The Stata Journal (*yyyy*) **Authors' final version**, Number *ii*, pp. 1–20

Multiple imputation of covariates by substantive model compatible fully conditional specification

Jonathan W. Bartlett
Department of Medical Statistics
London School of Hygiene & Tropical Medicine
London, UK
jonathan.bartlett@lshtm.ac.uk
Tim P. Morris
MRC Clinical Trials Unit at UCL
Institute of Clinical Trials and Methodology
London, UK
tim.morris@ucl.ac.uk

Abstract. Multiple imputation (MI) is a practical, principled approach to handling missing data. When used to impute missing values in covariates of regression models, imputation models may be mis-specified if they are not compatible with the substantive model of interest for the outcome. In this article we introduce the `smcfcs` command, which imputes covariates by substantive model compatible fully conditional specification (SMC–FCS). This modifies the popular FCS or chained equations approach to MI by imputing each covariate compatibly with a user-specified substantive model. The `smcfcs` command is compared to standard FCS imputation using `mi impute chained` in a simulation study and illustrative analysis of data from a study investigating time to tumour recurrence in breast cancer.

Keywords: st0001, multiple imputation, congenial, compatible, interactions, nonlinearities

1 Introduction

Missing data are a common issue in empirical research, reducing statistical power and potentially causing bias in parameter estimates. The method of multiple imputation (MI) has become one of the most popular approaches for handling missing data (van Buuren (2007)). For each missing value, MI creates a number of plausible imputations, based on a model for the conditional distribution of the variable being imputed given other variables, thus creating a number of completed, or imputed datasets. Each imputed dataset is then analysed separately and identically, giving estimates of parameters of interest and corresponding standard errors. These are then combined using rules derived by Rubin (1987). Virtually all implementations of MI in software packages assume data are missing at random (MAR), which states that the probability that data are missing is independent of the unobserved values, conditional on the observed values (Rubin (1976)).

 st0001

1.1 Fully conditional specification

As originally conceived, parametric MI involves specification of a joint model for the partially observed variables, conditional on any fully observed variables ('joint model MI'). A popular alternative to joint model MI is the fully conditional specification (FCS) or chained equations approach (White et al. (2011), van Buuren (2007)). FCS MI involves specifying a series of univariate models for the conditional distribution of each partially observed variable given the other variables. This approach permits a great deal of flexibility, since an appropriate regression model can be selected for each variable (e.g. linear regression for continuous variables, logistic regression for binary variables). Consequently, FCS MI is particularly appealing in settings in which a number of variables have missing data, some of which are continuous and some of which are discrete. In Stata the FCS approach was implemented by Royston (2005) as the user-written command `ice`, but as of version 12, has been available as part of official Stata through the `mi impute chained` command.

1.2 Multiple imputation of covariates

In this paper we focus on the setting in which some values are missing in the covariates of a substantive model of interest. Correctly specifying imputation models for covariates can be challenging, particularly when the substantive model relating the outcome to the covariates includes non-linear covariate effects or interactions between covariates. For example, Seaman et al. (2012) showed that for a linear regression substantive model with quadratic effects of a (marginally) normal covariate, imputation models which are implemented in existing MI software are mis-specified, and lead to biased estimates. They drew similar conclusions when the substantive model included an interaction. Even when the MAR assumption holds, mis-specification of the imputation model generally results in biased estimates of the substantive model parameters. In the aforementioned examples the mis-specification can be attributed to the fact that the imputation and substantive models are incompatible (sometimes referred to as uncongenial). Loosely speaking, the imputation and substantive models are compatible if there exists a joint model for covariate and outcome whose conditional distributions are equal to those given by the imputation and substantive models. While compatibility between the imputation and substantive models does not guarantee the former is correctly specified, provided the substantive model is correctly specified, incompatibility between the two (generally) implies the imputation model is mis-specified. This suggests that it is desirable that covariates are imputed using imputation models which are compatible with the substantive model.

1.3 Substantive model compatible FCS MI

Recently Bartlett et al. (2014) proposed substantive model compatible FCS (SMC–FCS). This modifies the FCS or chained equations MI approach by imputing each partially observed covariate using an imputation model which is compatible with a user-specified substantive model. In Section 2 we describe the SMC–FCS method in more detail. In Section 3 we describe the `smcfcs` command and its syntax, and in Section 4 we illustrate its use and compare its performance with standard FCS. In Section 5 we describe the results of a small simulation study comparing `smcfcs` to a standard approach using `mi impute chained`, in terms of bias and computational time. We conclude in Section 6 with some final remarks.

2 Substantive model compatible FCS

2.1 Setup

We consider the setting in which interest lies in fitting a model to a fully observed outcome Y with p partially observed covariates $X = (X_1, .., X_p)$ and q fully observed covariates $Z = (Z_1, .., Z_q)$. Let X^{obs} and X^{mis} denote the observed and missing components of X for a given subject, and let R be the vector of observation indicators whose elements are zero or one depending on whether the corresponding element of X is missing or observed respectively. We assume throughout that the data are missing at random (MAR) Rubin (1976). Here MAR means that $P(R|Y, X, Z) = P(R|Y, X^{\text{obs}}, Z)$. We assume that (Y_i, X_i, Z_i, R_i), $i = 1, .., n$ are independent and identically distributed. Lastly, we let $f(Y|X, Z, \psi)$ denote the 'substantive model', which is indexed by parameter ψ ($\psi \in \Psi$). We assume throughout that this substantive model is correctly specified. That is, there exists $\psi \in \Psi$ such that $f_0(Y|X, Z) = f(Y|X, Z, \psi)$, where $f_0(Y|X, Z)$ denotes the true conditional distribution of Y given X and Z.

2.2 Incompatibility and imputation model mis-specification

Suppose for the moment that there exists only a single partially observed covariate, denoted X. To impute X we must specify an imputation model $f(X|Z, Y, \omega)$, indexed by parameter $\omega \in \Omega$. Following Liu et al. (2013), this imputation model is said to be compatible with the substantive model $f(Y|X, Z, \psi)$, $\psi \in \Psi$, if there exists a joint model $g(Y, X|Z, \theta)$, $\theta \in \Theta$ and surjective maps $t_1 : \Theta \to \Omega$, $t_2 : \Theta \to \Psi$ such that:

1. for $\omega \in \Omega$, and $\theta \in t_1^{-1}(\omega) = \{\theta : t_1(\theta) = \omega\}$,

$$f(X|Z, Y, \omega) = g(X|Z, Y, \theta)$$

2. for $\psi \in \Psi$ and $\theta \in t_2^{-1}(\psi)$,

$$f(Y|X, Z, \psi) = g(Y|X, Z, \theta)$$

Again following Liu et al. (2013), the two models are said to be semi-compatible if they can be made compatible by setting certain parameters in either one or both models to zero. Lastly, if the two are semi-compatible and correctly specified, they are said to be valid semi-compatible. The imputation model is then correctly specified if and only if it is valid semi-compatible with the substantive model (Bartlett et al. (2014)).

Except in cases where the imputation and substantive models can be made compatible by restricting the parameter space Ω of the imputation model, incompatibility between the two implies that the imputation model is mis-specified, assuming the substantive model is correctly specified. This is because incompatibility means there exist no joint models which have the imputation and substantive models as its conditionals.

To illustrate this, suppose the substantive model is $Y|X \sim N(\psi_0+\psi_1 X+\psi_2 X^2, \sigma^2_\psi)$ and the imputation model is $X|Y \sim N(\omega_0 + \omega_1 Y, \sigma^2_\omega)$. These models are incompatible since there exists no joint model with conditionals corresponding to the substantive and imputation models. They are semi-compatible, by setting $\psi_2 = 0$, but unless $\psi_2 = 0$ in truth, the imputation model will not be valid semi-compatible with the substantive model, and so will necessarily be mis-specified. Figure 1 shows a plot of (Y, X) pairs simulated under this substantive model with $X \sim N(1, 1)$, and $Y|X \sim N(X+3X^2, 1.5^2)$, in which the missing X value has been imputed assuming the aforementioned linear imputation model. By virtue of the imputation model (wrongly) assuming linearity between Y and X, it is clear that the estimates of the quadratic substantive model will be biased. This example was investigated in detail through simulation by von Hippel (2009) and Seaman et al. (2012).

Now assume the substantive model is $Y|X \sim (\psi_0 + \psi_1 X, \sigma^2_\psi)$ and the imputation model is $X|Y \sim N(\omega_0 + \omega_1 Y + \omega_2 Y^2, \sigma^2_\omega)$, with each of the regression coefficients lying in $(-\infty, +\infty)$. These two models are again incompatible. However, they can be made compatible (and are hence semi-compatible) by restricting the parameter space of the imputation model by setting $\omega_2 = 0$. Here incompatibility does not imply mis-specification.

As a final example, suppose the substantive model $Y|X \sim (\psi_0 + \psi_1 X, \sigma^2_\psi)$ and the imputation model is $X|Y \sim N(\omega_0 + \omega_1 Y, \sigma^2_\omega)$. These models are compatible, with the joint model being the bivariate normal. We emphasize that compatibility does not guarantee that the imputation model is correctly specified.

Even when the substantive model only contains linear covariate effects without interactions, incompatibility may arise with default imputation models if the substantive model is non-linear. For example, for an exponential survival substantive model, Bartlett et al. (2014) describe how the recommended imputation model for continuous partially observed covariates is incompatible with the exponential model.

In conclusion, except in cases where the imputation and substantive models can be made compatible by restricting the parameter space Ω of the imputation model (i.e. a simpler model nested within the imputation model is compatible with the substantive model), incompatibility between the two implies the imputation model is mis-specified (assuming correct specification of the substantive model). Consequently, when choosing the covariate imputation model $f(X|Z, Y, \omega)$ we should (at least) ensure that it is either compatible with the substantive model, or a restriction of it is compatible with the substantive model.

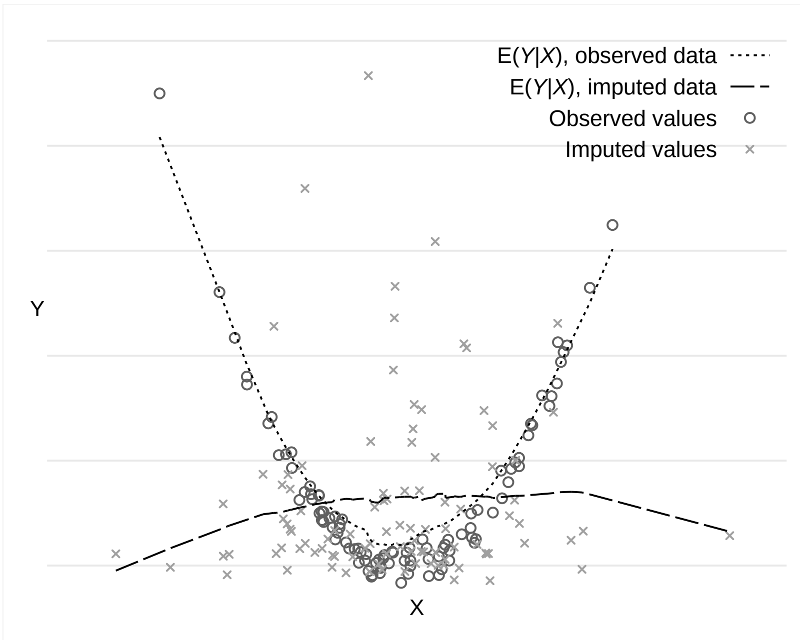

Figure 1: Plot of simulated (Y, X) data, in which $X \sim N(1, 1)$ and $Y|X \sim N(X + 3X^2, 1.5^2)$. Y is fully observed, whereas X is partially observed. 100 (Y, X) pairs in which X was observed are shown as circles. 100 (Y, X) pairs where X was imputed assuming $X|Y \sim N(\omega_0 + \omega_1 Y, \sigma^2_\omega)$ are shown by crosses. Conditional expectations were estimated non-parametrically using the `lowess` command

2.3 Substantive model compatible FCS

We now return to the setting of a vector of multiple partially observed covariates, $X = (X_1, .., X_p)$. To apply standard FCS MI (see van Buuren (2007) for further background on standard FCS MI) in the missing covariates setting, for each partially observed covariate X_j, $j = 1, .., p$, we specify a model for $f(X_j|X_{-j}, Z, Y)$, where $X_{-j} = (X_1, .., X_{j-1}, X_{j+1}, .., X_p)$. For the reasons described in the previous sub-section, common choices for this imputation model may be incompatible with the substantive model $f(Y|X, Z, \psi)$, implying mis-specification.

To motivate the SMC–FCS algorithm, note that the conditional distribution $f(X_j|X_{-j}, Z, Y)$ can be expressed as

$$f(X_j|X_{-j}, Z, Y) = \frac{f(Y, X_j|X_{-j}, Z)}{f(Y|X_{-j}, Z)} = \frac{f(Y|X_j, X_{-j}, Z)f(X_j|X_{-j}, Z)}{f(Y|X_{-j}, Z)} \propto f(Y|X, Z)f(X_j|X_{-j}, Z).$$

In substantive model compatible FCS (SMC–FCS), for X_j we specify a model $f(X_j|X_{-j}, Z, \phi_j)$, where ϕ_j is a vector of model parameters, and impute using the density proportional to

$$f(Y|X, Z, \psi)f(X_j|X_{-j}, Z, \phi_j). \tag{1}$$

For any given j, this imputation model will automatically be compatible with the substantive model $f(Y|X, Z, \psi)$. The model $f(X_j|X_{-j}, Z, \phi_j)$ can be chosen in the same way as models are selected for the standard FCS algorithm. For example, if X_j is binary, a default choice would be to use a logistic regression model. For discrete X_j which have a finite sample space (e.g. binary and categorical variables), samples can be drawn directly from the distribution proportional to equation 1. More generally, Bartlett et al. (2014) show that provided one can easily draw samples from $f(X_j|X_{-j}, Z, \phi_j)$, the Monte-Carlo method of rejection sampling can be used to draw samples from the imputation distribution when the substantive model is either a normal linear regression, a regression model for a discrete outcome Y (thereby including logistic and Poisson regression) or a proportional hazards model for a possibly censored time-to-event outcome. Rejection sampling involves repeatedly drawing from a candidate distribution (here $f(X_j|X_{-j}, Z, \phi_j)$) until a certain criterion is satisfied, which is therefore computationally intensive. To ensure reasonable run times the `smcfcs` command therefore uses Mata to perform rejection sampling.

The SMC–FCS algorithm initialises by imputing missing values in each variable by randomly observed values from the same variable. It then cycles through the imputation models for each partially observed variable, which here are the variables $X_1, .., X_p$, imputing each missing value. At the end of a suitable number of iterations, the current imputations form the first imputed dataset. The process is then repeated to create as many imputed datasets as desired.

In SMC–FCS the imputation model for X_j depends both on ϕ_j and the substantive model parameter ψ. Bartlett et al. (2014) derive a Gibbs sampler for the joint model (assuming it exists) defined by the substantive model and the models $f(X_j|X_{-j}, Z, \phi_j)$,

$j = 1, .., p$. At the tth iteration the SMC–FCS algorithm imputes missing values in X_j by performing the following draws

$$\psi^{(t,j)} \sim f(\psi)f(y|x_j^{\text{mis}(t-1)}, x_j^{\text{obs}}, x_{-j}^*, z, \psi)$$
$$\phi_j^{(t)} \sim f(\phi_j)f(x_j^{\text{mis}(t-1)}, x_j^{\text{obs}}|x_{-j}^*, z, \phi_j),$$

where $f(\psi)$ and $f(\phi_j)$ denote uninformative priors, y and z denote the (fully) observed values of Y and Z across the n subjects, x_{-j}^* denotes the observed and most recent imputed values of X_{-j} across all n subjects, x_j^{obs} denotes the observed values of X_j, and $x_j^{\text{mis}(t-1)}$ denotes the imputed values of X_j from the preceding iteration. The missing values in X_j are then imputed using rejection sampling from the density defined by equation 1 using $\psi^{(t,j)}$ and $\phi_j^{(t)}$.

Bartlett et al. (2014) give conditions, including that the models $f(X_j|X_{-j}, Z, \phi_j)$, $j = 1, .., p$ are mutually compatible, under which the SMC–FCS imputes from a well defined Bayesian joint model. When this joint model is correctly specified, application of Rubin's rules will result in valid inferences. There are however common model specifications (e.g. a combination of linear and logistic covariate models) for which SMC–FCS is not equivalent to MI from a Bayesian joint model. Bartlett et al. (2014) conjecture that when the models $f(X_j|X_{-j}, Z, \phi_j)$, $j = 1, .., p$ are 'semi-compatible valid', which means that there exists restrictions of these models which makes them mutually compatible, and that these are correctly specified, application of Rubin's rules to imputations generated by SMC–FCS will give consistent point estimates. Simulations by Bartlett et al. (2014) support this, and further suggest that confidence intervals based on Rubin's variance estimator may still perform well even when SMC–FCS is not equivalent to MI from a Bayesian joint model. Lastly, if the models $f(X_j|X_{-j}, Z, \phi_j)$, $j = 1, .., p$ are not compatible (and cannot be made so by restrictions of their parameter spaces), we cannot generally expect consistent point estimates to be obtained.

Bartlett et al. (2014) reported simulation results for a linear regression substantive model with quadratic covariate effects, a linear regression model with an interaction effect, and a Cox proportional hazards substantive model. Overall, their results suggest that SMC–FCS is an attractive approach for imputing missing values of covariates for substantive models which include non-linear covariate effects, interactions, or are themselves non-linear (e.g. Cox's proportional hazards model).

3 The `smcfcs` command

3.1 Syntax

`smcfcs` *smcmd smdepvar smindepvars*, `regress(`*varlist*`)` `logit(`*varlist*`)`
`poisson(`*varlist*`)` `nbreg(`*varlist*`)` `mlogit(`*varlist*`)` `ologit(`*varlist*`)`
`iterations(`*#*`)` `m(`*#*`)` `rjlimit(`*#*`)` `passive(`*string*`)` `eq(`*string*`)` `rseed(`*string*`)`
`chainonly savetrace(filename) noisily by(`*varlist*`)]`

3.2 Options

`regress(`*varlist*`)` specifies the names of the partially observed continuous variables (if any), which are to be imputed by normal linear regression.

`logit(`*varlist*`)` specifies the names of the partially observed binary variables (if any), which are to be imputed by logistic regression.

`poisson(`*varlist*`)` specifies the names of the partially observed Poisson variables (if any), which are to be imputed.

`nbreg(`*varlist*`)` specifies the names of the partially observed negative binomial variables (if any), which are to be imputed.

`mlogit(`*varlist*`)` specifies the names of the partially observed unordered categorical variables (if any), which are to be imputed.

`ologit(`*varlist*`)` specifies the names of the partially observed ordered categorcal variables (if any), which are to be imputed.

`iterations(#)` specifies the number of iterations to perform for each imputation; default is 10.

`m(#)` specifies number of imputations to generate; default is 5.

`rjlimit(#)` `smcfcs` uses rejection sampling to impute missing covariate values for variables which do not have a finite sample space. Rejection sampling involves repeatedly drawing from a distribution until a valid imputation is found. This option specifies the maximum number of attempts that `smcfcs` will make to find a valid draw for imputed values. If valid values have not been found for one or more subjects by the limit the command continues, using the last proposed draw for such subjects. The default limit is 1000.

`passive(`*string*`)` specifies a string of equations to update derived covariates (if any). Each expression within the string must be separated by a |. Derived covariates may appear either in the substantive model, in covariate models, or both.

`eq(`*string*`)` specifies a string of linear predictor sets for partially observed variables. Each expression within the string must be separated by a |. Each expression should be of the form varname: varlist, which specifies that the linear predictor of the covariate model for varname is given by varlist. If an expression is not specified for a given partially observed variable, the default is to impute using a covariate model which includes which includes any fully observed variables in the substantive model and all partially observed variables except the one being imputed.

`rseed(`*string*`)` sets Stata's random number seed to the given value.

`chainonly` perform iterations of SMC–FCS (as specified by iteration option) without creating imputations. Useful in conjunction with `savetrace` to assess convergence.

`savetrace(`*filename*`)` save means and SDs of imputed values from each iteration in filename.dta. Useful for checking convergence of SMC–FCS.

`noisily` runs SMC–FCS noisily. Useful for diagnosing errors.

`by(`*varlist*`)` imputes separately in groups defined by *varlist*.

3.3 Description

`smcfcs` multiply imputes missing values in covariates using the SMC–FCS algorithm. The substantive model is specified immediately following `smcfcs` by *smcmd smdepvar smindepvars*, giving the substantive model command, dependent variable, and independent variables respectively. Currently `smcfcs` supports `regress`, `logistic` and `stcox` substantive models. The independent variables of the substantive model can be fully observed, directly imputed variables, or passively imputed variables (i.e functions of imputed variables and possibly fully observed variables).

Partially observed variables can be imputed using linear, logistic, Poisson, negative binomial, multi-nomial logistic and ordered logistic regression models, which is controlled by passing the variables to the `regress`, `logit`, `poisson`, `nbreg`, `mlogit`, and `ologit` options respectively. By default, each partially observed variable is imputed from a model conditioning on all of the other partially observed variables and any fully observed independent variables in the substantive model. When they serve as predictors, partially observed variables are included by default as linear terms, except for partially observed categorical variables, which are included as factor variables. The `eq` option can be used should one wish to customise the models $f(X_j|X_{-j}, Z, \phi_j)$. Fully observed variables can be included as factor variables if desired by using Stata's i. notation.

If any of the covariates given as *smindepvars* are derived functions of the partially observed variables, the equations defining the covariates must be specified using the passive option. For example, if the substantive model includes `xsq` as a covariate, which is equal to the square of a partially observed variable `x`, we would pass `xsq=x^2` to the passive option. Further examples are given in the help file to `smcfcs`, and also see the illustrative example in Section 4.

Once the desired number of imputations have been generated, `smcfcs` imports the imputations to Stata `mi` format `flong`, and then fits the substantive model to the imputations using Stata's `mi estimate` command. The `mi estimate` command can then be used to fit alternative models for the outcome, although care should be taken to ensure that these are nested within the substantive model specified to generate the imputations.

The command will give a warning if valid draws are not obtained for one or more observations within the limit specified by the `rjlimit` option. If you receive this warning it is advisable to increase the limit until the warning no longer appears.

As with standard FCS MI, one should assess whether a sufficient number of iterations have been used for the algorithm to converge. Convergence can be assessed by using `chainonly` and `savetrace` options, as per Stata's `mi impute chained` command, and plotting the means and SDs of imputed values by iteration. Because, unlike standard FCS, SMC–FCS conditions on the last imputations of X_j when fitting the models

$f(X_j|X_{-j}, Z, \phi_j)$ and $f(Y|X, Z, \psi)$, SMC–FCS may be expected to require more iterations for convergence. However, Bartlett et al. (2014) obtained good performance in simulations with 10 iterations, which is therefore the default used by `smcfcs`.

The `by(`*varlist*`)` option can be used to impute separately in groups defined by the supplied *varlist*. In this case `smcfcs` fits the substantive model and covariate models, and imputes, entirely separately in each group. Following this, the imputations from each group are appended. Note that in this case `smcfcs` does not fit a single substantive model across all the groups - it is up to the user to select and fit an appropriate model using `mi estimate`.

When using standard FCS imputation, it is recommended that the outcome of the substantive model be included as a predictor in the imputation models for the substantive model covariates, to ensure that the covariates are (hopefully correctly) associated with the outcome. For the avoidance of doubt, when using `smcfcs`, the outcome should only be included as the *smdepvar* variable, and should not be included elsewhere in the command call.

4 Illustrative example

We illustrate the use of `smcfcs` using a dataset of 686 patients in Germany with positive node breast cancer, previously analysed by Royston (2004). The original data can be loaded with `webuse brcancer`. Royston (2004) previously developed a substantive Cox proportional hazards model for time to cancer recurrence, including five covariates: age (`age`) with a fractional polynomial transformation with powers −2 and −0.5, tumor grade 2/3 (`gradd1`), number of positive lymph nodes (`nodes`) with the exponential transformation `enodes` = exp(−0.12×`nodes`), progesterone receptors (`pgr`) with a fractional polynomial transformation with power 0.5, and hormonal therapy with tamoxifen (`tam`). The Cox model thus contains the following nonlinear transformations of three covariates:

$$\begin{aligned} \texttt{age\_1} &= (\texttt{age}/10)^{-2} \\ \texttt{age\_2} &= (\texttt{age}/10)^{-0.5} \\ \texttt{enodes} &= \exp(-0.12 \times \texttt{nodes}) \\ \texttt{pgr\_1} &= ((\texttt{pgr}+1)/1000)^{0.5} \end{aligned}$$

In the original dataset the covariates were fully observed in all 686 patients. However, Royston deleted 20% of values completely at random for each independent variable in the analysis model. As a sterner test, here we make 50% missing (completely randomly) in each independent variable, leaving just 25 complete cases (provided as partialdata.dta). For comparison with estimates based on MI, we first present results based on the full data, before data were made missing:

```
. use breastcancerfull, clear
(German breast cancer data)
. fracgen age -2 -0.5
```

```
-> gen double age_1 = X^-2
-> gen double age_2 = X^-0.5
   (where: X = age/10)

. fracgen pgr 0.5
-> gen double pgr_1 = X^0.5
   (where: X = (pgr+1)/1000)

. stcox age_1 age_2 gradd1 enodes pgr_1 tam, nohr nolog

         failure _d:  censrec
   analysis time _t:  rectime

Cox regression -- Breslow method for ties

No. of subjects =          686                   Number of obs   =        686
No. of failures =          299
Time at risk    =  2111.978093
                                                 LR chi2(6)      =     153.11
Log likelihood  =   -1711.6186                   Prob > chi2     =     0.0000
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age_1	43.55382	8.253433	5.28	0.000	27.37738	59.73025
age_2	-17.48136	3.911882	-4.47	0.000	-25.14851	-9.814212
gradd1	.5174351	.2493739	2.07	0.038	.0286713	1.006199
enodes	-1.981213	.2268903	-8.73	0.000	-2.425909	-1.536516
pgr_1	-1.84008	.3508432	-5.24	0.000	-2.52772	-1.15244
tam	-.3944998	.128097	-3.08	0.002	-.6455654	-.1434342

We first applied standard FCS MI to the partially observed dataset, using `mi impute chained` to create 100 imputations. Imputing variables ignoring non-linearities, and then passively imputing the non-linear covariates of the substantive model, is known to lead to biased estimates (Seaman et al. (2012)). Instead, we used the just another variable (JAV) approach for imputing non-linear and interaction terms proposed by von Hippel (2009). To do this, we directly imputed the non-linear terms involved in the substantive model, here given by `age_1`, `age_2`, `enodes` and `pgr_1` using normal linear regressions. Note that this ignores the deterministic relationship between `age_1` and `age_2`. We used logistic regression to impute the two binary variables, `gradd1` and `tam`. Following the advice of White and Royston (2009), we included the event indicator and the marginal Nelson–Aalen cumulative hazard estimate (generated using `sts gen`) as covariates in each imputation model:

```
. use partialdata, clear

. sts gen na = na

. mi set flong

. mi register imputed age_1 age_2 pgr_1 enodes gradd1 tam
(661 m=0 obs. now marked as incomplete)

. mi impute chained (reg) age_1 age_2 pgr_1 enodes (logit) gradd1 tam = na _d,
> add(100) rseed(6934)

Conditional models:
             age_1: regress age_1 age_2 enodes i.gradd1 i.tam pgr_1 na _d
             age_2: regress age_2 age_1 enodes i.gradd1 i.tam pgr_1 na _d
            enodes: regress enodes age_1 age_2 i.gradd1 i.tam pgr_1 na _d
            gradd1: logit gradd1 age_1 age_2 enodes i.tam pgr_1 na _d
               tam: logit tam age_1 age_2 enodes i.gradd1 pgr_1 na _d
```

```
            pgr_1: regress pgr_1 age_1 age_2 enodes i.gradd1 i.tam na _d
Performing chained iterations ...
Multivariate imputation                       Imputations =      100
Chained equations                                   added =      100
Imputed: m=1 through m=100                        updated =        0
Initialization: monotone                       Iterations =     1000
                                                  burn-in =       10
            age_1: linear regression
            age_2: linear regression
            pgr_1: linear regression
           enodes: linear regression
           gradd1: logistic regression
              tam: logistic regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
age_1	360	326	326	686
age_2	360	326	326	686
pgr_1	323	363	363	686
enodes	358	328	328	686
gradd1	350	336	336	686
tam	333	353	353	686

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
. mi estimate: stcox age_1 age_2 gradd1 enodes pgr_1 tam , nohr
Multiple-imputation estimates                   Imputations     =        100
Cox regression: Breslow method for ties         Number of obs   =        686
                                                Average RVI     =     1.2171
                                                Largest FMI     =     0.6489
DF adjustment:   Large sample                   DF:     min     =     237.25
                                                        avg     =     340.43
                                                        max     =     437.48
Model F test:       Equal FMI                   F(   6, 1955.8) =      11.40
Within VCE type:          OIM                   Prob > F        =     0.0000
```

_t	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age_1	33.24523	11.62389	2.86	0.004	10.39961	56.09085
age_2	-12.41139	5.569491	-2.23	0.026	-23.3611	-1.461675
gradd1	.299675	.343087	0.87	0.383	-.375485	.974835
enodes	-1.835314	.336044	-5.46	0.000	-2.496132	-1.174495
pgr_1	-2.287823	.5170807	-4.42	0.000	-3.306479	-1.269168
tam	-.4000215	.1941113	-2.06	0.040	-.7819716	-.0180715

We first note that the standard errors are all larger than those based on full data, as is to be expected with such large proportions of missingness. The coefficients of the two powers of `age` are in the same direction as the full data estimates, but there is the suggestion of attenuation (i.e. bias), with the estimates both being about 25% smaller in magnitude than the full data estimates. The coefficient of `gradd1` is also proportionately much smaller than the full data estimate. The coefficients corresponding to the number of nodes and `tam` are both quite close to their full data estimates, whilst the coefficient of pgr_1 is somewhat larger. Researchers may be uncomfortable using JAV because the normal imputation models used are not well specified. For example, in this data, for those patients with a negative value imputed for the enodes variable, one cannot take logarithms to obtain an imputed value for the original nodes variable, and for some patients for whom the back-transformation can be performed their nodes value is negative. Further, for each subject with age missing there is no single imputation of this variable, since the values imputed into age_1 and age_2 will not be consistent with a particular value of age.

One may argue that when interest lies in fitting a substantive model, we should only be concerned with the validity of inferences for the parameters of this model. Here there was some suggestion that some of the coefficients may be biased, although it is difficult here to distinguish between random variation and systematic bias. More importantly, although JAV can be shown to be unbiased for linear regression models under MCAR, it has been shown to be biased under MAR mechanisms, and also biased for logistic regression substantive models, even under MCAR (Seaman et al. (2012)). As far as we are aware there is no justification for its (even approximate) validity for Cox proportional hazards models.

For these reasons, SMC–FCS is an appealing alternative approach here, since we can impute each variable from an imputation model which is compatible with the assumed Cox proportional hazards model. Since the nodes and pgr variables are both integer valued and positively skewed, we chose to impute them using negative binomial regression. Since all patients have at least one node, we subtracted one from nodes and assumed this followed a negative binomial regression. The distribution of age had little skew, and so we chose a normal linear regression model. Since the nodes and pgr variables are so highly skewed, we deemed it implausible that they had linear effects in the covariate models $f(X_j|X_{-j}, Z, \phi_j)$. We therefore used the eq option to specify that when included as covariates, they should be included as log(nodes) and log(pgr+1) respectively. To do this we generated corresponding variables, and added expressions to the passive option (in addition to those required for the substantive model covariates) so that these were updated appropriately.

```
. use partialdata, clear
. gen nodesminusone = nodes-1
(328 missing values generated)
. gen logpgr = log(pgr+1)
(363 missing values generated)
. gen lognodes = log(nodesminusone+1)
(328 missing values generated)
. smcfcs stcox age_1 age_2 gradd1 enodes pgr_1 tam, reg(age) logit(gradd1 tam)
> nbreg(nodesminusone pgr) passive( age_1 = (age/10)^-2 | age_2 = (age/10)^-.5
> | enodes = exp(-0.12*(nodesminusone+1)) | pgr_1 = ( (pgr+1)/1000)^.5 | logpgr
>  = log(pgr+1) | lognodes = log(nodesminusone+1)) eq(age: gradd1 tam logpgr lo
> gnodes | gradd1: age tam logpgr lognodes | tam: gradd1 age logpgr lognodes |
> nodesminusone: tam gradd1 age logpgr | pgr: lognodes tam gradd1 age) rseed(59
> 13) m(100)
Covariate models:
reg age gradd1 tam logpgr lognodes
```

```
logistic gradd1 age tam logpgr lognodes, coef
logistic tam gradd1 age logpgr lognodes, coef
nbreg nodesminusone tam gradd1 age logpgr
nbreg pgr lognodes tam gradd1 age
Your passive statement(s) say:
age_1 = (age/10)^-2
age_2 = (age/10)^-.5
enodes = exp(-0.12*(nodesminusone+1))
pgr_1 = ( (pgr+1)/1000)^.5
logpgr = log(pgr+1)
lognodes = log(nodesminusone+1)
..............................................................................
> ....................
100 imputations generated
Fitting substantive model to multiple imputations
Multiple-imputation estimates                    Imputations     =        100
Cox regression: Breslow method for ties          Number of obs   =        686
                                                 Average RVI     =     1.2787
                                                 Largest FMI     =     0.6045
DF adjustment:   Large sample                    DF:     min     =     273.50
                                                         avg     =     329.79
                                                         max     =     453.10
Model F test:       Equal FMI                    F(   6, 1872.1) =      12.33
Within VCE type:          OIM                    Prob > F        =     0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age_1 |   37.86482   11.04905     3.43   0.001     16.15108    59.57857
       age_2 |  -13.96674   5.765115    -2.42   0.016    -25.30891   -2.624579
      gradd1 |   .4355209   .3483549     1.25   0.212    -.2502723    1.121314
      enodes |  -1.924758   .3263519    -5.90   0.000    -2.566525   -1.282991
       pgr_1 |  -2.996675    .577313    -5.19   0.000    -4.132892   -1.860458
         tam |  -.3652585    .204516    -1.79   0.075    -.7678841    .0373672
------------------------------------------------------------------------------
```

The command first gives a summary of the covariate models it will use. This shows that `log(nodes)` and `log(pgr+1)` will be used as covariates, rather than their untransformed versions, in the covariate models $f(X_j|X_{-j}, Z, \phi_j)$. The command then summarises the passive expressions which will be used. Next, the SMC–FCS algorithm runs, creating the desired imputations, and finally the substantive model is fitted to each imputation, and the results combined and displayed using `mi estimate`.

Comparing the estimates obtained using SMC–FCS with those from the full data and those using JAV, we see that all of the estimated coefficients from SMC–FCS are closer to those from the full data compared to those using JAV, with the exception of that of `pgr_1` (for which the SMC–FCS is quite a bit larger in magnitude) and `tam` (which is still fairly close to the full data estimate). Unlike the imputations generated from JAV, the distributions of the variables after imputation are similar to their full data distributions, and the values in the variables `age_1` and `age_2` are consistent with the imputed values of age.

5 Simulation study

In this section we present results of a small simulation study comparing the performance of `smcfcs` to standard FCS imputation with `mi impute chained`, in the case of a Cox proportional hazards substantive model. We include results on computational time to highlight the fact that `smcfcs` is more computationally demanding. Datasets were simulated for n subjects with two covariates, X_1 drawn from a Bernoulli distribution with probability 0.5, and $X_2|X_1 \sim N(X_1, 1)$. For each subject, we then simulated a survival time, with hazard function $h(t|X) = 0.002 \exp(\beta_1 X_1 + \beta_2 X_2)$ with $\beta_1 = \beta_2 = 1$. Censoring times were generated from an exponential distribution with hazard 0.002. Values in X_1 and X_2 were made (independently) missing completely at random with probability π in each simulation. We first investigated the impact of sample size, by performing simulations (100 per scenario) for $n = 100, 500, 1000, 2500, 5000$, with $\pi = 0.25$ (such that approximately 50% of subjects had at least one covariate missing). Next, for $n = 1000$, we performed simulations (100 per scenario) with varying proportions of missingness from $\pi = 0.05$ up to $\pi = 0.35$ in steps of 0.05.

For each simulated dataset, we first imputed the missing values in X_1 and X_2 using `mi impute chained`. X_1 and X_2 were imputed using logistic and linear regression models respectively, with the event indicator and Nelson–Aalen estimate of the (marginal) cumulative hazard as covariates. Next we imputed using `smcfcs`, again using logistic and linear models, but imputing compatibly with a Cox proportional hazards model for the survival time. For both methods 10 imputations were used, and in `smcfcs` the default setting for the rejection sampling limit of 1,000 was used. In this setting, `smcfcs` can directly sample from the imputation distribution for the binary covariate X_1, but rejection sampling is used for the continuous covariate X_2.

Figure 2 shows the distributions of the relative computation times taken by `smcfcs` compared with `mi impute chained`, for the different sample sizes considered. This shows that for $n = 100$, `smcfcs` typically takes the same time to complete as `mi impute chained`. However, As the sample size increases, the relative computational cost of `smcfcs` increases, with an approximately 6 fold increase in time taken for $n = 5000$. This additional computational cost is due to the fact that `smcfcs` uses rejection sampling to impute the continuous covariate X_2. As the sample size gets larger, in each dataset there is a larger probability of having at least one record with a very low acceptance probability, such that a large number of proposal draws are required before acceptance. Figure 3 shows the estimates of $\beta_2 = 1$ from the two imputation approaches, again for varying sample sizes. This shows that while `smcfcs` gives unbiased estimates, imputing the covariates directly, using the approximate approach proposed by White and Royston (2009), estimates are systematically biased towards the null, a bias which does not reduce with increasing sample size.

Figure 4 shows that as expected, for a fixed sample size, increasing levels of missingness lead to a modest increase in computation times for `smcfcs` relative to `mi impute chained`. Set against this however, Figure 5 illustrates that the bias in estimates of $\beta_2 = 1$ from `mi impute chained` steadily increase with increasing levels of missingness, whereas `smcfcs` continues to be unbiased.

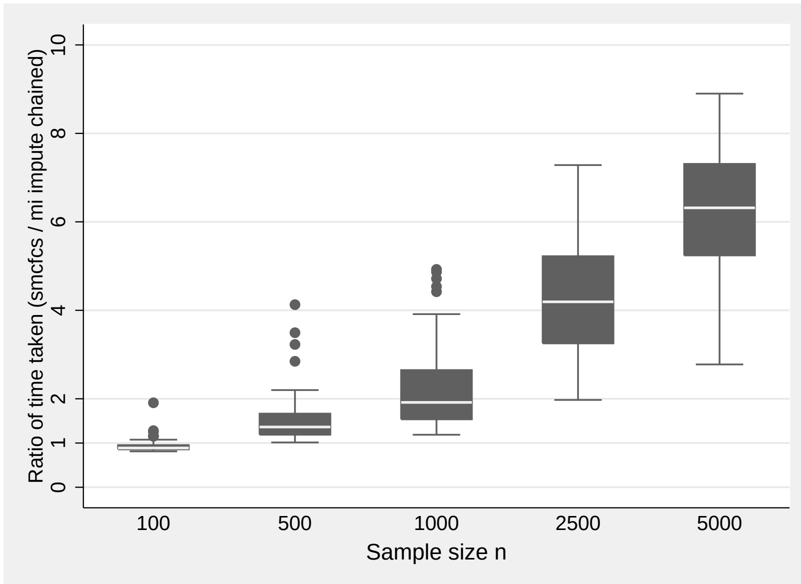

Figure 2: Plot showing ratio of time taken by `smcfcs` to `mi impute chained` for varying sample sizes, $\pi = 0.25$

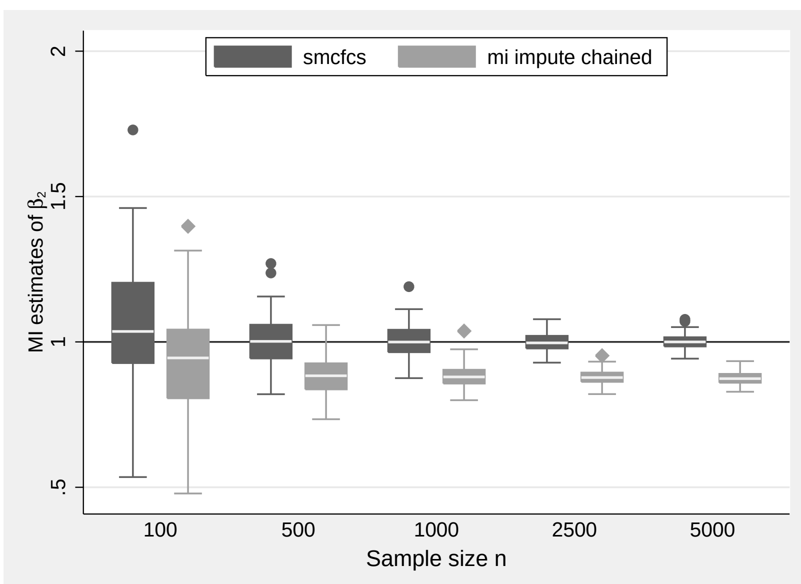

Figure 3: Plot showing estimates of $\beta_2 = 1$ from `smcfcs` and `mi impute chained` for varying sample sizes, $\pi = 0.25$

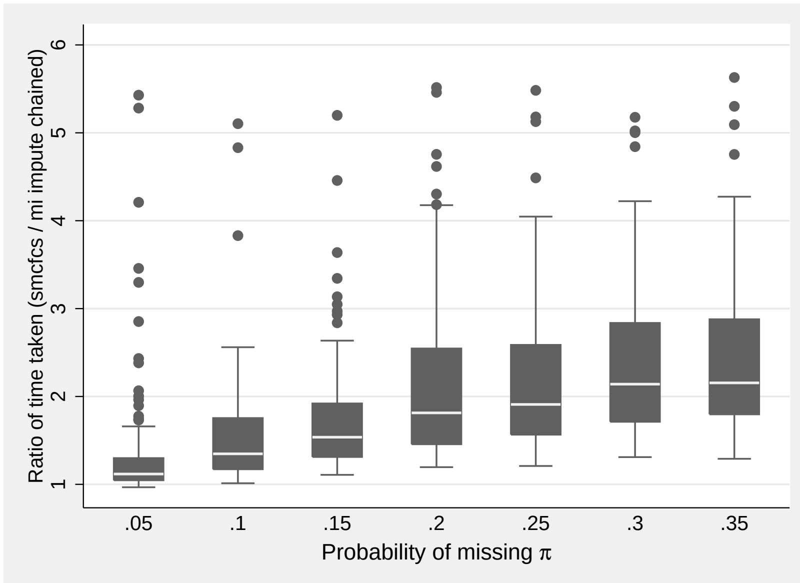

Figure 4: Plot showing ratio of time taken by `smcfcs` to `mi impute chained` for increasing probability of missingness, $n = 1000$

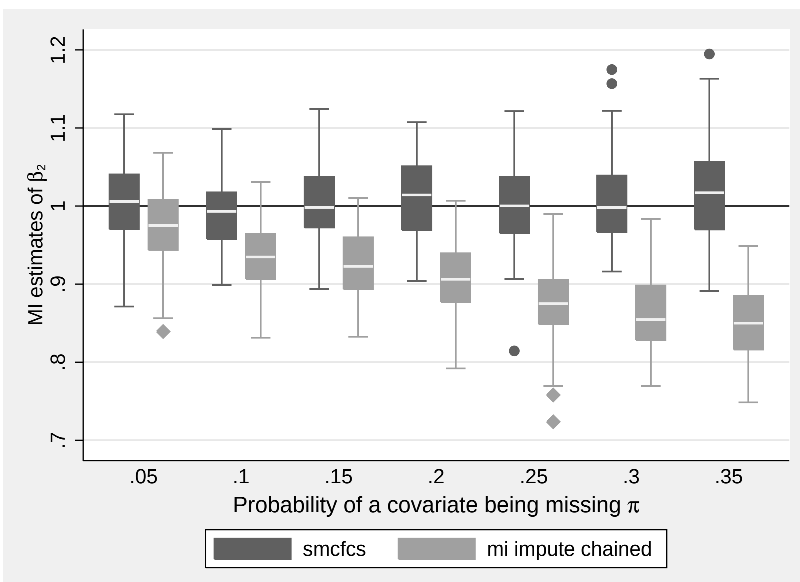

Figure 5: Plot showing estimates of $\beta_2 = 1$ from `smcfcs` and `mi impute chained` for increasing probability of missingness, $n = 1000$

In summary, the simulations demonstrate that `smcfcs` incurs an additional computational cost compared to using `mi impute chained`. Thus in settings where a substantive model compatible imputation model can be specified directly using `mi impute chained` (e.g. a linear regression outcome model with main effects only), use of `smcfcs` is not recommended. Outside of these settings however, use of `smcfcs` is expected to give estimates with less bias, by imputing compatibly with the assumed substantive model, and the increased computational cost would usually be deemed a small price to pay.

6 Final remarks

The `smcfcs` command allows one to impute covariates from imputation models which are compatible with a user-specific substantive model. In settings where the substantive model contains non-linear effects or interactions, and the variables involved in these contain missing values, we believe it offers material advantages relative to what can be achieved using standard FCS/ICE MI. That the algorithm forces the user to specify the substantive model (or rather a substantive model) at the imputation stage is, we believe, a strength of the approach, since it is clear that a set of multiple imputations can be generated which give reasonable results for certain analyses or substantive models but which may give biased estimates for others.

In practice one will typically not know the final substantive model at the imputation stage. A number of possible strategies could be employed. If the complete cases represent a reasonably large proportion of the sample, the substantive model could be chosen (using standard model selection strategies) using the complete cases. Alternatively, one could impute assuming a flexible substantive model, following which simpler, nested models for the outcome can be fitted to the imputations. Conversely, one should not fit substantive models which are not nested within the substantive model used to generate the imputations. For example, one should not impute assuming a substantive model which assumes no interactions and then fit alternative substantive models which allow for interactions.

A concrete example of the above advice is for fractional polynomial (FP) models. In section 4 we used FP transformations that had previously been selected by Royston (2004). If this was not the case we would have had to select our best FP model. To ensure each imputation model is semi-compatible with any FP model that might be selected, the following method could be used. For a partially observed X the transformations X^p under consideration will typically include $p = -2, -1, -0.5, 0, 0.5, 1, 2$, and 3, where X^0 is $\ln(X)$. *All* of these X^p should be included in the SMC–FCS specification of the substantive model. This ensures the imputation model for each partially observed variable (in particular X) is semi-compatible with any (to be subsequenty selected) FP model. An FP model for the outcome can then be selected using the imputed data. Note that for degree-2 FP models, repeated powers for X are possible. If this is a concern the variables $X^p \ln(X)$ should also be included in the 'substantive model' at the imputation stage.

The following fragment of code demonstrates how this strategy can be implemented in practice. Note that `fracgen` involves scaling and centering of `x_1`–`x_7`, so it is important to be aware of this in defining the `passive()` statement.

```
. smcfcs reg y x_1 x_2 x_3 x_4 x_5 x x_7 x_8, reg(x)
> passive(x_1 = x^-2 | x_2 = x^-1 | x_3 = x^-0.5 | x_4 = ln(x) | x_5 = x^0.5 |
> x_6 = x^2 | x_7 = x^3)
```

We believe imputing covariates from a model which is compatible with the substantive model is desirable, since, assuming the latter is correctly specified, unless the imputation model (or a restriction of it) is compatible with the substantive model, the imputation model is mis-specified. We emphasize that this compatibility does not ensure that the imputation model is correctly specified – if the covariate model $f(X_j|X_{-j}, Z, \phi_j)$ is mis-specified for a given value of j, the imputation model is mis-specified. Care should therefore be taken to ensure that the covariate models $f(X_j|X_{-j}, Z, \phi_j)$ are reasonable for the data in hand. Diagnostics which can be applied to multiple imputations should be applied, such as examining the distributions of imputed variables and comparing to the distribution of the observed values.

7 Acknowledgements

JWB was supported by UK Economic and Social Research Council grant RES-189-25-0103, UK Medical Research Council grant G0900724, and a UK Medical Research Council Fellowship MR/K02180X/1. TPM was supported by a UK Medical Research Council PhD studentship A731-5QL14. The authors are grateful to an anonymous referee, whose comments substantially improved the paper.

8 References

Bartlett, J. W., S. R. Seaman, I. R. White, and J. R. Carpenter. 2014. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research* epub: epub.

van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219–242.

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

Hosmer, D. W., S. Lemeshow, and S. May. 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. John Wiley and Sons Inc., New York.

Liu, J., A. Gelman, J. Hill, Y. S. Su, and J. Kropko. 2013. On the stationary distribution of iterative imputations. *Biometrika* 101: 155–173.

Royston, P. 2004. Multiple imputation of missing values. *The Stata Journal* 4: 227–241.

———. 2005. Multiple imputation of missing values: update. *The Stata Journal* 5: 188–201.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63: 581–592.

———. 1987. *Multiple imputation for nonresponse in surveys.* New York: Wiley.

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

White, I. R., and P. Royston. 2009. Imputing missing covariate values for the Cox model. *Statistics in Medicine* 28: 1982–1998.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine* 30: 377–399.