Causal Inference with Missing Data and Imputation Methods

Kun Qian, Stephane Remigereau, Xingzi Xu

November 2020

1 Introduction

Causal inference with missing data is a common challenge in many studies, notably in fields such as epidemiology, economics, computer science etc. The goal of using a statistical approach such as the potential outcomes framework is to leverage the outcomes that we observed and make meaningful interpretation of the "full schedule of potential outcomes", i.e. the Science. Under this framework, the causal effect is defined as the difference between the potential outcomes of units under different treatment conditions. Causal inference with complete data set can also be seen as a missing data problem. Because each unit is exposed to only one treatment condition, the Science can never be observed in full and causal inference is inherently a missing data problem. The issue is even harder to tackle when the outcome or one or several covariates are completely missing from the data set. In this case, assumptions and models must be made carefully to avoid introducing much bias and variance to inference, often yielding unsatisfactory results.

The paper at hand dives into the missing data theory and compare different methods to deal with inference on incomplete data sets. First, we focus on the missing data theory by underlining why such cases can happen in application and presenting different types of missing data mechanisms. Then, we shed lights on state of the art approaches and methods used by scientists to impute missing data, which can be separated in two categories, single imputation (SI) and multiple imputation (MI). Finally, we analyze and compare the results on the average treatment effect and variances for each method by applying them to the `Ozone` data set.

2 Missing Data Theory

2.1 Definition

Missing data is a common problem in data science and research that occurs when no data value is stored for an observable variable. For instance, it can happen when participants drop out of a study, or when individuals refuse to answer questions in a survey. For the survey case, some items are more likely to generate a nonresponse, notably those that are considered more private such as income.

The analysis of incompete data sets is usually based on either implicit or explicit assumptions, and thus ignore the processes that cause the missing data. The resulting inference could be biased or have inaccurate variance. This problem was first pointed out by Rubin in 1987.(1) Rubin classified missing data problems into three categories: missing completely at random(MSAR), missing at random(MAR), and missing not at random(MNAR). Each data point has certain likelihood to be missing. The missing data mechanism is the process that controls these probabilities. The missing model is the model of these processes.

To handle data correctly, it is essential to differentiate between these types: if the missing data is missing completely at random, the remaining data is representative of the whole population, yet if some data is missing systematically, it could lead to a bias in inference.

2.2 Missing Completely At Random (MCAR)

Values in a data set are said to be MCAR if the events that lead to missing data are random and independent of observable and unobservable variables, and occur entirely at random. In other words, the probability of missingness is the same for all entries. In this case, the remaining data is unbiased and representative of the whole data set, but such case rarely happens in practice.

If we consider a data set with a covariate X, a response Y, and an additional variable R to denote the missinngness where $R = 0$ means that Y is not observed and $R = 1$ means that Y is observed, then MCAR can be written as:

$$R \perp\!\!\!\perp X, Y$$

2.3 Missing At Random (MAR)

MAR occurs when the missingness is not random, but depends only on variables whose information is complete. MAR is an assumption that is impossible to verify statistically. It mainly depends on the analysis method. Regression on MAR data can still induce parameter bias if irrelevant variants are included. MAR assumption in missing-data problem is similar to ignorability in the casual framework, which assumes that all the confounding covariates are controlled in the regression model. Both require sufficient information to be collected for the assignment mechanism to be ignored.

In the simple case considered above, where X is a covariate, Y the response, and R the random variable denoting missingness, we can write MAR as:

$$P\left(R = 0|X, Y\right) \ = P\left(R = 0|X\right)$$

2.4 Missing Not At Random (MNAR)

The last case, and the hardest one to deal with, is MNAR (or nonignorable nonresponse). Based on the cause of missingness, MNAR can be further split into two categories: (1)missingness that depends on unobserved predictors, and (2)missingness that depends on the missing data itself. For the first kind, an accurate model can avoid bias in inference, but is almost impossible to construct in reality. For the second kind, including more predictors in the model can make the data set closer to MAR, and thus can mitigate the inference bias. The most common strategies to handle MNAR are to get more data to be more certain about the causes for the missingness, or to perform sensitive tests with various scenarios.

2.5 Dealing with missing data assumption in real cases

As discussed in 2.3, MAR is the easiest case to handle among all three missing data types but cannot be proved statistically. In practice, researchers usually make assumption of MAR and bring the data set closer to MAR through modeling or research design. Including more predictors in the models or doing follow-up survey on nonrespondents are both commonly used. In our project, we also use MAR assumption for the imputation and inference.

3 Imputation Methods

Missing-data methods can be divided into three categories: complete-case analysis (CC), single imputation (SI), and multiple imputation(MI). While CC and SI have advantages in computation time and complexity, these two methods are rarely used in the researches because they are likely to result in biased data sets and have stronger assumption on the missing data mechanism. On the other hand, MI properly reflects the uncertainty due to missing values and thus gives more accurate data. The results from MI can be easily analyzed by standard procedures, which makes MI more useful in scientific research.

We implement CC as an naive approach and a base line, two SI methods using mean and linear regression, and four MI methods including stochastic regression, predicted mean matching method(pmm), random forest, and principle components analysis (PCA). We compare the results with those from MI to show the advantages of MI in handling missing values.

3.1 Complete-case analysis (CC)

CC uses statistical procedures that exclude all incomplete cases - the observations with any missing data - from the analysis. This method is simple but ignores the possible missing values in other variables and the systematic difference between the complete and incomplete cases. CC also wastes a lot of useful information from observation. If data are MCAR, then CC works well without introducing any bias. However, since MCAR rarely happens in real case, the inference with CC may not be useful to draw conclusion for the population, especially for small population.

Using linear regression on the complete cases to estimate treatment effect is the most commonly used method in CC. We implement this method as the base line for comparison.

3.2 Single Imputation (SI)

In SI, each missing value is substituted with a value, which is imputed from all the complete cases in the data set, and the result is one filled-in data set. Standard statistical procedures for complete data analysis can then be used to generate conclusion on casual inference or prediction.

There are two main assumptions in SI. Firstly, data are assumed to be MAR. Missing values are treated as if they were known in the complete-data analyses. Secondly, the parameters of the data model for the complete cases and those for the missing data indicators are distinct. If both assumptions are satisfied, the missing-data mechanism is said to be ignorable and SI will give good result. However, if a complete case and an incomplete case have different values for a variable $Y3$ yet have the same values for variables $Y1$ and $Y2$, the assumption of MAR is violated. Then there will be a response bias for $Y3$ and the variance will be underestimated. A major disadvantage of SI is that it does not reflect the uncertainty in the prediction of missing values, and thus results in biased estimates and inaccurate standard errors. SI could give relatively good results when the incomplete data set is small and has only a few variables, which is usually not the case in a real-life study.

Using mean/median values, the most frequent value, k-NN and linear regression to fill in missing data are the most commonly used methods in SI. Considering the computation time and statistic power, we implement SI with mean values, deterministic linear regression imputation and random forest.

3.2.1 Mean, Median & Most frequent value Imputation

In imputation using mean/median values, the mean/median of the non-missing values in a column is first calculated. The missing values within each column are then replaced separately and independently. This method can only be used with numeric data and will give poor results on categorical features. On the other hand, imputation using the most frequent value in each column works well with categorical features, while introducing bias in the data.

3.2.2 k-NN Imputation

In imputation using k-NN, a basic mean is first imputed then the resulting complete list is used to construct a KD Tree. K nearest neighbours are found with the resulting KD tree and the weighted average of them will be the final imputation. The method will give the best result among the three most commonly used methods in SI, but it is computationally expensive and relatively sensitive to outliers in the data sets.

3.2.3 Random Forest Imputation

Random forest imputation is a combination of classification and regression trees. A new bootstrapped sample of observations and predictors are selected for each tree. Each tree is then subdivided recursively by a binary splitting approach based on the values of the predictor variables. (6) Eventually, a random forest consists of a large number of trees with different samples and predictors. The bootstrap aggregation of regression trees can reduce the risk of over-fitting, because combining results from all trees reflects the randomness and produce more accurate predictions.

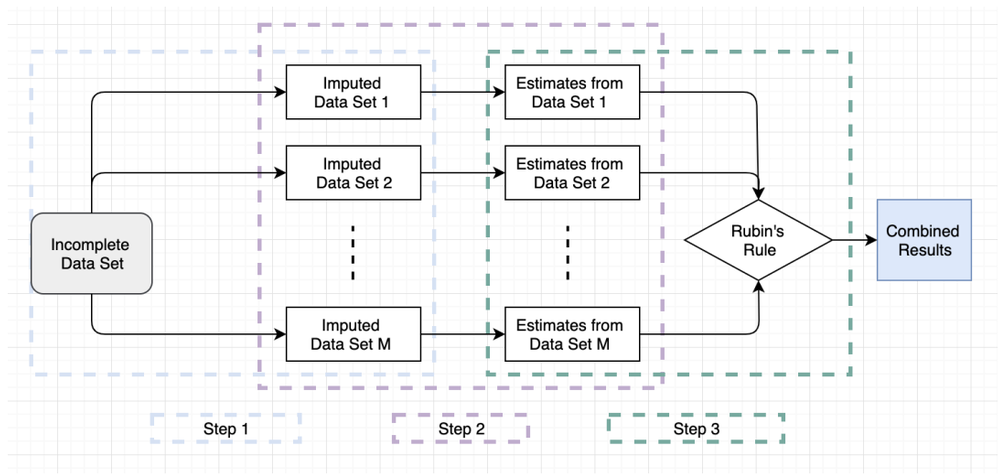
3.2.4 Deterministic Linear Regression Imputation

In deterministic linear regression imputation, a model is first built from the observed data and then the missing data are replaced with the values predicted with regression models. The process is repeated for each variable with missing values. The major disadvantage of this method is that the inherent uncertainty of the imputed variables are ignored, and thus the variance is underestimated.

3.3 Multiple Imputation (MI)

MI was first proposed by Rubin in 1987 as a powerful statistical technique to handle missing data.(13) Firstly, MI has looser assumptions on missing data mechanism. For both MAR and MCAR data, the pooled parameter estimates from several fill-in data sets are unbiased and standard errors are more accurate. When the missing data mechanism are close to MAR, MI outperforms CC and SI. Also, MI replaces each missing value with a set of plausible values instead of one plausible value in SI. Therefore, MI could properly reflect the uncertainty from the missing values and gives valid confidence intervals for parameters.(2) Just like in SI, standard procedures for complete data can be used on multiple imputed data sets from one incomplete data set and the combined results can be used for further analysis.

While MI is very powerful, users need to make many important decisions, which require a strong statistic background, to make fully use of MI. These decisions include the number of imputation, the choices of variables, and the method to impute data. The method to impute missing data depends on the assumption on the missing data pattern. For monotone missing data, either a regression method assuming multivariate normality or a non-parametric method using propensity scores give good results. For data sets with arbitrary missing patterns, a Markov chain Monte Carlo(MCMC) method can be used in two ways based on the assumption of multivariate normality. (3) The first method is to create multiple imputations with simulations from a Bayesian prediction distribution to generate normal data. The second method is to keep imputing until the missing data pattern become relatively monotone. (4) Three sets of variables should be included in imputation to produce high-quality result: variables used in models during further analysis, variables related to the missingness of the imputed variable, and variables correlated with the imputed variable. (3) Covariates having too many missing values for variables with missing values should be removed from the imputation model. Van Buuren, Boshuizen, and Knook also made similar suggestions on the choices of variables. The number of imputations should be determined by either efficiency or statistical power. 3 or 5 imputations are sufficient if efficiency is more important (3), while the number of imputation should equal the percentage of incomplete cases if statistical power is critical.(5)



:

Figure 1: Main steps in multiple imputation.

As shown in Figure 1, there are three main steps in MI.

1. `Imputation`: All missing data are filled in m times with the chosen method to generate m complete data sets. The distribution of the complete cases are used to estimate m sets of likely values of the missing data. The uncertainty due to the missing values is reflected in this step

2. `Analysis`: The m complete data sets are analyzed separately with standard procedures. The combination of results across the M data sets reduce the bias in estimators from each data set.

3. `Pooling`: m sets of results from the m complete data sets are combined and the estimated treatment effect and standard errors are calculated with Rubin's rules for the inference. Based on Rubin's suggestion, both the estimation of each missing data point and the standard errors should be the average of respective estimates across all imputed data sets. Then, the variance of imputation for each missing data point across all imputed data sets should be calculated and combined with the variance within each imputed data set to calculate the standard errors.

We implemented 3 different methods in Step 1 to generate imputed data sets: stochastic regression imputation, multiple imputation by chained equations(MICE) with predicted mean matching method(pmm), and principal component analysis (PCA) imputation. The number of imputed data for each method is set as 5 for comparison, which will be discussed more detailed in Part 4.

3.3.1 Stochastic Regression Imputation

Stochastic regression imputation is an extension of deterministic regression. Both methods predict the missing values by regression on other related variables in the same data set. In stochastic regression imputation, an extra error term is added to include the uncertainty and reduce the bias due to simple regression.

3.3.2 MICE with pmm

MICE with pmm is similar to the regression method because they both use simulated regression model on the sampling data sets. MICE algorithm is a iterative process.

In the imputation step with MICE, the missing data are filled in through iteration with varying predictive models. In each iteration, variables with missing data is imputed with other variables in the data set one by one. The iterations stop until convergence is met. Usually, the predictions for missing data converge in less than 5 iterations. Larger number of iteration is not always necessary because the accuracy of the imputations mainly depends on the information density and completeness of the data set. However, when all chosen variables are completely independent, the process might never converge and yield inaccurate imputations.

pmm is the method to select which values are imputed. For each missing data point, pmm first creates a matching set containing k matches with the observed data, and then does imputation by a random draw from the chosen set. The difference between these two methods is that for each missing value, pmm uses the observed data which result in the closest predicted value to that from the simulated regression model, instead of using all observed data. Therefore, pmm has better performance than the regression method if the data set is not normal. For example, if the original distribution is left-skewed, pmm will produce imputed data with similar distributional pattern. As shown in the pmm mechanism, this method is based on the assumption that the missing data mechanism is MAR. When the data set is far from MAR, pmm will give biased estimates. pmm is based on the observed values, and thus work well for both numerical and categorical data. Also, pmm prevents extrapolation beyond the range of the original data set. (7) Due to its dependence on the observed values, pmm could give good estimates with high speed for large data sets, but may not work well for small samples. The number of iterations, an additional parameter, must be set for MICE because MICE implements an iterative process, which is shown below. Ten iterations are usually performed and the imputed data set from the 10^{th} iteration will be reported. (8)

1. The variable with the least number of missing values is estimated with only complete data.

2. The variable with the second least number of missing values is imputed with the complete data and the imputed values from the first iteration.

3. The iteration is repeated with the complete data and the imputed values from the last iteration.

MICE and random forest imputation are both powerful technique with different pros and cons. MICE gives biased estimates to models with interactions and other non-linear terms, because these terms are omitted in MICE.(9). On the other hand, random forest imputation can handle these non-linear relationships well due to the randomness from the bootstrap aggregation. Random forest imputation also works well in two cases where MICE is not feasible: the number of predictor variables exceeds the number of observations and the imputation model includes highly correlated variables .(10) However, random forest imputation require a larger data set than MICE. MICE is also more flexible because it could handle different variable types by using different method to fill in missing data. Beside pmm, MICE can use linear regression for continuous variables, logistic regression for binary variables, and cardinal variables for data with Poisson distribution.

3.3.3 PCA-based Imputation

PCA-based imputation was first introduced by Dear, R.E in 1959 as a ML-based technique using an expectation-maximization (EM) algorithm to estimate values of missing data. (11) This method is commonly used in studies using traffic and transportation data. PCA-based imputation removes the relatively trivial factors so only the major variables are used to predict the probability distribution of the missing values through dimension reduction. This method also combines the merits of MICE and random forest imputation: high accuracy, reasonable speed, and robustness to abnormal data. We implemented the iterated PCA method using an iterative procedure:

1. missing values are first replaced by random values to generate a complete data set.

2. PCA is applied on this completed data set and missing values are updated by the fitted values using a predefined dimensions, which us 5 in our experiment .

3. The iteration is repeated until convergence.

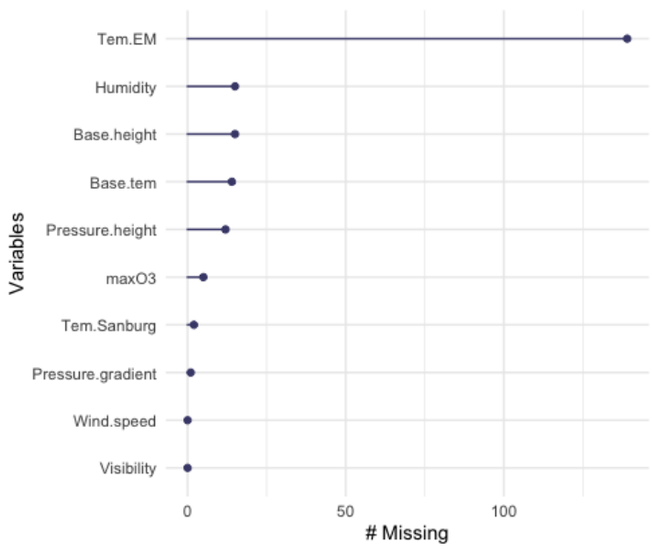
4 Demonstration with Ozone Data

We choose the `Ozone` data set which is widely used in the missing data world in order to demonstrate the methods of handling missing data in casual inference that we have discussed above. The `Ozone` data set is ozone pollution data in Los Angeles collected in 1976. In this section, we present what is the pattern for the data missing, when and why the issue of missing data is not negligible through exploratory data analysis. We then implement a range of imputation methods from naively dropping all the rows carrying missing data to mean imputation, from regression based methods to random forest and PCA methods as explained above. Finally, we apply causal inference framework on the imputed data sets, using difference in mean as the estimator for the average treatment effect estimand. Our goal is to compare the results of the estimated average treatment effect(ATE) and its variance, and then discuss which imputation method performs the best on this particular context.

4.1 Exploratory Data Analysis

`Ozone` data set describes ozone pollution in Los Angeles in 1976. This data set contains 366 observations of daily measurements of maximum one-hour-average ozone concentration `maxO3` and eight meteorological quantities including `Pressure.height`, `Wind.speed`, `Humidity`, `Tem.Sanburg`, `Tem.EM`, `Base.height`, `Pressure.gradient`, `Base.tem`, and `Visibility`. It can be found in R in the `mlbench` package. [1] The data set also specifies the Month, Day, and Day of Week of the measurement, yet for simplicity purposes we choose not to include these columns in our analysis, because the time series information might complicate causal relationships between the meteorological features and `maxO3`.

First, let us take a look at the column of `maxO3`, since we are particularly interested in if there is causal relationship between `Base.tem` and `maxO3` in this observational study. The mean of `maxO3` of the 366 observations is 11.53 without considering five missing data points. The highest `maxO3` is 38.00 and the lowest `maxO3` is 1.00. Since the number of missing `maxO3` is comparatively very small compared to the complete data set, we decide to delete these rows rather than impute the observed Y's.



:

Figure 2: Ozone Missing Data by Variables.

Second, we note that the percentage of missing value in the data set is 5.55% with certain columns having much more missing values than others. Specifically, `Tem.EM` has almost 37.98% missing values followed by `Humidity` (4.1%), `Base.height` (4.1%), `Pressure.height`(3.28%) and `Pressure.gradient` (0.27%). There are no missing values in `Wind.speed` and `Visibility`. Because the percentage of missing values is too high for `Tem.EM` and there are already two covariates related to temperature, we drop this column in the further analysis to avoid introducing additional inference bias.

Given the first glimpse of the missing values, we note that we would only have 203 observations if we delete all the rows with missing values, which is huge reduction from the original data set. In addition,

[1]
`maxO3`: Daily maximum one-hour-average ozone reading
`Pressure.height`: 500 millibar pressure height (m) measured at Vandenberg AFB
`Wind.speed`: Wind speed (mph) at Los Angeles International Airport (LAX)
`Humidity`: Humidity (%) at LAX
`Tem.Sanburg`:Temperature (degrees F) measured at Sandburg, CA
`Tem.EM`: Temperature (degrees F) measured at El Monte, CA
`Base.height`: Inversion base height (feet) at LAX
`Pressure.gradient`: Pressure gradient (mm Hg) from LAX to Daggett, CA
`Base.tem`: Inversion base temperature (degrees F) at LAX
`Visibility`: Visibility (miles) measured at LAX

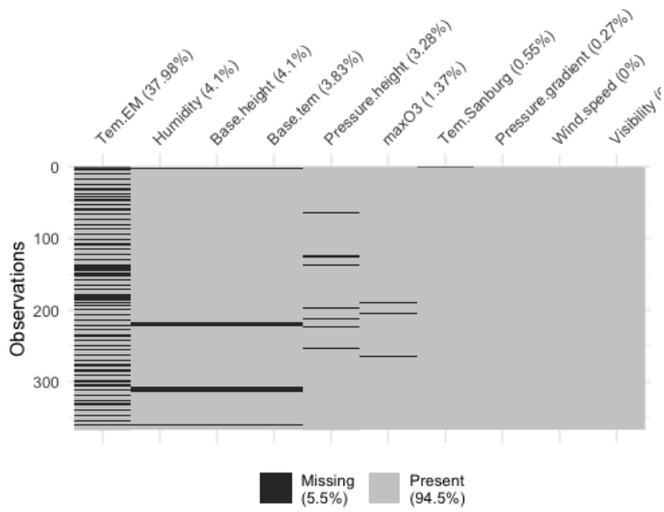


Figure 3: Ozone Missing Data Percentage.

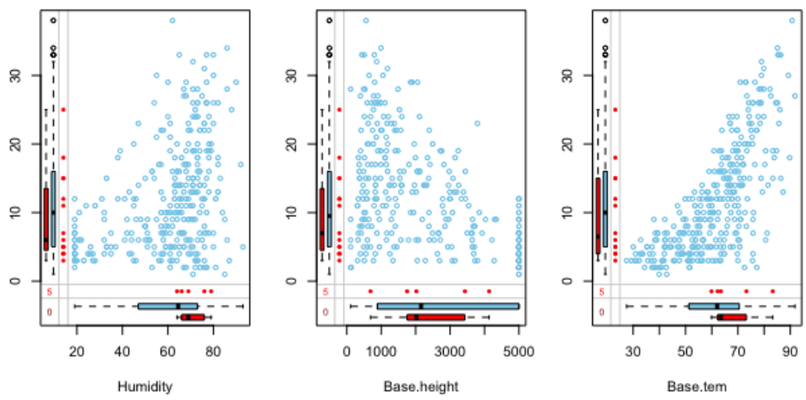


Figure 4: Ozone Missing Data Percentage. Blue box-plots describe data points without missing values. Red box-plots describe data points with missing values.

we notice that the missing values of `Humidity`, `Base.height`, `Pressure.height` then to occur at the same time probably due to some measurement failures, so these values are not MCAR. Figure 4.1 provides a more detailed view of the distribution of the missing data of the three most outstanding features with missing values. Apart from the standard scatter-plot for the data points without any missing values, box-plots of x and y with and without missing values are compared along the axes. For example, we can see that `Base.tem` and `maxO3` are positively correlated from the scatter-plot. The distributions of `maxO3` when `Base.tem` is observed and missing do not differ a lot. Therefore, we could assume that the missing values from `Humidity`, `Base.height`, `Pressure.height` are not likely due to MAR assumption. In short, the `Ozone` data set has a decent amount of missing values. On one hand, for the features where there are too many missing values, we choose to drop the columns completely. On the other hand, we examine the data is most likely missing at completely random because of measurement failures and we will handle these missing values through various ways of imputation. From Figure 4.1 and Figure 5, we know that `maxO3` and `Base.tem` are highly correlated. Is the correlation a result of causation? We will dive into the analysis after handling the missing values through imputation. Also, based on the rule mentioned in 3.3, the number of imputed data set for each method is set as 5, which is close to the percentage of missing data, to achieve both efficiency and statistical power.

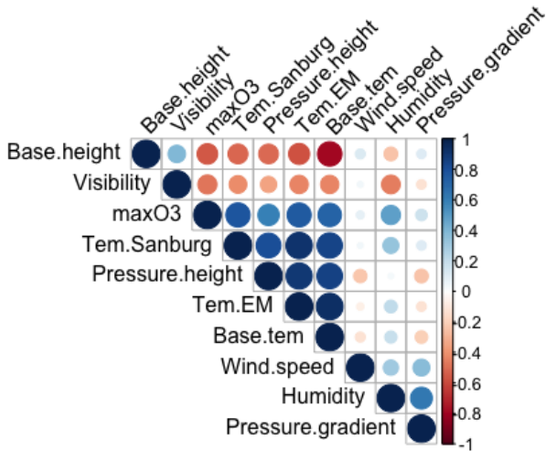


Figure 5: Ozone Missing Data Percentage.

4.2 Imputation

4.2.1 Naive Approach

The most straightforward way to handle missing data is to delete all the rows with missing values. Although we are interested in investigating the performance of different imputation method, we implement the naive approach in order for better comparison. Can imputation reduce the bias of simply deleting all the missing values?

4.2.2 Mean

Another basic imputation method used for comparison is the mean imputation. As stated by its name, we replace each missing value by the mean of the values available for other units. Although it is a fast method and can result in reducing variance in the data set, it can also induce a lot of bias in the results. We principally implement it here as a mean for comparison with other more complex and widely used methods for imputation.

4.2.3 Linear Regression

In addition, we use the linear regression imputation method, which, unlike the two previous methods introduced, leverages the relationships between covariates to produce better results in general. For that effect, it is usually a good step to identify the best predictors of the outcome variable that we analyze first. In practice, we can use an iterative process through which values for the missing variable are inserted and then all cases are used to predict the dependent variable using the regression equation until there is little difference between each step.

The first potential issue with this method is over-fitting the remaining data throughout the process, which would result in an unrealistically low variance. Another issue is that relationships with variables must be guessed, which can result in bias or unsatisfactory results depending on which variables are assumed to be predictive of the outcome.

We will use the package "ImputeTS" and notably the function na.interpolation of the package, which imputes missing value by interpolation using linear regression.

4.2.4 Random Forest

We use missForest to impute missing values. The package can impute mixed-type data. The `missForest()` algorithm makes no assumptions about the relationship between the features unlike MICE which assumes linearity. It is robust to noisy data and has high predictive power. The density distribution of the imputed data is extremely close to the original data according to Figure 10.

4.2.5 MICE with pmm

MICE using pmm is implemented with the R package `mice`. The function `mice()`, `with()`, and `pool()` match three main steps in MI imputation, analysis, and pooling respectively. As shown in Figure 5, variables in the `Ozone` data set are not completely independent, so MICE algorithm should converge in a few iterations. The number of iterations and the number of imputation are set as 5 following the general suggestion. (12)

MICE with pmm yield accurate estimates for either normal or non-normal distribution, as shown in Figure 6. The estimated distributions closely follow the original distribution. Therefore, the bias and the variance are both assumed to be relatively small.

τ estimator and variance				
	Difference in Means	Within Imputation Variance (for MI)	Between Imputation Variance(for MI)	Total Variance
Drop all NA rows	10.2762	NA	NA	0.4304
Mean	9.8143	NA	NA	0.4265
Random Fores	9.7356	NA	NA	0.3958
Linear Regression	9.7557	NA	NA	0.4004
PCA	9.7587	0.4048	0.0108	0.0131
pmm	9.7606	0.4028	0.0020	0.0100
Stochastic Regression	9.7537	0.0401	0.0133	0.0067

Table 1: Results of τ Estimator and Variance from Different Imputation Methods

4.2.6 Stochastic Regression

Stochastic regression imputation is implemented in R with the MICE function with the method `norm.nob`. To better compare the results among different MI method, the number of imputation is also set as 5.

Stochastic regression gives good results when the original data is normal-like, as shown in the result for `Pressure.height` in Figure 4.2.6. For other variables that are either skewed or far from normal distribution, stochastic regression still gives normal-like estimations and thus not yield accurate estimations for standard errors. However, the bias is not significant.

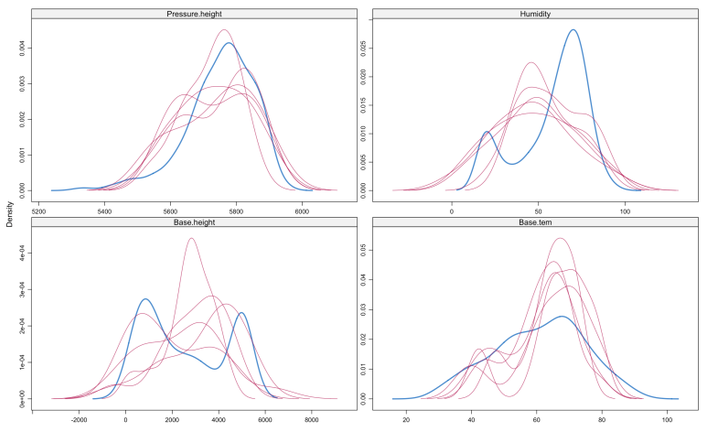


Figure 6: Imputation Result: Stochastic Regression. Blue line describes original data distribution with missing values. Red lines describe data distribution of 5 imputed data sets.

4.2.7 Multiple PCA

Here we use a PCA based model written in R package `missMDA`. This package allows principal component methods to predict the missing values for a predefined dimensions. The number of predefined dimensions can be tuned and we get 5 dimensions as the best number. Figure 4.2.7 visualizes how confident the imputed data. As we can see from the plot, the points are very concentrated after the iterative algorithm and we can interpret the PCA results with high confidence. From Figure 11, we can see that the imputed data has very similar distribution to the original data.

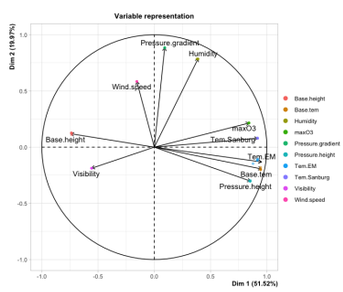


Figure 7: Variability of imputed data after MI PCA method

4.3 Inference

We choose to use difference in means $\hat{\tau}^{DIM}$ as the estimator for the estimand average treatment effect τ^{ATE}.

$$\tau = \tau^{ATE} = \bar{Y}(1) - \bar{Y}(0)$$

$$\hat{\tau} = \hat{\tau}^{DIM} = \bar{Y}^{obs}(1) - \bar{Y}^{obs}(0)$$

More specifically, for single imputation, we calculate the difference in means(DIM) and the conservative variance as provided in the lecture.

$$\hat{\tau} = \frac{1}{N_1(z)} \sum_{i=1}^{N} Z_i Y_i - \frac{1}{N_0(z)} \sum_{i=1}^{N} (1 - Z_i) Y_i$$

where

$$N_1(Z) = \sum_{i=1}^{N} Z_i$$

and

$$N_0(Z) = N - N_1(Z)$$

For the variance, we use the conservative variance:

$$\widehat{Var(\hat{\tau})} = \frac{V_1}{N_1} + \frac{V_0}{N_0}$$

that is derived from following formula.

$$Var(\tau) = \frac{V_1}{N_1} + \frac{V_0}{N_0} - \frac{V_{10}}{N}$$

$$v_a = \frac{1}{N-1} \sum_i \left(Y_i(a) - \bar{Y}_i(a) \right)^2$$

where $a = 0, 1$ and

$$v_{10} = \frac{1}{N-1} \sum_i (\tau_i - \tau)^2$$

In the case of MI with m imputations, we have m imputed data sets and for each data set we can get a $\hat{\tau}_i^{DIM}$ and we can get $\hat{\tau}^{DIM}$ as the mean of $\hat{\tau}_i^{DIM}$ for all $i = 1, 2, \ldots, m$ as

$$\hat{\tau}^{DIM} = \frac{1}{m} \sum_{i=1}^{m} \hat{\tau}_i^{DIM}$$

Since there are multiple imputations, we need to consider both within-imputation variance and between-imputation variance in order to get the total variance of the estimator $\hat{\tau}^{DIM}$.

We can have the within-imputation variance as followed where $\hat{Var}_i$ is the variance of $\hat{\tau}_i$ from each data set defined as above.(13)

$$\widehat{Var_{within}} = \frac{1}{m} \sum_{i=1}^{m} \widehat{Var}_i$$

For the between-imputation variance:

$$\widehat{Var_{between}} = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\tau}_i^{DIM} - \hat{\tau}^{DIM} \right)^2$$

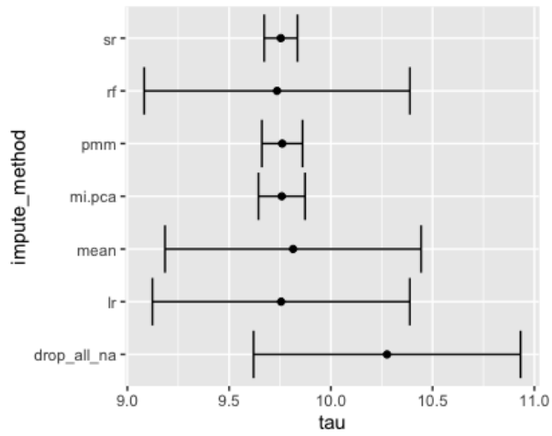
Finally, we can calculate the total variance of $\hat{\tau}$ as:

$$\widehat{Var_{total}} = \widehat{Var_{within}} + \left(1 + \frac{1}{m} \right) \widehat{Var_{between}}$$

In our context, we transform `Base.tem` to Z with $Z = 1$ when `Base.tem` is greater or equal than the mean of `Base.tem` after imputation, and $Z = 0$ when `Base.tem` is less than the mean of `Base.tem` after imputation. Such assignment is straightforward to learn whether there is difference between the potential outcome between the treatment and control groups-namely, if there is causal relationship between `Base.tem` and `max03`.

4.4 Discussion

As expected, the naive approach has large bias. As shown in Table 1, $\widehat{Var_{within}}$ of MI are close to $\widehat{Var}$ of SI. While SI is likely to underestimate variance, the difference in variance is not significant in our experiment due to the correlation among the covariants.



:

Figure 8: Comparison of DIM Average Treatment Effect of Different Imputation Methods

5 Conclusion and Future Work

In our project, we discussed three missing data mechanism, analyzed the pros and cons of three main kinds of imputation for the missing data from their mechanism and assumptions, and implemented 7 imputation methods on a real-life data set to confirm our analysis. The results from our experiment agreed with most of our analysis. MI introduces least bias and yield accurate standard errors among three main kinds of imputation. MI also works well on both normal and non-normal data sets. For the future, we would like to explore the following aspects:

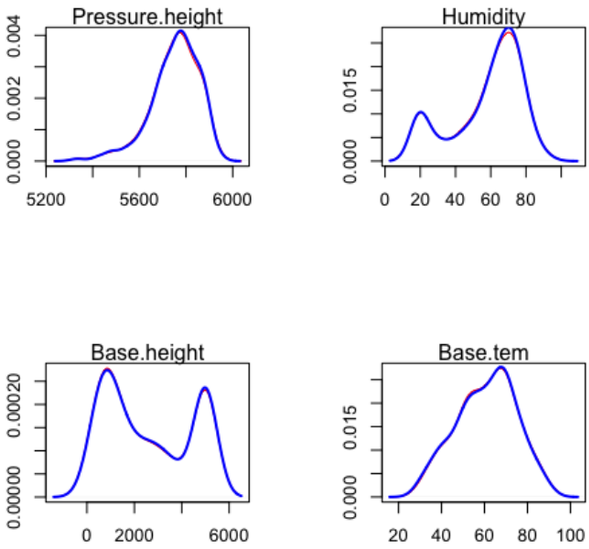
1. See how each kind of imputation perform on data sets with various kinds of variants and data sets based on time series.

2. Explore multilevel multiple imputation. This method is much more flexible than the imputation methods mentioned above due to the lack assumption of the consistent regression among all subsets.

3. Implement more dedicated model selection technique during imputation and more statistical tests for post-imputation analysis .

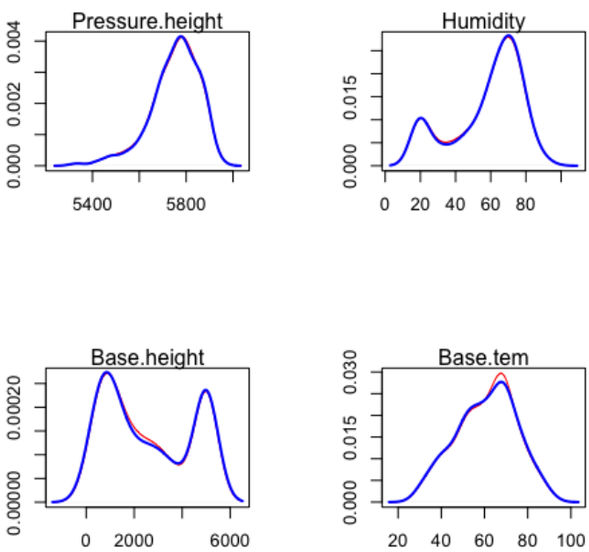
6 Appendices

The link to the project Github repository is:
https://github.com/xingzix/Causal-Inference-with-Missing-Data



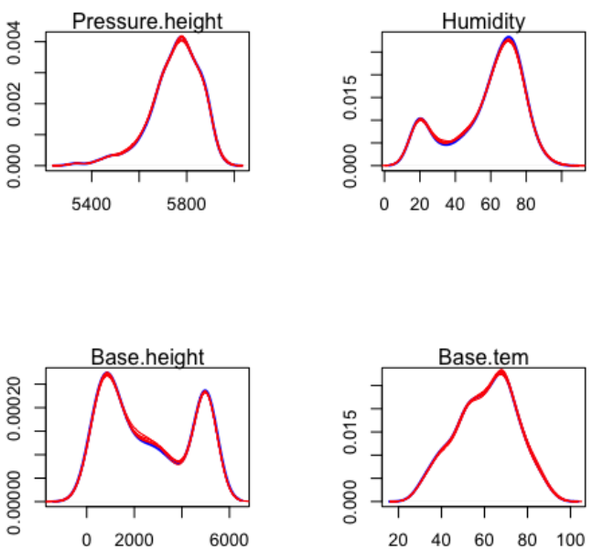
:

Figure 9: Imputation Result: Dropping all missing values. Blue line describes original data distribution with missing values. Red lines describe data distribution of 1 imputed data set.



:

Figure 10: Imputation Result: Random Forest. Blue line describes original data distribution with missing values. Red lines describe data distribution of 1 imputed data set.



:

Figure 11: Imputation Result: Multiple imputation with PCA model. Blue line describes original data distribution with missing values. Red lines describe data distribution of 5 imputed data sets.

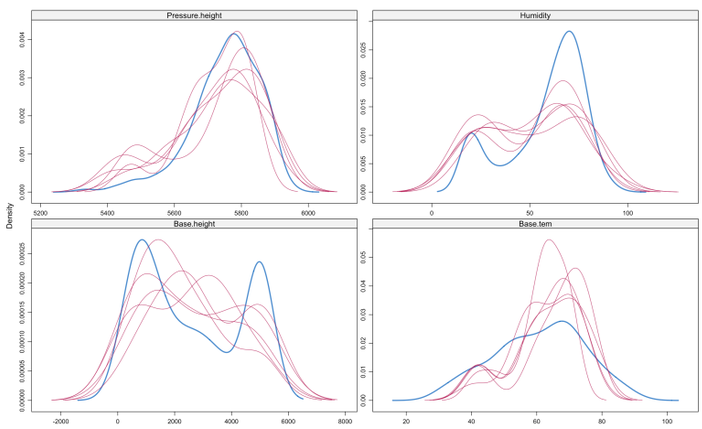


Figure 12: Imputation Result: MICE with pmm. Blue line describes original data distribution with missing values. Red lines describe data distribution of 5 imputed data sets.

References

[1] Rubin, D. Inference and Missing Data. Biometrika, 63(3), 581-592.1976, doi:10.2307/2335739

[2] Rubin, D., *Multiple Imputation After 18 Years. Journal of the American Statistical Association*, 91(434), 473-489, Dec. 1996, doi:10.2307/2291635

[3] Schafer, J. Analysis of Incomplete Multivariate Data. New York: Chapman and Hall/CRC, 1997, doi.org/10.1201/9780367803025

[4] C. Yuan, Y., 2020. Multiple Imputation For Missing Data: Concepts And New Development. 9th ed. [ebook] SAS Institute Inc.

[5] White, I. (n.d.). Multiple imputation using chained equations: Issues and guidance for practice. doi:10.1002/sim.4067

[6] Hastie, T., Tibshirani, R., amp; Friedman, J. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition. New York, NY: Springer New York. 2009

[7] Roderick J. A. Little, A Test of Missing Completely at Random for Multivariate Data with Missing Values, Journal of the American Statistical Association, 83:404, 1198-1202, 1988, doi: 10.1080/01621459.1988.10478722

[8] Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. Am J Epidemiol. 2009 May 1;169(9):1133-9. doi: 10.1093/aje/kwp026. Epub 2009 Mar 24. PMID: 19318618; PMCID: PMC2727238.

[9] Seaman, Shaun R et al. "Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods." BMC medical research methodology vol. 12 46. 10 Apr. 2012, doi:10.1186/1471-2288-12-46

[10] Hardt, Jochen et al. "Auxiliary variables in multiple imputation in regression with missing X: a warning against including too many in small sample research." BMC medical research methodology vol. 12 184. 5 Dec. 2012, doi:10.1186/1471-2288-12-184

[11] Dear, Robert Ernest. A principal-component missing-data method for multiple regression models. System Development Corporation, 1959.

[12] Buuren, Stef van, and Karin Groothuis-Oudshoorn. "Mice: Multivariate Imputation By Chained Equations Inr". Journal Of Statistical Software, vol 45, no. 3, 2011. Foundation For Open Access Statistic, doi:10.18637/jss.v045.i03.

[13] Rubin, D.B. Multiple Imputation for Nonresponse in Surveys, New York: John Wiley Sons, Inc, 1987.