Problem 1. Testing for completely random dropouts

Let P_{ij} denote the probability that the $i-$th unit drops out at time t_j, $j = 1, \ldots, m$.
Under the assumption of completely random dropouts, the probability P_{ij} may depend on time, treatment, or other explanatory variables, but cannot depend on the observed measurements $y_i = (y_{i1}, \ldots, y_{i\,m_i})$.

Testing Method:

(a) Choose the score function $h_k(y_1, \ldots, y_k)$ so that extreme values constitute evidence against completely random dropouts. A sensible choice is

$$h_k(y_1, \ldots, y_k) = \sum_{j=1}^{k} \omega_j y_j.$$

(b) For each of $k = 1, \ldots, (m-1)$, define

$$R_k = \{i : m_i \geq k\},$$
$$r_k = \{i : m_i = k\},$$

and compute the set of scores $h_{ik} = h_k(y_{i1\ldots,y_{ik}})$ 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 "populations" of R_k's scores.

—

Remark:

1. The implicit assumption that the separated $p-$values are mutually independent is valid precisely because once a unit drops out, it never returns.
2. A natural test statistics is $\overline{h}_k = \frac{1}{|r_k|} \sum_{\{j \in r_k\}} h_{jk}$. Under the assumption of completely random dropouts,

$$\overline{h}_k \sim N\left(\overline{H}_k, \frac{|R_k| - |r_k|}{(|R_k| - 1)|r_k|} \sum_{\{j \in R_k\}} (h_{jk} - \overline{H}_k)^2 / |R_k|\right),$$

where

$$\overline{H}_k = \frac{1}{|R_k|} \sum_{\{j \in r_k\}} h_{jk}.$$

- When $|R_k|$ or $|r_k|$ is small, evaluate the randomization distribution of $\overline{h}_k$ under the null hypothesis.
- Alternative method ...

3. The Final stage consists of analyzing the resulting set of $p-$values via
   (a) Empirical distribution of the $p-$values
   (b) Kolmogorov-Smirnov statistic $D_+ = \sup |\hat{F}_n(p) - p|$

Given a finite population of size N, with individual values $\{X_i\}_{i=1}^N$,

and a set of sample of size n, drawn from the population without replacement, with values $\{X_i\}_{i=1}^n$.

Let σ^2 be the population variance:

$$\sigma^2 = \mathbf{Var}[X_i] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu),$$

where $\mu = \sum_{i=1}^{N} X_i$ is the population mean.

Let $\bar{X} = \frac{1}{n}S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean based on the sample set.

Since every pair (X_i, X_j) for $i \neq j$ has the same joint distribution, we have

$$\mathbf{Var}[S_n] = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{Cov}[X_i, X_j],$$

where

$$\mathbf{Cov}[X_i, X_j] = \begin{cases} \sigma^2 & i = j \\ c & i \neq j \end{cases}.$$

Thus,

$$\mathbf{Var}[S_n] = n\sigma^2 + n(n-1)c.$$

which applies to the case $n = N$ as well. Notice that S_N is a constant (equal to the sum of all N values in the population). It follows that

$$0 = \mathbf{Var}[S_N] = N\sigma^2 + N(N-1)c.$$

Solve the equation above for

$$c = -\frac{\sigma^2}{N-1}.$$

Hence,

$$\mathbf{Var}[S_n] = n\sigma^2\left(1 - \frac{n-1}{N-1}\right) = \frac{N-n}{N-1}\cdot n\sigma^2$$

and

$$\mathbf{Var}[\bar{X}] = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}.$$

The factor $\dfrac{N-n}{N-1}$ is the Finite Population Correction Factor (FPC).

Problem 2. Generalized estimating equations under a random missing mechanism:

Suppose that at each occasion (visit) t, $t = 1, \ldots, T$, the marginal distribution of Y_{it} given X_i follows the regression model

$$\mathbb{E}(Y_{it}|X_i) = g_t(X_i, \beta)$$

for $i = 1, \ldots, n$, where $\mathbf{P}_0$ is a $p \times 1$ vector of unknown parameters and $g_t(\cdot, \cdot)$ are fixed functions[a].

—

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

—

Basic GEE method when dropouts are completely random:

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

Let $P = \mathrm{diag}(P)$,

$$P_{ij} = \prod_{k=1}^{j} \lambda_{ik},$$

with $\lambda_{ij} = P(R_{ij} = 1|R_{ij-1} = 1, H_{ij})$, $i = 1, \ldots, n$.
When P_i's are themselves estimated from the data using an assumed random dropout model, the estimators of $\mathbf{b}$ obtained from the following extended estimating equation are consistent.

$$S^*_\beta(\beta, \alpha) = \sum_i^n \left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1}(Y_i - \mu_i) = 0$$

—

Show that

$$\mathbb{E}[S^*_\beta(\beta, \alpha)] = 0.$$

[a] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data". In: *Journal of the American Statistical Association* 90.429 (Mar. 1995), pp. 106–121, p. 107.

$$\begin{aligned}
\mathbb{E}[S^*_\beta(\beta, \alpha)] &= \mathbb{E}\left[\sum_i^n \left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1}(Y_i - \mu_i)\right] \\
&= \sum_i^n \mathbb{E}\left[\left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1}(Y_i - \mu_i)\right].
\end{aligned}$$

Focus on one term of the summation

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1}(Y_i - \mu_i)\right] \\
&= \mathbb{E}\left[\left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1}\, \mathbb{E}(Y_i - \mu_i|X_i)\right] \quad \text{(Taking the expectation inside)} \\
&= \mathbb{E}\left[\left(\frac{\partial \mu}{\partial \beta}\right)^T \mathrm{Var}(Y_i)^{-1} P_i^{-1} \cdot 0\right] \quad \text{(Since } \mu_i = \mathbb{E}(Y_i|X_i)\text{)} \\
&= 0.
\end{aligned}$$

Summing over all i, we get:

$$\mathbb{E}[S^*_\beta(\beta, \alpha)] = \sum_i^n 0 = 0.$$

Hence, we have shown that $\mathbb{E}[S^*_\beta(\beta, \alpha)] = 0$.

.....

Paper[1] Summary

1. **Research Problem** The paper addresses the challenge of estimating the parameters of semiparametric regression models for repeated outcomes in the presence of missing data due to dropouts [2] . The paper aims to provide consistent and efficient estimators that do not require full specification of the likelihood or the joint distribution of the data.

2. **Methodology** The paper proposes a class of inverse probability of censoring weighted estimators that can handle missing data that are missing at random but not missing completely at random.

$$P(R_{it} = 1|R_{i,t-1} = 1, \bar{W}_{i,T+1}) = P(R_{it} = 1|R_{i,t-1} = 1, \bar{W}_{it}),$$

   where:

   - R_{it} is the response indicator, which equals 1 if the outcome is observed at time t for subject i, and 0 otherwise.
   - $\bar{W}_{it}$ is the vector of observed outcomes and covariates up to time t for subject i.

   The paper derives the asymptotic properties of the proposed estimators and compares them with other methods such as the G-computation algorithm, the sweep estimator[3].

3. **Data and Application** The paper illustrates the proposed methods with the analysis of the effect of zidovudine (AZT) treatment on the evolution of mean CD4 count with data from an AIDS clinical trial. The paper also conducts a simulation study to evaluate the performance of the proposed estimators under different scenarios of missing data mechanisms and model misspecification.

4. **Results and Conclusion** The paper shows that the proposed estimators can correct for dependent censoring and nonrandom noncompliance in randomized clinical trials and can be more efficient and robust than the existing methods. It also discusses the limitations of the proposed methods and suggests conducting a sensitivity analysis to assess the impact of possible violations of the assumptions.

[1] Robins, Rotnitzky, and Zhao, see n. *a*.

[2] In this study, nonresponse, dropout, and censoring are used interchangeably to refer to the situation when some subjects miss one or more visits and their outcome data are not observed. A monotone missing data pattern in this paper means that once a subject leaves the study, returning is not possible. The paper assumes this pattern until Section 6, where it generalizes the results to allow for arbitrary patterns of missing data.

[3] which I do not understand yet.

Problem 3.

Theorem 20.1

For longitudinal data with dropouts, MAR $\iff$ ACMV.[a]

[a] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Corr. 2. print. Springer Series in Statistics. New York Berlin Heidelberg: Springer, 2001. 568 pp., p. 334.

The MAR assumption states that

$$f(d = t+1 \mid y_1, \ldots, y_n) = f(d = t+1 \mid y_1, \ldots, y_t) \tag{B.9}$$

and the ACMV assumption that for all $t \geq 2$, $\forall j < t$,

$$f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) = f(y_t \mid y_1, \ldots, y_{t-1}, d > t). \tag{B.10}$$

First, a lemma will be established.

Lemma B.1 In a longitudinal setting with dropout, ACMV $\Leftrightarrow \forall t \geq 2, \forall j < t : f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) = f(y_t \mid y_1, \ldots, y_{t-1})$.

Proof. Take $t \geq 2, j < t$, then ACMV leads to

$$\begin{aligned}
&f(y_t \mid y_1, \ldots, y_{t-1}) \\
&= \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}, d = i+1) f(d = i+1) \\
&\quad + f(y_t \mid y_1, \ldots, y_{t-1}, d > t) f(d > t) \\
&= \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) f(d = i+1) \\
&\quad + f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) f(d > t) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) \left(\sum_{i=1}^{t-1} f(d = i+1) + f(d > t) \right) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1).
\end{aligned}$$

To show the reverse direction, take again $t \geq 2, j < t$:

$$\begin{aligned}
&f(y_t \mid y_1, \ldots, y_{t-1}, d > t) f(d > t) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}) - \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}, d = i+1) f(d = i+1) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}) - \sum_{i=1}^{t-1} f(y_t \mid y_1, \ldots, y_{t-1}) f(d = i+1) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}) \left(1 - \sum_{i=1}^{t-1} f(d = i+1) \right) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) \left(1 - \sum_{i=1}^{t-1} f(d = i+1) \right) \\
&= f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) f(d > t).
\end{aligned}$$

This completes the proof. We are now able to prove Theorem 20.1.

MAR $\Rightarrow$ ACMV

Consider the ratio Q of the complete data likelihood to the observed data likelihood. This gives, under the MAR assumption,

$$Q = \frac{f(y_1, \ldots, y_n) f(d = i+1 \mid y_1, \ldots, y_i)}{f(y_1, \ldots, y_i) f(d = i+1 \mid y_1, \ldots, y_i)} = f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i). \tag{B.11}$$

Further, one can always write,

$$\begin{aligned} Q &= f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i, d = i+1) \times \frac{f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1)}{f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1)} \\ &= f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i, d = i+1). \end{aligned} \tag{B.12}$$

Equating expressions (B.11) and (B.12) for Q, we see that

$$f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i, d = i+1) = f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i). \tag{B.13}$$

To show that (B.13) implies the ACMV conditions (B.10), we will use the induction principle on t. First, consider the case $t = 2$. Using (B.13) for $i = 1$, and integrating over $y_3, \ldots, y_n$, we obtain

$$f(y_2 \mid y_1, d = 2) = f(y_2 \mid y_1),$$

leading to, using Lemma B.1,

$$f(y_2 \mid y_1, d = 2) = f(y_2 \mid y_1, d > 2).$$

Suppose, by induction, ACMV holds for all $t \leq i$. We will now prove the hypothesis for $t = i+1$. Choose $j \leq i$. Then, from the induction hypothesis and Lemma B.1, it follows that for all $j < t \leq i$:

$$f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) = f(y_t \mid y_1, \ldots, y_{t-1}, d > t) = f(y_t \mid y_1, \ldots, y_{t-1}).$$

Taking the product over $t = j+1, \ldots, i$ then gives

$$f(y_{j+1}, \ldots, y_i \mid y_1, \ldots, y_j, d = j+1) = f(y_{j+1}, \ldots, y_i \mid y_1, \ldots, y_j). \tag{B.14}$$

After integration over $y_{i+2}, \ldots, y_n$, (B.13) leads to

$$f(y_{j+1}, \ldots, y_{i+1} \mid y_1, \ldots, y_j, d = j+1) = f(y_{j+1}, \ldots, y_{i+1} \mid y_1, \ldots, y_j). \tag{B.15}$$

Dividing (B.15) by (B.14) and equating the left- and right-hand sides, we find that

$$f(y_{i+1} \mid y_1, \ldots, y_i, d = j+1) = f(y_{i+1} \mid y_1, \ldots, y_i).$$

This holds for all $j \leq i$, and Lemma B.1 shows this is equivalent to ACMV.

ACMV $\Rightarrow$ MAR

Starting from the ACMV assumption and Lemma 1, we have

$$\forall t \geq 2, \forall j < t : f(y_t \mid y_1, \ldots, y_{t-1}, d = j+1) = f(y_t \mid y_1, \ldots, y_{t-1}). \tag{B.16}$$

We now factorize the full data density as

$$\begin{aligned} &f(y_1, \ldots, y_n, d = i+1) \\ &= f(y_1, \ldots, y_i, d = i+1) f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i, d = i+1) \end{aligned}$$

$$= f(y_1, \ldots, y_i, d = i+1) \prod_{t=i+1}^{T} f(y_t \mid y_1, \ldots, y_{t-1}, d = i+1).$$

Using (B.16), it follows that

$$\begin{aligned}
& f(y_1, \ldots, y_n, d = i+1) \\
&= f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1) \prod_{t=i+1}^{T} f(y_t \mid y_1, \ldots, y_{t-1}) \\
&= f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1) f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i) \\
&= f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1) \frac{f(y_1, \ldots, y_i) f(y_{i+1}, \ldots, y_n \mid y_1, \ldots, y_i)}{f(y_1, \ldots, y_i)} \\
&= f(y_1, \ldots, y_i \mid d = i+1) f(d = i+1) \frac{f(y_1, \ldots, y_n)}{f(y_1, \ldots, y_i)} \\
&= f(d = i+1 \mid y_1, \ldots, y_i) f(y_1, \ldots, y_n). \qquad \text{(B.17)}
\end{aligned}$$

An alternative factorization of $f(y, d)$ gives

$$f(y_1, \ldots, y_n, d = i+1) = f(d = i+1 \mid y_1, \ldots, y_n) f(y_1, \ldots, y_n). \qquad \text{(B.18)}$$

It follows from (B.17) and (B.18) that

$$f(d = i+1 \mid y_1, \ldots, y_n) = f(d = i+1 \mid y_1, \ldots, y_i),$$

completing the proof of Theorem 20.1.