Topic 7. Missing Values in Longitudinal Data

Missing values - One or more of the intended measurements from units within the study are not taken, are lost, or are unavailable.

Classification of missing value mechanisms:

Let $Y^* = (Y^{(o)}, Y^{(m)})$ with $Y^{(o)}$ denoting the measurements actually obtained, and $Y^{(m)}$ being the measurements which would have been available if they were not missing. Let R denote a set of indicator random variables, denoting which elements of the complete set of measurements fall into $Y^{(0)}$ and which into $Y^{(m)}$.

Missing value mechanism:

(a) Completely random: $R \perp (Y^{(0)}, Y^{(m)})$.

(b) Random: $R \perp Y^{(m)}$.

(c) Informative: $R \not\perp Y^{(m)}$.

Distinction between random and informative missing values:

Let $f(y^{(o)}, y^{(m)}, r)$ denote the probability density function of $(Y^{(o)}, Y^{(m)}, R)$. One has

$$f(y^{(o)}, y^{(m)}, r) = f_{Y^{(o)},Y^{(m)}}(y^{(o)}, y^{(m)}) f_{R|Y^{(o)},Y^{(m)}}(r \mid y^{(0)}, y^{(m)}),$$ which implies that

$$f_{(Y^{(o)},R)}(y^{(o)}, r) = \int f_{Y^{(o)},Y^{(m)}}(y^{(o)}, y^{(m)}) f_{R|Y^{(o)},Y^{(m)}}(r \mid y^{(o)}, y^{(m)}) dy^{(m)}.$$

1. Completely random: $f_{Y^{(0)},R}(y^{(o)}, r) = f_{Y^{(o)}}(y^{(o)}) f_R(r)$.

2. Random:
$$\begin{aligned} f_{Y^{(o)},R}(y^{(o)}, r) &= (\int f(y^{(o)}, y^{(m)}) dy^{(m)}) f_{R|Y^{(o)}}(r \mid y^{(o)}) \\ &= f_{Y^{(o)}}(y^{(o)}) f_{R|Y^{(o)}}(r \mid y^{(o)}). \end{aligned}$$

Remark.

(a) If $f_{R|Y^{(o)}}(r \mid y^{(o)})$ contains no information about the distribution of $Y^{(0)}$, it can be ignored for the purpose of making inferences about $Y^{(o)}$. Thus, both completely random and random missing mechanisms are sometimes referred to without distinction as ignorable.

(b) If there are parameters common to $f_{Y^{(o)}}(y^{(o)})$ and $f_{R|Y^{(o)}}(r \mid y^{(o)})$, ignoring $f(r \mid y^{(o)})$ will lead to a loss of efficiency.

(c) In some cases, it might be more sensible to make the distribution of time to survival and the conditional distribution of $Y^{(o)}$ given survival.

Intermittent missing values and dropouts:

Consider a sequence of measurements $Y_1, \cdots, Y_n$, missing values occur as dropouts if whenever Y_j is missing, so are Y_k for all $k \geq j$. Otherwise, the missing values are intermittent.

1. When intermittent missing values arise through a known censoring mechanism, an EM-algorithm provides a possible framework.
2. When intermittent missing values do not arise from censoring and missingness is unrelated to the measurement process, the resulting data can be analyzed by any method which can accommodate unbalanced data.
3. Dropouts are frequently lost to any form of follow-up and arise directly or indirectly connected to the measurement process.
4. When there is any relationship between the ~~measurement~~ measurements, the interpretation of apparently simple trends in the mean response over time can be problematic.

Example: Random dropouts (constant mean response): The empirical means might suggest a clearly increasing time-trend and the likelihood-based means is essentially constant.

Explaination 1. The likelihood-based analysis estimates the mean response which would have been observed if there were no dropouts, whereas the empirical means estimate the conditional mean response in the sub-population who have not dropped out by time t.

Explaination 2. The likelihood-based method recognizes the correlation in data and, in effect, imputes the missing values for a particular subject taking into account the same subject's observed values.

Simple solutions and their limitations:

1. Last observation carried forward - Extrapolating the lost observed ~~measurement~~ measurements for the subject in question to the remainder of their intended time-sequence.

Method:

a. Estimate the time trend, say $\mu_i(t)$, of the ith response at time t.

b. Let y_{ij} be the last observation of the ith subject and $r_{ij} = y_{ij} - \hat{\mu}(t_j)$.

c. Impute the missing values as $\hat{y}_{ik} = \hat{\mu}(t_k) + r_{ij}$ for $k > j$.

2. Complete case analysis – Discarding all incomplete sequences.

Limitations:

a. The complete cases might not be assumed to be a random sample.

b. The incomplete data provide additional information about the underlying measurement process conditional on completion if we are prepared to model the relationship between the measurement process and the dropout process.

Testing for completely random dropouts:

Property: Let P_{ij} denote the probability that the ith unit drops out at time t_j, $j = 1, \cdots, m$.

Under the assumption of completely random dropouts, the P_{ij} may depend on time, treatment, or other explanatory variables but cannot depend on the observed measurements $y_i = (y_{i1}, \cdots, y_{im_i})$.

Testing method:

a. Choose the score functions $h_k(y_1, \cdots, y_k)$ so that extreme values constitute evidence against completely random dropouts. A sensible choice is $h_k(y_1, \cdots, y_k) = \sum_{j=1}^{k} \omega_j y_j$.

b. For each of $k = 1, \cdots, (m-1)$, define $R_k = \{i : m_i \geq k\}$ and $r_k = \{i : m_i = k\}$, and compute the set of scores $h_{ik} = h_k(y_{i1}, \cdots, y_{ik})$ for $i \in R_k$.

c. If $1 \leq \#(r_k) \leq \#(R_k)$, test the hypothesis that the r_k's scores so defined are a random sample from the "population" of R_k's scores.

Remark.

1. The implicit assumption that the separate p-values are mutually independent is valid precisely because once a unit drops out it never returns.

2. A natural test statistic is $\bar{h}_k = \frac{1}{\#(r_k)} \sum_{\{j \in r_k\}} h_{jk}$. Under the assumption that dropouts occur completely at random,

$$\bar{h}_k \approx N(\bar{H}_k, \frac{(\#(R_k) - \#(r_k))}{\#(r_k)(\#(R_k) - 1)} \sum_{\{j \in R_k\}} (h_{jk} - \bar{H}_k)^2), \bar{H}_k = \frac{1}{\#(R_k)} \sum_{\{j \in R_k\}} h_{jk}.$$

(*) When $\#(R_k)$ or $\#(r_k)$ is small, evaluate the complete randomization distribution of $\bar{h}_k$ under the null hypothesis.

(*) Alternative method is to re-compute $(s-1)$ new $\bar{h}_k$'s based on a sample of size $\#(r_k)$ without replacement from $\{h_{ik} : i \in R_k\}$. Let x denote the rank of the original $\bar{h}_k$ among the recomputed values. The $p = \frac{x}{s}$ is the p-value of the exact Monte Carlo test.

3. Final stage consists of analyzing the resulting set of p-values via

$$\begin{cases} \text{Empirical distribution of the } p\text{-values} \\ \text{Kolmogorov-Smirnov statistic } D_+ = sup\,|\hat{F}_n(p) - p| \end{cases}$$

Generalized estimating equations under a random missing mechanism:

$$P(R_{ij} = 1 \mid R_{ij-1} = 1, H_{im}) = P(R_{ij} = 1 \mid R_{ij-1} = 1, H_{ij})$$

Basic GEE method when dropouts are completely random:

$$S_\beta(\beta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T Var(Y_i)^{-1} (Y_i - \mu_i) = 0$$

Let $P_i = diag(P_{ij})$, $diag(P_{ij}) = \prod_{k=1}^{j} \lambda_{ik}$ with $\lambda_{ij} = P(R_{ij} = 1 \mid R_{ij-1} = 1, H_{ij})$, $i = 1, \cdots, n$.

When P_i's are themselves estimated from the data using an assumed random dropout model, the estimators of β obtained from the following extended estimating equation are consistent.

$$S_\beta^*(\beta, \alpha) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T Var(Y_i)^{-1} P_i^{-1} (Y_i - \mu_i) = 0$$

inverse weight

Reference:

Robins, J.M., Rotnizky, A., Zhao, L.P. (1995). Analysis of semiparametric Regression Models for Repeated Outcomes in the presense of Missing Data. JASA, 106-121.

Informative Dropout -

Modeling the dropout process:

Let $Y^* = (Y_1^*, \cdots, Y_m^*)$ and $Y = (Y_1, \cdots, Y_m)$ denote ~~separately~~ the complete set of intended measurements and the observed measurements, respectively.

Drop-out time D obeys the relationship $Y_j = \begin{cases} Y_j^* & : j < D \\ 0 & : j \geq D \end{cases}$.

Note that $2 \leq D \leq m+1$ with $D = m+1$ indicating that the subject has not dropped out.

Model Selection:

$P(Y^*, D) = P(Y^*)P(D|Y^*)$: Dropouts are selected according to their measurement history.

Remark:

(a) Dropouts are completely random if $P(D \mid Y^*) = P(D)$.

(b) Dropouts are random if $P(D \mid Y^*) = P(D \mid Y_1^*, \cdots, Y_{D-1}^*)$.

Let $H_k = (Y_1, \cdots, Y_{k-1})$.

$f_k^*(y \mid H_k^*; \theta, \phi)$: the conditional p.d.f. of y_k^* given H_k^*.

$f_k(y \mid H_k; \theta, \phi)$: the conditional p.d.f. of y_k given H_k.

$P_k(H_k, y_k^*; \beta) = P(D = k \mid H_k, y_k^*; \beta)$.

Properties:

(p1) $P(Y_k = 0 \mid H_k, Y_{k-1} = 0) = 1$.

(p2) $P(Y_k = 0 \mid H_k, Y_{k-1} \neq 0) = \int P(D = k \mid H_k, y_k^*; \beta) f_k^*(y_k^* \mid H_k^*; \theta, \phi) dy_k^*$.

$$= \int P_k(H_k, y_k^*; \beta) f_k^*(y_k^* \mid H_k^*; \theta, \phi) dy_k^*.$$

(p3) $f_k(y \mid H_k; \theta, \phi) = (1 - P_k(H_k, y; \beta)) f_k^*(y \mid H_k^*; \theta, \phi)$ for $y \neq 0$.

$f_k(y \mid H_k; \theta, \phi) = P_k(H_k, y; \beta) I(y=0) + (1 - P_k(H_k, y; \beta)) f_k^*(y \mid H_k^*; \theta, \phi) I(y \neq 0)$.

$P_k(H_k, y; \beta) = f_k(0 \mid H_k, y_{k-1} \neq 0)$

Properties (p1)-(p3) determine the joint p.d.f. of a complete sequence Y as

$$f_Y(y) = f_1^*(y_1)\prod_{k=2}^{m} f_k(y_k|H_k) = f_{Y^*}(y)\prod_{k=2}^{m}(1-P_k(H_k,y_k;\beta)).$$

For an incomplete sequence $Y=(Y_1,\cdots,Y_{d-1},0,\cdots,0)$,

$$f_Y(y) = \{f_1^*(y_1)\prod_{k=1}^{d-1} f_k(y_k|H_k)\}P(Y_d=0|H_d,Y_{d-1}\neq 0)$$

$$= f^*_{Y_1^*,\cdots,Y_{d-1}^*}(y_1^*,\cdots,y_{d-1}^*)\prod_{k=2}^{d-1}(1-P_k(H_k,y_k^*;\beta))\left(\int P_d(H_d,y_d^*;\beta)f_d^*(y_d^*\,|\,H_d;\theta,\phi)dy_d^*\right).$$

Remark. Under either CRD or RD, $P_k(H_k,y_k^*;\beta)$ does not depend on y_k^*, it implies that $P(Y_k=0\,|\,H_k,Y_{k-1}\neq 0)=p(H_k;\beta)$ and the likelihood separates into two components, one for (θ,ϕ) and one for β. Let $y_i=(y_{i1},\cdots,y_{id_i-1})^T$, $i=1,\cdots,n$.

The log-likelihood function for (θ,ϕ,β) can be partitioned as

$$L(\theta,\phi,\beta)=L_1(\theta,\phi)+L_2(\beta)+L_3(\theta,\phi,\beta) \quad \text{with}$$

$L_1(\theta,\phi)=\sum_{i=1}^{n}\ln f^*_{Y_{i1},\cdots,Y_{id_i-1}}(y_1,\cdots,y_{d_i-1})$, $L_2(\beta)=\sum_{i=1}^{n}\sum_{k=2}^{d_i-1}\ln(1-P_k(H_{ik},y_{ik};\beta))$, and

$L_3(\theta,\phi,\beta)=\sum_{\{i:d_i\le m\}}\ln P(D_i=d_i\,|\,H_{d_i},y_{id_i-1}\neq 0)$.

Remark.

1. Under RD, $L_3(\theta,\phi,\beta)$ depends only on β and can therefore be absorbed into $L_2(\beta)$. Hence, the dropouts only affect $L_2(\beta)$ and are ignorable for likelihood-based inference about (θ,ϕ).

Similarly, $L_2(\beta)$ is the log-likelihood associated with the sub-model for D conditional on Y^*. It follows that the stochastic structure of the measurement process can be ignored for likelihood-based inference about β.

2. The general strategy for the informative dropout process is to model the relationship between an observed dropout event and an unobservable concomitant. Typically, it leads to poor identifiability of the model parameters and making it difficult or impossible to validate the assumed model from the observed data.

Pattern mixture models –

Pattern mixture models, introduced by Little (1993), work with the factorization of the joint distribution of Y^* and D into the marginal distribution of D and the conditional distribution of Y^* given D, thus $P(Y^*,D)=P(D)P(Y^*\,|\,D)$.

Rationale for pattern mixture model: Each subject's dropout time is somehow predestined, and that the measurement process varies between dropout cohorts.

The arguments in favor of pattern mixture modeling are usually of a more pragmatic kind.

1.Classification of subjects according to their dropout time provides separate inspection of sub-groups.

2. Pattern mixture factorization brings out very clearly those aspects of the models, which are assumption driven rather than data driven.

$$f(y^* \mid D=d) = f(y_1^* \mid d) f(y_2^* \mid y_1^*, d) \cdots f(y_d^* \mid y_1^*, \cdots, y_{d-1}^*, d) \cdots f(y_m^* \mid y_1^*, \cdots, y_{m-1}^*, d)$$

Assumption: ($d < m+1$ and $t \geq d$)

Complete case missing ~~variable~~ value restrictions: (CCMV)

$$f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, d) = f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, m).$$

Available case missing value restrictions: (ACMV)

$$f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, d) = f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, D > t).$$

Property 1: $f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, d) = f(y_t^* \mid y_1^*, \cdots, y_{t-1}^*)$.

Property 2: Missing at random (MAR): $f(D=d \mid y_1, \cdots, y_m) = f(D=d \mid y_1, \cdots, y_{d-1})$ $\Leftrightarrow$ ACMV.

Reference:

Verbeke, G. and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. Springer.

Random effects models -

Let $U = (U_1, U_2)$ be a bivariate random effect.

$$f_{Y^*,D,U}(y^*, d, u) = f_{Y^* \mid U_1}(y^* \mid u_1) f_{D \mid U_2}(d \mid u_2) f_U(u), \text{ i.e. } Y^* \perp D \mid U = u.$$

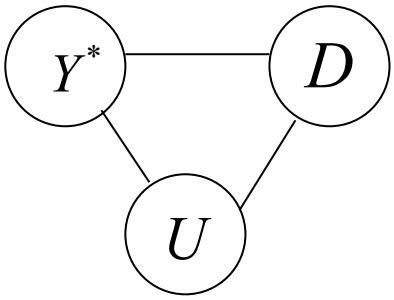
Remark.

(1) $U_1 \perp U_2 \Rightarrow$ the dropouts are completely random.

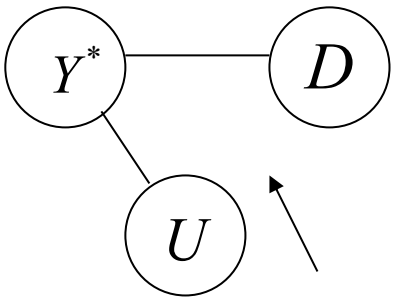
(2) $U_1 \not\perp U_2 \Rightarrow$ the dropouts are informative.

Contrasting assumptions:

Saturated model:



Selection model:

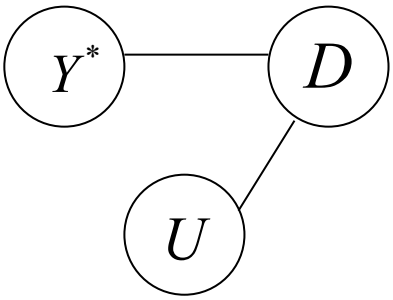


The absence of edge indicates the conditionally independent given the third.

$$f_{Y^*DU}(y^*,d,u) = f_{Y^*}(y^*) f_{D|Y^*}(d \mid y^*) f_{U|Y^*}(u \mid y^*).$$

$$= f_{Y^*|U}(y^* \mid u) f_{D|Y^*}(d \mid y^*) f_U(u)$$

Pattern mixture model:



$$f_{Y^*DU}(y^*,d,u) = f_D(d) f_{Y^*|D}(y^* \mid d) f_{U|D}(u \mid d) = f_{D|U}(d \mid u) f_{Y^*|D}(y^* \mid d) f_U(u).$$

Random effects dropout model: