BIOS 755: Missing Data in Longitudinal Studies

Alexander McLain

April 1, 2025

- Even in well-controlled situations, missing data invariably occur in longitudinal studies.
- Subjects can be missed at a particular measurement wave, with the result that these subjects provide data at some, but not all, study time points.
- Alternatively, subjects who are assessed at a given study time point might only provide responses to a subset of the study variables, again resulting in incomplete data.
- Finally, subjects might dropout of the study or be lost to follow-up, thus providing no data beyond a specific point in time.

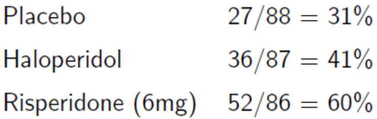
Missing data Mechanism

- In general, the situation may often be quite complex, with some missingness unrelated to either the observed or unobserved response, some related to the observed, some related to the unobserved, and some to both.
- A particular pattern of missingness that is common in longitudinal studies is 'dropout' or 'attrition.' This is where an individual is observed from baseline up until a certain point in time, thereafter no more measurements are made.
- Possible reasons for dropout:
  1. Recovery;
  2. Lack of improvement or failure;
  3. Undesirable side effects;
  4. External reasons unrelated to specific treatment or outcome;
  5. Death

An Example

Schizophrenia Treatment Trial

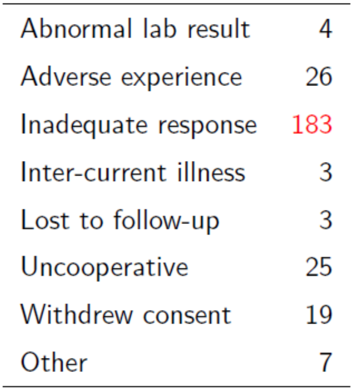
- A randomized longitudinal study of haloperidol and risperidone
- Primary outcome: Positive and Negative Scale for Schizophrenia (PANSS) score at 8 weeks, intermediate outcomes at 1, 2, 4, and 6 weeks



Placebo	$27/88 = 31\%$
Haloperidol	$36/87 = 41\%$
Risperidone (6mg)	$52/86 = 60\%$

An Example

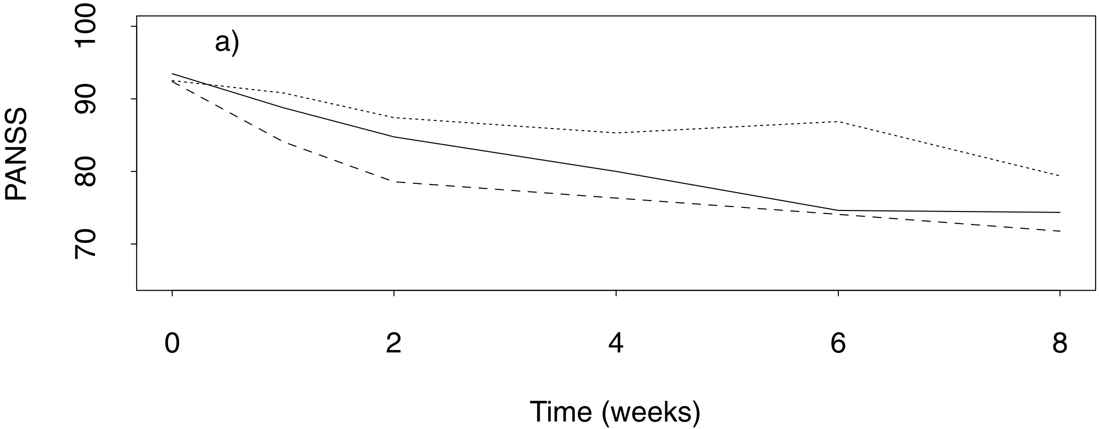
- Reasons for dropout:



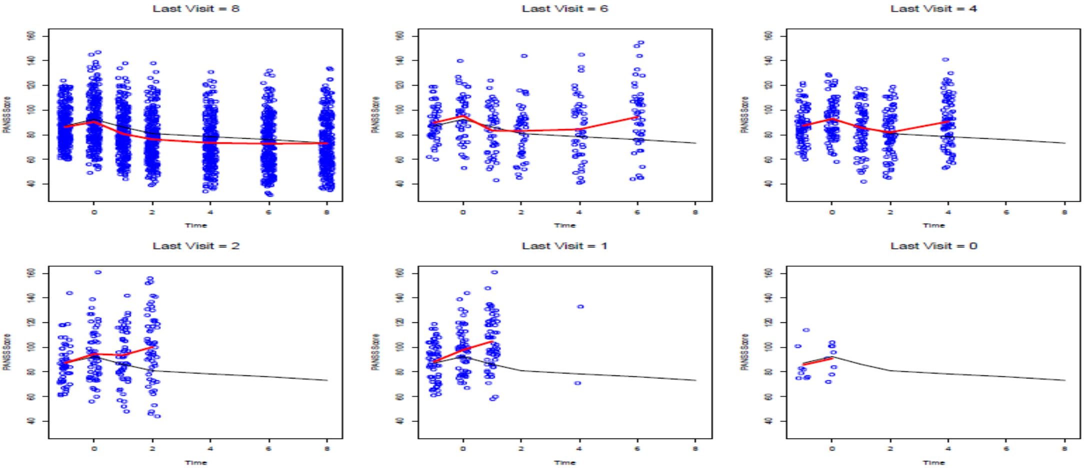
Abnormal lab result	4
Adverse experience	26
Inadequate response	183
Inter-current illness	3
Lost to follow-up	3
Uncooperative	25
Withdrew consent	19
Other	7

- “Inadequate response” is seen as potentially informative.

An Example



An Example


Last Visit = 8
Last Visit = 6
Last Visit = 4
Last Visit = 2
Last Visit = 1
Last Visit = 0
PANSS Score
40
60
80
100
120
140
160
Time
0
2
4
6
8

Another Example

Six Cities Study of Air Pollution and Health

- Longitudinal study designed to characterize lung growth as measured by changes in pulmonary function.
- A cohort of 13,379 children born on or after 1967 was enrolled in six communities across the U.S.
- Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up (1 to 12 measurements were taken).
- Major reason for late entry or attrition was moving in or out of the school district.

Another Example

Muscatine Coronary Risk Factor (MCRF)

- This study which had five cohorts of children, initially aged 5–7, 7–9, 9–11, 11–13, and 13–15.
- Goal: determine whether the risk for obesity increased with age.
- Measurements were taken in 1977, 1979 and 1981. Less then 40% had complete data.
- Major reasons for missing data:
  - failure to obtain consent from parents, and
  - the child was absent from school on the day of measurement.

Some Terminology

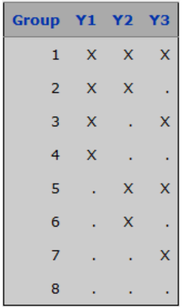
- Complete Data: The scheduled measurements. This is the outcome vector, $\boldsymbol{Y}_i$, that would have been recorded if no missing data occurred.
- Missing Data Indicators: The binary variables, $\boldsymbol{R}_i = \{R_{i1}, R_{i2}, \ldots, R_{in_i}\}$, that indicate whether Y_{ij} was observed, $R_{ij} = 1$, or unobserved, $R_{ij} = 0$. That is,

$$\begin{aligned} R_{ij} &= 1 \quad \text{if subject } i \text{ is observed at time } j \\ R_{ij} &= 0 \quad \text{if subject } i \text{ is not observed at time } j \end{aligned}$$

- Observed (Response) Data: $\boldsymbol{Y}_i^O$: Y_{ij} such that $R_{ij} = 1$.
- Missing Data: $\boldsymbol{Y}_i^M$: Y_{ij} such that $R_{ij} = 0$.

Missing Data Patterns and Mechanisms

- Consider the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{R}$ (both are random variables).
- The pattern of missingness concerns the distribution of $\boldsymbol{R}$.
- Mechanism concerns how the distributions of $\boldsymbol{R}$ and $\boldsymbol{Y}$ relate to one another.



Group	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	X
4	X	.	.
5	.	X	X
6	.	X	.
7	.	.	X
8	.	.	.

Types of Missing Data

- We will discuss three different types of missingness mechanisms:
  1. Missing completely at random (MCAR),
  2. Missing at random (MAR), and
  3. Missing not at random (MNAR).
- To formulate different missing data mechanisms suppose we only observe:
  - $\boldsymbol{Y}_i = \{Y_{i1}, Y_{i2}\}$ where Y_{i1} **is always observed and** Y_{i2} **is sometime missing.**
  - Covariate data $\boldsymbol{X}_i$.
  - $R_{i1} = 1$ for all individuals, R_{i2} can be either a 1 or 0.

MCAR

- Data are said to be MCAR if

  $$\Pr(R_{i2} = 1 | \boldsymbol{Y}_i, \boldsymbol{X}_i) = \Pr(R_{i2} = 1 | \boldsymbol{X}_i)$$

  This implies

  $$E(Y_{i2} | R_{i2} = 1, \boldsymbol{X}_i) = E(Y_{i2} | \boldsymbol{X}_i)$$

- This is sometimes called covariate dependent missingness.
- MCAR could occur if:
  - the dropout was a planned part of the study,
  - the dropout was completely unrelated to the outcome, or
  - the missingness was related to an observed variable (i.e., measured side effects of a drug).
- In general, the data are MCAR if

  $$\Pr(\boldsymbol{R}_i = 1 | \boldsymbol{Y}_i, \boldsymbol{X}_i) = \Pr(\boldsymbol{R}_i = 1 | \boldsymbol{X}_i)$$

MAR

- Missing at random (MAR) if

$$\Pr(R_{i2} = 1 | Y_{i1}, Y_{i2}, \boldsymbol{X}_i) = \Pr(R_{i2} = 1 | Y_{i1}, \boldsymbol{X}_i)$$

  Here the probability of missing data only depends on the observed value Y_{i1} and not the missing value Y_{i2}.
- MAR could occur if:
  - planned dropout was determined based on the previous outcomes,
  - e.g., in the six studies example if a child moved due to respiratory problems that could be predicted based on the data.
- If imputation is not used:
  - **Missing baseline variables usually mean all observations are missing** making MAR more doubtful
  - There is no MAR for cross-sectional data with complete case analysis.

MAR

- The trouble starts here since this implies

$$E(Y_{i2}|R_{i2}=1,\boldsymbol{X}_i) \neq E(Y_{i2}|\boldsymbol{X}_i) \ \ \text{(possibly)}$$

- In general MAR is stated as

$$\Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i^O,\boldsymbol{Y}_i^M,\boldsymbol{X}_i)=\Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i^O,\boldsymbol{X}_i).$$

- Under MAR the observed data are no longer a random sample of the complete data.
- However given the observed values $\boldsymbol{Y}_i^O$, the missing values $\boldsymbol{Y}_i^M$ do not have any type of “strange” or unexpected distribution. That is, their distribution is the same as the complete cases.

MAR with missing baseline data

- MAR is a “good” mechanism, but it doesn’t apply with missing baseline data unless imputation is used.
- MAR works by saying the missing data doesn’t add different any systematic trends based on the observed data.
  - i.e., the missing data is predictable
- However, if a person is completely missing from the analysis, they don’t contribute any observed data.
- If imputation is not used:
  - **Missing baseline variables usually mean all observations are missing** making MAR more doubtful
  - There is no MAR for cross-sectional data with complete case analysis
- Imputation or weighting should always be used with missing baseline data unless the amount of missing data is small.

MNAR

Missing not at random (MNAR)

- Also referred to as “nonignorable” missingness or “informative” dropout
- The data are MNAR if

$$\Pr(R_{i2} = 1 | Y_{i1}, Y_{i2}, \boldsymbol{X}_i) = \Pr(R_{i2} = 1 | Y_{i1}, Y_{i2}, \boldsymbol{X}_i)$$

  or

$$\Pr(\boldsymbol{R}_i | \boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M, \boldsymbol{X}_i) = \Pr(\boldsymbol{R}_i | \boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M, \boldsymbol{X}_i).$$

- Examples:
  1. In the MCRF study, if parent of obese children were more or less likely to sign the consent form.
  2. In the Schizophrenia study, subjects dropped out because PANSS increased.

Example of MCAR

- Lavange and Helms (1983) analyzed data from a longitudinal study of lung function conducted on 72 children age 3 to 12 years;
- A measure of maximum expiratory flow rate was measured annually, and differences in the resulting curve were related to between-subject covariates such as Race and Gender.
- The number of actual measurements recorded on each child ranged from 1 to 8, with an average of 4.2 per child.
- Some values were missing because the child was older than age 3 at the start of the study or younger than 12 at the end of the study.
- This mechanism depends on cohort, but plausibly does not depend on outcomes if cohort is a covariate in the model.

Example of MAR

- Murray and Findlay (1988) studied the effect of two drugs in a large multi-center clinical trial.
- The drugs were antihypertensive agents for patients with mild to moderate hypertension.
- Diastolic BP was the outcome of interest.
- The study lasted 12 weeks with measurements at 0, 2, 4, 8, and 12 weeks.
- The protocol stated that patients with DBP exceeding 110 mmHg at either the 4- or 8-week visit should “jump” to an open follow-up phase (i.e., drop-out).

Example of MNAR

- Diggle and Kenward (1994) analyzed a longitudinal milk protein trial discussed by Verbyla and Cullis (1990).
- Cows were randomly allocated to one of three diets: barley, mixed barley-lupins, or lupins, and the assayed protein content of milk samples was taken weekly for a period of 20 weeks.
- Drop-out corresponds to cows that stopped producing milk before the end of the experiment.

Some Simple Approaches

- We will now discuss some simple methods people use to deal with missing data.
- These methods usually create a single ‘complete’ dataset, which is analyzed as if it were the fully observed data.
- Unless certain, assumptions are true, the answers are biased or invalid.

Complete Case Analysis

- First, we will discuss complete case analysis:

Subject	Y_1	Y_2
1	2	-0.668
2	0	-2.118
3	1	-0.480
4	1	-1.058
5	2	??

Complete Case Analysis

- Completer case analysis (CCA) deletes all units with incomplete data (in the variables involved) from the analysis.
- Can be inefficient.
- Need to be careful comparing models when using CCA
- Need to make sure you are not changing the size of the data set, as you add/remove explanatory variables with missing observations, or we use the subset of the data with no missing values.
- When the missing observations are not MCAR or MAR (depends on the model, will discuss below), this analysis may give biased estimates and invalid inferences.
- The default in statistical packages.

Imputation of the mean

Subject	Y_1	Y_2
1	2	-0.668
2	0	-2.118
3	1	-0.480
4	1	-1.058
5	2	-0.883

- This approach is clearly inappropriate for categorical variables.
- It does not lead to proper estimates of measures of association or regression coefficients. Rather, associations tend to be diluted.
- In addition, variances will be wrongly estimated (typically under estimated) if the imputed values are treated as real. Thus inferences will be wrong too.

Regression mean imputation

- In this case we use those with complete data to build a model that relates all variables.
- Then we substitute in the *predicted mean* for each unit with a missing value.
- In this way we use information from the joint distribution of the variables to make the imputation.
- In our simple data set we might consider

$$Y_2 = \beta Y_1 + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

- This is a better idea than mean imputation, but not good in this form.
- This is the idea behind multiple imputation that we will discuss later on.

Creating an extra category

Subject	Y_1	Y_2
1	2	-0.668
2	0	-2.118
3	?→ 3	-0.480
4	1	-1.058
5	?→ 3	0.453

- The impact of this strategy depends on how missing values are divided among the real categories, and how the probability of a value being missing depends on other variables;
- Very dissimilar classes can be lumped into one group;
- Sever bias can arise, in any direction, and
- When used to correct for confounding the completed categorical variable will not do its job properly.

Last Observation Carried Forward (LOCF)

Subject	Y_1	Y_2	Y_3	Y_4	Y_5
1	-0.668	0.223	0.119	0.908	0.171
2	-2.118	-0.941	0.465	?→ 0.465	-0.554
3	-0.480	1.387	?→ 1.387	?→ 1.387	-0.218
4	-1.058	-0.413	-0.243	0.136	-0.106
5	-0.092	-0.377	1.115	0.215	-1.220

- Using LOCF, once the data set has been completed in this way it is analyzed as if it were fully observed.
- For full longitudinal data analyses this is clearly disastrous:
  - Means and covariance structure are seriously distorted.
  - For single time point analyses the means are still likely to be distorted, measures of precision are wrong and hence inferences are wrong.

Summary of ‘ad hoc’ methods

- Unless the proportion missing is so small as to be unlikely to affect inferences, these simple ad-hoc methods should be avoided.
- They usually conflict with the statistical model that underpins the analysis (however simple and implicit this might be) so they introduce bias.
- As the assumptions about the reason for the data being missing that they implicitly make are often difficult to describe (e.g. with LOCF), they can make it very hard to know what assumptions are being made in the analysis.
- **They do not properly reflect statistical uncertainty: data are effectively ’made up’ and no subsequent account is taken of this.**

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).
- **Accounts for the uncertainty in the missing-data.**
- Favor likelihood based approaches, but GEE's are also good.
  - Maximum Likelihood (ML) for large samples
  - Multiple Imputation/Bayes for small samples

MAR methods

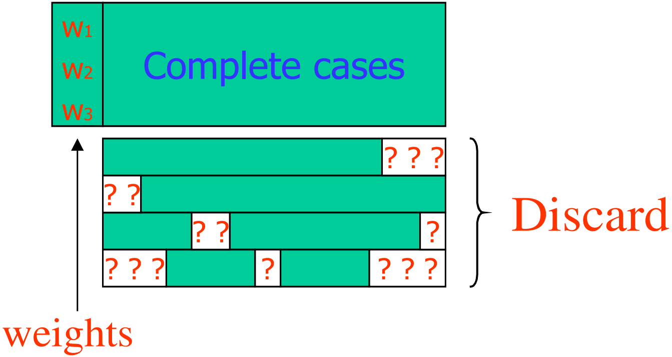
The main approaches to analyzing MAR data are:

- Weighting methods.
- Multiple imputation.
- Direct likelihood.
- Direct bayesian.

Weighting Methods

Introduction

- The complete case (CC) analysis is the default in statistical packages (list-wise deletion).
- Weighting methods using a CC where the complete cases are weighted.



Weights

- The weight corresponds to the inverse probability that the data point was collected.
$$w_i = \frac{1}{\Pr(i \text{ completely responds } |\boldsymbol{X})}$$
- In the analysis subject i represents w_i subjects of the population.
- That is, if $w_i = 2$ then subject i represents 2 people.
- People that were not likely to respond will be represented more heavily in the analysis.
- Since probability of response is not known, we need to estimate it.
  - This can be done using logistic regression.

Inference from Weighted Data

- Can use packages for computing estimates and standard errors with weighted data (e.g., weighted GEE) – but often these do not take into account sampling uncertainty in weights.
- Bootstrapping weighting procedures propagates uncertainty in weights – but weights need to be recalculated on each bootstrap sample.
- Weighting is a relatively simple device for reducing bias from complete-case analysis.
- Most useful when the patten of dropout is monotone.
- These methods includes:
  - Weighted GEE
  - Propensity Weighting
  - Weighted mean

Multiple Imputation

Introduction

- Create multiple data sets, each with a different set of imputed values
- Each imputed value includes stochastic variation based on the variability in the posterior distribution of the estimate and the residuals
    - Thus, the imputed values vary across each of the MI data set
- MI relies on the MAR assumption; a weaker assumption than MCAR

Actual data set

Subject	*PA*	*Age*	*SBP*	*BMI*	*Loss*
1	8	32	120	19	0
2	3	48	145	XX	1
3	4	25	XX	XX	0
⋮	⋮	⋮	⋮	⋮	⋮

- Let $Y_{im} \sim N(\hat{\beta}\boldsymbol{X}_i, \hat{\sigma}^2)$, which is from

$$Y_i = \beta\boldsymbol{X}_i + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

  is fitted using the complete cases.

- Here Y_i is any variable with missingness (usually not the outcome, but can be).

Imputed data sets

Actual data set

	m=1				
Subject	*PA*	*Age*	*SBP*	*BMI*	*Loss*
1	8	32	120	19	0
2	3	48	145	28.947	1
3	4	25	132.54	26.426	0
⋮	⋮	⋮	⋮	⋮	⋮
	m=2				
Subject	*PA*	*Age*	*SBP*	*BMI*	*Loss*
1	8	32	120	19	0
2	3	48	145	30.126	1
3	4	25	128.14	22.36	0
⋮	⋮	⋮	⋮	⋮	⋮

Multiple Imputation

- MI is NOT about finding “the best” value for each missing value.
- MI is about preserving unbiased estimates of population parameters:
  - means, variances, covariances, correlations, multivariate regression coefficients
  - We need to be right on average (MAR assumption) while accurately representing our uncertainty about missing values.
- MI uses one model to create the multiple imputed data sets, then another model to analyze the data.
- When we analyze the data there are methods to control for the imputation.

Multiple Imputation model

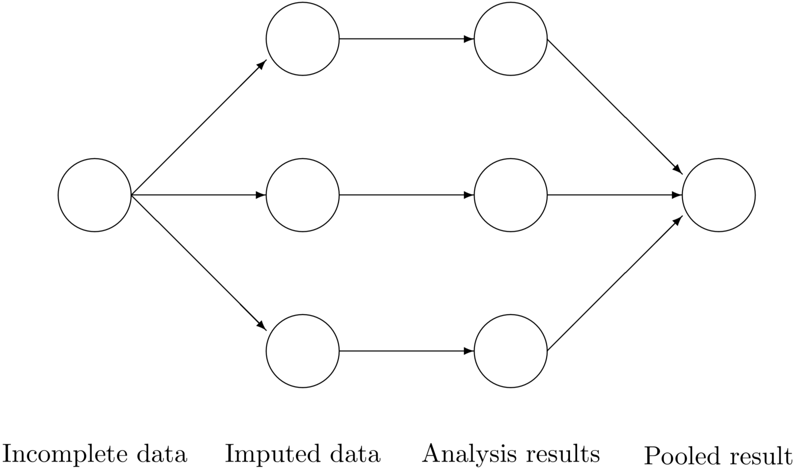
- Mostly people will use a Multivariate Normal (MVN) MI model for continuous data.
  - Strongest theoretical justification
  - Implemented in R, SAS, Stata, and stand alone packages
- To use this model you need to assume the variables follow a normal distribution.
- For this reason it is common to transform the variables, perform the MI, then back-transform the variables for analysis of the imputed data sets.

Using Multiple Imputation using fully conditional specification

- In `R`, either the Multivariate Imputation by Chained Equations, `mice`, or `Amelia` packages can be used to generate the imputed datasets.
  - their syntax in very similar.
- In SAS, there is the fully conditional specification (fcs) model which is similar to `mice`.
- Benefit of these approaches are that they can be used for binary, count, or multinomial data (really any kind).

Click here for more information about mice.

Using Multiple Imputation



Incomplete data Imputed data Analysis results Pooled result

Summary of Multiple Imputation

- Use at least 50 data sets. Use more to:
  - Achieve more stable estimates when missing rates are high
- Imputation model should be at least as comprehensive as the estimation model
  - The imputation model must include all of the variables & interactions you want to use in the estimation
  - Include any variables associated with the mechanism of missingness or with the variables to be imputed
- DO include the dependent variable in the imputation model
- Do NOT round imputed values (say, when using MVN model to predict a 0/1 variable)
- Impute at the lowest level of data possible.

Likelihood Based Inference

In 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 (technical condition on the parameter spaces)
- This also holds for Bayesian inference methods.
- See Little and Rubin (2002, chapter 6) for more on this.

In contrast, many ad-hoc methods require the stronger MCAR assumption

Summary methods that can be used in MNAR

- **Selection model** (Rubin, 1976)

  $$\Pr(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\theta) = \Pr(\boldsymbol{R}|\boldsymbol{Y}_{\text{obs}}, \theta_T)\Pr(\boldsymbol{Y}_{\text{obs}}|\theta_Y)$$

  model the outcome marginally and appended by a model on the dropout, conditioning on the outcome.

- **Pattern-mixture model** (Little, 1993)

  $$\Pr(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\theta) = \Pr(\boldsymbol{Y}_{\text{obs}}|\boldsymbol{R}, \theta_{Y|R})\Pr(\boldsymbol{R}|\theta_R)$$

  here we condition the outcome on the dropout structure (quality of life).

- **Shared parameter models** (Wu and Carroll, 1988, Follmann and Wu, 1995)

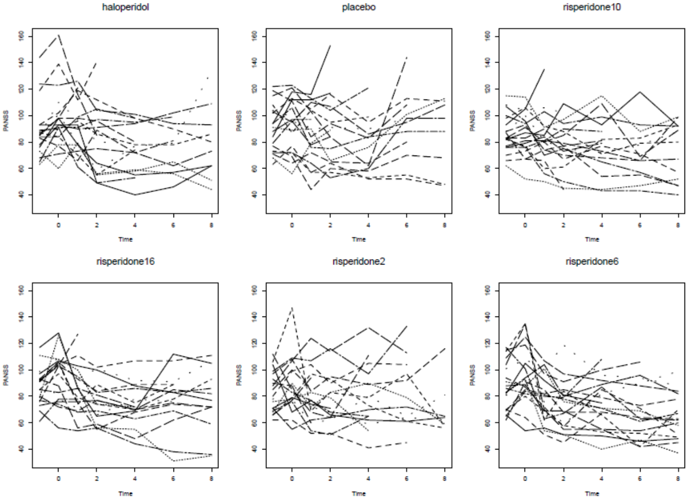
  $$\Pr(\boldsymbol{Y}_{\text{obs}}, \boldsymbol{R}|\theta) = \int \Pr(\boldsymbol{R}|b, \theta_{R|b})\Pr(\boldsymbol{Y}_{\text{obs}}|b, \theta_{Y|b})f(b)db$$

  can provide useful information on both models.

MNAR Conclusions

- All the MNAR methods have strong, untestable assumptions most useful in sensitivity analysis.
- 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.
- These are a fruitful area of statistical research.

PANSS Example



haloperidol

placebo

risperidone10

risperidone16

risperidone2

risperidone6

PANSS Example

	Final Outcome (8wks)		
	est.	(s.e.)	p-value
Haloperidol is reference group			
Risperidone (6mg)	-7.901	(4.148)	0.0580
Placebo	3.962	(4.844)	0.4142
	Longitudinal Analysis		
	est.	(s.e.)	p-value
Haloperidol is reference group			
Risperidone (6mg)	-13.670	(4.256)	0.0014
Placebo	13.085	(4.586)	0.0045

PANSS Example

- Analysis using available-case data at 8 weeks yields different conclusions from a longitudinal analysis.
- The result suggests dropout predicts the outcome:

$$E\{Y(8)|\text{completer}\} < E\{Y(8)|\text{drop-out}\}$$

- Therefore, the data are not MCAR.
- Q: What is the anticipated impact when analyzing data that are MAR?

General strategies

- Complete-case (CC) analysis (i.e., discard cases with incomplete data)
  - CC may be a biased sample
  - inefficient
  - weighting of complete cases can be used to correct for bias
- Imputation
  - naive imputations (e.g., means, LOCF) can be worse than CC analysis
  - multiple imputation good (Schafer & Graham, 2002)
- Analyze as incomplete
  - different methods have different mechanism assumptions

MCAR is sometimes a problem!

- Earlier, I swept over MCAR like it was no issue. Saying “it’s no problem as long as we adjust for the variables related to the missingness.”
- I said this because we’re doing regression models, and we can adjust for things.
- When you’re not doing regression models, biases can occur.
- For example, suppose you are counting votes, but you only count votes for those that have appropriate ID.
- If variables are related to having appropriate ID and related to the vote, the results will be a biased estimate of the voting proportions of the total population.
- They still can be an unbiased estimate of the voting proportions of the total population with appropriate ID.