y+\left\{\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}\right\}\left\{Y-m\left(V;\xi^*\right)\right\}\right]=\mu+E\left[\left\{\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}\right\}\left\{Y-m\left(V;\xi^*\right)\right\}\right]$$

- Consequently, $\widehat{\mu}^{\text{aipw}}$ is a consistent estimator for μ if we can show that

$$E\left[\left\{\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}\right\}\left\{Y-m\left(V;\xi^*\right)\right\}\right]=0$$

Double Robustness

- Using iterated conditional (on V) expectation, the previous expression can be written as

$$E\left(E\left[\left\{\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}\right\}\left\{Y-m\left(V;\xi^*\right)\right\}\mid V\right]\right)$$
$$=E\left[E\left\{\frac{C-\pi\left(V;\psi^*\right)}{\pi\left(V;\psi^*\right)}\mid V\right\}E\left\{Y-m\left(V;\xi^*\right)\mid V\right\}\right]$$

  where the last expression follows from MAR, so that C and Y are independent conditional on V.

- Consider two cases:

(a) $\pi(v;\psi)$ is correctly specified. Then $\widehat{\psi}$ converges in probability to the true value of ψ, so that

$$\pi\left(V;\psi^*\right)=\mathrm{pr}(C=1\mid V)$$

Double Robustness

- Under this condition,

$$E\left\{\frac{C-\pi\left(V;\psi^{*}\right)}{\pi\left(V;\psi^{*}\right)}\mid V\right\}=E\left\{\frac{E(C\mid V)-\mathrm{pr}(C=1\mid V)}{\mathrm{pr}(C=1\mid V)}\right\}=0$$

  using $E(C \mid V)=\mathrm{pr}(C=1 \mid V)$.

(b) $m(v;\xi)$ is correctly specified. Then $\widehat{\xi}$ converges in probability to the true value of ξ, and thus

$$m\left(V;\xi^{*}\right)=E(Y \mid V)$$

In this case, $E\left\{Y-m\left(V;\xi^{*}\right) \mid V\right\}=E\{Y-E(Y \mid V) \mid V\}=0$, and the result follows.

- The previous principles can be used to derive inverse probability weighted and doubly robust AIPW estimators for the regression parameters in a regression model of interest.
- For example, these ideas are at the foundation of weighted generalized estimating equations for longitudinal data subject to dropout.
- See Chapters 17-23 of the Handbook of Longitudinal Data Analysis

Handling NMAR data

- There is abundant literature devoted to modeling nonignorable longitudinal missing data in biostatistics. The primary focus of this literature is dropout in clinical trials.
- In these cases, the occurrence of missing data is predominantly due to reasons other than death and is closely related to outcomes being measured.
- Joint models are often used in longitudinal analyses to correct for non-ignorable non-response.
- Little and Rubin (1987) and Little (1993) identified two broad classes of joint models for the longitudinal data and response indicators:
  - selection models
  - pattern-mixture models

  The models need to specify a complete-data model with a missing data mechanism where the pattern of non-response is modeled conditional on the possibly unobserved outcomes

Selection models

- Focusing on the problem of non-ignorable dropouts, Wu and Carroll (1988) proposed a selection modeling approach used by many subsequent researchers.
- It assumes that the continuous responses follow a simple linear random-effects model and that the dropout process depends upon an individual's random intercept and slope.
- Models where the dropout probabilities depend indirectly upon the unobserved responses, via the random effects, are often referred to also as shared-parameter models
- In Shared-parameter models, a model for the longitudinal response measurements is linked with a model for the missing-data mechanism through a set of random effects that are shared between the two processes:

$$f(\boldsymbol{R}_i, \boldsymbol{Y}_i, \boldsymbol{b}_i | \boldsymbol{X}_i, \gamma, \phi) = \underbrace{f(\boldsymbol{b}_i | \boldsymbol{X}_i, \gamma_1)}_{\text{random effect distr}} \times \underbrace{f(\boldsymbol{Y}_i | X_i, \boldsymbol{b}_i, \gamma_2)}_{\text{likelihood}} \times \underbrace{f(\boldsymbol{R}_i | X_i, \boldsymbol{b}_i, \phi)}_{\text{Missing Data Mechanism}}$$

Example

- Consider a gaussian mixed effect model:

$$Y_{ij}|X_i, b_i = \beta_0 + \beta_1 X_i + b_i + \epsilon_{ij}$$

  The choice of the density for the missing data mechanism depends on the type of missing data being incorporated.

- E.g.: Probability of dropping out:

$$\Phi^{-1}\left\{\Pr\left(R_{ij} = 0 | R_{ij} = 1\right)\right\} = \alpha_0 + \alpha X_i + \theta b_i$$

  In a randomized clinical trial with no missing data, choice of baseline covariates to use in the LME is not crucial as randomization assures the validity of the test of treatment effect.

- With missing data, randomized trials become more like observational studies. The model must be correctly specified for correct inference.
- In particular, choice of covariates in the LME and MD model must be considered carefully.
- A sensitivity analysis might consider ML estimates of the parameters for a variety of plausible alternative choices of ϕ.

Pattern-mixture models

- The pattern mixture model is developed as a joint model, combining different patterns of missing data.
- More specifically, selection models specify the joint distribution of $\boldsymbol{R}_i$ and $\boldsymbol{Y}_i$ through models for the marginal distribution of $\boldsymbol{Y}_i$ and the conditional distribution of $\boldsymbol{R}_i$ given $\boldsymbol{Y}_i$:

$$f\left(\boldsymbol{R}_i, \boldsymbol{Y}_i | X_i, \boldsymbol{\gamma}, \boldsymbol{\phi}\right) = f_Y\left(\boldsymbol{Y}_i | X_i, \boldsymbol{\gamma}\right) f_{R|Y}\left(\boldsymbol{R}_i | X_i, \boldsymbol{Y}_i, \boldsymbol{\phi}\right)$$

  where $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{\delta})$.
- Pattern-mixture models specify the marginal distribution of $\boldsymbol{R}_i$ and the conditional distribution of $\boldsymbol{Y}_i$ given $\boldsymbol{R}_i$:

$$f\left(\boldsymbol{R}_i, \boldsymbol{Y}_i | X_i, \boldsymbol{\nu}, \boldsymbol{\delta}\right) = f_R\left(\boldsymbol{R}_i | X_i, \boldsymbol{\delta}\right) f_{Y|R}\left(\boldsymbol{Y}_i | X_i, \boldsymbol{R}_i, \boldsymbol{\nu}\right)$$

  where $\boldsymbol{\theta} = (\boldsymbol{\nu}, \boldsymbol{\delta})$.

Example

- A normal selection model for two repeated measures with non-MAR dropouts:

$$(Y_{i1}, Y_{i2}) \sim N(\boldsymbol{\mu}, \Sigma)$$

$$(R_i | Y_{i1}, Y_{i2}) \sim \text{Ber}\left(P\left(\phi\left(Y_{i1}, Y_{i2}\right)\right)\right.$$

$$\text{logit}\left\{P\left(\phi\left(Y_{i1}, Y_{i2}\right)\right)\right\} = \phi_0 + \phi_1 Y_{i1} + \phi_2 Y_{i2}$$

  Heckman (1976) selection model

- A normal pattern-mixture model for two repeated measures with non-MAR dropouts:

$$(Y_{i1}, Y_{i2} | R_i = k) \sim N\left(\boldsymbol{\mu}^{(k)}, \Sigma^{(k)}\right), \quad k = 0, 1$$

$$(R_i) \sim \text{Ber}(\delta)$$

Example

- The pattern-mixture model implies that the marginal mean of (Y_{i1}, Y_{i2}) averaged over patterns is

$$\boldsymbol{\mu} = (1-\delta)\boldsymbol{\mu}^{(0)} + \delta\boldsymbol{\mu}^{(1)},$$

  and the parameter of interest is

$$\mu_{\text{dif}} = (1-\delta)\left(\mu_2^{(0)} - \mu_1^{(0)}\right) + \delta\left(\mu_2^{(1)} - \mu_1^{(1)}\right)$$

  the weighted average of the differences in means in the two patterns.

- The parameter of interest is not a parameter of the pattern-mixture model, but is easily expressed as a function of the model parameters
- The model is not identifiable in a frequentist setting: it may be identified through prior distributions in Bayesian setting, or instead by placing restrictions on the model parameters, based on assumptions about the nature of the missing-data mechanism or the model for Y

Selection vs Pattern-Mixture models

- Selection models are a natural way of factoring the model, with f_Y the model for the data in the absence of missing values, and $f_{R|Y}$ the model for the missing-data
- If the MAR assumption is plausible, the selection model formulation leads directly to the ignorable likelihood - the distribution $f_{R|Y}$ for the missing-data mechanism is not needed for likelihood inferences, which can be based solely on the model for f_Y.
- In some NMAR situations, pattern-mixture models make more sense. E.g., if Y_{ij} is a measure of quality of life at age $\jmath$, and $R_{ij} = 1$ for survivors at age j and $R_{ij} = 0$ for individuals who die before age j, then it appears more meaningful to consider the distribution of Y_{ij} given $R_{ij} = 1$ (restricting the inference to the subpopulation of cases with values observed.) rather than the marginal distribution of Y_{ij}.

Selection vs Pattern-Mixture models

- From an imputation perspective , missing values $\boldsymbol{Y}_i^m$ should be imputed from their predictive distribution given the observed data including $\boldsymbol{R}_i$, that is, $f\left(\boldsymbol{Y}_i^m | \boldsymbol{Y}_i^o, \boldsymbol{R}_i, X_i\right)$.
- Pattern-mixture models may avoid specification of the model for the missing-data mechanism in NMAR situations, by using assumptions about the mechanism to yield restrictions on the model parameters.