Should Missing Values of the Outcome Variable Be Imputed for Regression Models?

Benjamin Bagozzi
Department of Political Science
1472 Social Sciences

University of Minnesota
267 19th Ave South
Minneapolis, MN 55455
bbagozzi@umn.edu

Jeff Gill
Department of Political Science
Division of Biostatistics
Department of Surgery
Washington University in St. Louis
One Brookings Drive, Campus Box 1063
Saint Louis, MO 63130
jgill@wustl.edu

ABSTRACT

While approaches to missingness in explanatory variables are now well understood, researchers are often confused about missingness in modeled outcomes. In this work we look at the question of whether missing outcome variables should be imputed with standard tools. We summarize the current state of practice in statistics and empirical social science, analytically derive the effects of imputation, and demonstrate properties with a Monte Carlo simulation. In general standard imputation (multiple imputation, random imputation, Bayesian stochastic imputation, etc.) are appropriate for missing Y-variables, but we also point out areas of caution. Our empirical example highlights the appropriate practice.

An earlier draft of this paper was presented at the Annual Meeting of the Southern Political Science Association, January 15, 2014. We thank Douglas Rice, Martin Johnson, and XXXX for their helpful comments and suggestions.

1 Introduction

The treatment of missing data in political science has improved dramatically over the last decade. Empirical researchers are generally aware that the practice of case-wise deletion for subjects with any missing values leads to biased results under all but the most optimistic circumstances. However, this was the practice for most of the history of quantitative political studies. Furthermore, easy-to-use software now exists such as `Amelia` (standalone and `R`), `mice` (`R`), `hot.deck` (`R`), `Solas` (standalone, commercial), `mitools` (`R`), the `mi impute` command in `Stata`, and more. Yet one missing data issue remains unresolved. When is it okay to impute missingness in the variable that is specified as the outcome in a subsequent statistical model? In this work we address this question technically and provide circumstances where such a procedure is appropriate and other where additional caution is warranted.

Contemporary trends in political science research strongly suggest that such guidance is needed. To gain an understanding of current imputation strategies with respect to outcome variables, we surveyed the past 12 years of the *American Political Science Review*, *American Journal of Political Science*, *Journal of Politics*, and *Political Analysis* and identified all articles that used some form of imputation during this time period.[1] Through this exercise, we found a total of 119 articles in these four journals (2003-2014) that discussed using *some* form of imputation during some stage of their analysis. Of these 119 articles, 73 (approximately 61%) appeared to impute their outcome variable, meaning that based upon our 'best guess,' the remaining 39% either erred against imputing the outcome variable or did not have to impute the outcome variable due to nonmissingness. Hence, current practices seem to suggest that substantial disagreement exists among political scientists when it comes to the imputation of outcome variables.

Equally as striking, however, was our difficulty in determining the above proportions. In fact, for the same 119 cases mentioned above, only 49 (41%)—that is, fewer than half of all articles using imputation in the discipline's top four journals over the past 12 years—explicitly mentioned whether or not they imputed the outcome variable within their methodology and imputation discussions. The majority of articles using imputation during this period gave no mention whatsoever to their handling of the outcome variable. We interpret this as a further sign that—in addition to outright disagreements about the handling of outcome variables in one's imputations—substantial uncertainties remain within the discipline as to whether one should impute outcome variables alongside one's explanatory variables.

Despite these admonitions, political science has recently recognized the need for treating missing data more appropriately (Cranmer and Gill 2013, Honaker and King 2010). Furthermore, particular models used by empirical researchers have received attention in this regard such as Enders *et al.* (2001) for structural equations models. Accessible discussions about theoretical aspects of imputation are given by Little (1988a, 1988b, 1992) and Mcknight *et al.* (2007).

When data are missing no statistical procedure can recover the actual missing values. Instead the goal is to deal with the missingness in such a way that the estimated quantities are unbiased by these absent values. So for a set of explanatory values, $\mathbf{X}$ ($n \times k$), and associated outcome variable n-length

[1] Relevant articles were identified based on searching for "imputation OR impute" in Google Scholar, with additional search fields set to each respective journal and the years of interest. The full list of uncovered articles is included in our Supplemental Appendix.

vector, $\mathbf{y}$, the likelihood function for the estimated coefficients, $L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})$, does not produce bias from the treatment of missingness. To clarify the structure of missingness, first segment the data into two groups:

$$Z_{mis} = (X_{mis}, Y_{mis}) \tag{1}$$

$$Z_{obs} = (X_{obs}, Y_{obs}) \tag{2}$$

such that there are no missing values in Z_{obs}. Now stipulate a $n \times k$ matrix, $\mathtt{R}$ corresponding to $\mathbf{X}$ that contains 0 when the $\mathbf{X}$ matrix data value is *not* missing, and 1 when it is missing. Additionally define ϕ as a parameter that "causes" missingness. This could be purposeful or inadvertant. The now standard terms from Rubin (1976, 1977) are:

MCAR	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\phi)$	missingness not related to observed or unobserved
MAR	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\mathbf{Z}_{obs}, \phi)$	missingness depends only on observed data
Non-Ignorable	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \phi)$	missingness depends on unobserved data

So for MCAR data there is no other data, observed or unobserved, that is related to the pattern of missingness. In this scenario, casewise deletion does not bias results, although it will decrease the degrees of freedom and the power of later tests. Note, however, that this is relatively rare occurrence in the social sciences where there is almost never a collection of independent variables in a data matrix. With MAR data there is information in the observed data that is related to the missing data and therefore can be useful in "filling-in" substitute values. The typologies of MCAR and MAR are collectively called "ignorable," meaning that there are tools for dealing with these conditions. This is where standard imputation procedures are useful. Finally, non-ignorable missing data depends on other missing data, which is the most challenging circumstance. There are no general statistical procedures for dealing with non-ignorable missing data, so researchers usually add assumptions or occasionally use specialized techniques (see Schafer 1997, pp.26-28). Importantly, there is no statistical procedure for determining which type you have: it is a substantive question. Therefore researchers need to make a defensible assumption about which type of missing data is present before proceeding to modeling.

2 Multiple Imputation

Multiple imputation (Rubin 1979) is the most general and most useful process for filling-in substitutes for the missing data and can reliably be applied to MCAR and MAR data. There are three general steps:

- impute values for the missing data (Z_{mis}) conditional on the observed values times to get M complete replicate datasets,
- analyze/regress each dataset separately,

- combine results with summary process.

The imputation step assumes a conditional *posterior* distribution for the missing data conditioning on observed values, and the more information in other variables the smaller the variance of this distribution. Rubin (1979) shows that $M = 5, \ldots, 10$ is usually sufficient. The combining process uses the output of the M separate regression (or other) models to produce a single result to report. The coefficients are simply averaged:

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m, \tag{3}$$

creating a single vector from the mean of the M $\hat{\theta}$ values. The total variance of these coefficient estimates is composed of the variation of the coefficient estimates *within* each imputed dataset and the variation of the coefficient estimates *between* the imputed datasets. The within imputation variance is the mean of individual coefficient variances across models:

$$W_M = \frac{1}{M}\sum_{m=1}^{M}\Sigma_m. \tag{4}$$

The between imputation variance is the variance of the M coefficient estimates:

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2. \tag{5}$$

The total variance for $\bar{\theta}_M$ is an ANOVA-style weighted sum:

$$T_M = W_M + \left(1 + \frac{1}{M}\right)B_M, \tag{6}$$

where $[1 + 1/M]$ is the adjustment for finite M and the new degrees of freedom are:

$$df_{MI} = (M-1)\left[1 + \frac{1}{M+1}\frac{W_M}{B_M}\right]. \tag{7}$$

Other revealing quantities from the imputation process in include: the *relative increase in variance due to nonresponse* (Rubin 1987): $r = \frac{m}{(m+1)}\frac{B}{W}$, where: B is the coefficient between variance and W is the average within imputation variance; the *fraction of information missing for the coefficient estimates* is $\lambda = \frac{r + \frac{2}{df+3}}{r+1}$. the *relative efficiency* of using m imputations rather an infinite number: $re = \left(1 + \frac{\lambda}{m}\right)^{-1}$.

Rubin and others have further refined the multiple imputation procedure (Rubin 1978, Rubin and Schenker 1986, Rubin 1986, Rubin and Schafer 1988, Schafer and Graham 2002, Rubin 1996), but the basic mechanism described above is still the same. Since multiple imputation in its variants and software implementations remains the general state-of-the-art procedure, all of the imputations discussed herein will be done with `mice` (van Burren *et al.* 2006, van Buuren 2007) or `hot.deck` (Cranmer and Gill 2013) in `R`.

3 What About Missingness On the Outcome Variable?

This issue of handling missing data for the outcome variable remains unresolved in the literature. Recall that the imputation process doesn't care about the eventual model, so including the $\mathbf{y}$ vector along with the $\mathbf{X}$ matrix has no effect on the imputation stage. When there is missingness in $\mathbf{y}$ but $\mathbf{X}$ is complete then the incomplete cases contribute no information to the regression of $\mathbf{y}$ on $\mathbf{X}$ (Little 1992, p.1227). When there is missingness in both $\mathbf{y}$ and $\mathbf{X}$ cases with missing values in $\mathbf{y}$ can have some information for the model by improving prediction of missing $\mathbf{X}$ values for cases with $\mathbf{y}$ present.

3.1 Some Recent Procedures

There have been some limited attempts to produce outcome variable imputation procedures. This section describes the two works that make the most progress. Some of these will appear to particular researchers, but none of them are general enough to be fully satisfying.

von Hipple (2007) introduces *Multiple Imputation Then Deletion* (MID), which has the following steps:

- run multiple imputation adding $\mathbf{y}$ on as an additional column of $\mathbf{X}$
- case-wise delete the rows that have imputed values for y
- run M models,
- combine results as usual.

Thus the rows with missing y still help in the imputation process, and MID can be more efficient than MI for small M and a large amount of missing data. However there are two relatively strict assumptions:

- missing $\mathbf{y}$ values are ignorable: unobserved $\mathbf{y}$ values are similar to observed $\mathbf{y}$ values from cases with similar values for $\mathbf{X}$
- missing $\mathbf{X}$ values are ignorable in cases with missing $\mathbf{y}$.

The first assumption states the outcome variable values that are missing are not related to each other and that there is an implied relationship between the $\mathbf{X}$ matrix rows and the corresponding $\mathbf{y}$ values. If that sounds like the assumption of a standard regression model, it is because that is the assumption of a standard regression model: $y_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) + e_i$. So one criticism of the von Hipple procedure is that conflates the imputation procedure with the modeling procedure. The second assumption is less distasteful: all of the missing $\mathbf{X}_i$ values for a case with missing $\mathbf{y}_i$ must be ignorable, so a missing $\mathbf{y}$ imposes restrictions on the rest of that case.

Little and Rubin (2002) show how least squares estimation of outcome variables in ANCOVA models with experimental data can be done effectively. They also give a solution for linear models using the EM algorithm when all of the missingness is confined to the outcome variable. Using starting values for $\boldsymbol{\beta}$ and σ^2, for $j = 1$ until convergence do the following iterations for each missing value:

- **E-Step**: $E[\mathbf{y}_i|\mathbf{X}, \mathbf{y}_{obs}, \boldsymbol{\beta}_j, \sigma_j^2] = \begin{cases} y_i & \text{if } y_i \text{ observed} \\ \mathbf{X}\boldsymbol{\beta}^{(j)} & \text{if } y_i \text{ not observed} \end{cases}$
  Produces a complete outcome vector: $y^{(j)}$.

- **M-Step**: $\boldsymbol{\beta}^{(j+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}^{(j)}$.
  Produces a new coefficient estimate: $\boldsymbol{\beta}^{(j+1)}$.

When there is also missingness in $\mathbf{X}$ they make a multivariate normal assumption for $[\mathbf{X}, \mathbf{y}]$ and use a slightly more involved version of the EM algorithm with a maximum likelihood M-Step. See also Lauritzen (1995). for the use of EM with missing data. Yet these solutions will be helpful to only a narrow slice of empirical political scientists.

Shao (2013) recently derived a method for missing outcome variable values restricted to linear regression modeling. For iid data $i = 1, \dots, n$ you have $\mathbf{x} = (1, \mathbf{b}, \mathbf{z})$ according to:

- $\mathbf{u}$: covariates with missing data, size $n \times k_u$,

- $\mathbf{z}$: covariates with no missing data, size $n \times k_z$.

We want to estimate $\boldsymbol{\beta} = (\alpha, \boldsymbol{\beta}_u, \boldsymbol{\beta}_z)$ from a linear model:

$$E(\mathbf{y}|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} = \alpha + \mathbf{u}\boldsymbol{\beta}_u + \mathbf{z}\boldsymbol{\beta}_z. \tag{8}$$

Assume $E(\mathbf{u}|\mathbf{z}) = \boldsymbol{\Gamma}\mathbf{z}$ where $\boldsymbol{\Gamma}$ is some unknown matrix with size $k_u \times n$ (row size the same as the columns of $\mathbf{u}$ and column size the same as the rows of $\mathbf{z}$). Further assume that conditional on $\mathbf{z}$ the components of u are independent. Now Create two indicator structures to keep track of missingness:

- $\mathbf{a}$ is a vector where $\mathbf{a}_i = 1$ if $\mathbf{y}_i$ is observed and $\mathbf{a}_i = 0$ otherwise.

- $\mathbf{b}$ is a matrix where $\mathbf{b}_{ij} = 1$ if $\mathbf{u}_{ij}$ is observed and $\mathbf{u}_{ij} = 0$ otherwise.

Assume:

$$p(\mathbf{a}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{u}, \mathbf{z}_i) = p(\mathbf{a}_i|\mathbf{u}_i, \mathbf{z}_i),\ \forall i \qquad \text{(non-ignorable)} \tag{9}$$

$$p(\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i) = p(\mathbf{u}_i|\mathbf{z}_i),\ \forall i \qquad \text{(MAR)} \tag{10}$$

Then:

$$p(\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i, \mathbf{a}_i) = \frac{p(\mathbf{a}_i|\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i\mathbf{b}_i)p(\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i)}{p(\mathbf{a}_i|\mathbf{y}_i, \mathbf{u}_i, \mathbf{z}_i)} = \frac{p(\mathbf{a}_i|\mathbf{b}_i, \mathbf{z}_i)p(\mathbf{b}_i|\mathbf{z}_i)}{p(\mathbf{a}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{z}_i)} = p(\mathbf{b}_i|\mathbf{z}). \tag{11}$$

We can impute missing values in a single covariate column from $\mathbf{u}$: $\mathbf{u}_j$ of length n. Notate γ_j as the corresponding n-length row of the $\boldsymbol{\Gamma}$ matrix, so that $E(\mathbf{u}_j|\mathbf{z}) = \gamma_j\mathbf{z}$. The n-length vector $\mathbf{b}_j$ contains the indicators for missingness in this single covariate. We don't know γ_j but we can estimate it by regressing through the origin $\mathbf{b}_j$ on the matrix $\mathbf{z}$:

$$\underset{(1\times k_z)}{\hat{\gamma}_j} = \left(\sum_{i=1}^{n} \underset{(1\times 1)}{\mathbf{b}_i} \underset{(1\times k_z)}{z_i} \underset{(k_z\times 1)}{z_i'} \right)^{-1} \sum_{i=1}^{n} \underset{(1\times 1)}{\mathbf{b}_i} \underset{(1\times 1)}{\mathbf{u}_{ij}} \underset{(1\times k_z)}{z_i}. \tag{12}$$

Using $E(\mathbf{u}|\mathbf{z}) = \mathbf{\Gamma z}$, we now have a filled-in dataset with asymptotically unbiased estimates for each case with missingness $\hat{\mathbf{x}}_i = (1, \hat{\mathbf{b}}, \mathbf{z})$ under the two assumptions given above. The estimate of $\boldsymbol{\beta} = (\alpha, \boldsymbol{\beta}_u, \boldsymbol{\beta}_z)$ is produced with another OLS step:

$$\underset{(1\times[1+k_u+k_z])}{\hat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{n} \underset{(1\times1)}{\mathbf{a}_i} \underset{(1\times[1+k_u+k_z])}{\hat{\mathbf{x}}_i} \underset{([k_u+k_z]\times1)}{\hat{\mathbf{x}}_i{}'} \right)^{-1} \sum_{i=1}^{n} \underset{(1\times1)}{\mathbf{a}_i} \underset{(1\times1)}{\mathbf{y}_i} \underset{(1\times[1+k_u+k_z])}{\hat{\mathbf{x}}_i}. \tag{13}$$

Unfortunately $\hat{\boldsymbol{\beta}}$ is asymptotically biased unless we replace the non-ignorable assumption above with ignorability in: $p(\mathbf{a}_i|\mathbf{y}_i, \mathbf{b}_i, \mathbf{u}, \mathbf{z}_i) = p(\mathbf{a}_i|\mathbf{u}_i, \mathbf{z}_i),\ \forall i$. Shao (p.365) solves this problem in a surprising way. For the data matrix given by $[\mathbf{y}, \mathbf{X}]$ define subsets where the $\mathbf{X}$ are missing on the same covariates irrespective of the status of the $\mathbf{y}$. Therefore there will be some missing outcome variable values and some observed outcome variable values in each subset. For the cases where there are no missing covariate values, $\mathbf{X}_{\mathsf{full}}$ but a missing $\mathbf{y}$ value, produce the estimate from $\hat{y} = \mathbf{X}_{\mathsf{full}}\hat{\boldsymbol{\beta}}$. For each of the other cases with the same pattern of covariate missingness, $\mathbf{X}_{\mathsf{miss}}$ calculate $\hat{\boldsymbol{\beta}} = (\alpha, \boldsymbol{\beta}_u, \boldsymbol{\beta}_z)$ and $\hat{\gamma}$ as done above, and stack the relevant $\hat{\gamma}$ row vectors into the matrix $\hat{\Gamma}$. Now dispense with the observed $\mathbf{y}$ values in the subset being processed so that all of the $\mathbf{y}_i$ are missing. Estimate all of the now missing $\mathbf{y}$ values from:

$$\underset{1\times1}{\hat{y}_i} = \underset{1\times1}{\alpha} + \underset{(1\times k_u)}{\hat{\gamma}} \underset{(k_u\times1)}{\hat{\boldsymbol{\beta}}_u} + \underset{(1\times k_z)}{\mathbf{z}_i} \underset{(k_z\times1)}{\hat{\boldsymbol{\beta}}_z}. \tag{14}$$

Finally the subsets are recombined to have a fully filled-in dataset of the original size. Shao shows that this procedure produces asymptotically unbiased coefficient estimates. Of course the limiting restriction is that this works only for linear modeling.

Missing data can also be computed in a Bayesian context using the Markov chain Monte Carlo process. Missing data are treated as unknown parameters and estimated alongside the regular unknown parameters. This means that a distribution for the missing data must be specified so that the sampler can draw from it on each iteration of the simulations. In practice a normal distribution works in many circumstances. A canonical example comes from data from Carlin and Gelfand (1991) are the length and age of 27 dugongs (a type of sea cow) captured in the ocean off of Queensland, which is a toy example used by many authors. The data can be specified in a `jags` formatted file as:

```
dugong <- list(age = c(1.0,  1.5,  1.5,  1.5,  2.5,  4.0,  5.0, 5.0,   7.0, 8.0,
                       8.5,  9.0,  9.5,  9.5, 10.0, 12.0, 12.0, 13.0, 13.0, 14.5,
                      15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5, 35.0, 40.0),
               length = c(1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47,
                          2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43,
                          2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, 2.72, 2.57,
                          NA,   NA),
               NUM.DUGONGS = 29)
```

The model commonly specified for these data is a nonlinear growth specification $\mu_i = \alpha - \beta \times \gamma^x$, which is specified with the following `jags` code run in `R`:

```
dugongs.linear <- function() {
    for(i in 1:NUM.DUGONGS) {
        length[i] ~ dnorm(mu[i], tau)
        mu[i] <- alpha - beta * pow(gamma,age[i])
    }
    mu35 <- alpha-beta*pow(gamma,35) # 35 IS AGE OF MISSING 1
    mu40 <- alpha-beta*pow(gamma,40) # 40 IS AGE OF MISSING 2

    alpha  ~ dunif(0, 10)
    beta   ~ dunif(0, 10)
    tau    ~ dgamma(1, 0.1)
    sigma  <- 1/sqrt(tau)
    gamma  ~ dunif(0.0, 1.0)
    miss35 ~ dnorm (mu35, tau) # PRIOR ON MISSING 1
    miss40 ~ dnorm (mu40, tau) # PRIOR ON MISSING 2
}
write.model(dugongs.linear,"Class.Multilevel/dugong.linear.jags")
params <- c("mu", "mu35", "mu40", "alpha", "beta", "gamma")
dugong.out <- jags(data = dugong, parameters.to.save = params,
                 model.file = "Class.Multilevel/dugong.linear.jags",
                 n.iter = 100000, n.burnin = 50000, n.thin = 1)
```

Notice that the two missing values for the outcome variable length, labeled `mu35` and `mu40`, are given normal prior distributions and inside the sampling loop are predicted with the latest updates of α, β, and γ. When the results from the Gibbs sampler output are summarized, we get the posterior means (standard errors) for the two values: $2.639(0.082)$, $2.644(0.088)$. One advantage of the Bayesian stochastic simulation is that uncertainty in the standard parameters and uncertainty in the missing data are simultaneously estimated and are conditional on each other. This is, in a sense, the oppositive of the standard multiple imputation process where there are two steps (imputation and statistical modeling) and the combined uncertainty is accounted for additively in (6)

3.2 A General Framework for Ignorable Missing Outcomes

This section gives a typology of ignorable missingness in $\mathbf{y}$ with the appropriate solutions. To consider the full structure of cases of missingness we now extend the Rubin (1979) matrix structure $Z_{mis} = (X_{mis}, y_{mis}), Z_{obs} = (X_{obs}, y_{obs})$, and causal mechanism ϕ. A generalized linear model specification for an arbitrary link function is expressed as follows:

$$E(\mathbf{Y}) = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \mathbf{X}_{ign}\boldsymbol{\Gamma} + \mathbf{X}_{non}\boldsymbol{\Delta}), \tag{15}$$

where $\mathbf{X}_{obs}$ is the matrix of fully observed column data, $\mathbf{X}_{ign}$ is matrix of ignorable missing column data, and $\mathbf{X}_{non}$ is the matrix of non-ignorable missing column data, and $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Delta}$ are the corresponding estimated coefficient vectors. Therefore there are six different typologies for ignorable missing $\mathbf{y}$, each with the following six prescription where $\tilde{\mathbf{X}}$ denotes a matrix with imputed values.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{obs}, \phi):}$ estimate $\hat{\mathbf{y}} = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \mathbf{X}_{ign}\boldsymbol{\Gamma})$.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{ign}, \phi):}$ run imputation on $[\mathbf{X}_{obs}, \mathbf{X}_{ign}]$, then estimate $\mathbf{y}_{mis}$ with $\hat{\mathbf{y}} = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \tilde{\mathbf{X}}_{ign}\boldsymbol{\Gamma})$.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{non}, \phi):}$ cannot impute $\mathbf{y}_{mis}$.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{obs}, \mathbf{X}_{ign}, \phi):}$ run imputation on $[\mathbf{X}_{obs}, \mathbf{X}_{ign}]$, then estimate $\mathbf{y}_{mis}$ with $\hat{\mathbf{y}} = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \mathbf{X}_{ign}\boldsymbol{\Gamma})$.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{obs}, \mathbf{X}_{non}, \phi):}$ assume $\mathbf{X}_{obs}$ contains information about $\mathbf{X}_{non}$, run imputation on $[\mathbf{X}_{obs}, \mathbf{X}_{ign}]$, then estimate $\mathbf{y}_{mis}$ with $\hat{\mathbf{y}} = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \mathbf{X}_{ign}\boldsymbol{\Gamma})$.

- $\boxed{p(\mathbf{y}_{mis}|\mathbf{X}_{ign}, \mathbf{X}_{non}, \phi):}$ cannot impute $\mathbf{y}_{mis}$.

The selected procedure above still needs to be repeated with the results combined in the standard manner.

The two problematic cases are $p(\mathbf{y}_{mis}|\mathbf{X}_{non}, \phi)$ and $p(\mathbf{y}_{mis}|\mathbf{X}_{ign}, \mathbf{X}_{non}, \phi)$ since missing outcome variable values are conditional on non-ignorable missing explanatory variable values. Since these values cannot be recovered and impose upon the missing outcome variable values, then there is direct way to effectively impute them. There are, however, indirect ways to solve this problem. If researchers are willing to impose restrictions for specific settings under the assumption that a combination of algorithm and other $\mathbf{X}$ information informs the nonignorable missingness. For instance, Little and Rubin (2002) provide nonignorable methods for cell means models, partially classified contingency models, and more. These are all model-based in the sense that an additional model for the missing process is fitted and predictions for the missing values are produced.

4 Multiple Imputation, Then Prediction

The prescriptions employing $\hat{\mathbf{y}} = g^{-1}(\mathbf{X}_{obs}\boldsymbol{\beta} + \tilde{\mathbf{X}}_{ign}\boldsymbol{\Gamma})$ above, when implemented via multiple imputation, are hereafter referred to as 'multiple imputation then prediction' (MIP) and involve the following steps:

1. Multiply impute [$\mathbf{y}_{mis}$,$\mathbf{X}_{ign}$] using $\mathbf{y}$ and all relevant $\mathbf{X}$'s (even if they are not in the final model).

2. Estimate M instances of $\mathbf{y}_{obs} = g^{-1}(\mathbf{X}_{obs}\hat{\boldsymbol{\beta}} + \tilde{\mathbf{X}}_{ign}\hat{\boldsymbol{\Gamma}})$, again using all relevant $\mathbf{X}$'s.

3. Calculate $\hat{\mathbf{y}}$ for each $\mathbf{y}_{mis}$, and replace all $\mathbf{y}_{mis}$ values with these predictions. Term these $\mathbf{y}_{MIP}$.

4. Conduct M final analyses using $\mathbf{y}_{MIP}$ and the subset of $(\tilde{\mathbf{X}}_{ign}, \mathbf{X}_{obs})$ of interest and combine.

The model combination component to step 4 follows a similar approach to that of standard multiple imputation (MI) techniques (Rubin 1987). For the regression applied to the mth imputed data set, let $\hat{\theta}_{MI}^{(m)}$ be one's parameter estimate, and $\sqrt{\hat{W}_{MI}^{(m)}}$ be that parameter's standard error. MIP parameter estimates are then obtained in an identical manner to standard MI: by averaging one's point estimates across the M imputed data sets according to equation 3 above. However, under MIP, we must adjust the standard MI approach to variance combination, so as to penalize our MIP technique for its understatement of variability in step 3 above. To do so, we multiply the MI formula for within imputation variance (equation 4) by the penalty parameter $(1+p)) / \left(1 - p^{\frac{1}{2}}\right)$, to yield the new within variance quantity:

$$W_M = \frac{1+p}{M(1-p^{\frac{1}{2}})} \sum_{m=1}^{M} \Sigma_m. \tag{16}$$

where p is the proportion of total $\mathbf{y}$ observations that are missing in the original dataset. The penalty accordingly accounts for the share of missing $\mathbf{y}$ observations, and thus the understatement of variability in our predicted $\mathbf{y}$'s. It also $\rightarrow \frac{1}{1}$ as the proportion of missing $\mathbf{y}$'s $\rightarrow 0$, thereby ensuring that our MIP within imputation variance formula collapses to the standard MI within imputation variance formula when applied to data with no missingness on $\mathbf{y}$. The remaining MIP formulas for between imputation variance, total variance, and degrees of freedom are identical to those of MI (equations 5-7 above).

5 Monte Carlo Simulations

To evaluate the performance of the MIP technique in finite samples, we conduct three Monte Carlo experiments. These experiments respectively examine three different types of missingness in both $\mathbf{X}$ and $\mathbf{y}$ (described below), in each case comparing MIP to a number of alternative regression approaches that could plausibly be used in situations where both $\mathbf{y}_{mis}$ and $\mathbf{X}_{mis}$ exist. The three experiments specifically compare MIP to (i) the analysis of the data before missingness is introduced, (ii) analysis of only the complete cases in our dataset (CC, i.e., case-wise deletion), (iii) multiple imputation, then deletion (MID), (iv) a strategy that multiply imputes $\mathbf{X}_{ign}$ using $(\mathbf{X}_{ign}, \mathbf{X}_{obs})$ but not $\mathbf{y}$ (i.e., MI$_\mathbf{X}$; which accordingly omits all cases with $\mathbf{y}_{mis}$ from one's final analysis due to case-wise deletion), and (v) full multiple imputation of both $\mathbf{y}_{mis}$ and $\mathbf{X}_{ign}$ (MI).

Our Monte Carlo experiments assign the number of $sims = 1,000$ and then compare the performance of each of the multiple imputation approaches listed above within finite samples of $N = 500$. All simulations were undertaken in `R` using `mice` for multiple imputation, `glm(...,gaussian(link = "identity"))` for estimation of $\hat{\mathbf{y}}$, and `glm.mids(...,gaussian(link = "identity"))` for final model estimation and combination. Each simulated dataset included two core explanatory variables $(\mathbf{X}_1, \mathbf{X}_2)$, which were drawn from a standard bivariate normal distribution with correlation coefficient ρ_{12}. The outcome variable $\mathbf{y}$ was then generated according to:

$$\mathbf{y} = \alpha + \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + e, \tag{17}$$

where $e \sim$ Normal$(0, \sigma_e^2)$ and $(\alpha, \beta_1, \beta_2)$=$(1, 1, 1)$. To parallel real-world social science circumstances, we also generated an auxiliary variable ($\mathbf{X}_3$), which was used in all imputations but withheld from the final estimation model, according to:

$$\mathbf{X}_3 = \mathbf{y}_s \gamma_s + e_s, \tag{18}$$

where $\mathbf{y}_s$ is a standardized version of $\mathbf{y}$, $e \sim$ Normal$(0, 1 - \gamma_s^2)$, and $\gamma_s = 0.5$. Following von Hippel (2007, pp. 101), Experiments 1-3 separately imposed three patterns of ignorable missingness upon $\mathbf{y}$ and $\mathbf{X}_2$: (i) *MCAR*, wherein $\mathbf{y}$ and $\mathbf{X}_2$ were each independently deleted with constant probability p; (ii) *coordinated missingness* (*COR*), where $\mathbf{y}$ and $\mathbf{X}_2$ were independently deleted with probability $2p\Phi(\mathbf{X}_1)$, and (iii) *complementary missingness* (*COMP*), where $\mathbf{y}$ was deleted with probability $2p\Phi(-\mathbf{X}_1)$ and $\mathbf{X}_2$ was deleted with probability $2p\Phi(-\mathbf{X}_1)$. $\Phi(\cdot)$ denotes the cumulative standard normal density. The imputation approaches under evaluation then imputed 5 copies of each incomplete dataset under a multivariate normal model.

In addition to varying the pattern of missingness, each of the Monte Carlo experiments also independently varied three additional factors. First, we varied the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$, $\rho_{1,2} = (.2, .5, .8)$. Second, we varied the proportion of variance explained in $\mathbf{y}$, $R^2 = (.2, .5, .8)$, by setting $\sigma_e^2 = 2(1 - R^2)(1 + \rho_{1,2})/R^2)$. Third, we varied the proportion of $\mathbf{y}$ and $\mathbf{X}_2$ observations that were deleted, $p = (.2, .4)$. These choices are consistent with past simulation studies of multiple imputation as applied to data with missing outcomes (von Hippel 2007). Altogether, our various experiments and experimental conditions generated 162 possible combinations of Monte Carlo results.

After implementing an imputation approach, each simulation then estimated a final model of interest: $\mathbf{y} = g^{-1}(\hat{\alpha} + \mathbf{X}_1\hat{\beta}_1 + \tilde{\mathbf{X}}_2\hat{\beta}_2)$. To make comparisons between MIP and our alternative approaches, we retained the point estimates, standard errors, and nominal 95% confidence intervals from these final models, and additionally calculated the following three quantities. First, and for each experiment and estimation approach, we calculated the mean square errors (MSEs) of our parameter estimates according to the formula:

$$MSE_{\hat{\theta}} = \frac{1}{sims} \sum_{i=1}^{sims} (\hat{\theta}_i - \theta_i)^2 \tag{19}$$

where $sims$ is the number of simulations (i.e., 1,000), $\hat{\theta}$ is a given approach's parameter estimate, and θ is the parameter value of interest. Second, and due to the adverse influence of outliers within our MSEs, we also derive and present the median absolute errors (MAEs) of our parameter estimates, $MAE_{\hat{\theta}} = Me\left(|\hat{\theta}_i - \theta|\right)$. Finally, and again for each estimation approach, we report and discuss the empirical coverage probabilities (CPs) of our estimates' confidence intervals, which correspond to the proportion of simulations for which a parameter estimate's 95% confidence intervals include the parameter of interest.

5.1 Results

Our Monte Carlo results appear in Table 1. For summary purposes, this table focuses on the 20% missingness case, and collapses our results across (i) the proportion of variance explained in $\mathbf{y}$ (i.e.,

$R^2 = .2, .5, .8$), and (ii) the correlation between $\mathbf{X}_1$ and $\mathbf{X}_2$ (i.e., $\rho_{1,2} = .2, .5, .8$), as our primary findings do not dramatically vary across either of these latter factors. For each experiment (and thus type of missingness described above, i.e., *MCAR, COR, COMP*), Table 1 then reports separate summary quantities for each parameter (α, β_1, β_2), across all estimation approaches evaluated (in columns).

Turning to the Experiment 1 (*MCAR*), we find that the MIP parameter estimates are clearly superior to those obtained after case-wise deletion (CC)—and are also superior (though sometimes only marginally) to the estimates derived from our three alternative imputation approaches (MI, MID, and $\text{MI}_{\mathbf{X}}$)—in terms of both bias and coverage. For instance, the Experiment 1 MIP estimates of α and β_1 consistently exhibit noticeably lower MAEs and MSEs than do the alternative approaches to handling missingness that are examined here, and are comparable to those achieved under our pre-deletion (Full) analysis. Further, the MIP 95% CIs for α and β_1 include our parameters of interest roughly 98% of the time, which is superior to each alternative estimation approach—although this comes at the cost of slightly larger standard errors under MIP.

Regarding β_2, which corresponds to the explanatory variable that included missingness, we again find in Experiment 1 that MIP exhibits a superior MAE to all alternative approaches, and a comparable CP to these alternatives. In addition, MIP also obtains a superior β_2 MSE to all alternate approaches, aside from $\text{MI}_{\mathbf{X}}$. Regarding the latter β_2 comparison, whereas $\text{MI}_{\mathbf{X}}$ does outperform both MIP and our other imputation approaches in MSE, it performs dramatically worse than *all* other alternatives in terms of both β_2 CP (where its CP of 80% is consistently over 10 percentage points lower than the CPs exhibited by all other approaches), and MAE (where the $\text{MI}_{\mathbf{X}}$ MAE is roughly 40% lager than the approaches that *do* impute the outcome variable in some form, including MIP). Hence, Experiment 1 suggests that while MIP generally exhibits less bias and better coverage than alternate imputation or case-wise deletion strategies when one's outcome variable is *MCAR*, approaches that case-wise delete missing observations in one's outcome variable, namely $\text{MI}_{\mathbf{X}}$ and CC, are by far the least optimal strategies under these circumstances.

Experiment 2 (*COR*) provides similar results to Experiment 1 (*MCAR*). MIP again exhibits advantages with respect to MAE, MSE, and CP over our alternative imputation approaches and case-wise deletion (CC). These advantages are most notable for (i) comparisons between MIP and CC, wherein the MSEs and MAEs obtained under CC are substantially larger than either MIP or alternate imputation approaches and (ii) comparisons between MIP and $\text{MI}_{\mathbf{X}}$ in the case of β_2, where $\text{MI}_{\mathbf{X}}$ yields dramatically worse CPs and MAEs than MIP, MI, MID, or even CC. There are also several instances where alternate imputation techniques exhibit equal or slightly less bias than does MIP, such as in the case of MID's α MAE relative to the comparable MAE obtained under MIP. However, these differences are very minor, and do not systematically favor any single imputation alternative over MIP across all parameters or summary quantities. Rather, among the five approaches that we applied to data with *COR* missingness, MIP appears to most systematically outperform alternative methods, albeit with several of these advantages being slight.

Experiment 3 (*COMP*)) is summarized in the bottom third of Table 1. Similar to Experiments 1 and 2, we find in Experiment 3 that the MIP approach almost uniformly outperforms our four alternate missing data strategies in terms of both MAE and MSE. The one notable exception is the MSE for β_2 under $\text{MI}_{\mathbf{X}}$, which is noticeably smaller that those of MI, MIP, and MID. As before, this advantage in

Table 1: Differences Between Imputation Approaches when N=500 and M=5

		Method					
Type of Missingness	Parameter	Full Data	CC	MI	$\text{MI}_{\mathbf{X}}$	MID	MIP
	$\alpha = 1$						
	$\hat{\alpha}$	1.002	1.002	1.003	1.003	1.002	1.002
	SE	0.235	0.294	0.259	0.266	0.265	0.323
	MSE	0.105	0.166	0.126	0.131	0.132	0.122
	MAE	0.073	0.094	0.082	0.085	0.084	0.080
	CP	0.949	0.950	0.952	0.958	0.950	0.985
	$\beta_1 = 1$						
	$\hat{\beta}_1$	0.998	1.005	0.994	0.996	0.996	0.997
MCAR	SE	0.357	0.447	0.392	0.404	0.401	0.489
	MSE	0.233	0.371	0.289	0.299	0.299	0.281
	MAE	0.109	0.135	0.121	0.123	0.122	0.121
	CP	0.953	0.953	0.951	0.960	0.954	0.986
	$\beta_2 = 1$						
	$\hat{\beta}_2$	0.994	0.992	0.979	0.85	0.985	0.982
	SE	0.357	0.447	0.442	0.454	0.454	0.532
	MSE	0.247	0.389	0.380	0.313	0.387	0.373
	MAE	0.115	0.142	0.144	0.202	0.144	0.140
	CP	0.949	0.944	0.939	0.802	0.942	0.977
	$\alpha = 1$						
	$\hat{\alpha}$	1.001	0.998	0.999	0.999	0.999	0.999
	SE	0.235	0.300	0.261	0.268	0.266	0.323
	MSE	0.108	0.177	0.136	0.137	0.138	0.132
	MAE	0.073	0.094	0.086	0.085	0.083	0.084
	CP	0.949	0.949	0.950	0.956	0.951	0.982
	$\beta_1 = 1$						
	$\hat{\beta}_1$	0.997	1.001	0.995	0.996	0.996	0.997
COR	SE	0.356	0.451	0.394	0.406	0.404	0.49
	MSE	0.240	0.387	0.291	0.307	0.307	0.288
	MAE	0.108	0.140	0.124	0.126	0.124	0.123
	CP	0.949	0.950	0.948	0.955	0.95	0.985
	$\beta_2 = 1$						
	$\hat{\beta}_2$	1.000	1.007	0.99	0.865	0.996	0.994
	SE	0.356	0.444	0.442	0.452	0.451	0.531
	MSE	0.234	0.364	0.353	0.304	0.363	0.347
	MAE	0.112	0.141	0.137	0.190	0.137	0.135
	CP	0.950	0.949	0.949	0.818	0.952	0.981
	$\alpha = 1$						
	$\hat{\alpha}$	1.001	0.999	1.000	0.998	0.998	0.999
	SE	0.235	0.296	0.261	0.268	0.266	0.324
	MSE	0.107	0.168	0.131	0.135	0.135	0.126
	MAE	0.072	0.09	0.084	0.086	0.085	0.081
	CP	0.952	0.954	0.949	0.958	0.952	0.985
	$\beta_1 = 1$						
	$\hat{\beta}_1$	1.004	1.011	1.004	1.009	1.009	1.008
COMP	SE	0.356	0.453	0.395	0.406	0.404	0.491
	MSE	0.240	0.389	0.296	0.311	0.311	0.29
	MAE	0.108	0.138	0.127	0.127	0.126	0.126
	CP	0.954	0.948	0.948	0.956	0.950	0.983
	$\beta_2 = 1$						
	$\hat{\beta}_2$	0.995	0.997	0.982	0.844	0.986	0.985
	SE	0.357	0.450	0.444	0.456	0.457	0.534
	MSE	0.237	0.381	0.370	0.309	0.378	0.362
	MAE	0.112	0.138	0.137	0.204	0.139	0.135
	CP	0.948	0.951	0.949	0.790	0.954	0.981

$\mathsf{MI}_{\mathbf{X}}$ is offset by its comparably poorer performance in MAE and CP for β_2, where the latter quantity in this case suggests that $\mathsf{MI}_{\mathbf{X}}$ 95% confidence intervals only include β_2 within 79% of our simulations, compared to CP's of 94.9%-98.1% for MI, MID, and MIP. More generally, we can also note that our MIP CP's are consistently superior to those of CC, MI, MID, and $\mathsf{MI}_{\mathbf{X}}$ in Experiment 3, albeit with slightly standard errors than these alternatives. Under cases of *COMP* missingness, we can thus conclude that MIP, and to slightly lessor degrees MI and MID, provide one with parameter estimates that exhibit consistently better accuracy and higher coverage than do approaches such as CC or $\mathsf{MI}_{\mathbf{X}}$.

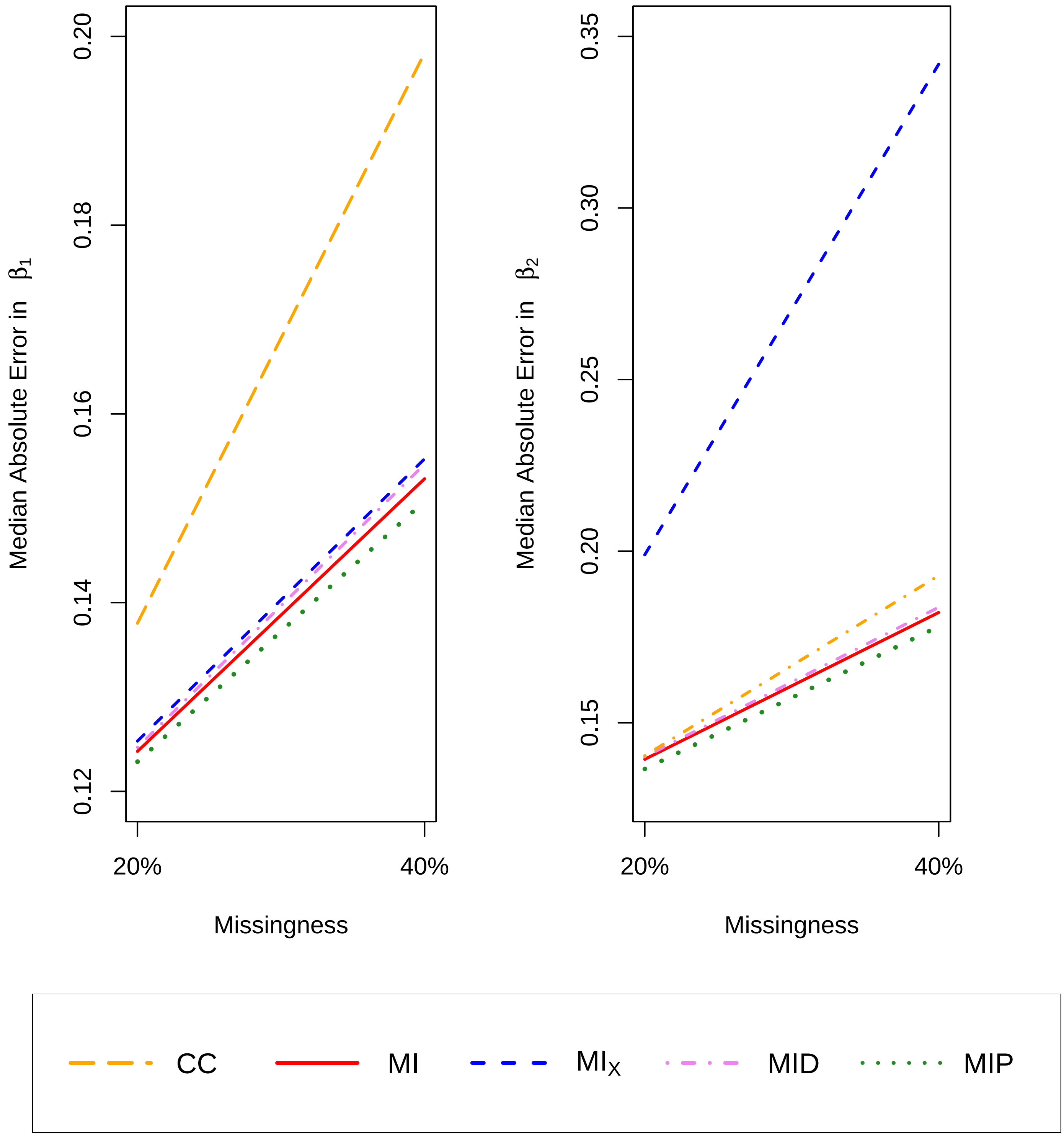
The results discussed above pertain to the case of 20% missing data. To further summarize our findings, we now extend and compare this analysis to instances of 40% missingness. To do so, we combine our experimental results (i.e., *COMP*, *COR*, and *MCAR*) and (re)calculate each estimation approach's MAEs separately for these cases of 20% and 40% missingness.[2] These plots appear in Figure 1 and reinforce several of the conclusions drawn above. Beginning first with β_1, we find that—while all estimates generally worsen in MAE as missingness is increased from 20% to 40%—our MIP point estimates are consistently closer to β_1 than are the estimates obtained under MI, MID, and $\mathsf{MI}_{\mathbf{X}}$. In this respect, MI performs second best, followed by MID, and then $\mathsf{MI}_{\mathbf{X}}$. Case-wise deletion (CC) then exhibits by far the highest levels of bias in $\hat{\beta}_1$, with MAEs that are several orders of magnitude higher than our imputation approaches within both the 20% or 40% missingness cases. Hence, for parameter estimates associated with an explanatory variable that itself does not have missingness, case-wise deletion of missing observations on Y or other X's can dramatically increase bias. Imputation strategies help to minimize this problem, with MIP performing best when missingness in Y is present.

Turning to the MAEs for β_2, which appear on the right-hand side of Figure 1, we again find that raising the level of missingness from 20% to 40% increases bias for all estimation approaches. Even so, MIP provides lower levels of bias, relative to all alternative strategies, at either the 20% or 40% missingness levels. These MIP advantages are closely followed by those of MI and MID, whereas in this case, CC now performs fourth best, with an MAE that is slightly worse than MIP, MID, and MI in the 20% missingness case, and notably worse than these alternatives in the 40% missingness case. Most dramatically in the case of β_2, however, are the MAEs for $\mathsf{MI}_{\mathbf{X}}$, which suggest that this imputation strategy yields significantly higher levels of bias in $\hat{\beta}_2$ than does any of our alternatives, and increasingly so as the proportion of missingness is increased—with an MAE in the 40% missingness case that is roughly double that of even the second worst alternative. In sum, when missingness exists in both one's outcome and explanatory variable, estimates of the latter's effects are most accurate under MIP, MI, and MID, whereas choosing to only impute one's explanatory variables can potentially do dramatically more harm than even simple case-wise deletion.

Given missingness in one's outcome and explanatory variables along the lines analyzed here (i.e., MCAR or MAR) and the potential choice of using CC, MI, MID, MIP, or $\mathsf{MI}_{\mathbf{X}}$, which strategy provides the most accurate estimates? Experiments 1-3 suggest that in such instances, MIP will be preferable to the other four approaches (MID, MIP, CC, or $\mathsf{MI}_{\mathbf{X}}$) with respect to the accuracy of one's point estimates and empirical coverage, followed perhaps by MI and MID. In these contexts, using CC or $\mathsf{MI}_{\mathbf{X}}$ can be

[2]The plots and findings for α_1 are very similar to those of β_1, and are hence not reported and discussed here.

Figure 1: Median Absolute Errors (MAEs) in Point Estimates (Experiments 1-3)



much more problematic in terms of both variance and bias. Even so, it is worth emphasizing that the Monte Carlo findings and insights discussed here are dependent upon our specific choices of sample size, replicate datasets, missingness proportion, and parameters. While some additional experimentation on our part (as well as past simulation studies of similar phenomena) suggests that introducing more missingness than that employed here will favor the MIP over the others examined here even more dramatically than it does in Experiment 1-3, this remains an active area of research.

6 Conclusion

Here we address the questions of when and how to impute missing values in the outcome variable. Some scholars feel that one should not use various methods to fill-in such missingness, since the subsequent model specification is providing fitted values given levels of covariates and this would then be a redundant process. This is misguided since the imputation process differs from the modeling process, with the exception of EM methods which combine them. Political Scientists frequently encounter missingness in the response variable and we know that case-wise deletion produces biased coefficient estimates, even though it is the default with many empirical researchers. Our method improves on existing approaches in that it gives a systematic, principled set of steps that produce efficient results. Furthermore, the process is easily implemented in `R` and other languages. We hope that this work clears-up confusion about imputing outcomes *and* provides tools for doing so.

7 References

Cranmer, Skyler and Gill, Jeff. 2013. "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data." *British Journal of Political Science*, 43, 425-449.

Enders, Craig K., and Bandolos, Deborah L. 2001. "The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models." *Structural Equation Modeling*, 8, 430-457.

Honaker, James Honaker and King, Gary. 2010. "What to do about Missing Values in Time-Series Cross-Section Data", *American Journal of Political Science*, 54, 561–581.

Kalton, Graham and Kish, Leslie. 1984. "Some Efficient Random Imputation Methods." *Communications in Statistics - Theory and Methods*, 13, 1919–1939.

Lauritzen, Steffen L. 1995. "The EM Algorithm for Graphical Association Models with Missing Data." *Computational Statistics & Data Analysis*, 19, 191-201.

Little, Roderick J. A. 1988a. "Approximately Calibrated Small Sample Inference about Means from Bivariate Normal Data with Missing Values." *Computational Statistics & Data Analysis*, 7, 161–178.

Little, Roderick J. A. 1988b. "A Test of Missing Completely At Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association*, 83, 1198-1202.

Little, Roderick J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association*, 87, 1227–1237.

Little, R. J. A., and Rubin, D. B. 2002. *Statistical Analysis with Missing Data.* Second Edition. New York: John Wiley & Sons.

Mcknight, Patrick E., McKnight, Katherine M., Sidani, Souraya, and Figueredo, Aurelio Jose. 2007. *Missing Data: A Gentle Approach.* New York: The Guilford Press.

Molenberghs, Geert and Kenward, Michael G. 2007. *Missing Data in Clinical Studies.* New York: Wiley.

Rubin, Donald B. 1976. "Inference and Missing Data (with Discussion)." *Biometrika*, 63, 581-592.

Rubin, Donald B. 1977. "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys.", *Journal of the American Statistical Association*, 72, 538–543.

Rubin, Donald B. 1978. "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse." *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20–34.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Rubin, D. B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91, 473-489.

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall.

Schafer, Joseph L., and John W. Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7, 147-177.

Tsiatis, Anastasios A. 2010. *Semiparametric Theory and Missing Data.* New York: Springer.

van Buuren, Stef. 2007. "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." *Statistical Methods in Medical Research*, 16, 219-242.

von Hippel, Paul T. 2007. "Regression With Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data." *Sociological Methodology*, 37, 83-117.