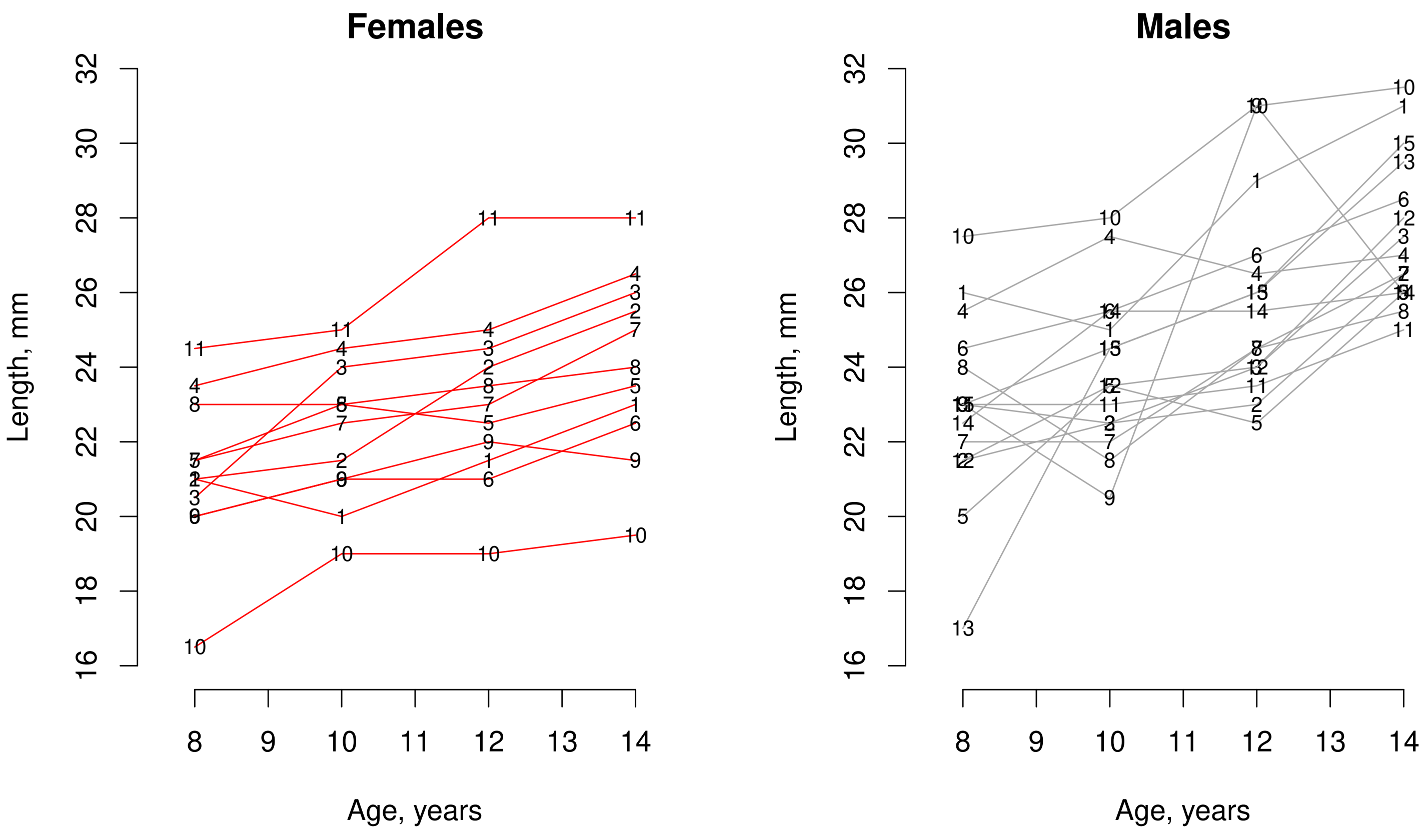
Part VI:

Missing data

- Recall the dental growth data:



- An important feature of this dataset is that all K=26 children have *complete* data
  - ⋆ complete in the sense that dental length is observed at each of four prescribed follow-up times

Psychiatric trial data

- Schizophrenia Collaborative Study
  - ⋆ treatment trial for schizophrenia conducted by NIMH
  - ⋆ three active drugs vs placebo
  - ⋆ see Hedeker and Gibbons (*Psychological Methods*, 1997)

- Response is the *Inpatient Multidimensional Psychiatric Scale item 79*, an ordinal measure of disease severity:

  1: normal, not ill at all

  2: borderline mentally ill

  ⋮

  7: among the most extremely ill

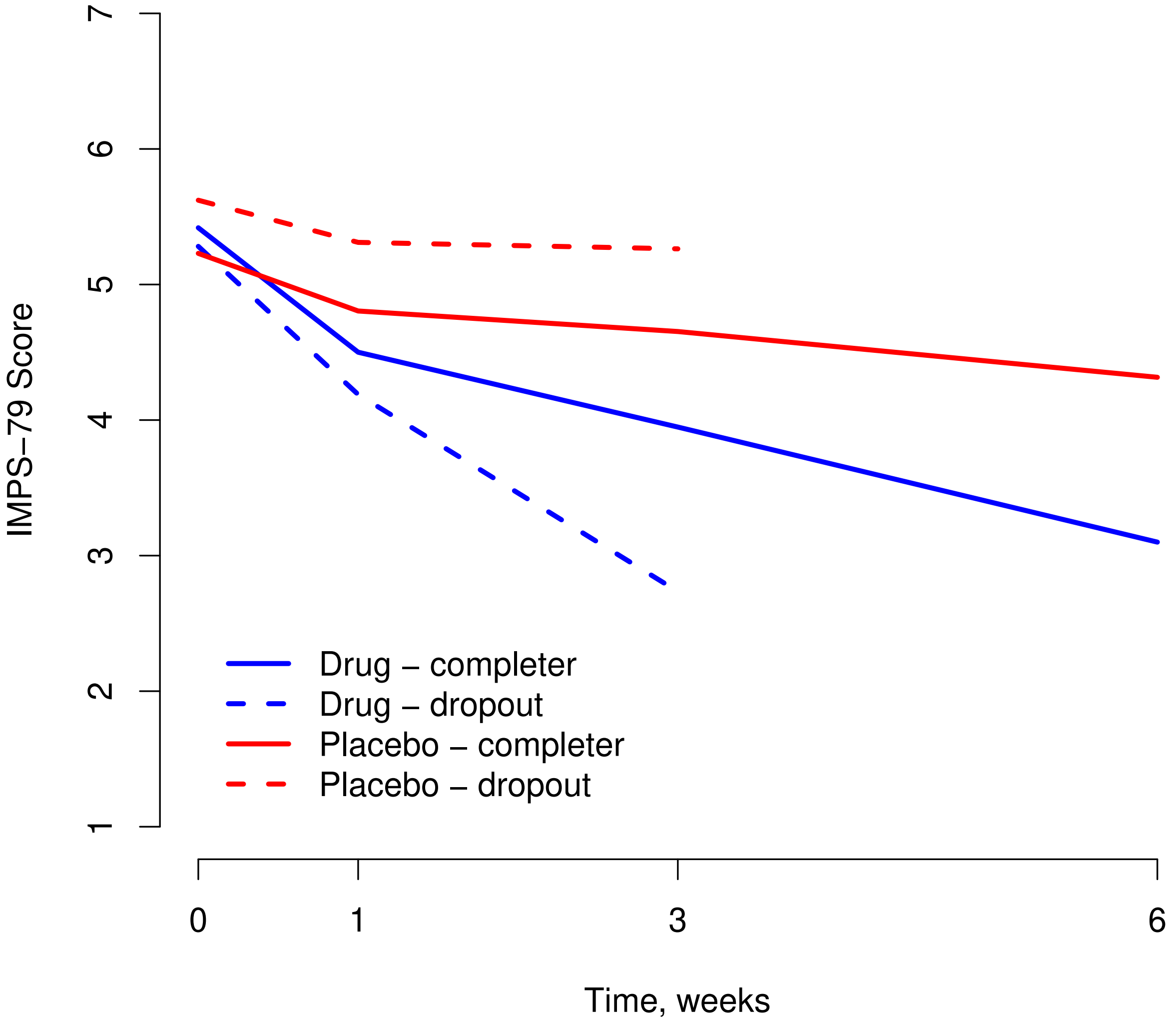
- Data on K=437 patients
  - ⋆ up to four measurements at weeks 0, 1, 3, and 6

```
>
> ##
> load("nimh.Rdata")
>
> ##
> head(nimh)
    id IMPS79 week drug sex
1 1103    5.5    0    1   1
2 1103    3.0    1    1   1
3 1103    2.5    3    1   1
4 1103    4.0    6    1   1
5 1104    6.0    0    1   1
6 1104    3.0    1    1   1
>
> ##
> dim(nimh)
[1] 1603    5
>
> ##
> ids <- unique(nimh$id)
> length(ids)
[1] 437
```

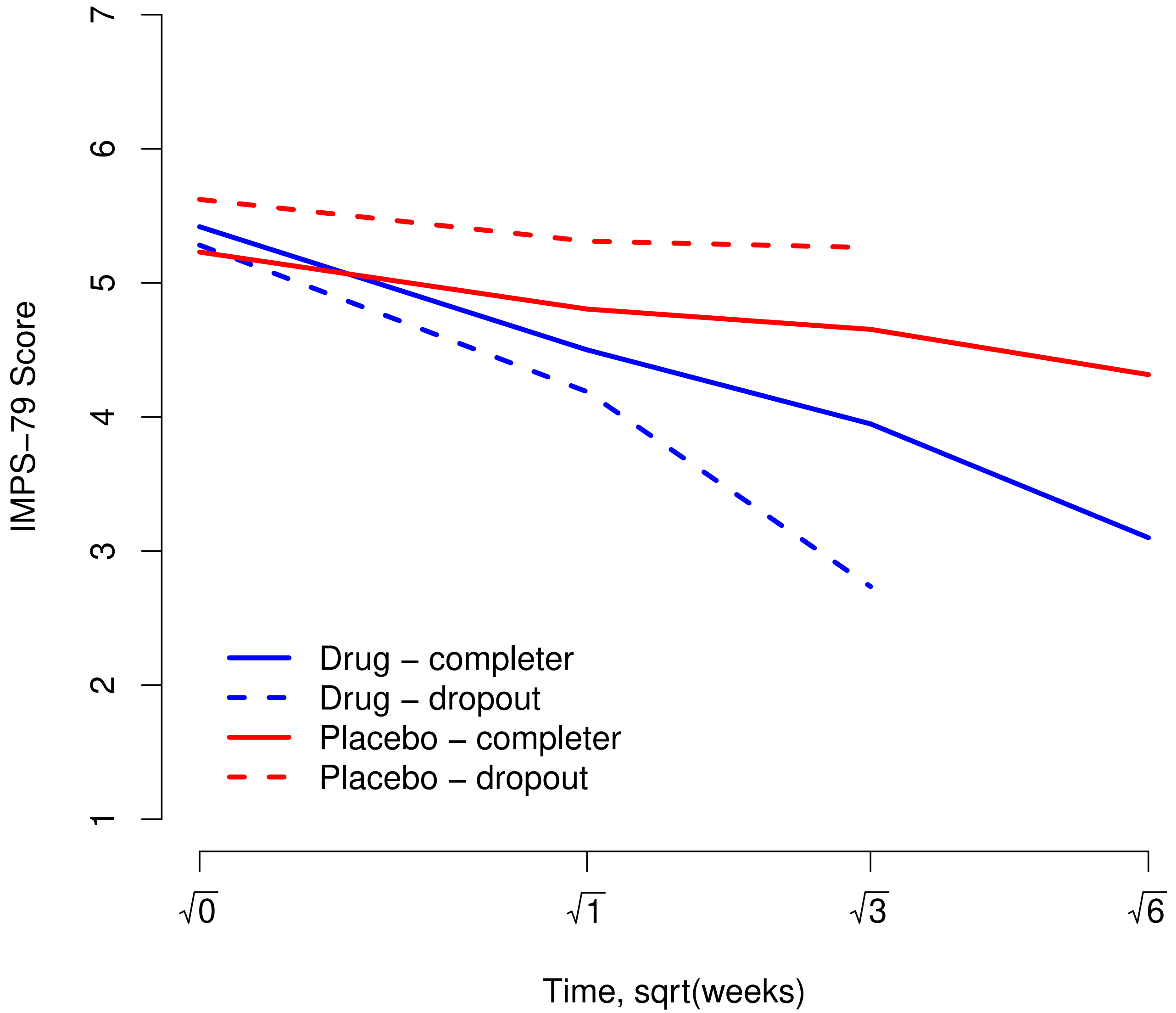
```
>
> ##
> table(nimh$drug, nimh$week)

      0   1   2   3   4   5   6
  0 107 105   5  87   2   2  70
  1 327 321   9 287   9   7 265
>
> nimh$week[nimh$week == 2] <- 3
> nimh$week[nimh$week == 4] <- 3
> nimh$week[nimh$week == 5] <- 6
>
> table(nimh$drug, nimh$week)

      0   1   3   6
  0 107 105  94  72
  1 327 321 305 272
>
> ## Define a "completer" as anyone who makes it to week 6
> ##
> idsC <- unique(nimh$id[nimh$week == 6])
> nimh$completer <- as.numeric(is.element(nimh$id, idsC))
```

```
>
> ## Overall drop-out rate by treatment group
> ##
> nimh0 <- nimh[nimh$week == 0,]
> round(1 - tapply(nimh0$completer, list(nimh0$drug), FUN=mean), 3) * 100
   0    1
33.6 17.7
>
> ##
> nimh$type <- NA
> nimh$type[nimh$drug == 1 & nimh$completer == 1] <- "Drug-Comp"
> nimh$type[nimh$drug == 1 & nimh$completer == 0] <- "Drug-Drop"
> nimh$type[nimh$drug == 0 & nimh$completer == 1] <- "Plac-Comp"
> nimh$type[nimh$drug == 0 & nimh$completer == 0] <- "Plac-Drop"
>
> tabY <- tapply(nimh$IMPS79, list(nimh$type, nimh$week), FUN=mean)
```

- Average response, as a function of time (weeks)



- Average response, as a function of time (square-root of weeks)



- Focusing on the 335 patients who completed the study, the figures suggest a beneficial effect of treatment
  - ⋆ referred to as a ‘complete-case analysis’

- Comparing those who dropped out to those who completed the study:
  - ⋆ patients who dropped out from the placebo arm tended to have **higher** IMPS-79 values (prior to dropout) than those patients in the placebo arm that completed the study
  - ⋆ patients who dropped out from the drug arm tended to have **lower** IMPS-79 values (prior to dropout) than those patients in the drug arm that completed the study

- Had the patients who dropped out been retained in the study, it’s plausible to suppose that the treatment effect would be even greater than what is observed when one focuses on the completers

Taxonomy for missing data

- Except in highly controlled settings, missing data in longitudinal studies will be inevitable
  - ⋆ generally, complete data is the exception rather than the rule

- Until now, however, we have not paid any specific attention to the problem of missing data

Q: What is the potential impact of missing data?

- The main impact is bias, sometimes referred to as *selection bias*
  - ⋆ some patients are 'selected' to have complete data
  - ⋆ if these patients are not representative of the population of interest, then results based on solely on their information may not generalize to the population of interest

- Before thinking about methods to resolve selection bias we need:

  (1) a taxonomy to help distinguish the various types of missingness one might encounter, and

  (2) a means of characterizing when missing data is a problem and when its not a problem

  ∗ it seems clear that we cannot ignore the fact that some patients in the schizophrenia treatment trial dropped out, although this will not always be the case

Patterns of missing data

- As a first step, it is useful to distinguish patterns of missing data

Dropout

Pattern	Y_{k1}	Y_{k2}	Y_{k3}
# 1	✓	✓	✓
# 2	✓	✓	
# 3	✓		

⋆ if Y_{ki} is missing then so are $Y_{k(i+1)}, Y_{k(i+2)}, \ldots$

⋆ *monotone missingness*

Intermittent missingness

Pattern	Y_{k1}	Y_{k2}	Y_{k3}
# 1	✓	✓	✓
# 2	✓	✓	
# 3	✓		✓
# 4	✓		
# 5		✓	✓
# 6		✓	
# 7			✓
# 8			

⋆ missing data can occur anywhere

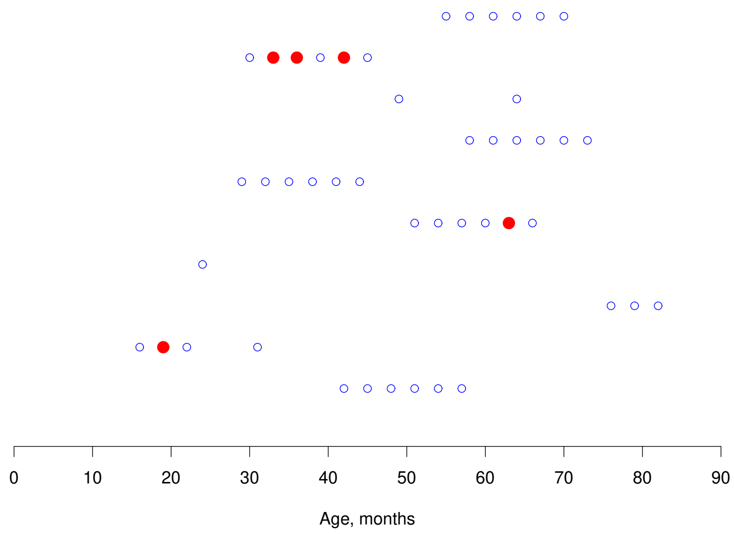
⋆ *non-monotone missingness*

- As an example of intermittent missingness, recall the ICHS data
  - ⋆ relationship between risk of respiratory infection and vitamin A deficiency

- These data are *incomplete* in the sense that not all children have all six measurements that were 'prescribed' by the design:

# measurements	1	2	3	4	5	6
# children	22	32	29	55	15	122

- It is also clear that children with missing data have varying patterns of missingness
  - ⋆ miss a particular visit but then subsequently come back
  - ⋆ i.e. intermittent missingness



Notation

- In order to formally define the notion of ‘missing’ data we have to first define what we mean by ‘complete’ data
  - ⋆ i.e. the number and timing of the measurements in a longitudinal setting

- For simplicity, we are going to suppose that the design of the longitudinal study was such that n follow-up measurements were to be collected on each subject at some prescribed sequence of time points

- Given this, complete data for the k^{th} subject would consist of n measurements:

$$\begin{aligned}\boldsymbol{Y}_k &= (Y_{k1},\ Y_{k2},\ \ldots,\ Y_{kn})^T \\ \boldsymbol{X}_k &= (\boldsymbol{X}_{k1},\ \boldsymbol{X}_{k2},\ \ldots,\ \boldsymbol{X}_{kn})^T\end{aligned}$$

where $\boldsymbol{X}_{ki} = (X_{ki,1},\ X_{ki,2},\ \ldots,\ X_{ki,p})$

- Moving forward we are going to assume $\boldsymbol{X}_k$ is fully observed
  - reasonable when $\boldsymbol{X}_{ki}$ consists solely of baseline covariates or when the measurement of exposures do not rely on the same (logistical) process as the outcome
    - e.g. air pollution in environmental studies
    - e.g. diagnoses from an EHR

- In addition, we are going to assume that data may be available on additional covariates that are themselves not of substantive interest but may be relevant to missingness

  Ancillary variable
  - measured at the same time as Y_{ki} and, therefore, could be time-varying
  - denoted $\boldsymbol{Z}_{ki}$

- An example of when such data might be available is in a randomized trial where information on the patient is collected post-randomization

- To distinguish which responses are observed and which are missing, one can partition the response for the k^{th} patient as $\boldsymbol{Y}_k = \{\boldsymbol{Y}_k^o, \boldsymbol{Y}_k^m\}$
  - ⋆ if $\boldsymbol{Y}_k$ is completely observed then $\boldsymbol{Y}_k^o = \boldsymbol{Y}_k$ and $\boldsymbol{Y}_k^m = \emptyset$

- To identify which specific elements of $\boldsymbol{Y}_k$ are observed, define:

missing indicator

$$R_{ki} = \begin{cases} 1 & \text{if } Y_{ki} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

and let $\boldsymbol{R}_k = (R_{k1}, \ldots, R_{kn})$

- In the special case of dropout, we can simplify the notation by letting

$$D_k = 1 + \sum_{i=1}^{n} R_{ki}$$

denote the first time at which Y_{ki} is missing in the k^{th} cluster

only useful when your data is monotony missing

- Simple example with $n = 3$:

Pattern	Response vector, $\boldsymbol{Y}_k$			Missingness indicators, $\boldsymbol{R}_k$			Dropout time
	Y_{k1}	Y_{k2}	Y_{k3}	R_{k1}	R_{k2}	R_{k3}	D_k
# 1	✓	✓	✓	1	1	1	4
# 2	✓	✓		1	1	0	3
# 3	✓			1	0	0	2

- Finally, it will eventually be convenient to characterize the totality of information that is potentially observable at a particular time point

  - ⋆ let $\boldsymbol{W}_{ki} = (\boldsymbol{Z}_{ki}, Y_{ki})$, for $i = 1, \ldots, n$, and

    $$\overline{\boldsymbol{W}}_{ki} \;=\; (\boldsymbol{W}_{k0},\ \boldsymbol{W}_{k1}, \ldots, \boldsymbol{W}_{k,i-1})$$

    where $\boldsymbol{W}_{k0} = \boldsymbol{X}_k$

  - ⋆ $\overline{\boldsymbol{W}}_{ki}$ is the k^{th} subjects' "history" up to the i^{th} time point

Missing data mechanisms

- As one considers missingness, and its potential impact, one can think of $\boldsymbol{R}_k$ or D_k as the 'response' that could be modelled
  - ⋆ e.g. a logistic regression for R_{ki} as a function of $\boldsymbol{Y}_k$, $\boldsymbol{X}_k$ and/or $\boldsymbol{Z}_k$

- Often refer to such a model as representing the *missingness mechanism*, although it might be better referred to as the *observance mechanism* or as the *selection mechanism*

- Taxonomy for missing data mechanisms due to Rubin (1976):
  - ⋆ missing completely at random (MCAR) toss a coin to determine whether to go class or not
  - ⋆ missing at random (MAR) low temp less likely to go to the morning class
  - ⋆ missing not at random (MNAR)

Missing completely at random (MCAR)

- Missingness does not depend on either the outcome or on covariates that are of substantive interest

- Examples:
  - loss-to-follow-up because of unexpected life events that are unrelated to the study
  - follow-up ends due to funding considerations
  - data from a particular file is corrupted

- While it is, of course, possible that data is missing due to innocuous reasons, it is very hard to characterize systematically (i.e. in a model)

- MCAR is sometimes referred to as *ignorable* or *non-informative* missingness

Missing at random (MAR)

- Missingness depends on the outcome and/or covariates that are of substantive interest solely through what is actually observed

- Examples:
  - ⋆ older patients have a higher chance of dropping out
  - ⋆ patients under a particular treatment have a higher chance of dropping out because of the increased risk of adverse events
  - ⋆ patients selected to have ‘complete’ data in a case-control study

- The extent to which $\boldsymbol{R}_k$ depends on observed quantities can be evaluated empirically

- MAR is also sometimes referred to as *ignorable* or *non-informative* missingness

Missing not at random (MNAR)

- Missingness depends, in part at least, on unobserved outcomes and/or covariates that are of substantive interest

- Examples:
  - ⋆ patients who are experiencing poor outcomes are more likely to drop out
  - ⋆ patients refuse to be weighed because they recently lost/gained a significant amount of weight

- The extent to which $\boldsymbol{R}_k$ depends on quantities that are not observed cannot be evaluated empirically

- MNAR is sometimes referred to as *nonignorable* or *informative* missingness

Possible analysis strategies

- After consideration of the nature of the missing data (i.e. why some patients might have complete data and others not), the statistical literature provides a huge number of options for performing an analysis, including:

  1. Complete-case analysis
  2. Imputation
  3. Inverse-probability weighting (IPW)
  4. Likelihood-based methods
  5. Other *ad hoc* methods, such as last observation carried forward or including an indicator for missing value

- In the remainder of Part VI, we are going to focus on IPW and likelihood-based methods
  - ⋆ actually only focus on select methods within these classes of methods

Inverse-probability weighting

- Suppose that if complete data on all K study participants had been available, the analysis would proceed by:

  (1) specifying a marginal model for the mean, and

  (2) performing estimation/inference via GEE

- Recall the estimating equations for $\boldsymbol{\beta}$ from GEE 1.0 (and GEE 1.5):

$$\sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \;=\; \mathbf{0}$$

- Assuming the mean model is correctly specified:

CAN: consistent and asymptotically normal

  ⋆ the solution, $\widehat{\boldsymbol{\beta}}$, is consistent and asymptotically Normal

  ⋆ the sandwich variance estimator is (asymptotically) valid regardless of whether the working covariance structure is correct

Complete-case GEE

- Suppose only a sub-sample of study participants have complete outcome data
  - ⋆ assume information on $\boldsymbol{X}_k$ is available $\forall$ K participants
  - ⋆ those with incomplete data could be either drop-outs or have intermittent missing observations

- One way forward is a *complete-case* analysis
  - ⋆ i.e. restrict to those with $\boldsymbol{R}_k = \mathbf{1} = (1, \ldots, 1)$ or, equivalently, for whom $D_k = n + 1$

- Represent this analysis via the *complete-case GEE*:

$$\sum_{k=1}^{K} I(\boldsymbol{R}_k = \mathbf{1}) \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} (\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \ = \ \mathbf{0}$$

where $I(\boldsymbol{R}_k = \mathbf{1})$=0/1 indicates whether the k^{th} subject has complete data

- ⋆ also referred to as *unweighted complete-case GEE*

Q: Are the complete-case GEE unbiased?

Only randomness

- Consider the expectation with respect to the joint distribution of $(\boldsymbol{Y}, \boldsymbol{R})$
  - ⋆ note, conditioning on $\boldsymbol{X}_k$ is implicit throughout

- For the k^{th} subject, we have:

Check whether the estimating equations has mean zero

$$
\begin{aligned}
\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[I(\boldsymbol{R}_k = \mathbf{1})\boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)]& \\
&= \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[I(\boldsymbol{R}_k = \mathbf{1})(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&= \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}[\mathsf{E}_{\boldsymbol{R}|\boldsymbol{Y}}\{I(\boldsymbol{R}_k = \mathbf{1})(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)\}] \\
&= \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}[\mathsf{E}_{\boldsymbol{R}|\boldsymbol{Y}}\{I(\boldsymbol{R}_k = \mathbf{1})\}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&= \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}[P(\boldsymbol{R}_k = \mathbf{1}|\,\boldsymbol{Y}_k)(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)]
\end{aligned}
$$

If this part dndo Y

- To continue, we need to say something about $P(\boldsymbol{R}_k = \mathbf{1}|\,\boldsymbol{Y}_k)$

- If the missingness mechanism is MCAR then

$$\Pr(\boldsymbol{R}_k = \mathbf{1} | \, \boldsymbol{Y}_k) \;=\; \Pr(\boldsymbol{R}_k = \mathbf{1})$$

so that, if the mean model is correctly specified, we have:

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[I(\boldsymbol{R}_k = \mathbf{1})\boldsymbol{D}_k^T \boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \mathsf{E}_{\boldsymbol{Y}}[P(\boldsymbol{R}_k = \mathbf{1} | \, \boldsymbol{Y}_k)(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \mathsf{E}_{\boldsymbol{Y}}[P(\boldsymbol{R}_k = \mathbf{1})(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} P(\boldsymbol{R}_k = \mathbf{1}) \underbrace{\mathsf{E}_{\boldsymbol{Y}}[(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)]}_{= \mathbf{0}} \;=\; \mathbf{0}
\end{aligned}$$

- Consequently, if missingness is MCAR the complete-case analysis yields:
  - ⋆ point estimates are consistent
  - ⋆ robust standard error estimates are valid

Q: What happens when estimating functions do not have zero mean?
A; No consistency, no worries about the standard error
Big data paradox

- As with full data GEE, this holds regardless of whether the dependence structure is correctly specified

- If the missingness is either MAR or MNAR, however, then

$$\begin{aligned}\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[I(\boldsymbol{R}_k = \mathbf{1})\boldsymbol{D}_k^T \boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\ &= \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}[P(\boldsymbol{R}_k = \mathbf{1}|\, \boldsymbol{Y}_k)(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\ &\neq \mathbf{0}\end{aligned}$$

  ⋆ i.e. the complete-case estimating equations will not, in general, be unbiased

- Consequently, a complete-case GEE analysis will, in general, yield point estimates of $\boldsymbol{\beta}$ that are not guaranteed to be consistent

  ⋆ i.e. one cannot naïvely forge ahead with a GEE analysis if there is missing data, unless the missingness is completely innocuous

IPW complete-case GEE

- If the missingness is MAR then one can use IPW to construct an unbiased estimating equation

- Specifically, consider the following *IPW complete-case GEE*:

This whole part is 1

$$\sum_{k=1}^{K} \frac{I(\boldsymbol{R}_k = \mathbf{1})}{\pi_k} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} (\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \;=\; \mathbf{0}$$

  where π_k is the probability of being a completer

  ⋆ note, this is only well-defined if $\pi_k > 0 \;\forall\; k$

- Under MAR, π_k is taken to depend on the response and covariates of substantive interest solely through what is observed:

$$\pi_k \;=\; \Pr(\boldsymbol{R}_k = \mathbf{1} |\; \boldsymbol{Y}_k,\; \boldsymbol{X}_k,\; \boldsymbol{Z}_k) \;=\; \Pr(\boldsymbol{R}_k = \mathbf{1} |\; \boldsymbol{Y}_k^o,\; \boldsymbol{X}_k,\; \boldsymbol{Z}_k)$$

  ⋆ dependence on $\boldsymbol{Z}_k$ is included to provide the most general setting

- We then have:

$$
\begin{aligned}
& \mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}\left[\frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k}\boldsymbol{D}_k^T\ \boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)\right] \\
&\qquad = \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}\left[\frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k}(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)\right] \\
&\qquad = \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}\left[\mathsf{E}_{\boldsymbol{R}|\ \boldsymbol{Y}}\left\{\frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k}(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)\right\}\right] \\
&\qquad = \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}\left[\mathsf{E}_{\boldsymbol{R}|\ \boldsymbol{Y}}\left\{\frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k}\right\}(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)\right] \\
&\qquad = \boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}\mathsf{E}_{\boldsymbol{Y}}[(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)] \\
&\qquad = \mathbf{0}
\end{aligned}
$$

- Consequently, under MAR, the IPW complete-case estimating equations are unbiased
  - ⋆ yields point estimates that are consistent and sandwich-based standard errors that are valid (asymptotically)

- To see how/why IPW works, suppose, for example, that $\pi_k = 0.5$:
  - ⋆ only expect, on average, half of subjects with the same observed data profile as the k^{th} subject to complete the study
  - ⋆ weighting the (observed) contribution of the k^{th} subject by $\pi_k^{-1} = 2$ amounts to saying that we would have seen another person with this profile (i.e. a total of 2 such subjects) in the complete data

- Weighting therefore serves make up for the fact that some subjects are missing by replicating the contributions from those who do have complete data, in an effort to mimic the results that one would have obtained if all K subjects had complete data (on average, at least)

- Note, since $\pi_k \in (0, 1)$, it will always be the case that $\pi_k^{-1} > 1$, so that each subject is *up-weighted*
  - ⋆ if π_k is small then the weight is (relatively) large
  - ⋆ if π_k is large then the weight is (relatively) small

IPW available-data GEE

- Although IPW complete-case GEE yields valid estimation/inference under MAR, it is often inefficient because it ignores the partial information from subjects who drop out part way through the study

- Recall the simple example with $n=3$:

Pattern	Response vector, $\boldsymbol{Y}_k$			Missingness indicators, $\boldsymbol{R}_k$			Dropout time
	Y_{k1}	Y_{k2}	Y_{k3}	R_{k1}	R_{k2}	R_{k3}	D_k
# 1	✓	✓	✓	1	1	1	4
# 2	✓	✓		1	1	0	3
# 3	✓			1	0	0	2

⋆ only subjects with pattern #1 are included in a complete-case analysis

⋆ information from subjects with patterns #2 or #3 are ignored and yet they do provide some information Not efficient

- We can make use of all of the available data using a framework developed by Robins, Rotnitzky and Zhao (JASA, 1995)

- The essential idea is to move beyond *subject-specific* weighting (i.e. by π_k^{-1}) to *observation-specific weighting*

- Before getting into the details, recall that $\overline{\boldsymbol{W}}_{ki} = (\boldsymbol{W}_{k0},\ \boldsymbol{W}_{k1}, \ldots, \boldsymbol{W}_{k,i-1})$, with $\boldsymbol{W}_{k0} = \boldsymbol{X}_k$ and $\boldsymbol{W}_{ki} = (\boldsymbol{Z}_{ki}, Y_{ki})$, denotes a subjects 'history' up to the i^{th} time point

- Moving forward, we make two key assumptions:

(A1) $\Pr(R_{ki} = 1 |\ R_{k,i-1} = 1, \overline{\boldsymbol{W}}_{ki}, \boldsymbol{Y}_k) \ = \ \Pr(R_{ki} = 1 |\ R_{k,i-1} = 1, \overline{\boldsymbol{W}}_{ki})$

$\ast$ the missing data process is MAR or ignorable

(A2) $\Pr(R_{ki} = 1 |\ R_{k,i-1} = 1, \overline{\boldsymbol{W}}_{ki}) \ > \ 0$

$\ast$ the probability of remaining in the study is positive

- Notationally, it will be convenient to let

$$\lambda_{ki} \;=\; \Pr(R_{ki}=1|\; R_{k,i-1}=1, \overline{\boldsymbol{W}}_{ki})$$

and

$$\pi_{ki} \;=\; \lambda_{k1} \;\times\; \ldots \;\times\; \lambda_{ki}$$

⋆ π_{ki} is the conditional probability of observing subject k at the i^{th} time point, given their observed history up to that time

- Consider the *IPW available-data GEE*:

$$\sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{\Delta}_k (\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \;=\; \mathbf{0}$$

where $\boldsymbol{\Delta}_k = \mathcal{R}_k \mathcal{W}_k$ Weighting matrix

⋆ $\mathcal{R}_k$ is an $n \times n$ diagonal matrix with elements $I(R_{ki}=1)$ on the diagonal indicating whether or not the i^{th} measurement is observed

⋆ $\mathcal{W}_k$ is an $n \times n$ diagonal matrix with elements π_{ki}^{-1} on the diagonal

- Using the same arguments as before, assuming the mean model is correctly specified and that both (A1) and (A2) hold, it is straightforward to see that the IPW available-data estimating equations are unbiased:

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[\boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{\Delta}_k(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \mathsf{E}_{\boldsymbol{Y},\boldsymbol{R}}[\boldsymbol{\Delta}_k(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \mathsf{E}_{\boldsymbol{Y}}[\mathsf{E}_{\boldsymbol{R}|\ \boldsymbol{Y}}\{\boldsymbol{\Delta}_k(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)\}] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \mathsf{E}_{\boldsymbol{Y}}[\underbrace{\mathsf{E}_{\boldsymbol{R}|\ \boldsymbol{Y}}\{\boldsymbol{\Delta}_k\}}_{=\boldsymbol{I}_n}(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)] \\
&\qquad = \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \underbrace{\mathsf{E}_{\boldsymbol{Y}}[(\boldsymbol{Y}_k - \boldsymbol{\mu}_k)]}_{=\mathbf{0}} = \mathbf{0}
\end{aligned}$$

- Consequently, while the estimating function has been constructed specifically to use all available information it retains the crucial property that the estimating equations are unbiased

- Under the conditions laid out, therefore, IPW available-case GEE can be used as a basis for valid estimation/inference in the presence of ignorable dropout while enjoying improved efficiency properties

Estimation of π_{ki}

- In most settings, the observance probabilities

$$\lambda_{ki} \;=\; \Pr(R_{ki} = 1 |\; R_{k,i-1} = 1, \overline{\boldsymbol{W}}_{ki})$$

  are unknown and, therefore, need to be estimated

- Typically, however, it is assumed that the λ_{ki} are known up to finite vector of unknown parameters, say $\boldsymbol{\theta}$
  - ⋆ write $\lambda_{ki} = \lambda_{ki}(\boldsymbol{\theta})$

- For example, one could assume a logistic regression for the binary R_{ki}
  - ⋆ elements of $\boldsymbol{\theta}$ are, as usual, the baseline log-odds and odds ratios
  - ⋆ include covariates that are deemed relevant to dropout process, perhaps via hypothesis testing
  - ⋆ may include time-specific intercepts
    - $*$ analogous to a discrete time survival analysis

- For a given model specification, estimates of $\boldsymbol{\theta}$ can be obtained via maximization of the partial likelihood:

$$\mathcal{L}^p(\boldsymbol{\theta}) \;=\; \prod_{k=1}^{K} \mathcal{L}_k^p(\boldsymbol{\theta}) \;=\; \prod_{k=1}^{K}\prod_{i=1}^{n} \left[\lambda_{ki}(\boldsymbol{\theta})^{R_{ki}} \left\{1 - \lambda_{ki}(\boldsymbol{\theta})\right\}^{1-R_{ki}}\right]^{R_{k,i-1}}$$

  with $R_{k,0} \equiv 1$

  $\star$ denote the corresponding estimate as $\widehat{\boldsymbol{\theta}}$

- Plugging in $\widehat{\boldsymbol{\theta}}$ into $\lambda_{ki}(\boldsymbol{\theta})$, an estimate of π_{ki} is

$$\widehat{\pi}_{ki} \;=\; \lambda_{k1}(\widehat{\boldsymbol{\theta}}) \;\times\; \ldots \;\times\; \lambda_{ki}(\widehat{\boldsymbol{\theta}})$$

Q: Can we just plug these estimates into the available-case estimating equations? Are there any consequences?

Robins, Rotnitzky and Zhao (1995)

- Consider the following IPW available-case estimating equations that use estimated weights:

$$\boldsymbol{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}) \;=\; \sum_{k=1}^{K} \boldsymbol{U}_k(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}) \;=\; \sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{\Delta}_k(\widehat{\boldsymbol{\theta}})(\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \;=\; \boldsymbol{0}$$

  where $\boldsymbol{\Delta}_k(\widehat{\boldsymbol{\theta}}) = \mathcal{R}_k \widehat{\mathcal{W}}_k$, with $\mathcal{R}_k$ as before and $\widehat{\mathcal{W}}_k$ is an $n \times n$ diagonal matrix with elements $\widehat{\pi}_{ki}^{-1}$ on the diagonal

- Theorem 1 of RRZ (1995) states that, assuming that various regularity conditions hold, that the mean model is correctly specified and that both (A1) and (A2) hold:
  - ⋆ there exists a unique solution, $\widehat{\boldsymbol{\beta}}$, to the equation $\boldsymbol{U}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}) = \boldsymbol{0}$
  - ⋆ $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically Normally distributed

- Theorem 1 also gives the form of the asymptotic variance as well as a consistent estimate, specifically:

$$\widehat{\mathsf{Cov}}[\widehat{\boldsymbol{\beta}}] \;=\; \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{A}}^{-1}$$

where

$$\widehat{\boldsymbol{A}} \;=\; \sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{\Delta}_k(\widehat{\boldsymbol{\theta}}) \boldsymbol{D}_k$$

$$\widehat{\boldsymbol{B}} \;=\; \sum_{k=1}^{K} \boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}) \boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}})^T$$

with $\boldsymbol{U}_k^*(\boldsymbol{\beta},\boldsymbol{\theta})$ given as:

Projection Matrix

$$\boldsymbol{U}_k(\boldsymbol{\beta},\boldsymbol{\theta}) \;-\; \left(\sum_{k=1}^{K} \boldsymbol{U}_k(\boldsymbol{\beta},\boldsymbol{\theta}) \boldsymbol{S}_k(\boldsymbol{\theta})^T\right) \left(\sum_{k=1}^{K} \boldsymbol{S}_k(\boldsymbol{\theta}) \boldsymbol{S}_k(\boldsymbol{\theta})^T\right)^{-1} \boldsymbol{S}_k(\boldsymbol{\theta})$$

where

$$\boldsymbol{S}_k(\boldsymbol{\theta}) \;=\; \frac{\partial}{\partial \boldsymbol{\theta}} \log \mathcal{L}^p(\boldsymbol{\theta})$$

- It's interesting to contrast the asymptotic variance with the form of the variance that might naïvely be used if we treated the $\widehat{\boldsymbol{\theta}}$ as fixed and 'known':

$$\widehat{\text{Cov}}_{\text{naïve}}[\widehat{\boldsymbol{\beta}}] \;=\; \widehat{\boldsymbol{A}}^{-1}\widehat{\boldsymbol{B}}_{\text{naïve}}\widehat{\boldsymbol{A}}^{-1}$$

where

$$\begin{aligned} \widehat{\boldsymbol{A}} &= \sum_{k=1}^{K} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{\Delta}_k(\widehat{\boldsymbol{\theta}}) \boldsymbol{D}_k \\ \widehat{\boldsymbol{B}}_{\text{naïve}} &= \sum_{k=1}^{K} \boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) \boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})^T \end{aligned}$$

- So, the key difference between this naïve expression and the correct expression is in the adjustment made to the 'cheese' part of the sandwich
  - ⋆ i.e. $\boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ versus $\boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$

- By inspecting the form of $\boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$, we can see that it corresponds to the residual from a multivariate linear regression analysis with $\boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ as the 'outcome' and $\boldsymbol{S}_k(\widehat{\boldsymbol{\theta}})$ as the 'covariates'

- By definition, the *residual sum of squares* given by

$$\widehat{\boldsymbol{B}} = \sum_{k=1}^{K} \boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) \boldsymbol{U}_k^*(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})^T$$

must be smaller (or at least no larger) than the *sum of squares* given by

$$\widehat{\boldsymbol{B}}_{\text{naïve}} = \sum_{k=1}^{K} \boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) \boldsymbol{U}_k(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})^T$$

so that $\text{Cov}[\widehat{\boldsymbol{\beta}}]$ will be smaller (or at least no larger) than $\text{Cov}_{\text{naïve}}[\widehat{\boldsymbol{\beta}}]$

- Consequently, and somewhat counter-intuitively, use of the naïve standard error estimates (that fail to account for the fact that $\boldsymbol{\theta}$ is estimated) will generally lead to standard errors that are too large
  - ⋆ i.e. inference is conservative
  - ⋆ estimating the weights, plugging them in and using the standard formula only leads to a loss of precision/power

    Do not use naive version of standard error

Augmented IPW GEE

- So far we have focused on settings in which the missingness is MAR (or ignorable)

Q: What if the missingness is in fact or is suspected to be MNAR (or nonignorable)?

- Consider, for example, the following model for observance:

$$\text{logit } \lambda_{ki} \; = \; \theta_{0i} \; + \; \theta_1 Y_{k,i-1} \; + \; \theta_2 Y_{ki}$$

  - ⋆ $\{\theta_{0i}; i = 1, \ldots, n\}$ represent variation in the probability of observance over time
  - ⋆ θ_1 and θ_2 represent the extent to which whether or not a subject drops out depends on the *previous* and *current* responses, respectively

- Unfortunately we do not observe Y_{ki} for patients who drop out at the i^{th} time point, so θ_2 is not identifiable

Q: How do we move forward?

- Rotnitzky, Robins and Scharfstein (1998) and Scharfstein, Rotnitzky and Robins (1999) proposed a framework for sensitivity analyses when missingness is potentially nonignorable, based on a class of *augmented IPW estimators*
  - ⋆ can be used for both intermittent missingness and dropout

- Towards formalizing estimation/inference when missingness is MNAR, recall that $\boldsymbol{W}_{k0} = \boldsymbol{X}_k$ and $\boldsymbol{W}_{ki} = (\boldsymbol{Z}_{ki}, Y_{ki})$ and let

$$\overline{\boldsymbol{W}}_k \;=\; (\boldsymbol{W}_{k0},\; \boldsymbol{W}_{k1}, \ldots, \boldsymbol{W}_{kn})$$

denote the totality of potentially observable information on the k^{th} subject

- Furthermore, let $\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}$ denote the subvector of $\overline{\boldsymbol{W}}_k$ that is actually observed when $\boldsymbol{R}_k = \boldsymbol{r}$
  - ⋆ e.g. if $\boldsymbol{r}$ = (1, 1, 1, 0, ..., 0), then $\overline{\boldsymbol{W}}_{(\boldsymbol{r})k} = (\boldsymbol{W}_{k0},\; \boldsymbol{W}_{k1}, \boldsymbol{W}_{k2}, \boldsymbol{W}_{k3})$

- Let $\pi_k(\boldsymbol{r})$ denote the conditional probability that the k^{th} subject experiences observance pattern $\boldsymbol{r}$, given information on all their covariates at all time points:

$$\pi_k(\boldsymbol{r}) \;=\; \Pr(\boldsymbol{R}_k = \boldsymbol{r} |\, \overline{\boldsymbol{W}}_k)$$

- Letting $\overline{\boldsymbol{R}}_{ki} = (R_{k1}, \ldots, R_{k,i-1})$, consider the factorization of $\pi_k(\boldsymbol{r})$ given by:

$$\begin{aligned} \pi_k(\boldsymbol{r}) \;&=\; \prod_{i=1}^{n} \Pr(R_{ki} = 1 |\, \overline{\boldsymbol{R}}_{ki}, \overline{\boldsymbol{W}}_k)^{r_i} \\ &\qquad\qquad \times\; \Pr(R_{ki} = 0 |\, \overline{\boldsymbol{R}}_{ki}, \overline{\boldsymbol{W}}_k)^{1-r_i} \\ &=\; \prod_{i=1}^{n} \overline{\lambda}_{ki}{}^{r_i} \;\times\; (1 - \overline{\lambda}_{ki})^{1-r_i} \end{aligned}$$

- As with specification of λ_{ki} in the factorization of π_{ki} in IPW available-data GEE (see slide 446), it is typical to assume some parametric model for the $\overline{\lambda}_{ki}$, as a function of some finite vector $\boldsymbol{\theta}$

- As in the example on slide 456, if $\overline{\lambda}_{ki}$ depends on current (or future) outcomes, then $\boldsymbol{\theta}$ is not identifiable

- Arguably, therefore, the most reasonable way forward is to conduct a sensitivity analysis
  - ⋆ fix components of $\boldsymbol{\theta}$ that are not identified and estimate the remaining components of $\boldsymbol{\theta}$
    - ∗ i.e. fix θ_2 in the example on slide 456, and estimate the θ_{0i} and θ_1
  - ⋆ perform the IPW analysis and record the results
  - ⋆ examine sensitivity across a range of plausible values for θ_2

- Before considering this, recall the IPW complete-case GEE:

$$\sum_{k=1}^{K} \frac{I(\boldsymbol{R}_k = \mathbf{1})}{\pi_k(\mathbf{1})} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} (\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \ = \ \mathbf{0}$$

  - ⋆ only subjects with complete data contribute

- RRS (1998) consider the *augmented IPW GEE*:

$$\sum_{k=1}^{K}\left[\frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k(\mathbf{1})}\boldsymbol{D}_k^T\boldsymbol{V}_k^{-1}(\boldsymbol{Y}_k-\boldsymbol{\mu}_k)\ +\ \boldsymbol{A}_k\right]\ =\ \mathbf{0}$$

where the vector $\boldsymbol{A}_k$ is the *augmentation* term:

$$\boldsymbol{A}_k\ =\ \sum_{\boldsymbol{r}\neq\mathbf{1}}\left\{I(\boldsymbol{R}_k=\boldsymbol{r})\ -\ \frac{I(\boldsymbol{R}_k=\mathbf{1})}{\pi_k(\mathbf{1})}\pi_k(\boldsymbol{r})\right\}\phi_{\boldsymbol{r}}\left(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\right)$$

with $\phi_{\boldsymbol{r}}\left(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\right)$, an arbitrary investigator-chosen function

- Notice that subjects with complete data (i.e. with $\boldsymbol{R}_k=\mathbf{1}$) have:

$$\boldsymbol{A}_k\ =\ -\sum_{\boldsymbol{r}\neq\mathbf{1}}\frac{\pi_k(\boldsymbol{r})}{\pi_k(\mathbf{1})}\phi_{\boldsymbol{r}}\left(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\right)$$

while subjects with incomplete data have $\boldsymbol{A}_k=\phi_{\boldsymbol{r}}\left(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\right)$

  ⋆ i.e. subjects with incomplete data contribute to estimation/inference

- A key result is that $\boldsymbol{A}_k$ has mean zero for all values of $\boldsymbol{\beta}$
- This follows since, for each $\boldsymbol{r}$:

$$\begin{aligned}
\mathrm{E}\left[\left\{I(\boldsymbol{R}_k = \boldsymbol{r}) \; - \; \frac{I(\boldsymbol{R}_k = \mathbf{1})}{\pi_k(\mathbf{1})}\pi_k(\boldsymbol{r})\right\}\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big) \mid \overline{\boldsymbol{W}}_k\right] \\
&= \left\{\pi_k(\boldsymbol{r}) \; - \; \frac{\pi_k(\mathbf{1})}{\pi_k(\mathbf{1})}\pi_k(\boldsymbol{r})\right\}\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big) \\
&= \mathbf{0}
\end{aligned}$$

- Consequently, the augmented IPW GEE is unbiased and the solution is consistent and asymptotically Normal, regardless of the choice of $\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$
  - ⋆ the specific choice of $\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$ only impacts efficiency
- RRS (1998) provide some discussion on good choices for $\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$
  - ⋆ highlight theoretical considerations
  - ⋆ also go through a detailed example

Comments

- In the paper, RRS (1998) actually lay out a broader class of estimating equations
  - ⋆ the equations on slide 459:

$$\sum_{k=1}^{K} \left[\frac{I(\boldsymbol{R}_k = \mathbf{1})}{\pi_k(\mathbf{1})} \boldsymbol{D}_k^T \boldsymbol{V}_k^{-1} (\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \; + \; \boldsymbol{A}_k \right] \; = \; \mathbf{0}$$

can be generalized to:

$$\sum_{k=1}^{K} \left[\frac{I(\boldsymbol{R}_k = \mathbf{1})}{\pi_k(\mathbf{1})} d(\boldsymbol{X}_k; \boldsymbol{\beta})(\boldsymbol{Y}_k - \boldsymbol{\mu}_k) \; + \; \boldsymbol{A}_k \right] \; = \; \mathbf{0} \qquad (1)$$

where $d(\boldsymbol{X}_k; \boldsymbol{\beta})$ is a matrix of functions of the covariates and $\boldsymbol{\beta}$ chosen by the investigator

  - ⋆ class of estimators defined by the functions $d(\boldsymbol{X}_k; \boldsymbol{\beta})$ and $\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$

- It turns out that any regular and asymptotically linear (RAL) estimator of $\boldsymbol{\beta}$ is asymptotically equivalent to the solution of an equation in the class of estimators defined by (1)
  - ⋆ that is, for any RAL estimator $\widetilde{\boldsymbol{\beta}}$, there exist functions $d(\boldsymbol{X}_k; \boldsymbol{\beta})$ and $\phi_{\boldsymbol{r}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$ such that if one was to solve (1) to obtain an estimator $\widehat{\boldsymbol{\beta}}$, then $\sqrt{K}(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})$ converges to 0 in probability and $\sqrt{K}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{K}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ have the same asymptotic distribution

- Furthermore, while the form adopted on slide 459 was chosen to be consistent with the development in the notes, it is possible to achieve efficiency gains by choosing alternative $d(\boldsymbol{X}_k)$
  - ⋆ indeed, there exist functions $d_{\text{opt}}(\boldsymbol{X}_k; \boldsymbol{\beta})$ and $\phi_{\boldsymbol{r},\text{opt}}\big(\overline{\boldsymbol{W}}_{(\boldsymbol{r})k}\big)$ such that the resulting estimator $\widehat{\boldsymbol{\beta}}_{\text{opt}}$ has variance that attains the semiparametric efficiency bound for regular estimators
  - ⋆ see Section 6 of RRS (1998)

Likelihood-based methods

- Suppose that if complete data on all K study participants were available (i.e. all n measurements), the analysis would proceed by:

  (1) specifying a GLMM for the mean, and

  (2) performing estimation/inference via maximum likelihood

- Recall the integrated likelihood for a GLMM:

$$\mathcal{L}(\boldsymbol{Y}_k; \boldsymbol{\beta}^*, \boldsymbol{\alpha}) = \prod_{k=1}^{K} \int \left\{ \prod_{i=1}^{n} f_{Y|\gamma}(Y_{ki} | \boldsymbol{\beta}^*, \boldsymbol{\alpha}, \boldsymbol{\gamma}_k) \right\} f_{\gamma}(\boldsymbol{\gamma}_k | \boldsymbol{\alpha}) \, \partial \boldsymbol{\gamma}_k$$

  ⋆ $\boldsymbol{\beta}^*$ are the regression parameters for the fixed effects

  ⋆ $\boldsymbol{\alpha}$ are the parameters that index the random effects distribution

  ⋆ integrate over the distribution of the unknown $\boldsymbol{\gamma}_k$

- Intuitively, when there is no missing information, the likelihood is the joint distribution of $\boldsymbol{Y}_k$ conditional on $\boldsymbol{X}_k$, viewed as a function of the unknown parameters
  - ⋆ typically adopt some independence assumptions to simplify the form

- Its worth noting that by conditioning on $\boldsymbol{X}_k$, only the $\boldsymbol{Y}_k$ are taken to be 'random'
  - ⋆ at the outset, while $\boldsymbol{Y}_k$ is well-defined, we don't know what the observed response will be

- In the missing data context, the observance indicators are also random
  - ⋆ i.e. the R_{ki}
  - ⋆ at the outset, while $\boldsymbol{R}_k$ is well-defined, we don't know who will have complete data and who will not

- When there is missing data, therefore, the observed data likelihood is the joint distribution of $(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k)$ conditional on $\boldsymbol{X}_k$, viewed as a function of the unknown parameters

- Following the strategy used in GLMMs, the observed data likelihood can be obtained by integrating the full data likelihood over the distribution of the missing data:

$$\begin{aligned}\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k) &= \int \mathcal{L}(\boldsymbol{Y}_k,\ \boldsymbol{R}_k)\ \partial \boldsymbol{Y}_k^m \\ &= \int \mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m,\ \boldsymbol{R}_k)\ \partial \boldsymbol{Y}_k^m\end{aligned}$$

- How we move forward depends on:

(1) the assumptions we are willing to make regarding the missingness mechanism

$*$ i.e. MCAR, MAR or MNAR

(2) the specification for $\mathcal{L}(\boldsymbol{Y}_k,\ \boldsymbol{R}_k)$

$*$ in most settings this can be an unwieldy task, especially as n_k increases

Estimation/inference under MAR

- Under MAR, we have

$$\Pr(\boldsymbol{R}_k \mid \boldsymbol{Y}_k,\ \boldsymbol{X}_k) \;=\; \Pr(\boldsymbol{R}_k \mid \boldsymbol{Y}_k^o,\ \boldsymbol{X}_k)$$

so that the observed data likelihood can be factored as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k) &= \int \mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m,\ \boldsymbol{R}_k)\ \partial \boldsymbol{Y}_k^m \\
&= \int \mathcal{L}(\boldsymbol{R}_k \mid \boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m; \boldsymbol{\theta})\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m; \boldsymbol{\beta})\ \partial \boldsymbol{Y}_k^m \\
&= \int \mathcal{L}(\overset{\text{MAR}}{\boldsymbol{R}_k \mid \boldsymbol{Y}_k^o}; \boldsymbol{\theta})\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m; \boldsymbol{\beta})\ \partial \boldsymbol{Y}_k^m \\
&= \mathcal{L}(\boldsymbol{R}_k \mid \boldsymbol{Y}_k^o; \boldsymbol{\theta})\ \int \mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m; \boldsymbol{\beta})\ \partial \boldsymbol{Y}_k^m \\
&= \mathcal{L}(\boldsymbol{R}_k \mid \boldsymbol{Y}_k^o; \boldsymbol{\theta})\ \mathcal{L}(\boldsymbol{Y}_k^o; \boldsymbol{\beta})
\end{aligned}$$

- Note, $\mathcal{L}(\boldsymbol{Y}_k^o;\boldsymbol{\beta})$ is the marginal likelihood for the observed response data
  - ⋆ i.e. the likelihood that would be used if one simply ignored the fact that some data are missing

- Taking the log we have:

$$\ell(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k)\ =\ \ell(\boldsymbol{R}_k|\ \boldsymbol{Y}_k^o;\boldsymbol{\theta})\ +\ \ell(\boldsymbol{Y}_k^o;\boldsymbol{\beta})$$

which, when differentiated, gives the score:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k)\ =\ \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{Y}_k^o;\boldsymbol{\beta})$$

- Consequently, as long as MAR holds, one can perform valid estimation/ inference based on the marginal likelihood for the observed responses
  - ⋆ i.e. one can forge ahead with an analysis based on the observed data

- This is extremely appealing because it means that we don't even need to consider the form of $\mathcal{L}(\boldsymbol{R}_k |\ \boldsymbol{Y}_k^o; \boldsymbol{\theta})$!
  - ⋆ as long as MAR holds, the missingness mechanism can be anything and estimation/inference will still be valid

- This is in contrast to complete-case GEE for which estimation/inference is only valid, in general, when the data are MCAR
  - ⋆ IPW is needed if the missing data are MAR
  - ⋆ requires (correct) specification of the missingness mechanism

- Additional details/comments:
  - ⋆ requires the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ to be distinct
  - ⋆ otherwise ignoring the first term will lead to a loss of efficiency

Estimation/inference under MNAR

- Under MNAR, the observed data likelihood

$$\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k) = \int \mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m,\ \boldsymbol{R}_k)\, \partial \boldsymbol{Y}_k^m$$

  does not factor conveniently, as it does under MAR, so that an explicit specification for $\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m,\ \boldsymbol{R}_k)$ is required

- Two general frameworks for structuring $\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m,\ \boldsymbol{R}_k)$:

  (1) selection models

  (2) pattern mixtures models

- Note, given either approach, the observed data likelihood is typically computed via numerical approximations to the integral

Selection models

- In the selection models framework, the observed data likelihood is decomposed as:

MNAR: Rk depends on Y^m_k

$$\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k) \ = \ \int \underbrace{\mathcal{L}(\boldsymbol{R}_k|\ \boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m)}_{(1)}\ \underbrace{\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m)}_{(2)}\ \partial \boldsymbol{Y}_k^m$$

(1) the conditional distribution of observance, given the response (and $\boldsymbol{X}_k$)

(2) the marginal distribution of the response (given $\boldsymbol{X}_k$)

- Intuitively, the idea is that subjects are 'selected' to have complete data potentially on the basis of their response

- Attractive because specification of a model in (2) corresponds to the model that one would consider in the (ideal) situation when one has complete data
  - ⋆ e.g. a GLMM

- Also attractive because the decomposition lends itself to the Little and Rubin taxonomy as applied to $\mathcal{L}(\boldsymbol{R}_k |\ \boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m)$:
  - ⋆ i.e. MCAR, MAR, MNAR applied to the conditional distribution of $\boldsymbol{R}_k |\ \boldsymbol{Y}_k, \boldsymbol{X}_k$

- As an example, suppose interest lies in some continuous response based on the following LMM:

$$\begin{aligned} Y_{ki} &= \boldsymbol{X}_{ki}^T\boldsymbol{\beta} + \gamma_{0k} + \epsilon_{ki} \\ \gamma_{0k} &\sim \text{Normal}(0,\ \sigma_\gamma^2) \\ \epsilon_{ki} &\sim \text{Normal}(0, \sigma_\epsilon^2) \end{aligned}$$

- The induced marginal distribution of $\boldsymbol{Y}_k$ is a MVN with mean vector $\boldsymbol{\mu}_k(\boldsymbol{\beta}) = (\boldsymbol{X}_{k1}^T\boldsymbol{\beta}, \ldots, \boldsymbol{X}_{kn}^T\boldsymbol{\beta})$ and an exchangeable variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha} = (\sigma_\gamma^2,\ \sigma_\epsilon^2)$
  - ⋆ basis for $\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m) \equiv \mathcal{L}(\boldsymbol{Y}_k; \boldsymbol{\beta}, \boldsymbol{\alpha})$

- Now suppose that whether a subject drops out at the i^{th} time point is dictated by the following model for $\lambda_{ki} = \Pr(R_{ki} = 1 | \ R_{k,i-1} = 1)$:

$$\text{logit } \lambda_{ki} \ = \ \theta_{0i} \ + \ \theta_1 Y_{k,i-1} \ + \ \theta_2 \underset{\text{Missing}}{Y_{ki}}$$

  - ⋆ that this depends, in part, on the current response indicates that missingness is MNAR
  - ⋆ $\boldsymbol{\theta} = (\theta_{01}, \ldots, \theta_{0n}, \theta_1, \theta_2)$

- Consider the contribution $\mathcal{L}(\boldsymbol{R}_k | \ \boldsymbol{Y}_k^o, \ \boldsymbol{Y}_k^m)$ from a subject who completes the study

  - ⋆ drop out time is $D_k = n + 1$
  - ⋆ probability of not dropping out is given by:

$$\pi_k(\boldsymbol{\theta}) \ = \ \prod_{i=1}^{n} \lambda_{ki}(\boldsymbol{\theta}) \ \equiv \ \prod_{i=1}^{D_k - 1} \lambda_{ki}(\boldsymbol{\theta})$$

- Now consider the contribution $\mathcal{L}(\boldsymbol{R}_k | \; \boldsymbol{Y}_k^o, \; \boldsymbol{Y}_k^m)$ from a subject who does not complete the study
  - drop out time is $D_k < n+1$
  - probability of dropping out at time D_k:

$$\pi_k(\boldsymbol{\theta}) \;=\; \left\{ \prod_{i=1}^{D_k-1} \lambda_{ki}(\boldsymbol{\theta}) \right\} \{1 - \lambda_{k,D_k}(\boldsymbol{\theta})\}$$

- The observed data likelihood can therefore be succinctly written as:

$$\mathcal{L}_k(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) \;=\; \int \left\{ \prod_{i=1}^{D_k-1} \lambda_{ki}(\boldsymbol{\theta}) \right\} \{1 - \lambda_{k,D_k}(\boldsymbol{\theta})\}^{I(D_k<n+1)} \, \mathcal{L}(\boldsymbol{Y}_k; \boldsymbol{\beta}, \boldsymbol{\alpha}) \, \partial \boldsymbol{Y}_k^m$$

  - multivariate integral with respect to a MVN normal distribution
  - dimension of the integral depends on how much data is missing for the k^{th} subject

- Given a specification of the response model and the observance probabilities, $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ can be maximized using any of the techniques we typically use

- Sometimes this will return an estimate of $\boldsymbol{\theta}$ and others not
  - ⋆ depends on the shape of the likelihood surface

- For the selection model we are currently considering:

$$\text{logit } \lambda_{ki} \ = \ \theta_{0i} \ + \ \theta_1 Y_{k,i-1} \ + \ \theta_2 Y_{ki}$$

  suppose an estimate of θ_2 is returned

Q: If θ_2 is, in principle, not identified, what information is being used to estimate it?

- ⋆ in most settings, it will be unclear
- ⋆ structural assumptions imposed by the model will likely play a role
- ⋆ e.g. the assumption of normality for the residuals when the response exhibits skewness

- Some folks use this lack of transparency as an argument to suggest that one should not even attempt to estimate the components of $\boldsymbol{\theta}$ that are not estimable

- Instead, one should either:
  - ⋆ remain within the selection models framework but fix the components that are not estimable at a range of values and examine sensitivity
  - ⋆ consider the pattern mixture models framework (see later)

- Note, there are numerous alternative strategies for modeling λ_{ki} in the selection models framework
  - ⋆ e.g. random coefficient selection models and shared parameter models
  - ⋆ see Lab this week

Pattern mixture models

- In the pattern mixture models (PMM) framework, the observed data likelihood is decomposed as:

$$\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{R}_k) \;=\; \int \underbrace{\mathcal{L}(\boldsymbol{Y}_k^o,\ \boldsymbol{Y}_k^m |\ \boldsymbol{R}_k)}_{(1)}\ \underbrace{\mathcal{L}(\boldsymbol{R}_k)}_{(2)}\ \partial \boldsymbol{Y}_k^m$$

(1) the conditional distribution of the response, given a particular observance pattern (and $\boldsymbol{X}_k$)

(2) the marginal distribution of observance (given $\boldsymbol{X}_k$)

- Intuition is that each subjects drop-out time is predestined, leading to variation across drop-out cohorts
  - ⋆ classification according to drop-out time provides a stratification to examine whether outcomes of interest vary across drop-out strata
  - ⋆ Little (1993, 1994)

- Consider a trial designed to collect two measurements per subject
  - ⋆ if all subjects have, at least, a baseline measurement, then there are two possible observance patterns:

Pattern	Y_{k1}	Y_{k2}	R_{k1}	R_{k2}	D_k
# 1	✓	✓	1	1	3
# 2	✓		1	0	2

- An implication of the PMM decomposition is that the distribution of the response vector varies across observance patterns
  - ⋆ if $\boldsymbol{Y}_k$ is continuous one might, for example, assume:

$$\begin{aligned} &\text{Pattern \#1:} \quad && \boldsymbol{Y}_k^{(1)} \sim \text{MVN}\left(\boldsymbol{\mu}_k^{(1)}, \boldsymbol{\Sigma}_k^{(1)}\right) \\ &\text{Pattern \#2:} \quad && \boldsymbol{Y}_k^{(2)} \sim \text{MVN}\left(\boldsymbol{\mu}_k^{(2)}, \boldsymbol{\Sigma}_k^{(2)}\right) \end{aligned}$$

  - ⋆ or, more specifically, that $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ differ across the patterns

- Since subjects with pattern #1 have complete data, $\boldsymbol{\mu}_k^{(1)}$ and $\boldsymbol{\Sigma}_k^{(1)}$ are estimable from the observed data

- For subjects with pattern #2, however, only $\mu_{k1}^{(2)}$ and $\Sigma_{k,11}^{(2)}$ are estimable
  - ⋆ $\mu_{k2}^{(2)}$, $\Sigma_{k,12}^{(2)}$ and $\Sigma_{k,22}^{(2)}$ are not estimable
  - ⋆ there is no information in the data on the conditional distribution of $Y_{k2} | Y_{k1}$ for patients with pattern #2

- In order to move forward, one has to place additional *identifying assumptions* on the pattern-specific parameters
  - ⋆ e.g. assume that $\boldsymbol{\Sigma}_k^{(1)} = \boldsymbol{\Sigma}_k^{(2)}$ or that certain slope parameters are the same

- Some folks find it appealing that the assumptions being made are being made with respect to the distribution of the response
  - ⋆ clear delineation of what about the response distribution is estimable and what is not

- If one is willing to make necessary identifying assumptions, then the marginal distribution of the response is a mixture of the distributions across the patterns:
  - ⋆ in the current example, the mixture is a mixture of two normals with marginal mean:

$$\mathsf{E}[\boldsymbol{Y}_k] \;=\; \sum_{i=1}^{2} P(D_k = i+1)\boldsymbol{\mu}_k^{(i)}$$

    where $P(D_k = i+1)$ is modeled, possibly, as a function of the components of $\boldsymbol{X}_k$

- This notion can also be extended to give the *marginal covariate effect*:

$$\widehat{\boldsymbol{\beta}} \;=\; \sum_{i=1}^{2} P(D_k = i+1)\widehat{\boldsymbol{\beta}}^{(i)}$$

  - ⋆ standard errors via the delta method
  - ⋆ also fold in uncertainty about the mixture probabilties

- An important drawback of pattern mixture models, however, is that the interpretation of regression coefficients is unclear
  - requires conditioning on the missing data pattern, which is unknown at the outset
  - may, therefore, be less clinically useful

- Another drawback is that as the n increases, the number of patterns can become large
  - computational burden
  - conceptual burden as one attempts to make sense of different parameters across the patterns

- Finally, although presented as alternatives to each other, it is always possible (in theory, at least) to express a PMM as a selection model
  - alternative factorizations of the same joint distribution
  - differences lie in the simplifying assumptions one as to make in order to accommodate missingness that is MNAR