University of Waterloo

Stat 454 - Final Project

Comparing Imputation Strategies on Point Estimation

Author:

Po-Han Lin - 20470883

Professor:

Dr. ChangBao Wu

April 23, 2017

Abstract

Missing data is a common problem that arises in many modern applications. In terms of survey data, missing data is commonly due to non-response of the chosen unit for the sample. Missing data can be classified as unit nonresponse or item nonresponse. If left untreated, this can lead to nonresponse bias. In the case of item nonresponse, missing mechanisms can be further classified, and solutions such as imputation can be used to correct for nonresponse bias. The simulation presented in this paper attempts to investigate the effectiveness of imputation in terms of point estimation, under each missing mechanisms.

Introduction

Missing data can be a common occurrence in many applications in the modern world. In the context of survey data, missing data is predominately due to nonresponse from participants selected for our sample. The main issue with nonresponse in survey is that it can significantly impact statistical analysis and inference, by introducing nonresponse bias when analyzing estimates. For instance, if we let N = $N_1 + N_2$, where N_1 = total population respondents and N_2 = total population nonrespondents, then our sample can also be divided by n_1 = total respondents in our sample and n_2 = total nonrespondents in our sample. If we simply assume n_1 and use it as an estimate to represent the entire population, this will introduce **nonresponse bias**.

Nonresponse patterns can be classified into two cagtegories: unit nonresponse and item nonresponse. In unit nonresponse, all information related to the unit of interest is completely unavailable (i.e the individual is selected but will chooses not to participate). In item nonresponse, however, some information is observed, while other information is missing (i.e individual chooses not to disclose very private information such as income, or individual is under response burden). For item nonresponse, missing patterns can be further explained by classifying missing mechanisms such as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Rubin, 1976). A common solution to help solve the missing data problem under item nonresponse is known as imputation. The main assumption under imputation is that we can use other non-missing covariates as well as other information from the survey to compute a "synthetic" answer to the missing values (Bethlehem, Cobben, & Schouten, 2011). In this project, we will investigate the impact of various imputation techniques on point estimation, while also considering all different missing mechanisms proposed by Rubin.

Main Section

In order to understand the basic premise of imputation simulation presented below, we need to start by further explaining Rubin's missing mechanisms and the various different imputation techniques that are explored.

Missing Mechanisms

In any type of survey, the units in the population of interest all have a probability propensity measure of non response defined as:

$$p_i = P(R_i = 1|y_i, x_i) \tag{1}$$

Here, y_i represents the variable of interest that is subject to missing, while x_i represent a given covariate in the survey that is complete. R_i is an indicator variable indicating whether the unit selected will respond given its observed information is defined as:

$$R_i = \begin{cases} 1, & \text{if } y_i \text{ is missing} \\ 0, & \text{if } y_i \text{ is observed} \end{cases} \tag{2}$$

As mentioned earlier, Rubin defined and classified three different missing mechanisms:

i **Missing Completely at Random (MCAR)** $\implies$ $p_i = p$, some probability constant: this mechanism describes that nonresponse is not correlated to a specific covariate or information that exist in the survey, including itself. Nonresponse is thus due to complete random chance and every observation in our sample is subject to the same probability propensity of not responding.

ii **Missing at Random (MAR)** $\implies$ $p_i = p(x_i)$, some probability function based on x_i: this mechanism describes that nonresponse is correlated to a specific observed covariate(s) in our survey, but not including itself. For example, suppose that our variable of interest, y_i represents the income level of respondents, and is subject to missing. If the relationship between its missing values is associated with a certain covariate such as occupation (i.e certain

profession such as doctors, lawyers, accountants tend to not report their income), then we say that the missing mechanisms is due to MAR.

iii **Not Missing at Random (NMAR)** $\implies p_i = p(y_i)$, some probability function based on y_i: this mechanism describes that nonresponse is correlated to itself, the variable of interest. Again, suppose that our variable of interest, y_i represents the income level of respondents, and is subject to missing. If the relationship between its missing values is associated with with their income levels (i.e people with less income are too embarrassed to report the figure), then we say that the missing mechanism is due to NMAR.

Imputation Methods

There exists a wide variety of methods, when it comes to imputation. Each imputation methods can be thought of as their own prediction model for missing data. In this project, we will be investigating three common imputation methods: Mean, Random Hot-Deck by Class, and Regression. Again, let us suppose that our sample can be divided by $S = S_R + S_N$, where S_R = number of respondents in our sample and S_N = number of nonrespondents in our sample.Finally, the variable that is subject to missing is y_i.

i **Mean Imputation:** under mean imputation, we first calculate the mean under our sample of respondents. That is, we compute:

$$\bar{y}_R = \sum_{i \in S_R} y_i \tag{3}$$

Once we have obtained a value for equation (3), we substitute that mean value for all missing values in our variable of interest such that it satisfies: $\forall$ j $\in S_N$, $y_j = \bar{y}_R$.

ii **Random Hot-Deck by Imputation Class:** The idea behind random hot-deck is very similar to mean imputation. The key difference is, instead of replacing all missing values with a specific deterministic value, we randomly select a value from the pool of respondents, known as the donor pool, to take its place. Thus, $\forall$ j $\in S_N$, $y_j = y_k$, where y_k is randomly selected with replacement from S_R. Furthermore, we can combine hot-deck imputation with the concept of imputation classes. Imputation classes divides our sample into groups that

have similar characteristics. The idea of hot-deck imputation by class is that if the missing variable of interest belongs to a specific class with available donors, we randomly select select from that pool of donors. The main assumption is of course that the donors will exhibit similar values for the missing variable of interest, since they belong in the same class (i.e individuals in the same age groups (classes) should have similar height (missing data)). Thus, random hot-deck by imputation class would look like the following if we have two classes:

$$y_j = \begin{cases} y_i, & \text{if i is in class 1} \\ y_k, & \text{if k is in class 2} \end{cases} \tag{4}$$

Here, both i and k are selected randomly with replacement from their respective donor pools.

iii **Deterministic Regression Imputation:** under regression imputation, we predict the missing values via a fitted regression model. This method specifically, requires the use of a covariate variable, x_i that is not missing for all observation under our sample. Furthermore, since the covariate is used to predict the missing values of interest, it should be strongly correlated to y_i. Thus, our model assumption would be:

$$y_i = \beta x_i + \varepsilon \tag{5}$$

$\forall i \in N_R$. This implies that we can estimate $\hat{\beta}$ using our sample data: $(y_i,\, x_i) \in S_R$. Thus, our estimated missing values can be predicted by: $\hat{y}_j = \hat{\beta} x_j$, where $j \in S_N$

Simulation Study

Data

The data that is used for the simulation study comes from McElreath's github repository, under the name of Howell1 (McElreath, 2016). According to McElreath, the data contained in Howell1 are partial census data for the Dobe area !Kung San people, in Namibia. The original survey was conducted by Nancy Howell in the late 1960s. The dataset includes the variables: height (in cm), weight (in kg), age, sex, and accounts for 544 total observations.

Purpose

For the purpose of our simulation, we are setting weight as our variable of interest, which will be subjected to the three missing mechanisms. The goal of the simulation is to find the most efficient imputation method, in terms of point estimation efficiency, under the three different missing mechanisms that best predicts the population weight of the !Kung San people. The metrics considered by the simulation in terms of simulation will be the estimate itself, the MSE, Relative Bias, Average Interval Length, and Coverage Probability.

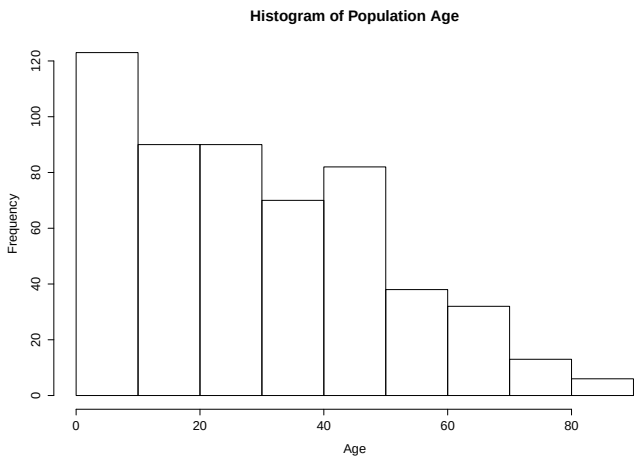
Imputation Setting

Mean Method Setup

The mean method is perhaps the most simple and straightforward imputation procedure in the simulation. It will simply be done by taking the average of all observed units in the sample, store that value, and finally replace all missing cases with the estimate for every iteration.

Random Hot-Deck by Class Setup

The variable age will be used to construct two simple imputation classes: individuals whose age is 35 and above, and those under the age of 35. If you look at the distribution of the frequency histogram below, you can see that the age 35 is the approximate median. The assumption here is that individual who falls in the same age class should also display similar weights.



Thus, our Random Hot-Deck imputation by class becomes the following:

$$y_j = \begin{cases} y_i, & \text{if the age of i is less than 35} \\ y_k, & \text{if the age of unit is 35 and above} \end{cases} \tag{6}$$

Here, both i and k are selected randomly with replacement from their respective donor pools.

Regression Setup

The height variable will be mainly used as a covariate under regression imputation, in order to predict simulated missing height details. Under each iteration, the simulation will fit a regression model according to all observed y_i's (weight) and x_i's (height). After a model is fitted, it will be used to predict all missing variables. Thus, our missing value estimate under regression looks like the following:

$$\hat{y}_i = \hat{\beta} x_i \tag{7}$$

For for all units missing. The assumption here is that we have a complete knowledge of all height for each unit in our population of 544.

Missing Mechanism Setup

Since MCAR is fixed, the user can simply control the overall non-response rate for users by setting line 41 of the R script to a decimal proportion that represents the non-response rate. The Simulation results below used MCAR at rates 10% and 40%.

For MAR a custom propensity function had to be created to recreate an overall population response rate of 10% and 40% for result comparison purposes.

The propensity function is as follows:

$$p2 = \begin{cases} \frac{e^{-4.5+0.00951x_1+0.027x_2}}{1+e^{-4.5+0.00951x_1+0.027x_2}}, & \text{for nonresponse level of 10\%} \\ \frac{e^{-2.5+0.00951x_1+0.02469x_2}}{1+e^{-2.5+0.00951x_1+0.02469x_2}}, & \text{for nonresponse level of 40\%} \end{cases} \tag{8}$$

where, x_1 is the unit's height in cm and x_2 is the unit's age.

For NMAR, a custom propensity function had to be created to recreate an overall population response rate of 10% and 40% for result comparison purposes.

The propensity function is as follows:

$$p3 = \begin{cases} \frac{e^{-4.75+0.0636y_1}}{1+e^{-4.75+0.0636y_1}}, & \text{for nonresponse level of 10\%} \\ \frac{e^{-2.75+0.0629y_1}}{1+e^{-2.75+0.0629y_1}}, & \text{for nonresponse level of 40\%} \end{cases} \tag{9}$$

where, y_1 represents the unit's weight.

How the Simulation Works

The simulation starts by generating probability propensity score columns for the 544 units in the population, which corresponds to the three missing mechanisms. p1 corresponds to MCAR, p2 corresponds to MAR and p3 corresponds to NMAR. Next we use the rbinom() function in R to generate a binary outcome, using the probability generated above (outcome of 1 indicates missing, and 0 for observed). Thus we generate the columns I1, I2, I3, for MCAR, MAR, and NMAR respectively. Based on the method chosen by the user on line 35 ("MCAR", "MAR", "NMAR"), the variable of interest will be subject to the respective missing/observed dictated by the indicator columns (I1, I2, I3).

For Example, if the user chose "MAR", the dataframe looks like the following in R:

	height	weight	age	male	p1	p2	p3	I1	I2	I3	y1	y2	y3
1	151.7650	47.825606	63.00	1	0.1	0.6221667	0.15338546	0	0	1	47.825606	47.825606	47.825606
2	139.7000	36.485807	63.00	0	0.1	0.5948416	0.08095046	1	1	0	NA	NA	NA
3	136.5250	31.864838	65.00	0	0.1	0.5994569	0.06160683	0	1	0	NA	NA	NA
4	156.8450	53.041914	41.00	1	0.1	0.5009715	0.20156619	1	1	0	NA	NA	NA
5	145.4150	41.276872	51.00	0	0.1	0.5354620	0.10671052	1	1	0	NA	NA	NA
6	163.8300	62.992589	35.00	1	0.1	0.4805531	0.32220185	0	1	0	NA	NA	NA
7	149.2250	38.243476	32.00	0	0.1	0.4278104	0.08966640	0	1	0	NA	NA	NA
8	168.9100	55.479971	27.00	1	0.1	0.4434836	0.22767719	0	1	0	NA	NA	NA
9	147.9550	34.869885	19.00	0	0.1	0.3489091	0.07362631	0	1	0	NA	NA	NA
10	165.1000	54.487739	54.00	1	0.1	0.5994949	0.21677165	0	1	0	NA	NA	NA
11	154.3050	49.895120	47.00	0	0.1	0.5319242	0.17126756	1	1	0	NA	NA	NA
12	151.1300	41.220173	66.00	1	0.1	0.6380213	0.10636727	0	0	0	41.220173	41.220173	41.220173
13	144.7800	36.032215	73.00	0	0.1	0.6635663	0.07882999	0	0	0	36.032215	36.032215	36.032215

The columns y1, y2, y3 will be subject to missing or observed based on the I2 indicator column, which correspond to the MAR missing mechanism. Three variable of interest columns (y1,y2, y3) will correspond to each different imputation process (Mean, Hot-Deck, and Regression). The simulation will repeat based on the number of repetition specified (reps = 500). Finally, the user will see the results printed in the R console, or you can access it by clicking the "Summary" dataframe created.

How to Run the Simulation

First off, make sure you have the data file on a specific directory for the computer you are running the simulation on. Either save the data in the default directory or change the file directory code to the specific directory, where the dataset (csv file) is stored on lines 12 and 71 in the R script.

In order to run the simulation, the user must make manual changes to lines 35, 41, 78, and 81, depending on the missing mechanism chosen. There is a total of 6 tables, meaning you have to make 6 separate simulation runs in order to replicate the same results. The seed was set at 23 for the results below and a total of 500 iterations were ran for each of the 6 results (reps = 500, line 24).

Steps for MCAR:

1. set line 35 to "MCAR" (like a string)

2. set line 41 to non-response rate desired (default is 0.1)

3. Run the Whole Script, including the seed at the start.

Steps for MAR:

1. set line 35 to "MAR" (like a string)

2. There are already customized propensity score functions defined as variables: e (40%) and e3 (10%) for MAR in the R script.

3. Simply modify line 78 to the desired propensity score function, by adding or removing the number 3 next to the e. For example, if you want the results in table 5, set line 78 to read like the following: pop data$p2 = e3/(1 + e3), (e3 is score propensity function corresponding to non-response for MAR at an overall level of 10%.

4. Run the Whole Script, including the seed at the start.

Steps for NMAR:

1. set line 35 to "NMAR" (like a string)

2. There are already customized propensity score functions defined as variables: e2 (40%) and e4 (10%) for NMAR in the R script.

3. Simply modify line 81 to the desired propensity score function, by changing the number next to the "e" to 2 or 4. For example, if you want the results in table 6, set line 81 to read like the following: pop data$p3 = e4/(1 + e4), (e4 is score propensity function corresponding to non-response for NMAR at an overall level of 10%.

4. Run the Whole Script, including the seed at the start.

Simulation Results

The true population weight is $\mu_y = 35.61062$ kg

MCAR:

To obtain these results, set line 35 to "MCAR" and line 41 to 0.4

Imputation Estimate under MCAR					
$p = 0.4$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.54408	2.812032	-0.0018684231	3.823657	0.750
Hot-Deck by Class	35.58731	3.137407	-0.0006544217	4.930010	0.832
Regression	35.52110	1.874904	-0.0025136904	4.859806	0.920

MAR:

To obtain these results, set line 35 to method = "MAR" and line 78 to *pop_data*\$*p*2 = e/(1 + e).

Imputation Estimate under MAR					
$p = 0.4000472$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.63443	4.975482	0.0006687114	3.0096567	0.478
Hot-Deck by Class	35.59934	4.759867	-0.0003166367	4.893753	0.738
Regression	35.47504	2.064888	-0.0038073560	4.811667	0.902

NMAR:

To obtain these results, set line 35 to method = "NMAR", line 81 to *pop_data*\$*p*3 = e2/(1 + e2).

Imputation Estimate under NMAR					
$p = 0.400495$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.60921	4.458560	-3.964403e-05	3.222287	0.542
Hot-Deck by Class	35.67610	4.159384	1.838922e-03	4.870242	0.746
Regression	35.56640	2.212723	-1.241834e-03	4.790271	0.892

Simulation Results Continued

This time, we will see what happens when nonresponse rate is 10 % for MCAR, and approximately 10% for both MAR and NMAR.

MCAR:

To obtain these results, set line 35 to "MCAR" and line 41 to 0.1

Imputation Estimate under MCAR					
$p = 0.1$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.43621	1.943373	-0.004897661	4.678693	0.906
Hot-Deck by Class	35.47329	1.959348	-0.003856336	4.928398	0.922
Regression	35.49662	1.717967	-0.003201134	4.903183	0.942

MAR:

To obtain these results, set line 35 to method = "MAR" and line 78 to $pop\_data\$p2 = e3/(1 + e3)$.

Imputation Estimate under MAR					
$p = 0.1017226$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.62484	2.295332	0.0003994712	4.398555	0.846
Hot-Deck by Class	35.62529	2.484200	0.0004120832	4.938985	0.876
Regression	35.58928	1.866574	-0.0005993308	4.903760	0.914

NMAR:

To obtain these results, set line 35 to method = "NMAR", line 81 to *pop_data*$*p*3 = e4/(1 + e4).

Imputation Estimate under NMAR					
$p = 0.1008514$	Estimate	MSE	Relative Bias	Average Length	Coverage Probability
Mean	35.54817	1.812933	-0.0017536655	4.550644	0.910
Hot-Deck by Class	35.55398	2.057959	-0.0015903952	4.931293	0.922
Regression	35.57801	1.555769	-0.0009156423	4.894004	0.956

Analysis

The simulation results showed that the final estimates of all six runs were estimated very close with relatively low bias. Furthermore, when the nonresponse was approximately 10% for the overall population under all three mechanisms, we could see that the estimates and the coverage probability do not vary by much for all three imputation methods. The reason behind this, is due to the fact that most of our sampled data are observed, and thus most of the variance for our variable of interest is still preserved, even after accounting for the imputed values in the mixture. This leads to a much higher probability coverage, regardless of the missing mechanism chosen or imputation method performed. On the other hand, when the overall nonresponse rate was set to approximately 40% for the overall population under all three missing mechanisms, we could clearly see that the coverage probability drops, for each imputation method performed. Again, this is due to the fact that the simulation calculates the sample variance using both the observed data and the imputed values. As the number of missing data increases, we rely more and more on the estimated values from our imputation models, which in turn leads to an underestimation of the true variance, as the imputated values have a more significant impact on variance. We can see this effect was present in all three imputation methods, under all three missing mechanisms.

In terms of comparing the missing mechanisms, when we considered MCAR, all units had equal chance of not responding, which implies that we could still expect our sample to be representative of the population, despite missing observations. Thus, all imputation methods performed reasonably well, even under the scenario where the average population nonresponse was 40%. When MAR

was chosen, (equation 8), the equation gave a higher nonresponse score to observations who were taller and older, due to the covariates used. In turn, this leads our expected sample to have more shorter and younger observation (as they have a lower probability propensity score of not responding). Similarly, under NMAR observations who are heavier were less likely to respond. Since our imputation methods cannot account for this lack of information, we can clearly see that the MSE under all three methods increased, going from MCAR to MAR and NMAR. Note that under regression imputation, our model captures the weight and height relationship, so the MSE did not increase by much, despite not accounting for the observation's age.

Conclusion

Imputation methods were considered as a solution to correct for nonresponse bias in item nonresponse. However, several factors need to be considered, when choosing the most suitable imputation method for missing data in a given survey for point estimation. When we have a relatively low frequency of unobserved data for our variable of interest, we can choose any of the three methods presented to estimate the missing data, without jeopardizing the estimation. However, in the presence of high count of missing data, missing mechanisms must be considered before choosing a specific imputation method. If the missing mechanism is under the assumption of MCAR, then regression or hot-deck imputation by class could work well, if you have some complete set of covariates that correlates with the variable of interest. Using mean imputation may not be suitable, as it may severely underestimate or overestimate the true value. If the missing mechanisms detected is under MAR or NMAR, all three imputation methods will not work well if the imputation class or covariate does not capture the underrepresented units related to the response propensity model.

Acknowledgements

I would like to thank Prof. Changbao Wu for all his guidance and support. This project could not have been accomplished without his help. The details and explanations presented in this project was taught and motivated in lectures. The simulation approach and design was also heavily under his guidance.

References

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys* (Vol. 568). John Wiley & Sons.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan* (Vol. 122). CRC Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581–592.