Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model

International Conference of the ERCIM WG on Computational and Methodological Statistics

Jonathan Bartlett
www.missingdata.org.uk

London School of Hygiene and Tropical Medicine

16th December 2013

Acknowledgements

This is work with

- Shaun Seaman and Ian White (MRC Biostatistics Unit), supported by MRC grant (MC_US_A030_0015) and unit programme U105260558
- James Carpenter (LSHTM), supported by ESRC Fellowship RES-063-27-0257

Support for myself from ESRC Follow-On Funding scheme RES-189-25-0103, MRC grant G0900724, MRC fellowship MR/K02180X/1.

Outline

Outline

The setting

- Suppose we have an outcome of interest Y, partially observed variables $X_1, X_2, .., X_p$, and fully observed covariates Z.
- We specify a substantive model (SM) for $f(Y|X_1, .., X_p, Z, \psi)$, with parameters ψ.
- e.g. linear regression of Y, with covariate vector some function of $X_1, .., X_p$ and Z.
- e.g. covariates include $X_1 \times X_2$, or X_1^2, or X_1/X_2^2...
- We assume **throughout** that the SM is correctly specified.
- The variables $X_1, .., X_p$ have missing values, and we will assume the missing at random assumption holds.

Full conditional specification (FCS) multiple imputation

- Multiple imputation by full conditional specification (FCS) (sometimes called chained equations) has become very popular in recent years.
- FCS involves specifying univariate models for each partially observed variable, conditional on all other variables: $f(X_j|X_{-j}, Z, Y, \theta_j)$, $j = 1, .., p$.
- Missing values are imputed in X_j, conditional on observed values and most recent imputation of X_{-j} and Z, Y.
- We then cycle through each of the partially observed variables, imputing from each univariate model, in a Gibbs sampling approach.
- Since each univariate model can be of a different type, FCS is particularly appealing for datasets with mixtures of continuous and categorical variables.

Full conditional specification (FCS) multiple imputation

For imputation $m = 1, .., M$:

1. Initially impute missing values in X using some ad-hoc approach.
2. For iteration $t = 1, .., T$:
   - 2.1 Impute from $f(X_1|X_{-1}, Z, Y, \theta_1)$:
     - Fit model $f(X_1|X_{-1}, Z, Y, \theta_1)$ using subjects for whom X_1 was observed.
     - Draw $\theta_1^{(t)}$ from posterior for θ_1 corresponding to this fit.
     - Impute missing values in X_1 (once) from $f(X_1|X_{-1}, Z, Y, \theta_1^{(t)})$.
   - 2.2 Impute from $f(X_2|X_{-2}, Z, Y, \theta_2)$
   - 2.3 ...
   - 2.4 Impute from $f(X_p|X_{-p}, Z, Y, \theta_p)$
3. Current imputed values of missing values used to form mth imputed dataset.

Existing imputation approaches

- If the SM contains non-linear terms, interactions, or is non-linear (e.g. Cox), using FCS for covariates becomes tricky.
- i.e. difficult to directly specify $f(X_j|X_{-j}, Z, Y, \theta_j)$ from standard models families which are compatible with $f(Y|X_j, X_{-j}, Z, \psi)$
- As described in the preceding talk, existing approaches (at least those which are available to researchers in software) in general lead to biased estimates and invalid inferences.

Compatibility

- Loosely speaking, an imputation model (IM) $f(X_j|X_{-j}, Z, Y, \omega)$ is said to be compatible with the SM $f(Y|X_j, X_{-j}, Z, \psi)$ if there exists a joint model

$$f(Y, X_j|X_{-j}, Z, \theta)$$

  which has conditionals which match the IM and SM.
- e.g. suppose the SM is $Y|X \sim N(\psi_0 + \psi_1 X + \psi_2 X^2, \sigma^2_\psi)$.
- Suppose the IM is $X|Y \sim N(\omega_0 + \omega_1 Y, \sigma^2_\omega)$.
- Then the SM and IM are incompatible.

The implications of incompatibility

- Unless the IM, or a restricted version of it, is compatible with the SM, incompatibility implies the IM is mis-specified (assuming of course the SM is correct).
- When the SM contains non-linear terms or interactions, common choices of IMs for covariates are incompatible, and are hence mis-specified.
- It is therefore desirable to use an IM which is compatible with the SM.
- Note that compatibility is necessary but not sufficient for the IM to be correctly specified (remembering we are assuming the SM is always correct).

Ensuring compatibility

- The natural way to ensure the IM is compatible with the SM is to specify a model $f(X_1, .., X_p|Z, \phi)$ and impute from

$$f(X|Z, Y, \psi, \phi) = \frac{f(Y, X|Z, \psi, \phi)}{f(Y|Z, \psi, \phi)} \propto \quad f(Y|X, Z, \psi)f(X|Z, \phi)$$

- This depends on specifying a joint model $f(X_1, .., X_p|Z, \phi)$.
- In practice specifying such joint models is challenging - this is partly why FCS is so popular.

Outline

Substantive model compatible FCS

- We propose a modification of FCS, which ensures each univariate IM is compatible with the assumed SM.
- For each $j = 1, .., p$, we specify a model for $f(X_j|X_{-j}, Z, \phi_j)$ and then impute from the distribution proportional to

$$f(Y|X_j, X_{-j}, Z, \psi)f(X_j|X_{-j}, Z, \phi_j)$$

- The first density in this product is just the SM $f(Y|X, Z, \psi)$.

Drawing imputations

- The implied imputation model(s) $f(X_j|X_{-j}, Z, Y, \phi_j, \psi)$ usually do not belong to standard model families.
- We appeal to the Monte-Carlo method of rejection sampling to generate draws.
- Rejection sampling involves drawing from an easy-to-sample (candidate) distribution until a particular criterion/bound is satisfied.
- Deriving this bound is relatively easy if we use our model for $f(X_j|X_{-j}, Z)$ as the candidate distribution.

The SMC-FCS algorithm

- Substantive model compatible FCS (SMC-FCS) approach modifies standard FCS as follows.
- To impute X_j, we (assuming independence of priors for ψ and ϕ_j)
  1. Draw $\psi^{(t)}$ from the posterior for ψ conditional on observed data and current imputations.
  2. Draw $\phi_j^{(t)}$ from the posterior for ϕ_j conditional on observed data and current imputations.
  3. Impute X_j from density proportional to $f(Y|X_j, X_{-j}, Z, \psi^{(t)})f(X_j|X_{-j}, Z, \phi_j^{(t)})$, using rejection sampling.
- More details in [1].

Statistical properties

- With only a single covariate partially observed, SMC-FCS is equivalent to traditional 'joint model' MI, and thus inherits the latter's statistical properties.
- With multiple partially observed covariates, under certain conditions regarding compatibility between the covariate models $f(X_j|X_{-j}, Z)$ and priors, SMC-FCS is equivalent to 'joint model MI'.
- As with standard FCS MI, it is possible to specify models $f(X_j|X_{-j}, Z)$ that are mutually incompatible [2, 3].

Outline

Simulation study

Data for $n = 1,000$ subjects were simulated according to:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon,$$

with $\epsilon \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$ and σ_ϵ^2 chosen to give $R^2 = 0.5$.

X_1 and X_2 were generated as (correlated):

- Bivariate normal
- X_1 Bernoulli, $X_2|X_1$ normal with constant variance

Values of X_1 and X_2 were each made MAR with probability of observation expit$(\alpha_0 + \alpha_1 Y)$ where $\alpha_1 = -1/\text{SD}(Y)$ and α_0 such that 30% of values were missing.

Estimation methods

The parameters of the SM were estimated using:

- Passive imputation (assuming $X_j|Y, X_{-j}$ is normal/logistic, with interaction of Y and X_{-j})
- Just another variable (JAV) (assuming (X_1, X_2, X_1X_2, Y) is multivariate normal)
- SMC-FCS (assuming $X_j|X_{-j}$ normal or logistic)

10 imputations were used for each method.

Results

Mean (empirical SD) of estimates of $\beta_1 = 1$ and $\beta_3 = 1$ based on 1,000 simulations.

X_1, X_2 distribution		Passive	JAV	SMC-FCS
X_1, X_2 bivariate normal	$\beta_1 = 1$	1.63 (0.37)	1.31 (0.60)	1.03 (0.46)
	$\beta_3 = 1$	0.64 (0.12)	0.96 (0.30)	0.97 (0.19)
X_1 Bernoulli $X_2 \mid X_1$ normal	$\beta_1 = 1$	1.11 (0.21)	1.14 (0.22)	1.00 (0.22)
	$\beta_3 = 1$	0.78 (0.15)	0.97 (0.22)	0.98 (0.17)

CI coverage (not shown here) was poor for passive and JAV, but was close to 95% for SMC-FCS

Outline

Conclusions - 1

- We think SMC-FCS is an attractive approach for imputing covariates, particulary when the SM contains non-linear/interaction terms.
- Analogous to standard FCS MI, one should be wary of the possibility of incompatibility between the models $f(X_j|X_{-j}, Z)$.
- To some, the requirement to specify the SM when imputing is a drawback.
- We argue one should always bear in mind the SM when imputing.
- In practice, one could impute assuming a general SM, and then fit nested SMs to the imputed data.

Conclusions - 2

- SMC-FCS may be useful in allowing for skewed distributions while retaining desired/assumed dependence between outcome and covariate.
- Also useful in situations when SM depends on a particular function of variables, e.g.

  `BMI=weight/height^2`
- Stata command `smcfcs` can be downloaded from www.missingdata.org.uk.
- Preprints of methods paper available on arXiv [1] and Stata journal paper (under review) at www.missingdata.org.uk

References I

[1] J. W. Bartlett, S. R. Seaman, I. R. White, and J. R. Carpenter.

Multiple imputation of covariates by fully conditional specification: accommodating the substantive model.

arXiv:1210.6799 [stat.ME], 2012.

[2] R A Hughes, I R White, S Seaman, J Carpenter, K Tilling, and J A C Sterne.

Joint modelling rationale for chained equations imputation.

Under review.

[3] J Liu, A Gelman, J Hill, Y Su, and J Kropko.

On the stationary distribution of iterative imputations.

Biometrika, epub:1–19, 2013.