Honest Inference with Missing Data*

Naijia Liu
naijial@princeton.edu

July, 2021

Abstract

This paper proposes a method to achieve valid model inference with missing data and a new way to evaluate performance of missing data inference, integrating missing value imputation step and model inference step. The proposed method uses a bias correction term to offset the difference between missing and complete observations. For a parametric regression model, the method relaxes the conventional "missing at random" assumption and distributional assumptions. Simulation and validation results show the superior performance of the proposed method, as compared with more conventional imputation methods. I conclude the paper with an application using a survey dataset showing a substantive change of model estimation before and after imputation.

*Replication code and R-package `CausalImpute` are available upon request.

1 Introduction

Social science data often contains missing values. Simply deleting any observation with missing value (listwise deletion) has potentially pernicious consequences (Lall, 2016). Its alternatives, multiple imputation (MI) [1] presents its own challenges, such as dependence on highly restrictive assumptions. In this paper, I develop a method to obtain honest confidence intervals with missing data. Model honesty defined by Li et al. (1989) states that a confidence interval should have nominal coverage, which I present formal definition later in equation (2).

Conventional multiple imputation methods separate imputation step from inference step, thus cause unreliable inference results when certain assumptions are violated. Selection model (Heckman, 1976) relies heavily on functional form, thus creates inflexible model dependency. This paper proposes a method that integrates imputation step with the inference step, producing more reliable inference results without making highly-stringent model-dependent assumptions. The proposed method is able to relax some of the conventional assumptions made by other imputation methods. As presented later in the paper, I relax the missing at random (MAR) assumption by incorporating the parametric model structure into the imputation process. I also relax distributional assumptions made by some existing methods by adopting machine learning procedure. Here, I advocate evaluating imputation methods on the basis of inference rather than accuracy, which is more fundamental to social science data analysis.

Table 1 presents a simple example with 1000 observations. This example demonstrates how it can be challenging for both listwise deletion and multiple imputation methods to produce a valid inference in linear regression setting. I also provide the intuition behind proposed method and compare it with existing methods.

[1] Multiple imputation (MI) refers to the procedure of replacing each missing value by a vector of more than two imputed values, each of which is a random draw from the predictive distribution of the missing values (Little and Rubin, 2014).

Let us start with a random variable $X_1 \sim N(3,1)$ and the outcome variable $Y = X_1 + e$ where the error term $e \sim N(0,1)$. Then we add another random variable $X_2 \sim N(-3,1)$, which is independent of both X_1 and Y. The complete dataset consists of Y, X_1 and X_2, with 1000 observations. I then make X_2 to be missing if $X_1 + X_1^2 \leq 12$. In other words, the probability of missing for X_2 is determined by X_1. When running the linear regression as $Y = \beta_1 X_1 + \beta_2 X_2$, researchers aim at recovering β_1 and β_2. In this case, with complete data one should get $\beta_1 = 1$ and an insignificant β_2 at around zero.

As demonstrated in table 1, the first column confirms the above claim, with the complete dataset. The second column shows that, by deleting observations with missing (with 470 observations remaining), we get biased estimates for β_1. This happens because missingness is not completely random. Rather, the probability of missing for X_2 is a function of X_1. By excluding observations with $X_1 + X_1^2$ below a certain threshold, we no longer have a representative sample of the original dataset.

The third column shows the regression result out of a conventional multiple imputation method `mice` (White et al., 2011). The algorithm is able to fill in all the missingness in the dataset, by assuming missing at random which will be discussed in greater detail in section 1.1. However, with the imputed data, we wrongfully induce significance to the variable X_2 as shown in table 1. Assuming missing at random, `mice` imputes missing values based on the conditional distribution of all other observed variables. The imputed values of X_2 are thus determined by both X_1 and Y. When we first regress X_1 onto Y, the error terms will be correlated with the imputed values of X_2 and hence a wrongfully significant coefficient.

Heckman (1976) proposed models to tackle with selection bias due to missing dependent variables. The proposed method is different from the selection model, since it deals with missing independent variables. But both methods focus on the selection bias when data is censored.

The intuition of the proposed method is to add a bias-correction term to the imputation result. I demonstrate this in the forth column of table 1. I first mean-impute the missingness

in $X_2 = \bar{X}_2$ as place holders, then I added an extra column in the dataset called bias-correct term to offset the difference between missing and complete observations. Again, the bias-correction term η is determined by the missing mechanism that X_2 goes missing if $X_1 + X_1^2 \leq 12$. Bias correction term is then added to cancel out this correlation.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 \hat{\eta} + e$$

In other words, the bias correction term is to make sure that observations with missing X_2 and with observed X_2 share same expected value of Y. In most of the cases, one does not know the exact mechanism, and thus impractical to fill in the bias-correction term as we did in this toy example. This paper in later section proposes a new method to estimate the bias correction term. And I also provide further proof that under certain assumptions, this method will produce unbiased estimates of coefficients, or β_1 and β_2 in the example above.

I start the paper with a review on missing data literature in section 1.1. Then, I present the model setup and asymptotic properties in section 2. I present simulation results and a validation study using real data in section 3. Finally, I conclude the paper with an application on Gilens (2001) by imputing all missingness in a survey dataset.

1.1 Existing literatures

Missing data imputation has received a great amount of attention from social scientists. Little and Rubin (2014) proposed that there were three possible missing mechanisms. Missing completely at random (MCAR) indicates that the missingness does not depend on the the observed or missing data.[2] Under MCAR, conducting listwise deletion will not induce any bias. However, MCAR is a rather strong assumption for researchers to make. Furthermore, King et al. (1998) showed that MCAR does not apply to most datasets of interests

[2] $f(M|X,\phi) = f(M|\phi), \quad \forall X, \phi$.

Table 1

	Dependent variable: Y			
	Complete	Listwise	MICE	Proposed
X1	1.004***	0.836***	0.900***	1.001***
	(0.033)	(0.077)	(0.035)	(0.051)
X2	0.029	−0.038	−0.238***	0.086
	(0.031)	(0.043)	(0.034)	(0.066)
Constant	0.067	0.528*	−0.294***	0.234
	(0.139)	(0.301)	(0.106)	(0.242)
Observations	1,000	480	1,000	500
R^2	0.485	0.205	0.509	0.456
Adjusted R^2	0.484	0.202	0.508	0.453
Residual Std. Error	1.002 (df = 997)	0.956 (df = 477)	0.979 (df = 997)	1.033 (df = 496)
F Statistic	468.817***	61.510***	515.761***	138.710***
	(df = 2; 997)	(df = 2; 477)	(df = 2; 997)	(df = 3; 496)
Note:				*p<0.1; **p<0.05; ***p<0.01

to social scientists, and that listwise deletion will in general lead to biased inference.

Missing not at random (MNAR) indicates that the missingness depends either on the missing values in the data X, or unobserved variables.[3] The response in the literature is to resort to data imputation, but the imputed values are contingent on both the model of missingness and on the underlying model of inference. Pepinsky (2018) addressed the differences in performance between listwise deletion and multiple imputation methods (under missing not at random) for regression, in which he "recommend caution when comparing the results from multiple imputation and listwise deletion, when the true data generating process is unknown". The scope of the proposed method includes MNAR under a certain condition, which I will discuss in greater detail in assumption 2.1.

Missing at random (MAR) indicates that missingness depends only on the components of X that are observed.[4] The proposed method can solve imputation problems under MAR, where missingness in a variable of an observation can be fully explained by other observed covariates. MAR is a more reasonable assumption to make compared to MCAR, while it presents a more tractable problem then does MNAR.[5]

There are several existing multiple imputation methods that are popular among social scientists. Multiple imputation by chained equations (`MICE`) is a repetitive algorithm that conducts regressions on each values with missing values (White et al., 2011). This method is problematic when the conditional distribution of outcome Y on covariates is wrongly specified. Below I present a toy example to show that `MICE` potentially suffers from overfitting problem, see table 1. The missing dataset will consist of column Y, X_1, and X_2(MAR). Notice that the missingness in X_2(MAR) is taken out of X_2, following missing at random. It is easy for the regression model to recognize that $Y = \beta_1 X_1 + \beta_2 X_2$ with $\beta_1 = \beta_2 = 1$, and thus impute the missing observations following $X_2 = Y - X_1$ (see "Imputed X_2" column). However, the true data might not follow this rule. Table **??** shows the regression result

[3] $f(M|X, \phi) = f(M|X_{\text{mis}}, \phi), \quad \forall X_{\text{mis}}, \phi.$
[4] $f(M|X, \phi) = f(M|X_{\text{obs}}, \phi), \quad \forall X_{\text{mis}}, \phi$.
[5] MCAR can be seen as a special case of MAR.

from imputed dataset and the true dataset, with wrongfully induced statistical significance for imputed "X2" [6]. I provide more evidence of this issue in the simulation section (section 3).

Building on the concept of multiple imputation, King et al. (2001); Honaker and King (2010) developed a faster and easier to use software `Amelia`, which implements an Expectation Maximization (EM) algorithm to impute missing values. The `Amelia` algorithm assumes a joint multivariate normal distribution for the data and implement an EM algorithm until the sufficient statistics of the data matrix converge. However, as the simulation results show, when a dataset contains non-continuous variables (such as binary and categorical values), one may not get a perfect multivariate normal distribution, and thus less ideal imputation performance of the method.

Another school of thought on missing data adopts a causal inference perspective for the problem. By modeling on the probability of observations being missing, Bang and Robins (2005) introduced a doubly-robust estimator for missing values. Based on the idea of inverse probability weighting (IPW) method in causal inference, they adopted a sequential regression method to compute the doubly-robust estimator. Sun and Tchetgen (2014) also proposed a similar method using inverse propensity weighting to impute under the missing at random situation. One constraint of a IPW method is that estimated propensity scores of 0 or 1 cannot be inverted and so we cannot construct the weights. Seaman and White (2013) pointed out that, compared to MI, IPW methods is often less efficient and require more assumptions on the missing pattern.

The proposed method of this paper shares similarity with the Bang and Robins (2005)'s method, but also differs in fundamental ways: Instead of the propensity of being missing, I focus on the heterogeneous difference in mean between the missing and observed. By

[6] This table presents the most extreme result of a false imputation. I then applied the toy dataset with missingness to `MICE` in `R`. One still gets biased regression results, both in terms of point estimate and uncertainty estimation. When set imputation parameter m greater or equal to 500, i.e. to take average of more than 500 imputed datasets, the regression model finally gets closer to truth, still with bias in point estimations.

leveraging the parametric regression structure, I am able to adopt the machine learning approach to calculate heterogeneous treatment effects under the ignorability condition. [7]

1.2 Honesty of model inference

Consider a parametric model with an additive error with conditional mean of zero:

$$Y = f_{\theta_0}(X) + e, \quad E(e|X) = 0. \tag{1}$$

Y is the outcome variable, X is the covariate matrix and one is interested in making inference between the two by estimating the function f_{θ_0}.

Under equation (1), researchers are interested in both the point estimate $\hat{\theta}_n$ and uncertainty measure around θ_0. We want an honest inference of model uncertainty in the sense that the confidence interval covers the truth with a probability equal to or higher than the specified probability $1-\alpha$. Mathematically, this is called asymptotic nominal coverage (also model honesty) by Li et al. (1989).

$$\lim_{n\to\infty} \inf_{f\in F} P\left(\|\hat{\theta}_n - \theta_0\|_n \leq s_n\right) \geq 1-\alpha \tag{2}$$

Li et al. (1989) also show that for a nonparametric model with convergence faster than $n^{1/4}$, one should be able to achieve model honesty. Furthermore, under the set up in equation 1, Van der Vaart (2000) shows a proof for asymptotic normality in theorem 2.1: Under mild condition one can have:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N\left(0, V_{\theta_0}^{-1} P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T V_{\theta_0}^{-1}\right) \tag{3}$$

where m_{θ_0} is an objective function for the model and $\dot{m}$ is the first derivative matrix while

[7] Here, I used `causal tree` (Athey and Imbens, 2016) to calculate this value of interest. `causal tree` is an algorithm that provides honest inference for treatment effect by partitioning data into subpopulations.

V_{θ_0} is the second derivative matrix of function m. In this context, m function is the sum of squared errors. The asymptotic normality of parametric regression estimators makes sure that the confidence intervals achieve asymptotic nominal coverage.

However, when missing data occurs in the dataset, one can only observe an incomplete list of covariates for each observation X_{obs}. The parametric regression result may be biased, i.e $\theta \neq \theta'$:

$$Y = \begin{cases} f_\theta(X) + e, & E(e|X) = 0 \quad \text{if fully observed} \\ f_{\theta'}(X_{\text{obs}}) + e', & E(e'|X_{\text{obs}}) = 0 \quad \text{if missing covariate} \end{cases} \tag{4}$$

When one gets biased estimation of θ, both in terms of point estimate and / or inference, model honesty defined in equation 2 will be violated.

2 The Proposed Method

I start this section with an introduction to notations and assumptions for the proposed method. Then I explain the general setup as well as implementation of the method by step. In section 2.3, I present a proof for asymptotic normality for the proposed estimator.

2.1 The setup

Let X denote a p dimensional vector of covariates and Y denote the outcome. Suppose that researchers are interested in estimating the relationship between X and Y using the following regression model,

$$Y = f_\theta(X) + e, \quad E(e \mid X) = 0 \tag{5}$$

where $f : \mathbb{R}^p \to \mathbb{R}$ is a parametric function indexed by parameter θ. For example, in the linear regression, f can be written as $f_\theta(X) = X^T\theta$.

Now, we consider a scenario where some of the covariates have missing values. Let

M_{ij} denote a binary missing indicator that takes 1 if the value of variable j is missing for unit i and takes 0 otherwise. I also let X_{obs} denote the observed covariates and X_{mis} as the missing covariates. By construction, we have that $X_i = (X_{i,\text{obs}}, X_{i,\text{mis}})$. Finally, X_{imp} denotes the imputed value of the missing covariates.

The goal is to make statistical inference on θ using X_{obs} and X_{imp}.

Assumption 2.1 (Missingness independent of outcome variable)**.** *Conditional on full set of covariates, the outcome variable is independent of missingness.*

$$Y \perp\!\!\!\perp M \mid X \tag{6}$$

This assumption makes sure the ignorability assumption hold for the estimation of heterogeneous treatment effect in the later step. This is different from the MAR assumption that most MI methods adopt: $M|X_{\text{obs}} \perp\!\!\!\perp X_{\text{mis}}$. Since we don't get to observe the missing part of the dataset X_{mis}, this assumption is not verifiable (just like MAR). However, this means the proposed method potentially works for some cases of MNAR, since the assumption allows $M|X_{\text{obs}} \not\perp\!\!\!\perp X_{\text{mis}}$ as long as there does not exist an unobserved confounder for M and Y. Formally, there should not exist extra variables $U \notin X$, such that $Y \not\perp\!\!\!\perp M|U$. I present simulation results in section 3.2 supporting this claim.

I further assume that researchers use imputation methods to replace X_{mis} with X_{imp}. Then, after multiple imputation, one will get a new dataset consist of Y and $X' = (X_{\text{obs}}, X_{\text{imp}})$. When certain assumptions are violated for multiple imputation methods, X' may not share similar distribution as the true X. This implies that $f_{\theta_0}(X) \neq f_{\theta_0}(X')$ and estimate $\widehat{\theta}$ based on X' does not converge to θ_0.

Hence, I propose a bias-correction term to deal with the biases. In this setup, the bias-correction term will be added as an additional covariate to the parametric regression.

Assumption 2.2 (Linearity of bias term)**.** *The parametric model in Equation (1) can be*

partitioned as

$$Y = f_{\theta_0}(X') + g_0(\eta) + e', \quad E(e'|X', \eta) = 0 \tag{7}$$

where g is a linear function of η.

The proposed method aims to calculate the the biased correction term η to restore the correct error term for the parametric model. After computing η, we can estimate θ by regressing Y on X' and $\widehat{\eta}$, which restores model honesty and asymptotic normality of $\widehat{\theta}$.[8]

The bias term for observation i can be written as

$$\eta_i = E(Y_i|T_i = 0, X_i) - E(Y_i|T_i = 1, \{X_{\text{obs},i}, X_{\text{mis},-i}\}) \tag{8}$$

where T_i denote the indicator variable that takes 1 if observations with missing indicator $M_{ij} = 1$ for at least one j ("missing" group) and zero otherwise ("complete" group), namely $T_i = \mathbf{1}\{\sum_j M_{ij} \geq 1\}$.

Equation (8) can be interpreted conceptually as the difference between the mean of outcome conditional on complete covariates, and the mean of outcome conditional on observed covariates for observations with missingness. The tree-based method has an advantage here because in equation (8), the first term is conditional on the complete covariates (which we cannot observe for the missing variables), while the second term only utilizes the observed covariates for observation i and missing covariates from all other observations. After partitioning based on all the observed covariates, tree will place the observation into one of the most likely terminal leaf. [9]

[8] A straightforward example is that in OLS setting, the η term corrects for the intercept and thus leads to better estimation and inference of the coefficients. Below, I show a decomposition of biases for OLS regression. I denote the true missing values as X_{mis} and the imputed ones as $_{\text{imp}}$.
$Y = X_{\text{obs}}\beta_{\text{obs}} + X_{\text{imp}}\beta_{\text{imp}} + X_{\text{mis}}(\beta_{\text{mis}} - \beta_{\text{imp}}) + (X_{\text{mis}} - X_{\text{imp}})\beta_{\text{imp}} + e$
This shows that the bias comes from two parts: the difference between imputed covariates and the real values, and the difference between the coefficients for the two. By design, both parts of the biases will be included in the estimated η term. Thus, we have a corrected intercept for the OLS regression.

[9] Tree model often drops observations with missingness in outcome variable Y, which is not included in the scope of this paper by assumption 2.3. Some other tree methods adopt an auxiliary method where missing indicators are created for each variable, to aid the partition process.

In other words, we are trying to calculate the difference between $E(Y|X)$ if we were to observe complete covariates and $E(Y|X')$ if we were to observe partial covariates and use other observations to fill in the missing entries. If MAR assumption holds, the two part should be equal, i.e the expectation η should be zero under MAR.

The first term $E(Y|X, T = 0)$ is nonparametrically identified by Assumption 2.1, that is, we can estimate $E(Y|X)$ using the complete observations, $E(Y|X) = E(Y|X, T = 0)$.

Finally, I assume that the outcome variable is observed without any missing values.

Assumption 2.3 (Complete outcome variable)**.** *No missingness in outcome variable Y.*

2.2 Implementation

This section describes how to estimate the bias term η in Equation (8) and then the parameter of interest θ in Equation (1). One major drawback of the existing imputation methods is their cyclical use of the data, which affects the statistical inference on θ. To avoid the problem, I split the data into the training set and the testing set (Step 1) where the bias term η is estimated using the training set (Step 2) and then the parameter θ is estimate using the testing set together with the estimated bias term $\widehat{\eta}$ (Step 3).

Step 1 Sample splitting.

Instead of using the whole dataset, I will split sample for the proposed method. This can prevent cyclic usage of data and thus prevent potential overfitting problem. The loss of statistical power due to sample split can be compensated through cross fitting multiple times.

Suppose researchers start with a raw dataset with missing data: $\{Y, X_{\text{obs}}\}$. We will first add a missing indicator M for each observation. So the new dataset now consists of $\{Y, X_{\text{obs}}, M\}$. Then, we will randomly split the sample into training S_{train} and testing sets S_{test}.

Step 2 Heterogeneous bias correction on observation i on training set S_{train}.

The quantity of interest in this step is the bias term η defined in Equation (8).

For observations in the training set S_{train}, let T_i denote the indicator variable that takes 1 if observations with $M_{ij} = 1$ for at least one j ("missing" group) and zero otherwise ("complete" group), namely $T_i = \mathbf{1}\{\sum_j M_{ij} \geq 1\}$.

The model will be trained in training set S_{train}. All demonstration in this paper adopts causal tree (Athey and Imbens, 2016) for this step.[10]

To implement causal tree procedure, $T = 1$ for all missing group ($M_{ij} = 1$ for at least one j) and $T = 0$ for the complete group ($M_{ij} = 0\ \forall j$). Y is the outcome variable.

Step 3 Inference on θ using the testing set S_{test}.

First, I predict $\hat{\eta}$ term for all observations in S_{test}. Since one needs a complete dataset to conduct parametric regression, I conduct mean imputation for the missing entries as place holder X_{imp}. Now, the S_{test} is complete with one extra column for $\hat{\eta}$ (the term will be zero for observations who did not contain missingness).

I then adopt any parametric regression model that a researcher decides to fit the data with, with the bias correction term $\hat{\eta}$. One can repeat step 2 and 3 for K times and take the average, to compensate for the loss of statistical power.

$$Y = \begin{cases} f_\theta(X) + 0 + e, & E(e|X) = 0 \quad \text{if fully observed} \\ f_\theta(X') + g(\hat{\eta}) + e', & E(e'|X', \hat{\eta}) = 0 \quad \text{if missing covariate} \end{cases} \tag{9}$$

Here $X' = (X_{\text{obs}}, X_{\text{imp}})$

[10] Any machine learning method for heterogeneous treatment effects, assuming conditional ignorability with a convergence rate equals to or faster than $n^{-1/4}$ works in this step.

2.3 Theoretical property

To prove theoretically why one can get nominal coverage, we should look at both the point estimate and standard error of the proposed estimator. This section will show the proposed estimator can achieve asymptotic normality, following a framework by Van der Vaart (2000).

Theorem 2.1. *(Van der Vaart, 2000)*[11] *For each θ in an open subset of Euclidean space let $x \mapsto m_\theta(x)$ be a measurable function such that $\theta \mapsto m_\theta(x)$ is differentiable at θ_0 for $P-$almost every x with derivative $\dot{m}_{\theta_0}(x)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $P\dot{m}^2 < \infty$*

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\| \tag{10}$$

Furthermore, assume that the map $\theta \mapsto Pm_\theta$ admits a second-order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If $P_n m_{\hat{\theta}_n} \geq \sup_\theta P_n m_\theta - op(n^{-1})$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_1^n \dot{m}_{\theta_0}(X_i) + op(1) \tag{11}$$

In particular, we have:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N\left(0, V_{\theta_0}^{-1} P\dot{m}_{\theta_0} \dot{m}_{\theta_0}^T V_{\theta_0}^{-1}\right) \tag{12}$$

where $V_{\theta_0}^{-1}$ is the second derivative matrix of function m_θ.

Proof. The proof will be done in a parametric least square setting. Thus, the objective function for the model is: $m_{\theta_0,\eta} = (Y - f_\theta(X') - g(\hat{\eta}))^2$. I will derive the term $\dot{m}_{\theta_0}$ and V_{θ_0} separately. First, I want to check if the objective function follows the condition stated in

[11] This is theorem 5.23 in the book. Van der Vaart denotes $E\dot{m}$ as $P\dot{m}$.

equation (10).

$$
\begin{aligned}
E(m_{\theta_0,g_0}) &= E(Y - f_\theta(X') - g(\hat{\eta}))^2 \\
&= E(\underbrace{f_{\theta_0}(X') + g_0(\eta) + e'}_{\text{By equation (7)}} - f_\theta(X') - g(\hat{\eta}))^2 \\
&= E(f_{\theta_0}(X') - f_\theta(X'))^2 + E(e'^2) + E(g_0(\eta) - g(\hat{\eta}))^2 \\
&\quad + \underbrace{2E(e'(f_{\theta_0}(X') - f_\theta(X')))}_{E(e'|X',\hat{\eta}=0)} + \underbrace{2E(e'(g_0(\eta) - g(\hat{\eta})))}_{E(e'|X',\hat{\eta}=0)} + \underbrace{2E((f_{\theta_0}(X') - f_\theta(X'))(g_0(\eta) - g(\hat{\eta})))}_{\text{By sample splitting}} \\
&= E(f_{\theta_0}(X') - f_\theta(X'))^2 + E(e'^2) + E(g_0(\eta) - g(\hat{\eta}))^2 \\
&\approx E[(\theta - \theta_0)\dot{f}_{\theta_0}]^2 + E[(\hat{\eta} - \eta)\dot{g}_0]^2 + E(e'^2)
\end{aligned}
$$

Since $E(m_{\theta_0,\eta})$ follows quadratic functional form, it satisfies the condition in equation (10). Note that the cross terms disappear based on $E(e|X', \hat{\eta}) = 0$ and sample splitting procedure. Also note that, $\dot{g}_0$ is a constance since g is a linear function by assumption.

$$
\begin{aligned}
&E(e'(f_{\theta_0}(X') - f_\theta(X'))) \\
&= E(E(e'(f_{\theta_0}(X') - f_\theta(X')))|X', \hat{\eta}) \\
&= E(E(e'|X', \hat{\eta})E(f_{\theta_0}(X') - f_\theta(X')|X', \hat{\eta})) \\
&= 0
\end{aligned}
$$

This is also true for the second cross term $2E(e'(g_0(\eta) - g(\hat{\eta})))$. And since $\hat{\eta}$ is estimated from a model that is trained by another subset of the data, we also have:

$$
E((f_{\theta_0}(X') - f_\theta(X'))(g_0(\eta) - g(\hat{\eta}))) = 0
$$

The first order derivative is:

$$\begin{aligned}\dot{m}_{\theta_0} &= -2(Y - f_\theta(X') - g(\hat{\eta}))\dot{f}_{\theta_0} \\ &= -2e\dot{f}_{\theta_0}\end{aligned}$$

The Hessian of objective function is:

$$\begin{aligned}E\ddot{m}_{\theta_0} &= 2E\dot{f}_{\theta_0}\dot{f}_{\theta_0} - 2E\ddot{f}_{\theta_0}(Y - f_\theta(X') - g(\hat{\eta})) \\ &= 2E\dot{f}_{\theta_0}\dot{f}_{\theta_0} - 2E(E(\ddot{f}_{\theta_0}(Y - f_\theta(X') - g(\hat{\eta}))|X', \hat{\eta})) \\ &= 2E\dot{f}_{\theta_0}\dot{f}_{\theta_0} \\ &= EV_{\theta_0}\end{aligned}$$

□

Asymptotic normality of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ with a mean of zero ensures that the estimator $\hat{\theta}_n$ possess nominal coverage. Li et al. (1989) further discuss the lower rate of convergence for nonparametric estimators in their paper (Theorem 2.1) and how to attain the best rate of convergence $n^{-1/4}$. This is tangential to the proposed method, since we are not interested in the properties of the coefficient for $\hat{\eta}$ term.

Thus, we have:

$$\begin{aligned}&E(V_{\theta_0}^{-1}P\dot{m}_{\theta_0}\dot{m}_{\theta_0}^T V_{\theta_0}^{-1}) \\ &\quad = (2E\dot{f}_{\theta_0}\dot{f}_{\theta_0})^{-1}E(-2e'\dot{f}_{\theta_0})(-2e'\dot{f}_{\theta_0})^T(2E\dot{f}_{\theta_0}\dot{f}_{\theta_0})^{-1} \\ &\quad = (E\dot{f}_{\theta_0}\dot{f}_{\theta_0})^{-1}E(e'e'^T)\end{aligned}$$

Taking a further look at the variance of error term[12]:

$$
\begin{aligned}
E(e'e'^T) &= E(Y - f_\theta(X') - g(\hat{\eta}))(Y - f_\theta(X) - g(\hat{\eta}))^T \\
&= E(Y - f_\theta(X'))(Y - f_\theta(X'))^T + Eg(\hat{\eta})g(\hat{\eta})^T - \underbrace{2E(Y - f_\theta(X'))g(\hat{\eta})}_{\text{by sample splitting}} \\
&= E(Y - f_\theta(X'))(Y - f_\theta(X'))^T + Eg(\hat{\eta})g(\hat{\eta})^T
\end{aligned}
$$

3 Simulation and Validation Studies

In this section, I compare performances in different settings among the proposed method, `Amelia` and `MICE`. Specifically, I consider three scenarios for the simulation study. In Section 3.1, covariates are simulated by the mixture of Poisson and normal distribution, while in Section 3.2 and 3.3, covariates are simulated using the multivariate normal distribution following the design of Pepinsky (2018). In the first two simulations, I consider the linear regression model while the last scenario consider the non-linear relationship between covariates and the outcome. I show that when certain assumptions are violated, `Amelia` and `MICE` fail to achieve model honesty – their confidence intervals fail to maintain the nominal coverage of the target parameter.

Furthermore, in Section 3.4 I conduct a validation study based on a real-world dataset. Starting with a complete dataset from Broockman (2013)'s study, I introduced missing values to the dataset under MAR. I show that the proposed method outperforms the other two methods, in terms of replicating the original study result.

[12]The most extreme case is that none of the observations contains missing variables. Then the term will become $E(Y - f_\theta(X))(Y - f_\theta(X))^T$. Thus, the proposed method tends to overestimate the standard error, as shown in simulations. I argue that, due to the proven asymptotic behavior of the estimator, one should not be concerned much of the issue. For a small sample size, if researchers are concerned about the standard error estimation, numerical correction (i.e. subtract the standard error from the overestimated part) cannot be done since we don't get to observe the true value X_{mis}. A more straightforward way to obtain an estimation is to bootstrap for the standard error. I provide code in the R-package for users to do so, when worried about the small sample size.

3.1 Linear regression under Poisson mixture: Missing at random

The first simulation set up is to assess how methods perform when the multivariate normal assumption is violated.

I consider a linear DGP under two covariates,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.5)$$

where

$$X_1 \sim \mathcal{N}(2, 1) \times \mathcal{N}(0, 5) \quad \text{and} \quad X_2 \sim \text{Pois}(50) + \mathcal{N}(0, 5).$$

The coefficients β_1 and β_2 are drawn from the binomial distribution as $\beta_1 \sim \text{Bern}(0.5)$ and $\beta_2 = 1 - \beta_1$.

I then introduce the missingness in X_1 and X_2 under missing at random: the missingness in X_1 is a function of X_2 and the missingness in X_2 is a function of X_1.

The performance of the method is evaluated with respect to coefficients. Specifically, I consider the empirical coverage of confidence intervals and bias with respect to β_1. For `Amelia` and `MICE` I first impute the dataset under default setting of the package. Then, I estimated parameters using OLS regression.

Figure 1 shows the results. The left panel shows the coverage where x-axis is the level of the confidence interval and y-axis is the empirical coverage. Everything that is above the 45 degree line is a conservative estimation of the confidence interval, while area below the 45 degree line is a "dishonest" inference defined earlier since one fails to achieve the designated confidence level. The figure shows that the proposed method (blue dots) maintains the nominal coverage while other two methods, `Amelia` (gray square) and `MICE` (gray triangle), fail to have a proper coverage. The panel in the middle shows the bias where the proposed method has the least bias among the three methods. Lastly, the right panel shows the distribution of standard error estimates over 2000 simulations. In this

plot, in addition to the three methods, I also plot the standard error estimate by fitting the model on the complete data (in red). The figure shows that the coverage results of the proposed method is not driven by the large variance estimate. It also demonstrates that poor coverage of `Amelia` and `MICE` is largely due to their large bias, while they sometimes underestimate the uncertainty.

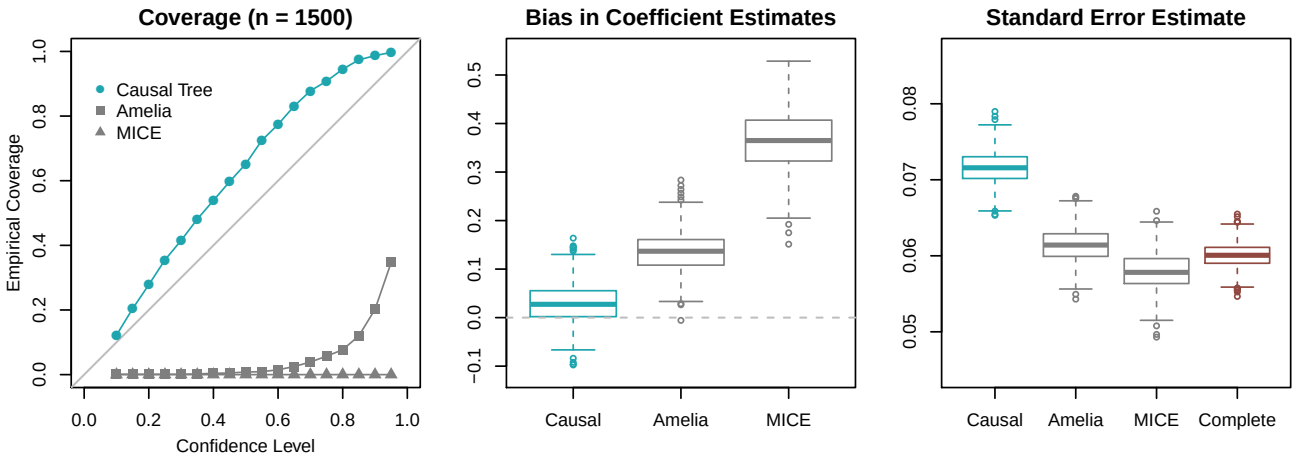

Figure 1: MAR - Poisson mixture

I further tested the asymptotic behavior of the proposed method by varying the sample size from 500 to $9,500$. The result is presented in Appendix A.1.

3.2 Linear regression under multivariate normal distribution: Missing not at random

To show that the proposed method works in MNAR, I simulated a dataset with a multivariate normal joint distribution. Then, I induced MNAR in both covariates – the probability of missing for a variable is a function of its own value.

Below is the data generation process for a multivariate normal distribution, which is similar to the simulation setup of Pepinsky (2018):

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

where the first covariate simply follows the standard normal distribution $X_1 \sim \mathcal{N}(0, 1)$ while the second covariate is simulated as

$$X_2 \sim \mathcal{N}(0.5u_1 + 0.5u_2, \sqrt{q}), \quad u_1, u_2 \sim \mathcal{N}(0, \sqrt{2(1-q)})$$

with $q = 0.2$. I set the coefficients $\beta_1 = \beta_2 = 5$ and assess the performance of methods on β_1 as in the previous section.

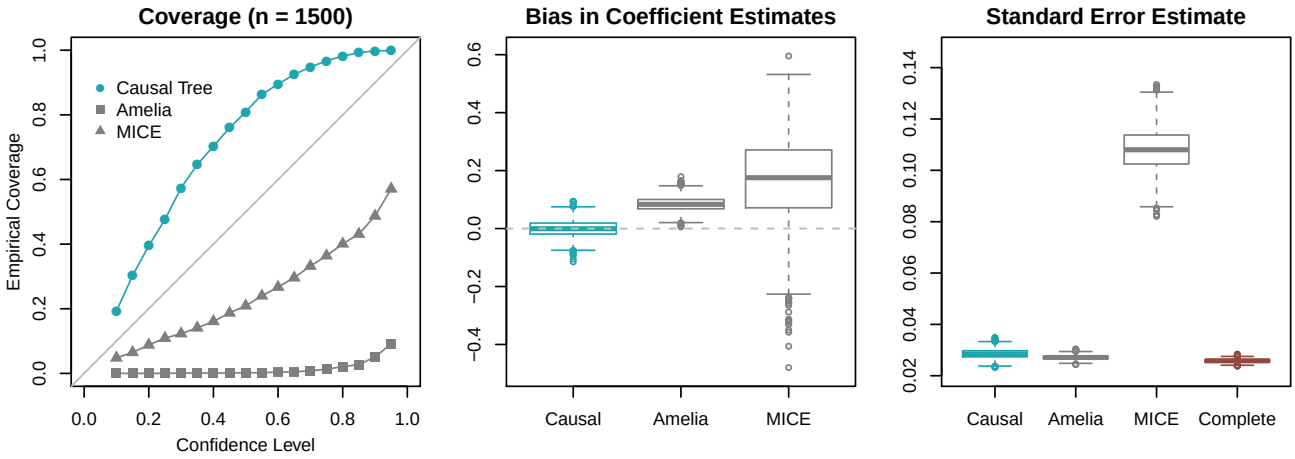

Figure 2: MNAR

Figure 2 shows the results. Similar to the previous section, I consider the coverage and the bias. As is shown in Figure 2, proposed method achieves honesty while other two fail. The missing not at random design means that a covariate is missing due to its own

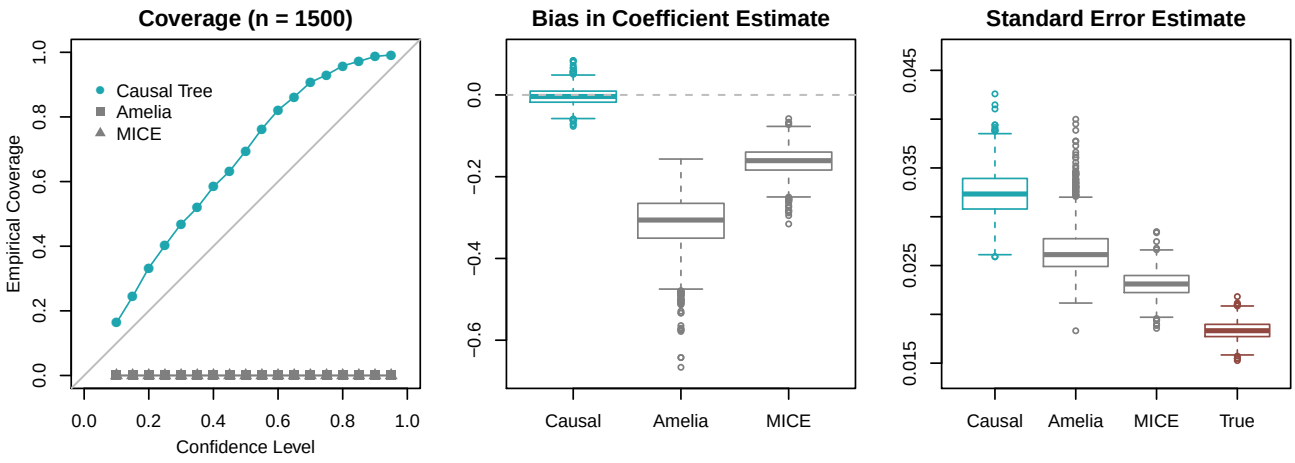

Figure 3: Polynomial Regression

value. However, since the missingness still satisfies conditional independence with regard to outcome variable: $Y \perp\!\!\!\perp M \mid X$ (Assumption 2.1), the proposed method still produce valid estimates, while other two methods should satisfy a stronger assumption of MAR, $M \mid X_{\text{obs}} \perp\!\!\!\perp X_{\text{mis}} \; \forall X_{\text{mis}}$.

3.3 Polynomial regression: Missing at random

To further check the performance of the proposed model in non-linear regression settings, I conduct simulations for polynomial regression. Keeping the covariates generating process the same as in Section 3.2, I modify the regression model to make it a non-linear relation.

$$Y = \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1)$$

I set the coefficients $\beta_2 = \beta_4 = 1$, $\beta_1 = 2$ and $\beta_3 = 5$.

I present the simulation result on β_2. The same comparison result (see Figure 3) holds

in polynomial setting: only the proposed method achieves model honesty. Both `Amelia` and `MICE` suffer from large biases and thus improper coverage.

3.4 Validation

I use a real-world dataset to validate the results above. Broockman (2013) conducted an experimental study to test legislators' responsiveness towards voters, by checking if they respond to voters' emails. 5593 emails were sent and Broockman received 2365 replies to them. He found that in general, legislators are less likely to respond to out of district emails. Furthermore, African American politicians are less sensitive than non-African Americans to their political incentives for responding.

I use this dataset for the purpose of validating the simulation results in a real world setting. The experimental dataset is complete, with no missing data in it. Thus, I use the original paper result as baseline for comparison, which is a replication to table 2 column "3-OLS" in the original paper, as the truth. Then, I make some of the continuous variables to be missing under MAR. Specifically, I made African American median household income and White median household income missing as a nonlinear function of state squire index. In total, 4743 missing entries were created and around 20% observations contain NA.

Table 2 shows the results based on the complete dataset (first two columns) and listwise deletion (last two columns). Variable names in the bold face are the coefficients of interest of the original author and ones with underline have missing values. In both types of variables, after listwise deletion, both the point estimates and standard errors are far off from the original result. Results using the proposed method, `Amelia` and `MICE` are shown in Table 3. The proposed method replicates the original result very well. For the three main covariates of interests, the estimates are almost identical.

Table 2: Comparison between listwise deletion and complete data

	True coef	True st.e	Listwise coef	Listwise st.e
(Intercept)	0.442	0.025	0.400	0.058
Treatment	−0.276	0.013	−0.249	0.026
Black legislator	−0.112	0.045	−0.313	0.101
Interaction	0.128	0.051	0.346	0.109
Non black minority	−0.035	0.031	0.027	0.068
Democratic legislator	−0.051	0.014	−0.051	0.027
Senator	0.089	0.016	0.100	0.032
South	−0.004	0.017	0.022	0.034
Black percentage	0.084	0.067	0.353	0.151
Black median household income	−0.0002	0.007	0.002	0.009
White median household income	0.021	0.010	0.011	0.018
State Squire index	0.489	0.071	0.897	0.186
Total population	−0.004	0.001	−0.002	0.003
Urban percentage	0.014	0.023	−0.050	0.049

Table 3: Comparisons among methods under MAR

	True coef	True st.e	Proposed coef	Proposed st.e	`MICE` coef	`MICE` st.e	`Amelia` coef	`Amelia` st.e
(Intercept)	0.442	0.025	0.451	0.041	0.443	0.030	0.457	0.032
Treatment	-0.276	0.013	-0.278	0.014	-0.208	0.013	-0.282	0.013
Black legislator	-0.112	0.045	-0.109	0.044	-0.106	0.045	-0.127	0.045
Interaction	0.128	0.051	0.136	0.053	0.104	0.052	0.137	0.051
Non black minority	-0.035	0.031	-0.030	0.031	-0.038	0.031	-0.036	0.030
Democratic legislator	-0.051	0.014	-0.049	0.014	-0.051	0.014	-0.052	0.014
Senator	0.089	0.016	0.089	0.016	0.088	0.016	0.089	0.016
South	-0.004	0.017	-0.008	0.017	-0.006	0.017	-0.008	0.017
Black percentage	0.084	0.067	0.090	0.069	0.080	0.068	0.095	0.067
Black median household income	-0.0002	0.007	-0.0003	0.012	−0.003	0.006	0.002	0.006
White median household income	0.021	0.010	0.018	0.013	0.003	0.010	0.006	0.011
State Squire index	0.489	0.071	0.459	0.073	0.507	0.073	0.539	0.074
Total population	-0.004	0.001	-0.003	0.001	-0.004	0.001	-0.003	0.001
Urban percentage	0.014	0.023	0.019	0.024	0.030	0.024	0.027	0.024
Bias correction			0.001	0.014				

4 Application: "Don't know" answers in ANES political knowledge questions

The missing data problem appears in many settings of empirical studies, such as survey datasets with non-ignorable non-response (Kuha et al., 2018; Dahlgaard et al., 2019; Wing, 2019). In this section, I also present an application on survey nonresponse (American National Election Study, ANES) using the proposed method. Both MAR and MNAR are possible mechanisms for missing data in ANES. For example, certain missingness could be the result of data collection issue and could be imputed by other variables. Other missingness can be survey refusal due to the answer itself, such as sensitive questions (election turnout) and difficult questions (political knowledge test). Both MAR and MNAR due to variable's value itself is within the scope of the proposed method. Thus, one should feel more comfortable adopting it, compared to other MI methods assuming MAR only.

Gilens (2001) found that raw policy-specific facts have a significant influence on the public's political judgment, and the general knowledge questions do not measure respondents' policy specific knowledge. Part of the analysis in the original paper was conducted using the American National Election Study (ANES) 1988. Logistic regression shows that both policy-specific and general political information significantly affect respondent's policy preferences. All missing answers are listwise deleted. All "don't know" answers are treated as lacking the knowledge and thus coded the same as wrong answer.

In this section, I first replicate the same model using ANES 2012 dataset. Then, I adopt the proposed method to bias-correct for all the missing answers. Comparison results are presented in Table 4 and Table 5. In both tables, the first column replicates exactly the coding from Gilens (2001). The second column shows the result of the proposed method imputing all "NA" entries in the dataset.

The general knowledge variable is constructed as the sum of five general knowledge questions in ANES 2012: the majority of House, the majority of Senate, the definition of

medicare, Democratic presidential candidate's religion and Republican presidential candidate's religion. The score ranges from 5 (highest, answers are all correct) to 0 (lowest, answers are all wrong). There are also "Don't know" and NA answers included in the dataset.

Table 4 presents results for environmental policy preference. The policy-specific information is the change in spending on environment related areas for 2012, with "decrease" as the correct answer. The outcome is a binary variable of opposing or favoring US offshore drilling. Table 5 presents results in tax raise policy preference. The policy-specific information is the change in deficit size for 2012, with "increase but almost the same" as the correct answer. The outcome is a binary variable of favoring tax raises.

Table 4

	Dependent variable:		
	(Opposing offshore drill)		
	Gilens coding	DK not imputed	DK imputed
Policy specific (environmental spending)	0.195	-0.693^{***}	-0.048
	(0.365)	(0.092)	(0.132)
General knowledge	0.167	0.159	0.209^{***}
	(0.134)	(0.137)	(0.051)
Interaction	-0.273	0.070	-0.271^{***}
	(0.194)	(0.126)	(0.044)
bias		0.088	0.244
		(0.087)	(0.173)
Observations	731	2,830	2,830
Note:			*p<0.1; **p<0.05; ***p<0.01
			Control Variables included.

Both tables replicate the original results in the paper: correct political knowledge leads to the correct policy preference. However, with additional imputation, the sign or the

Table 5

	Dependent variable:		
	(Taxraise)		
	Gilens coding	DK not imputed	DK imputed
Policy specific (deficit size)	1.047***	0.629***	0.867***
	(0.374)	(0.135)	(0.159)
General knowledge	0.140	0.113	0.115
	(0.185)	(0.175)	(0.083)
Interaction	-0.074	0.259	-0.084
	(0.221)	(0.159)	(0.072)
bias		0.170*	0.547*
		(0.097)	(0.293)
Observations	734	2,830	2,830
Note:		*p<0.1; **p<0.05; ***p<0.01	
		Control Variables included.	

significance of the coefficients changed. This is a warning message that ignoring the missing data might lead to biased estimation. Tables 6 and 7 in the Appendix show the full results of the regression. For example, after imputation the party variables gained significance. This may indicate that the impact of knowledge on policy preference could be confounded by party affiliation. This is not discussed in the original paper.

In tables 4 and 5, the third column shows the results with the "Don't know" answer treated as missing and imputed. In his original study, Gilens (2001) treated them as wrong answers. However, "Don't know" answers do not necessarily indicate lack of memory from the respondent, instead it only shows that the respondent chooses not to provide an answer after deliberation (Beatty et al., 1998). In his original paper, Gilens (2001) coded "Don't know" as wrong answers. This may be an over-simplification of the problem since people might also answer "don't know" if they are uncertain about their answer and avoiding giving a wrong answer. Potentially, this falls in the realm of MNAR. To further illustrate the "Don't know" respondents are different from the wrong answer group, I plotted (Figure 4), for the "don't know" group (in green) and wrong answer group (in black), the distribution of age, income group and political interest level. And I show that there may exist some systematic differences between the two types of answers. Simply coding all the "don't know" answer as wrong may lead to biased estimation.

5 Discussion

Sample splitting is not a common practice for missing data imputation: Other multiple imputation methods compared in this paper make use of the whole dataset. However, when MAR is violated, utilizing the whole dataset may cause over-fitting by constructing from the dependent variables. Social scientists have been arguing against selection on the dependent variables (King et al., 1994; Montgomery et al., 2018). In other words, when all the missing covariates are imputed (partly) from the dependent variable, one should

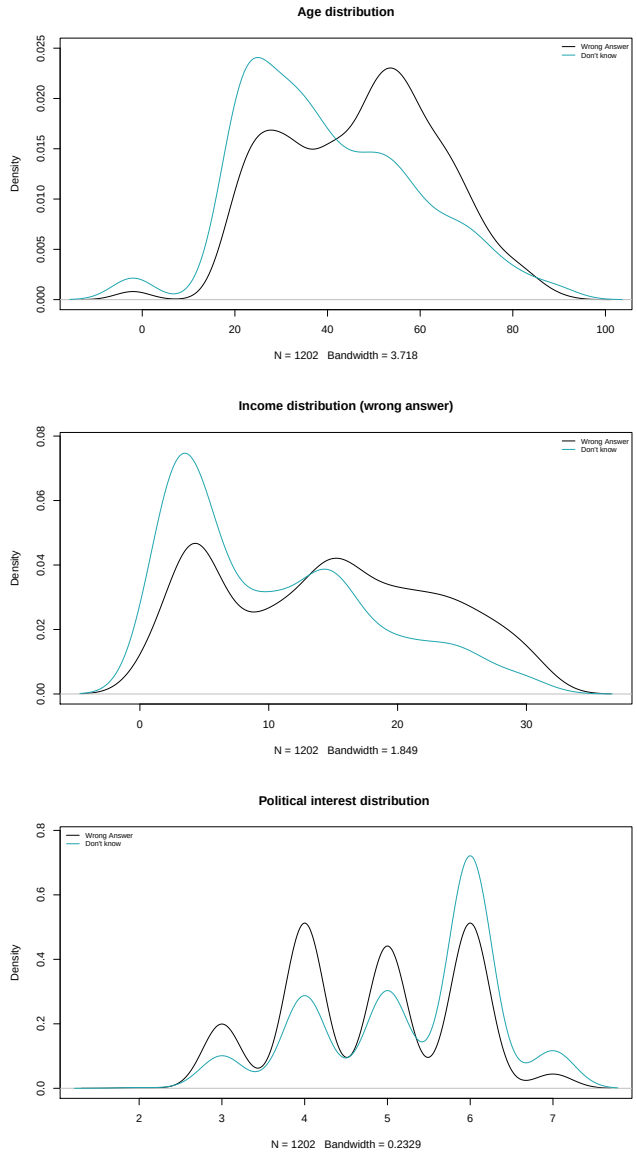

Figure 4

take more caution before using the imputed value to run any model to predict the same dependent variable. In figure 5, I used the `Amelia` and `MICE` **without** dependent variable and then conducted the OLS model (the simulation set up is exactly the same as in figure 8). As it is shown, without dependent variable included in imputation step, both multiple imputation methods fail to achieve similar performance as in figure 8. As a matter of fact, Bao et al. (2019) show a proof in their paper that, under MAR, excluding outcome variables

will lead to bias in multiple imputation.

The author acknowledges the potential loss of statistical power by sample splitting. One consequence is that the proposed method tends to over-estimate standard errors. To mitigate the problem, all the results presented in this paper by the proposed method are averages of cross-fitting among different training and testing splits.

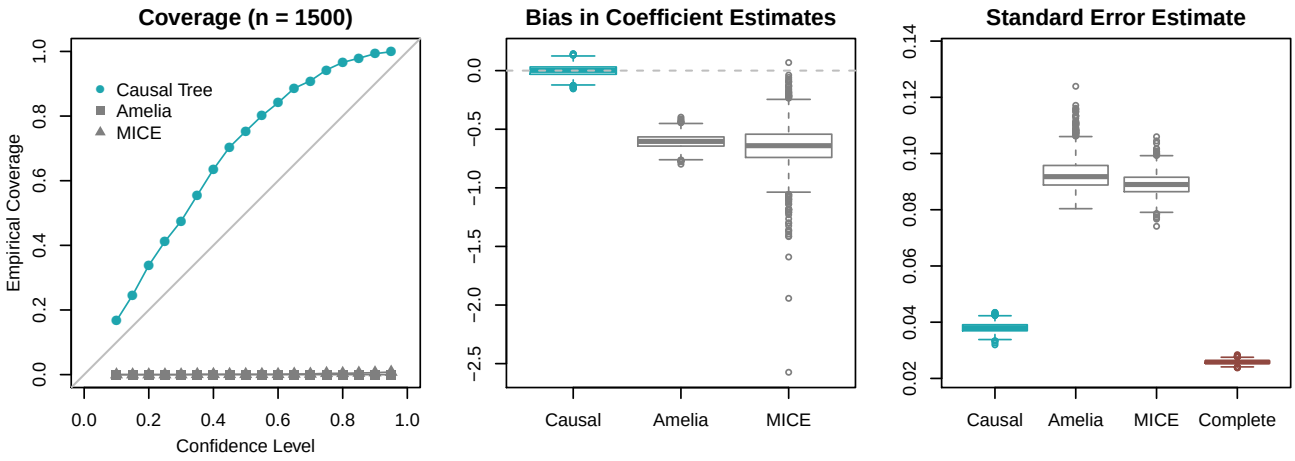

Figure 5: Imputation without Y under MCAR

I also acknowledges the fact that the proposed method makes compromise between prediction accuracy (such as MSE of predicted Ys) and model inference. Figure 6 shows the mean squared error of the predicted Ys for each method, with sample size increasing. Although achieving better model inference, the proposed method produces higher MSE than the other two. This happens because placeholders are the results of mean-imputation and thus may not recover the individual characteristics in the dataset.

Furthermore, this comparison also corresponds to the previous criticism of the multiple imputation methods. By using all observations and all variables in the dataset, the regression result from imputed dataset may be an overfit to the observed data. In other words,

high performance in MSE does not necessarily guarantee a valid model inference for imputation methods. And researchers should take caution in choosing appropriate methods for specific goals. Arel-Bundock and Pelc (2018) show that, only under limited conditions will multiple imputation improve regression estimates.

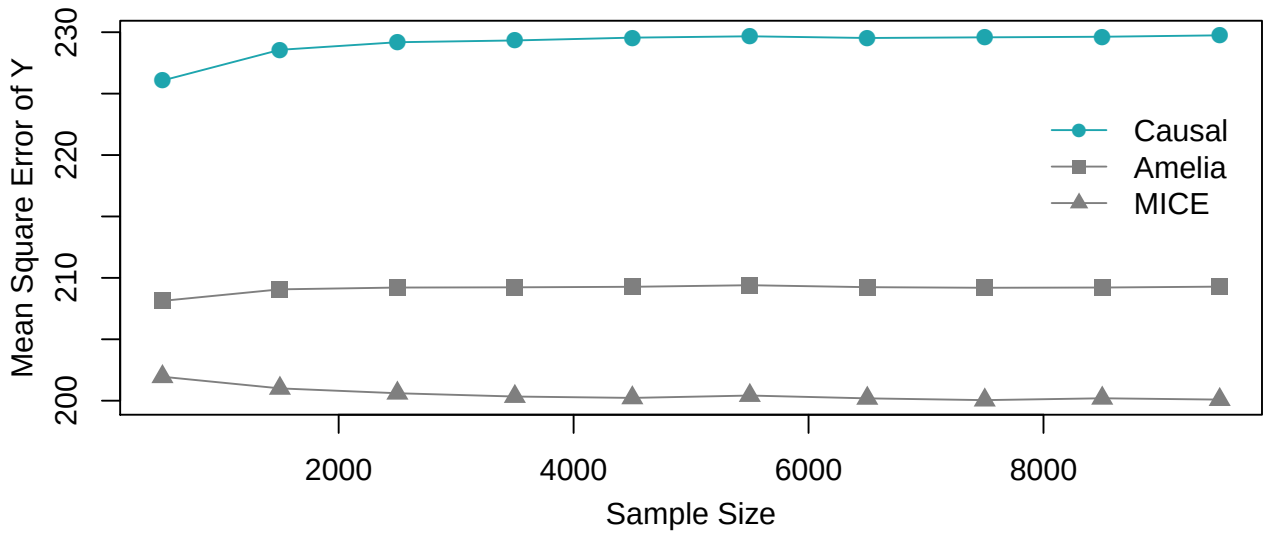

Figure 6: Mean squared error of $\overline{Y}$

6 Conclusion

In most social science applications, the ultimate goal of missing data imputation is to produce a complete dataset for researchers to analyze. Existing imputation models create artificial separation between imputation step and analysis step. In this paper, I show that this is problematic for inference in the sense that the confidence intervals may not have nominal coverage. The proposed method integrates data imputation with the downstream inference step, thus ensuring a more reliable inference result.

Valid model inference enables empirical scholars to claim theories that can be applied to a broader population that shares similar traits with the sample, while a minimized prediction error only shows goodness of fit to the specific sample. For example, in a ordinary least square setting, the coefficients (and their confidence intervals) of variables are the inferences that we draw from the dataset, while the MSE of prediction measures the goodness of fit to this one specific dataset. Thus, imputation methods should also be evaluated in terms of the property of the model estimator coming out of it. Especially for studies designed to prove or disprove a certain scientific claim, researchers conduct imputation not just to obtain a complete dataset, but also to use it for models to test their theories.

The proposed method can potentially be applied to a wide range of studies to tackle with the missing data problem, when researchers rely on the data for honest model inferences. In addition to survey data analysis, we see missing data in other observational datasets such as trade datasets, climate records and government reports. Possible methodological extension includes expanding the applicable model class, such as the exponential family model and to all the maximum likelihood estimators.

References

Arel-Bundock, V. and Pelc, K. J. (2018). When can multiple imputation improve regression estimates? *Political Analysis*, 26(2):240–245.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Bao, L., Bagozzi, B., and Gill, J. (2019). Should missing values of the outcome variable be imputed for regression models? *working paper*.

Beatty, D., Herrmann, C., Puskar, J., and Kerwin, P. (1998). 'don't know'responses in surveys: Is what i know what you want to know and do i want you to know it? *Memory*, 6(4):407–426.

Broockman, D. E. (2013). Black politicians are more intrinsically motivated to advance blacks' interests: A field experiment manipulating political incentives. *American Journal of Political Science*, 57(3):521–536.

Dahlgaard, J. O., Hansen, J. H., Hansen, K. M., and Bhatti, Y. (2019). Bias in self-reported voting and how it distorts turnout models: Disentangling nonresponse bias and overreporting among danish voters. *Political Analysis*, 27(4):590–598.

Gilens, M. (2001). Political ignorance and collective policy preferences. *American Political Science Review*, 95(2):379–396.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models.

In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER.

Honaker, J. and King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2):561–581.

King, G., Honaker, J., Joseph, A., and Scheve, K. (1998). List-wise deletion is evil: what to do about missing data in political science. In *Annual Meeting of the American Political Science Association, Boston*.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review*, 95(1):49–69.

King, G., Keohane, R. O., and Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press.

Kuha, J., Butt, S., Katsikatsou, M., and Skinner, C. J. (2018). The effect of probing "don't know" responses on measurement quality and nonresponse in surveys. *Journal of the American Statistical Association*, 113(521):26–40.

Lall, R. (2016). How multiple imputation makes a difference. *Political Analysis*, 24(4):414–433.

Li, K.-C. et al. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*, volume 333. John Wiley & Sons.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, 62(3):760–775.

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. *Political Analysis*, 26(4):480–488.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295.

Sun, B. and Tchetgen, E. J. T. (2014). On inverse probability weighting for nonmonotone missing at random data. *arXiv preprint arXiv:1411.5310*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Wing, C. (2019). What can instrumental variables tell us about nonresponse in household surveys and political polls? *Political Analysis*, 27(3):320–338.

A Additional Simulation Results

A.1 Missing at random with varying sample size: Poisson Mixture

To test the asymptotic behavior of all methods, I present figure 7 which shows the simulation result under Poisson-Normal mixture (missing at random). As sample size increases, both `Amelia` and `MICE` cover 0% for 95% confidence interval while the proposed method provides relatively stable performance.

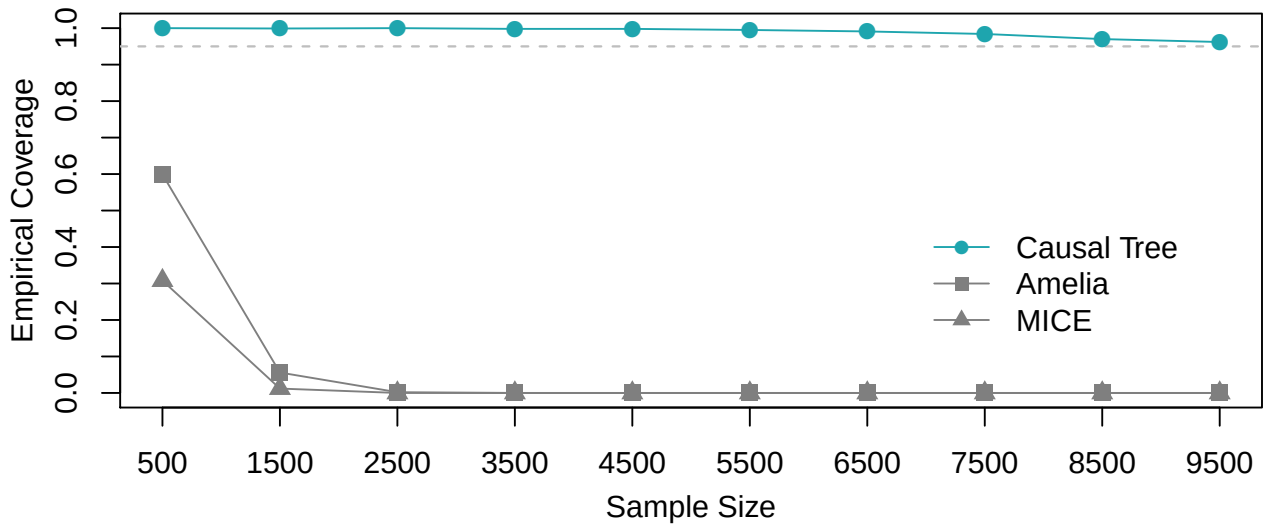

Figure 7: Simulation with variation in sample size

A.2 Missing at random - Multivariate normal

I borrow the simulation setup in Pepinsky (2018). Keeping data generating process the same, I only modify the missing assumptions for the purpose of this paper.

$$
\begin{aligned}
x_1 &\sim N(0,1) \\
u_1, u_2 &\sim N(0, \sqrt{V}) \\
V &= 2(1-q) \\
x_2 &= 0.5u_1 + 0.5u_2 + N(0, \sqrt{q}) \\
q &= 0.2 \\
e &\sim N(0,1) \\
y &= e + 5x_1 + 5x_2
\end{aligned}
$$

The scatter plot shows the coverage of confidence intervals produced by each method after 2000 simulations, compared to the ideal nominal coverage line $y = x$ in red. Anything that is above the redline is a conservative estimation of the confidence interval, while everything below the redline area is a "dishonest" inference defined earlier since one fails to achieve the designated confidence level. I also plotted the 2000 point estimates (second box plot) and their standard deviations (third box plot). Figure 8 and figure 9 present the detailed results under MCAR and MAR. Under multivariate normal setup, all three methods of interest give out very similar performance.

B Causal tree

Since we are interested in the learning about the difference between each group:

$$\eta = E(Y|X) - E(Y|X_{\text{obs}})$$

where the causal forest model is defined as $\Pi = \{l_1, \cdots, l_{\#\Pi}\}$, with $\cup_{j=1}^{\#\Pi} l_j = X$. Π denotes partition and l denotes each leaf. Here, we assume ignorability , i.e. $Y \perp\!\!\!\perp M|X$. This

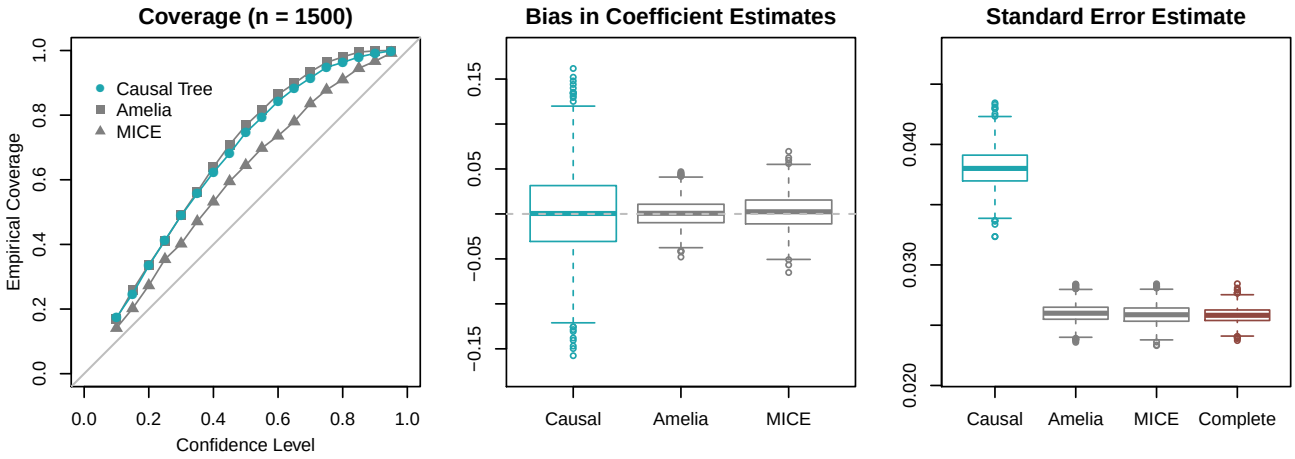

Figure 8: MCAR - multivariate normal

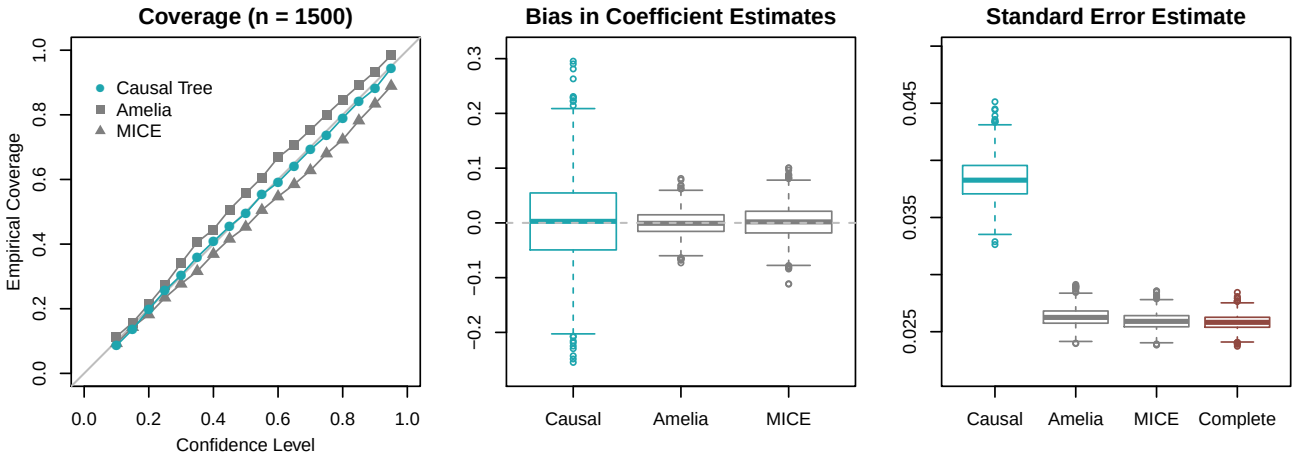

Figure 9: MAR - multivariate normal

assumption can be interpreted as the missing scheme is random with regard to Y, when conditional on all the observed data. The estimated counterparts for η:

$$\hat{\eta} = \hat{\mu}(M = 1, X_{\text{obs}}, \Pi) - \hat{\mu}(M = 0, X, \Pi)$$

The objective is to minimize the in sample mean squared error defined as follow (Athey and Imbens, 2016) :

$$\begin{aligned}
\widehat{\text{MSE}}(S_{\text{train}}, \Pi) &= \frac{1}{\#(S_{\text{train}})} \sum_{i \in S_{\text{train}}} \left((\eta_i - \hat{\eta}_i(X_i, S_{\text{train}}, \Pi))^2 - \eta_i^2 \right) \\
&= -\frac{2}{N_{\text{train}}} \sum_{i \in S_{\text{train}}} \hat{\eta}_i(X_i, S_{\text{train}}, M) \hat{\eta}_i(X_i, S_{\text{train}}, \Pi) + \frac{1}{N_{\text{train}}} \sum_{i \in S_{\text{train}}} \hat{\eta}_i^2(X_i, S_{\text{train}}, \Pi) \\
&= -\frac{1}{N_{\text{train}}} \hat{\eta}_i^2(X_i, S_{\text{train}}, \Pi)
\end{aligned}$$

C ANES full results

Table 6: ANES full result on offshore drilling

	Dependent variable: offshore drill		
	(1)	(2)	(3)
Enviro	0.195	-0.693***	-0.048
	(0.365)	(0.092)	(0.132)
general	0.167	0.159	0.209***
	(0.134)	(0.137)	(0.051)
interaction	-0.273	0.070	-0.271***
	(0.194)	(0.126)	(0.044)
gender	0.587***	0.607***	0.615***
	(0.174)	(0.087)	(0.088)
age	-0.020***	-0.014***	-0.016***

	(0.007)	(0.004)	(0.004)
edu	-0.030	0.030	0.022
	(0.037)	(0.018)	(0.019)
black	0.588**	0.098	0.117
	(0.236)	(0.121)	(0.122)
income	-0.043***	-0.012*	-0.014**
	(0.012)	(0.006)	(0.007)
marital	0.019	-0.233**	-0.231**
	(0.188)	(0.095)	(0.096)
party1	-0.041	0.385***	0.364***
	(0.196)	(0.097)	(0.097)
party2	-0.492*	-0.904***	-0.897***
	(0.256)	(0.120)	(0.121)
homeown	-0.250	-0.055	-0.060
	(0.189)	(0.102)	(0.102)
union	0.073	0.216*	0.198*
	(0.238)	(0.118)	(0.119)
ch18	0.074	0.021	0.051
	(0.187)	(0.103)	(0.103)
worseoff	-0.203**	0.136***	0.120***
	(0.089)	(0.045)	(0.046)
religion1	-0.570***	-0.298***	-0.274***
	(0.209)	(0.101)	(0.101)
religion2	-0.225	-0.263**	-0.214**
	(0.218)	(0.108)	(0.109)

region2. North central	-0.050	-0.034	-0.042
	(0.268)	(0.137)	(0.138)
region3. South	-0.194	-0.290**	-0.288**
	(0.247)	(0.124)	(0.125)
region4. West	0.275	-0.015	-0.019
	(0.276)	(0.134)	(0.135)
retire	0.348	-0.035	-0.035
	(0.311)	(0.127)	(0.128)
homemaker	0.041	-0.150	-0.176
	(0.303)	(0.152)	(0.152)
bias		0.088	0.244
		(0.087)	(0.173)
Constant	1.692***	0.564*	0.533*
	(0.561)	(0.330)	(0.302)
Observations	731	2,830	2,830
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 7: ANES full result on Tax-raise

	Dependent variable: taxraise		
	(1)	(2)	(3)
defsize	1.047***	0.629***	0.867***
	(0.374)	(0.135)	(0.159)
general	0.140	0.113	0.115
	(0.185)	(0.175)	(0.083)
interaction	-0.074	0.259	-0.084
	(0.221)	(0.159)	(0.072)

gender	-0.146	-0.061	-0.062
	(0.187)	(0.088)	(0.088)
age	0.022***	0.015***	0.015***
	(0.008)	(0.004)	(0.004)
edu	0.044	0.008	0.007
	(0.039)	(0.019)	(0.019)
black	-0.232	0.036	0.039
	(0.253)	(0.135)	(0.135)
income	0.028**	0.006	0.005
	(0.014)	(0.006)	(0.006)
marital	0.052	-0.166*	-0.163*
	(0.207)	(0.096)	(0.096)
party1	1.027***	0.819***	0.813***
	(0.219)	(0.107)	(0.107)
party2	-0.492*	-0.789***	-0.771***
	(0.259)	(0.103)	(0.103)
homeown	0.017	-0.206*	-0.191*
	(0.203)	(0.106)	(0.106)
union	0.034	0.003	0.013
	(0.261)	(0.121)	(0.121)
ch18	-0.118	0.048	0.055
	(0.197)	(0.104)	(0.104)
worseoff	0.003	0.192***	0.190***
	(0.095)	(0.046)	(0.046)
religion1	0.206	-0.127	-0.122

	(0.222)	(0.101)	(0.101)
religion2	0.115	0.011	0.021
	(0.236)	(0.110)	(0.110)
region2. North central	0.377	0.149	0.134
	(0.284)	(0.139)	(0.139)
region3. South	0.335	-0.034	-0.043
	(0.260)	(0.125)	(0.125)
region4. West	0.496*	-0.033	-0.058
	(0.295)	(0.135)	(0.135)
retire	-0.952***	-0.351***	-0.361***
	(0.339)	(0.128)	(0.128)
homemaker	-0.314	-0.134	-0.150
	(0.312)	(0.148)	(0.148)
bias		0.170*	0.547*
		(0.097)	(0.293)
Constant	-2.145***	-0.724**	-0.589*
	(0.611)	(0.336)	(0.314)
Observations	734	2,830	2,830
Note:		*p<0.1; **p<0.05; ***p<0.01	