COMPLETING THE INCOMPLETE

Estimating Functions, Likelihood, Imputation, and Bayes

Don McLeish and Cyntha Struthers

University of Waterloo

Dec 5, 2015

Methods for missing data: A more general approach

We have discussed some common methods for dealing with missing data and their drawbacks.

- Mean imputation
- Listwise deletion (LD) or complete-case analysis
- Pairwise deletion (PD) or available-case analysis
- Regression imputation
- Stochastic regression imputation
- Inverse probability weighting (IPW)
- Imputation (hot deck imputation, predictive mean matching, multiple imputation (RMI))

We now discuss a general and more intuitive approach to dealing with missing data.

The Mother[1] of all[2] estimators....

IS AN ESTIMATING EQUATION $\psi(\theta, Y) = 0$ where Y is the data. (unless you are Bayes!)

[1] or father

[2] (almost)

Estimating equations

- The estimate of θ is found by solving the estimating equation $\psi(\theta, Y) = 0$.
- $\psi(\theta, Y)$ is *an unbiased estimating function* if $E_\theta[\psi(\theta, Y)] = E[\psi(\theta, Y)|\theta] = 0$ for all θ.

  Why is this good?
- Example: If $E(Y_i) = \theta$ then $\psi(\theta, Y) = \sum_{i=1}^{n}(Y_i - \theta)$ is an unbiased estimating function for the mean.
- The score function $S(\theta, Y) = \frac{\partial}{\partial\theta} \ln[f_\theta(y)]$ always provides an unbiased estimating function (some authors denote S by U).

Estimating equations

- If $\psi(\theta, Y)$ is *an unbiased estimating function* such that

$$E_\theta[\psi(\theta, Y)] = E[\psi(\theta, Y)|\theta] = 0 \quad \text{for all } \theta$$

  does this imply that the estimators obtained by solving $\psi(\theta, Y) = 0$ for θ are unbiased?

- The asymptotic unbiasedness of the estimator is inherited from this property.

Estimating equations and the maximum likelihood (ML) estimator

- To find the maximum likelihood (ML) estimator we maximize the likelihood function

$$L(\theta) = f_\theta(y_1) \times f_\theta(y_2) \times \cdots \times f_\theta(y_n) = \prod_{i=1}^{n} f_\theta(y_i)$$

- The ML estimator is usually obtained by solving $\sum_{i=1}^{n} S_1(\theta; Y_i) = 0$ where $S_1(\theta; y) = \frac{\partial}{\partial\theta} \ln [f_\theta(y)]$.
- Since $E_\theta \left[\sum_{i=1}^{n} S_1(\theta; Y_i) \right] = 0$, $\sum_{i=1}^{n} S_1(\theta; Y_i)$ is an unbiased estimating function.

Estimating equations

- Note that the function $\sum_{i=1}^{n} S_1(\theta; Y_i)$ is a function of the observations $Y_1, Y_2, \ldots, Y_n$ and the parameter θ.
- Most sensible estimators, like the ML estimator, can be described easily through an estimating function.
- **Example:**

  If $Var_\theta(Y_i) = \theta$ for independent identically distributed Y_i, then we can use the estimating function

$$\psi(\theta, Y) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - (n-1)\theta$$

  to estimate the parameter θ, without any other knowledge of the distribution, its density, mean, etc.

Estimating equations and their asymptotic behaviour

Theorem

Suppose we estimate the parameter θ *using* $\hat{\theta}$ *the solution to the estimating equation* $\psi(\theta, Y) = 0$. *Suppose also that* $\psi(\theta, Y) = \sum_{i=1}^{n} \psi_1(\theta, Y_i)$ *where* $Y_1, Y_2, \ldots, Y_n$ *are* n *independent observations and* $E_\theta[\psi_1(\theta, Y_i)] = 0$. *Then under regularity conditions,* $\hat{\theta}$ *has an asymptotic Normal distribution with mean* θ *and variance*

$$\frac{Var_\theta[\psi(\theta, Y)]}{\left\{E_\theta\left[\frac{\partial}{\partial\theta}\psi(\theta, Y)\right]\right\}^2} = \frac{1}{n}\frac{Var[\psi_1(\theta, Y_i)]}{\left\{E_\theta\left[\frac{\partial}{\partial\theta}\psi_1(\theta, Y_i)\right]\right\}^2}$$

Godambe information

Definition

By analogy with the relationship between the asymptotic variance of the ML estimator and the Fisher information, we call the **reciprocal of the asymptotic variance**

$$J(\psi, \theta) = \frac{\left\{ E_\theta \left[\frac{\partial}{\partial \theta} \psi(\theta, Y) \right] \right\}^2}{Var_\theta \left[\psi(\theta, Y) \right]}$$

the *Godambe information* of the estimating function.

Multivariate case: $J(\psi, \theta)$

$$= \left[E_\theta \left\{ \frac{\partial}{\partial \theta} \psi(\theta, Y) \right\} \right] \left\{ Var_\theta \left[\psi(\theta, Y) \right] \right\}^{-1} \left\{ E_\theta \left[\frac{\partial}{\partial \theta} \psi(\theta, Y) \right] \right\}^T$$

Godambe information

Godambe (1960) proved that, among all unbiased estimating functions satisfying the usual regularity conditions, the (essentially unique[3]) estimating function which maximizes

$$J(\psi, \theta) = \frac{\left\{ E_\theta \left[\frac{\partial}{\partial \theta} \psi(\theta, Y) \right] \right\}^2}{Var_\theta [\psi(\theta, Y)]}$$

is the score function $S(\theta; Y)$.

For $\psi(\theta, Y) = S(\theta; Y)$, $J(\psi, \theta) =$ Fisher information $= E_\theta \left[-\frac{\partial}{\partial \theta} S(\theta, Y) \right] = Var_\theta [S(\theta, Y)]$

[3] up to multiplication by a non-random $c(\theta)$

Missing data paradigm

Algorithm for solving a missing data problem:

1. Write down the estimating function $\psi(\theta, Y)$ you would use if you had $Y =$ ALL the data (complete).

   For example, for $E(Y_i) = \theta$, $\psi(\theta, Y) = \sum_{i=1}^{n} (Y_i - \theta) = 0$.

2. Condition this estimating function (or project it) on Y_{obs}, the data you did observe, and solve $E_\theta [\psi(\theta, Y) | Y_{obs}] = 0$.

This simple algorithm preserves maximum likelihood!

If $\psi(\theta, Y)$ is the complete data score then $E_\theta [\psi(\theta, Y) | Y_{obs}]$ is called the incomplete data score.

The Fundamental Theorem for missing data: The Power of Projection!

Theorem (Fisher, 1925)

If $S(\theta; Y) = \frac{\partial}{\partial\theta} \ln f_\theta(y)$ *is the score function for ALL the data, and the data we actually observe is* $Y_{obs} = T(Y)$*, a function of* Y*, (e.g. coarsened, censored, or rounded data, data MAR, etc.) then the score function for* Y_{obs} *is* $E_\theta\left[S(\theta; Y)|Y_{obs}\right]$.

The Fundamental Theorem for missing data: The Power of Projection

- The information in Y_{obs} is (see Rao (1973) for proof)

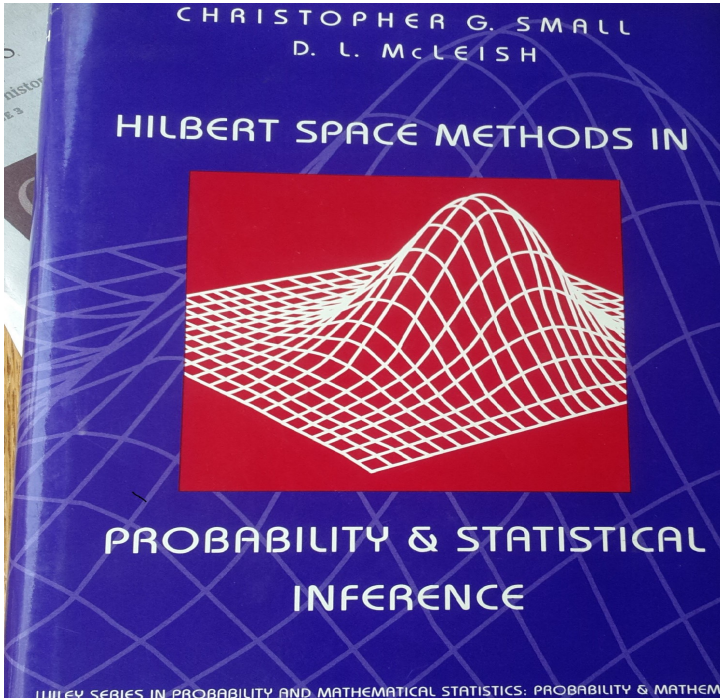
$$
\begin{aligned}
& Var\left\{E_\theta[S(\theta; Y)|Y_{obs}]\right\} \\
&= Var\left[S(\theta; Y)\right] - E_\theta\left\{Var\left[S(\theta; Y)|Y_{obs}\right]\right\} \\
&= E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ln f_\theta(Y)\right] - E_\theta\left[-\frac{\partial^2}{\partial\theta^2}\ln f_\theta(Y|Y_{obs})\right] \\
&= \text{Information in } Y - \text{Information in } (Y|Y_{obs})
\end{aligned}
$$

- $Var\left\{E_\theta[S(\theta; Y)|Y_{obs}]\right\}$ is smaller than the complete data information $Var\left[S(\theta; Y)\right]$.
- For the estimating function $\psi(\theta, Y)$, $\frac{1}{J(\psi,\theta)}$ is an underestimate of the standard error of the estimator if data are missing.

Projecting estimating functions

Theorem

Suppose $\psi(\theta; Y)$ is an optimal estimating function for ALL the data Y. If we observe $Y_{obs} = T(Y)$, a function of Y, then $E_\theta[\psi(\theta; Y)|Y_{obs}]$ is the optimal estimating function for the observation T. (See Small and McLeish)



Example: Normal mean model with data MCAR

Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\theta, 1)$ random variables and $R_1, R_2, \ldots, R_n$ are i.i.d. $Bernoulli(p)$ random variable so the data are MCAR.

- The complete data estimating equation based on ALL the data (Y, R) for θ is $\sum_{i=1}^{n} (Y_i - \theta) = 0$.
- Observed data Y_{obs} : $R_i Y_i$, $i = 1, 2, \ldots, n$.
- The estimating equation for the observed data is
$$E\left[\sum_{i=1}^{n} (Y_i - \theta) | Y_{obs}\right] = \sum_{i=1}^{n} R_i (Y_i - \theta) = 0.$$
- Solution: $\hat{\theta} = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$ = average of observed values (i.e. LD).

Example: Normal mean model with data MAR

- Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\theta, 1)$ random variables.
- Suppose $R_i = 1$ if $Y_i \leq 1$ and $R_i \backsim Bernoulli(p)$ if $Y_i > 1$, $i = 1, 2, \ldots, n$ so the data are MAR.
- Observed data Y_{obs} : $R_i Y_i$, $i = 1, 2, \ldots, n$.
- Estimating equation for observed data is

$$0 = \sum_{i=1}^{n} R_i(Y_i - \theta) + \sum_{i=1}^{n} (1 - R_i)(\hat{Y}_i - \theta)$$

  where[4] $\hat{Y}_i = E(Y_i | Y_i > 1)$ (Mill's ratio).
- Solution: $\hat{\theta}$ = average of values of observed Y_i's and values $\hat{Y}_i$ for those not observed.

[4] φ and Φ are N(0,1) p.d.f. and c.d.f. respectively

Example: Normal mean model with data MNAR

- Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\theta, 1)$ random variables.
- Suppose $R_i | Y_i$ are independent *Bernoulli* $\left(\frac{1}{1+Y_i^2}\right)$ random variables, so large values of $|Y_i|$ are less likely to be observed.
  (Rarely would we know this much.)
- The estimating function $\sum_{i=1}^{n} R_i(Y_i - \theta) = \sum_{i \in obs} (Y_i - \theta)$ is **biased** and the estimator is **inconsistent**.

Example: Normal mean model with data MNAR

- The estimating function based on ALL the data is

$$\psi(\theta; Y) = \sum_{i=1}^{n} (Y_i - \theta) + \sum_{i=1}^{n} \left(R_i - \frac{1}{1 + Y_i^2} \right)$$

so

$$E_\theta[\psi(\theta; Y) | Y_{obs}] \neq \sum_{i=1}^{n} R_i (Y_i - \theta).$$

- Could use the estimating function
$\sum_{i \in obs} [Y_i - E_\theta (Y_i | R_i = 1)]$.

EM Algorithm: Dempster, Laird and Rubin (1977)

- EM algorithm is a general technique for finding ML estimates for a parametric model when the data are not fully observed.
- Basic idea: Y_{mis} contains information relevant to estimating θ and θ helps to find likely values of Y_{mis}.
- Suggests the following algorithm for estimating θ: "Fill in" the missing data Y_{mis} based on an initial estimate of θ (e.g. estimate based on complete-cases). Then
  1. Update the estimate of θ using Y_{obs} and the filled-in Y_{mis}.
  2. "Fill in" the missing data Y_{mis} based on the updated estimate of θ (how?).
- Iterate steps 1-2 until convergence.

Question www.socrative.com (calgary) Incomplete data quiz

The "EM" algorithm obtains its name from the acronym for

A. ETHEL MERMAN (There's no business like show business!)
B. EXPECTATION-MAXIMUM
C. An algorithm for ElectroMagnetic propulsion
D. A cheap Russian vodka
E. A dyslexic narcissist's "ME" algorithm

Using the Fundamental Theorem to find estimates: A version of the EM algorithm

1. Assuming there are no missing data choose an estimating equation for the parameter θ: e.g. $\psi(\theta; Y) = 0$.
2. Use the complete cases to get a preliminary estimator $\hat{\theta}_1$ of θ.
3. Find $E_{\hat{\theta}_1}[\psi(\theta; Y)|Y_{obs}]$.

   For example, if $\psi(\theta; Y)$ is linear in Y_6 and Y_6 is missing, replace Y_6 by $E_{\hat{\theta}_1}(Y_6|Y_{obs})$.
4. Solve the equation $E_{\hat{\theta}_1}[\psi(\theta; Y)|Y_{obs}] = 0$ to find $\hat{\theta}_2$.
5. Repeatedly solve the equation $E_{\hat{\theta}_k}[\psi(\hat{\theta}_{k+1}; Y)|Y_{obs}] = 0$ for $\hat{\theta}_{k+1}$ as $k = 1, 2, \ldots$.

If $\hat{\theta}_{k+1}$ converges, it converges to a solution of $E_\theta[\psi(\theta; Y)|Y_{obs}] = 0$.

This kind of two-step algorithm is the basis of many

Example: Censored exponential with data MAR

- Suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $Exponential(\theta)$ random variables.
- Suppose for $a_i < b_i$ (either may be infinite) we observe Y_i if $Y_i \notin [a_i, b_i]$ and otherwise we only observe that $Y_i \in [a_i, b_i]$. (This setup includes right censoring, left censoring, interval censoring.)
- Let $R_i = 1$ if the observation Y_i is observed exactly (so $Y_i \notin [a_i, b_i]$), otherwise $R_i = 0$.
- Observed data Y_{obs} : $R_i Y_i$, $i = 1, 2, \ldots, n$.
- Estimating equation if there are no missing data:
$$\psi(\theta; Y) = \sum_{i=1}^{n} (Y_i - \theta) = 0$$

Example: Censored exponential data cont'd

Estimating equation if there are no missing data:

$$\psi(\theta; Y) = \sum_{i=1}^{n} (Y_i - \theta) = 0$$

- Let $\hat{Y}_i(\theta) = \begin{cases} E_\theta(Y_i | a_i < Y_i < b_i) & \text{if } R_i = 0 \\ Y_i & \text{if } R_i = 1 \end{cases}$
  where $E_\theta(Y_i | a \leq Y_i \leq b) = \theta + \frac{ae^{-a/\theta} - be^{-b/\theta}}{e^{-a/\theta} - e^{-b/\theta}}$.
- Then $E_\theta[\psi(\theta; Y) | Y_{obs}] = \sum_{i=1}^{n} \left[\hat{Y}_i(\theta) - \theta\right]$.
- Solve $\sum_{i=1}^{n} \left[\hat{Y}_i(\hat{\theta}_1) - \theta\right] = 0$ for $\hat{\theta}_2$ where $\hat{\theta}_1$ = mean of observed Y_i's.
- The ML estimate is found iteratively using $\hat{\theta}_{k+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i(\hat{\theta}_k)$.

Using the Fundamental Theorem to find estimates:

To use this algorithm we need $E_{\hat{\theta}}[\psi(\theta; Y)|Y_{obs}]$ which requires calculating conditional expectations.

- If the estimating function for the parameter of interest (e.g. variance) requires Y_i^2 and Y_i is missing then we replace Y_i^2 by $E_\theta\left(Y_i^2|Y_{obs}\right)$.
  (NOT $\left[E_\theta\left(Y_i|Y_{obs}\right)\right]^2$)
- If $E_\theta[\psi(\theta; Y)|Y_{obs}]$ is difficult to determine, then a possible solution is to use stochastic imputation, that is, we "fill-in" Y_{mis} by randomly generating Y_{mis} from the conditional distribution of Y_{mis} given Y_{obs}.
  - Such an algorithm is referred to as a stochastic EM algorithm.

A simulation-based version EM algorithm: Stochastic EM

The "correct" imputation of the missing data is from **the true distribution** $f_\theta\left(y_{mis}|y_{obs}\right)$, where θ is the true value.

- Since θ is unknown, we could impute Y_{mis} using the current estimate of θ.
- Let $Y^*_{\hat{\theta}_k}$ be the completed data consisting of Y_{obs} and the values for Y_{mis} imputed using $\hat{\theta}_k$, the current estimate of θ.
- Use $Y^*_{\hat{\theta}_k}$ to obtain updated estimate $\hat{\theta}_{k+1}$.

A simulation-based version EM algorithm: Stochastic EM

Impute Y_{mis}	Estimate θ assuming complete data
Impute Y_{mis} using $f(y_{mis}/\theta_0, y_{obs})$, θ_0 obtained from complete data	Solve for θ_1, $\psi(\theta;Y)=0$, where $Y=(Y_{mis}, Y_{obs})$
Impute Y_{mis} using $f(y_{mis}/\theta_1, y_{obs})$	Solve for θ_2, $\psi(\theta;Y)=0$, where $Y=(Y_{mis}, Y_{obs})$
Impute Y_{mis} using $f(y_{mis}/\theta_2, y_{obs})$	Solve for θ_3, $\psi(\theta;Y)=0$, where $Y=(Y_{mis}, Y_{obs})$
...	...

A simulation-based version EM algorithm: Stochastic EM

The updated estimate $\hat{\theta}_{k+1}$ inherits "noise" from the imputed $Y_{mis}|\hat{\theta}_k$.

- How else can we obtain an updated estimate $\hat{\theta}_{k+1}$?
- Ideally we want to solve the equation $E_\theta[\psi(\theta; Y^*_\theta)|Y_{obs}] = 0$ for θ.
- If we solve $\psi(\hat{\theta}_{k+1}; Y^*_{\hat{\theta}_k}) = 0$ for $\hat{\theta}_{k+1}$, then each imputation is random so solution depends on generated $Y^*_{\hat{\theta}_k}$.
- We can use averaging to deal with the noise, that is, generate m imputations $Y^{*(1)}_{\hat{\theta}_k}, Y^{*(2)}_{\hat{\theta}_k}, \ldots, Y^{*(m)}_{\hat{\theta}_k}$ and find $\hat{\theta}_{k+1}$ by solving

$$\frac{1}{m}\sum_{i=1}^{m}\psi(\hat{\theta}_{k+1}; Y^{*(i)}_{\hat{\theta}_k}) = 0.$$

A simulation-based version EM algorithm: Stochastic EM

An alternative to averaging is to use an algorithm suggested by Robbins-Monro (1951) for solving $E[M(\theta)] = 0$ where $M(\theta)$ is a random function:

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{c}{n} M\left(\hat{\theta}_n\right), \text{ where } c \text{ is a constant.}$$

- Using this idea, $\hat{\theta}_{k+1}$ is found by solving

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{c}{n} \psi\left(\hat{\theta}_k; Y^*_{\hat{\theta}_k}\right).$$

- If $\hat{\theta}_k$ converges, it converges to the solution of $E_\theta[\psi(\theta; Y)|Y_{obs}] = 0$.
- In particular if $\psi(\theta; Y)$ is the score function then $\hat{\theta}_k$ converges to the ML estimate of θ.

A simulation-based version EM Algorithm: Stochastic EM

At convergence, the iterations provide imputations Y_θ^* of the missing values for the ML estimate of the distribution.

- This is the gold standard for imputations.
- **We still need to use the appropriate (incomplete data) information matrix for confidence intervals.**

What to do when your confidence is lacking...

- Let $L(\theta, Y)$ be the likelihood. For MAR, the ML estimate is the solution to

$$\frac{\partial}{\partial \theta} \ln f_\theta(y_{obs}) = E_\theta \left[S(\theta; Y) | Y_{obs} \right] = 0$$

$$\text{where } S(\theta; Y) = \frac{\partial}{\partial \theta} \ln L(\theta, Y)$$

  The ML estimator has asymptotic covariance matrix given by the inverse of the Fisher Information matrix, $J(S, \theta) = Var_\theta \left[\frac{\partial}{\partial \theta} \ln f_\theta(y_{obs}) \right] = E_\theta \left[-\frac{\partial^2}{\partial \theta^2} \ln f_\theta(y_{obs}) \right]$.

- $J(S, \theta)$ is often difficult to obtain
- Observed information $J_o = -\frac{\partial^2}{\partial \theta^2} \ln f_\theta(y_{obs})$ is asymptotically equivalent, easier to obtain and better for many purposes (Wald tests, CI's, etc.).

The confidence game

For the regular exponential family (Normal, Poisson, etc.) with canonical parameter θ, $f_\theta(y) = e^{\theta^T t(y) - a(\theta)} h(y)$.

- Score function: $S(\theta, y) = t(y) - a(\theta)$
- Observed Information: $J_o(\theta) = a'(\theta)$
- For most missing data ML software, J_o is a by-product of the EM or SEM algorithm.
- Asymptotic variance for original parametrization, say $\eta = \eta(\theta)$, is obtainable by the delta method.
- For the Normal distribution with monotone missingness, explicit parameter estimators and Fisher information are available.

The importance of the MAR assumption

Data which are MNAR require a model for the joint distribution (Y, R). Why not always model (Y, R)?

- Since we observe no information about $P(Y|R = 0)$, we cannot assess the fit of the assumed model for $P(Y = y, R = 0)$.
- Statisticians prefer models that allow some assessment of fit.
- We do observe information about $P(Y|R = 1)$.
- We extrapolate properties of $P(Y|R = 0)$ from those that are estimable based on our observed information about $P(Y|R = 1)$.
- The MAR assumption allows us to do this.

When can we ignore the distribution of R?

Suppose the pdf of the complete data Y is $f_\theta(y)$ and the conditional pdf of $R|Y$ is $f_\gamma(r|y)$ which may depend on a parameter γ.

- The joint likelihood for (Y_{obs}, R) is

$$L(\theta, \gamma, y_{obs}, r) = \int f_\theta\left(y_{obs}, y_{mis}\right) f_\gamma\left(r|y_{obs}, y_{mis}\right) dy_{mis} \qquad (*)$$

- If the observations are MAR, then $f_\gamma\left(r|y_{obs}, y_{mis}\right) = f_\gamma\left(r|y_{obs}\right)$ and (*) equals

$$f_\gamma\left(r|y_{obs}\right) \int f_\theta\left(y_{obs}, y_{mis}\right) dy_{mis} = f_\gamma\left(r|y_{obs}\right) f_\theta\left(y_{obs}\right).$$

- Since the right hand side is a function of γ times a function of θ, maximizing $f_\theta\left(y_{obs}\right)$ is equivalent to maximizing the joint likelihood of θ and γ over θ.

When can we ignore the distribution of R?

Fact

If the observations are MAR, maximizing $f_\theta(y_{obs})$ leads to consistent, fully efficient estimators.

Example: ML estimates for bivariate normal (BVN)

- $Y_1 = \log(\text{head size})$ completely observed
  $Y_2 = \log(\text{brain weight})$ sometimes missing[5]
- Likelihood under MAR assumption:

$$L(\mu_1, \mu_2, \Sigma) = \prod \left[f(y_1, y_2 | \mu, \Sigma) \right]^R \left[f(y_1 | \mu_1, \sigma_1) \right]^{(1-R)}$$

  where $f(y_1, y_2 | \mu, \Sigma)$ is $BVN(\mu, \Sigma)$ density with $\mu = (\mu_1, \mu_2)$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ and $f(y_1 | \mu_1, \sigma_1)$ is $N(\mu_1, \sigma_1^2)$ density.
  $f(y_1, y_2 | \mu, \Sigma)$ is the contribution to L if y_1, y_2 observed.
  $f(y_1 | \mu_1, \sigma_1)$ is the contribution to L if only y_1 observed.
- Observed data information: $-\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \ln(L)$ can be obtained explicitly (5×5 matrix).

[5] (sometimes the head is there but the brain appears to be missing)

Example: ML estimation for BVN: the brainweight data

- $Y_1 = \log(\text{head size})$ completely observed
  $Y_2 = \log(\text{brain weight})$ sometimes missing
- ML estimates: $\hat{\mu}_1 = \bar{y}_1$, $\hat{\sigma}_1^2 = \hat{var}(Y_1)$, $\hat{\mu}_2 = 7.1582$

parameter	μ_1	μ_2	σ_1	σ_2	σ_{12}
estimate	8.193	7.1582	0.1005	0.0938	0.0076

- Observed data information

$$J_o = -\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta}\ln(L) \text{ is a } 5 \times 5 \text{ matrix}$$

- A 95% CI for μ_2 is $\hat{\mu}_2 \pm 1.96\sqrt{\left(J_o^{-1}\right)_{22}}$

Question-Bayesian approach to missing data (www.socrative.com (calgary))

Are you familiar with the following terms:

A. prior distribution
B. posterior distribution
C. conjugate prior
D. noninformative prior, improper prior, Jeffreys' prior
E. Bayes estimator, credible region (Bayesian confidence interval)

Bayes estimation

- In a Bayesian approach to estimation the parameter θ is assumed to be a realization of some larger random experiment generated according to the distribution $\pi(\theta)$ called the *prior distribution*.
- The observations $y_1, y_2, \ldots, y_n$ are assumed to be independent realizations drawn from $f_\theta(y)$, the conditional distribution of y given θ.
- $\pi(\theta)$ quantifies the information about θ prior to the data $y_1, y_2, \ldots, y_n$ being observed.
- $\pi(\theta)$ can be constructed using past data or subjective beliefs based on expert opinion. The form is sometimes chosen to convey little prior knowledge about the distribution of θ or for mathematical convenience.

Posterior distribution

- Suppose a value of θ is drawn at random from $\pi(\theta)$ and then given this value of θ the i.i.d. observations $y_1, y_2, \ldots, y_n$ are drawn from the conditional distribution $f_\theta(y)$.
- The *posterior distribution of the parameter* is the conditional distribution of θ given the data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$

$$\pi(\theta|\mathbf{y}) = c\pi(\theta)\prod_{i=1}^{n} f_\theta(y_i) = c\pi(\theta)L(\theta)$$

where

$$c = \left[\int_{-\infty}^{\infty} \pi(\theta)L(\theta)d\theta\right]^{-1}$$

Choosing a prior: Conjugate priors

If a prior distribution has the property that the posterior distribution is in the same family of distributions as the prior then the prior is called a *conjugate prior*.

- Suppose $(Y_1, \ldots, Y_n)$ is a random sample from the exponential family $f_\theta(y) = C(\theta)\exp[q(\theta)T(y)]h(y)$.
- Suppose also that θ is assumed to have the prior distribution with parameters a, b given by

$$\pi(\theta) = k[C(\theta)]^a \exp[bq(\theta)]$$

where

$$k = \left[\int_{-\infty}^{\infty} [C(\theta)]^a \exp[bq(\theta)]d\theta\right]^{-1}$$

Choosing a prior: Conjugate priors

The posterior distribution of θ, given the data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is

$$\pi(\theta|x) = c\,[C(\theta)]^{a+n} \exp\left\{ q(\theta) \left[b + \sum_{i=1}^{n} T(y_i) \right] \right\}$$

where

$$c = \frac{1}{\int_{-\infty}^{\infty} [C(\theta)]^{a+n} \exp\left\{ q(\theta) \left[b + \sum_{i=1}^{n} T(y_i) \right] \right\} d\theta}$$

- Note that the posterior distribution is in the same family of distributions as $f_\theta(y)$ and thus $\pi(\theta)$ is a conjugate prior.
- The parameters a and b of the posterior distribution reflect the choice of parameters in the prior.

Conjugate priors

Conjugate prior distribution for a random sample from each of the following distributions:

Distribution	**Parameter(s)**	**Conjugate Prior**
Uniform$(0, \theta)$	θ	Pareto(a, b)
Poisson(θ)	θ	Gamma
N(θ, σ^2) σ^2 known	θ	Normal(μ, σ_0^2)
N(μ, θ) μ known	θ	Inverse (reciprocal) Gamma
Gam(α, θ) α known	θ	Inverse (reciprocal) Gamma
N$(\mu, \frac{1}{\theta})$	μ, θ	$c\theta^{b_1/2} \exp e^{-\frac{\theta}{2}\left[a_1 + b_2(a_2-\mu)^2\right]}$ Normal-Gamma

Noninformative prior distributions

- The conjugate prior is often motivated by mathematical convenience.
- The prior should accurately represent the preliminary uncertainty about the plausible values of θ, and this may not translate easily into a conjugate prior distribution.
- Noninformative priors provide standard representation of ignorance about θ. A noninformative prior is arguably more objective than a subjectively assessed prior distribution since the latter may contain personal bias as well as background knowledge.
- The amount of information in the prior is always far less than the information contained in the data. (Little point in worrying about a precise specification of the prior distribution.)

Ignorance is bliss: Noninformative prior dist'ns

- If θ only takes values on the interval $[0, 1]$ then choosing the Uniform$(0, 1)$ distribution as the prior reflects the fact no value of θ is preferred over another.
- If θ takes values on the interval $(-\infty, \infty)$ and we assume the prior$\pi(\theta) = c$ for $\theta \in (-\infty, \infty)$ then

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta = \infty.$$

  which is not a proper density.
- Such prior densities are called improper priors.

Improper prior distributions

Even though the prior is improper, the posterior may be proper.

For example if $\pi(\theta) = 1$, and Y is $\mathrm{N}(\theta, 1)$. Then

$$\pi(\theta|y) = L(\theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2}, \text{ the } N(y, 1) \text{ density.}$$

Fact

The likelihood function is proportional to the posterior distribution of the parameter when using a uniform improper prior on the whole real line.

For a scale parameter (e.g. variance) which takes on positive values, often assume the logarithm of the parameter is uniform.

Independent improper priors for the Normal

- Let $(Y_1, \ldots, Y_n)$ be a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution. Assume that the prior distributions of μ, and $\log(\sigma^2)$ are independent improper uniform distributions.
- The marginal posterior distribution of μ given the data $\mathbf{y} = (y_1, \ldots, y_n)$ is such that

$$\sqrt{n}(\mu - \bar{y})/s \backsim t\,(n-1)\,.$$

- The marginal posterior distribution of σ^2 given the data $\mathbf{y}$ is such that

$$\frac{1}{\sigma^2} \backsim \mathit{Gamma}\left(\frac{n-1}{2},\ \frac{2}{(n-1)S^2}\right)$$

Jeffreys' prior

- A problem with noninformative prior distributions is whether the prior distribution should be uniform for θ or some function of θ, such as θ^2 or $\log(\theta)$.
- It is common to use a uniform prior for $\tau = h(\theta)$ where $h(\theta)$ is the function of θ whose Fisher information is constant.
- This idea is due to Jeffreys and leads to a prior distribution which is proportional to the square root of the determinant of the Fisher Information $|J(\theta)|^{1/2}$.
- Such a prior is referred to as a *Jeffreys' prior*.

Bayesian inference

- Bayesian inference is based on the posterior distribution $\pi(\theta|\mathbf{y})$ which depends on the data $\mathbf{y}$ through the likelihood function since $\pi(\theta|\mathbf{y}) \propto \pi(\theta)L(\theta;\mathbf{y})$.
- For example the *Bayes estimator for squared error loss* is the mean of the posterior distribution

$$\hat{\theta}_B = E(\theta|\mathbf{Y}) = \int_{-\infty}^{\infty} \theta\pi(\theta|\mathbf{y})d\theta$$

which minimizes

$$E\left[\left(\theta - \hat{\theta}_B\right)^2\right] = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \left(\theta - \hat{\theta}_B\right)^2 f_\theta(\mathbf{y})d\mathbf{y}\right] \pi(\theta)d\theta$$

Asymptotic Normality of the posterior distribution

Recall

$$\pi(\theta|\mathbf{y}) \propto \pi(\theta) \prod_{i=1}^{n} f_\theta(y_i) = \pi(\theta) L(\theta).$$

Taking the logarithm of the right side and expanding we have

$$\ln \pi(\theta) + \sum_{i=1}^{n} \ln f_\theta(y_i) \approx \ln \pi(\theta) - \frac{1}{2}(\theta - \hat{\theta})^2 I_n(\hat{\theta})$$

where $\hat{\theta}$ is the ML estimate and I_n is the observed information.

- This means that the posterior distribution is asymptotically Normal with mean $\hat{\theta}$ and variance $\left[I_n(\hat{\theta})\right]^{-1}$.
- The ML estimator and the Bayes estimator are asymptotically equivalent, i.e. $\hat{\theta}_B - \hat{\theta} = o_p(n^{-1/2})$.

The Fundamental Theorem for missing data: Bayesian version

Theorem

If $f(\theta|Y)$ is the posterior distribution function for the complete data Y and the data we actually observe is $Y_{obs} = T(Y)$, a function of Y, (e.g. coarsened, censored, or rounded data, data MAR, etc.) then the posterior distribution based Y_{obs} is $E[f(\theta|Y)|Y_{obs}]$.

Similarly if $\hat{\theta}_B$ is the Bayes estimator based on complete data then $E\left(\hat{\theta}_B|Y_{obs}\right)$ is the Bayes estimator based Y_{obs}.

It's not about U (the score function), it's all about MI (multiple imputation)

Recall the steps of multiple imputation:

- A random sample is drawn from *some* distribution to replace the missing values. (How?)
- This is done m times to form m completed datasets.
- Each of the m completed datasets are analysed and the results are combined using Rubin's Rules.

Rubin's rules for combining completed datasets

Let $\hat{Q}$ denote an estimate of a parameter θ and let U be its estimated variance **assuming complete data**. For example, $\hat{Q}$ could be an estimated regression coefficient and U its squared standard error.

1. Impute m datasets to obtain estimates $\hat{Q}_1, \ldots, \hat{Q}_m$ with corresponding variances $U_1, \ldots, U_m$.
2. Obtain the estimate $\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$ and let $\bar{U} = \frac{1}{m} \sum_{i=1}^{m} U_i$.
3. Estimate the variance of $\bar{Q}$ using $T = \bar{U} + \left(1 + \frac{1}{m}\right) B$, where $B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2$ is the between-imputation variance.

Making sense of Rubin's rules for combining completed datasets

- $\sqrt{T}$ is the overall standard error associated with $\bar{Q}$ where $T = \bar{U} + \left(1 + \frac{1}{m}\right) B$.
- If there were no missing data, then $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ would be identical, B would be 0 and $T = U$.
- The size of B relative to U is a reflection of how much information is contained in the missing part of the data relative to the observed.
- $r = \frac{B}{\bar{U}}(1 + \frac{1}{m}) =$ "relative increase in variance due to missingness".
- $\lambda = \frac{B}{T}(1 + \frac{1}{m}) =$ "proportion of variance due to missingness".

Making sense of Rubin's rules for combining completed datasets cont'd

Suppose $Q = Q(y_{obs}, y_{mis})$ is the statistic we would have used to estimate θ, had the complete data all been available.

We use $\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i(y_{obs}, \dot{y}_{mis}^{(i)})$ instead where $\dot{y}_{mis}^{(i)}$ is imputed.

Then the error is

$$\bar{Q} - \theta \quad = (Q - \theta) \quad + [E(Q|y_{obs}) - Q] \quad + [\bar{Q} - E(Q|y_{obs})]$$

and $Var(\bar{Q} - \theta)$

$$= \underset{\approx}{Var(Q - \theta)} + \underset{\bar{U}}{Var(E(Q|y_{obs}) - Q)} + \underset{+B}{Var(\bar{Q} - E(Q|y_{obs}))}$$
$$\underset{+\frac{1}{m}B}{}$$

since $\bar{U}$ estimates $Var(Q)$ and B, the between-imputation variance, estimates $Var[Q - E(Q|y_{obs})]$ and $Var[\hat{Q}_i - E(\hat{Q}_i|y_{obs})]$.

Confidence intervals based on Rubin's rules for combining data analyses

Rubin gives an approximate 95% confidence interval as

$$\bar{Q} \pm a\sqrt{T}$$

where a is the value such that $P\left(-a \leq X \leq a\right) = 0.95$ where $X \backsim t\left(\nu\right)$ and

$$\nu = (m-1)\left[1 + \frac{m\overline{U}}{(m+1)B}\right]^2.$$

Bayesian approach to imputing the missing data

Sample missing values from the *posterior predictive distribution*.

How?

1. Generate $\tilde{\theta}$ from its posterior distribution given the observed data: $E[f(\theta|Y)|Y_{obs}]$.
2. Impute the missing data using $f_{\tilde{\theta}}(y_{mis}|y_{obs})$.

Toy example of sampling from the posterior distribution

- Y_i are independent $N(\theta, 1)$, $i = 1, \ldots, 20$
- Y_i, $i = 1, \ldots, 10$ observed
- Y_i, $i = 11, \ldots, 20$ missing.
- Assume improper uniform prior distribution for θ.
- Posterior distribution for θ given the observed data is $N(\bar{y}_{obs}, 1/10)$ where $\bar{y}_{obs} = \frac{1}{10} \sum_{i=1}^{10} y_i$.
- $Q = \frac{1}{20} \sum_{i=1}^{20} Y_i$ and $U = \frac{1}{20}$.

Toy example of sampling from the posterior distribution

Repeat m times.
Draw $\dot{\theta} \sim N(\bar{y}_{obs}, 1/10)$.
Draw $\dot{y}_{mis} = \{y_i,\ i = 11, \ldots, 20\}$ from $N(\dot{\theta}, 1)$ to fill in missing data.

Obtain $\hat{Q} = \hat{Q}(y_{obs}, \dot{y}_{mis}) = \frac{1}{20} \sum_{i=1}^{20} y_i$ with $U = \frac{1}{20}$.

Then estimate θ with $\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$ and

$Var\,(\bar{Q})$ with $T = \bar{U} + \left(1 + \frac{1}{m}\right) \sum_{i=1}^{m} \left(\hat{Q}_i - \bar{Q}\right)^2$ where

$\bar{U} = \frac{1}{20}$.

Frequentist viewpoint

From a frequentist viewpoint we have done two things wrong:

1. Instead of imputing from the "correct" distribution $f_{\theta_0}(y_{mis}|y_{obs})$, we simulated from the predictive distribution, i.e. we generated $\tilde{\theta}$ from the posterior distribution and then imputed from the distribution $f_{\tilde{\theta}}(y_{mis}|y_{obs})$.
2. We are using the "complete data" variances U_i instead of the observed data formula, but then inflating these with the "combining" rule to get T.

Frequentist viewpoint

Do two wrongs really make a right???

- Rubin shows that under a number of (rather strong) conditions, the coverage of the interval $\bar{Q} \pm a\sqrt{T}$ equals or exceeds its nominal value.
- To this is added substantial positive testimony from Rubin and colleagues on less tractable parametric problems.

Rubin knows best? When is "proper" imputation improper?

- Suppose $Q = Q(y_{obs}, y_{mis})$ is the statistic we would have used to estimate θ, had the complete data all been available.
- We use $\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i(y_{obs}, \dot{y}_{mis}^{(i)})$ instead where $\dot{y}_{mis}^{(i)}$ are imputed.
- $\bar{Q}$ estimates $E(\hat{Q}_i | y_{obs})$.
- $\bar{Q}$ is not the ML estimate (close asymptotically). Rubin's confidence intervals are often wider than likelihood intervals.

"Proper" imputation

Definition

The imputation must be such that (roughly)

1. The imputation is consistent with the complete data model: i.e. $E[\hat{Q}_i(y_{obs}, \dot{y}_{mis}^{(i)})|y_{obs}] = E[Q(y_{obs}, y_{mis})|y_{obs}]$.

2. $\bar{U}$ estimates the complete-data variance, that is $U = U(y_{obs}, y_{mis})$ satisfies $Var(Q) \approx U$ and $E[U(y_{obs}, \dot{y}_{mis})|y_{obs}] = E[U(y_{obs}, y_{mis})|y_{obs}]$.

3. B estimates $Var[Q - E(Q|y_{obs})] = Var[\hat{Q}_i - E(\hat{Q}_i|y_{obs})]$.

These conditions are not easy to verify in practice.

Imputation estimation schemes behaving badly

Many have observed (e.g. Nielsen, 2003) that some degree of consistency and efficiency is required between the estimation method and the imputation method. The requirements for a proper imputation are not sufficient to ensure stated coverage of the confidence interval.

Meng adds the condition of self-efficiency:

Definition

(Meng, 1994): Let Y be a data set, and let Y_0 be a subset of Y created by a selection mechanism. A statistical estimation procedure $\hat{\theta}(Y)$ for 0 is self-efficient (with respect to the selection mechanism) if *the estimator applied to the complete data set* $\hat{\theta}(Y)$ *is better in terms of mean squared error than when applied to a portion thereof*, i.e. $\hat{\theta}(Y_0)$.

Multiple Imputation: Rules and pitfalls

1. The imputation model should be *at least as general as that of the analyst*. Any variable which may effect the missingness or any variable which might be used in the complete data model should be included in the imputation. This is to ensure that the estimating function is still unbiased. The imputation model should be consistent with (*congenial* with) the analyst's model.
2. Failure to include variables in the imputation model that are correlated with those in the analyst's model and which help to predict missingness may lead to bias.
3. The MAR assumption may not hold.
4. Convergence problems may occur especially when the model of interest contains non-linear relationships or interactions are not included in the imputation model.

Question www.socrative.com (calgary)

We wish to impute missing Y values in a dataset. In modelling the missing indicator R, we should

A. Include only those variables that we know effect whether or not data are missing.

B. Include only those variables that may be used in a subsequent analysis.

C. Include all variables that may be used in a subsequent analysis and all variables that might effect whether or not data are missing.

D. Introduce an indicator random variable as covariate in our regression for observations that are missing.

E. Use independent Bernoulli(p) trials.

Example: Bayesian imputation and the Normal linear model

Suppose the observed data are

$$\begin{bmatrix} y_1 & X_1 \\ \vdots & \vdots \\ y_n & X_n \\ ? & X_{n+1} \\ \vdots & \vdots \\ ? & X_N \end{bmatrix} = \begin{bmatrix} y_{obs} & X_{obs} \\ y_{mis} & X_{mis} \end{bmatrix}$$

There are ONLY missing values for the response y.

X_{obs} ($n \times q$ matrix) and X_{mis} have NO missing values.

Bayesian imputation using the Normal linear model cont'd

Assume $Y_i \backsim N\left(X_i\beta, \sigma^2\right)$, $i = 1, \ldots, N$ so the parameters are β ($q \times 1$ vector) and σ^2.

- Assume improper prior distributions for β and $\ln\left(\sigma^2\right)$.
- Let $\hat{\beta}$ and $\hat{\sigma}^2$ be the usual linear regression estimates of β and σ^2 based on the n complete cases.
- The posterior distribution for β given σ^2 is $MVN\left(\hat{\beta}, \sigma^2 V_{obs}\right)$ where $V_{obs} = \left(X_{obs}^T X_{obs}\right)^{-1}$.
- The posterior distribution for σ^2 is $\hat{\sigma}^2\left(n - q\right)$ (a constant) divided by a $\chi^2\left(n - q\right)$ random variable.

Bayesian imputation using the Normal linear model cont'd

To impute the missing y's:

1. Draw w from a $\chi^2(n-q)$ distribution. Then $\dot{\sigma}^2 = \hat{\sigma}^2 / w$ is a draw from the posterior distribution of σ^2.
2. Draw z_1, a random sample of size q, from the $N(0,1)$ distribution. Then $\dot{\beta} = \hat{\beta} + \dot{\sigma} z_1 V_{obs}^{1/2}$ is a draw from the posterior distribution of β.
3. Draw z_2, a random sample of size $N-n$, from the $N(0,1)$ distribution. Then $\dot{y} = X_{mis}\dot{\beta} + z_2\dot{\sigma}$ are the imputed y values.
4. Repeat steps $1-3$, m times.

Brainweight data: MICE and imputations (in R)

```
library("mice", lib.loc="~/R/win-library/2.15")
brainweight_mis3 <- read.csv("..../brainweight_mis3.csv")
imp3<-
mice(brainweight_mis3,method="norm",m=5,maxit=1,seed=1)
fit3<- with(imp3,lm(lweight~lsize+fem))
round(summary(pool(fit3)),3)
```

	est	se	t	df	Pr(>|t|)	nmis
(Intercept)	1.577	0.531	2.971	14.979	0.010	NA
lsize	0.679	0.064	10.592	15.034	0	130
fem	-0.003	0.013	-0.250	16.219	0.805	0

Brainweight data: Results using LD

> fit4 <- with(brainweight_mis3,lm(formula = lweight ~lsize +fem))

> summary(fit4)

Call: lm(formula = lweight ~lsize + fem)

Coefficients:

	est	se	t	Pr(>|t|)
(Intercept)	1.27401	0.56785	2.244	0.027
lsize	0.71763	0.06862	10.458	<2e-16
fem	0.00932	0.01383	0.674	0.502

Residual standard error: 0.05815 on 104 degrees of freedom
(130 observations deleted due to missingness)
Multiple R-squared: 0.5716, Adjusted R-squared: 0.5633
F-statistic: 69.37 on 2 and 104 DF, p-value: < 2.2e-16

Brainweight data: MICE versus LD

	True	est (MICE)	se	est (LD)	se
(Intercept)	1.284	1.577	0.531	1.27401	0.56785
lsize	0.717	0.679	0.064	0.71763	0.06862
fem	-0.014	-0.003	0.013	0.00932	0.01383

What if your posterior is hard to find?

If the missing data pattern is complex then the posterior given the observed data, $E[f(\theta|Y)|Y_{obs}]$, may be difficult to determine. If the posterior is easy to obtain for completed data sets, then the following two steps can be used:

1. Impute the missing values.
2. Generate θ from the posterior given the completed data.

When these two steps are repeated, the result is a Markov Chain. This algorithm is called "Data Augmentation" or Gibb's Sampling. More on this in last session.

The do's of multiple imputation (van Buuren)

1. Find out the reasons for the missing data; Include factors that govern the missingness in the imputation model;
2. Include the outcome variable(s) in the imputation model;
3. Impute categorical response data by techniques for categorical data;
4. Impute by proper imputation methods;
5. Inspect the imputed data; Evaluate whether the imputed data could have been real data;
6. Describe potential departures from MAR; Specify simple MNAR models for sensitivity analysis;

The don't's of multiple imputation (van Buuren)

Do not:

1. Use multiple imputation if simpler methods are valid;
2. Impute blindly;
3. Put too much faith in the defaults for multiple imputation software;
4. Create imputations using a model that is more restrictive than needed;
5. Uncritically accept imputations that are very different from the observed data;