Multiple Longitudinal Markers

- So far we have focused on a single continuous marker
- But often there may be many biomarker we want to study that may be associated with prognosis
- Some of these biomarkers may be categorical
- In our PBC study, in addition to serum bilirubin they also collected:
  - Serum cholesterol (continuous)
  - Edema (3 categories)
  - Ascites (2 categories)
  - And more

Multiple Longitudinal Markers

- We need to extend the basic joint model!
- To handle multiple longitudinal markers of different types we use generalized linear mixed models
- We assume that we have longitudinal outcomes $Y_{i1}, \ldots, Y_{iJ}$ for each subject, where each of them has a distribution in the exponential family, with expected value

$$m_{ij}(t) = E(y_{ij}(t)|b_{ij}) = g_j^{-1}\{x'_{ij}(t)\beta_j + z'_{ij}(t)b_{ij}\}$$

where $g(\cdot)$ is a link function

Multiple Longitudinal Markers

- Correlation between outcomes is built by assuming a multivariate normal distribution for the random effects

$$b_i = (b'_{i1}, ..., b'_{iJ})' \sim N(0, D)$$

- The expected value of each longitudinal marker is incorporated in the linear predictor of the survival submodel

$$h_i(t) = h_0(t) \exp\{\gamma' w_i + \sum_{j=1}^{J} \alpha_j m_{ij}(t)\}$$

Multiple Longitudinal Markers

- **Full conditional independence:** Given the random effects
  - The repeated measurements in each outcome are independent
  - The longitudinal outcomes are independent of each other
  - Longitudinal outcomes are independent of the time-to-event outcomes

$$p(y_{ij}|b_{ij}) = \sum_{k=1}^{n_{ij}} p(y_{ij,k}|b_{ij})$$

$$p(y_i|b_i) = \prod_j p(y_{ij}|b_{ij})$$

$$p(y_i, T_i, \delta_i|b_i) = \prod_j p(y_{ij}|b_{ij}) p(T_i, \delta_i|b_i)$$

Multiple Longitudinal Markers

- With the conditional independence assumption, the extensions of the joint models to multiple longitudinal outcomes is straightforward
- Can use the same estimation procedure
- However, computationally much more intensive due to requirement for high dimensional numerical integrations with respect to the random effects

Multiple Longitudinal Markers

- In the PBC study, we are going to also consider a binary time-dependent variable for blood vessel malformations in the skin ("spiders")
- Going to use "mvglmer" and "mvJointModelBayes" in the "JMBayes" package

```
multMixedFit <- mvglmer(list(log(serBilir) ~ year + (year | id),
                             spiders ~ year + (1 | id)), data = pbc2,
                        families = list(gaussian, binomial))

CoxFit <- coxph(Surv(Time, event) ~ drug + age, data = pbc2.id,
                model = TRUE)

multJMFit <- mvJointModelBayes(multMixedFit, CoxFit, timeVar = "year")
summary(multJMFit)
```

Multiple Longitudinal Markers

```
Survival Outcome:
                     PostMean  StDev  StErr    2.5%  97.5%     P
drugD-penicil         -0.0720 0.1754 0.0052 -0.4149 0.2613 0.698
age                    0.0633 0.0087 0.0003  0.0471 0.0803 0.000
log(serBilir)_value    1.3168 0.1106 0.0028  1.1101 1.5339 0.000
spiders_value          0.0712 0.0501 0.0014 -0.0214 0.1772 0.142

Longitudinal Outcome: log(serBilir) (family = gaussian, link = identity)
            PostMean  StDev  StErr   2.5%  97.5% P
(Intercept)   0.4935 0.0589 0.0019 0.3700 0.6066 0
year          0.1792 0.0130 0.0004 0.1538 0.2049 0
sigma         0.3495 0.0068 0.0002 0.3359 0.3627 0

Longitudinal Outcome: spiders (family = binomial, link = logit)
            PostMean  StDev  StErr    2.5%   97.5% P
(Intercept)  -1.6786 0.2247 0.0178 -2.1371 -1.2688 0
year          0.1839 0.0299 0.0032  0.1246  0.2399 0
```

Other Extensions

- Latent class joint models
- Multiple failure times (competing risks, recurrent events)
- Alternative modeling frameworks (accelerated failure time (AFT) model)

Missing Data

Day 8

- Missing data mechanisms

PBC Study

Research Question:

- Longitudinal Outcome: Investigate the longitudinal evolution of serum bilirubin correcting for dropout

Missing Data in Longitudinal Studies

- Missing data is a major challenge in the analysis of longitudinal data
- Studies are often designed to collect data on every subject at a set of prespecified follow-up times
- Subjects sometimes miss these planned measurements
- We can have different patterns of missing data

Implications of Missing Data

- **Loss of efficiency:** We collect less data than originally planned
- **Unbalanced datasets:** Not all subjects have the same number of measurements
- **Potential bias:** Missingness may depend on outcome

Missing Data in Longitudinal Studies

- Introduce some terminology to describe the missing data mechanisms
- Suppose we have a missing data indictor for each subject *i* at each time point *j* that we expect to collect a measurement at

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- Then we can partition the complete response vector $y_i = (y_i^o, y_i^m)$
  - Observed data y_i^o containing those y_{ij} for which $r_{ij} = 1$
  - Missing data y_i^m containing those y_{ij} for which $r_{ij} = 0$
- **Missing data process:** the vector $r_i = (r_{i1}, \ldots, r_{in_i})$ and the process generating r_i

Missing Data Mechanisms

- Describes the probabilistic relationship between the measurement and missingness processes
- Rubin (1976) introduced 3 mechanisms

1. Missing completely at random (MCAR)
2. Missing at random (MAR)
3. Missing not at random (MNAR)

Missing Data Mechanisms

- **Missing Completely at Random (MCAR):** The probability that responses are missing is unrelated to both y_i^o and y_i^m

$$p(r_i|y_i^o, y_i^m) = p(r_i)$$

- **Examples**
  - Subjects exit the study after providing a pre-determined number of measurements
  - Lab measurements are lost due to equipment malfunction

Missing Data Mechanisms

Features of MCAR:

- Observed data y_i^o is a subset of the complete data y_i
- **Analysis**
  - Any statistical procedure that is valid for complete data
  - Sample averages per time point, linear regression ignoring correlation (consistent), t-test

Missing Data Mechanisms

- **Missing at Random (MAR):** The probability that responses are missing is related to y_i^o, but is unrelated to y_i^m

$$p(r_i|y_i^o, y_i^m) = p(r_i|y_i^o)$$

- **Examples**
  - Study protocol requires patients whose response value exceeds a threshold to be removed from the study
  - Physicians give rescue medication to patients who do not respond to treatment

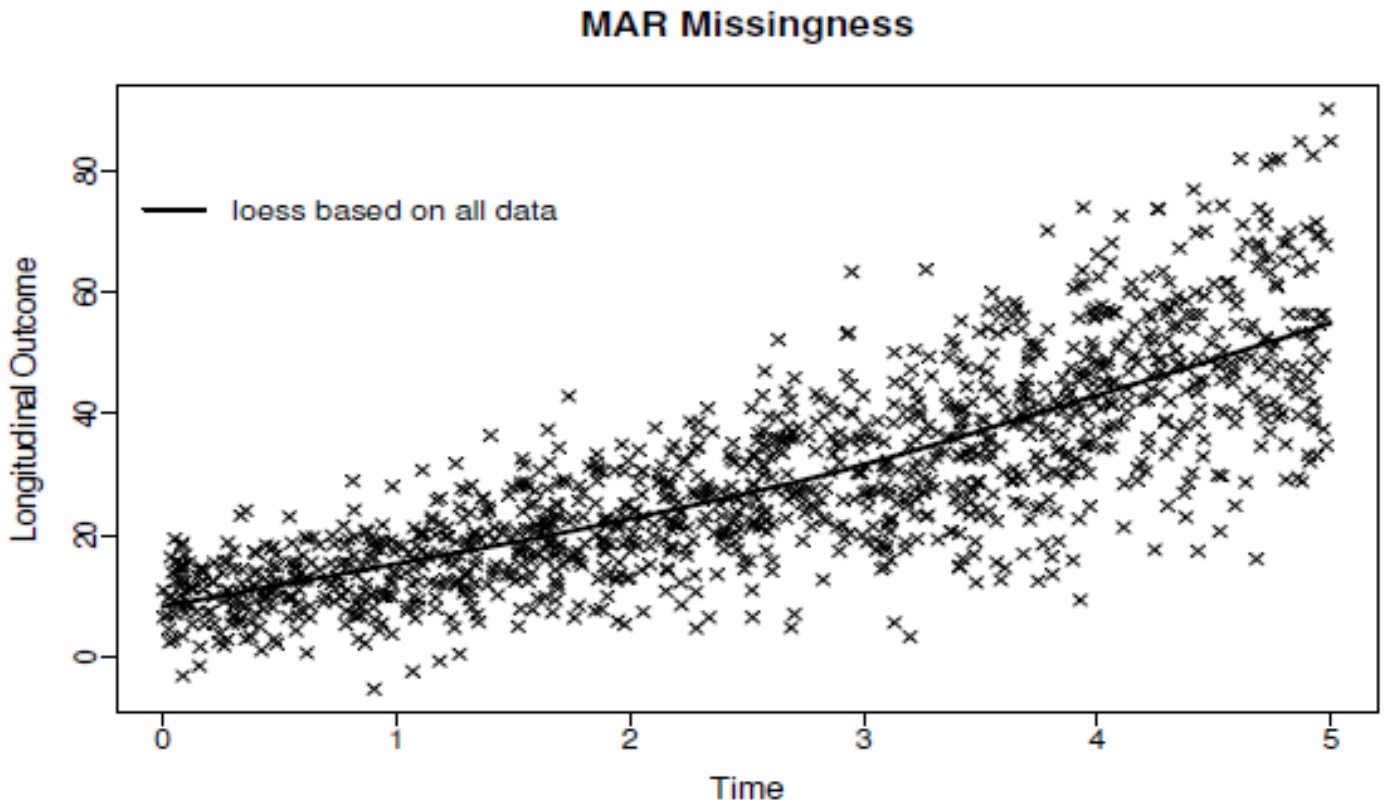
Missing Data Mechanisms

- **Features of MAR**
  - The observed data cannot be considered a random sample from the target population
  - Not all statistical procedures provide valid results

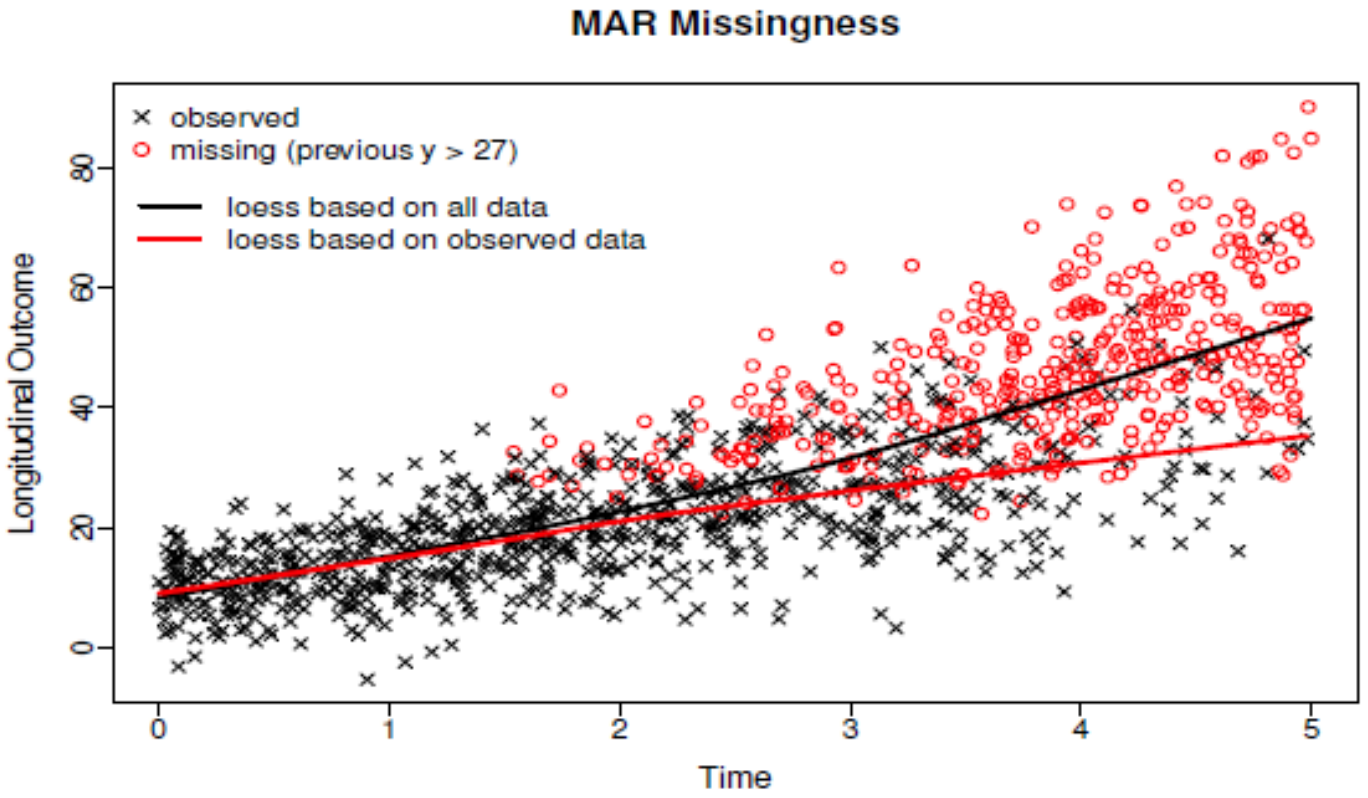
$$p(y_i^m | y_i^o, r_i) = p(y_i^m | y_i^o)$$

Not valid under MAR	Valid under MAR
Sample marginal evolutions	Sample subject-specific evolutions
Methods based on moments, GEE	Likelihood based inference
Mixed models with misspecified correlation structure	Mixed models with correctly specified correlation structure
Marginal residuals	Subject-specific residuals

Missing Data Mechanisms



Missing Data Mechanisms



Missing Data Mechanisms

- **Missing Not at Random (MNAR):** The probability that responses are missing is related to y_i^m, and possibly also to y_i^o

$$p(r_i|y_i^m) \quad \text{or} \quad p(r_i|y_i^o, y_i^m)$$

- Examples
  - In studies on drug users, people who return to drugs are less likely than other to report their status
  - In longitudinal studies for quality-of-life, patients may fail to complete the questionnaire at occasions when their quality-of-life is compromised

Missing Data Mechanisms

Features of MNAR

- The observed data cannot be considered a random sample from the target population
- Only procedures that explicitly model the joint distribution $\{y_i^o, y_i^m, r_i\}$ provide valid inferences
- Analyses that are valid under MAR will not be valid under MNAR
- Selection models, Pattern mixture models (Little, 1995; Molenberghs and Kenward, 2007)
- **Shared-parameter models**

Missing Data Mechanisms

- We can't tell from the data whether the missing data mechanism is MAR or MNAR
- We can distinguish between MCAR and MAR

$$\text{logit}(P(D_i = j | D_i \geq j)) = \alpha_{0j} + \alpha_1 f(y_{ij}^o) + \alpha_2 x_{ij} + \alpha_3 (f(y_{ij}^o) \times x_{ij})$$

- D_i is time of dropout (last measured timepoint)
- $f\left(y_{ij}^o\right)$ is some function of the history of the observed marker measurements
- MCAR is rejected if H_0: $\alpha_1 = \alpha_3 = 0$ is rejected

Connection with Missing Data

- So far we have looked at the problem from the survival point of view
- However, often we may also be interested in the longitudinal outcome
- **Issue:** When patients experience the event, they dropout from the study
- Dropout must be taken into account when deriving inferences for the longitudinal outcome

Connection with Missing Data

- Implications of nonrandom dropout
  - Observed data do not constitute a random sample from the target population

- This feature complicates the validation of the joint model's assumptions using standard residual plots
  - Residual plots may show systemic behavior due to dropout and not because of model misfit

Connection with Missing Data

- What about censoring?
  - Censoring also corresponds to a discontinuation of the data collection process for the longitudinal outcome
- Likelihood-based inferences for joint models provide valid inferences when censoring is MAR
  - A patient relocates to another country (MCAR)
  - A patient is excluded from the study when her longitudinal response exceeds a prespecified threshold (MAR)
  - Censoring depends on random effects (MNAR)

Connection with Missing Data

Frameworks for MNAR data

- **Shared Parameter Models**

$$p(y_i^o, y_i^m, T_i^*) = \int p(y_i^o, y_i^m | b_i) p(T_i^* | b_i) p(b_i) db_i$$

- **Selection Models**

$$p(y_i^o, y_i^m, T_i^*) = p(y_i^o, y_i^m) p(T_i^* | y_i^o, y_i^m)$$

- **Pattern Mixture Models**

$$p(y_i^o, y_i^m, T_i^*) = p(y_i^o, y_i^m | T_i^*) p(T_i^*)$$

Selection Models

$$p(y_i^o, y_i^m, r_i|\theta, \psi) = p(y_i^o, y_i^m|\theta)p(r_i|y_i^o, y_i^m, \psi)$$

- First factor is the marginal density of the measurement process (e.g., linear mixed model)
- Second factor is the density of the drop-out process, conditional on the outcomes (e.g., logistic, probit)
- Example:

$$y_i = x_i'\beta + z_i'b_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), b_i \sim N(0, D)$$

$$\text{logit}[g(h_{ij}, y_{ij})] = \text{logit}[P(D_i = j|D_i \geq j)] = h_{ij}\psi_0 + y_{ij}\psi_r$$

- D_{ij} is the measurement time following the last observed measurement
- h_{ij} is a vector containing all responses observed up to but not including occasion j, and all other relevant covariates

Pattern-Mixture Models

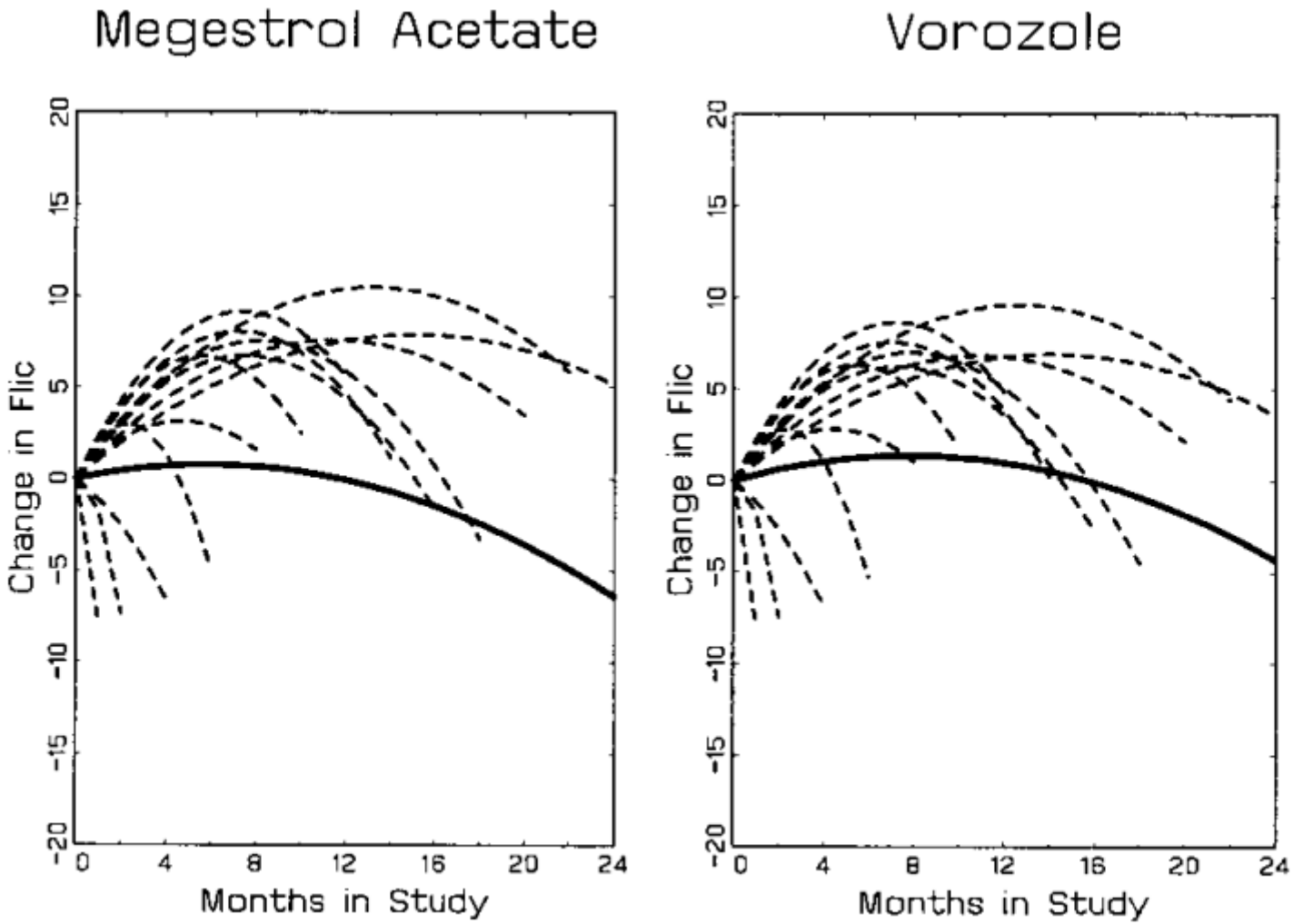
$$p(y_i^o, y_i^m, r_i|\theta, \psi) = p(y_i^o, y_i^m|r_i, \theta)p(r_i|\psi)$$

- First factor is the conditional density of the measurements given the drop-out pattern
- Second factor is the marginal density of the drop-out mechanism
- The measurement model has to reflect dependence on drop-out and the parameters are allowed to change with drop-out pattern

$$y_i = x_i'\beta(d_i) + z_i'b_i + \epsilon_i$$
$$b_i \sim N(0, D(d_i))$$
$$\epsilon_i \sim N(0, \Sigma_i(d_i))$$

- Where d_i represents a missingness pattern

Selection and Pattern-Mixture Models



Michiels et al. 2002

Shared Parameter Models

- A nice feature of shared random effects models is that they can "automatically" handle intermittent missing data
- The observed data likelihood contributions take the form

$$
\begin{aligned}
p(y_i^o, T_i^*) &= \int p(y_i^o, y_i^m, T_i^*) dy_i^m \\
&= \int \int p(y_i^o, y_i^m | b_i) p(T_i^* | b_i) p(b_i) db_i dy_i^m \\
&= \int \left\{ \int p(y_i^o, y_i^m | b_i) dy_i^m \right\} p(T_i^* | b_i) p(b_i) db_i \\
&= \int p(y_i^o | b_i) p(T_i^* | b_i) p(b_i) db_i
\end{aligned}
$$

- This is not the case for selection and pattern mixture models!

Connection with Missing Data

- In the PBC data the association parameter α is highly significant, suggesting nonrandom dropout
- A comparison between:
  - Linear mixed-effect model -> MAR
  - Joint model -> MNAR
- MAR assumes that missingness mechanism depends only on the observed data

$$p(T_i^*|y_i^o, y_i^m) = p(T_i^*|y_i^o)$$

- MNAR depends on unobserved data

$$p(T_i^*|y_i^o, y_i^m) = \int p(T_i^*|b_i)p(b_i|y_i^o, y_i^m)db_i$$

Sensitivity Analysis

- Using observed data alone cannot distinguish between a MAR and MNAR dropout mechanism
- Every MNAR model has a MAR counterpart that provides exactly the same fit to the data (i.e., same likelihood value), but inferences may be different *(Molenberghs et al. 2008)*
- So identification of the non-ignorability parameters in a MNAR model is provided through modelling assumptions
- Thus, need to assess violation of assumptions using a sensitivity analysis
- Compare MNAR model to corresponding MAR model
- Compare to other types of MNAR models

Joint Modeling Framework

- To account for possible MNAR dropout, we need to postulate a model that relates the longitudinal marker with time to dropout
- Intuitive idea behind the joint model:
  - Use an appropriate model to describe the evolution of the marker in time
  - Use the estimated evolutions in the Cox model

Example: MNAR Analysis of PBC Study

- We are going to fit the following joint model to the PBC Study data

$$
\begin{aligned}
h_i(t) &= h_0(t)\exp\{\gamma_1 \text{D-penecil}_i + \alpha m_i(t)\} \\
y_i(t) &= m_i(t) + \epsilon_i(t), \qquad \epsilon_i(t) \sim N(0, \sigma^2) \\
&= \beta_0 + \beta_1 t + \beta_2\{t \times \text{D-penecil}_i\} + b_{i0} + b_{i1}t + \epsilon_i(t) \\
b_i &\sim N(0, D)
\end{aligned}
$$

- Where we assume that $h_0(t)$ is piecewise-constant

Example:

	LMM (MAR) Est (SE)	JM (MNAR) Est (SE)
Intercept	0.496 (0.058)	0.492 (0.058)
Time	0.176 (0.018)	0.183 (0.018)
Time:Drug	0.003 (0.024)	0.003 (0.025)

- Minimal sensitivity in parameter estimate and standard errors
- This does not mean that this is always the case!

Breakout Room #8

1. How would you simulate longitudinal marker trajectories that have the following missing data mechanisms?
   a) MCAR
   b) MAR
   c) MNAR

   Hint: Start by thinking about the situation where you might have these missing mechanisms and how you would replicate them.