Supporting Information for "A stacked approach for chained equations multiple imputation incorporating the substantive model"

Lauren J. Beesley*[1] **and Jeremy M G Taylor**[1]
[1]University of Michigan, Department of Biostatistics
*Corresponding Author: lbeesley@umich.edu

1 Derivation of proposed variance estimator

In this section, we motivate our estimator for standard errors after analyzing data from stacked multiple imputation (with weights). This estimator can be applied when we routinely impute data using MICE and then stack with weights defined as 1 over the number of times each subject appears in the stacked dataset or when we define imputations and weights as in **Figure 1**.

We first observe that we are interested in estimating the observed data information matrix, I_{obs}. Following Louis (1982), we can express this as follows:

$$I_{obs}(\theta) = I_{com}(\theta) - I_{mis}(\theta)$$

where I_{com} is the expected complete data information given the observed data and I_{mis} is the expected missing information due to the missing data given the observed data. Let J_{com} be the negative of the second derivative matrix of the complete data log-likelihood function. Let U_{com} be the first derivative matrix of the complete data log-likelihood function. Following Louis (1982), we can rewrite this expression as follows:

$$\begin{aligned} I_{obs}(\theta) &= E_\theta(J_{com}(\theta)|X^{obs}, Y) - \left[E_\theta(U_{com}(\theta)^{\otimes 2}|X^{obs}, Y) - E_\theta(U_{com}(\theta)|X^{obs}, Y)^{\otimes 2}\right] \\ &= E_\theta(J_{com}(\theta)|X^{obs}, Y) - Var_\theta(U_{com}(\theta)|X^{obs}, Y) \end{aligned}$$

where the expectations are with respect to the distribution of the missing data. Now, we assume that data are independent across i. In this case, we can rewrite the above as

$$I_{obs}(\theta) = \sum_i E_\theta(J^i_{com}(\theta)|X^{obs}, Y) - \sum_i Var_\theta(U^i_{com}(\theta)|X^{obs}, Y)$$

where $J^i_{com}(\theta)$ and $U^i_{com}(\theta)$ are the contributions to the complete data information matrix and score matrix for subject i respectively.

In practice, these conditional expectations and variances are not simple to calculate. However, we can approximate these expression as averages of these expressions evaluated across imputed datasets, which were imputed by drawing from distributions for the missing data given the observed data. A similar approach is used in the context of Monte Carlo log-likelihood maximization in Wei and Tanner (1990).

Suppose first that we give equal weight to multiple imputations within a particular subject. Let X_{im} be the m^{th} imputation of the missing covariates for subject i. For subjects without missing values, define X_{im} to be all equal to fully-observed X_i, and suppose we apply data analysis using the "tall stack" where each fully-observed subject appears in the stacked dataset M times. A similar expression is also applicable for the "short stack" formulation, where each fully-observed subject appears only once in the stacked data. We can approximate the above expression as follows:

$$I_{obs}(\theta) \approx \sum_i \frac{1}{M} \sum_m J^i_{com}(X_{im}, Y_i; \theta) - \sum_i \frac{1}{M} \sum_m \left[U^i_{com}(X_{im}, Y_i; \theta) - \frac{1}{M} \sum_j U^i_{com}(X_{ij}, Y_i; \theta) \right]^{\otimes 2}$$

where M is the number of multiple imputations. Now, suppose we give multiple imputations within subject i unequal weight, where imputation m for subject i is given weight w_{im}, where $\sum_m w_{im} = 1$. We propose the following reformulation of the above approximation with unequal weights across multiple imputations within subjects:

$$I_{obs}(\theta) \approx \sum_i \sum_m w_{im} J^i_{com}(X_{im}, Y_i; \theta) - \sum_i \sum_m w_{im} \left[U^i_{com}(X_{im}, Y_i; \theta) - \bar{U}^i_{com}(X_i, Y_i; \theta) \right]^{\otimes 2}$$

where $\bar{U}^i_{com}(X, Y_i; \theta) = \sum_j w_{im} U^i_{com}(X_{ij}, Y_i; \theta)$. We can evaluate this expression at the maximum likelihood estimator for θ, $\hat{\theta}$, obtained previously from fitting the model for $Y|X$ to the weighted, stacked dataset to obtain an estimate of the observed data information matrix. Inverting this matrix will provide the estimate for the observed data covariance matrix for $\hat{\theta}$ in *Eq. 3*.

2 Imputation strategy for linear regression

In order to make this estimation strategy clearer, we demonstrate how it works for linear regression. Suppose we are interested in fitting a linear regression model for outcome Y using covariates X and variance parameter σ^2. Suppose further that we have missing data in X, and we multiply impute these missing values using only other information in X to obtain X_{im} for each subject i and imputations $m = 1, \ldots, M$.

Suppose we stack the M imputed datasets on top of each other to create a dataset of size $Mn \times p$, where p is the dimension of X_i. Using $\hat{\theta}_{cc}$ from fitting a linear regression model for $Y|X$ on the complete case data (subjects with X fully observed), we define weights

$$w_{im} = \frac{\frac{1}{\sqrt{2\pi\sigma^2_{cc}}} e^{-\frac{(Y_i - X_{im}\beta_{cc})^2}{2\sigma^2_{cc}}}}{\sum_{j=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2_{cc}}} e^{-\frac{(Y_i - X_{ij}\beta_{cc})^2}{2\sigma^2_{cc}}}} = \frac{e^{-\frac{(Y_i - X_{im}\beta_{cc})^2}{2\sigma^2_{cc}}}}{\sum_{j=1}^{M} e^{-\frac{(Y_i - X_{ij}\beta_{cc})^2}{2\sigma^2_{cc}}}}$$

For subjects with fully-observed X_i, this expression will equal $1/M$ for all m. Now, we define the complete data log-likelihood, score and information matrices (just focusing on the part based on β as follows:

$$\begin{aligned} l^i_{com}(X_{im}, Y_i; \theta) &= -\frac{(Y_i - X_{im}\beta)^2}{2\sigma^2} - \log\left[\sqrt{2\pi\sigma^2}\right] \\ U^i_{com}(X_{im}, Y_i; \theta) &= \frac{Y_i - X_{im}\beta}{\sigma^2} X_{im} \\ J^i_{com}(X_{im}, Y_i; \theta) &= \frac{X_{im}X^T_{im}}{\sigma^2} \end{aligned}$$

so we have that

$$I_{obs}(\theta) \approx \sum_i \sum_m w_{im} \frac{X_{im}X^T_{im}}{\sigma^2} - \sum_i \sum_m w_{im} \left[\frac{Y_i - X_{im}\beta}{\sigma^2} X_{im} - \bar{U}^i_{com}(X_i, Y_i; \theta)\right]^{\otimes 2}$$

where $\bar{U}^i_{com}(X_i, Y_i; \theta) = \sum_j w_{ij} \frac{Y_i - X_{ij}\beta}{\sigma^2} X_{ij}$ and we then plug in the final maximum likelihood estimates for β and σ^2 into the above expression.

3 Example R code for implementation

In this section, we provide some example R code to demonstrate how we can implement our proposed imputation approach. First, we provide some code to simulate outcome Y and covariates X and B from a multivariate normal distribution. We then generate missingness in B under missing completely at random (MCAR) assumptions with a 50% missingness rate.

We use *mice* in R to impute missing values of B, but we impute B from a distribution that does *not* condition on Y. We then take these 50 imputed datasets and stack them. Weights are obtained by first fitting the outcome model (linear regression for $Y|X, B$) to the complete case dataset. We use the resulting parameter estimates to obtain weights proportional to $Y|X, B$. Weights are then scaled to sum to 1 across imputed datasets but within individuals. In the final estimation step, we fit a weighted version of the same regression model to the stacked data. We estimate corresponding standard errors using *Eq. 3* available in the R package *StackImpute*.

```
### Download R package from GitHub
devtools::install_github("lbeesleyBIOSTAT/StackImpute", build_vignettes = TRUE, build_
    opts = c("--no-resave-data", "--no-manual"))
library(StackImpute)

### Simulate Data
Nobs = 2000
DAT = MASS::mvrnorm(n = Nobs, mu = c(0,0,0), Sigma = rbind(c(1, 0.18, 0.42), c(0.18,
    0.09, 0.12),c(0.42, 0.12, 0.49 )))
Y = DAT[,1]
B = DAT[,2]
X = DAT[,3]
S = sample(x=c(0,1), size = Nobs, prob = c(0.5,0.5), replace = TRUE)
complete_cases = data.frame(Y, X, B, S)[S == 1,] #complete case subjects only
observed_data = data.frame(Y, X, B, S) #data with missingness in B
observed_data[S==0,’B’] = NA

### Step 1: Impute B|X
imputes = mice::mice(observed_data, m=50, method="norm", maxit = 1)
pred = imputes$predictorMatrix
pred[pred != 0] = 0
pred["B","X"] = 1
imputes = mice::mice(observed_data, m=50, predictorMatrix=pred, method="norm")

### Step 2: Stack imputed datasets
stack = mice::complete(imputes, action="long", include = FALSE)

### Step 3: Obtain weights
library(dplyr)
fit_cc = glm(Y ˜ X + B, family=’gaussian’, data= complete_cases)
stack$wt = dnorm(stack$Y,mean = predict(fit_cc, newdata = stack), sd = sqrt(summary(fit_
    cc)$dispersion))
stack = as.data.frame(stack %>% group_by(.id) %>% mutate(wt = wt / sum(wt)))

### Step 4: Estimation
fit = glm(Y ˜X + B, data=stack, family=gaussian(), weights = stack$wt)
Info = StackImpute::Louis_Information(fit, stack, M = 50, IMPUTED = unique(stack$.id[
    stack$S==0]))
VARIANCE = diag(solve(Info))
```

Alternatively, one can perform analysis using a short stack, where subjects with complete case data only appear once and subjects with missing data appear M times as follows:

```
### Step 2: Stack imputed datasets
cc = unique(stack$.id[stack$S == 1])
stack_short = rbind(stack[stack$S==0,], stack[stack$S==1 & !duplicated(stack$.id),])

### Step 3: Obtain weights
stack_short$wt = dnorm(stack_short$Y,mean = predict(fit_cc, newdata = stack_short), sd =
     sqrt(summary(fit_cc)$dispersion))
stack_short = as.data.frame(stack_short %>% group_by(.id) %>% mutate(wt = wt / sum(wt)))

### Step 4: Estimation
fit = glm(Y ~X + B, data=stack_short, family=gaussian(), weights = stack_short$wt)
Info = StackImpute::Louis_Information(fit, stack_short, M = 50, IMPUTED = unique(stack_
    short$.id[stack_short$S==0]))
VARIANCE = diag(solve(Info))
```

4 Theoretical justification for stacking and $f(Y|X)$-weighting

Let $\boldsymbol{R}$ be a **random variable** corresponding to data completeness. We make a distinction between this *random variable* and the *observed value*, R, which corresponds to whether each subject is a complete case in the data. This distinction between theoretical missingness, $\boldsymbol{R}$, and the data realization, R, will be important later on. We will assume for now that there is no missingness in Y.

In order to impute missing values of X (denoted $X^{(mis)}$) we want to draw $X^{(mis)}$ from $f(X^{(mis)}|X^{(obs)}, Y, \boldsymbol{R})$, where $X^{(obs)}$ denotes the observed part of X (Little and Rubin, 2002). Conceptualizing these distributions in terms of $X^{(obs)}$ and $X^{(mis)}$ can be confusing, but we will ground these results in practical implementation and in terms of the assumed data models later on. For now, we will continue with this notation. Under missing at random (MAR) dependent on $X^{(obs)}$ and/or Y, we have that

$$\begin{aligned} f(X^{(mis)}|X^{(obs)}, Y, \boldsymbol{R}) &= \frac{f(\boldsymbol{R}|X, Y)}{f(\boldsymbol{R}|X^{(obs)}, Y)} f(X^{(mis)}|X^{(obs)}, Y) \\ &= f(X^{(mis)}|X^{(obs)}, Y). \qquad (\textit{SuppEq. 1}) \end{aligned}$$

This result shows that we can ignore the mechanism generating MAR missingness when imputing $X^{(mis)}$. This argument serves as the backbone justifying standard multiple imputation strategies, where missing values of X are imputed from distributions approximating $f(X^{(mis)}|X^{(obs)}, Y)$ and ignoring the missingness mechanism.

4.1 Imputation and weighting as a two-step procedure

Now, we focus on the structure of $f(X^{(mis)}|X^{(obs)}, Y)$ in the setting where the outcome model for $Y|X$ is of interest. In the main paper, we note that

$$f(X^{(mis)}|X^{(obs)}, Y) = \frac{f(Y|X)}{f(Y|X^{(obs)})} f(X^{(mis)}|X^{(obs)}) \qquad (\textit{SuppEq. 2})$$

One approach to make inference about the distribution of $Y|X$ is to (1) draw multiple imputations from the distribution $f(X^{(mis)}|X^{(obs)}, Y)$ that is proportional to $f(Y|X)f(X^{(mis)}|X^{(obs)})$, (2) analyze the resulting imputed datasets, and (3) apply Rubin's combining rules. This is the approach taken in Bartlett et al. (2014).

An alternative approach grounded in the importance sampling literature is to draw multiple imputations from $f(X^{(mis)}|X^{(obs)})$ and then weight these multiple imputations proportional to $\frac{f(Y|X)}{f(Y|X^{(obs)})}$. This will result in a sample representing approximate draws from $f(X^{(mis)}|X^{(obs)}, Y)$ (Tanner, 1993; Little and Rubin, 2002). We can equivalently weight these draws proportional to $f(Y|X)$. This is the main idea behind our proposed imputation and weighting scheme.

Now, for the analysis. Let U^i_{com} represent the contribution of person i to the complete data score matrix for our $Y|X$ model likelihood. Let X_{ik} represent the k^{th} multiple imputation of X for person i and define unit-scaled weights $w_{ik} = \frac{f(Y_i|X_{ik})}{\sum_{j=1}^{M} f(Y_i|X_{ij})}$. For subjects with fully-observed covariates, define X_{ik} to equal X_i and w_{ik} to equal $1/M$. Following importance sampling logic, we can then estimate a function $h(X_i, Y_i)$ as $h(X_i, Y_i) \approx \frac{1}{M}\sum_{k=1}^{M} w_{ik}h(X_{ik}, Y_i)$. Equivalently, we can estimate the complete data score matrix for the i^{th} subject as

$$U^i_{com}(\theta) \approx \frac{1}{M}\sum_{k=1}^{M} w_{ik} U^i_{com}(X_{ik}, Y_i; \theta),$$

which corresponds to a weighted average of the score matrix evaluated across the M multiple imputations. We can then estimate θ, the parameter of interest, using the overall complete data score matrix

$$U_{com}(\theta) \propto \sum_{i=1}^{n} \sum_{k=1}^{M} w_{ik} U_{com}^{i}(X_{ik}, Y_i; \theta).$$

Solving this score equation for θ is equivalent to solving the weighted score equation using a dataset obtained by stacking the weighted multiply imputed datasets on top of each other. Corresponding standard errors for the maximum likelihood estimate of θ can be estimated as in **Section 1**.

The above stacking approach involves obtaining a stacked dataset with $n \times M$ rows, where each subject with fully-observed data is repeated across M rows with weights $1/M$. We call this approach the "long stack" approach. We could equivalently include only a single row for each subject with fully-observed data and set corresponding weights to be 1. We call this approach the "short stack" approach. The above logic applies in this latter case as well with M, the number of rows in the stacked dataset, varying across subjects.

4.2 Practical Implementation

The proprosed two-step approach involves (1) drawing multiple values of the missing data from $f(X^{(mis)}|X^{(obs)})$ and (2) weighting these draws proportional to $f(Y|X)$. We note that these distributions *do not* condition on their corresponding parameter values. Instead, they are defined integrating over the parameter value. Therefore, some thinking is needed to determine how we perform these steps in practice.

Step 1: Obtaining multiple imputations

First, we consider the imputation step, which is equivalent to performing multiple imputation of missing X using only the observed data in X. Our approach for obtaining these imputations is standard practice in the statistical literature for handling missing data (see Little and Rubin (2002) for a good reference), but we will briefly highlight the main points for readers less familiar with this literature. In practice, it is difficult to conceptualize the distribution $f(X^{(mis)}|X^{(obs)})$, the distribution of the missing data given the observed data. Instead, we often apply a Gibbs Sampling approach where we impute each variable with missingness one-by-one in an iterative algorithm. Missing values of the p^{th} covariate, X_p, can be imputed from an assumed distribution for X_p given X_{-p}, which consists of all covariates except the p^{th}. When each distribution $f(X_p|X_{-p})$ is specified independently and may not correspond to a valid joint distribution, this approach is called chained equations imputation. Covariates with missingness are singly imputed one-by-one from their corresponding conditional distributions, and the entire iterative algorithm is repeated multiple times to obtain multiple imputations.

At each step of the iterative imputation algorithm, we want to impute missing values of X_p from $f(X_p|X_{-p})$. However, this distribution does not condition on its corresponding parameter, ϕ_p. A common approach for obtaining an *approximate* draw from is to first draw parameter ϕ_p using the subjects with complete data on X_p (treating the most recent imputations of X_{-p} as if they had been observed) and then drawing X_p from $f(X_p|X_{-p}; \phi_p)$ evaluated using the drawn value of ϕ_p.

Step 2: Obtaining weights

Now, we consider the weighting step. In this step, we want to evaluate $f(Y_i|X_{ik})$ for each subject. Again, this distribution does not condition on the parameter value. Instead, we can draw a value of parameter θ related to the $Y|X$ distribution and then define weights based on $f(Y_i|X_{ik}; \theta)$ using the drawn θ. The question is then how we go about drawing θ. As a first pass, one might imagine drawing θ by fitting models to the imputed data. However, we must remember that we have imputed missing covariate values from the "wrong" distribution. By this, we mean that we have imputed from distributions that do not condition on Y. As a result, draws of θ using the unweighted imputations will often produce bias.

Instead, we propose to draw θ using a fit of the $Y|X$ model to the overall complete case data consisting of subjects with fully-observed data. We can obtain a *draw* of θ by either a) fitting this model to a bootstrap sample of the complete case data or by b) drawing from a normal distribution using the point estimate and variance-covariance matrix of θ obtained by a complete case fit to the original data (Little and Rubin, 2002). We note that complete case analysis will produce bias in estimating θ when missingness depends on Y. We will address this challenge later on.

Previously, we have discussed drawing θ for use in the weights, but a simple approach would be to use the maximum likelihood estimate from the complete case data rather than a draw to calculate the weights. In practice, we have found that this approach produces good results as seen in our simulations in the main paper. Therefore, we present this strategy in our algorithm below.

Algorithm for stacked and $f(Y|X)$-weighted imputation

(A) Obtain a single imputation of missing X values as follows. After randomly initializing the missing values, iterate the following for each covariate X_p with missingness

- Draw ϕ_p using fit of $X_p|X_{-p}$ model to subjects with fully observed X_p, treating the most recently filled-in values of X_{-p} as if they were observed
- Draw missing X_p from the assumed distribution of $X_p|X_{-p}$ (e.g. logistic regression model) evaluated at the drawn ϕ_p.

(B) Repeat (A) M times to obtain M multiple imputations.
(C) Define weights proportional to $f(Y|X;\theta)$ using the complete case estimate of θ.
(D) Re-estimate θ by fitting the $Y|X$ model to a stacked, weighted version of the data.
(E) Estimate standard errors for $\hat{\theta}$ using *Eq. 3*.

4.3 MAR missingness dependent on Y

The proposed approach involves imputing missing values of X conditioning on observed values of X but not conditioning on Y. When missingness depends on Y and we ignore Y in the imputation models, we induce missing not at random (MNAR) missingness in X. This presents a problem, since standard multiple imputation methods assume that missingness depends only on observed data (MAR).

The primary challenge comes from how we draw parameters during the imputation and weighting steps. In both steps, we perform parameter draws using some type of complete case fit to the observed data. When missingness depends on Y, complete case analysis for the $Y|X$ model will often produce biased estimates of θ. Additionally, when we ignore Y in the imputation model $X_p|X_{-p}$, we induce a dependence between missingness and X_p. Fitting a model for $X_p|X_{-p}$ using only subjects with complete X_p will often result in biased parameter estimates. Therefore, complete case-based parameter draws in both steps of the proposed imputation and weighting procedure will be biased when missingness depends on Y.

Key Takeaway: Complete case analysis will produce biased estimates of $\phi = (\phi_1, \ldots, \phi_P)$ and θ when missingness depends on Y

However, simulations provide little evidence of bias in the overall estimate of θ, and there is a theoretical reason for this. Recall the distinction between $\boldsymbol{R}$, the random variable corresponding to missingness, and R, the realization $\boldsymbol{R}$ takes in the observed data. We noted in *SuppEq. 1* that $f(X^{(mis)}|X^{(obs)}, Y, \boldsymbol{R}) = f(X^{(mis)}|X^{(obs)}, Y)$ under MAR dependent on $X^{(obs)}$ and/or

Y. This result means that the *distribution* of $X^{(mis)}|X^{(obs)}, Y$ is the same as the *distribution* given $\boldsymbol{R} = 1$ (again, this is the random variable $\boldsymbol{R}$, not the data realization). However, we have that

$$f(X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1) \neq f(X^{(mis)}|X^{(obs)})$$
$$f(Y|X, \boldsymbol{R} = 1) \neq f(Y|X)$$

Therefore, missingness based on Y is not ignorable for either of these individual distributions. We have that

$$f(X^{(mis)}|X^{(obs)}, Y) = f(X^{(mis)}|X^{(obs)}, Y, \boldsymbol{R} = 1) = \frac{f(Y|X, \boldsymbol{R} = 1)}{f(Y|X^{(obs)}, \boldsymbol{R} = 1)} f(X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1) \qquad (\textit{SuppEq. 3})$$

This last equality takes the same form as *SuppEq. 2* except that it conditions on $\boldsymbol{R} = 1$. Following the same logic as in **Section 4.1**, we can impute missing covariate values from $f(X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1)$ and weight these imputations proportional to $f(Y|X, \boldsymbol{R} = 1)$.

Let θ_R and ϕ_R be the parameters related to $Y|X, \boldsymbol{R} = 1$ and $X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1$ respectively. As before, these distributions do not condition on parameters ϕ_R and θ_R. Recall, these parameters come from distributions that condition on the random variable $\boldsymbol{R} = 1$. In practice, we can obtain valid draws of ϕ_R and θ_R using the data realization of $\boldsymbol{R}$, R. In other words, complete case analysis will produce unbiased estimates of ϕ_R and θ_R.

Key Takeaway: Complete case analysis will produce unbiased estimates of ϕ_R and θ_R when missingness depends on Y

Next, suppose we can approximate $f(Y|X, \boldsymbol{R} = 1; \theta_R) \approx f(Y|X; \theta_R)$ and $f(X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1; \phi_R) \approx f(X^{(mis)}|X^{(obs)}; \phi_R)$ where θ_R and ϕ_R are the parameters from the distribution conditioning on $\boldsymbol{R}$. In other words, suppose we can approximate the distribution conditional on $\boldsymbol{R} = 1$ with the unconditional distribution evaluated at the parameter from the conditional distribution. This will be reasonable if these distributions have similar structures (e.g. both are linear regression) but with different parameter values.

As an example, suppose $Y|X$ is a logistic regression. Unless missingness depends only on Y, the distribution of $Y|X, \boldsymbol{R} = 1$ does **not** follow a standard logistic regression model structure. Instead, we have

$$\text{logit}(P(Y = 1|X, \boldsymbol{R} = 1)) = \theta^T(1, X) + \log\left[\frac{P(\boldsymbol{R} = 1|X, Y = 1)}{P(\boldsymbol{R} = 1|X, Y = 0)}\right]$$

where θ is the parameter in the unconditional logistic regression for $Y|X$. However, suppose we approximate the $Y = 1|X, \boldsymbol{R} = 1$ distribution with a logistic regression. In this case, we can define weights using the logistic regression model structure for $Y|X$ evaluated at the complete case point estimate for θ_R. We can make similar substitutions for complete case draws of ϕ_R in the distribution for $X^{(mis)}|X^{(obs)}$. We note that this assumes we draw parameters using global complete case analysis (only subjects with fully-observed data) rather than partial complete case analysis (using subjects with fully-observed X_p for imputing X_p) as is done in MICE. We address this issue in **Section 5.2**.

Assuming we can make these approximations, we can impute missing data using the algorithm in **Section 4.2** involving complete case parameter draws even though the individual distributions for $Y|X$ and $X^{(mis)}|X^{(obs)}$ are not equal to the distributions for $Y|X, \boldsymbol{R} = 1$ and $X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1$ respectively. This ultimately comes from the fact that $f(X^{(mis)}|X^{(obs)}, Y) \propto f(Y|X, \boldsymbol{R} = 1) f(X^{(mis)}|X^{(obs)}, \boldsymbol{R} = 1)$. All of these subtle distinctions and approximations boil down to the following.

Key Takeaway: If we can reasonably approximate the *structures* (e.g. logistic regression, linear regression with main effects only, etc.) of the distributions for the outcome and missing covariates conditional on $\boldsymbol{R} = 1$ with the *structures* of the unconditional distributions, then we can apply the imputation and weighting algorithm in **Section 4.2** even with covariate missingness dependent on Y.

4.4 Handling missingness in Y

Up until now, we have assumed that Y is fully observed. However, it may be that Y is only partially observed. Here, we restrict our attention to settings where missingness is independent of Y. Now, subjects with Y entirely missing do not really provide information regarding the relationship between Y and X, so we may simply perform our analysis and imputation entirely ignoring subjects with missing Y. However, there are many cases where we may have partial information on Y. For example, suppose Y represents observations across multiple time-points for each subject. If some subjects have missing observations at only some time-points, it is inefficient to entirely drop those subjects from analysis. Instead, we may want to impute the missing values of Y to include in our final analysis.

Our proposed imputation and weighting strategy can be easily adapted to handle missingness in Y. Let $Y^{(mis)}$ and $Y^{(obs)}$ denote the missing and observed parts of Y respectively. We want to impute missing Y and X from the joint distribution

$$\begin{aligned} f(X^{(mis)}, Y^{(mis)}|X^{(obs)}, Y^{(obs)}) &= f(Y^{(mis)}|X^{(obs)}, Y^{(obs)}, X^{(mis)})f(X^{(mis)}|X^{(obs)}, Y^{(obs)}) \\ &= f(Y^{(mis)}|X, Y^{(obs)})f(X^{(mis)}|X^{(obs)}, Y^{(obs)}) \\ &\propto \left[f(Y^{(obs)}|X)\right]\left[f(Y^{(mis)}|X, Y^{(obs)})f(X^{(mis)}|X^{(obs)})\right] \end{aligned}$$

We can apply the same logic as before and obtain approximate draws from $f(X^{(mis)}, Y^{(mis)}|X^{(obs)}, Y^{(obs)})$ by obtaining (1) imputations of X from $f(X^{(mis)}|X^{(obs)})$, (2) imputations of Y from $f(Y^{(mis)}|X, Y^{(obs)})$, and (3) corresponding analysis weights from $f(Y^{(obs)}|X)$. Step 1 can proceed as in **Section 4.2**. We can implement Step 2 by imputing missing values of Y using the distribution of $Y|X$ fit to the global complete case data. Then, we calculate weights proportional to $f(Y|X)$ with the parameter obtained from the global complete case fit.

5 Additional notes about missingness in covariates dependent on Y

5.1 An illustrative example

An interesting feature of the proposed stacking and imputation approach involving weighting by $f(Y|X)$ is that complete case analysis in both the imputation and weighting stages produces bias. However, the point estimates in the final data analysis show very little bias.

In **Figure S1**, we show the estimated parameters for the $Y|X$ and $X_2|X_1$ model across 500 simulated datasets under simulation Scenario 1 and covariate missingness in X_2 dependent on Y. Under this missingness model, we expect complete case analysis to be biased for estimating both the $Y|X$ model (**Figure S1a**) and the $X_2|X_1$ model (**Figure S1b**). For both distributions, we can see that complete case analysis (red) clearly results in substantial bias relative to analysis of the full data (green) without missingness. When we are going to impute missing X_2, we are drawing parameters associated with the imputation model from the red distribution when we perform complete case analysis. Similarly, estimation of the weights in stacked data analysis based on complete case analysis uses the red distribution under complete case analysis. When we put these pieces together and perform the proposed stacked and $f(Y|X)$-weighted analysis, however, we obtain the yellow estimates in **Figure S1a**. These estimates (yellow) are centered near the full data distribution (green). As motivated by our theoretical development, when we put the imputation and weight stages together (both of which involve biased distributions from complete case analysis as in **Web Appendix Section 4.2**) and perform our final data analysis on the stacked and $f(Y|X)$-weighted data, the resulting outcome model parameters show very little bias.

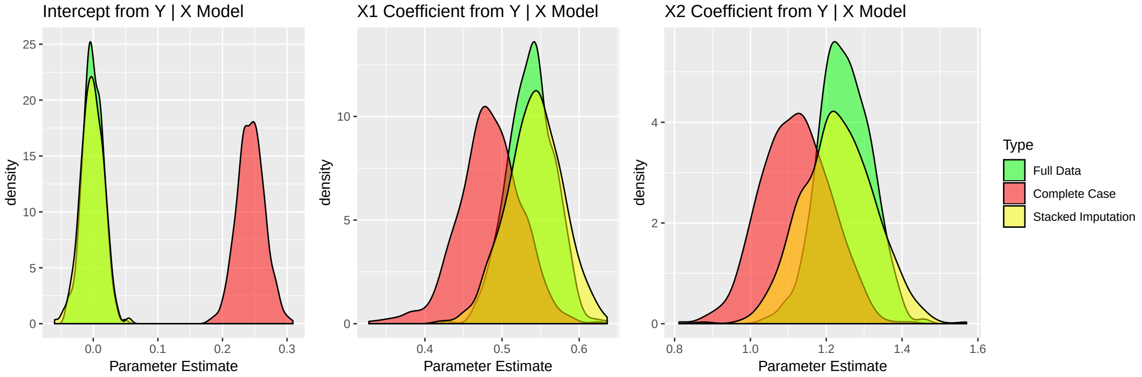
5.2 A note on parameter draws

Our theoretical justification for applying the proposed imputation and weighting approach under MAR dependent on Y assumes that parameters used for covariate imputation are drawn using the global complete case data (subjects with $R_i = 1$). However, users may often want to apply existing MICE software for performing the covariate imputation. When imputing each covariate X_p with missingness, most MICE imputation algorithms obtain parameter draws using the partial complete case data with respect to X_p (data from all subjects with X_p observed) rather than the global complete case data. In practice, we do not expect this to substantially impact the performance of our proposed approach.

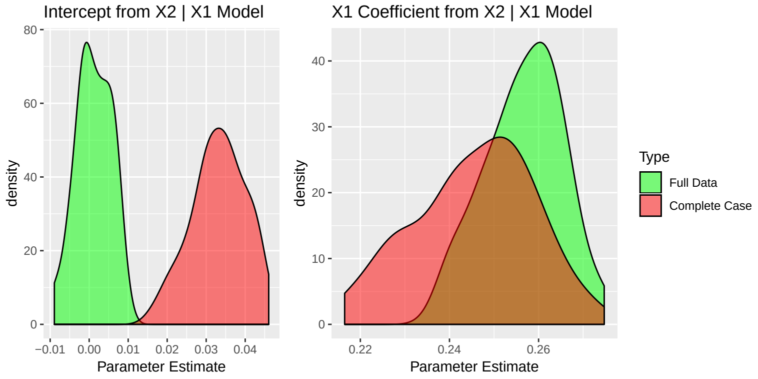
However, there is still a potential that parameter draws as implemented in the R package *mice* could result in some small bias when missingness depends on Y, since the theoretical justification in this setting relies on global complete case analysis for parameter draws. In **Table S1**, we compare point estimates obtained from our $f(Y|X)$-weighted stacking approach when imputation model parameters are drawn using the global complete case data vs. the covariate-specific partial complete case data (termed MICE complete case draws here). We compare the outcome model point estimates across 500 simulated datasets for two outcome model scenarios: logistic and linear regression. In the logistic regression case, we do not see any evidence of bias in parameter estimates due to the choice of parameter draws. In the linear regression case, we see some small bias in estimating the intercept of the linear regression model when parameters are drawn as in MICE rather than using the overall complete case data. However, these biases are very small, and the covariate effects of interest appear unaffected. We see this phenomenon throughout our simulations, where at most negligible bias can be seen to result from the use of MICE-type parameter draws rather than parameter draws based on the overall complete case data.

Figure S1: Estimation of $Y|X$ model parameters using full data, complete case data, and stacked imputation with $f(Y|X)$ weighting across 500 simulated datasets*

(a) Distribution of estimated $Y|X_1, X_2$ model parameter estimates



(b) Distribution estimated $X_2|X_1$ model parameter estimates



* Estimation using full data without missingness, complete case data, or estimated parameters after applying our proposed stacked data analysis approach.

Table S1: Impact of parameter drawing strategy on bias of outcome point estimates for the Stacked, $f(Y|X)$-weighted approach across 500 simulated datasets.*

Missingness	Method	Intercept	X_1	X_2	X_3
	Scenario 2: Logistic regression with missingness in X_2 and X_3				
MAR X, Y	MICE complete case draws**	-0.007	0.003	-0.005	0.002
MAR X, Y	Overall complete case draws	0.001	-0.001	-0.005	0.003
MAR Y	MICE complete case draws	0.006	0.001	0.001	-0.003
MAR Y	Overall complete case draws	0.001	0.001	0.001	-0.003
	Scenario 5: Linear regression with missingness in X_2 and X_3†				
MAR X, Y	MICE complete case draws	-0.023	0.005	-0.007	-0.004
MAR X, Y	Overall complete case draws	0.002	0.001	-0.005	-0.001
MAR Y	MICE complete case draws	0.015	-0.001	-0.003	-0.005
MAR Y	Overall complete case draws	-0.002	0.004	-0.002	-0.003

* True values = 0.50 for all parameters. Covariate imputation used correctly-specified linear regression imputation models.

† X_1, X_2, and X_3 simulated as in Scenario 2. Y generated under linear regression model $N(0.5+0.5X_1+0.5X_2+0.5X_3, 1)$, and missingness generated as in Scenario 2.

** "MICE complete case draws" refers to the practice of drawing the parameter for imputing covariate X_p using the partial complete case data with respect to X_p as implemented in *mice* in R. "Overall complete case draws" corresponds to drawing the parameter for imputing covariate X_p using the subjects with global complete case data for all X.

References

Jonathan W Bartlett, Shaun R Seaman, Ian R White, and James R Carpenter. Multiple imputation of covariates by fully conditional specification: accomodating the substantive model. *Statistical Methods in Medical Research*, 24(4):462–487, 2014.

Roderick J A Little and Donald B Rubin. *Statistical analysis with missing data.* John Wiley and Sons, Inc, Hoboken, NJ, 2nd edition, 2002.

Thomas A Louis. Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, 44(2):226–233, 1982.

Martin A Tanner. *Methods for the Exploration of Posterior Distributions and Likelihood Functions.* Springer, 2nd edition, 1993. ISBN 9781468401943.

Greg C G Wei and Martin A Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.