Statistical Analysis with Missing Data

Module 5

Likelihood theory for complete and incomplete data

Likelihood methods

- Statistical model + data $\Rightarrow$ Likelihood
- Two general approaches based on likelihood
  - maximum likelihood (ML) inference for large samples
  - Bayesian inference for small samples:

    Posterior = prior x likelihood
- Methods can be applied to incomplete data
  - Because they do not require rectangular data sets
- First review main ideas for complete data
  - Little and Rubin (2019, chapter 6)

Parametric Likelihood

- Data y
- Statistical model yields probability density $f_Y(y \mid \theta)$

  for y with unknown parameters θ
- Likelihood function is then the density as a function of θ

$$L(\theta \mid y) = \text{const} \times f_Y(y \mid \theta)$$

- Loglikelihood is often easier to work with:

$$\ell(\theta \mid y) = \log L(\theta \mid y) = \text{const}' + \log\{f_Y(y \mid \theta)\}$$

  Constants can depend on data but not on parameter θ

Example: Normal sample

- univariate iid normal sample

Data $y = (y_1, ..., y_n)$

Parameters $\theta = (\mu, \sigma^2)$, μ = mean, σ^2 = variance

Normal density: $f_Y(y \mid \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$

Likelihood: $L(\mu, \sigma^2 \mid y) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2 \right)$

Loglikelihood: $\ell(\mu, \sigma^2 \mid y) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mu\right)^2$

Example: Multinomial sample

- Univariate *K*-category multinomial sample $y = (y_1, \ldots, y_n)$

  n_j = number of y_i equal to j (j=1,…,K)

$$\theta = (\pi_1, \ldots, \pi_{K-1}); \;\; \pi_K = 1 - \pi_1 - \ldots - \pi_{K-1}$$

$$f_Y(y \mid \pi_1, \ldots, \pi_{K-1}) = \left(\frac{n!}{n_1! \ldots n_K!} \right) \left(\prod_{j=1}^{K-1} \pi_j^{n_j} \right) (1 - \pi_1 - \ldots - \pi_{K-1})^{n_K}$$

$$L(\pi_1, \ldots, \pi_{K-1} \mid y) = \left(\prod_{j=1}^{K-1} \pi_j^{n_j} \right) (1 - \pi_1 - \ldots - \pi_{K-1})^{n_K}$$

Maximum Likelihood Estimate

- The maximum likelihood (ML) estimate $\hat{\theta}$ of θ maximizes the likelihood

$$L(\hat{\theta} \mid y) \geq L(\theta \mid y) \text{ for all } \theta$$

- The ML estimate is the "value of the parameter that makes the data most likely"
- The ML estimate is <u>not always unique:</u>
  - it is for many regular problems with complete data
  - Multiple maxima and nonunique ML estimates are more common when there are missing data

Computing the ML estimate

- In regular problems, the ML estimate can be found by solving the score equation

$$S(\theta \mid y) \equiv \frac{\partial \log L(\theta \mid y)}{\partial \theta} = 0$$

where *S* is the score function.

Explicit solutions for some models (normal regression, multinomial, …)

Iterative methods – e.g. Newton-Raphson, Scoring, EM algorithm -- required for other problems (logistic regression, repeated measures models, non-monotone missing data)

Normal Examples

- Univariate Normal sample $y = (y_1, ..., y_n)$ $\theta = (\mu, \sigma^2)$

$$\hat{\mu} = \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

(Note the lack of a correction for degrees of freedom)

- Multivariate Normal sample

$$\hat{\mu} = \bar{y},\ \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$$

- Normal Linear Regression (possibly weighted)

$$(y_i \mid x_{i1}, ..., x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}, \sigma^2 / w_i)$$

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) = \text{ weighted least squares estimates}$$

$$\hat{\sigma}^2 = (\text{weighted residual sum of squares})/n$$

Multinomial Example

$$y = (y_1, ..., y_n);\ y_i \sim \text{MNOM}(\pi_1, ..., \pi_K)$$

n_j = number of y_i equal to j $(j = 1, ..., K)$

Score equations:

$$\frac{\partial \ell}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_K}{1 - \pi_1 - ... - \pi_{K-1}} = 0,\ j = 1, ..., K-1$$

Hence ML estimate is

$$\hat{\pi}_j = n_j / n,\ j = 1, ..., K$$

More complex problems require iterative methods

Properties of ML estimates

- Under assumed model and regularity conditions, ML estimate is:
  - Consistent (not necessarily unbiased)
  - Efficient for large samples
  - not necessarily the best for small samples
- ML estimate is transformation invariant

$$\text{If } \hat{\theta} \text{ is the ML estimate of } \theta, \text{ then}$$

$$\phi(\hat{\theta}) \text{ is the ML estimate of } \phi(\theta)$$

Role of assumptions

Consider balanced repeated-measures data:

$$\left(y_i = (y_{i1},\ldots,y_{iT}), x_i = (x_{i1},\ldots,x_{iT}), i = 1,\ldots,n\right)$$

where it is assumed that

$$E(y_{it} \mid x_{i1},\ldots,x_{iT}) = \beta_0 + \beta x_{it}$$

$$\text{cov}(y_{is}, y_{it} \mid x_{i1},\ldots,x_{iT}) = \sigma_{st}$$

Consider the following analysis methods:

ML assuming y_i given x_i is multivariate normal with the above mean and variance structure

GEE with the mean structure above and an unstructured covariance matrix

Which of these statements is true?

A. ML is less robust because unlike GEE it assumes normality.

B. ML is more efficient if the normal assumption is true

C. Both methods have the same properties

D. None of the above

E. All of the above

Modes of Inference

- Frequentist inference: inference based on sampling distribution of statistics in repeated sampling of (Y, M)
  - Tricky because M varies in repeated sampling
  - Frequentist ML inference: inference based on asymptotic distribution of the ML estimate (see next slide)
- Pure likelihood inference: inference that data only through likelihood ratios for pairs of parameter values (θ, θ^*)
  - Does not involve sampling distribution of ML estimate

By far the most common form of pure likelihood inference is:

- Bayesian inference: add a prior distribution for parameters,base inferences on the posterior distribution of θ

Large-sample ML Inference

- Basic large-sample approximation:

  for regular problems,

$$\theta - \hat{\theta} \sim N(0, C)$$

  where C is a covariance matrix estimated from the sample

  – Frequentist ML inference: treats $\hat{\theta}$ as random, θ as fixed; equation defines the asymptotic sampling distribution of $\hat{\theta}$

  – Bayesian inference treats θ as random, $\hat{\theta}$ as fixed; equation defines large-sample approximation of the posterior distribution of θ

Forms of precision matrix

- The precision of the ML estimate is measured by C^{-1} Some forms for this are:
  - Observed information

$$C^{-1} = I(\hat{\theta} \mid Y) = -\left. \frac{\partial^2 \log L(\theta \mid Y)}{\partial\theta \, \partial\theta} \right|_{\theta=\hat{\theta}}$$

  - Expected information

$$C^{-1} = J(\hat{\theta}) = E\left[I(\theta \mid Y, \theta) \right]\Big|_{\theta=\hat{\theta}}$$

  - Sandwich estimator, C = covariance matrix of the bootstrap distribution of $\hat{\theta}$

Interval estimation

- 95% (confidence, probability) interval for scalar θ is:

$\hat{\theta} \pm 1.96\, C^{1/2}$,where 1.96 is 97.5 pctile of normal distribution

- Example: univariate normal sample

$$I = J = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4) \end{bmatrix} \Rightarrow C = \begin{bmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 2\hat{\sigma}^4/n \end{bmatrix}$$

Hence some 95% intervals are:

$$\bar{y} \pm 1.96\, s/\sqrt{n} \quad \text{for } \mu$$

$$s^2 \pm 1.96\, s^2/\sqrt{n/2} \quad \text{for } \sigma^2$$

$$\ln(s) \pm 1.96\, \sqrt{2/n} \quad \text{for } \ln(\sigma)$$

Significance Tests

Tests based on likelihood ratio (LR) or Wald (W) statistics:

$\theta = (\theta_{(1)}, \theta_{(2)})$; $\theta_{(1)0}$ = null value of $\theta_{(1)}$; θ_2 = other parameters

$\hat{\theta}$ = unrestricted ML estimate

$\tilde{\theta} = (\theta_{(1)0}, \tilde{\theta}_{(2)})$; $\tilde{\theta}_{(2)}$ = ML estimate of $\theta_{(2)}$ given $\theta_{(1)} = \theta_{(1)0}$

LR statistic: $\mathrm{LR}(\hat{\theta}, \tilde{\theta}) = 2\left[\ell(\hat{\theta} \mid Y) - \ell(\tilde{\theta} \mid Y)\right]$

Wald statistic: $W(\hat{\theta}, \tilde{\theta}) = (\theta_{(1)0} - \hat{\theta}_{(1)})^T C_{(11)}^{-1} (\theta_{(1)0} - \hat{\theta}_{(1)})$

$C_{(11)}$ = covariance matrix of $(\theta_{(1)} - \hat{\theta}_{(1)})$

yield P-values $P = pr\left(\chi_q^2 > D(\hat{\theta}, \tilde{\theta})\right)$

D = LR or Wald statistic; q = dimension of θ_0

χ_q^2 = Chi-squared distribution with q degrees of freedom

Bayes inference

- Given a prior distribution $\pi(\theta)$ for the parameters, inference can be based on the posterior distribution using Bayes' theorem:

$$p(\theta \mid Y) = \text{const.} \times \pi(\theta) \times L(\theta \mid Y)$$

- For small samples, I prefer Bayes' inference based on the posterior to the large sample ML approximation.
  - In important standard problems with non-informative priors, Bayes yields inference comparable to small-sample frequentist inference
  - In many non-standard problems, Bayes yields answers where no exact frequentist answer exists

Example: linear regression

The normal linear regression model:

$$(y_i \mid x_{i1},\ldots,x_{ip}) \sim N(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2)$$

with non-informative "Jeffreys'" prior:

$$\pi(\beta_0,\ldots,\beta_p,\sigma^2) \propto 1/\sigma^2$$

yields the posterior distribution of $(\beta_0,\ldots,\beta_p)$ as multivariate *T* with mean given by the least squares estimates $(\hat{\beta}_0,\ldots,\hat{\beta}_p)$, covariance matrix $(X^T X)^{-1} s^2$, where *X* is the design matrix, and degrees of freedom *n* - *p* - 1.

Resulting posterior credibility intervals are equivalent to standard t confidence intervals.

Simulating Draws from Posterior Distribution

- With problems with high-dimensional θ, it is often easier to draw values from the posterior distribution, and base inferences on these draws
- For example, if

$$(\theta_1^{(d)} : d = 1, \ldots, D)$$

is a set of draws from the posterior distribution for a scalar parameter θ_1, then

$\bar{\theta}_1 = D^{-1} \sum_{d=1}^{D} \theta_1^{(d)}$ approximates posterior mean

$s_\theta^2 = (D-1)^{-1} \sum_{d=1}^{D} (\theta_1^{(d)} - \bar{\theta}_1)^2$ approximates posterior variance

$(\bar{\theta}_1 \pm 1.96 s_\theta)$ or 2.5th to 97.5th percentiles of draws approximates 95% posterior credibility interval for θ

- Particularly useful for incomplete data, as we'll see…

Example: Posterior Draws for Normal Linear Regression

$(\hat{\beta}, s^2) =$ ls estimates of slopes and resid variance

$\sigma^{(d)2} = (n-p-1)s^2 / \chi^2_{n-p-1}$

$\beta^{(d)} = \hat{\beta} + A^T z \sigma^{(d)}$

χ^2_{n-p-1} = chi-squared deviate with $n-p-1$ df

$z = (z_1, ..., z_{p+1})^T$, $z_i \sim N(0,1)$

$A =$ upper triangular Cholesky factor of $(X^T X)^{-1}$:

$A^T A = (X^T X)^{-1}$

- Easily extends to weighted regression: see Example 6.17

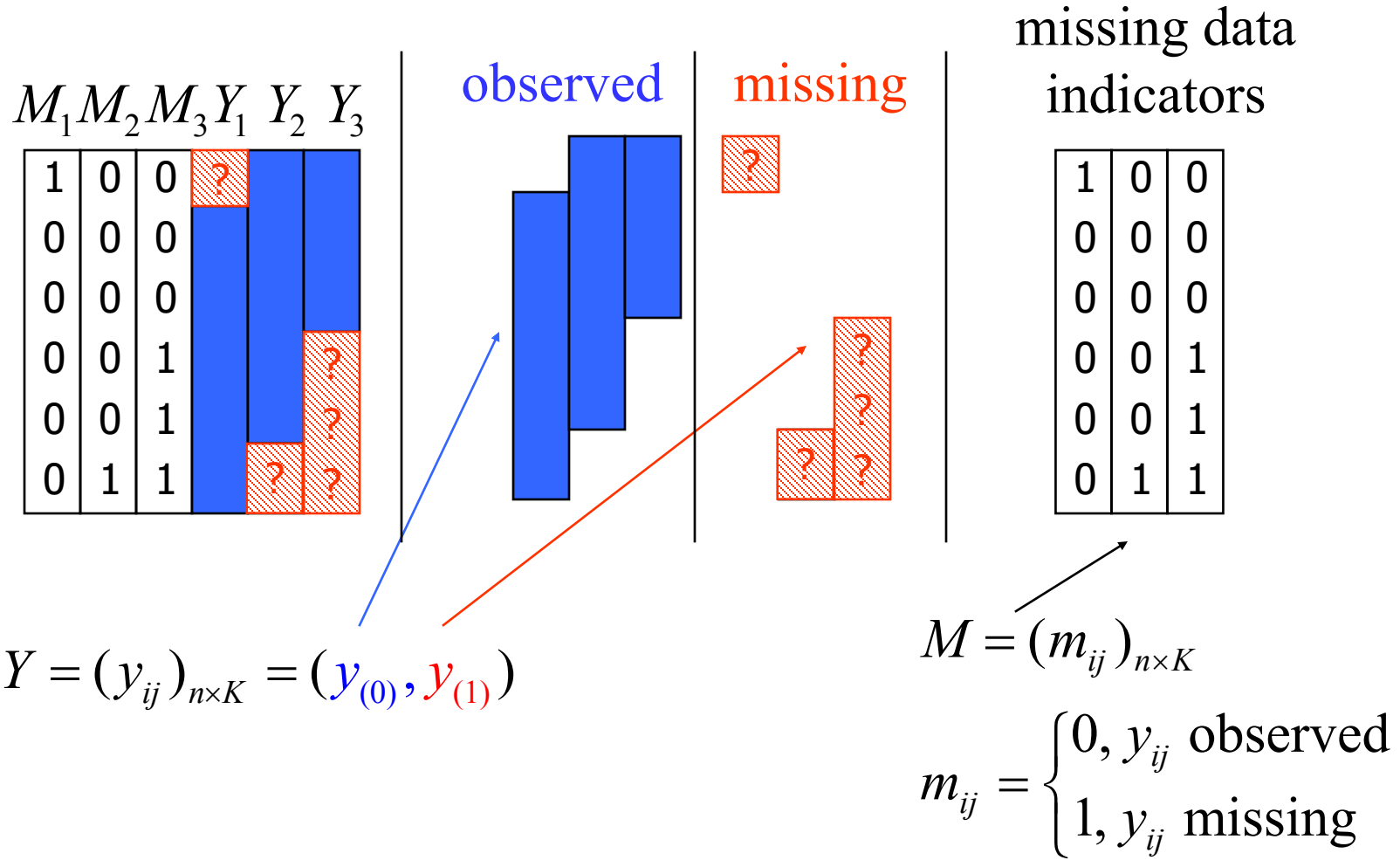
Summary

- We have reviewed basic ML results in the context of standard complete-data problems
- Bayes particularly useful for small sample problems
- Next consider applications to incomplete data

Likelihood methods with incomplete data

- Statistical model + incomplete data $\Rightarrow$ Likelihood
- Statistical models needed for:
  - data without missing values
  - missing-data mechanism
- Model for mechanism not needed if it is ignorable (to be defined later)
- With likelihood, proceed as before:
  - ML estimates, large sample standard errors
  - Bayes posterior distribution
  - Little and Rubin (2002, chapter 6)

The Observed Data



Model for Y and M

$$f_{Y,M}(y,m \mid \theta,\psi) = f_Y(y \mid \theta) \times f_{M|Y}(m \mid y,\psi)$$

Complete-data model

model for missingness mechanism

Example: bivariate normal monotone data

<u>complete-data model:</u>

$$(y_{i1}, y_{i2}) \sim_{\text{iid}} N_2(\mu, \Sigma)$$

<u>model for missingness mechanism:</u>

$$(m_{i2} \mid y_{i1}, y_{i2}) \sim_{\text{ind}} \text{Bern}\left[\Phi\left(\psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2}\right)\right]$$

Φ = Normal cumulative distribution function

M_1	M_2	Y_1	Y_2
0	0		
0	0		
0	0		
0	1		?
0	1		?

Two likelihoods

- *Full* likelihood - involves joint model for Y and M

$$f(y_{(0)}, m \mid \theta, \psi) = \int f_Y(y_{(0)}, y_{(1)} \mid \theta) f_{M|Y}(m \mid y_{(0)}, y_{(1)}, \psi) dy_{(1)}$$

$$\Rightarrow L_{\text{full}}(\theta, \psi \mid y_{(0)}, m) = \text{const} \times f(y_{(0)}, m \mid \theta, \psi)$$

- Likelihood *ignoring the missingness mechanism*
  - does not involve a model for M *given* Y

$$f(y_{(0)} \mid \theta) = \int f_Y(y_{(0)}, y_{(1)} \mid \theta) dy_{(1)}$$

$$\Rightarrow L_{\text{ign}}(\theta \mid y_{(0)}) = \text{const} \times f_Y(y_{(0)} \mid \theta)$$

Ignoring the missingness mechanism

- If at observed values $(\tilde{y}_{(0)}, \tilde{m})$ of $(y_{(0)}, m)$:

$$L_{\text{full}}(\theta, \psi \mid \tilde{y}_{(0)}, \tilde{m}) = L_{\text{rest}}(\psi \mid \tilde{m}, \tilde{y}_{(0)}) \times L_{\text{ign}}(\theta \mid \tilde{y}_{(0)})$$

where $L_{\text{rest}}(\psi \mid \tilde{m}, \tilde{y}_{(0)})$ does not depend on θ

then pure likelihood inference about θ can be based on

$L_{\text{ign}}(\theta \mid \tilde{y}_{(0)})$

• The missingness mechanism is then called *ignorable* for pure likelihood inference

Ignoring the missingness mechanism continued

- Sufficient conditions for ignoring the missingness mechanism for pure likelihood inference are:

(A) Missing at Random (MAR):

$$f_{M|Y}(m \mid \tilde{y}_{(0)}, y_{(1)}, \psi) = f_{M|Y}(m \mid \tilde{y}_{(0)}, \psi) \text{ for all } y_{(1)}, \psi$$

(B) Distinctness:

θ and ψ have distinct parameter spaces

(Bayes: prior distributions of θ and ψ are independent)

- If MAR holds but not distinctness, ML based on ignorable likelihood is "valid" but not fully efficient, so MAR is the key condition

Ignoring the missingness mechanism for frequentist inference

In general, frequentist inference requires that MAR also holds for future samples with different patterns of missing data, specifically the following stronger "missing always at random" assumption.

(A') Missing Always at Random (MAAR):

$$f_{M|Y}(m \mid y_{(0)}, y_{(1)}, \psi) = f_{M|Y}(m \mid y_{(0)}, \psi) \text{ for all } m, y_{(0)}, y_{(1)}, \psi$$

MAAR in most cases means missingness can only depend on fully observed variables.

However, I argue in Little (2020) that for asymptotic frequentist ML inference, the MAR assumption remains sufficient for ignoring the missingness mechanism

Example of MAAR vs MAR

- Consider comparing the means of two independent normal samples with the same variance. The hypothetical complete data consist of

- $(y_i, x_i), i = 1, \ldots, n$, y_i = outcome, $x_i = 1$ or 2 for two groups

  $(y_i \mid x_i = j, \theta) \sim_{\text{ind}} N(\mu_j, \sigma^2), \theta = (\mu_1, \mu_2, \sigma^2)$

  With no missing data, classical frequentist inference for difference in means $\delta = \mu_2 - \mu_2$ is based on pivotal quantity

  $$t = \left((\bar{y}_2 - \bar{y}_1) - \delta\right) / \left(s\sqrt{1/n_1 + 1/n_2}\right) \sim t_\nu, \nu = n_1 + n_2 - 2$$

  e.g. 95% CI is $I_{0.95}(\delta) = \bar{y}_2 - \bar{y}_1 \pm t_{\nu, 0.975}\left(s\sqrt{1/n_1 + 1/n_2}\right)$

Example of MAAR vs MAR

Suppose X is always observed, but Y potentially has missing values, so $m_i = 0$ if y_i is observed and $m_i = 1$ if y_i is missing.We assume that

$$\Pr(m_i = 0 \mid x_i, y_i, \psi) = h_1(x_i, \psi)$$

- Thus missingness is allowed to depend on X, that is, the response rates in the two groups can differ, but missingness does not depend on the value of Y.
- missingness is MAAR (and therefore also MAR), since the probability of response depends on the group indicator, which is always observed
- usual t inference methods can be applied to the set of units where Y is observed, provided $\nu > 0$

Example of MAAR vs MAR

Suppose on the other hand that

$\Pr(m_i = 0 \mid x_i, y_i, \psi) = h_1(x_i, y_i, \psi)$, where

$$h(x_i, y_i, \psi) = \begin{cases} 1, \text{ if } x_i = 1 \\ 1, \text{ if } x_i = 2 \text{ and } y_i \le \psi, \\ 0, \text{ if } x_i = 2 \text{ and } y_i > \psi. \end{cases}$$

For example, the values of Y in group 2 are measured using a flawed instrument that does not return values greater than an unknown ψ.

This mechanism is not MAAR, but it is MAR *if* all the values of Y happen to be observed in the realized data set.

The standard t confidence interval is then not valid in a frequentist sense, though it is a valid posterior credible interval for Bayes inference with a Jeffreys prior

Weakening MAAR for frequentist inference

Lemma 3 (Little, 2020). MAR and Distinctness are sufficient for ignoring the missingness mechanism for asymptotic frequentist ML inference with precision based on the inverse of the observed information.

Proof: asymptotic distribution for full model for Y and M is:

$$\begin{pmatrix} \theta - \hat{\theta} \\ \psi - \hat{\psi} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} I_{\text{full}(\theta\theta)} & I_{\text{full}(\theta\psi)} \\ I_{\text{full}(\theta\psi)} & I_{\text{full}(\psi\psi)} \end{pmatrix}^{-1} \right], \text{where}$$

$$I_{\text{full}(\theta\theta)} = -D_{\theta\theta}\left(\log L_{\text{full}}(\theta,\psi \mid \tilde{y}_{(1)}, \tilde{m})\right)\Big|_{\theta=\hat{\theta},\psi=\hat{\psi}},$$

$$I_{\text{full}(\theta\psi)} = -D_{\theta\phi}\left(\log L_{\text{full}}(\theta,\psi \mid \tilde{y}_{(1)}, \tilde{m})\right)\Big|_{\theta=\hat{\theta},\psi=\hat{\psi}}, \text{ and}$$

$$I_{\text{full}(\psi\psi)} = -D_{\phi\phi}\left(\log L_{\text{full}}(\theta,\psi \mid \tilde{y}_{(1)}, \tilde{m})\right)\Big|_{\theta=\hat{\theta},\psi=\hat{\psi}},$$

Weakening MAAR for frequentist inference

Suppose MAR holds but not MAAR. Because data are MAR, *for the realized sample*, the full likelihood factors as

$$L_{\text{full}}(\theta,\psi \mid \tilde{y}_{(0)},\tilde{m}) = L_{\text{rest}}(\psi \mid \tilde{m},\tilde{y}_{(0)}) \times L_{\text{ign}}(\theta \mid \tilde{y}_{(0)})$$

So $\hat{\theta}=\hat{\theta}_{\text{ign}}$ and $I = \begin{pmatrix} I_{\text{ign}(\theta\theta)} & 0 \\ 0 & I_{(\psi\psi)} \end{pmatrix}^{-1}$

Thus,tests and confidence intervals for θ are the same for full or ignorable model, when evaluated using the observed data. This is despite the fact that the sampling distribution of θ is only ignorable under the stronger MAAR assumption; for repeated samples where MAR is violated, the observed information matrix would not be block diagonal for θ and ψ and the full model would be needed.

Bayes: add prior distributions

$$p(\theta,\psi \mid y,m) = \pi(\theta,\psi) \times f_Y(y \mid \theta) \times f_{M|Y}(m \mid y,\psi)$$

Prior dist — Complete-data model — model for mechanism

(A) Posterior based on full model for Y and M :

$$p(\theta,\psi \mid \tilde{y}_{(0)},\tilde{m}) \propto \pi(\theta,\psi) \times f(\tilde{y}_{(0)},\tilde{m} \mid \theta,\psi), \text{ where}$$

$$f(\tilde{y}_{(0)},\tilde{m} \mid \theta,\psi) = \int f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) f_{M|Y}(\tilde{m} \mid \tilde{y}_{(0)}, y_{(1)},\psi) dy_{(1)}$$

(B) Posterior ignoring the missingness mechanism

(does not involve a model for M)

$$p(\theta \mid \tilde{y}_{(0)}) \propto \pi(\theta) \times f(\tilde{y}_{(0)} \mid \theta), \text{ where}$$

$$f(\tilde{y}_{(0)} \mid \theta) = \int f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) dy_{(1)}$$

Summary of Theory

- Likelihood Inference ignoring the missingness mechanism is valid if
  - Model for Y is correctly specified
  - Data are MAR
  - Fully efficient if distinctness condition holds
- In contrast, many ad-hoc methods require the stronger MCAR assumption

Extensions to semiparametric and nonparametric likelihood

example 1: nonparametric estimation of a distribution function

example 2: semiparametric model with missing covariates