Multiple imputation of covariates in the presence of interactions and nonlinearities

2013 UK Stata Users Group meeting

Jonathan Bartlett
www.missingdata.org.uk

London School of Hygiene and Tropical Medicine

12th September 2013

Acknowledgements

Stata program `smcfcs`

- Tim Morris (MRC Clinical Trials Unit), supported by MRC PhD studentship A731-5QL14

Methodological development

- 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 and MRC grant G0900724

Outline

Imputing covariates and compatibility

Substantive model compatible FCS

The `smcfcs` command

Simulations

Conclusions

Outline

The setting

- Suppose we have an outcome of interest Y, partially observed covariates $X_1, X_2, .., X_p$, 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 $X1X2$, or $X1^2$, or $X1/X2^2$...
- The covariates $X_1, .., X_p$ have missing values.

Full conditional specification (FCS)

- Multiple imputation by full conditional specification (FCS) has become very popular in recent years.
- FCS involves specifying univariate models for each partially observed variable, conditional on all other variables: $f(X_j|X_{-j}, Z, Y)$.
- 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.
- Since each univariate model can be of a different type, FCS is particularly appealing for datasets with mixtures of continuous and categorical variables.

Multiple imputation of covariates

- If the SM contains non-linear terms, interactions, or is non-linear (e.g. Cox), MI for covariates becomes tricky.
- One option is to use a standard imputation model (IM) choice followed by passive imputation of higher order terms.
- Another is to impute each higher order term as if it were just another variable (JAV) [1].
- As shown by Seaman *et al* [2], both in general lead to biased estimates and inferences.

Compatibility

- Loosely speaking, an 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 does not ensure the IM is correctly specified, but merely that it does not conflict with the SM.

Outline

Imputing covariates and compatibility

Substantive model compatible FCS

The `smcfcs` command

Simulations

Conclusions

Substantive model compatible FCS

- We propose a modification of FCS, which ensures each univariate IM is compatible with the assumed SM.
- We must impute from a model for $f(X_j|X_{-j}, Z, Y)$.
- This can be expressed as

$$\frac{f(Y|X_j, X_{-j}, Z)f(X_j|X_{-j}, Z)}{\int f(Y|X_j^*, X_{-j}, Z)f(X_j^*|X_{-j}, Z)dX_j^*}.$$

- The SM is a model for $f(Y|X_j, X_{-j}, Z)$.
- We can thus specify an IM for X_j which is compatible with the SM by additionally specifying a model for $f(X_j|X_{-j}, Z)$.

Drawing imputations

- Having specified a model for $f(X_j|X_{-j}, Z)$, the implied imputation model $f(X_j|X_{-j}, Z, Y)$ will in general not belong to a standard distributional family.
- 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.

Statistical properties

- With only a single covariate partially observed, the algorithm 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.
- In this case it is not clear which (if any) joint distribution the algorithm will converge to.

Outline

Imputing covariates and compatibility

Substantive model compatible FCS

The `smcfcs` command

Simulations

Conclusions

The `smcfcs` command

- `smcfcs` implements the SMC-FCS approach.
- Linear, logistic and Cox SMs are currently supported.
- `regress`, `logistic`, `ologit`, `mlogit`, `poisson`, `nbreg` covariate imputation models are supported.
- The SM can contain essentially any function of the variables, e.g. squares, cubes, interactions, logarithms of variables, etc etc.

Performance issues

- `smcfcs` is slower than standard chained/FCS imputation, due to the rejection sampling.
- This is mitigated somewhat by using Mata code for the sampling.
- e.g. I have used it with a dataset of ∼10,000 individuals with a complex Cox SM, with missingness in many covariates.
- 10 imputations can be generated in ∼30 mins.

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 $\text{expit}(\alpha_0 + \alpha_1 Y)$ 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)
- `smcfcs` (assuming $X_j|X_{-j}$ normal or logistic)

10 imputations were used for each method.

smcfcs syntax for the example

smcfcs, ctsmiss(x1 x2) smcmd("reg") smout(y) smcov(x1 x2 x1x2) passive(x1x2=x1*x2)m(10)

smcfcs, binmiss(x1) ctsmiss(x2) smcmd("reg") smout(y) smcov(x1 x2 x1x2) passive(x1x2=x1*x2) m(10)

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	$\beta_1 = 1$	1.61 (0.37)	1.36 (0.60)	1.02 (0.45)
normal	$\beta_3 = 1$	0.79 (0.24)	0.93 (0.30)	0.99 (0.19)
X_1 Bernoulli	$\beta_1 = 1$	1.11 (0.21)	1.15 (0.22)	1.00 (0.22)
$X_2 \| X_1$ normal	$\beta_3 = 1$	0.79 (0.14)	0.97 (0.22)	0.98 (0.17)

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.
- But perhaps one should always bear in mind the SM when imputing. What is a good IM for one SM may be a poor IM for another SM.
- In practice, one could impute assuming a general SM, and then fit nested SMs to the imputed data.

Conclusions - 2

- May also be useful to allow for tricky distributions. e.g. suppose X is skewed, but $\log(X)$ is approximately normal.
- `smcfcs` permits imputation of $\log(X)$ using normal linear regression, but SM can still contain X (or some other transformation) in the linear predictor.
- Also useful in situations when SM depends on a particular function of variables, e.g.

  `BMI=weight/height^2`
- `smcfcs` can be downloaded from www.missingdata.org.uk, and will be made available on SSC soon.
- For preprints of methods and Stata journal papers (both under review), see www.missingdata.org.uk

References I

[1] P T von Hippel.
How to impute interactions, squares, and other transformed variables.
Sociological Methodology, 39:265–291, 2009.

[2] S. R. Seaman, J. W. Bartlett, and I. R. White.
Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods.
BMC Medical Research Methodology, 12:46, 2012.

[3] 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.