Incomplete Data Analysis for Economic Statistics

By
Masayoshi Takahashi

経済統計のための不完全データ解析

高橋　将宜

A Dissertation
submitted to
the Department of Computer and Information Science
Seikei University
in partial fulfillment of the requirements
for the degree of
Doctor of Science and Technology
2017

成蹊大学大学院　理工学研究科　理工学専攻情報科学コース
博士論文

Dissertation Committee:
Professor Manabu Iwasaki, Committee Chair
Professor Kimio Oguchi
Professor Yukiko Nakano
Professor Jinfang Wang

Abstract

Incomplete Data Analysis for Economic Statistics
By
Masayoshi Takahashi

Incomplete data are ubiquitous in social sciences; as a consequence, available data are inefficient and often biased. This dissertation deals with the problem of missing data in official economic statistics. Building on the practices of the United Nations Economic Commission for Europe (UNECE), the first half of the dissertation focuses on single imputation methods. After revealing that single ratio imputation is often used for economic data in the current practices of official statistics, this study unifies the three ratio imputation models under the framework of weighted least squares and proposes a novel estimation strategy for selecting a ratio imputation model based on the magnitude of heteroskedasticity. After showing that multiple imputation is suited for public-use microdata, the latter half of the dissertation focuses on multiple imputation methods. From a new perspective, this dissertation compares the three computational algorithms for multiple imputation: Data Augmentation (DA), Fully Conditional Specification (FCS), and Expectation-Maximization with Bootstrapping (EMB). It is found that EMB is a confidence proper multiple imputation algorithm without between-imputation iterations, meaning that EMB is more user-friendly than DA and FCS. Based on these findings, the current study proposes a novel application of the EMB algorithm to ratio imputation in order to create multiple ratio imputation, the new multiple imputation version of ratio imputation, providing brand-new software *MrImputation* implemented in *R*. Combining all of these findings, this dissertation will be an important addition to the literature of missing data analysis and official economic statistics.

Keywords: Missing data; multiple imputation; ratio imputation; official statistics

キーワード：欠測データ；欠損；多重代入法；比率代入法；補完；補定；公的統計

Acknowledgments

I deeply appreciate Dr. Manabu Iwasaki for supervising the dissertation. In the past two years, I attended the Seikei Statistics Laboratory Seminar held by Dr. Iwasaki, where I presented all of the materials related to this dissertation. I acquired a tremendous amount of expertise on missing data analysis from Dr. Iwasaki in this seminar. Without his support and suggestions, this dissertation would never come to fruition. I also would like to thank Dr. Kimio Oguchi (Seikei University), Dr. Yukiko Nakano (Seikei University), and Dr. Jinfang Wang (Chiba University) for being the members of the dissertation committee.

The origin of this dissertation can be traced back to the Advanced Political Methodology Seminar by Dr. Valentina Bali (Michigan State University) and the ICPSR Applied Bayesian Modeling Seminar by Dr. Ryan Bakker (University of Georgia), where I first learned about multiple imputation in the context of political science research. Since then, I have been fascinated by the beauty of multiple imputation.

Part of this dissertation is related to my work at the National Statistics Center (NSTAC), an umbrella organization under the Statistics Bureau of Japan. My research at NSTAC revolved around the issue of data editing (error detection and correction) in the context of official statistics. Special thanks go to Dr. Michiko Watanabe (Keio University) for constantly encouraging me to conduct research on imputation methods when she was NSTAC Vice-President. Also, I would like to thank Mr. Nobuyuki Sakashita (SRTI: Statistical Research and Training Institute), Mr. Yoshiyuki Kobayashi (SRTI), Mr. Tatsuo Noro (SRTI), and Mr. Takayuki Ito (NSTAC) for exchanging ideas and thoughts on imputation methods in official statistics when we were colleagues at the National Statistics Center.

I wish to thank Dr. Shinsuke Ito (Chuo University) for inviting me to the Japan Society of Economic Statistics, where the work related to Chapter 2 was developed. Special thanks go to the participants of the UNECE Work Session on Statistical Data Editing, who responded to my surveys for Chapter 2. I also wish to thank Dr. Hiroe Tsubaki (NSTAC) for his insightful idea

about ratio imputation, which helped me to complete the research related to Chapter 3. Earlier versions of Chapters 4 and 5 benefited from the feedback of Dr. Takayuki Abe (Keio University) and Dr. Tetsuto Himeno (Shiga University). Special thanks go to Mr. Yutaka Abe (NSTAC) for reviewing the mathematical proofs in the Appendix of Chapter 3 and for his valuable comments on an earlier version of Chapter 6.

The main chapters of this dissertation are based on the peer-reviewed articles that I previously published. The permission to use these articles was explicitly obtained from the publishers. I would like to thank the Japan Society of Economic Statistics for permission to use "Missing data treatments in official statistics: Imputation methods for aggregate values and public-use microdata" (*Statistics*, no.112, 65-83), IOS Press for permission to use "Imputing the mean of a heteroskedastic log-normal missing variable: A unified approach to ratio imputation" (coauthored with Iwasaki, M. and Tsubaki, H., *Statistical Journal of the IAOS*, vol.33, no.3, in press), Ubiquity Press for permission to use "Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations" (*Data Science Journal*, in press), and JMASM Inc. for permission to use "Multiple ratio imputation by the EMB algorithm: Theory and simulation" (*Journal of Modern Applied Statistical Methods*, vol.16, no.1, 630-656) and "Implementing multiple ratio imputation by the EMB algorithm (R)" (*Journal of Modern Applied Statistical Methods*, vol.16, no.1, 657-673).

Last but not least, my heartfelt gratitude goes to my family. My parents, Shigeyoshi and Miyuki, have always been supportive and understanding throughout the long and daunting journey to the completion of the dissertation. My grandmother, Tsuyako Tachibana, has cared about me in many respects while I grew up. My wife, Yoko, has constantly given me unwavering love that boosted the completion of this dissertation. Without their love and support, this dissertation would never become complete. I dedicate this dissertation to my family.

Table of Contents

1 Introduction

Social science data are often incomplete, which leads to inefficient and biased analyses. This dissertation is a study of the missing data problem in official economic statistics. Specifically, this dissertation deals with single and multiple imputation methods. This chapter explains why missing data are problematic. It also shows the structure of the dissertation.

Many surveys in official statistics are based on samples because of the time and budgetary constraints. A sample is usually chosen as a subset of the population, via some type of random sampling techniques. In so doing, a random sample may be considered unit nonresponse, where some respondents do not answer in the survey, meaning that the rows for some respondents are all blank (de Waal et al., 2011, p.223). As long as sampling is random, the sampling error can be numerically assessed. Therefore, King et al. (2001, p.49) argue that unit nonresponse in social sciences generally does not introduce much bias into analyses. In fact, as the sample size increases, the law of large numbers predicts, the sampling error tends to become small (Weiss, 2005, p.329; Ross, 2006, p.443). This may lead us to believe that, if we increase the number of observations, there will be no error in data.

A census is said to be a method to acquire information on the entire population of interest (Weiss, 2005, p.11). Once in a while, official statistical agencies have the luxury of conducting censuses in order to establish the population framework. One such example is the Economic Census for Business Activity, which aims at covering all of the enterprises and establishments in Japan. Since the census aims at obtaining information on the entire population of interest, there are no sampling errors in the census.

However, non-sampling errors and missing data occur in the measurement process of official statistics (de Waal et al., 2011, p.2). Measurement is said to be the process where numbers are assigned to objects in meaningful ways, where measurement errors occur when the attributes of empirical objects are assigned to numerical values due to the imperfect functional translation (Jacoby, 1991; Jacoby, 1999). If no numerical values are assigned to empirical objects, missingness occurs due to measurement errors. In light of this, it is highly unlikely that the observations in the Economic Census will be complete. In fact, the second-term Master Plan Concerning the Development of Official Statistics (adopted by the Japanese Government in 2014) points out that it is generally difficult to obtain values for accounting items from enterprises and establishments. In other words, non-accounting items may be answered, but accounting items may not be answered by some enterprises and establishments. Although the census aims at obtaining information on the entire population of interest with no sampling error, it is likely that the census is incomplete data due to missingness.

This situation represents item nonresponse, where partial responses are available from some respondents, meaning that missing values are scattered in a data matrix (de Waal et al., 2011, p.223). To be clear about the topic of interest, this dissertation is about item nonresponse. King et al. (2001, p.49) contend that item nonresponse is more serious than unit nonresponse. In fact, whether the survey is a census or a sample does not change the fact that there are almost always some respondents who do not answer some questions in the survey. In other words, incomplete data are ubiquitous in official economic statistics, whether it is a census or a sample. When some values are missing, available data are inefficient at best and often biased at worst, without explicitly taking missing values into account. Unfortunately, this bias does not disappear when the number of observations is increased.

In light of this, the current study deals with the problem of missing data in official economic statistics, where most variables are continuous, rather than categorical. Specifically, this dissertation is a study of imputation methods, which are known to be able to rectify the missing data problem under certain assumptions. By way of organization, Chapters 2 to 6 are the body of the dissertation. Chapters 2 and 3 are mainly concerned with single imputation methods (part of Chapter 2 also with multiple imputation). Chapters 4, 5, and 6 deal with multiple imputation methods. Below is the synopsis of each chapter.

Chapter 2 reveals the status quo of official statistics around the world. For this purpose, Chapter

2 builds on the practices of the United Nations Economic Commission for Europe (UNECE). Chapter 2 shows that, in the current practices of imputation among the national statistical institutes, ratio imputation is often used and indeed suitable for economic data. Also, Chapter 2 reveals that hot deck imputation is often used and indeed suitable for household data. Both of these methods are deterministic single imputation. Furthermore, Chapter 2 assesses deterministic single imputation, stochastic single imputation, and multiple imputation, demonstrating that multiple imputation is suited for public-use microdata. Therefore, this chapter indicates that the future practice of official economic statistics would need to be changed from single imputation to multiple imputation.

As is made clear in Chapter 2, ratio imputation is often used to treat missing values in official economic statistics. However, there are three competing estimators in the literature: Ordinary least squares; ratio of means; and mean of ratios. A natural question arises. Under what circumstances, which method should we use? Chapter 3 answers this question by unifying ratio imputation models under the framework of weighted least squares. Furthermore, Chapter 3 proposes a novel estimation strategy for selecting a ratio imputation model based on the magnitude of heteroskedasticity. The results in the Monte Carlo simulation give a strong support for the proposed method. This chapter should be not only academically important, but also practically useful, in choosing the best imputation method for a given dataset in an economic survey.

As Chapter 2 indicated, the future course for official statistics would be multiple imputation. Therefore, Chapter 4 shifts gears from single imputation to multiple imputation, and compares the three computational algorithms for multiple imputation: Data Augmentation (DA), Fully Conditional Specification (FCS), and Expectation-Maximization with Bootstrapping (EMB). In the literature, many comparative studies are available from the perspectives of joint modeling (DA, EMB) and conditional modeling (FCS), which shows that joint modeling is computationally more efficient and conditional modeling is more flexible. However, little is known about the relative superiority between the MCMC algorithms (DA, FCS) and the non-MCMC algorithm

(EMB), where MCMC stands for Markov chain Monte Carlo. Based on simulation experiments, Chapter 4 contends that, while DA and FCS are not confidence proper without between-imputation iterations, EMB is confidence proper even without between-imputation iterations; thus, EMB is more user-friendly than DA and FCS.

Chapters 2 and 3 demonstrate that ratio imputation is often employed to deal with missing values in the practices of official economic statistics. Chapter 4 demonstrates that EMB is a useful multiple imputation algorithm. Since the literature is devoid of multiple ratio imputation, Chapter 5 proposes a novel application of the EMB algorithm to ratio imputation, so as to create the multiple imputation version of ratio imputation. Chapter 5 presents the mechanism of multiple ratio imputation and assesses the performance compared to traditional imputation methods using Monte Carlo simulation to establish the usefulness of multiple ratio imputation. Furthermore, Chapter 6 outlines a concrete code in the *R* statistical environment to execute multiple ratio imputation by the EMB algorithm and provides brand-new software, *MrImputation* implemented in *R*. Readers can use this software by simply copying and pasting these codes in *R*. Thus, this chapter should be practically useful. Finally, Chapter 7 summarizes the central findings in the dissertation and indicates the possible future courses of research. Combining all of these five chapters together, the author believes that this dissertation will be an important addition to the literature of missing data in particular and official statistics in general.

All of the chapters in the body of this dissertation are based on the peer-reviewed articles written by the author. The permission to use these articles was explicitly obtained from the publishers (See Acknowledgements). Chapter 2 is based on Takahashi (2017a), a peer-reviewed article in *Statistics*, which is the official journal of the Japan Society of Economic Statistics, a corporative science and research body of the Science Council of Japan. Chapter 3 is based on Takahashi et al. (2017), a peer-reviewed article in the *Statistical Journal of the IAOS*, which is the flagship journal of the International Association for Official Statistics (IAOS) under the umbrella of the International Statistical Institute (ISI). Chapter 4 is based on Takahashi (2017d), a peer-reviewed article in the *Data Science Journal*, which is sponsored by CODATA (Committee on Data for Science and Technology), an interdisciplinary scientific committee of the International Council for Science (ICSU). Chapters 5 and 6 are based on Takahashi (2017b) and Takahashi (2017c), both of which are peer-reviewed articles in the *Journal of Modern Applied Statistical Methods*, which is operated by the Wayne State University Library System, classified as one of the top 115 libraries in the United States by the Association for Research Libraries (Kyrillidou et al., 2015). Note that the *Statistical Journal of the IAOS*, the *Data Science Journal*, and the *Journal of Modern Applied Statistical Methods* are indexed in Scopus by Elsevier as of April 2017.

2 Imputation Methods in Official Statistics: Current and Future Perspectives

This chapter derived from Takahashi (2017a), a peer-reviewed article in *Statistics* (112), which is the official journal of the Japan Society of Economic Statistics, a corporative science and research body of the Science Council of Japan. The author would like to thank the Japan Society of Economic Statistics for permission to use "Missing data treatments in official statistics: Imputation methods for aggregate values and public-use microdata" (*Statistics*, no.112, 65-83).

2.1 Introduction

About half of the respondents generally do not answer at least one question in social surveys (King et al., 2001, p.49). Especially, the response rate tends to be low in sensitive items such as the income of individuals and the turnovers of enterprises (Schenker et al., 2006). Furthermore, respondents may unintentionally overlook or forget to answer a question. Also, if respondents change their addresses or an enterprise goes bankrupt, then values in longitudinal surveys will be inevitably missing (Allison, 2002; de Waal et al., 2011).

Thus, it is quite difficult to collect all data points in a social survey, which implies that the statistical treatment of missing values is an indispensable process in the official statistical production. While missing values in official statistics are often dealt with by imputation (de Waal et al., 2011, Ch.7), the methodological importance of imputation has rarely been discussed in Japan. On the other hand, the origin of imputation methods in official statistics can be traced back to the 1950s (U.S. Bureau of the Census, 1957, p.XXIV), which amounts to a huge body of the literature. For example, in the context of the statistical production process of microdata in official statistics, imputation methods have been the topic of debate at international conferences, such as the Work Session on Statistical Data Editing by the United Nations Economic Commission for Europe (UNECE).

In light of the findings reported at UNECE, this chapter reviews the methodological development for missing value treatments in official statistics. The first half of the chapter surveys the methods used by the UNECE member states, examining the characteristics of the imputation methods for specific types of surveys such as economic surveys and household surveys in the traditional aggregation-based imputation methods.

Furthermore, while analyses were often based on macro-level data in the 20th century, it is recognized that more and more empirical analyses are based on micro-level data in the 21st century (Sakata, 2006, p.31). Under such a circumstance, there is an increasing demand that the survey data collected by official statistics should be made openly available as public-use microdata. On the supply side as well, the second-term Master Plan Concerning the Development of Official Statistics was adopted by the Japanese Government in 2014, which mentions the utilization of official statistics, where microdata will be experimentally made available at some on-site facilities (Nakamura and Hirasawa, 2016, pp.36-37). This shows that Japan has just taken a step forward in the path of public-use microdata. Thus, the latter half of the chapter discusses the future challenges about how the imputation methods for public-use microdata are different from the ones used in the traditional official statistics production process.

Note that the term “public-use microdata” in this dissertation does not imply a specific way of providing a microdata service. Unlike the traditional analysis of relying on the aggregated values tabulated by official statistical agencies, this dissertation assumes the environment where the analysts can analyze data at their discretion. In other words, the term “public-use microdata” in this dissertation refers to a larger conception that includes “public-use microdata sample,” “anonymized microdata,” and “original data (at the individual level).” Generally, the term “public” in this context is used for open data supplied to the general public, not for microdata supplied to the scholars who meet the terms and conditions of using microdata. However, the important point of the discussion in this chapter is that the imputers (survey organization) and the analysts (the general public and the scholars) are different entities. This chapter does not differentiate the general public and the scholars. Therefore, the term “public-use microdata” in this chapter includes the cases where the scholars are the analysts using the microdata provided by the national statistical agencies.

Also, note that the discussion in this chapter is limited to the topic of missing values in public-use microdata, assuming that enough level of disclosure limitation is already taken care of. For a detailed discussion on confidentiality and usability of anonymized data, see Ito and Hoshino (2014).

2.2 Problems of Missing Data

Tables 2.1, 2.2, and 2.3 are simulated data of income and age for four people. All data are continuous in Table 2.1. Age is categorical in Table 2.2. Income is categorical in Table 2.3. Black numbers are observed values, and white numbers in gray cells are the true values of the missing values. Let us assume that the estimands in Tables 2.1 and 2.2 are the mean of income, and the estimand in Table 2.3 is the mode of income.

Table 2.1
Quantitative
Data

ID	Income	Age
1	239	26
2	421	38
3	505	47
4	650	54

Table 2.2
Quantitative/Qualitative
Data

ID	Income	Age
1	239	1
2	421	1
3	505	2
4	650	2

Table 2.3
Qualitative/Quantitative
Data

ID	Income	Age
1	1	26
2	2	38
3	3	47
4	3	54

Note: The unit of income is 10,000 yen, and the unit of age is a year. In Table 2.2, 1 = less than 40 years old, 2 = 40 years or above. In Table 2.3, 1 = 0 and 2.49 million, 2 = 2.5 million and 4.99 million, 3 = 5 million or above. Tables 2.2 and 2.3 will be used in the next sections.

In Table 2.1, if all the data are observed, then the mean income of the four people can be easily calculated as 453.75 in equation (2.1).

$$\overline{\text{Income}}_{\text{true}} = \frac{1}{4}\sum_{i=1}^{4} \text{Income}_i = \frac{239 + 421 + 505 + 650}{4} = 453.75 \tag{2.1}$$

On the other hand, as we can see in equation (2.2), even one missing value makes the calculation of the mean impossible. The fact that the mean cannot be calculated implies that the standard statistical analyses such as the computation of the standard deviation, correlation, regression coefficient, and standard error are also impossible. In other words, the primary problem of missing values is the impossibility of statistical analysis without treating missing values.

$$\overline{\text{Income}}_{\text{missing}} = \frac{1}{4}\sum_{i=1}^{4} \text{Income}_i = \frac{239 + 421 + 505 + \text{Income}_4}{4}$$
$$= \frac{1155 + \text{Income}_4}{4} = ? \tag{2.2}$$

When a cell is missing in a row, then the row is deleted in the default setting of statistical software such as SAS, SPSS, and STATA. In this way, the data will be seemingly "complete," making statistical analysis possible. This method is called listwise deletion, also known as complete case analysis and casewise deletion (Baraldi and Enders, 2010, p.10). In Table 2.1, we treat ID4 as if it were not there; thus, the mean of income is 388.33 as in equation (2.3). However, this example shows that the true mean value is 453.75, which is underestimated due to bias in missing data. Furthermore, the valuable information about $\text{Age}_4 = 54$ is not utilized, but thrown away. The secondary problem of missing data is that the analyses based on missing data may be biased and inefficient, where bias means the difference between the expected value of the estimator and the true parameter value and efficiency means the size of variance for the estimator that becomes large as the sample size n becomes small.

$$\overline{\text{Income}}_{\text{listwise}} = \frac{1}{3}\sum_{\text{i}=1}^{3} \text{Income}_{\text{i}} = \frac{239 + 421 + 505}{3} = 388.33 \tag{2.3}$$

2.3 Assumptions of Missing Data Mechanisms

Little and Rubin (2002) proposed the three classifications of missing data mechanisms: Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR). NMAR is sometimes referred to as MNAR (Missing Not At Random), but these two are exactly the same concepts. For a more detailed discussion on the missing data mechanisms, also see Sections 3.2, 4.4, and 5.3.

Under MCAR, missing data can be considered a subsample of the population, leading to no bias, but reducing the efficiency. Under MAR, missing data may be biased. As Allison (2002, p.5) points out, we can ignore the parameters of the missing data mechanism under MCAR and MAR;

thus, MCAR and MAR are ignorable. As a result, imputation can rectify the bias in missing data. On the other hand, under NMAR, the missing data mechanism is non-ignorable. The selection model and pattern-mixture model can be used to tackle the issue of non-ignorable missing data mechanisms, but these models require strong assumptions (Allison, 2002, ch.7; Enders, 2010, ch.10). As we will see later in this chapter, these methods are useful in sensitivity analysis, which evaluates how the results based on the MAR assumption would change under the assumption that NMAR is correct (Abe, 2016, p.160). If the results are not drastically different, then the confidence will be enhanced about the results based on the MAR assumption, while if the results are drastically different, then the confidence is low, so that we would need to handle the situation by including more auxiliary variables to make the assumption of MAR more relevant.

The true missing data mechanism is often unknown, but there is an occasion where the missing data mechanism is obvious through the planned missing design (Enders, 2010). For example, generally in official economic statistics, the actual turnover values among large enterprises are collected by follow-ups even if they are missing at first; then, only the missing values among small-and-medium enterprises are imputed by statistical methods (de Waal et al., 2011, pp.245-246). In this case, the missing rate of turnover changes according to the size of enterprises such as the number of employees; thus, this can be thought of as MAR. Scheuren (2005) estimates that the proportion of MCAR is about 10% to 20%, MAR about 50%, and NMAR about 10% to 20% in official statistics.

2.4 Current Imputation Methods in Official Statistics: Deterministic Single Imputation

Traditionally, the main goal in official statistics is to compute the total (or the mean) of survey data, not the analysis of the distribution and variance (de Waal et al., 2011, p.225). Deterministic single imputation uses the predicted values based on the imputation model without adding random error. It is customary to use deterministic single imputation in official statistics, because it is an unbiased point estimator of the mean. Among the deterministic single imputation methods, it is said that the frequently used methods are regression imputation, ratio imputation, mean imputation, and hot deck imputation (Hu et al., 2001; de Waal et al., 2011, ch.7). This section briefly introduces the mechanisms of these four methods.

2.4.1 Regression Imputation

In regression imputation, parameters β_0 and β_1 in equation (2.4) are estimated by Ordinary Least Squares (OLS) based on observed data (Takahashi et al., 2015, pp.11-14). Observed data refer to the data based on listwise deletion. Using the data in Table 2.1, we can estimate $\beta_0 = -85.33$ and $\beta_1 = 12.80$. Since the age of ID 4 is 54, the estimated income of ID 4 is 605.87 as in equation (2.5). If we use this value instead of Income_4 in equation (2.2), then the mean income is computed as 442.72. This implies that the mean value is now closer to the truth than the one based on listwise deletion. For a detailed discussion on regression imputation, also see Chapter 3 of this dissertation.

$$\widehat{\mathrm{Income}}_{\mathrm{i}} = \beta_0 + \beta_1 \mathrm{Age}_{\mathrm{i}} \tag{2.4}$$

$$\mathrm{Income}_4 = -85.33 + 12.80 \times 54 = 605.87 \tag{2.5}$$

2.4.2 Ratio Imputation

In ratio imputation, parameter β_1 in equation (2.6) is estimated by the ratio of means based on observed data (Takahashi et al., 2015, pp.18-22). Using the data in Table 2.1, the mean of income in observed data is 388.33, and the mean of age in observed data is 37. These are the values based on listwise deletion. Therefore, we can estimate $\beta_1 = 388.33/37 = 10.50$. Since the age of ID 4 is 54, the estimated income of ID 4 is 567.00 as in equation (2.7). If we use this value instead of Income_4 in equation (2.2), then the mean income is computed as 433.00. This implies that the mean value is now closer to the truth than the one based on listwise deletion. For a detailed discussion on ratio imputation, also see Chapter 3 of this dissertation.

$$\widehat{\mathrm{Income}}_{\mathrm{i}} = \beta_1 \mathrm{Age}_{\mathrm{i}} \tag{2.6}$$

$$\mathrm{Income}_4 = 10.50 \times 54 = 567.00 \tag{2.7}$$

2.4.3 Mean Imputation

In mean imputation, the mean of observed data is used as imputed values for missing values. Generally, mean imputation is not useful except rare circumstances (Takahashi and Ito, 2013a, pp.27-28; Takai et al., 2016, p.6). However, in Table 2.2, the value of age is not numerical but categorical. In this situation, we may use group mean imputation, which computes the mean in each age group (de Waal et al., 2011, pp.246-249). If we stratify the data in Table 2.2 by age, then we can classify ID 1 and ID 2 to group 1, and ID 3 and ID 4 to group 2. In order to estimate the income value of ID 4, we may use the mean of group 2, which is 505. If we use this value instead of Income_4 in equation (2.2), then the mean income is computed as 417.50. This implies that, unlike simple mean imputation, the mean value using group mean imputation is now closer to the truth than the one based on listwise deletion.

2.4.4 Hot Deck Imputation

Just as in Table 2.3, let us suppose that age is numeric, but income is categorical. If the estimand is categorical, we may use hot deck imputation, where we find a donor whose value in an auxiliary variable is close to that of the recipient, and the donor's value is used as an imputation. The age of ID 4 is 54 which is close to the age 47 of ID 3 in Table 2.3. Therefore, ID 3 is the donor for ID 4. We use the income of ID 3 for the value of ID 4; thus, it will be 3. In this case, the mode of income is 3, and we can see that this value matches the true value in complete data. In the actual application, the nearest neighbor method is often used to find a suitable donor by defining the distance function, which is essentially the same as matching. For a detailed discussion on hot deck and matching, see Abe (2016, pp.57-59), Takai et al. (2016, pp.110-113), and Kurihara (2015). *R* Package HotDeckImputation can be used for this purpose (Joenssen, 2015). Hot deck is a non-parametric method that can be used even when all data are categorical.

2.5 Current Practice of Data Editing Across the UNECE Member States

The Work Session on Statistical Data Editing is an international conference hosted every 18 months by UNECE, where national statistical agencies from Europe, North America, and Oceania meet together to exchange their ideas and information concerning the methods of handling missing values and error. The author attended the Norway conference (September, 2012), the France conference (April, 2014), and the Hungary conference (September, 2015). Questionnaires were sent to those participants who presented research papers at least in one of the above mentioned three conferences. All of them are the national statistical agencies that internationally lead official statistics. The results of the survey are summarized in Table 2.4.

Population: Twenty three national statistical agencies

Survey Period: July to September, 2016

Survey Method: A questionnaire sent via email to a staff who specializes in data editing

Response Rate: 87.0% (as of September 6, 2016)

Table 2.4 Results of UNECE Survey (Multiple Answers)

	Regression Imputation	Ratio Imputation	Mean Imputation	Hot Deck Imputation
Question 1	95.0%	95.0%	95.0%	100.0%
Question 2	40.0%	60.0%	35.0%	65.0%
Question 3	30.0%	80.0%	35.0%	30.0%
Question 4	10.0%	10.0%	25.0%	80.0%

Question 1: Does your organization use all of the four methods in practice?
Question 2: Generally speaking, which of the four methods is most often used in practice?
Question 3: In economic data, where the unit is enterprises and establishments, which of the four methods is most often used in practice?
Question 4: In household data, which of the four methods is most often used in practice?

Question 1 reveals that all of the four methods are used in practice among almost all of the twenty national statistical agencies, where mean imputation is used more frequently than expected. Question 2 shows that ratio imputation (60.0%) and hot deck imputation (65.0%) are deemed important. Question 3 reveals that ratio imputation (80.0%) is often used in economic data, and that regression imputation is not used very often. Incidentally, regression imputation covers a wider variety of models than ratio imputation, such as multiple regression, polynomials, logistic regression; thus, regression imputation is sometimes employed in those situations (de Waal et al., 2011, pp.233-235). Question 4 shows that hot deck imputation (80.0%) is often used in household data, and that the numerical items in household data are occasionally dealt with by group mean imputation (25.0%).

Question 5 in Table 2.5 reveals the fact that, in the current practice of data editing, stochastic single imputation is used by fourteen national statistical agencies (70.0%), multiple imputation by eight national statistical agencies (40.0%), and fractional imputation by one national statistical agency (5.0%). Note that stochastic single imputation is a method that adds random components to each imputed value, so that the dispersion of data is adjusted (Takahashi et al., 2015, pp.15-18).

Table 2.5 Results of UNECE Survey (Multiple Answers)

	Stochastic Single Imputation	Multiple Imputation	Fractional Imputation
Question 5	70.0%	40.0%	5.0%

Question 5: Does your organization use any of the following methods in practice? If yes, which method(s)?

This chapter does not deal with fractional imputation, but briefly, fractional imputation is a repeated imputation method just as multiple imputation. It is different from multiple imputation in the following three points (de Waal et al., 2011, p.272): (1) Fractional imputation can be considered improper multiple imputation based on the frequentist perspective; (2) the purpose of fractional imputation is to minimize the inflation of the variance in multiple imputation; and (3) fractional imputation relies on a version of hot deck; thus, it can handle qualitative data. Interested readers are referred to de Waal et al. (2011, pp.271-272).

2.6 Simulation Studies on Deterministic Single Imputation Methods

As we saw in Section 2.5, all of the four methods of mean imputation, ratio imputation, regression imputation, and hot-deck imputation are utilized by the national statistical agencies around the world. This section conducts a series of Monte Carlo simulation experiments for these four methods, assuming the following data-generation processes.

(1) Economic Data: Numerical (quantitative) data following the log-normal distribution

(2) Qualitative Economic Data: Numerical (quantitative) data following the log-normal distribution and categorical (qualitative) auxiliary variable

(3) Household Data: Categorical (qualitative) data and numerical (quantitative) auxiliary variable

Monte Carlo simulation is an analytic method that repeatedly draws random numbers. We assume a certain probability distribution based on observed data, and generate pseudo random numbers by a computer in order to quantitatively analyze random variables that follow probability distributions (Ono and Ikawa, 2015). In other words, Monte Carlo simulation is a method to use a computer as an experimental laboratory, where the researcher has the control over the experiments and can measure the effects by observing the results based on different laboratory environments (Carsey and Harden, 2014). To be specific, Monte Carlo simulation is conducted by the following five steps (Mooney, 1997). All of the computations in Chapter 2 were done in *R* 3.2.4.

(1) Define the pseudo population on the computer.

(2) Draw a sample from the pseudo population.

(3) Estimate the parameter.

(4) Repeat steps (2) and (3), many times, say 1000 times.

(5) Calculate the relative frequency of the parameter estimates.

The results of the experiments are evaluated through the Mean Squared Error (MSE) in equation (2.8). The MSE of an estimate $\hat{\theta}$ can be computed by generating a vector of true values θ, taking the differences from a vector of $\hat{\theta}$, and dividing the sum of the squared differences by the number of simulation runs (Mooney, 1997; Carsey and Harden, 2014). A smaller MSE value means that the method is comparatively good.

$$MSE = E\left[\left(\hat{\theta} - \theta\right)^2\right] \tag{2.8}$$

In the actual applications below, following Di Zio and Guarnera (2013, p.549), the Relative Root Mean Squared Error (RRMSE) is used as in equation (2.9), which normalizes the MSE by the true value and takes the square root.

$$RRMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\theta} - \theta}{\theta}\right)^2} \tag{2.9}$$

The design of the simulation is as follows. The population model is equation (2.10), where the estimand is $\bar{y}$.

$$y_i = \beta_1 x_{1i} + \varepsilon_i \tag{2.10}$$

where

$$x_{1i} \sim LN(logmean = 0, logsd = 1)$$

$$\varepsilon_i \sim N\left(mean = 0, sd = \sigma\sqrt{x_i}\,\right)$$

The number of iterations in Monte Carlo simulation is set to 1000, in each of which sample data with $n = 1000$ are generated. The missingness in y_i mimics the planned missing design (Enders, 2010) mentioned in Section 2.3. Specifically, let $u_i \sim U(0,1)$. Also, let $\text{med}(x_{1i})$ be the median of x_{1i}. When $x_{1i} < \text{med}(x_{1i})$ and $u_i < 0.6$, the value of y_i is made missing, which creates the MAR missingness given the values of x_{1i}. The missing rate is set to about 30%. This setting is realistic, because the mean missing rates of income and wage in the National Health Interview Survey from 1997 to 2004 are about 30%, respectively (Schenker et al., 2006, p.925). Also, the variance of the error term ε_i increases in proportion to the values of x_{1i}, which means that the variance is heteroskedastic. The values of β_1 are randomly drawn from $U(1.1,2.0)$, and the values of σ are randomly drawn from $U(1.0,2.0)$. In other runs, not reported here, where these values were changed, similar results are obtained. $LN(\cdot)$ is R-function `rlnorm`, $N(\cdot)$ is R-function `rnorm`, and $U(\cdot)$ is R-function `runif`, respectively.

Table 2.6 simulates the treatment of missing values in economic data, an example of which is Table 2.1. In log-normally distributed data where the variance is heteroskedastic, all of the imputation methods have smaller RRMSE compared to listwise deletion. The performance of ratio imputation (RRMSE = 0.048) is best in relation to regression imputation (RRMSE = 0.050) and hot deck imputation (RRMSE = 0.050). As is discussed in Cochran (1977, p.158) and Takahashi et al. (2017), ratio imputation is the best linear unbiased estimator (BLUE) under the heteroskedastic error, $\varepsilon_i \sim N(0, \sigma\sqrt{x_i})$.

Table 2.6 RRMSE for Missing Value Treatments in Economic Data

Complete Data	Listwise Deletion	Regression Imputation	Ratio Imputation	Hot Deck
0.047	0.302	0.050	0.048	0.050

Table 2.7 simulates the treatment of missing values in economic data that include qualitative items, an example of which is Table 2.2. Two groups are defined by the different values of x_{1i}, 0 and 1, where the mean and the missing rate are set to different values in each group. Other settings are exactly the same as in Table 2.6. Group mean imputation (RRMSE = 0.055) outperforms listwise deletion (RRMSE = 0.081) when the auxiliary variable is qualitative.

Table 2.7 RRMSE for Missing Value Treatments in Qualitative Economic Data

Complete Data	Listwise Deletion	Mean Imputation
0.043	0.081	0.055

Table 2.8 simulates the treatment of missing values in household data that include qualitative items, an example of which is Table 2.3. The values of y_i are transformed into three unordered categories, while x_{1i} is kept numerical. The estimand is the proportion of the values that are categorized into the mode of y_i. Other settings are exactly the same as in Table 2.6. When the target variable for computation is qualitative, the performance of hot deck (RRMSE = 0.056) is best, and regression imputation (RRMSE = 0.381) and ratio imputation (RRMSE = 0.381) are useless in these situations.

Table 2.8 RRMSE for Missing Value Treatments in Household Data

Complete Data	Listwise Deletion	Regression Imputation	Ratio Imputation	Hot Deck
0.038	0.123	0.381	0.381	0.056

2.7 Public-Use Microdata and Multiple Imputation

Up to this point, our discussion assumes that the total (or the mean) is the target for computation. As we saw in Section 2.5, regression imputation, ratio imputation, group mean imputation, and hot deck imputation are used by the national statistical agencies around the world. As we examined in Section 2.6, these methods are used for appropriate data types. The advantage of deterministic single imputation is that it is unbiased for point estimation of the mean (or the total), but the disadvantage is that the distribution and variance are not correctly estimated (Abe, 2016, p.55). The estimand in the analysis using public-use microdata is not limited to the mean and the total.

If we want to make the analysis valid not only for the mean, but also for the variance and the standard error, we need to use multiple imputation (Schafer and Graham, 2002; Donders et al., 2006; Baraldi and Enders, 2010; Cheema, 2014). Multiple imputation, in theory, randomly draws several values from the distribution of missing data. However, missing data are unobserved; thus, the distribution of missing data is also unobserved. In real applications, we estimate the predictive posterior distribution of missing values given observed data by using Bayesian statistics, and we randomly draw the mean vector and the variance-covariance matrix from the posterior distribution. In this way, we can implement imputation that can take into account the fact that the parameter of the imputation model is estimated (King et al., 2001). For a detailed discussion on multiple imputation, see Chapter 4 of this dissertation. Also see Iwasaki (2002, Ch.10), Takahashi and Ito (2014), Takahashi et al. (2015), and Abe (2016, Ch.5).

Table 2.9 presents a concrete example of multiply-imputed data. The empty cell in income is a missing value, and the white numbers in gray cells for income1, income2, and income3 are the imputed values by multiple imputation. The mean of income1 is estimated as 388.25, the mean of income2 as 439.75, and the mean of income3 as 475.25. The point estimate of the mean income is the mean of the three means, i.e., 428.42. Single imputation in Section 2.4 deterministically estimated one value for a missing value, seeing no estimation uncertainty. However, in Table 2.9, the imputed values change every time we impute the missing value, showing estimation uncertainty, which makes the standard error valid.

Table 2.9 Example of Multiply-Imputed Data ($M = 3$)

ID	Income	Age	Income1	Income2	Income3
1	239	26	239	239	239
2	421	38	421	421	421
3	505	47	505	505	505
4		54	388	594	664

Rubin (1987) proposes that if M versions of multiply-imputed data ($M > 1$) are released by data providers, then the analysts can conduct a variety of statistical analyses regardless of their statistical literacy concerning missing data analysis. Therefore, it is suggested that multiple imputation is suitable for public-use microdata. By copying and pasting the code presented in Appendix 2.1, the analysts can perform statistical analyses without being annoyed by a practical difficulty of how to combine the analyses based on multiply-imputed data once they download public-use microdata (assuming that the public-use microdata are available online).

Table 2.10 presents concrete examples of public-use microdata provided by the U.S. government using multiple imputation.

Table 2.10 Example of Public-Use Microdata by Multiple Imputation (U.S. Government)

Organization	Survey	Target Variable	Number of Multiple Imputation	Publication Date
Centers for Disease Control and Prevention[1]	2015 National Health Interview Survey	Income, earnings	$M = 5$	June 30, 2016
Federal Reserve System[2]	2013 Survey of Consumer Finances	Almost all missing variables	$M = 5$	September 25, 2014
Department of Transportation[3]	2014 Fatality Analysis Reporting System	Blood alcohol concentration	$M = 10$	December 1, 2015
Bureau of Labor Statistics[4]	2014 Consumer Expenditure Survey	Income	$M = 5$	September 3, 2015

[1] http://www.cdc.gov/nchs/nhis/nhis_2015_data_release.htm
[2] http://www.federalreserve.gov/econresdata/scf/scfindex.htm
[3] http://www.nber.org/data/fars.html
[4] http://www.bls.gov/cex/csxmicrodoc.htm

An additional survey questionnaire was sent to those twenty national statistical agencies which answered Questions 1 to 5 in Section 2.5, and the responses were obtained from eighteen agencies (response rate = 90.0% as of September 6, 2016). The results are presented in Table 2.11.

Table 2.11 Results of UNECE Survey (Multiple Answers)

	Incomplete Data	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation	None of the above
Question 6	22.0%	50.0%	61.1%	44.4%	22.2%

Question 6: Hypothetically speaking, if survey data are to be made open as “public-use microdata,” which of the following imputation methods do you think should be used?

In an open field, it was stated that public-use microdata should be imputed data so that all citizens would be able to conduct statistical analysis without being concerned with missing data. This concurs with Rubin’s (1987) suggestion. However, consensus has not been achieved about what type of imputed data should be used, such as deterministic single imputation (50.0%), stochastic single imputation (61.1%), or multiple imputation (44.4%).

2.8 Multiple Imputation and Microdata Analysis

As we saw in Section 2.7, there is no consensus of choosing listwise deletion, deterministic single imputation, stochastic single imputation, and multiple imputation as a method to treat missing values in public-use microdata among the national statistical agencies around the world. This section evaluates the accuracy of the mean and the regression coefficient using these four methods by Monte Carlo simulation. The number of multiply-imputed data is set to 5. This section utilizes *R* Package AMELIA II, a general purpose multiple imputation software program (Honaker et al., 2011). For different multiple imputation algorithms, see Chapter 4 of this dissertation.

2.8.1 Regression Analysis: Missing Independent Variable

The design of simulation is as follows (for detailed information, also see section 2.6). The population model is equation (2.11). $\bar{x}_1$ and β_1 are the estimands. The number of Monte Carlo simulation runs T is set to 1000, in each of which sample data of $n = 1000$ are generated. The missingness in x_{1i} is generated by MAR, conditional on y_i, setting the missing rate at about 30%. Specifically, just as in Section 2.6, if $y_i < \text{med}(y_i)$, $Pr(x_{1i} = missing) = 0.6$. In order to mimic the analysis based on log-normal, economic data, we assume that the data are log-transformed, so that the data were generated based on the normal distribution. The values of β_1

are randomly drawn from $U(1.1,2.0)$, and the values of σ are randomly drawn from $U(1.0,2.0)$. In other runs, not reported here, where these values were changed, similar results are obtained.

$$y_i = \beta_1 x_{1i} + \varepsilon_i \tag{2.11}$$

where

$$x_{1i} \sim N(mean = 0, sd = 1)$$

$$\varepsilon_i \sim N(mean = 0, sd = \sigma)$$

Table 2.12 shows the RMSE for $\bar{x}_1$, the RRMSE for β_1, and the coverage rate of the nominal 95% confidence interval when missingness occurs in an independent variable.

Table 2.12 Estimation of $\bar{x}_1$ and β_1 when Independent Variable is Missing

	Complete Data	Listwise Deletion	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation
RMSE ($\bar{x}_1$)	0.076	0.618	0.085	0.090	0.087
RRMSE (β_1)	0.026	0.062	0.139	0.031	0.030
95% CI Coverage	94.9	61.8	0.1	90.5	94.7

Note: Since the true value of $\bar{x}_1$ is zero, I used RMSE instead of RRMSE. CI stands for confidence interval. The 95% CI coverage means the proportion of the times the true β_1 was included in the 95% confidence interval in 1,000 Monte Carlo experiments.

As for the RMSE of $\bar{x}_1$, both single imputation and multiple imputation are unbiased, but listwise deletion is biased. The performances of deterministic single imputation (RMSE = 0.085), multiple imputation (RMSE = 0.087), and stochastic single imputation (RMSE = 0.090) are almost equal, but the performance of listwise deletion (RMSE = 0.618) is quite low.

As for the RRMSE of β_1, the performance of multiple imputation (RRMSE = 0.030) is best, followed by stochastic single imputation (RRMSE = 0.031) and listwise deletion (RRMSE = 0.062). The performance of deterministic single imputation (RRMSE = 0.139) is quite low (Allison, 2002, p.53; Carpenter and Kenward, 2013, p.28).

As for the nominal 95% confidence interval for β_1, the CI by multiple imputation contains the true parameter with the probability of 94.7%, which is quite accurate. The CI by stochastic single imputation contains the true parameter with the probability of 90.5%, meaning that the nominal 5% Type I error is about double, which is a serious concern (Enders, 2010, pp.53-54). The CI by listwise deletion contains the true parameter with the probability of 61.8%, meaning that the nominal 5% Type I error is about eight-fold, which is a very serious concern. The CI by deterministic single imputation contains the true parameter with the probability of 0.1%, meaning that the nominal 5% Type I error is about twenty-fold, which is an extremely serious concern. When the independent variable has missing values and the estimands are the regression coefficient and the mean, then this analysis shows that multiple imputation should be used. Also see Chapter 4 of this dissertation about more detailed analyses, including other versions of multiple imputation.

2.8.2 Regression Analysis: Missing Dependent Variable

The design of simulation is as follows. The population model is equation (2.11). $\bar{y}$ and β_1 are the estimands. The missingness in y_i is generated by MAR conditional on x_{1i}, setting the missing rate at about 30%. Other settings follow Section 2.8.1. Table 2.13 shows the RMSE for $\bar{y}$, the RRMSE for β_1, and the coverage rate of the nominal 95% confidence interval when missingness occurs in the dependent variable.

Table 2.13 Estimation of $\bar{y}$ and β_1 when Dependent Variable is Missing

	Complete Data	Listwise Deletion	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation
RMSE ($\bar{y}$)	0.067	0.609	0.073	0.075	0.074
RRMSE (β_1)	0.021	0.027	0.027	0.029	0.028
95% CI Coverage	94.8	95.0	80.0	83.9	94.2

Note: Since the true value of $\bar{y}$ is zero, I used RMSE instead of RRMSE. CI stands for confidence interval. The 95% CI coverage means the proportion of the times the true β_1 was included in the 95% confidence interval in 1,000 Monte Carlo experiments.

As for the RMSE of $\bar{y}$, both single imputation and multiple imputation are unbiased, but listwise deletion is biased. The performances of deterministic single imputation (RMSE = 0.073), multiple imputation (RMSE = 0.074), and stochastic single imputation (RMSE = 0.075) are almost equal, but the performance of listwise deletion (RMSE = 0.609) is quite low.

As for the RRMSE of β_1, the performances of listwise deletion (RRMSE = 0.027), deterministic single imputation (RRMSE = 0.027), multiple imputation (RRMSE = 0.028), and stochastic single imputation (RRMSE = 0.029) are almost the same. When the dependent variable has missing values and the estimand is the regression coefficient, then imputation does not change the result from listwise deletion. This is because the incomplete cases do not contribute to the computation of regression coefficients under MAR when missingness occurs in the dependent variable (Little, 1992; Carpenter and Kenward, 2013, pp.24-28; Raghunathan, 2016, p.99).

As for the nominal 95% confidence interval for β_1, the CI by listwise deletion contains the true parameter with the probability of 95.0%, which is quite accurate. The CI by multiple imputation contains the true parameter with the probability of 94.2%, which is also quite accurate. The CI by stochastic single imputation contains the true parameter with the probability of 83.9%, meaning that the nominal 5% Type I error is more than triple, which is a serious concern. The CI by deterministic single imputation contains the true parameter with the probability of 80.0%, meaning that the nominal 5% Type I error is about quardruple, which is an extremely serious concern.

If the estimands are the regression coefficient and the mean, then multiple imputation, though it comes in second in each situation, may be the best method overall. This analysis shows that single imputation should not be used at all in this situation.

2.9 Multiple Imputation, Microdata Analysis, and Congeniality

When the imputation model and the analysis model have exactly the same variables estimating the same number of parameters, the two models are said to be congenial (Enders, 2010, p.227; Abe, 2016, p.118; Takai et al., 2016, p.123). The models we used so far are all congenial. However, in real applications, there can be occasions under which the imputation model is different from the analysis model. If this is the case, there is no guarantee in theory about the consistency of the parameter estimates by multiple imputation. In this section, we will examine the two uncongenial cases: (1) The analysis model is the subset of the imputation model; and (2) the imputation model is the subset of the analysis model.

2.9.1 Analysis Model Subset of Imputation Model

The design of simulation is as follows. The imputation model is equation (2.12), and the analysis model is equation (2.13). The esimands are $\bar{x}_1$ and β_1. To be technically correct, since missingness occurs in x_{1i}, the imputation model in a strict sense is $x_{1i} = \gamma_0 + \gamma_1 y_i + \gamma_2 x_{2i} + \epsilon_i$, and $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$ is the population model of y_i. Equation (2.12) is equation (2.11) that contains the bivariate distribution of X. The number of Monte Carlo simulation runs T is set to 1000, in each of which sample data of $n = 1000$ are generated. The missingness in x_{1i} is generated by MAR conditional on y_i, setting the missing rate at about 30%. $MN(\cdot)$ refers to R-function `mvrnorm`. The values of β_1 are randomly drawn from $U(1.1,1.5)$, and the values of σ are randomly drawn from $U(1.1,1.5)$. In other runs, not reported here, where these values were changed, similar results are obtained.

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{2.12}$$

$$y_i = \beta_1 x_{1i} + \varepsilon_i \tag{2.13}$$

where

$$X \sim MN(mean = 0, sd = 1)$$

$$X = (x_{1i}, x_{2i})$$

$$cor(X) = \begin{pmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{pmatrix}$$

$$\varepsilon_i \sim N(0, \sigma)$$

When the analysis model is the subset of the imputation model, the two models are strictly speaking uncongenial. However, as is clear in Table 2.14, there is no problem in the performance of multiple imputation (Enders, 2010, pp.228-229; Carpenter and Kenward, 2013, pp.64-65). What this implies is that the official statistical agencies as data providers can include as many auxiliary variables in the imputation model as possible, and they can make the data available after removing the variables that contain sensitive information (Takai et al., 2016, p.124). In the current practice of imputation among the national statistical agencies, the variables used for imputation are only a subset of the variables that are included in microdata; however, due to the problem of congeniality, missing values must be imputed by using all of the available variables that are contained in microdata. See Section 2.9.2 below.

Table 2.14 The Analysis Model is the Subset of the Imputation Model

	Complete Data	Listwise Deletion	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation
RMSE ($\bar{x}_1$)	0.074	0.633	0.080	0.083	0.081
RRMSE (β_1)	0.026	0.058	0.084	0.029	0.028
95% CI Coverage	95.6	64.8	14.4	91.5	95.6

Note: Since the true value of $\bar{x}_1$ is zero, I used RMSE instead of RRMSE. CI stands for confidence interval. The 95% CI coverage means the proportion of the times the true β_1 was included in the 95% confidence interval in 1,000 Monte Carlo experiments.

2.9.2 Imputation Model Subset of Analysis Model

The design of simulation is as follows. The imputation model is equation (2.14), and the analysis model is equation (2.15). Other settings follow Section 2.9.1.

$$y_i = \beta_1 x_{1i} + \varepsilon_i \tag{2.14}$$

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{2.15}$$

When the imputation model is the subset of the analysis model, as is clear in Table 2.15, the performances of all the imputation models are quite low. In other words, we should not use the analysis model that is larger than the imputation model (Enders, 2010, p.229; Carpenter and Kenward, 2013, p.64).

Table 2.15 The Imputation Model is the Subset of the Analysis Model

	Complete Data	Listwise Deletion	Deterministic Single Imputation	Stochastic Single Imputation	Multiple Imputation
RMSE ($\bar{x}_1$)	0.087	0.739	0.093	0.098	0.094
RRMSE (β_1)	0.036	0.063	0.119	0.117	0.115
95% CI Coverage	95.3	82.0	5.6	8.9	13.7

Note: Since the true value of $\bar{x}_1$ is zero, I used RMSE instead of RRMSE. CI stands for confidence interval. The 95% CI coverage means the proportion of the times the true β_1 was included in the 95% confidence interval in 1,000 Monte Carlo experiments.

2.10 Conclusion

This chapter showed that the imputation methods were adopted according to the types of data in the current practice of official statistics among the UNECE member states. Specifically, ratio imputation is used for economic data, and hot deck imputation for household data. Also, the estimand in public-use microdata is, by its nature, not limited to the mean and the total, unlike in the current practice of official statistics. If the regression coefficient along with the standard error is the estimand, this chapter showed that multiple imputation would be best suited for use.

Appendix 2.1: Example Code of Generating and Analyzing Multiply-Imputed Data

This appendix presents *R*-code to generate multiply-imputed data by *R*-package AMELIA II (Honaker et al., 2011) and analyze the multiply-imputed data by *R*-package Zelig (Imai et al., 2008).

First, the imputer treats missing values by multiple imputation, where $M = 5$ (Takahashi and Ito, 2013a, pp.48-49). In the example below, five multiply-imputed data files will be created.

```
library(Amelia)
set.seed(6997582)
a.out < -amelia(data, m = 5)
write.amelia(obj = a.out, file.stem = "outdata", orig.data = F,
             separate = T, row.names = F)
```

Next, the imputer prepares the following code, and publishes it along with the five multiply-imputed data files. The analyst only needs to download the five multiply-imputed data files and paste the following code in *R*. Note that, in order to use this code, *R*-package hot.deck (Cranmer and Gill, 2013) must be installed by the analyst. This code is necessary when the multiply-imputed data will be first outputted, and then they will be inputted on another computer. This procedure is not explained in the *R* manuals for AMELIA II and Zelig.

```
data1<-read.csv("outdata1.csv",header=T)
data2<-read.csv("outdata2.csv",header=T)
data3<-read.csv("outdata3.csv",header=T)
data4<-read.csv("outdata4.csv",header=T)
data5<-read.csv("outdata5.csv",header=T)
idata<-list(imp1 = data1, imp2 = data2, imp3 = data3, imp4 = data4,
            imp5 = data5)
idata<-list(imputations=idata)
library(hot.deck)
midata<-hd2amelia(idata)
```

Finally, the analyst uses *R*-package Zelig for statistical analysis (Takahashi and Ito, 2013a, p.49). The analyst only needs to specify the variables for use such as `x1~x2+x3` and the analysis model `model = "ls"`. Multiple results based on multiply-imputed data will be automatically combined by Zelig.

```
library(Zelig)
z.out <- zelig(x1~x2+x3, data = midata, model = "ls", cite = F)
summary(z.out)
```

3 A Unified Approach to Ratio Imputation for Heteroskedastic Missing Variables

This chapter derived from Takahashi et al. (2017), a peer-reviewed article in the *Statistical Journal of the IAOS* 33(3), which is the flagship journal of the International Association for Official Statistics (IAOS) under the umbrella of the International Statistical Institute (ISI). The *Statistical Journal of the IAOS* is indexed in Scopus by Elsevier as of April 2017. The author would like to thank IOS Press for permission to use “Imputing the mean of a heteroskedastic log-normal missing variable: A unified approach to ratio imputation” (coauthored with Iwasaki, M. and Tsubaki, H., *Statistical Journal of the IAOS*, vol.33, no.3, in press).

3.1 Introduction

When data are collected through surveys, some values are almost always missing, making the survey data incomplete. As Rubin (1987) pointed out, incomplete data are inefficient at best and biased at worst when there are systematic differences between respondents and non-respondents. Little and Rubin (2002) demonstrate that if the missing mechanism is at random (MAR), then this bias can be rectified by imputations. Under the assumption that data are multivariate normal, many standard modern imputation methods have been developed, such as regression imputation, EM algorithm, and multiple imputation (Schafer, 1997; Donders, 2006; Baraldi and Enders, 2010; Cheema, 2014).

However, the non-normal distribution complicates the issue of imputation, especially when our goal is to estimate the mean of the raw data in the original scale. Many economic data are highly skewed to the right, examples including, but not limited to, income, earnings, turnover, and GDP. These variables also tend to create heteroskedasticity, because the larger values of economic variables allow the units to have more discretion for the decisions to make. It is true that if a variable is right-skewed and heteroskedastic, taking the log can mitigate the problems (Wooldridge, 2009), but this is only valid if the goal is to estimate the mean of log (Y_i), not the mean of Y_i (See Appendix 3.1). To tackle this problem, ratio imputation has been often used to treat missing values in official economic statistics, where the goal of surveys is to estimate the mean or the total of a heteroskedastic log-normally distributed variable (Hu et al., 2001; de Waal et al., 2011; Thompson and Washington, 2012; Office for National Statistics, 2014).

In the literature, however, there are three ratio estimators: Ordinary least squares (OLS); ratio of means (RoM); and mean of ratios (MoR) (Snowdon, 1992; Eisenhauer, 2003). It is not quite obvious which of the estimators best perform when our goal is to estimate the mean of a heteroskedastic log-normal variable; thus, leading to a gap between theory and practice. The purpose of this chapter is to fill in this gap by assessing which of them should be employed in official economic statistics in a unifying manner.

By way of organization, Section 2 introduces the notations used in the current study and the three common assumptions of missing mechanisms. Section 3 explains the three competing ratio imputation models. Section 4 discusses how the three ratio estimators can be unified under the weighted least squares (WLS) model. Section 5 shows the results of Monte Carlo simulation for ratio imputation models. Section 6 presents a novel estimation strategy for imputation model selection followed by Monte Carlo simulation to assess the proposed method. Section 7 applies the proposed method to real data. Section 8 concludes.

3.2 Notations and Assumptions of Missing Mechanisms

The notations used in this chapter are as follows. Let $\mathbf{D}$ be an $n \times p$ dataset, where n is the number of observations and p is the number of variables. If there are no missing data, we assume that $\mathbf{D}$ is log-normally distributed with the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{D} \sim LN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. An observation index is denoted i, where $i = 1, \ldots, n$. Ratio imputation involves two variables; therefore, $\mathbf{D} = \{Y_i, X_i\}$, where Y_i is the incomplete variable (target variable for imputation) and X_i is the complete variable (auxiliary variable). Also, let $\mathbf{M}$ be a missing indicator matrix, the dimension of which is the same as $\mathbf{D}$. Whenever $\mathbf{D}$ is observed $\mathbf{M} = 1$, and whenever $\mathbf{D}$ is not observed $\mathbf{M} = 0$. Furthermore, $\mathbf{D_{obs}}$ refers to the observed part of data, and $\mathbf{D_{mis}}$ refers to the missing part of data, i.e., $\mathbf{D} = \{\mathbf{D_{obs}}, \mathbf{D_{mis}}\}$.

Next, let us briefly introduce the three common assumptions of missingness (Schafer, 1997; King et al., 2001; Allison, 2002; Little and Rubin, 2002). The first assumption is Missing Completely At Random (MCAR), where the probability of missingness is independent of the data for the unit, i.e., $P(\mathbf{M}|\mathbf{D}) = P(\mathbf{M})$. The second assumption is the case where missingness is conditionally at random, traditionally known as Missing At Random (MAR), where the conditional probability of missingness given data is equal to the conditional probability of missingness given observed data, i.e., $P(\mathbf{M}|\mathbf{D}) = P(\mathbf{M}|\mathbf{D_{obs}})$. The third assumption is Non-Ignorable (NI), where the missingness probability of a variable depends on the variable's value itself, and we cannot break this relationship conditional on observed data, i.e., $P(\mathbf{M}|\mathbf{D}) \neq P(\mathbf{M}|\mathbf{D_{obs}})$. The current study assumes that the missing mechanism is MAR.

3.3 Competing Ratio Imputation Models

This section outlines the three competing ratio imputation models. Suppose that Y_i is missing in the current data and that X_i is fully observed in the previous data. For example, Y_i is turnover in year 2016 and X_i is turnover in year 2015. The missing values of Y_i may be imputed by equation (3.1), where the value of β reflects the trend between the two time points.

$$\hat{Y}_i = \beta X_{i,obs} \tag{3.1}$$

Note that cold deck imputation is a special case of equation (3.1) (de Waal et al., 2011). While the value of β in cold deck imputation is assumed to be 1.0, the value of β in ratio imputation is not assumed to be known and must be estimated from the observed part of data. Since ratio imputation is a combination of cold deck and hot deck, some scholars call it warm deck (Shao, 2000).

Equation (3.1) takes the form of a simple regression model without an intercept, where the value of β can be estimated by the following three methods.

$$\hat{\beta}_{OLS} = \frac{\sum X_{i,obs} Y_{i,obs}}{\sum X_{i,obs}^2} \tag{3.2}$$

$$\hat{\beta}_{RoM} = \frac{\bar{Y}_{i,obs}}{\bar{X}_{i,obs}} \tag{3.3}$$

$$\hat{\beta}_{MoR} = \frac{1}{n}\sum \frac{Y_{i,obs}}{X_{i,obs}} \tag{3.4}$$

Equation (3.2) is the regression through the origin by OLS (ordinary least squares) (Eisenhauer, 2003), which we call the OLS imputation model. Equation (3.3) is the ratio-of-means imputation model (Rao, 2002; Liang et al., 2008), which we call the RoM imputation model. Equation (3.4) is the mean-of-ratios imputation model, which we call the MoR imputation model (Rao, 2002; Liang et al., 2008).

What complicates the matter is the fact that there are opposing views in the literature as to which estimator outperforms the others (Table 3.1).

Table 3.1. Proponents of Various Methods

Methods	Proponents
Ordinary Least Squares (OLS)	Eisenhauer (2003), Gujarati (2003), Wooldridge (2009)
Ratio of Means (RoM)	Hu et al. (2001), de Waal et al. (2011), Gupta and Kabe (2011), Thompson and Washington (2012), Office for National Statistics (2014)
Mean of Ratios (MoR)	Hoenig et al. (1997) , Zarnoch and Bechtold (2000), Liu et al. (2005) , Zou et al. (2010), Larivière and Gingras (2011)

Standard textbooks only discuss the case of OLS, mentioning nothing about RoM and MoR (Eisenhauer, 2003; Gujarati, 2003; Wooldridge, 2009). Some scholars recommend RoM as a suitable imputation method for economic data, but no comparisons are made with OLS and MoR (Hu et al., 2001; de Waal et al., 2011; Thompson and Washington, 2012). Some scholars also argue that RoM is less biased than MoR (Gupta and Kabe, 2011; Office for National Statistics, 2014). Yet, other scholars contend that MoR is less biased than RoM (Hoenig et al., 1997; Zarnoch and Bechtold, 2000; Liu et al., 2005; Zou et al., 2010; Larivière and Gingras, 2011).

Therefore, there is no clear consensus as to which imputation method should be generally used for a heteroskedastic log-normal variable. The literature shows that the superiority of the methods differs on a case-by-case basis, but what are the conditions that dictate the superiority of the methods? How do we know which condition is relevant to the data we have? The current study answers these questions and proposes a solution to how we can choose the best ratio imputation model, given the nature of the data in hand.

3.4 Unifying the Competing Ratio Estimators

Egghe (2012) contends that the ratio of means (RoM) and the mean of ratios (MoR) are equivalent if $b = 0$ under the model of $z = a + bx$, where $z = y/x$. Apparently, this model is the same as the ratio imputation model described in equation (3.1), where $\beta = a$.

$$Z_i = a + bX_i$$

$$\frac{Y_i}{X_i} = a + bX_i$$

$$\frac{Y_i}{X_i} = a + 0X_i$$

$$Y_i = aX_i \tag{3.5}$$

In fact, OLS is also equivalent to RoM and MoR under this model. From equations (3.6), (3.7) and (3.8), it can be easily shown that OLS, RoM, and MoR are exactly the same constant a under equation (3.5). This is true regardless of the underlying distributions of Y_i and X_i.

$$\text{OLS} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{\sum X_i a X_i}{\sum X_i^2} = \frac{\sum a X_i^2}{\sum X_i^2} = \frac{a \sum X_i^2}{\sum X_i^2} = a \tag{3.6}$$

$$\text{MoR} = \frac{1}{n}\sum \frac{Y_i}{X_i} = \frac{1}{n}\sum \frac{aX_i}{X_i} = \frac{1}{n}\sum a = \frac{1}{n} na = a \tag{3.7}$$

$$\text{RoM} = \frac{\sum Y_i}{\sum X_i} = \frac{\sum aX_i}{\sum X_i} = \frac{a\sum X_i}{\sum X_i} = a \tag{3.8}$$

However, what is missing in the above argument is the disturbance component. In other words, equation (3.5) specifies the perfect linear relationship between Y_i and X_i. On the other hand, in real data, ratio imputation assumes that Y_i and X_i are not expected to be perfectly related, unlike in cold deck imputation.

Now, suppose that the population model is equation (3.9), where Y_i and X_i are log-normally distributed and $\varepsilon_i \sim N\left(0, \sigma X_i^\theta\right)$, i.e., the mean is zero and the standard deviation is σX_i^θ. If we estimate the slope β by OLS, we obtain $\hat{\beta}_{OLS}$ in equation (3.10).

$$Y_i = \beta X_i + \varepsilon_i \tag{3.9}$$

$$\hat{\beta}_{OLS} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{3.10}$$

The fact that Y_i and X_i are log-normally distributed implies that the standard deviation of ε_i may not be constant conditional on X_i, i.e., $\mathrm{sd}(\varepsilon|X_i) = \sigma X_i^\theta$, meaning that ε_i is heteroskedastic (Wooldridge, 2009). As is shown in Appendix 3.1, log-transformation may be inappropriate to take care of heteroskedasticity if the goal is to estimate the mean of raw data. Instead, following Royall (1970) and Cochran (1977), we will use weighted least squares (WLS) to transform heteroskedastic errors ε_i into homoskedastic errors γ_i, where $\gamma_i \sim N(0,\ \sigma)$. Since X_i^θ is a function of X_i, $\varepsilon_i / X_i^\theta$ has the expected value of zero conditional on X_i. Also, the standard deviation of $\varepsilon_i / X_i^\theta$ is σ conditional on X_i. Therefore, to correct for heteroskedasticity, equation (3.9) is transformed into equation (3.11).

$$\frac{Y_i}{X_i^\theta} = \frac{\beta X_i}{X_i^\theta} + \frac{\varepsilon_i}{X_i^\theta} \tag{3.11}$$

For simplicity, let $\gamma_i = \varepsilon_i / X_i^\theta$. Since $X_i / X_i^\theta = X_i^{1-\theta}$, we have equation (3.12). This means that the WLS estimate of β can be obtained from equation (3.13).

$$\frac{Y_i}{X_i^\theta} = \beta X_i^{1-\theta} + \gamma_i \tag{3.12}$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-\theta} \frac{Y_i}{X_i^\theta}}{\sum \left(X_i^{1-\theta}\right)^2} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}} \tag{3.13}$$

Also, multiplying both sides of equation (3.12) by X_i^{θ} yields equation (3.14). This clearly shows that each of the three competing ratio imputation models is a special case of this generalized model. The actual value of β depends on the parameter value of θ. Also see Appendix 3.2.1 for proof.

$$Y_i = \beta X_i + X_i^{\theta} \gamma_i \tag{3.14}$$

When $\theta = 0$, equation (3.14) is reduced to equation (3.15), where β can be estimated by OLS in equation (3.16). This means that, OLS is the best linear unbiased estimator (BLUE) under the assumption of homoskedasticity (classical Gauss-Markov theorem). Also see Appendix 3.2.2 for proof.

$$Y_i = \beta X_i + \gamma_i \tag{3.15}$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}} = \frac{\sum X_i^{1-2\times 0} Y_i}{\sum X_i^{2(1-0)}} = \frac{\sum X_i Y_i}{\sum X_i^2} = \hat{\beta}_{OLS} \tag{3.16}$$

When $\theta = 0.5$, equation (3.14) is reduced to equation (3.17), where β can be estimated by the ratio of means (RoM) in equation (3.18). This means that RoM is BLUE under the assumption of heteroskedasticity that the standard deviation of γ_i is proportional to $\sqrt{X_i}$. Also see Appendix 3.2.3 for proof.

$$Y_i = \beta X_i + \sqrt{X_i} \gamma_i \tag{3.17}$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}} = \frac{\sum X_i^{1-2\times 0.5} Y_i}{\sum X_i^{2(1-0.5)}} = \frac{\sum X_i^{0} Y_i}{\sum X_i} = \frac{\sum Y_i / n}{\sum X_i / n} = \frac{\bar{Y}}{\bar{X}} = \hat{\beta}_{RoM} \tag{3.18}$$

When $\theta = 1.0$, equation (3.14) is reduced to equation (3.19), where β can be estimated by the mean of ratios (MoR) in equation (3.20). This means that MoR is BLUE under the assumption of heteroskedasticity that the standard deviation of γ_i is proportional to X_i. Also see Appendix 3.2.4 for proof.

$$Y_i = \beta X_i + X_i \gamma_i \tag{3.19}$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}} = \frac{\sum X_i^{1-2\times 1} Y_i}{\sum X_i^{2(1-1)}} = \frac{\sum X_i^{-1} Y_i}{\sum X_i^{0}} = \frac{\sum X_i^{-1} Y_i}{\sum 1} = \frac{\sum X_i^{-1} Y_i}{n} = \frac{1}{n} \sum X_i^{-1} Y_i \tag{3.20}$$

$$= \frac{1}{n} \sum \frac{Y_i}{X_i} = \hat{\beta}_{MoR}$$

Therefore, all of the three estimators are BLUE given an appropriate value of θ. This fact is what Cochran (1997) called model-unbiasedness. Furthermore, all of the three estimators are BLUE under any sampling plan specified by Royall (1970, p.380). Thus, they are sampling-unbiased as well.

3.5 Monte Carlo Evidence for Ratio Imputation Models

Using the simulated datasets, this section compares the Relative Root Mean Square Errors (RRMSE) of the estimators for the mean across different missing data handling techniques. The Monte Carlo experiments here are based on 1,000 iterations, each of which is a random draw from the following multivariate log-normal distribution with $n = 1000$.

$Y_i = 1.1X_i + \varepsilon_i$, where

$X_i \sim LN(\text{meanLog} = 0, \text{sdLog} = 1)$

$\varepsilon_i \sim N(\text{mean} = 0, \text{sd} = X_i^{\theta})$

The value of θ is changed from -1.0 to 2.0 with a 0.1 increment, thus creating 31 different patterns of 1,000 datasets, i.e., a total of 31,000 datasets. Note that in other few runs, not reported here, the parameter values of β were changed, and the conclusions were very similar. Computations are done in R 3.2.4, where X_i is generated by the `rlnorm` function and ε_i is generated by the `rnorm` function.

Furthermore, following King et al. (2001), each of these 31,000 datasets is made incomplete using the MAR data generation process that was introduced in Section 2. Note that Variable y is the target incomplete variable for imputation, Variable x is completely observed in all of the situations to be used as the auxiliary variable, and Variable u is 1,000 sets of continuous uniform random numbers ranging from 0 to 1 for the missingness mechanism. Under the assumption of MAR, the missingness of y depends on the values of x and u. In other words, y is missing if x is smaller than the median of x and u is larger than 0.5, i.e., y missing if $x < \text{median}(x)$ and $u > 0.5$. We assume that the missing values are scattered among small-and-medium size enterprises, because the missing values of turnover for large enterprises are collected through recontacts in official statistics (de Waal et al., 2011, pp.245-246). The average missing rates are set to 25%.

The overall performance can be assessed by the Mean Square Error (MSE), which is defined as equation (3.21), where ϑ is the true population parameter and $\hat{\vartheta}$ is an estimator. The MSE measures the spread around the true value of the parameter, suggesting that an estimator with the smallest MSE is the best of a competing set of estimators (Gujarati, 2003).

$$\text{MSE}(\hat{\vartheta}) = \text{E}(\hat{\vartheta} - \vartheta)^2 \tag{3.21}$$

For the ease of interpretation, following Di Zio and Guarnera (2013), this study uses the Relative Root Mean Square Error (RRMSE), which is defined as equation (3.22), where ϑ is the truth, $\hat{\vartheta}$ is an estimator, and T is the number of trials, i.e., 1,000. In our specific example, ϑ is $\bar{Y}$. (Note that ϑ is a generic parameter notation which is different from θ as a specific parameter for the magnitude of heteroskedasticity.)

$$\text{RRMSE}(\hat{\vartheta}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\vartheta} - \vartheta}{\vartheta}\right)^2} \tag{3.22}$$

The results are presented in Figure 3.1. As was theoretically expected, in terms of the comparison among the ratio imputation models, it boils down to the assumption of $X_i^{\theta}\gamma_i$. If θ is less than 0.5, the MoR model consistently performs worse than the RoM and OLS models, but the performance of the RoM and OLS models is almost indistinguishable. When θ is larger than

0.6, the performance of the OLS model gets rapidly worse while the RoM model performs best. Up to the point where $\theta = 1.0$, the RoM model well competes against the MoR model. Only when θ is larger than 1.0, can the MoR model consistently perform better than the RoM model.

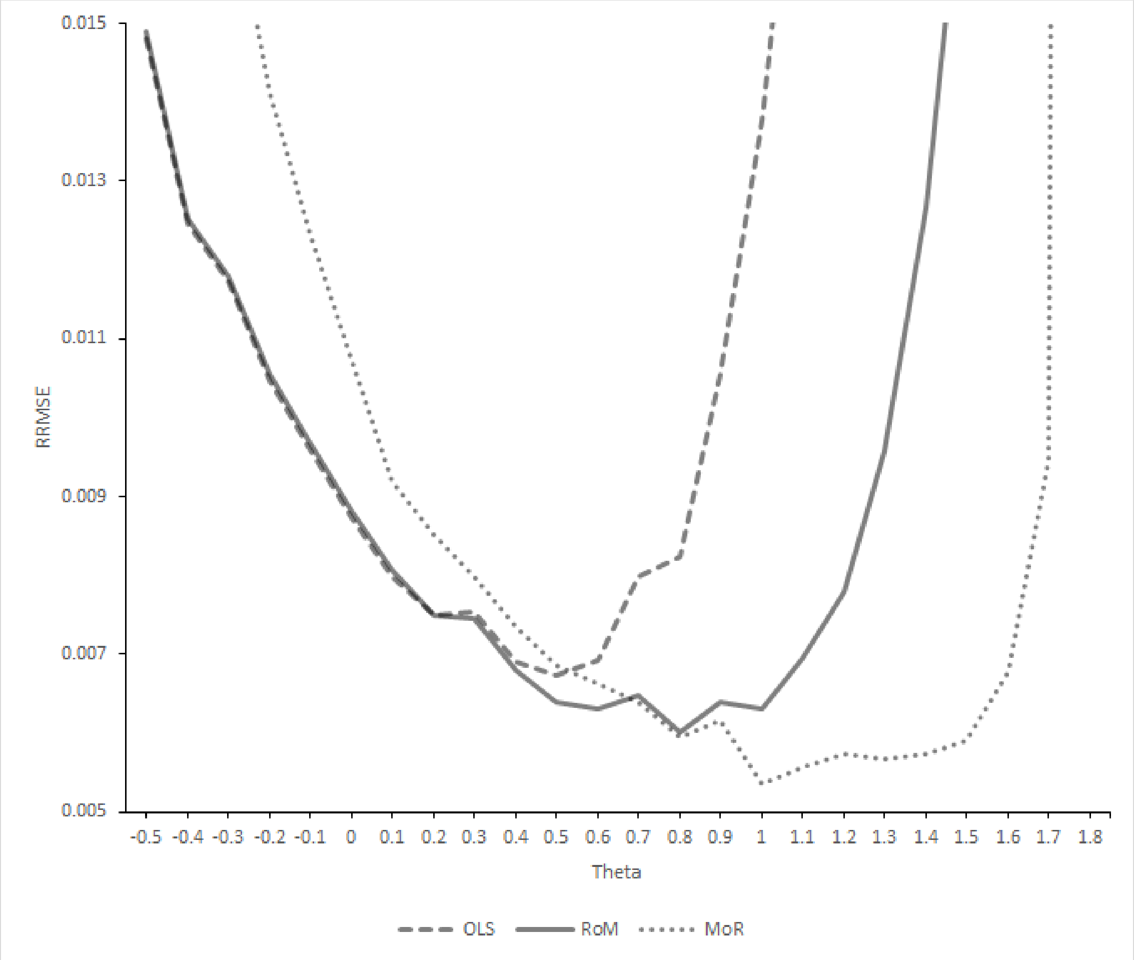


Figure 3.1. Relative root mean square error comparisons for different values of θ among the three competing ratio imputation models

In fact, when θ is 0.0 the OLS model is the best of imputing the mean of Y_i, when θ is 0.5 the RoM model is the best of imputing the mean of Y_i, and when θ is 1.0 the MoR model is the best of imputing the mean of Y_i. The results in Figure 3.1 imply that, if we could get to know the value of θ, we would be better off in imputing the mean of a heteroskedastic log-normal variable.

3.6 Estimating the Value of Theta

As Gujarati (2003) notes, “heteroskedasticity may be a matter of intuition, educated guesswork, prior empirical experience, or sheer speculation” (p.401). The literature is devoid of the method of estimating θ. This section presents an attempt to estimate the value of θ, trying to go beyond intuition, educated guesswork, prior empirical experience, or sheer speculation. In this section, two methods are presented: One is a graphical method of guesstimating θ; and the other is a numerical method of estimating θ.

3.6.1 Graphical Method of Guesstimating Theta

Cochran (1977) recommends, "When we are trying to decide what kind of estimate to use, a graph in which the sample values of y_i are plotted against those of x_i is helpful" (pp.159-160). This serves as an informal method of guessing the value of θ. Zarnoch and Bechtold (2000) followed Cochran's (1977) advice when they decided which ratio estimator to use in their analysis.

Following this recommendation, Figure 3.2 presents scatterplots between Y_i and X_i, where the values of θ are changed from -1.0 to 1.5 in a 0.5 increment, using the same specification in Section 3.5. These are the theoretical graphs we can expect to see with different θ values and may be used as a quick guide for the diagnostic purposes. A complete R-code is included in Appendix 3.3.

3.6.2 Numerical Method of Estimating Theta

Before proceeding to the presentation of the proposed method, our journey should start with the Breusch-Pagan test for heteroskedasticity (Gujarati, 2003; Wooldridge, 2009). If the result in this test turns out to be statistically significant, then data are heteroskedastic; otherwise, data are assumed to be homoskedastic, meaning that $\theta = 0$. Therefore, we should first apply this method. If the test result is statistically significant, the proposed method in this section will be helpful in determining the value of θ. An interested reader may be referred to Hothorn et al. (2015), who present the R-function to perform this test.

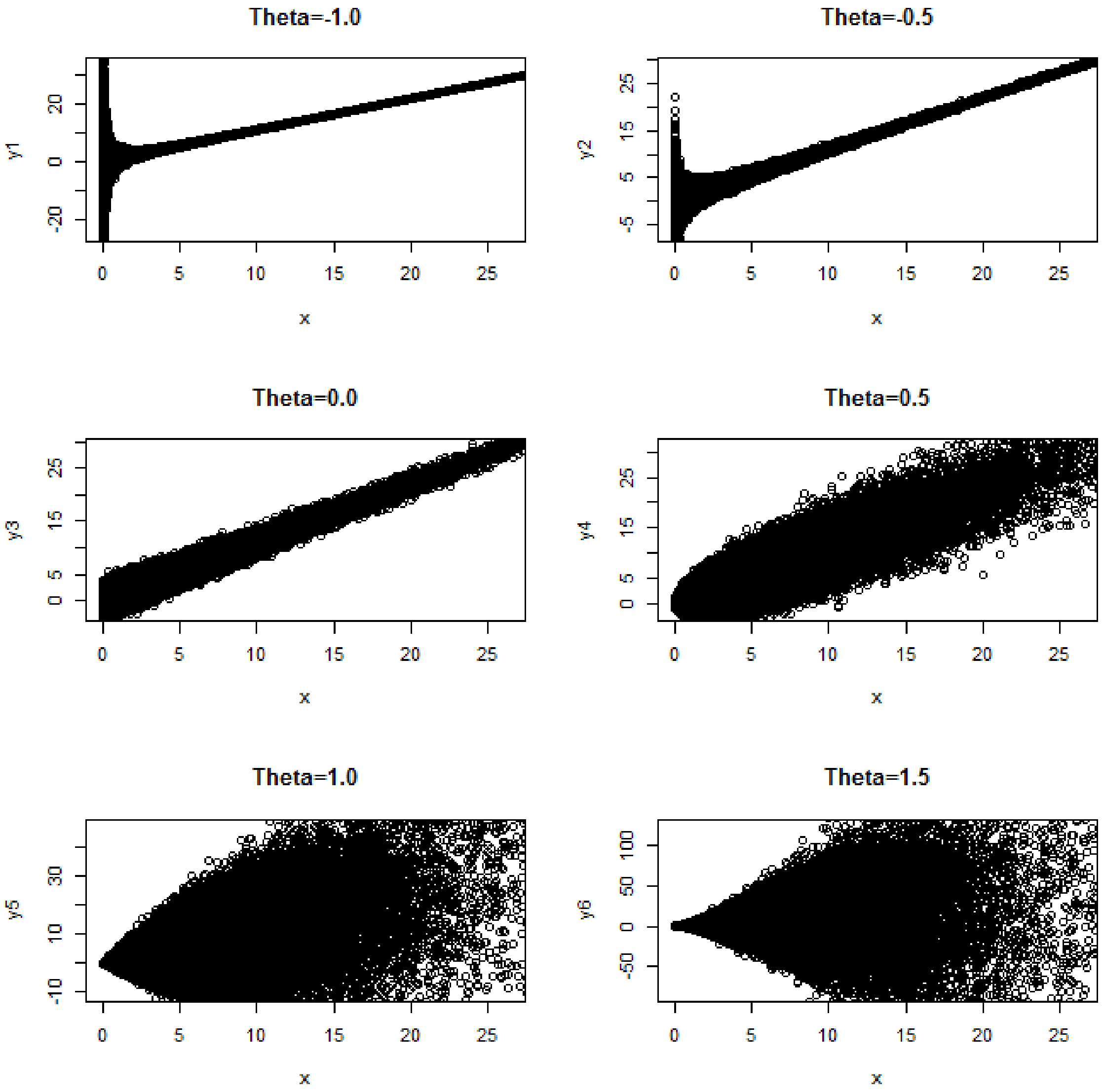


Figure 3.2. Theoretical Scatterplots between Y_i and X_i based on different values of θ from -1.0 to 1.5, where $n = 1{,}000{,}000$.

Suppose that the population model is equation (3.9). An assumption required for the proposed method of estimation is that the functional form of the population model is known. Y_i and X_i are given in the data. Remember that $\mathrm{sd}(\varepsilon|X) = \sigma X^{\theta}$. Therefore, what is unknown here is the values of β and θ. In order to correctly estimate the value of β, we need to estimate the value of θ.

$$Y_i = \beta X_i + \varepsilon_i \tag{3.9}$$

Our method tries to estimate the value of θ with no assumptions on the value of β. Cochran's

(1977) recommendation of the graph of y_i plotted against x_i is insightful, on which our strategy will be based as in equation (3.23), where $Q1_X$, $Q2_X$, and $Q3_X$ represent the 1st quartile of X_i, 2nd quartile of X_i, and 3rd quartile of X_i, respectively; $\mathrm{sd}(Y_i)$ represents the standard deviation of Y_i; thus, $\mathrm{sd}(Y_i)_{Q1_X}$ is the standard deviation of Y_i between the minimum and the 25th percentile of X_i, $\mathrm{sd}(Y_i)_{Q2_X}$ is the standard deviation of Y_i between the 25th percentile and the 50th percentile of X_i, and $\mathrm{sd}(Y_i)_{Q3_X}$ is the standard deviation of Y_i between the 50th percentile and the 75th percentile of X_i.

$$\hat{\theta} = \frac{1}{2}\left(\frac{\mathrm{sd}(Y_i)_{Q2_X} - \mathrm{sd}(Y_i)_{Q1_X}}{\mathrm{sd}(Y_i)_{Q1_X}} + \frac{\mathrm{sd}(Y_i)_{Q3_X} - \mathrm{sd}(Y_i)_{Q2_X}}{\mathrm{sd}(Y_i)_{Q2_X}}\right)\frac{\exp(1)}{\exp\left(\mathrm{sd}\big(\log(X_i)\big)\right)} \tag{3.23}$$

First, we categorize the data into four groups based on the values of X_i, i.e., the 25th percentile, 50th percentile, and 75th percentile, where we calculate the standard deviation of Y_i in each category. We take the growth rate from the first category to the second category, and the second category to the third category. We will ignore the fourth category because a large number of outliers exist in this area. Then, we take the average of the two growth rates. Since our model is two-variable, if the standard deviation of $\log(X_i)$ is 1.0, this shows the dispersion of ε_i; in other words, this approximately estimates the value of θ. However, the variance of Y_i depends on both the variances of X_i and ε_i. When the standard deviation of $\log(X_i)$ is not 1.0, we need a correction factor. This can be attained through dividing $\exp(1)$ by the exponent of the standard deviation of $\log(X_i)$, which takes the dispersion of X_i into account. Note that if the standard deviation of $\log(X_i)$ is 1.0, this correction factor will be $\exp(1)/\exp(1) =$ 1.0. A complete *R*-code is included in Appendix 3.4.

With only one equation in hand, it is not possible to analytically solve the two unknowns, β and θ. Therefore, we cannot analytically evaluate equation (3.23). Instead, we use simulation to assess equation (3.23) across different parameter settings. As before, the Monte Carlo experiments here are based on 1,000 iterations, each of which is a random draw from the following multivariate log-normal distribution with $n = 1000$.

$Y_i = \beta X_i + \varepsilon_i$, where

$X_i \sim LN(\text{meanLog} = \mu, \text{sdLog} = \sigma)$

$\varepsilon_i \sim N(\text{mean} = 0, \text{sd} = X_i^{\theta})$

The value of β is from 0.6 to 1.6 with a 0.5 increment, the value of μ is from 0.0 to 1.0 with a 0.5 increment, the value of σ is from 1.0 to 2.0 with a 0.5 increment, and the value of θ is from -0.5 to 1.5 with a 0.5 increment, thus creating 135 different patterns of 1,000 datasets, i.e., a total of 135,000 datasets. Computations are done in R 3.2.4, where X_i is generated by the `rlnorm` function and ε_i is generated by the `rnorm` function. The reported values are the average value of estimated θ based on 1,000 iterations followed by RMSE in parentheses. Note that the values in parentheses are the Root Mean Squared Error (RMSE), not the Relative Root Mean Squared Error (RRMSE), because in case of $\theta = 0$, RRMSE will be undefined.

The results are presented in Table 3.2. Regardless of the parameter values of β, μ, and σ, the performance of the proposed method is quite high in estimating the value of θ, when the true value of θ is 0.0, 0.5, or 1.0. Based on the estimated value of θ after rounding, we can easily choose among equations (3.2), (3.3), and (3.4). When the true value of θ is -0.5, there is some difficulty, where the estimated value is between -0.338 and 0.146. However, our method first applies the Breusch-Pagan test for heteroskedasticity. Thus, if the true state of the world is homoskedasticity, then the Breusch-Pagan test would not reject the null hypothesis. Therefore, if we reject the Breusch-Pagan test and our method estimates θ being less than 0.0, then the true value of θ is assumed to be negative. Also, when the true value of θ is 1.5, there is some difficulty, where the estimated value is between 1.541 and 2.443; however, we can be still confident that if the estimated value is larger than 1.5, then the true value of θ is not 0.0, 0.5, or 1.0. After all, our goal is to choose one of the ratio imputation models from equations (3.2), (3.3), and (3.4).

Table 3.2. Results of estimated θ values with different parameter settings

			θ				
			-0.5	0.0	0.5	1.0	1.5
$\beta = 0.6$	$\mu = 0.0$	$\sigma = 1.0$	-0.338 (0.164)	0.001 (0.008)	0.466 (0.061)	1.070 (0.110)	1.860 (0.386)
		$\sigma = 1.5$	-0.233 (0.286)	0.000 (0.004)	0.467 (0.052)	1.155 (0.175)	2.154 (0.672)
		$\sigma = 2.0$	-0.018 (0.486)	0.000 (0.003)	0.417 (0.093)	1.124 (0.157)	2.261 (0.786)
	$\mu = 0.5$	$\sigma = 1.0$	-0.317 (0.185)	0.000 (0.007)	0.467 (0.060)	1.069 (0.108)	1.893 (0.416)
		$\sigma = 1.5$	-0.203 (0.310)	0.001 (0.006)	0.469 (0.052)	1.149 (0.170)	2.235 (0.7523)
		$\sigma = 2.0$	-0.013 (0.489)	0.001 (0.009)	0.423 (0.092)	1.123 (0.168)	2.356 (0.887)
	$\mu = 1.0$	$\sigma = 1.0$	-0.240 (0.262)	0.002 (0.012)	0.467 (0.061)	1.070 (0.109)	1.908 (0.432)
		$\sigma = 1.5$	-0.106 (0.397)	0.002 (0.014)	0.479 (0.050)	1.155 (0.178)	2.273 (0.794)
		$\sigma = 2.0$	-0.003 (0.497)	0.002 (0.016)	0.440 (0.080)	1.121 (0.160)	2.443 (0.972)
$\beta = 1.1$	$\mu = 0.0$	$\sigma = 1.0$	-0.323 (0.179)	0.001 (0.008)	0.466 (0.059)	1.032 (0.084)	1.711 (0.241)
		$\sigma = 1.5$	-0.214 (0.300)	0.002 (0.009)	0.483 (0.045)	1.111 (0.134)	1.900 (0.424)
		$\sigma = 2.0$	-0.014 (0.488)	0.001 (0.009)	0.449 (0.063)	1.079 (0.122)	1.910 (0.440)
	$\mu = 0.5$	$\sigma = 1.0$	-0.262 (0.240)	0.003 (0.014)	0.469 (0.059)	1.030 (0.085)	1.788 (0.314)
		$\sigma = 1.5$	-0.130 (0.375)	0.002 (0.016)	0.498 (0.040)	1.109 (0.140)	2.021 (0.542)
		$\sigma = 2.0$	-0.005 (0.495)	0.003 (0.022)	0.475 (0.053)	1.080 (0.124)	2.080 (0.604)
	$\mu = 1.0$	$\sigma = 1.0$	-0.088 (0.414)	0.004 (0.021)	0.471 (0.057)	1.033 (0.085)	1.848 (0.374)
		$\sigma = 1.5$	0.037 (0.539)	0.005 (0.033)	0.519 (0.047)	1.105 (0.132)	2.133 (0.650)
		$\sigma = 2.0$	0.011 (0.512)	0.008 (0.044)	0.512 (0.052)	1.087 (0.117)	2.210 (0.734)
$\beta = 1.6$	$\mu = 0.0$	$\sigma = 1.0$	-0.303 (0.198)	0.003 (0.013)	0.468 (0.061)	0.987 (0.081)	1.541 (0.117)
		$\sigma = 1.5$	-0.184 (0.326)	0.002 (0.012)	0.505 (0.043)	1.056 (0.094)	1.665 (0.205)
		$\sigma = 2.0$	-0.009 (0.492)	0.002 (0.015)	0.493 (0.049)	1.030 (0.098)	1.656 (0.220)
	$\mu = 0.5$	$\sigma = 1.0$	-0.195 (0.307)	0.004 (0.021)	0.471 (0.058)	0.981 (0.079)	1.666 (0.207)
		$\sigma = 1.5$	-0.059 (0.443)	0.004 (0.027)	0.529 (0.052)	1.058 (0.094)	1.817 (0.343)
		$\sigma = 2.0$	0.003 (0.503)	0.005 (0.035)	0.535 (0.060)	1.031 (0.110)	1.822 (0.356)
	$\mu = 1.0$	$\sigma = 1.0$	0.039 (0.542)	0.006 (0.037)	0.475 (0.057)	0.981 (0.081)	1.751 (0.279)
		$\sigma = 1.5$	0.146 (0.651)	0.008 (0.050)	0.565 (0.082)	1.056 (0.094)	1.955 (0.477)
		$\sigma = 2.0$	0.020 (0.524)	0.008 (0.053)	0.582 (0.105)	1.031 (0.099)	1.986 (0.518)

Note: The reported values are the average value of estimated θ based on 1,000 iterations followed by RMSE in parentheses.

Therefore, no matter what the parameter values of β, μ, and σ, our proposed method allows us to choose an appropriate imputation model from equations (3.2), (3.3), and (3.4), when the true value of θ is 0.0, 0.5, or 1.0.

3.7 Example: Application to Real Economic Data

The issue under investigation is of particular importance in official economic statistics because the primary goal of surveys in official statistics is to calculate the total (or the mean) of a heteroskedastic log-normal variable. However, the current issue may be also relevant in other areas of social statistics, such as a cross-national comparison of GDP.

This section illustrates the application of the proposed method to a concrete real dataset, by utilizing Penn World Table 9.0 (Feenstra et al., 2016), where we obtained CGDPe in 2011 and CGDPo in 2005. CGDPe is "expenditure-side real GDP at current PPPs, to compare relative living standards across countries at a single point in time" and CGDPo is "output-side real GDP at current PPPs, to compare relative productive capacity across countries at a single point in time" (Feenstra et al., 2016). The number of observations is 167 countries. Table 3.3 presents the summary statistics of the two variables.

Table 3.3. Summary Statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.Dev.
CGDPe2011	488	20690	69060	556000	357000	15590000	1756865
CGDPo2005	369	14950	42970	415300	239200	14710000	1419322

Figure 3.3 graphically displays the distributions of the two variables, which shows that CGDPe2011 and CGDPo2005 are both highly skewed to the right, but their distributions are close to normality after log-transformation. Therefore, these variables can be considered log-normal.

The correlation between CGDPe2011 and CGDPo2005 is 0.975, but as Figure 3.4 graphically presents, the relationship between the two variables may not be homoskedastic. Therefore, these variables can be considered heteroskedastic log-normal.

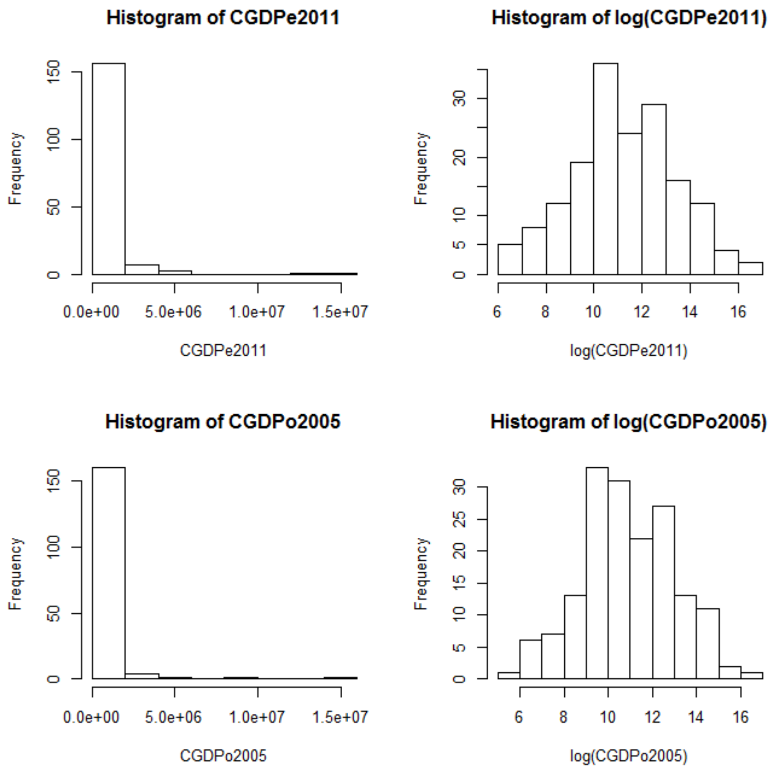


Figure 3.3. Histograms of CGDPe2011 and CGDPo2005.

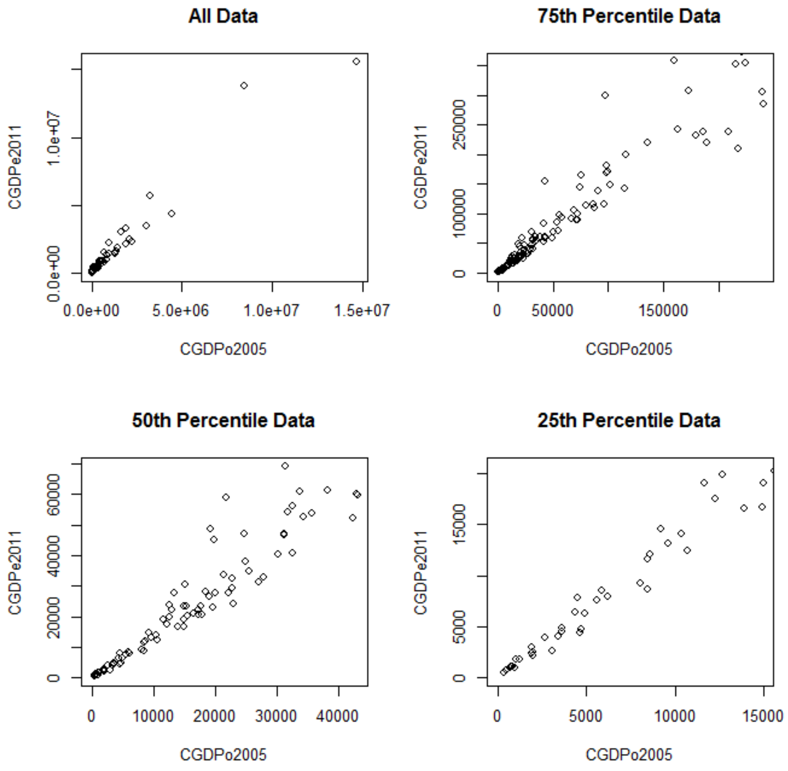


Figure 3.4. Scatterplot of CGDPe2011 and CGDPo2005.

Following the simulation settings in Section 3.5, the missing mechanism is introduced as MAR, meaning that the missingness of CGDPe2011 depends on the values of CGDPo2005 and uniform random numbers. In other words, CGDPe2011 is missing if CGDPo2005 is smaller than the median of CGDPo2005 and a uniform random number is larger than 0.5; thus, the missing rate is again set to 25%. We will impute the missing values in CGDPe2011 by using CGDPo2005 as an auxiliary variable in equation (3.24), where $\hat{\beta}$ is estimated by OLS, RoM, and MoR.

$$\widehat{\mathrm{CGDPe2011}}_i = \hat{\beta} \times \mathrm{CGDPo2005}_i \tag{3.24}$$

If we apply the proposed method to this dataset, the estimated value of θ is 0.93. Therefore, our method predicts that the imputations by MoR would be best in comparison with listwise deletion, RoM, and OLS. The results are summarized in Table 3.4. As was predicted by our method, the difference between the truth and MoR is smallest (Difference = 1071), followed by RoM (Difference = 1463) and OLS (Difference = 1873).

Table 3.4. Results of Example Data Analysis

	Truth	Listwise	OLS	RoM	MoR
Mean	555982	695702	554109	554518	554911
Difference		140791	1873	1463	1071

3.8 Conclusions

The method proposed in the current study calculates the standard deviations of Y_i below the first, second, and third quartiles, and estimates the magnitude of heteroskedasticity, with a correction factor for the variance of X_i. The results in the Monte Carlo simulation give a strong support for the method.

If the estimated value of θ is less than 1.0, then the ratio of means (RoM) should be used as a ratio imputation model. If the estimated value of θ is larger than 1.0, then the mean of ratios (MoR) should be used as a ratio imputation model. If the estimated value of θ is less than 0.3, OLS may be used as a ratio imputation model; however, our simulation results suggest that the performance of RoM is quite similar to that of OLS even when θ is less than 0.3.

The proposed method should be regarded as a first step toward the estimation of the θ value.

Future research should expand this method, in order to make it more rigorous and robust. One course of future research may be to apply this method to a variety of real economic data.

Appendix 3.1

Suppose that our goal is to estimate the population mean. Also, suppose that the population model is equation (3.9) with no intercept and the slope β, where Y_i and X_i are log-normally distributed.

$$Y_i = \beta X_i + \varepsilon_i \tag{3.9}$$

Since Y_i and X_i are log-normally distributed, log-transformation will produce normally distributed data. However, often times in official statistics, the estimation of the mean of $\hat{Y}_i$ is the goal, and the estimation of the mean of $\widehat{\log(Y_i)}$ is hardly the goal. If the estimated model after log-transformation is equation (3.25), the associated true model is equation (3.26). This means that the estimated model in raw data is equation (3.27) and the true model in raw data is equation (3.28) (Gujarati, 2003; Wooldridge, 2009).

$$\widehat{\log(Y_i)} = \hat{\delta}\log(X_i) \tag{3.25}$$

$$\log(Y_i) = \delta\log(X_i) + \varepsilon_i \tag{3.26}$$

$$\hat{Y}_i = X_i^{\hat{\delta}} \tag{3.27}$$

$$Y_i = X_i^{\delta}\exp(\varepsilon_i) \tag{3.28}$$

Let μ be the mean and σ^2 be the variance. Then, the expected value of log-normal variable Y_i is equation (3.29) (DeGroot and Schervish, 2002). This clearly shows that the exponent of $\widehat{\log(Y_i)}$ systematically underestimates the expected value of Y_i by the order of $\exp(\sigma^2/2)$. Since σ^2 is unknown, there are no universal ways of adjusting for this error (Also see Wooldridge, 2009, pp.210-212).

$$\mathrm{E}(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(\mu)\exp\left(\frac{\sigma^2}{2}\right) \tag{3.29}$$

This echoes von Hippel's (2013) finding that the safest way to impute a skewed variable using a normal model is to do so with no transformations.

Appendix 3.2

This appendix shows the expectations and the variances of the three estimators. First, the generalized outcome by weighted least squares (WLS) will be shown, followed by the specialized outcomes by ordinary least squares (OLS), ratio of means (RoM), and mean of ratios (MoR). We assume that the population model is equation (9), where $\varepsilon_i \sim N\left(0, \sigma X_i^\theta\right)$, i.e., the mean is zero and the standard deviation is σX_i^θ.

$$Y_i = \beta X_i + \varepsilon_i \tag{3.9}$$

In the following derivations, the usual assumptions in the classical Gauss Markov theorem apply (Gujarati, 2003, pp.66-71; Wooldridge, 2009, pp.157-158), except for the assumption of homoskedasticity. Assumption 1 states that the model is linear in the parameters as shown in equation (3.9). Assumption 2 states that n observations are randomly sampled from the population model of equation (3.9). Assumption 3 states that the values of X_i are fixed in repeated sampling, i.e., $\mathrm{E}(X_i)$ is nonstochastic; thus, this can be treated as constant. Assumption 4 states that the error term ε_i has an expected value of zero given X_i, i.e., $\mathrm{E}(\varepsilon_i | X_i) = 0$. Assumption 5 states zero covariance between ε_i and X_i, i.e., $\mathrm{E}(\varepsilon_i X_i) = 0$.

In the derivations below, we will extensively use the following three properties of expected values (Wooldridge, 2009, p.724). Property 1 states that for any constant c, $\mathrm{E}[c] = c$. Property 2 states that for any constants a and b, $\mathrm{E}[aX + b] = a\mathrm{E}[X] + b$. Property 3 states that if a_i are constants and X_i are random variables, where $i = 1, \dots n$, then $\mathrm{E}[\sum a_i X_i] = \sum a_i \mathrm{E}[X_i]$.

Note that, in all equations below, the sums are taken from $i = 1 \dots n$, that is, across all observations in the sample. Therefore, the limits of summation are not shown.

$$\sum X_i = \sum_{i=1}^{n} X_i$$

Appendix 3.2.1

The expected value of $\hat{\beta}_{WLS}$ can be proven to be β; therefore it is unbiased under the population model of equation (3.9).

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} Y_i}{\sum X_i^{2(1-\theta)}}$$

$$\hat{\beta}_{WLS} = \frac{\sum X_i^{1-2\theta} (\beta X_i + \varepsilon_i)}{\sum X_i^{2(1-\theta)}}$$

$$\hat{\beta}_{WLS} = \frac{\beta \sum X_i^{2(1-\theta)} + \sum X_i^{1-2\theta}\, \varepsilon_i}{\sum X_i^{2(1-\theta)}}$$

$$\hat{\beta}_{WLS} = \beta + \frac{\sum X_i^{1-2\theta}\, \varepsilon_i}{\sum X_i^{2(1-\theta)}}$$

$$\mathrm{E}\big[\hat{\beta}_{WLS}\big] = \mathrm{E}\left[\beta + \frac{\sum X_i^{1-2\theta}\, \varepsilon_i}{\sum X_i^{2(1-\theta)}}\right]$$

$$\mathrm{E}\big[\hat{\beta}_{WLS}\big] = \mathrm{E}[\beta] + \mathrm{E}\left[\frac{\sum X_i^{1-2\theta}\, \varepsilon_i}{\sum X_i^{2(1-\theta)}}\right]$$

$$\mathrm{E}\big[\hat{\beta}_{WLS}\big] = \beta$$

The last term on the right-hand side drops out, because we assume that the value of the covariance between the independent variable X_i and the error term ε_i is zero. Also see Gujarati (2003, p.100-101) and Wooldridge (2009, pp.114-116) as a reference guide.

The variance of $\hat{\beta}_{WLS}$ can be shown to be $\sigma^2 / \sum X_i^{2(1-\theta)}$, which is guaranteed to be the minimum variance. If the original equation meets the five Gauss-Markov assumptions (except homoskedasticity), then $\hat{\beta}_{WLS}$ meets all of the six Gauss-Markov assumptions including homoskedasticity (Gujarati, 2003, p.415; Wooldridge, 2009, p.278).

$$\mathrm{V}(\hat{\beta}_{WLS}) = \mathrm{E}\left[(\hat{\beta}_{WLS} - \beta)^2\right]$$

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sum X_i^{2(1-2\theta)}}{\left(\sum X_i^{2(1-\theta)}\right)^2}\sigma^2 X_i^{2\theta}$$

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sigma^2 \sum X_i^{2(1-\theta)}}{\left(\sum X_i^{2(1-\theta)}\right)^2}$$

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sigma^2}{\sum X_i^{2(1-\theta)}}$$

When $\theta = 0$, the following is the minimum variance.

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sigma^2}{\sum X_i^{2(1-\theta)}} = \frac{\sigma^2}{\sum X_i^2}$$

When $\theta = 0.5$, the following is the minimum variance.

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sigma^2}{\sum X_i}$$

When $\theta = 1.0$, the following is the minimum variance.

$$\mathrm{V}(\hat{\beta}_{WLS}) = \frac{\sigma^2}{n}$$

Appendix 3.2.2

The expected value of $\hat{\beta}_{OLS}$ can be proven to be β; therefore it is unbiased under the population model of equation (3.9).

$$\hat{\beta}_{OLS} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

$$\hat{\beta}_{OLS} = \frac{\sum X_i(\beta X_i + \varepsilon_i)}{\sum X_i^2}$$

$$\hat{\beta}_{OLS} = \frac{\beta \sum X_i^2 + \sum X_i\,\varepsilon_i}{\sum X_i^2}$$

$$\hat{\beta}_{OLS} = \beta + \frac{\sum X_i\,\varepsilon_i}{\sum X_i^2}$$

$$\mathrm{E}[\hat{\beta}_{OLS}] = \mathrm{E}\left[\beta + \frac{\sum X_i\,\varepsilon_i}{\sum X_i^2}\right]$$

$$\mathrm{E}[\hat{\beta}_{OLS}] = \mathrm{E}[\beta] + \mathrm{E}\left[\frac{\sum X_i\,\varepsilon_i}{\sum X_i^2}\right]$$

$$\mathrm{E}[\hat{\beta}_{OLS}] = \beta$$

The last term on the right-hand side drops out, because we assume that the value of the covariance between the independent variable X_i and the error term ε_i is zero. Also see Gujarati (2003, p.100-101) and Wooldridge (2009, pp.114-116) as a reference guide.

The variance of $\hat{\beta}_{OLS}$ can be shown to be $\sigma^2 \sum X_i^{2(1+\theta)}/\left(\sum X_i^2\right)^2$, which is equal to the minimum variance of $\mathrm{V}(\hat{\beta}_{WLS})$ when $\theta = 0$.

$$\mathrm{V}(\hat{\beta}_{OLS}) = \mathrm{E}\left[\left(\hat{\beta}_{OLS} - \beta\right)^2\right]$$

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{1}{\left(\sum X_i^2\right)^2}\sum X_i^2 \sigma^2 X_i^{2\theta}$$

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{\sigma^2 \sum X_i^2 X_i^{2\theta}}{\left(\sum X_i^2\right)^2}$$

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{\sigma^2 \sum X_i^{2(1+\theta)}}{\left(\sum X_i^2\right)^2}$$

When $\theta = 0$,

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{\sigma^2 \sum X_i^{2(1+\theta)}}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2 \sum X_i^2}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2}{\sum X_i^2} = \mathrm{V}(\hat{\beta}_{WLS})$$

When $\theta = 0.5$,

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{\sigma^2 \sum X_i^{2(1+\theta)}}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2 \sum X_i^3}{\left(\sum X_i^2\right)^2}$$

When $\theta = 1.0$,

$$\mathrm{V}(\hat{\beta}_{OLS}) = \frac{\sigma^2 \sum X_i^{2(1+\theta)}}{\left(\sum X_i^2\right)^2} = \frac{\sigma^2 \sum X_i^4}{\left(\sum X_i^2\right)^2}$$

Therefore, when $\theta = 0$, $\mathrm{V}(\hat{\beta}_{OLS})$ is equal to $\mathrm{V}(\hat{\beta}_{WLS})$, which proves that $\hat{\beta}_{OLS}$ is BLUE with θ being 0.

Appendix 3.2.3

The expected value of $\hat{\beta}_{RoM}$ can be proven to be β; therefore it is unbiased under the population model of equation (3.9).

$$\hat{\beta}_{RoM} = \frac{\sum Y_i}{\sum X_i}$$

$$\hat{\beta}_{RoM} = \frac{\sum(\beta X_i + \varepsilon_i)}{\sum X_i}$$

$$\hat{\beta}_{RoM} = \frac{\beta \sum X_i + \sum \varepsilon_i}{\sum X_i}$$

$$\hat{\beta}_{RoM} = \beta + \frac{\sum \varepsilon_i}{\sum X_i}$$

$$\mathrm{E}[\hat{\beta}_{RoM}] = \mathrm{E}\left[\beta + \frac{\sum \varepsilon_i}{\sum X_i}\right]$$

$$\mathrm{E}[\hat{\beta}_{RoM}] = \mathrm{E}[\beta] + \mathrm{E}\left[\frac{\sum \varepsilon_i}{\sum X_i}\right]$$

$$\mathrm{E}[\hat{\beta}_{RoM}] = \beta + \frac{\mathrm{E}[\sum \varepsilon_i]}{\sum X_i}$$

$$\mathrm{E}[\hat{\beta}_{RoM}] = \beta + \frac{\sum \mathrm{E}[\varepsilon_i]}{\sum X_i}$$

$$\mathrm{E}[\hat{\beta}_{RoM}] = \beta$$

The last term on the right-hand side drops out, because we assume that the expected value of ε_i is zero.

The variance of $\hat{\beta}_{RoM}$ can be shown to be $\sigma^2 \sum X_i^{2\theta}/(\sum X_i)^2$, which is equal to the minimum variance of $\mathrm{V}(\hat{\beta}_{WLS})$ when $\theta = 0.5$.

$$\mathrm{V}(\hat{\beta}_{RoM}) = \mathrm{E}\left[\left(\hat{\beta}_{RoM} - \beta\right)^2\right]$$

$$\mathrm{V}(\hat{\beta}_{RoM}) = \frac{1}{(\sum X_i)^2}\sum \sigma^2 X_i^{2\theta}$$

$$\mathrm{V}(\hat{\beta}_{RoM}) = \frac{\sigma^2 \sum X_i^{2\theta}}{(\sum X_i)^2}$$

When $\theta = 0$,

$$\mathrm{V}(\hat{\beta}_{RoM}) = \frac{\sigma^2 \sum X_i^{2\theta}}{(\sum X_i)^2} = \frac{\sigma^2 \sum X_i^{0}}{(\sum X_i)^2} = \frac{\sigma^2 \sum 1}{(\sum X_i)^2} = \frac{n\sigma^2}{(\sum X_i)^2}$$

When $\theta = 0.5$,

$$\mathrm{V}(\hat{\beta}_{RoM}) = \frac{\sigma^2 \sum X_i^{2\theta}}{(\sum X_i)^2} = \frac{\sigma^2 \sum X_i}{(\sum X_i)^2} = \frac{\sigma^2}{\sum X_i} = \mathrm{V}(\hat{\beta}_{WLS})$$

When $\theta = 1.0$,

$$\mathrm{V}(\hat{\beta}_{RoM}) = \frac{\sigma^2 \sum X_i^{2\theta}}{(\sum X_i)^2} = \frac{\sigma^2 \sum X_i^{2}}{(\sum X_i)^2}$$

Therefore, when $\theta = 0.5$, $\mathrm{V}(\hat{\beta}_{RoM})$ is equal to $\mathrm{V}(\hat{\beta}_{WLS})$, which proves that $\hat{\beta}_{RoM}$ is BLUE with θ being 0.5.

Appendix 3.2.4

The expected value of $\hat{\beta}_{MoR}$ can be proven to be β; therefore it is unbiased under the population model of equation (3.9).

$$\hat{\beta}_{MoR} = \frac{1}{n}\sum \frac{Y_i}{X_i}$$

$$\hat{\beta}_{MoR} = \frac{1}{n}\sum \frac{\beta X_i + \varepsilon_i}{X_i}$$

$$\hat{\beta}_{MoR} = \beta + \frac{1}{n}\sum \frac{\varepsilon_i}{X_i}$$

$$\mathrm{E}[\hat{\beta}_{MoR}] = \mathrm{E}\left[\beta + \frac{1}{n}\sum \frac{\varepsilon_i}{X_i}\right]$$

$$\mathrm{E}[\hat{\beta}_{MoR}] = \mathrm{E}[\beta] + \mathrm{E}\left[\frac{1}{n}\sum \frac{\varepsilon_i}{X_i}\right]$$

$$\mathrm{E}[\hat{\beta}_{MoR}] = \beta + \frac{1}{n}\mathrm{E}\left[\sum \frac{\varepsilon_i}{X_i}\right]$$

$$\mathrm{E}[\hat{\beta}_{MoR}] = \beta + \frac{1}{n}\sum \frac{\mathrm{E}[\varepsilon_i]}{X_i}$$

$$\mathrm{E}[\hat{\beta}_{MoR}] = \beta$$

The last term on the right-hand side drops out, because we assume that the expected value of ε_i is zero.

The variance of $\hat{\beta}_{MoR}$ can be shown to be $1/n^2 \sum \sigma^2 X_i^{2\theta}/X_i^2$, which is equal to the minimum variance of $\mathrm{V}(\hat{\beta}_{WLS})$ when $\theta = 1.0$.

$$\mathrm{V}(\hat{\beta}_{MoR}) = \mathrm{E}\left[\left(\hat{\beta}_{MoR} - \beta\right)^2\right]$$

$$\mathrm{V}(\hat{\beta}_{MoR}) = \frac{1}{n^2}\sum \frac{\sigma^2 X_i^{2\theta}}{X_i^2}$$

When $\theta = 0$,

$$\mathrm{V}(\hat{\beta}_{MoR}) = \frac{1}{n^2}\sum \frac{\sigma^2 X_i^{2\theta}}{X_i^2} = \frac{1}{n^2}\sum \frac{\sigma^2}{X_i^2}$$

When $\theta = 0.5$,

$$\mathrm{V}(\hat{\beta}_{MoR}) = \frac{1}{n^2}\sum \frac{\sigma^2 X_i^{2\theta}}{X_i^2} = \frac{1}{n^2}\sum \frac{\sigma^2 X_i}{X_i^2} = \frac{1}{n^2}\sum \frac{\sigma^2}{X_i}$$

When $\theta = 1.0$,

$$\mathrm{V}(\hat{\beta}_{MoR}) = \frac{1}{n^2}\sum \frac{\sigma^2 X_i^{2\theta}}{X_i^2} = \frac{1}{n^2}\sum \frac{\sigma^2 X_i^2}{X_i^2} = \frac{\sum \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} = \mathrm{V}(\hat{\beta}_{WLS})$$

Therefore, when $\theta = 1.0$, $\mathrm{V}(\hat{\beta}_{MoR})$ is equal to $\mathrm{V}(\hat{\beta}_{WLS})$, which proves that $\hat{\beta}_{MoR}$ is BLUE with θ being 1.0.

Appendix 3.3

This appendix presents the *R*-code to graphically assess the value of θ. We assume that the target variable for imputation is stored in the first column of the data and the auxiliary variable is stored in the second column of the data. Also, the name of the dataset is “data”. For the values of min and max, specify appropriate numbers between 0 and 1.

```
attach(data)
min<-0.001
max<-0.999
plot(data[,2],data[,1],
      xlim=c(as.numeric(quantile(data[,2],min,na.rm=TRUE)),
              as.numeric(quantile(data[,2],max,na.rm=TRUE))),
      ylim=c(as.numeric(quantile(data[,1],min,na.rm=TRUE)),
              as.numeric(quantile(data[,1],max,na.rm=TRUE))),
  )
```

Appendix 3.4

This appendix presents the *R*-code to numerically assess the value of θ. We assume that the target variable for imputation is stored in the first column of the data and the auxiliary variable is stored in the second column of the data. Also, the name of the dataset is “data”.

```
attach(data); matdata<-matrix(NA,nrow(data),4); require(lmtest)
for(i in 1:nrow(data)){
  if (data[i,2]<as.numeric(quantile(data[,2])[2])){
  matdata[i,1]<-data[i,1]
  }
  else if (data[i,2]<as.numeric(quantile(data[,2])[3])){
  matdata[i,2]<-data[i,1]
  }
  else if (data[i,2]<as.numeric(quantile(data[,2])[4])){
  matdata[i,3]<-data[i,1]
  }
  else{matdata[i,4]<-data[i,1]
  }
}
data2<-data.frame(matdata)
 s1<-sd(data2[,1],na.rm=TRUE); s2<-sd(data2[,2],na.rm=TRUE)
 s3<-sd(data2[,3],na.rm=TRUE); r1<-(s2-s1)/s1; r2<-(s3-s2)/s2
 estimation1<-(r1+r2)/2*exp(1)/exp(sd(log(data[,2])))
pvalue<-as.numeric(bptest(data[,1]~data[,2])$p.value)
if (pvalue>0.05){
  estimation2<-0.0
  }else{
    estimation2<-estimation1
  }
estimation2
```

4 Comparison of MCMC and Non-MCMC Multiple Imputation Algorithms

This chapter derived from Takahashi (2017d), a peer-reviewed article in the *Data Science Journal*, which is sponsored by CODATA (Committee on Data for Science and Technology), an interdisciplinary scientific committee of the International Council for Science (ICSU). The *Data Science Journal* is indexed in Scopus by Elsevier as of April 2017. The author would like to thank Ubiquity Press for permission to use "Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations" (*Data Science Journal*, in press).

4.1 Introduction

Generally, it is quite difficult to obtain complete data in social surveys (King et al., 2001, p.49). Consequently, available data are not only inefficient due to the reduced sample size, but also biased due to the difference between respondents and non-respondents, thus making statistical inference invalid. Since Rubin (1987), multiple imputation has been known to be the standard method of handling missing data (Graham, 2009; Baraldi and Enders, 2010; Carpenter and Kenward, 2013; Raghunathan, 2016).

While the theoretical concept of multiple imputation has been around for decades, the implementation is difficult because making a random draw from the posterior distribution is a complicated matter. Accordingly, there are several computational algorithms in software (Schafer, 1997; Honaker and King, 2010; van Buuren, 2012). The most traditional algorithm is Data Augmentation (DA) followed by the other two new algorithms, Fully Conditional Specification (FCS) and Expectation-Maximization with Bootstrapping (EMB). Although an abundant literature exists on the comparisons between joint modeling (DA, EMB) and conditional modeling (FCS), no comparisons have been made about the relative superiority between the MCMC algorithms (DA, FCS) and the non-MCMC algorithm (EMB), where MCMC stands for Markov chain Monte Carlo. This study assesses the effects of between-imputation iterations on the performance of the three multiple imputation algorithms, using Monte Carlo experiments.

By way of organization, Section 4.2 introduces the notations in this chapter. Section 4.3 gives a motivating example of missing data analysis in social sciences. Section 4.4 presents the assumptions of imputation methods. Section 4.5 shows the traditional methods of handling missing data. Section 4.6 introduces the three multiple imputation algorithms. Section 4.7 surveys the literature on multiple imputation. Sections 4.8 gives the results of the Monte Carlo experiments, showing the impact of between-imputation iterations on multiple imputation. Section 4.9 concludes with the findings and the limitations in the current research.

4.2 Notations

$\mathbf{D}$ is $n \times p$ data, where n is the sample size and p is the number of variables. The distribution of $\mathbf{D}$ is multivariate-normal with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{D} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where all of the variables are continuous. Let i refer to an observation index ($i = 1, \dots, n$). Let j refer to a variable index ($j = 1, \dots, p$). Let $\mathbf{D} = \{\mathbf{Y_1}, \dots, \mathbf{Y_p}\}$, where $\mathbf{Y_j}$ is the j-th column in $\mathbf{D}$ and $\mathbf{Y_{-j}}$ is the complement of $\mathbf{Y_j}$, i.e., all columns in $\mathbf{D}$ except $\mathbf{Y_j}$. Depending on the model specification, $\mathbf{Y_{-j}}$ are denoted $\mathbf{X_1}, \dots, \mathbf{X_{p-1}}$, where the p-th variable is $\mathbf{Y}$. Also, let $\mathbf{Y_{obs}}$ be observed data and $\mathbf{Y_{mis}}$ be missing data: $\mathbf{D} = \{\mathbf{Y_{obs}}, \mathbf{Y_{mis}}\}$.

Let $\mathbf{R}$ be a response indicator matrix that has the same dimension as $\mathbf{D}$. Whenever $\mathbf{D}$ is observed, $\mathbf{R} = 1$; otherwise, $\mathbf{R} = 0$. Note, however, that italicized R refers to the R statistical environment. In the multiple imputation context, M refers to the number of imputations and T refers to the number of between-imputation iterations. In general, θ is an unknown parameter.

4.3 Motivating Example: Missing Economic Data

Social scientists have long debated the determinants of economic development across countries (Barro, 1997; Feng, 2003; Acemoglu et al., 2005). Using the data from the Central Intelligence Agency (CIA, 2016) and Freedom House (2016), we may estimate a multiple regression model, in which the dependent variable is GDP per capita and the independent variables include social, economic, and political variables. The problem is that the data are incomplete (Table 4.1), where the median missing rate is 22.4% and the total missing rate is 62.3%.

Table 4.2 presents multiple regression models; however, the conclusions are susceptible to how we deal with missing data. The coefficients for central bank and public debt are statistically significant at the 5% error level using incomplete data, while they are not significant using multiply-imputed data. On the other hand, the coefficients for education and military are not significant using incomplete data, while they are significant using multiply-imputed data. Therefore, the issue of missing data is of grave concern in applied empirical research.

Variables	Missing Rates
GDP per capita (purchasing power parity)	0.0%
Freedom House index	15.4%
Central bank discount rate	32.9%
Life expectancy at birth	2.6%
Unemployment rate	10.5%
Distribution of family income: Gini index	37.3%
Public debt	22.4%
Education expenditures	24.6%
Taxes and other revenues	6.1%
Military expenditures	43.0%

Table 4.1: Variables and Missing Rates
Data sources: CIA (2016) and Freedom House (2016)

	Incomplete Data		Multiply-Imputed Data	
Variables	Coef.	Std. Err.	Coef.	Std. Err.
Intercept	-7.323	3.953	-11.545 *	3.495
Freedom	-0.321 *	0.127	-0.362 *	0.127
Central Bank	**-0.118** *	**0.041**	-0.107	0.049
Life Expectancy	3.922 *	0.794	4.908 *	0.655
Unemployment	-0.205 *	0.087	-0.214 *	0.070
Gini	0.114	0.253	-0.018	0.363
Public Debt	**-0.198** *	**0.092**	-0.002	0.093
Education	0.035	0.164	**-0.488** *	**0.154**
Tax	0.357 *	0.174	0.471 *	0.151
Military	0.123	0.085	**0.299** *	**0.109**
Number of obs.	86		228	

Table 4.2: Multiple Regression Analyses on GDP Per Capita
Note: *significant at the 5% error level. Coef. stands for coefficient. Std. Err. stands for standard error. All of the variables are log-transformed.

4.4 Assumptions of Imputation Methods

Missing data analyses always involve assumptions (Raghunathan, 2016, p.12). In order to judge the appropriateness of missing data methods, it is vital to comprehend the assumptions for the methods. Imputation involves the following four assumptions. These assumptions will play important roles in simulation studies (Section 4.8).

4.4.1 Assumptions of Missing Data Mechanisms

There are three common assumptions of missing data mechanisms in the literature (King et al., 2001, pp.50-51; Little and Rubin, 2002; Carpenter and Kenward, 2013, pp.10-21). The first assumption is Missing Completely At Random (MCAR), which is $Pr(\mathbf{R}|\mathbf{D}) = Pr(\mathbf{R})$. If respondents are selected to answer their income values by throwing dice, this is an example of MCAR. The second assumption is Missing At Random (MAR), which is $Pr(\mathbf{R}|\mathbf{D}) = Pr(\mathbf{R}|\mathbf{Y_{obs}})$. If older respondents are more likely to refuse to answer their income values and if the ages of the respondents are available in the data, this is an example of MAR. The third assumption is Not Missing At Random (NMAR), which is $Pr(\mathbf{R}|\mathbf{D}) \neq Pr(\mathbf{R}|\mathbf{Y_{obs}})$. If respondents with higher values of incomes are more likely to refuse to answer their income values and if the other variables in the data cannot be used to predict which respondents have high amounts of income, this is an example of NMAR.

4.4.2 Assumption of Ignorability

To be strict, the missing data mechanism is ignorable if both of the following conditions are satisfied: (1) The MAR condition; and (2) the distinctness condition, which stipulates that the parameters in the missing data mechanism are independent of the parameters in the data model (Schafer, 1997, p.11).

However, the MAR condition is said to be more relevant in real data applications (Allison, 2002, p.5; van Buuren, 2012, p.33). Thus, for all practical purposes, NMAR is Non-Ignorable (NI). The current study assumes that the missing data mechanism is MAR and thus ignorable.

4.4.3 Assumption of Proper Imputation

Imputation is said to be Bayesianly proper if imputed values are independent realizations of

$Pr(\mathbf{Y_{mis}}|\mathbf{Y_{obs}})$, which means that successive iterates of $\mathbf{Y_{mis}}$ cannot be used because of the correlations between them (Schafer, 1997, pp.105-106). Between-imputation convergence relies on a number of factors, but the fractions of missing information are one of the most influential factors (Schafer, 1997, p.84; van Buuren, 2012, p.113).

van Buuren (2012, p.39) introduces a slightly simplified version of proper imputation, which he calls confidence proper. Let $\bar{\theta}$ be the multiple imputation estimate, $\hat{\theta}$ be the estimate based on the hypothetically complete data, $\bar{V}$ be the estimate of the sampling variance of the estimate based on the hypothetically complete data, and $\hat{V}$ be the sampling variance estimate based on the hypothetically complete data. An imputation procedure is said to be confidence proper if all of the following three conditions are satisfied: (1) $\bar{\theta}$ is equal to $\hat{\theta}$ when averaged over the response indicators sampled under the assumed response model; (2) $\bar{V}$ is equal to $\hat{V}$ when averaged over the response indicators sampled under the assumed response model; and (3) the extra inferential uncertainty due to missingness is correctly reflected. In order to check whether an imputation method is confidence proper, van Buuren (2012, p.47) recommends to use bias, coverage, and confidence interval length as the evaluation criteria (See Section 4.8.2).

4.4.4 Assumption of Congeniality

Congeniality means that the imputation model is equal to the substantive analysis model. It is widely known that the imputation model can be larger than the substantive analysis model, but the imputation model cannot be smaller than the substantive analysis model (Enders, 2010, pp.227-229; Carpenter and Kenward, 2013, pp.64-65; Raghunathan, 2016, pp.175-177).

4.5 Traditional Methods of Handling Missing Data

This section introduces listwise deletion, deterministic single imputation, and stochastic single imputation, which are used as baseline methods for comparisons in Section 8.

Listwise deletion (LD), also known as complete-case analysis, throws away any rows that have at least one missing value (Allison, 2002, pp.6-8; Baraldi and Enders, 2010, pp.10). Although it is simple and convenient, LD is less efficient due to the reduced sample size and may be biased if the assumption of MCAR does not hold (Schafer, 1997, p.23).

Deterministic single imputation (D-SI) replaces a missing value with a reasonable guess. The most straightforward version calculates predicted scores for missing values based on a regression model (Allison, 2002, p.11; Baraldi and Enders, 2010, p.12). If the goal of analysis is to estimate the mean of an incomplete variable, D-SI produces an unbiased estimate under the assumptions of MCAR and MAR. However, D-SI tends to underestimate the variation in imputed data (de Waal et al., 2011, p.231). D-SI is available as R-function `norm.predict` in MICE (van Buuren, 2012, p.57).

Stochastic single imputation (S-SI) also utilizes a regression model to predict missing values, but it adds to imputed values random components drawn from the residual distribution (Baraldi and Enders, 2010, p.13). S-SI is likely to recover the variation of an incomplete variable under the assumptions of MCAR and MAR; thus, compensating for the disadvantage of D-SI (de Waal et al., 2011, p.231). S-SI is available as R-function `norm.nob` in MICE (van Buuren, 2012, p.57).

However, both D-SI and S-SI tend to underestimate the standard error in imputed data because imputed values are treated as if they were real (Raghunathan, 2016, p.77).

4.6 Competing Multiple Imputation Algorithms

Multiple imputation was made widely known by Rubin (1987) and concise history can be found in Scheuren (2005). In theory, multiple imputation replaces a missing value by M simulated values ($M > 1$) independently and randomly drawn from the distribution of missing data. The variation among M simulated values reflects uncertainty about missing data; thus, making the standard error valid. In practice, missing data are by definition unobserved; therefore, the distribution of missing data is also unobserved. Instead, under the assumption of MAR (or MCAR), multiple imputation constructs the posterior predictive distribution of missing data, conditional on observed data. Then, a random draw is independently made from this posterior distribution (Rubin, 1987, p.75; King et al., 2001, pp.53-54; Carpenter and Kenward, 2013, pp.38-

39).

However, using the analytical methods, it is not easy to randomly draw sufficient statistics from the posterior distribution (Allison, 2002, pp.33; Honaker and King, 2010, pp.564). In order to solve this problem, three computational algorithms have been proposed in the literature.

4.6.1 Data Augmentation

The traditional algorithm of multiple imputation is the Data Augmentation (DA) algorithm, which is a Markov chain Monte Carlo (MCMC) technique (Takahashi and Ito, 2014, pp.46-48). DA improves parameter estimates by repeated substitution conditional on the preceding value, forming a stochastic process called a Markov chain (Gill, 2008, p.379).

The DA algorithm works as follows (Schafer, 1997, p.72). Equation (4.1) is the imputation step that generates imputed values from the predictive distribution of missing values, given the observed values and the parameter values at iteration t. Equation (4.2) is the posterior step that generates parameter values from the posterior distribution, given the observed values and the imputed values at iteration $t+1$.

$$\mathbf{Y}_{\mathbf{mis}}^{(\mathbf{t+1})} \sim Pr\left(\mathbf{Y}_{\mathbf{mis}} | \mathbf{Y}_{\mathbf{obs}}, \theta^{(t)}\right) \tag{4.1}$$

$$\theta^{(t+1)} \sim Pr\left(\theta | \mathbf{Y}_{\mathbf{obs}}, \mathbf{Y}_{\mathbf{mis}}^{(\mathbf{t+1})}\right) \tag{4.2}$$

These two steps are repeated T times until convergence is attained. The convergence of MCMC is stochastic because it converges to probability distributions (Schafer, 1997, p.80). Therefore, it is hard to judge the convergence in MCMC.

There are two ways of generating multiple imputations by DA (Schafer, 1997, p.139; Enders, 2010, pp.211-212). In the first method, a single chain is run for $M \times T$ iterations, taking every t-th iteration of Y_{mis}. In the second method, M parallel chains of length T are run, and the final values of Y_{mis} from M chains are taken as the imputations. The current study adopts the second method.

The software using this algorithm is R-Package NORM2, which was originally developed by Schafer (1997) and is currently maintained by Schafer (2016).

4.6.2 Fully Conditional Specification

An alternative algorithm to DA is the Fully Conditional Specification (FCS) algorithm, which specifies the multivariate distribution by way of a series of conditional densities, through which missing values are imputed given the other variables (Takahashi and Ito, 2014, pp.50-53).

The FCS algorithm works as follows (van Buuren and Groothuis-Oudshoorn, 2011, pp.6-7; van Buuren, 2012, p.110; Zhu and Raghunathan, 2015). Equation (4.3) draws the unknown parameters of the imputation model, given the observed values and the t-th imputations, where $\widetilde{\mathbf{Y}}_{\mathbf{-j}}^{(\mathbf{t})} = \left(\widetilde{\mathbf{Y}}_{\mathbf{1}}^{(\mathbf{t})}, \dots, \widetilde{\mathbf{Y}}_{\mathbf{j-1}}^{(\mathbf{t})}, \widetilde{\mathbf{Y}}_{\mathbf{j+1}}^{(\mathbf{t-1})}, \dots, \widetilde{\mathbf{Y}}_{\mathbf{p}}^{(\mathbf{t-1})}\right)$, where tilde denotes a random draw. Equation (4.4) draws imputations, given the observed values, the t-th imputations, and the t-th parameter estimates. These two steps are repeated for $j = 1, \dots, p$.

$$\tilde{\theta}_j^{(t)} \sim Pr\left(\theta_j^{(t)} | \mathbf{Y}_{\mathbf{j,obs}}, \widetilde{\mathbf{Y}}_{\mathbf{-j}}^{(\mathbf{t})}\right) \tag{4.3}$$

$$\widetilde{\mathbf{Y}}_{\mathbf{j}}^{(\mathbf{t})} \sim Pr\left(\mathbf{Y}_{\mathbf{j,mis}} | \mathbf{Y}_{\mathbf{j,obs}}, \widetilde{\mathbf{Y}}_{\mathbf{-j}}^{(\mathbf{t})}, \tilde{\theta}_j^{(t)}\right) \tag{4.4}$$

The entire process is repeated for $t = 1, \dots, T$ until convergence is attained. FCS can be considered an MCMC method, because FCS is a Gibbs sampler under the compatible conditionals (van Buuren and Groothuis-Oudshoorn, 2011, p.6; van Buuren, 2012, p.109). This means that the convergence of FCS is stochastic. Therefore, it is hard to judge the convergence in FCS.

The software using this algorithm is R-Package MICE (van Buuren and Groothuis-Oudshoorn, 2011), which stands for Multivariate Imputation by Chained Equations and is currently maintained by van Buuren et al. (2015). The FCS algorithm is also known as Sequential Regression Multivariate Imputation (Raghunathan, 2016, p.76).

4.6.3 Expectation-Maximization with Bootstrapping

Another emerging algorithm is the Expectation-Maximization with Bootstrapping (EMB) algorithm, which combines the Expectation-Maximization (EM) algorithm with the nonparametric bootstrap to create multiple imputation (Takahashi and Ito, 2014, pp.55-57).

The EMB algorithm works as follows (Honaker and King, 2010, p.565; Honaker et al., 2011, p.4). Suppose that a random sample of size n is drawn from a population, where some values are missing in the sample. Bootstrap resamples of size n are randomly drawn from the sample data with replacement M times (Horowitz, 2001, pp.3163-3165; Carsey and Harden, 2014, p.215). The variation among the M resamples represents uncertainty about estimation. The EM algorithm is applied to each of these M bootstrap resamples to refine M point estimates of parameter θ. Equation (4.5) is the expectation step that calculates the Q-function by averaging the complete-data log-likelihood over the predictive distribution of missing data. Equation (4.6) is the maximization step that finds parameter values at iteration $t+1$ by maximizing the Q-function.

$$Q\left(\theta|\theta^{(t)}\right) = \int l(\theta|\mathbf{Y})\, Pr\left(\mathbf{Y_{mis}}|\mathbf{Y_{obs}}, \theta^{(t)}\right) d\mathbf{Y_{mis}} \tag{4.5}$$

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta|\theta^{(t)}\right) \tag{4.6}$$

These two steps are repeated until convergence is attained, where the converged value is a Maximum Likelihood Estimate (MLE) under well-behaved conditions (Schafer, 1997, pp.38-39; Do and Batzoglou, 2008). The convergence of EM is deterministic because it converges to a point in the parameter space (Schafer, 1997, p.80). Therefore, it is straightforward to judge the convergence in EM. The substitution of MLEs from bootstrap resamples is asymptotically equal to a sample from the posterior distribution (Little and Rubin, 2002, pp.216-217).

The software using this algorithm is R-Package AMELIA II (Honaker et al., 2011), which was originally developed by King et al. (2001) and is currently maintained by Honaker et al. (2016).

4.6.4 Relationships among the Three Algorithms

The three algorithms share certain characteristics with each other, but not exactly the same as summarized in Table 4.3.

	Joint Modeling	Conditional Modeling
MCMC	DA	FCS
Non-MCMC	EMB	

Table 4.3: Relations among DA, EMB, and FCS

DA and EMB are joint modeling while FCS is conditional modeling (Kropko et al., 2014). Joint modeling specifies a multivariate distribution of missing data while conditional modeling specifies a univariate distribution on a variable-by-variable basis (van Buuren, 2012, pp.105-108). Conditional modeling is more flexible and joint modeling is computationally more efficient (van Buuren, 2012, p.117; Kropko et al., 2014).

DA and FCS are different versions of MCMC techniques. On the other hand, EMB is not an MCMC technique. It is said that DA and FCS require between-imputation iterations to be confidence proper (Schafer, 1997, p.106; van Buuren, 2012, p.113) while EMB does not need iterations to be confidence proper (Honaker and King, 2010, p.565). However, as is clear in Section 4.7, whether EMB is confidence proper when DA and FCS are improper, this is an open question that has not been tested in the literature.

4.7 Comparative Studies on Multiple Imputation in the Literature

Table 4.4 presents the literature that compared imputation methods. Nine studies compared multiple imputation with other missing data methods, such as listwise deletion, single imputation, and maximum likelihood. Among these nine studies, four studies focused on DA (Schafer and Graham, 2002; Abe and Iwasaki, 2007; Lee and Carlin, 2012; von Hippel, 2016), four studies on FCS (Donders et al., 2006; Stuart et al., 2009; Cheema, 2014; Deng et al., 2016), and one study on an unknown algorithm (Shara et al., 2015).

Four studies investigated specialized situations for multiple imputation, such as small-sample degrees of freedom in DA (Barnard and Rubin, 1999), Likert-type data in DA (Leite and Beretvas, 2010), non-parametric multiple imputation (Cranmer and Gill, 2013), and variance estimators (Hughes et al., 2016).

Authors	MI Algorithms	Sample Size	Number of Variables	Number of Imputations	Number of Iterations	Missing Rate
Barnard and Rubin (1999)	DA	10, 20, 30	2	3, 5, 10	Unknown	10%, 20%, 30%
Horton and Lipsitz (2001)	DA, FCS	10000	3	10	200	50%
King et al. (2001)	DA, EMis	500	5	10	1000	17%, 22%, 50%
Schafer and Graham (2002)	DA	50	2	20	Unknown	73%
Donders et al. (2006)	FCS	500	2	10	Unknown	40%
Abe and Iwasaki (2007)	DA	100	4	5	100	20%, 30%
Horton and Kleinman (2007)	**DA, EMB, FCS**	133774	10	10	**5**	41%
Stuart et al. (2009)	FCS	9186	400	10	10	18%
Lee and Carlin (2010)	DA, FCS	1000	8	20	10	33%
Leite and Beretvas (2010)	DA	400	10	10	Unknown	10%, 30%, 50%
Hardt, Herke, and Leonhart (2012)	**DA, EMB, FCS**	50, 100, 200	3, 13, 23, 43, 83	20	**Unknown**	20%, 50%
Lee and Carlin (2012)	DA	1000	8	20	Unknown	10%, 25%, 50%, 75%, 90%
Cranmer and Gill (2013)	EMB, MHD	500	5	Unknown	NA	20%, 50%, 80%
Cheema (2014)	FCS	10, 20, 50, 100, 200, 500, 1000, 2000, 5000, 10000	4	Unknown	Unknown	1%, 2%, 5%, 10%, 20%
Kropko et al. (2014)	**DA, EMB, FCS**	1000	8	5	**30**	25%
Shara et al. (2015)	Unknown	2246	8	Unknown	Unknown	20%, 30%, 40%
Deng et al. (2016)	FCS	100	200, 1000	10	20	40%
von Hippel (2016)	DA	25, 100	2	5	Unknown	50%
Hughes, Sterne, and Tilling (2016)	Unknown	100, 1000	5	50	Unknown	40%, 60%
McNeish (2017)	DA, FCS	20, 50, 100, 250	4	5, 25, 100	Unknown	10%, 20%, 30%, 50%

Table 4.4: Summary of the 20 Studies on Multiple Imputation

Note: DA stands for Data Augmentation, EMis for Expectation-Maximization with Importance Sampling, FCS for Fully Conditional Specification, EMB for Expectation-Maximization with Bootstrapping, and MHD for Multiple Hot Deck. Unknown means that information is unavailable. NA means Not-Applicable.

Seven studies compared different multiple imputation algorithms (King et al., 2001; Horton and Lipsitz, 2002; Horton and Kleinman, 2007; Lee and Carlin, 2010; Hardt et al., 2012; Kropko et al., 2014; McNeish, 2017). The comparative perspective in most of the seven studies, except King et al. (2001), is based on the difference between joint modeling and conditional modeling. Thus, the perspective from MCMC vs. non-MCMC is generally lacking in the literature.

Ten studies did not explicitly state the number of iterations T. Furthermore, Horton and Kleinman (2007) used the default setting in software for T, and the information in Kropko et al. (2014) can be only found in their computer codes, not in the article.

Thus, no studies in Table 4.4 have systematically investigated the effects of convergence on the three multiple imputation algorithms.

4.8 Monte Carlo Simulation

Section 4 introduced MAR, proper imputation, and congeniality as crucial assumptions. To make the assumptions of MAR and congeniality realistic, an inclusive analysis strategy is recommended in the literature (Enders, 2010, pp.16-17; Raghunathan, 2016, p.73), which contains any auxiliary variables that can increase the predictive power of the imputation model or any variables that may be related to the missing data mechanism. What complicates the matter, however, is that auxiliary variables themselves are often incomplete. This creates a dilemma in multiple imputation. Including many auxiliary variables makes it more likely for MAR and congeniality to be satisfied, but including many incomplete variables leads to a higher total missing rate, which further makes it more difficult for convergence in MCMC to be attained.

When assumptions do not hold in statistical methods, analytical mathematics does not often provide answers about the properties of the methods (Mooney, 1997, p.1). Monte Carlo simulation converts the computer into an experimental laboratory, where the researcher can control various conditions in the environment to observe the outcomes (Carsey and Harden, 2014, p.4). Thus, Monte Carlo simulation is a powerful method of assessing the performance of statistical methods under various settings especially when assumptions are violated.

Abbreviations in this section are explained in Table 4.5, where MI stands for multiple imputation and SI for single imputation.

Abbreviations	Missing Data Methods
CD	Complete data without missing values
LD	Listwise deletion
EMB	MI by AMELIA II
DA1	MI by NORM2 with no iterations
DA2	MI by NORM2 with 2*EM iterations
FCS1	MI by MICE with no iterations
FCS2	MI by MICE with 2*EM iterations
D-SI	Deterministic SI by `norm.predict` in MICE
S-SI	Stochastic SI by `norm.nob` in MICE

Table 4.5: Abbreviations and the Missing Data Methods

4.8.1 Monte Carlo Simulation Designs

The current study prepares two versions of simulation data, (1) theoretical and (2) realistic. Auxiliary variables $\mathbf{X}$ are generated by R-Function `mvrnorm`. All of the computations are done in R 3.2.4. The computer used in the current study is HP Z440 Workstation (Windows 7 Professional, processor: Intel Xeon CPU E5-1603 v3), with the processor speed of 2.80 GHz and the memory (RAM) of 32.0 GB under the 64 bit operating system. The number of Monte Carlo simulation runs is set to 1000.

The first setting is theoretical. The number of observations is 1000, which is the 75th percentile in Table 4.4. The number of variables p is changed from 2, 3, 4, 5, 6, 7, 8, 9, to 10, which covers the 70th percentile in Table 4.4. Note that in another simulation run, not reported here, p was changed to 20, and the conclusions were similar. Auxiliary variables x_j are multivariate-normal with the mean of 0 and the standard deviation of 1, i.e., $\mathbf{X} \sim N_{p-1}(0,1)$, where the number of auxiliary variables is $p-1$. The correlation among x_j is randomly generated in R as follows: `r<-matrix(runif(9^2,-1,1),ncol=9)` and `Cor<-cov2cor(r%*%t(r))`. The generated correlation matrix is shown in equation (4.7). The p-th variable y_i is a linear combination of x_j such that $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1i} + \varepsilon_i$, where $\beta_j \sim U(-2.0, 2.0)$ and $\varepsilon_i \sim N(0, \sigma)$. Note that β_j includes β_0 and $\sigma \sim U(0.5, 2.0)$.

$$Cor_1 = \begin{bmatrix} 1.000 & -0.231 & 0.335 & 0.401 & -0.276 & 0.247 & -0.120 & 0.327 & -0.068 \\ -0.231 & 1.000 & 0.074 & -0.761 & 0.041 & -0.623 & -0.083 & -0.432 & -0.183 \\ 0.335 & 0.074 & 1.000 & 0.183 & -0.323 & 0.254 & -0.458 & 0.434 & -0.801 \\ 0.401 & -0.761 & 0.183 & 1.000 & 0.007 & 0.639 & -0.094 & 0.676 & 0.169 \\ -0.276 & 0.041 & -0.323 & 0.007 & 1.000 & -0.547 & 0.357 & -0.025 & 0.081 \\ 0.247 & -0.623 & 0.254 & 0.639 & -0.547 & 1.000 & 0.024 & 0.204 & 0.023 \\ -0.120 & -0.083 & -0.458 & -0.094 & 0.357 & 0.024 & 1.000 & -0.486 & 0.373 \\ 0.327 & -0.432 & 0.434 & 0.676 & -0.025 & 0.204 & -0.486 & 1.000 & -0.153 \\ -0.068 & -0.183 & -0.801 & 0.169 & 0.081 & 0.023 & 0.373 & -0.153 & 1.000 \end{bmatrix} \quad (4.7)$$

The second setting is realistic. The number of observations is 228, which is the full sample size of the real data in Table 4.2. The number of variables p is again changed from 2, 3, 4, 5, 6, 7, 8, 9, to 10. Auxiliary variables x_j are multivariate-normal with the means and standard deviations based on the empirical data (log-transformed), where x_j consist of the nine independent variables in Table 4.2 (CIA, 2016; Freedom House, 2016). Furthermore, the correlation matrix is based on the empirical data (log-transformed) as in equation (4.8). The p-th variable y_i is a linear combination of x_j such that $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1i} + \varepsilon_i$, where β_j (including β_0) reflects the coefficients in multiple regression models using the empirical data and $\varepsilon_i \sim N(0, \sigma_{resid})$, where σ_{resid} is the residual standard deviation from the empirical regression model.

$$Cor_2 = \begin{bmatrix} 1.000 & 0.646 & -0.500 & -0.007 & 0.376 & -0.354 & -0.378 & -0.534 & 0.312 \\ 0.646 & 1.000 & -0.531 & 0.021 & 0.371 & -0.305 & -0.150 & -0.427 & 0.049 \\ -0.500 & -0.531 & 1.000 & -0.474 & -0.512 & 0.278 & 0.092 & 0.280 & -0.086 \\ -0.007 & 0.021 & -0.474 & 1.000 & 0.205 & 0.079 & 0.014 & 0.086 & 0.161 \\ 0.376 & 0.371 & -0.512 & 0.205 & 1.000 & -0.204 & -0.089 & -0.370 & 0.220 \\ -0.354 & -0.305 & 0.278 & 0.079 & -0.204 & 1.000 & 0.106 & 0.212 & -0.180 \\ -0.378 & -0.150 & 0.092 & 0.014 & -0.089 & 0.106 & 1.000 & 0.578 & -0.128 \\ -0.534 & -0.427 & 0.280 & 0.086 & -0.370 & 0.212 & 0.578 & 1.000 & -0.134 \\ 0.312 & 0.049 & -0.086 & 0.161 & 0.220 & -0.180 & -0.128 & -0.134 & 1.000 \end{bmatrix} \quad (4.8)$$

In both settings, x_j are incomplete variables for imputation, y_i is completely observed in all of the situations, and u_{ji} are a set of $p - 1$ continuous uniform random numbers ranging from 0 to 1 for the missing data mechanism. As was introduced in Section 4.4.1, under the assumption of MAR, the missingness of x_{ji} depends on the values of y_i and u_{ji}, i.e., x_{ji} is missing if $y_i < \text{median}(y_i)$ and $u_{ji} < 0.5$, and x_{ji} is missing if $y_i > \text{median}(y_i)$ and $u_{ji} > 0.9$. This creates approximately 30% missing values in each x_j. This is realistic, because the average missing rates of income and earnings are 30% on a variable basis in the National Health Interview

Survey (Schenker et al., 2006, p.925) and the median missing rate is 30.0% in Table 4.4. Note that the above setting may be translated into the following statement. Variable y_i is age and x_{1i} is income. The missingness of income depends on age and some random components. Income is missing if age is less than the median of age and uniform random numbers are less than 0.5. Also, income is missing if age is larger than the median of age and uniform random numbers are larger than 0.9.

Although the literature (Graham et al., 2007; Bodner, 2008; Takahashi and Ito, 2014, pp.68-71) recommends to use relatively large M, the simulation studies in Table 4.4 use relatively small M. This is due to the computational burden of Monte Carlo simulation for multiple imputation. Considering this practical issue, the current study sets M to 20, which is the 75th percentile in Table 4.4.

As for T, there is no consensus in the literature (Table 4.4). There are no clear-cut rules for determining whether MCMC algorithms attained convergence (Schafer, 1997, p.119; King et al., 2001, p.59; van Buuren and Groothuis-Oudshoorn, 2011, p.37). Though not perfect, doubling the number of EM iterations is a rule of thumb for a conservative estimate about convergence speed for MCMC (Schafer and Olsen, 1998; Enders, 2010, p.204). Since it is not possible to check convergence in each of the 1000 simulation runs, the current study relies on the rule of thumb to set T.

4.8.2 Criteria for Judging Simulation Results

The estimand in all of the simulation runs is β_1 in $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1i} + \varepsilon_i$. The purpose of multiple imputation is to find an unbiased estimate of the population parameter that is confidence valid (van Buuren, 2012, pp.35-36).

Unbiasedness can be assessed by equation (4.9), because an estimator $\hat{\theta}$ is an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to the true θ (Mooney, 1997, p.59; Gujarati, 2003, pp.899).

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{4.9}$$

Unbiasedness and efficiency can be simultaneously assessed by the Root Mean Square Error (RMSE), defined as equation (4.10). The RMSE measures the spread around the true value of the parameter, placing slightly more emphasis on efficiency than bias (Gujarati, 2003, p.901; Carsey and Harden, 2014, pp.88-89).

$$\text{RMSE}(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2} \tag{4.10}$$

Confidence validity can be assessed by the coverage probability of the nominal 95% confidence interval (CI), which 'is the proportion of simulated samples for which the estimated confidence interval includes the true parameter' (Carsey and Harden, 2014, p.93). The formula of the standard error for proportions is equation (4.11), where π is the proportion and s is the number of simulation runs.

$$\text{SE}(\pi) = \sqrt{\frac{\pi(1-\pi)}{s}} \tag{4.11}$$

The standard error of the 95% CI coverage over 1000 iterations is $\sqrt{0.95 \times 0.05/1000} \approx 0.007$ which is 0.7%. Therefore, with 95% confidence, the estimated coverage probability should be between 93.6% and 96.4% (Abe and Iwasaki, 2007, p.10; Lee and Carlin, 2010, p.627; Carsey and Harden, 2014, pp.94-95; Hughes et al., 2016).

4.8.3 Results of the Simulation: Theoretical Case

This section presents the results of the Monte Carlo simulation for the theoretical case, where the correlation matrix and the regression coefficients are randomly generated.

Table 4.6 shows the Bias and RMSE values for the regression coefficient β_1. The Bias and RMSE values for listwise deletion and single imputation methods indicate that these methods are not recommended at all. All of the Bias and RMSE values from EMB, DA1, DA2, and FCS2 are almost identical, showing that they are generally unbiased. However, FCS1 is rather biased, quite similar to S-SI. Therefore, when between-imputation iterations are ignored, there are no discernible effects on bias and efficiency in EMB and DA, but FCS may suffer from some bias.

		Number of Variables								
		2	3	4	5	6	7	8	9	10
CD	Bias	0.001	0.003	0.001	0.002	0.001	0.001	0.001	0.002	0.001
	RMSE	0.040	0.047	0.038	0.039	0.058	0.026	0.046	0.039	0.047
LD	Bias	**0.032**	**0.135**	**0.105**	**0.104**	**0.332**	**0.085**	**0.129**	**0.210**	**0.116**
	RMSE	0.059	0.153	0.122	0.121	0.349	0.103	0.160	0.228	0.155
EMB	Bias	0.000	0.004	0.002	0.000	0.005	0.001	0.005	0.005	0.002
	RMSE	0.046	0.053	0.050	0.051	0.075	0.041	0.069	0.059	0.072
DA1	Bias	0.001	0.002	0.003	0.001	0.001	0.000	0.003	0.003	0.002
	RMSE	0.046	0.053	0.050	0.051	0.074	0.041	0.069	0.058	0.072
DA2	Bias	0.002	0.001	0.005	0.002	0.001	0.000	0.001	0.003	0.000
	RMSE	0.046	0.053	0.050	0.051	0.074	0.041	0.069	0.058	0.072
FCS1	Bias	0.002	0.001	**0.082**	**0.040**	**0.090**	**0.047**	**0.093**	**0.027**	**0.233**
	RMSE	0.047	0.053	0.097	0.062	0.116	0.065	0.109	0.052	0.239
FCS2	Bias	0.001	0.002	0.004	0.002	0.001	0.000	0.001	0.002	0.001
	RMSE	0.046	0.053	0.050	0.051	0.075	0.041	0.069	0.058	0.071
D-SI	Bias	**0.186**	**0.242**	**0.174**	**0.093**	**0.187**	**0.098**	**0.231**	**0.070**	**0.163**
	RMSE	0.192	0.248	0.182	0.110	0.207	0.109	0.248	0.099	0.189
S-SI	Bias	0.002	0.000	**0.081**	**0.038**	**0.090**	**0.047**	**0.091**	**0.029**	**0.230**
	RMSE	0.050	0.057	0.102	0.066	0.124	0.076	0.119	0.062	0.241

Table 4.6: Bias and RMSE (Theoretical Data)
Note: Biased results are in boldface, i.e., Bias > 0.010.

Table 4.7 gives the coverage probability of the 95% CI for β_1. The CIs for listwise deletion and single imputation methods are not confidence valid. When the number of auxiliary variables is small (and hence the overall missing rate is small), the between-imputation iterations may be ignored, where all of the multiple imputation CIs are confidence valid. However, as the number of auxiliary variables becomes large, DA1 and FCS1 drift away from the confidence validity. EMB, DA2, and FCS2 are confidence valid regardless of the number of variables and the missing rate. This shows that EMB is confidence proper even if it does not iterate. This is an important finding in the current study.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
CD	95.3	94.9	94.2	94.0	96.0	96.0	95.3	94.9	94.6
LD	**88.5**	**47.9**	**54.6**	**56.7**	**10.8**	**65.1**	**69.2**	**32.1**	**78.1**
EMB	95.0	95.1	94.2	95.5	94.9	94.4	94.3	94.1	95.0
DA1	94.6	94.9	**93.2**	**93.1**	94.1	**91.8**	**92.9**	**92.4**	**92.9**
DA2	94.3	95.8	95.1	94.1	94.8	94.3	94.2	**93.2**	94.9
FCS1	94.2	95.0	**75.0**	**91.6**	**84.4**	95.5	**84.5**	**96.8**	**6.8**
FCS2	94.7	95.6	94.4	93.9	95.4	94.5	94.2	95.0	95.0
D-SI	**0.8**	**0.2**	**2.2**	**37.8**	**22.2**	**16.9**	**8.3**	**51.0**	**22.5**
S-SI	**88.9**	**89.6**	**47.8**	**75.0**	**62.3**	**64.4**	**48.9**	**76.0**	**3.7**

Table 4.7: Coverage of the 95% CI (Theoretical Data)
Note: Confidence invalid results are in boldface, i.e., outside of 93.6 and 96.4.

Table 4.8 shows the CI lengths. The CI length by listwise deletion is generally too long, reflecting inefficiency due to the reduced sample size. The CI lengths by single imputation methods are ‘correct’ in the sense that they are quite similar to those of complete data analysis; however, this means that single imputation methods ignore estimation uncertainty associated with imputation. This is the cause of confidence invalidity of single imputation methods in Table 4.7. The CI length by DA1 is too short and the CI length by FCS1 is too long. The CI lengths by EMB, DA2, and FCS2 are essentially equal, reflecting the correct level of estimation uncertainty associated with imputation.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
CD	0.157	0.184	0.144	0.148	0.236	0.102	0.184	0.151	0.180
LD	0.189	0.259	0.226	0.235	0.384	0.213	0.358	0.339	0.390
EMB	0.178	0.209	0.196	0.200	0.301	0.160	0.275	0.229	0.281
DA1	0.176	0.207	0.187	0.192	0.293	0.145	0.256	0.208	0.253
DA2	0.177	0.208	0.194	0.198	0.298	0.158	0.271	0.223	0.274
FCS1	0.178	0.209	0.237	0.211	0.324	0.248	0.306	0.223	0.299
FCS2	0.178	0.209	0.197	0.201	0.302	0.161	0.275	0.228	0.281
D-SI	0.143	0.174	0.133	0.149	0.244	0.103	0.205	0.150	0.188
S-SI	0.157	0.184	0.161	0.155	0.238	0.145	0.188	0.149	0.186

Table 4.8: Lengths of the 95% CI (Theoretical Data)

Table 4.9 displays the computational time required to generate multiple imputations. When the number of auxiliary variables is small (and hence the overall missing rate is small), DA2 is fastest among the three confidence proper multiple imputation algorithms. On the other hand, as the number of auxiliary variables becomes large, EMB becomes fastest. As is known in the literature (van Buuren, 2012, p.117; Kropko et al., 2014), FCS2 is at least 5 times slower and can be more than 50 times slower than EMB and DA2. However, the difference in computational time is not substantial, given that all of the computations can be done within a few minutes.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
EMB	0.46	0.53	0.53	0.59	0.71	**0.78**	**0.97**	**1.27**	**1.69**
DA2	**0.10**	**0.16**	**0.29**	**0.42**	**0.55**	1.09	1.39	2.22	3.63
FCS2	2.47	5.98	14.48	21.33	25.40	54.71	59.14	85.69	133.17

Table 4.9**:** Computational Time (Theoretical Data)

Note: Reported values are the time in seconds to perform multiple imputation, which is averaged over 1,000 simulation runs. The fastest results are in boldface.

4.8.4 Results of the Simulation: Realistic Case

This section presents the results of the Monte Carlo simulation for the realistic case, where the correlation matrix and the regression coefficients are based on the real data (CIA, 2016; Freedom House, 2016). The results in this section reinforce the findings in Section 8.3.1.

Table 4.10 shows the Bias and RMSE values for the regression coefficient β_1. The overall conclusions are similar to Table 4.6. When between-imputation iterations are ignored, there are no discernible effects on bias and efficiency in EMB and DA, but FCS may occasionally suffer from small bias.

Table 4.11 gives the coverage probability of the 95% CI for β_1. The overall conclusions are similar to Table 4.7, except that DA1 is confidence invalid even when $p = 3$. This implies that we cannot ignore between-imputation iterations in MCMC-based approaches even when the number of variables is small. On the other hand, EMB is confidence valid and we can safely ignore between-imputation iterations in EMB. Again, this is an important finding in the current study.

		Number of Variables								
		2	3	4	5	6	7	8	9	10
CD	Bias	0.003	0.002	0.002	0.002	0.001	0.002	0.000	0.002	0.002
	RMSE	0.074	0.086	0.068	0.067	0.066	0.065	0.070	0.069	0.075
LD	Bias	**0.034**	**0.047**	**0.037**	**0.054**	**0.082**	**0.099**	**0.083**	**0.072**	**0.085**
	RMSE	0.095	0.128	0.104	0.118	0.141	0.154	0.157	0.159	0.188
EMB	Bias	0.001	0.002	0.002	0.005	0.001	0.000	0.000	0.002	0.006
	RMSE	0.084	0.113	0.091	0.090	0.089	0.092	0.102	0.099	0.110
DA1	Bias	0.006	0.001	0.003	0.003	0.001	0.001	0.001	0.001	0.002
	RMSE	0.084	0.112	0.090	0.089	0.087	0.091	0.100	0.096	0.105
DA2	Bias	0.009	0.000	0.002	0.004	0.002	0.004	0.000	0.001	0.001
	RMSE	0.084	0.111	0.089	0.088	0.086	0.090	0.098	0.094	0.102
FCS1	Bias	0.007	**0.013**	0.006	0.005	0.002	0.008	0.006	**0.012**	0.000
	RMSE	0.084	0.106	0.081	0.081	0.080	0.081	0.086	0.083	0.088
FCS2	Bias	0.007	0.001	0.002	0.002	0.003	0.005	0.002	0.003	0.005
	RMSE	0.084	0.112	0.088	0.088	0.086	0.090	0.097	0.093	0.100
D-SI	Bias	**0.188**	**0.075**	**0.011**	**0.035**	**0.037**	**0.047**	**0.023**	**0.034**	**0.059**
	RMSE	0.207	0.163	0.115	0.118	0.118	0.123	0.130	0.127	0.151
S-SI	Bias	0.005	**0.014**	0.007	0.006	0.002	0.006	0.005	0.009	0.006
	RMSE	0.089	0.116	0.096	0.095	0.091	0.094	0.100	0.102	0.105

Table 4.10: Bias and RMSE (Realistic Data)
Note: Biased results are in boldface, i.e., Bias > 0.010.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
CD	94.6	95.3	95.8	94.7	95.2	96.4	94.6	95.3	94.8
LD	**92.2**	**91.6**	**92.8**	**91.5**	**86.8**	**85.0**	**89.8**	**90.0**	**90.8**
EMB	94.3	94.1	94.7	93.9	96.1	94.2	94.0	94.4	94.7
DA1	94.1	**92.2**	94.4	**93.4**	95.7	**92.2**	**93.1**	**92.9**	**93.1**
DA2	94.0	94.0	94.8	94.4	95.9	94.5	93.8	95.0	95.0
FCS1	94.6	94.7	96.3	**96.7**	**97.0**	**97.0**	**96.7**	**96.9**	**97.7**
FCS2	94.7	93.8	95.5	95.7	96.4	94.3	94.8	95.2	96.1
D-SI	**32.7**	**74.5**	**79.2**	**77.6**	**77.7**	**74.1**	**75.3**	**75.1**	**68.8**
S-SI	**87.9**	**83.2**	**82.3**	**82.5**	**84.2**	**82.1**	**81.0**	**80.3**	**81.2**

Table 4.11: Coverage of the 95% CI (Realistic Data)
Note: Confidence invalid results are in boldface, i.e., outside of 93.6 and 96.4.

Table 4.12 shows the CI lengths. The overall conclusions are similar to Table 4.8. One difference is that the CI length by FCS1 is slightly short.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
CD	0.279	0.334	0.268	0.266	0.267	0.261	0.278	0.274	0.289
LD	0.333	0.441	0.389	0.412	0.436	0.457	0.516	0.543	0.631
EMB	0.314	0.429	0.364	0.356	0.362	0.359	0.397	0.396	0.432
DA1	0.313	0.414	0.348	0.342	0.343	0.337	0.370	0.364	0.390
DA2	0.315	0.423	0.356	0.351	0.353	0.351	0.383	0.380	0.410
FCS1	0.315	0.416	0.353	0.348	0.350	0.350	0.382	0.380	0.406
FCS2	0.316	0.429	0.359	0.355	0.358	0.352	0.389	0.386	0.413
D-SI	0.288	0.380	0.292	0.289	0.291	0.278	0.302	0.294	0.315
S-SI	0.281	0.325	0.262	0.257	0.259	0.255	0.269	0.267	0.277

Table 4.12: Lengths of the 95% CI (Realistic Data)

Table 4.13 displays the computational time required to generate multiple imputations. The overall conclusions are similar to Table 4.9.

	Number of Variables								
	2	3	4	5	6	7	8	9	10
EMB	0.14	0.15	0.16	0.20	0.23	0.28	0.36	**0.44**	**0.53**
DA2	**0.04**	**0.05**	**0.06**	**0.10**	**0.15**	**0.22**	**0.33**	0.47	0.67
FCS2	1.05	2.55	4.22	8.92	12.02	15.59	20.82	26.78	35.95

Table 4.13: Computational Time (Realistic Data)
Note: Reported values are the time in seconds to perform multiple imputation, which is averaged over 1,000 simulation runs. The fastest results are in boldface.

4.9 Conclusions

This chapter assessed the relative performance of the three multiple imputation algorithms (DA, FCS, and EMB). In both theoretical and realistic settings (Table 4.7 and Table 4.11), if between-imputation iterations were ignored, the MCMC algorithms (DA and FCS) did not attain confidence validity. The nominal 95% CIs by DA and FCS without iterations were different from 95% coverage beyond the margin of error in 1,000 simulation runs. This is because the CI lengths by DA without iterations were generally too short, and the CI lengths by FCS are generally too long (Table 4.8 and Table 4.12). Based on Schafer (1997, p.139), this can be explained by choices for starting values. DA uses EM as a single starting value for M chains that understates missing data uncertainty (Schafer, 2016, p.22) while FCS uses random draws as M over-dispersed starting values that overstates missing data uncertainty (van Buuren and Groothuis-Oudshoorn, 2011, p.6). Without iterations, imputed values depend on the choice of starting values.

DA and FCS can be both confidence valid under the large number of iterations; however, the assessment of convergence in MCMC is notoriously difficult. Furthermore, the convergence properties of FCS are currently under debate due to possible incompatibility (Li et al., 2012; Zhu and Raghunathan, 2015). On the other hand, the current study found that EMB was confidence valid regardless of the situations. Therefore, EMB is a confidence proper imputation algorithm without iterations, which allows us to avoid a painful decision-making process of how to judge the convergence to generate confidence proper multiple imputations. This finding is useful in the missing data literature. For example, while ratio imputation is often used in official statistics (Takahashi et al., 2017), multiple ratio imputation does not exist in the literature. The EMB algorithm was applied to ratio imputation to create multiple ratio imputation (Takahashi, 2017b; Takahashi, 2017c).

No simulation studies can include all the patterns of relevant data (Kropko et al., 2014, p.511). Therefore, the current study focused on two types of data, (1) theoretical and (2) realistic. Although the author believes that the two data generation processes cover data types relevant to many social research situations, the results in any simulation studies must be read with caution (Hardt et al., 2014, p.11). Future research should delve into other data types, such as small-n data, large-p data, categorical data, to name a few.

5 Multiple Ratio Imputation by the EMB Algorithm: Theory and Simulation

This chapter derived from Takahashi (2017b), a peer-reviewed article in the *Journal of Modern Applied Statistical Methods* 16(1), which is operated by the Wayne State University Library System, classified as one of the top 115 libraries in the United States by the Association for Research Libraries (Kyrillidou et al., 2015). The *Journal of Modern Applied Statistical Methods* is indexed in Scopus by Elsevier as of April 2017. The author would like to thank JMASM Inc. for permission to use "Multiple ratio imputation by the EMB algorithm: Theory and simulation" (*Journal of Modern Applied Statistical Methods*, vol.16, no.1, 630-656).

5.1 Introduction

In survey data, missing values are prevalent. At best, missing data are inefficient because the incomplete dataset does not contain as much information as is expected. At worst, missing data can be biased if non-respondents are systematically different from respondents (Rubin, 1987). The best solution to the missing data problem is to collect the true data, by resending questionnaires or by calling respondents. Nevertheless, there are two problems to this ideal solution. First, data users often have no luxury of collecting more data to take care of missingness. Second, facing a world-wide trend of resource reduction in official statistics, data providers such as national statistical agencies need to make the statistical production as efficient as possible. From these two perspectives for both data users and data providers, parametric imputation models, if used properly, may help to reduce bias and inefficiency due to missing values. In fact, if the missing mechanism is at random (MAR), it has been demonstrated that imputation can ameliorate the problems associated with incomplete data (Little and Rubin, 2002; de Waal et al., 2011).

Among others, ratio imputation is often used to treat missing values in practice (de Waal et al., 2011; Thompson and Washington, 2012; Office for National Statistics, 2014). When there is an auxiliary variable that is a *de facto* proxy for the target incomplete variable, ratio imputation is assumed to produce high quality data (Hu et al., 2001). On the other hand, proponents of multiple imputation have long argued that single imputation generally ignores estimation uncertainty by treating imputed values as if they were true values (Rubin, 1987; Schafer, 1997; Little and Rubin, 2002). Multiple imputation is indeed known to be the standard method of handling missing data (Baraldi and Enders, 2010; Cheema, 2014). In the literature, however, there is no such thing as multiple ratio imputation, leading to a gap between theory and practice. Therefore, this chapter fills in this gap by proposing a novel application of the Expectation-Maximization with Bootstrapping (EMB) algorithm to ratio imputation, where multiply-imputed values will be created for each missing value.

This chapter describes the standard single ratio imputation techniques and their limitations, illustrates the mechanism and advantages of multiple ratio imputation, and assesses the performance of multiple ratio imputation using a total of the 45,000 simulated datasets based on a variety of sample sizes, missing rates, and missingness mechanisms. This research shows that the fit of multiple ratio imputation is generally as good as or better than that of traditional imputation methods such as single ratio imputation and regular multiple imputation if the assumption holds. Thus, multiple ratio imputation will be a valuable option for treating missing data problems. Also, Software *MrImputation* is provided in Chapter 6.

5.2 Notations

Let us take a moment to review the notations used throughout this chapter. $\mathbf{D}$ is an $n \times p$ dataset, where n is the number of observations and p is the number of variables. If no data are missing, the distribution of $\mathbf{D}$ is assumed to be multivariate normal, with the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, i.e., $\mathbf{D} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let i be an observation index, $i = 1, \dots, n$. Let j be a variable index, $j = 1, \dots, p$. Thus, $\mathbf{D} = \{\mathbf{Y_1}, \dots, \mathbf{Y_p}\}$, where $\mathbf{Y_j}$ is the *j*-th column in $\mathbf{D}$, and $\mathbf{Y_{-j}}$ is the complement of $\mathbf{Y_j}$. Generally, $\mathbf{Y_{-j}}$ refers to all of the columns in $\mathbf{D}$ except $\mathbf{Y_j}$. Especially, this chapter deals with a two-variable imputation model; thus, $\mathbf{Y_1}$ is the incomplete variable (target variable for imputation) and $\mathbf{Y_2}$ is the complete variable (auxiliary variable). Thus, $\mathbf{D} = \{\mathbf{Y_{i1}}, \mathbf{Y_{i2}}\}$.

Also, let $\mathbf{R}$ be a response indicator matrix, whose dimension is the same as $\mathbf{D}$. Whenever $\mathbf{D}$

is observed $\mathbf{R} = 1$, and whenever $\mathbf{D}$ is not observed $\mathbf{R} = 0$. Note, however, that R in Italics refers to the *R* statistical environment. Furthermore, $\mathbf{D_{obs}}$ refers to the observed part of data, and $\mathbf{D_{mis}}$ refers to the missing part of data, i.e., $\mathbf{D} = \{\mathbf{D_{obs}}, \mathbf{D_{mis}}\}$.

Finally, β is the slope in the complete model, $\hat{\beta}$ is the slope estimated by the observed model, and $\tilde{\beta}$ is the estimated slope by multiple imputation.

5.3 Assumptions of Missing Mechanisms

This section briefly explains the three common assumptions of missingness (Little and Rubin, 2002, pp.11-13, pp.312-313; King et al., 2001, pp.50-51). This is an important issue, because the results of statistical analyses depend on the type of missing mechanisms (Iwasaki, 2002, pp.7-8).

The first assumption is Missing Completely At Random (MCAR), which means that the missingness probability of a variable is independent of the data for the unit. In other words, $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R})$. Take an economic survey as an example. If enterprises choose to answer their turnover values by tossing a coin, this is a perfect example of MCAR. This is the easiest case to take care of, because MCAR is simply a case of random subsampling from the intended sample; thus, subsamples may be inefficient, but unbiased. Note that the assumption of MCAR can be tested by entering dummy variables for each variable, scoring it 1 if the data are missing and 0 otherwise.

The second assumption is the case where missingness is conditionally at random. Traditionally, this is known as Missing At Random (MAR), which means that the conditional probability of missingness given data is equal to the conditional probability of missingness given observed data. In other words, $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R}|\mathbf{D_{obs}})$. If enterprises with the smaller number of employees are more likely to refuse to answer their turnover values, then this is an example of MAR, assuming that there is a column in the dataset that has values on the number of employees. If the missing mechanism is at random, imputation can rectify the bias due to missingness. Note that the assumption of MAR (unlike MCAR) cannot be tested.

The third assumption is Non-Ignorable (NI), where the missingness probability of a variable depends on the variable's value itself, and this relationship cannot be broken conditional on observed data. In other words, $P(\mathbf{R}|\mathbf{D}) \neq P(\mathbf{R}|\mathbf{D_{obs}})$. An example of NI is that enterprises with lower values of turnover are more likely to refuse to answer their turnover values and the other variables in the dataset cannot be used to predict which enterprises have small amounts of turnover. If the missing mechanism is NI, a general-purpose imputation method may not be appropriate. Instead, a special technique should be developed to take care of the unique nature of non-ignorable missing mechanisms.

Note that, to be strict, for the missingness mechanism to be ignorable, both of the MAR and distinctness conditions need to be met (Little and Rubin, 2002, pp.119-120). However, under many practical conditions, the missingness data model is often regarded as ignorable if the MAR condition is satisfied (Allison, 2002, p.5; van Buuren, 2012, p.33). This practically means that NI is Not Missing At Random (NMAR).

Also, as Carpenter and Kenward (2013, p.12) nicely put it, MAR actually means that the probability of observing a variable's value often depends on its own value, but the dependence can be eliminated, given observed data. NI means that the probability of observing a variable's value not only depends on its own value, but also the dependence cannot be eliminated, given observed data. However, the exact meaning of MAR differs from researchers to researchers (Seaman et al., 2013); thus, there is some ambivalence to this terminology.

5.4 Existing Algorithms and Software for Multiple Imputation

Before moving on to the discussion of multiple ratio imputation, this section is a concise review of the existing multiple imputation algorithms and software programs. As of today, there are three major algorithms for multiple imputation.

The first traditional algorithm is based on Markov chain Monte Carlo (MCMC). This is the original version of Rubin's (1978, 1987) multiple imputation. *R*-Package NORM currently implements this version of multiple imputation (Schafer, 1997; Fox, 2015). The second major algorithm is called Fully Conditional Specification (FCS), also known as chained equations by van Buuren (2012). R-Package MICE currently implements this version of multiple imputation (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren and Groothuis-Oudshoorn, 2015). The FCS algorithm is known to be flexible. The third relatively new algorithm is the Expectation-Maximization with Bootstrapping (EMB) algorithm by Honaker and King (2010). R-Package AMELIA II currently implements this version of multiple imputation (Honaker et al., 2011; Honaker et al., 2015). The EMB algorithm is known to be computationally efficient.

Assessing the superiority among the different multiple imputation algorithms is beyond the scope of the current study. Suffice it to say that, according to Takahashi and Ito (2013b), if the underlying distribution can be approximated by a multivariate normal distribution with the MAR condition, all of the three algorithms essentially give the same answers. As for the performance of the EMB algorithm, Honaker and King (2010) contend that the estimates of population parameters in bootstrap resamples can be appropriately used instead of random draws from the posterior. In fact, Rubin (1987, p.124) argues that the approximately Bayesian bootstrap method is proper imputation because it incorporates between-imputation variability. Also, Little and Rubin (2002, pp.216-217) assure that the substitution of Maximum Likelihood Estimates (MLEs) from bootstrap resamples is proper because the MLEs from the bootstrap resamples are asymptotically identical to a sample drawn from the posterior distribution. Therefore, multiple imputation by the EMB algorithm can be considered to be proper imputation in Rubin's sense (1987, pp.118-119). Also, according to van Buuren (2012, p.58), the bootstrap method is computationally efficient because there is no need to make a draw from the χ^2 distribution, unlike the other traditional algorithms of multiple imputation. This means that it is not necessary to resort to the Cholesky decomposition (a.k.a. the Cholesky factorization), the property of which is that if $\mathbf{A}$ is a symmetric positive definite matrix, i.e., $\mathbf{A} = \mathbf{A}^{\mathbf{T}}$, then there is a matrix $\mathbf{L}$ such that $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$, which means that $\mathbf{A}$ can be factored into $\mathbf{L}\mathbf{L}^{\mathbf{T}}$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements (Leon, 2006, p.389). Chapter 4 of this dissertation demonstrated that the EMB algorithm would be more useful than DA and FCS.

Nonetheless, R-Package AMELIA II does not allow us to estimate the multiple ratio imputation model. In fact, none of the existing multiple imputation software programs mentioned above have an option to perform multiple ratio imputation. Lee et al. (1994, p.234) developed an application-specific multiple ratio imputation model, but to the author's knowledge, a general-purpose multiple ratio imputation model has not been developed and implemented. This chapter contributes to the literature by applying the EMB algorithm to ratio imputation; thus, the new multiple ratio imputation is born.

5.5 Single Ratio Imputation

This section outlines the logic and mechanism behind ratio imputation to see why multiple ratio imputation is necessary and useful. Suppose that the population model is equation (5.1). Under the following special case, the ratio $\bar{Y}_1/\bar{Y}_2$ is an unbiased estimator of β, where ε_i is independent of Y_{i2} with the mean of 0 and the unknown variance of $Y_{i2}\sigma^2$ (Cochran, 1977, p.158; Shao, 2000, p.79; Liang et al., 2008, p.2). Under the general case, the ratio $\bar{Y}_1/\bar{Y}_2$ is a consistent but biased estimator of β, and the mean of ε_i is 0 with unknown variance. However, as the sample size increases, this bias tends to be negligible. Also, the distribution of the ratio estimate is known to be asymptotically normal (Cochran, 1977, p.153).

$$Y_{i1} = \beta Y_{i2} + \varepsilon_i \tag{5.1}$$

Suppose that Y_{it} is missing in our survey and that Y_{it-1} is fully observed in a previous dataset, where Y_{it} is the current value of the variable and Y_{it-1} is the value of the same variable at an earlier moment. The missing values of Y_t may be imputed by equation (5.2), where the value of β reflects the trend between the two time points.

$$\hat{Y}_{it} = \beta Y_{it-1} \tag{5.2}$$

A special case of equation (5.2) is cold deck imputation (de Waal et al., 2011, p.245), an example of which is that a missing value for unit i in an economic survey at t is replaced with an observed value for unit i in another highly reliable dataset such as tax data at $t-1$. This model implies that the imputer is confident that β is always 1. Thus, there will be no estimation uncertainty whatsoever. A general case of equation (5.2) is ratio imputation (de Waal et al., 2011, p.250), an example of which is that a missing value for unit i of an economic survey at t is replaced with an observed value for unit i of the same economic survey at $t-1$, assuming that unit i answered at $t-1$. In this case, the imputer is not confident that β is always 1. Thus, there will be estimation uncertainty.

Therefore, in the general case of equation (5.2), the value of β is not known and must be estimated from the observed part of data. For this purpose, ratio imputation takes the form of a simple regression model without an intercept, whose slope coefficient is calculated not by OLS, but by the ratio between the means of the two variables. In other words, the ratio imputation model is equation (5.3), where $\hat{\beta} = \bar{Y}_{1,obs}/\bar{Y}_{2,obs}$. Also, ratio imputation can be made stochastic by adding a disturbance term as in equation (5.4) (Hu et al., 2001, pp.15-16).

$$\hat{Y}_{i1} = \hat{\beta} Y_{i2} \tag{5.3}$$

$$\hat{Y}_{i1} = \hat{\beta} Y_{i2} + \hat{\varepsilon}_i \tag{5.4}$$

This study uses Table 5.1 for illustration, where the simulated data on income among 10 people are recorded.

Table 5.1. Example Data
(Simulated Weekly Income in U.S. Dollars)

ID	Income0	Income1	Income2
1	543	543	514
2	272	272	243
3	**797**	**NA**	597
4	239	239	264
5	415	415	350
6	371	371	346
7	**650**	**NA**	545
8	495	495	475
9	553	553	564
10	**710**	**NA**	558

Note. Income0 is the true complete variable. Income1 is the observed incomplete variable with NA = missing. Income2 is the auxiliary variable.

Income0 is the unobserved truth, Income1 is the current value, and Income2 is the previous value. The mean of Income0 is 504.500, the mean of income1 is 412.571, and the mean of Income2 is 445.600.

Table 5.2 presents the imputed dataset by both deterministic ratio imputation and stochastic ratio imputation. The true model is $\widehat{Income}_0 = \beta \times Income_2$, where $\beta = mean(Income_0)/mean(Income_2) = 1.132$. On the other hand, the imputation model is $\widehat{Income}_1 = \hat{\beta} \times Income_2$, where $\hat{\beta} = mean(Income_{1,obs})/mean(Income_{2,obs}) = 1.048$. This clearly means that the imputation model consistently underestimates the true model due to missing values.

Table 5.2. Example of Imputed Data
(Simulated Weekly Income in U.S. Dollars)

ID	Income0	Income1	Deterministic Ratio Imputation	Stochastic Ratio Imputation
1	543	543	543.000	543.000
2	272	272	272.000	272.000
3	**797**	**NA**	**625.594**	**586.441**
4	239	239	239.000	239.000
5	415	415	415.000	415.000
6	371	371	371.000	371.000
7	**650**	**NA**	**571.103**	**575.654**
8	495	495	495.000	495.000
9	553	553	553.000	553.000
10	**710**	**NA**	**584.756**	**621.730**

Note. Income0 is the true complete variable. Income1 is the observed incomplete variable with NA = missing.

The deterministic imputations are the exact predicted values by the imputation model. The stochastic imputations deviate from the predictions, reflecting fundamental uncertainty captured by $\hat{\varepsilon}_i$. Nevertheless, both types of ratio imputation models suffer from the lack of mechanism to incorporate estimation uncertainty, i.e., both models share the same deterministically calculated value of $\hat{\beta} = 1.048$, which is clearly different from the true $\beta = 1.132$.

Ratio imputation is considered to be an important tool in official statistics, because the model is supposed to be intuitively easy to verify for the practitioners (Bechtel et al., 2011). As a result, many national statistical agencies use ratio imputation in their statistical production processes, such as the U.S. Census Bureau (Thompson and Washington, 2012), the UK Office for National Statistics (2014), Statistics Netherlands (de Waal et al., 2011, pp.244-246), to name a few. However, this section demonstrated that the standard single ratio imputation models ignored estimation uncertainty. Thus, multiple ratio imputation comes for the rescue on this point.

5.6 Theory of Multiple Ratio Imputation

As the literature has demonstrated, if the missing mechanism is MAR, imputation can ameliorate the bias due to missingness (Little and Rubin, 2002; de Waal et al., 2011). Caution is that imputed values are not the complete reproduction of the true values, and that the goal of imputation is generally not to replicate the truth for each missing value, but to make it possible to have a valid statistical inference. For this purpose, it is necessary to evaluate the error due to missingness, for which Rubin (1978, 1987) proposed multiple imputation as a solution. Indeed, Baraldi and Enders (2010) and Cheema (2014) demonstrate that multiple imputation is superior to listwise deletion, mean imputation, and single regression imputation. Furthermore, Leite and Beretvas (2010) contend that multiple imputation is robust to violations of continuous variables and the normality assumption. Thus, multiple imputation is the standard method of treating missing data. The current study extends the utility of ratio imputation by transforming it to multiple imputation by way of the EMB algorithm described in this section.

Multiple imputation in theory is to randomly draw several imputed values from the distribution of missing data. However, missing data are by definition unobserved; as a result, the true distribution of missing data is always unknown. A solution to this problem is to estimate the posterior distribution of missing data based on observed data, and to make a random draw of imputed values. Honaker and King (2010) and Honaker et al. (2011) suggested the use of the EMB algorithm for the purpose of drawing the mean vector and the variance-covariance matrix from the posterior density, and presented a general-purpose multiple imputation software program called AMELIA II, which is a computationally efficient and highly reliable multiple imputation program. Nevertheless, as presented above, AMELIA II does not allow us to estimate the ratio imputation model.

The previous section demonstrated that the value of β was estimated by $\hat{\beta} = \bar{Y}_{1,obs}/\bar{Y}_{2,obs}$. Therefore, in order to create multiple ratio imputation, the mean vector is what needs to be randomly drawn from the posterior distribution of missing data given observed data. This chapter applies the EMB algorithm to ratio imputation to create multiple ratio imputation. In this section, let us review the bootstrap method and the Expectation-Maximization (EM) algorithm, in order to illustrate how the EMB algorithm works for the purpose of generating multiple ratio imputation.

5.6.1 Nonparametric Bootstrap

The first step for multiple ratio imputation is to randomly draw vectors of means from an appropriate posterior distribution to account for the estimation uncertainty. The EMB algorithm replaces the complex process of random draws from the posterior by nonparametric bootstrapping, which uses the existing sample data (size = n) as the pseudo-population and draws resamples (size = n) with replacement M times (Horowitz, 2001). If data $Y_1, \dots, Y_n$ are independently and identically distributed from an unknown distribution F, this distribution is estimated by $\hat{F}(y)$, which is the empirical distribution F_n defined in equation (5.5), where $I(Y)$ is the indicator function of the set Y.

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) \qquad (5.5)$$

Based on equation (5.5), bootstrap resamples are generated. The distribution $\hat{F}$ can be any estimator in order to generate the bootstrap resamples of F based on $Y_1, \dots, Y_n$. A nonparametric estimator of F is the empirical distribution F_n defined by equation (5.5) (Shao and Tu, 1995, pp.2-4, pp.9-11; DeGroot and Schervish, 2002, pp.753-754).

This section uses Table 5.3 for illustration. The incomplete data in Table 5.3 are the original missing data in Table 5.1. When listwise deletion is applied to this dataset, the mean of income1 is 412.571. The Bootstrap 1 and Bootstrap 2 in Table 5.3 refer to the bootstrap resamples, where

M = 2. When listwise deletion is applied to these bootstrap datasets, the mean of incomeB11 is 366.000 and the mean of incomeB21 is 391.286. The variation between these estimates is the essential mechanism of capturing estimation uncertainty due to imputation.

Table 5.3. Bootstrap Data (*M* = 2)

Incomplete Data		Bootstrap 1		Bootstrap 2	
Income1	Income2	IncomeB11	IncomeB12	IncomeB21	IncomeB22
543	514	NA	545	495	475
272	243	272	243	272	243
NA	597	239	264	371	346
239	264	NA	597	415	350
415	350	272	243	NA	597
371	346	553	564	543	514
NA	545	272	243	272	243
495	475	495	475	NA	545
553	564	553	564	371	346
NA	558	272	243	NA	545

Note. NA represents missing values.

However, when incomplete data are bootstrapped, the chance is that each bootstrap resample is also incomplete. Therefore, the information from incomplete bootstrap resamples is biased and inefficient. The EM algorithm refines bootstrap estimates in the next section.

5.6.2 EM Algorithm

MLEs are the parameter estimates that maximize the likelihood of observing the existing data (Long, 1997, p.26), which have the NICE properties of asymptotic Normality, Invariance, Consistency, and asymptotic Efficiency (Greene, 2003, p.473). Nevertheless, it is difficult to directly calculate MLE in missing data. Making incomplete data complete requires information about the distribution of the data, such as the mean and the variance-covariance; however, these incomplete data are used to estimate the mean and the variance-covariance, which is a chicken and egg problem. Therefore, it is not straightforward to analytically solve this problem. For the purpose of dealing with this problem, iterative methods such as the EM algorithm were proposed to estimate such quantities of interest (Allison, 2002, pp.17-20).

The EM algorithm first assumes a certain distribution and tentative starting values for the mean and the variance-covariance. Using these starting values, an expected value of model likelihood is calculated, the likelihood is maximized, model parameters are estimated that maximize these expected values, and then the distribution is updated. The expectation and the maximization steps are repeated until the values converge, whose properties are known to be an MLE (Schafer, 1997, pp.37-39; Iwasaki, 2002, pp.285-288; Do and Batzoglou, 2008). Formally, the EM algorithm can be summarized as follows. Starting from an initial value θ_0, repeat the following two steps:

E-step: $Q(\theta|\theta_t) = \int l(\theta|Y)\, P(Y_{mis}|Y_{obs}; \theta_t) dY_{mis}$, where $l(\theta|Y)$ is log likelihood.

M-step: Maximize $\theta_{t+1} = \arg\max_{\theta} Q(\theta|\theta_t)$ with respect to θ.

Under certain conditions, it is proven that $\theta_t \to \hat{\theta}\ (t \to \infty)$.

The values in Table 5.3 were incomplete. If the EM algorithm is used to refine these values, the EM mean for incomeB11 is 405.741 and the EM mean for incomeB12 is 398.100; also, the EM mean for incomeB21 is 450.912 and the EM mean for incomeB22 is 420.400. Using these values, the ratio will be estimated as 1.019 and 1.072, respectively. Thus, in this small example, the ratio is estimated as 1.046 on average, ranging from 1.019 to 1.072. This variation captures the estimation uncertainty due to missingness, which is called the between-imputation variance (Little and Rubin, 2002, p.211). Obviously, real applications require a much larger value of *M* (Graham et al., 2007; Bodner, 2008).

5.6.3 Application of the EMB Algorithm to Multiple Ratio Imputation

The multiple ratio imputation model is defined by equation (5.6), where tilde means that these values are drawn from an appropriate posterior distribution of missing data. In other words, $\tilde{\beta}$ is a vector of ratios drawn from the appropriate posterior taking estimation uncertainty into account and $\tilde{\varepsilon}_i$ is the disturbance term taking fundamental uncertainty into account (King et al., 2001, p.54).

$$\tilde{Y}_{i1} = \tilde{\beta} Y_{i2} + \tilde{\varepsilon}_i, \text{where}$$
$$\tilde{\beta} = \frac{\tilde{\bar{Y}}_1}{\tilde{\bar{Y}}_2} \tag{5.6}$$

Table 5.4 presents the result of multiple ratio imputation, where $M = 2$, using the same example data as in Table 5.1. The model is $\widetilde{Income}_1 = \tilde{\beta} \times Income_2 + \tilde{\varepsilon}_i$. If $M = 100$, the mean of $\tilde{\beta}$ is 1.050 with the standard deviation of 0.048, ranging from 0.903 to 1.342. This variation captures the stability of the imputation model, which serves as a diagnostic method for imputation, because the simulation standard error (between-imputation variance) can be appropriately used for assessing the likeliness of the simulation estimator being close to the true parameter of interest (DeGroot and Schervish, 2002, p.704). Note that, in Table 5.4, the values of Imputation1 and Imputation2 for ID 3, 7, and 10 change over columns Imputation1 to Imputation2, because the values in these rows are imputed values. Also, note that the values in the other rows do no change over columns, because these are observed values.

Table 5.4. Multiple Ratio Imputation Data ($M = 2$)

ID	Income1	Income2	Imputation1	Imputation2
1	543	514	543.000	543.000
2	272	243	272.000	272.000
3	**NA**	597	**620.917**	**662.732**
4	239	264	239.000	239.000
5	415	350	415.000	415.000
6	371	346	371.000	371.000
7	**NA**	545	**571.100**	**600.655**
8	495	475	495.000	495.000
9	553	564	553.000	553.000
10	**NA**	558	**597.406**	**637.115**

Just as in regular multiple imputation (Little and Rubin, 2002, p.86), the estimates by multiple ratio imputation can be combined as follows. Let $\hat{\theta}_m$ be an estimate based on the m-th multiply-imputed dataset. The combined point estimate $\bar{\theta}_M$ is equation (5.7).

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{5.7}$$

The variance of the combined point estimate consists of two parts. Let v_m be the estimate of the variance of $\hat{\theta}_m$, $\mathrm{var}(\hat{\theta}_m)$, let $\bar{W}_M$ be the average of within-imputation variance, let $\bar{B}_M$ be the average of between-imputation variance, and let T_M be the total variance of $\bar{\theta}_M$. Then, the total variance of $\bar{\theta}_M$ is equation (5.8), where $(1 + 1/M)$ is an adjustment factor because M is not infinite. If M is infinite, $\lim_{M\to\infty}\left(1+\frac{1}{M}\right)\tilde{v}_M = \tilde{v}_M$. In short, the variance of $\bar{\theta}_M$ takes into account within-imputation variance and between-imputation variance.

$$T_M = \bar{W}_M + \left(1+\frac{1}{M}\right)\bar{B}_M = \frac{1}{M}\sum_{m=1}^{M} v_m + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}_M\right)^2\right] \qquad (5.8)$$

Figure 1 graphically outlines a schematic overview of multiple ratio imputation (M = 5). In summary, multiple ratio imputation replaces missing values by M simulated values, where M > 1. Conditional on observed data, the imputer constructs a posterior distribution of missing data, draws a random sample from this distribution, and creates several imputed datasets. Then, researchers and analysts conduct standard statistical analysis, separately using each of the M multiply-imputed datasets, and combine the results of the M statistical analyses in the above manner to calculate a point estimate just as in regular multiple imputation.

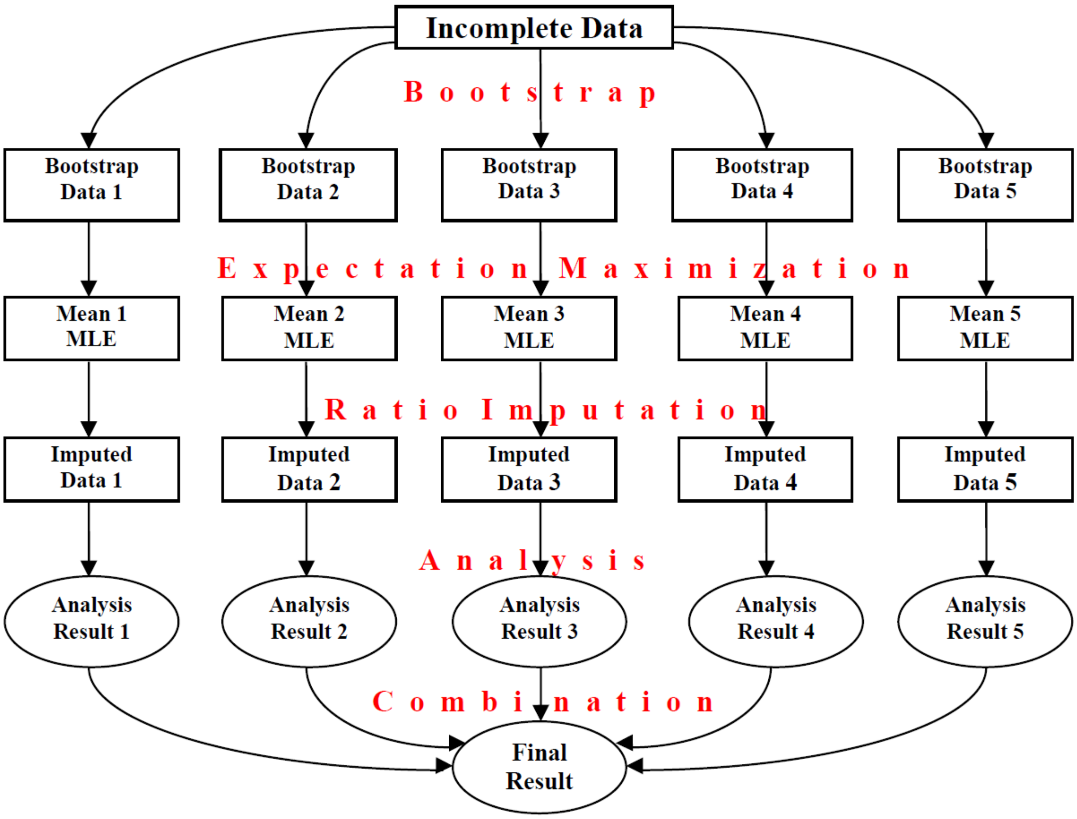


Figure 5.1. Schematic of Multiple Ratio Imputation by the EMB Algorithm (M = 5)

5.7 Monte Carlo Evidence

Using a total of the 45,000 simulated datasets with various characteristics, this section compares the Relative Root Mean Square Errors (RRMSE) of the estimators for the mean, the standard deviation, and the *t*-statistics in regression across different missing data handling techniques. The data used in this section are a modified version of the simulated data used by King et al. (2001, p.61). The Monte Carlo experiments here are based on 1,000 iterations, each of which is a random draw from the following multivariate normal distribution: Variables y1 and y2 are normally distributed with the mean vector (6, 10) and the standard deviation vector (1, 1), where the correlation between y1 and y2 is set to 0.6 (Note that the value of 0.6 was chosen because this is approximately the correlation value among the variables in official economic statistics which is the target of the current study. Also, in other few runs, not reported, the parameter values were changed, and the conclusions were very similar). Each set of these 1,000 data is repeated for *n* = 50, *n* =100, *n* =200, *n* =500, and *n* =1,000; thus, there are 5,000 datasets of five different data sizes. Our simulated data assume that the population model is equation (5.9).

$$\begin{gathered} Y_{i1} = \beta Y_{i2} + \varepsilon_i, where \\ \beta = \frac{\bar{Y}_1}{\bar{Y}_2} = 0.6, \varepsilon_i \sim N(0, 0.64) \end{gathered} \tag{5.9}$$

Furthermore, following King et al. (2001, p.61), each of these 5,000 datasets is made incomplete using the three data generation processes of MCAR, MAR, and NI as in Table 5.5.

Table 5.5: Missingness Mechanisms and Missing Rates

MCAR	Missingness of y1 is a function of u. 15%: y1 is missing if u > 0.85. 25%: y1 is missing if u > 0.75. 35%: y1 is missing if u > 0.65.
MAR	Missingness of y1 is a function of y2 and u. 15%: y1 is missing if y2 > 10 and u > 0.7. 25%: y1 is missing if y2 > 10 and u > 0.5. 35%: y1 is missing if y2 > 10 and u > 0.3.
NI	Missingness of y1 is a function of y1, x, and u. 15%: y1 is missing if y1 > 6 and u > 0.7. 25%: y1 is missing if y1 > 6 and u > 0.5. 35%: y1 is missing if y1 > 6 and u > 0.3.

Under the assumption of MCAR, the missingness of y1 randomly depends on the values of u (uniform random numbers). Under the assumption of MAR, the missingness of y1 depends on the values of y2 and u. Under the assumption of NI, the missingness of y1 depends on the observed and unobserved values of y1 itself and the values of u.

Note that Variable y1 is the target incomplete variable for imputation, Variable y2 is completely observed in all of the situations to be used as the auxiliary variable, and Variable u in Table 5.5 is 1,000 sets of continuous uniform random numbers ranging from 0 to 1 for the missingness mechanism. The average missing rates are set to 15%, 25%, and 35%. These missing rates approximately cover the range from 10% to 40% missingness.

Therefore, there is a total of 45,000 datasets, i.e., 1,000 datasets multiplied by five sample sizes, three missing mechanisms, and three missing rates.

The overall performance can be captured by the Mean Square Error (MSE), which is defined as equation (5.10), where θ is the true quantity of interest and $\hat{\theta}$ is an estimator. The MSE measures the dispersion around the true value of the parameter, suggesting that an estimator with the smallest MSE is the best of a competing set of estimators (Gujarati, 2003, p.901).

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{5.10}$$

For the ease of interpretation, following Di Zio and Guarnera (2013, p.549), this study uses the Relative Root Mean Square Error (RRMSE), which is defined as equation (5.11), where θ is the truth, $\hat{\theta}$ is an estimator, and *T* is the number of trials. For example, θ in the following analyses is the mean, the standard deviation, and the *t*-statistic based on complete data. $\hat{\theta}$ is the estimated quantity based on imputed data. *T* is 1,000.

$$RRMSE(\hat{\theta}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\theta} - \theta}{\theta}\right)^2} \tag{5.11}$$

The complete results based on the 45,000 datasets are presented in Tables 5.6, 5.8, and 5.9. In the following analyses, the multiple ratio imputation model sets the number of multiply-imputed datasets (M) to 100, based on the recent findings in the multiple imputation literature (Graham et al., 2007; Bodner, 2008).

5.7.1 RRMSE Comparisons for the Mean

Table 5.6 presents the RRMSE comparisons for the mean among listwise deletion, deterministic single ratio imputation, and multiple ratio imputation ($M = 100$), where the RRMSE is averaged over the 1,000 simulations. For multiple ratio imputation, the 100 mean values are combined using equation (5.7) in each of the 1,000 simulations.

The standard recommendation (de Waal et al., 2011, p.245) is that if the goal is to calculate a point estimate, the choice is deterministic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of deterministic single ratio imputation, which is known to be a preferred method for the estimation of the mean. If multiple ratio imputation equally performs well compared to deterministic single ratio imputation, this means that multiple ratio imputation attains the highest performance in estimating the mean.

In 42 of the 45 patterns, deterministic ratio imputation and multiple imputation both outperform listwise deletion with 3 ties. Even when the missing mechanism is MCAR, the results by imputation are almost always better than those of listwise deletion. Between the ratio imputation methods, deterministic ratio imputation slightly performs better than multiple ratio imputation in 14 out of the 45 patterns with 31 ties. However, the largest difference is only 0.002 in terms of the RRMSE. Thus, there are no significant differences between deterministic ratio imputation and multiple ratio imputation. Furthermore, this difference is expected to completely disappear as M approaches infinity. In general, under the situations where the model is correctly specified and the assumption of MAR is satisfied, both single imputation and multiple imputation ($M = \infty$) would be unbiased and agree on the point estimation (Donders et al., 2006, p.1089).

Table 5.6. RRMSE Comparisons for the Mean (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
50	15%	MCAR	0.009	0.008	0.008
		MAR	0.017	0.008	0.008
		NI	0.026	0.017	0.018
	25%	MCAR	0.014	0.011	0.011
		MAR	0.030	0.010	0.011
		NI	0.048	0.032	0.033
	35%	MCAR	0.017	0.014	0.014
		MAR	0.045	0.012	0.014
		NI	0.075	0.050	0.052
100	15%	MCAR	0.007	0.006	0.006
		MAR	0.016	0.005	0.005
		NI	0.024	0.016	0.016
	25%	MCAR	0.010	0.008	0.008
		MAR	0.028	0.007	0.008
		NI	0.046	0.030	0.030
	35%	MCAR	0.012	0.010	0.010
		MAR	0.044	0.008	0.010
		NI	0.073	0.048	0.050
200	15%	MCAR	0.005	0.004	0.004
		MAR	0.015	0.004	0.004
		NI	0.024	0.016	0.016
	25%	MCAR	0.007	0.005	0.005
		MAR	0.028	0.005	0.005
		NI	0.045	0.029	0.030
	35%	MCAR	0.009	0.007	0.007
		MAR	0.043	0.006	0.007
		NI	0.072	0.048	0.049
500	15%	MCAR	0.003	0.003	0.003
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	MCAR	0.004	0.003	0.003
		MAR	0.027	0.003	0.003
		NI	0.045	0.029	0.029
	35%	MCAR	0.006	0.004	0.004
		MAR	0.043	0.004	0.005
		NI	0.072	0.047	0.048
1000	15%	MCAR	0.002	0.002	0.002
		MAR	0.014	0.002	0.002
		NI	0.024	0.015	0.015
	25%	MCAR	0.003	0.003	0.003
		MAR	0.027	0.002	0.002
		NI	0.044	0.029	0.029
	35%	MCAR	0.004	0.003	0.003
		MAR	0.043	0.002	0.003
		NI	0.072	0.047	0.048

Note. Average over the 1,000 simulations for each data type. $M = 100$ for multiple ratio imputation

The results in Table 5.6 assure that this general relationship also applies to the relationship between single ratio imputation and multiple ratio imputation. Therefore, on average, multiple ratio imputation can be expected to give essentially the same answers as to the estimation of the mean, compared to deterministic ratio imputation.

On top of this, multiple ratio imputation can be more useful than deterministic single ratio imputation in the estimation of the mean, because multiple ratio imputation has more information in its output. Recall that there are three sources of variation in multiple imputation (van Buuren, 2012, p.38). One is the conventional measure of statistical variability (also known as within-imputation variance). Another is the additional variance due to missing values in the data (also known as between-imputation variance). The last one is simulation variance by the finite number of multiply-imputed data captured by $\bar{B}_M/M$ in equation (5.8). Among these, the between-imputation variance is particularly important, because it reflects the uncertainty associated with missingness (Honaker et al., 2011, p.23).

To demonstrate how multiple ratio imputation provides additional information on the between-imputation variance, Table 5.7 presents the mean of y1 when the missing data mechanism is MAR with the average missing rate of 35%, where the reported values are the average over the 1,000 simulations. In Table 5.7, when the missing data mechanism is MAR, both of the imputation methods are almost equally accurate, in terms of estimating the mean. Additionally, multiple ratio imputation has more rows in Table 5.7 for BISD and CI (95%). BISD stands for the Between-Imputation Standard Deviation, and CI (95%) stands for the Confidence Interval associated with estimation error due to missingness at the 95% level. BISD is the square-root of the between-imputation variance and measures the dispersion of the 100 mean values based on multiple ratio imputation (M = 100). In other words, BISD is the variation in the distribution of the estimated mean, which is usually called the standard error (Baraldi and Enders, 2010, p.16). Thus, based on BISD, the imputer can be approximately 95% confident that the true mean value of complete data is somewhere between 5.941 and 6.057, after taking the error due to missingness into account.

Furthermore, the imputer can be approximately 95% confident that the imputed mean value (6.00) is meaningfully different from the listwise deletion estimate (5.74), which is outside the 95% confidence interval (5.94, 6.06). Single ratio imputation (both deterministic and stochastic) lacks this mechanism of assessing estimation uncertainty.

Table 5.7. Mean of y1 (MAR-35%)

	Complete	Listwise	Deterministic	Multiple
	Data	Deletion	Ratio Imputation	Ratio Imputation
Mean	6.000	5.741	6.000	5.999
BISD	NA	NA	NA	0.029
CI (95%)	NA	NA	NA	5.941, 6.057
n	500	325	500	500

Note. NA means Not-Applicable. Average over the 1,000 simulations. $M = 100$ for multiple ratio imputation

5.7.2 RRMSE Comparisons for the Standard Deviation

Table 5.8 presents the RRMSE comparisons for the standard deviation among listwise deletion, stochastic single ratio imputation, and multiple ratio imputation ($M = 100$), where the RRMSE is averaged over the 1,000 simulations. For multiple ratio imputation, the 100 standard deviation values are combined using equation (5.7) in each of the 1,000 simulations.

The standard recommendation (de Waal et al., 2011, p.245) is that if the goal is to estimate the variation of data, the choice is stochastic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of stochastic ratio imputation, which is known to be a preferred method to estimate the standard deviation. Note that, in other simulation runs, the EM algorithm was applied to the imputed data by the deterministic ratio imputation model, in order to compute the standard deviation. However, these results were not good and thus omitted here.

Table 5.8. RRMSE Comparisons for the Standard Deviation (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Stochastic Ratio Imputation	Multiple Ratio Imputation
50	15%	MCAR	0.042	0.048	0.037
		MAR	0.045	0.047	0.038
		NI	0.048	0.052	0.043
	25%	MCAR	0.059	0.062	0.049
		MAR	0.066	0.062	0.054
		NI	0.079	0.074	0.067
	35%	MCAR	0.075	0.075	0.058
		MAR	0.088	0.071	0.067
		NI	0.146	0.117	0.118
100	15%	MCAR	0.029	0.035	0.026
		MAR	0.031	0.034	0.026
		NI	0.035	0.037	0.031
	25%	MCAR	0.040	0.044	0.033
		MAR	0.046	0.044	0.037
		NI	0.064	0.058	0.054
	35%	MCAR	0.052	0.052	0.040
		MAR	0.067	0.054	0.047
		NI	0.121	0.097	0.098
200	15%	MCAR	0.021	0.025	0.018
		MAR	0.022	0.025	0.019
		NI	0.025	0.027	0.023
	25%	MCAR	0.028	0.030	0.023
		MAR	0.036	0.032	0.027
		NI	0.049	0.044	0.042
	35%	MCAR	0.037	0.037	0.028
		MAR	0.053	0.038	0.034
		NI	0.109	0.086	0.088
500	15%	MCAR	0.014	0.016	0.012
		MAR	0.014	0.016	0.012
		NI	0.018	0.019	0.016
	25%	MCAR	0.018	0.020	0.015
		MAR	0.024	0.020	0.017
		NI	0.042	0.038	0.036
	35%	MCAR	0.022	0.023	0.018
		MAR	0.043	0.024	0.021
		NI	0.106	0.083	0.084
1000	15%	MCAR	0.010	0.012	0.008
		MAR	0.010	0.011	0.008
		NI	0.014	0.015	0.013
	25%	MCAR	0.013	0.014	0.011
		MAR	0.019	0.014	0.011
		NI	0.040	0.037	0.033
	35%	MCAR	0.017	0.017	0.013
		MAR	0.038	0.016	0.014
		NI	0.100	0.080	0.079

Note. Average over the 1,000 simulations for each data type. *M* = 100 for multiple ratio imputation

In all of the 45 patterns, multiple ratio imputation always outperforms listwise deletion. Even when the missing mechanism is MCAR, the results by multiple ratio imputation are always better than those of listwise deletion. In contrast, stochastic ratio imputation outperforms listwise deletion in only 20 out of the 45 patterns. Especially, when the missing mechanism is MCAR, listwise deletion often outperforms stochastic ratio imputation in 11 out of the 15 patterns with 4 ties, although the difference is minimal. This implies that when missing data are suspected to be MCAR, there is a chance that using stochastic ratio imputation may make the situation worse than simply using listwise deletion. When the missing mechanism is MAR or NI, stochastic ratio imputation indeed outperforms listwise deletion in 20 out of the 30 patterns.

Between the ratio imputation methods, multiple ratio imputation often performs better than stochastic ratio imputation, 41 out of the 45 patterns. Therefore, this study contends that multiple ratio imputation is the preferred method for the estimation of the standard deviation. Table 5.8 implies that, regardless of missing mechanisms, multiple ratio imputation should be used for the purpose of estimating the standard deviation.

Just as in the case of estimating the mean, let us take the case of 35% missingness with the MAR condition as an example. Based on BISD, the imputer can be approximately 95% confident that the true standard deviation value of complete data is somewhere between 0.960 and 1.040, after taking the error due to missingness into account.

5.7.3 RRMSE Comparisons for the *t*-Statistics in Regression

The comparisons in this section are particularly important because up to today, even if the intercept should be zero and the slope should be estimated by the ratio between two variables, there are no other choices but to stick to regular multiple imputation for the computation of the *t*-statistics in regression. The regression model in Table 5.9 is y2 = a + b*y1. The quantity of interest is the *t*-statistic of b, i.e., $t_b = b/\mathrm{se}(b)$. The RRMSE reported here measures the average distance between the true t_b based on complete data and the estimated t_b based on imputed data. Table 5.9 presents the RRMSE comparisons for the *t*-statistics in regression among listwise deletion, regular multiple imputation (AMELIA II), and multiple ratio imputation, where $M = 100$ for both regular multiple imputation and multiple ratio imputation, and the RRMSE is averaged over the 1,000 simulations. For regular multiple imputation and multiple ratio imputation, the 100 coefficient values are combined using equation (5.7), the 100 standard error values are combined using equation (5.8), and the *t*-statistics are calculated using these two values in each of the 1,000 simulations. Remember that the multiple ratio imputation model is equation (5.6). On the other hand, multiple imputation by AMELIA II is equation (5.12), where the coefficients are random draws of the mean vectors and the variance-covariance matrices from the posterior distribution (Honaker and King, 2010).

$$\begin{gathered} \tilde{Y}_{i1} = \tilde{\beta}_0 + \tilde{\beta}_1 Y_{i2} + \tilde{\varepsilon}_i, \text{where} \\ \tilde{\beta}_1 = \frac{\widetilde{cov}(Y_{i1}, Y_{i2})}{\widetilde{var}(Y_{i2})} \\ \tilde{\beta}_0 = \tilde{\bar{\bar{Y}}}_1 - \tilde{\beta}_1 \tilde{\bar{\bar{Y}}}_2 \end{gathered} \tag{5.12}$$

The standard recommendation (van Buuren, 2012, pp.16-18; Hughes et al., 2016) is that if the goal is to obtain valid inferences with standard errors, the choice is multiple imputation which is a superior variance-estimation method. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is better than that of regular multiple imputation in terms of estimating the *t*-statistics. The comparison of the *t*-statistics in regression is appropriate, because it is the quantity of interest for many applied researchers in disputing whether an independent variable has some impact on a dependent variable. According to Cheema (2014, p.58), comparisons of *t* statistics are fair because the complete sample and the imputed sample are identical in all respects including power, except for the fact that no values were missing in the complete sample while some values were missing in the imputed values. Therefore, the differences in the observed value of statistics are caused by the differences between imputed values and their true counterparts.

Table 5.9. RRMSE Comparisons for *t*-statistics (45,000 Datasets)

Sample Size	Average Missing Rate	Missing Mechanism	Listwise Deletion	Multiple Imputation AMELIA II	Multiple Ratio Imputation
50	15%	MCAR	0.126	0.103	0.087
		MAR	0.137	0.107	0.093
		NI	0.141	0.114	0.099
	25%	MCAR	0.185	0.144	0.113
		MAR	0.220	0.173	0.135
		NI	0.222	0.175	0.138
	35%	MCAR	0.242	0.189	0.134
		MAR	0.317	0.247	0.171
		NI	0.328	0.269	0.179
100	15%	MCAR	0.104	0.075	0.066
		MAR	0.113	0.080	0.071
		NI	0.111	0.081	0.072
	25%	MCAR	0.159	0.109	0.087
		MAR	0.192	0.127	0.101
		NI	0.194	0.136	0.108
	35%	MCAR	0.218	0.153	0.107
		MAR	0.294	0.191	0.131
		NI	0.297	0.224	0.147
200	15%	MCAR	0.091	0.059	0.052
		MAR	0.101	0.064	0.056
		NI	0.101	0.066	0.060
	25%	MCAR	0.145	0.092	0.075
		MAR	0.181	0.106	0.085
		NI	0.177	0.117	0.095
	35%	MCAR	0.208	0.136	0.097
		MAR	0.282	0.159	0.113
		NI	0.282	0.199	0.133
500	15%	MCAR	0.084	0.050	0.044
		MAR	0.094	0.053	0.047
		NI	0.093	0.058	0.051
	25%	MCAR	0.141	0.086	0.066
		MAR	0.171	0.092	0.069
		NI	0.170	0.107	0.083
	35%	MCAR	0.202	0.127	0.086
		MAR	0.279	0.144	0.097
		NI	0.282	0.193	0.121
1000	15%	MCAR	0.080	0.046	0.041
		MAR	0.089	0.046	0.043
		NI	0.091	0.048	0.049
	25%	MCAR	0.137	0.053	0.063
		MAR	0.167	0.084	0.067
		NI	0.168	0.105	0.083
	35%	MCAR	0.198	0.122	0.084
		MAR	0.275	0.132	0.092
		NI	0.275	0.186	0.120

Note. Average over the 1,000 simulations for each data type. *M* = 100 for multiple imputation

The comparison of multiple ratio imputation and AMELIA II is appropriate, because the algorithm is the same EMB under the same platform of the *R* statistical environment. In all of the 45 patterns, regular multiple imputation and multiple ratio imputation both outperform listwise deletion. Furthermore, multiple ratio imputation almost always outperforms regular multiple imputation 43 out of the 45 patterns under the condition where the true population model is equation (5.9). Thus, when the true model is a ratio model such as equation (5.9), multiple ratio imputation is more accurate and efficient than regular multiple imputation.

Therefore, multiple ratio imputation adds an important option for the tool kit of imputing and analyzing the mean, the standard deviation, and the *t*-statistics. If the true model is equation (5.9), multiple ratio imputation is at least as good as and in many cases better than the other traditional imputation methods for the three quantities of interest, regardless of the missingness mechanisms. To be fair, this chapter never claims that multiple ratio imputation is always superior to regular multiple imputation. If the true model is not a ratio model such as equation (5.9), the superiority shown in this section is not guaranteed.

5.8 Conclusion

This chapter proposed a novel application of the EMB algorithm to ratio imputation and presented the mechanism and the usefulness of multiple ratio imputation. For this purpose, Monte Carlo evidence was presented, where the newly-developed *R*-function called *MrImputation* (See Chapter 6 of this dissertation) for multiple ratio imputation was applied to the 45,000 simulated data.

This research showed that the fit of multiple ratio imputation was generally as good as or sometimes better than that of single ratio imputation and regular multiple imputation if the assumption holds. Specifically, for the purpose of estimating the mean, the performance of deterministic ratio imputation and multiple ratio imputation are essentially equally good, with multiple ratio imputation having additional information on estimation uncertainty. For the purpose of estimating the standard deviation, multiple ratio imputation outperforms stochastic ratio imputation. For the purpose of estimating the *t*-statistics in regression, multiple ratio imputation clearly outperforms regular multiple imputation when the population model is equation (5.6).

These findings are important because researchers are recommended to use different ways of imputation depending on the type of statistical analyses, meaning that there are no one-size-fit-for-all imputation methods (Poston and Conde, 2014, p.476). Thus, multiple ratio imputation will be a valuable addition for treating missing data problems, so that multiple ratio imputation will expand the choice of missing data treatments.

This said, the current research is only a starting point for multiple ratio imputation. As noted in Chapter 4, there are three multiple imputation algorithms. The version of multiple ratio imputation introduced in this research utilized the Expectation-Maximization with Bootstrapping algorithm. However, multiple ratio imputation is a generic imputation model; thus, future research may apply the other two multiple imputation algorithms to expand the scope and the applicability of the method.

6 Implementing Multiple Ratio Imputation by the EMB Algorithm in R

This chapter derived from Takahashi (2017c), a peer-reviewed article in the *Journal of Modern Applied Statistical Methods* 16(1), which is operated by the Wayne State University Library System, classified as one of the top 115 libraries in the United States by the Association for Research Libraries (Kyrillidou et al., 2015). The *Journal of Modern Applied Statistical Methods* is indexed in Scopus by Elsevier as of April 2017. The author would like to thank JMASM Inc. for permission to use "Implementing multiple ratio imputation by the EMB algorithm (R)" (*Journal of Modern Applied Statistical Methods*, vol.16, no.1, 657-673).

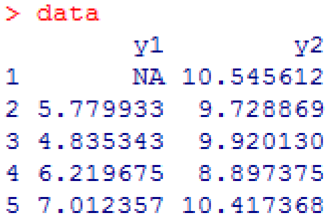
6.1 Introduction

Code is presented for multiple ratio imputation step by step in the imputation stage, followed by the analysis stage. The Appendix combines these *R*-codes to present Software *MrImputation* as a collection of *R*-functions `mrimpute` and `mranalyze`. *R*-function `mrimpute` performs multiple ratio imputation. *R*-function `mranalyze` allows us to conduct statistical analyses using the multiply-imputed data by *R*-function `mrimpute`.

6.2 Preparation Stage

As an illustration, let us use the following dataset named `data`. Note that, in the complete code presented in Appendix, the name of `data` can be defined by option `data=`. Thus, it can be named any way an imputer wants it to be. This small example dataset contains two variables and five units as displayed in Figure 6.1. The observation for unit 1 in y1 is missing (NA). Thus, y1 is the target incomplete variable for imputation, and y2 is the auxiliary complete variable. Also, y1 is stored in `data[,1]` and y2 in `data[,2]`. This chapter will use this small dataset for illustration. As this dataset implies, the target variable for imputation needs to be stored in the first column of data, i.e., `data[,1]`, in order to execute the code shown in this chapter.

```
data<-read.csv("data.csv",header=T)
attach(data)
```



```
> data
        y1        y2
1       NA 10.545612
2 5.779933  9.728869
3 4.835343  9.920130
4 6.219675  8.897375
5 7.012357 10.417368
```

Figure 6.1: Example of Incomplete Data

The number of multiply-imputed data is set by `M`, where `M` > 1. In this example, it is set to 2 so that the outputs can be visually presented below. To allow reproducibility, the random number seed value needs to be set by function `set.seed`. This step is necessary, because multiple imputation relies on pseudo-random numbers; thus, without setting a seed, there will be no way of reproducing the same results.

```
M<-2
set.seed(1223)
```

Many economic data are skewed to the right in the distribution, i.e., the distribution is not multivariate normal, but multivariate log-normal. If this is the case, a sensible option to deal with such a variable is to use log-transformation, and the imputed values will be unlogged after imputations are completed (Allison, 2002, p.39; Honaker et al., 2011, p.15). In the complete code shown in Appendix, if `log=TRUE`, then the following code log-transforms the data. The default setting is that `log=FALSE`. Obviously, if data are multivariate normal to begin with, this option should be set to `FALSE`.

```
if(log){
  data<-log(data)
}
```

6.3 Imputation Stage

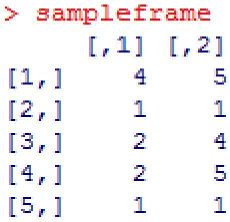
6.3.1 Nonparametric Bootstrap

The first step to perform multiple ratio imputation is to implement random draws of $\boldsymbol{\mu}$ from an appropriate posterior distribution to account for estimation uncertainty. The EMB algorithm substitutes the complex process of drawing $\boldsymbol{\mu}$ from the posterior distribution with a nonparametric bootstrapping algorithm, which is a resampling method, where the observed sample is used as the pseudo-population. In other words, a resample of size *n* is randomly drawn from this observed sample of size *n* with replacement, and this process is repeated *M* times (Shao and Tu, 1995; Horowitz, 2001).

R-function `sample(x,size,replace=TRUE)` can be used for this purpose, where `x` is a vector from which to sample, `size` is the number of items to sample, and `replace=TRUE` specifies sampling with replacement. Unfortunately, this function randomly draws a vector, not a matrix. In the process of imputation, the imputer must keep a pair of observations for the two variables. Thus, our code first creates `sampleframe` to randomly draw the row number of data, which is an `nrow(data)` by `M` matrix, where `nrow(data)` is the number of rows in `data`.

```
sampleframe<-matrix(sample(nrow(data),
                           nrow(data)*M,
                           replace=TRUE),
                           nrow=nrow(data),
                           ncol=M)
```

The resulting matrix obtained from the above code is displayed in Figure 6.2, where each column contains a vector of the row numbers randomly drawn from the original data. For example, `sampleframe[1,1]` is 4, meaning that this cell refers to row number 4 in the original data, i.e., y1 = 6.219675 and y2 = 8.897375, `sampleframe[2,1]` is 1, meaning that this cell refers to row number 1 in the original data, i.e., y1 = NA and y2 = 10.545612, and so on.



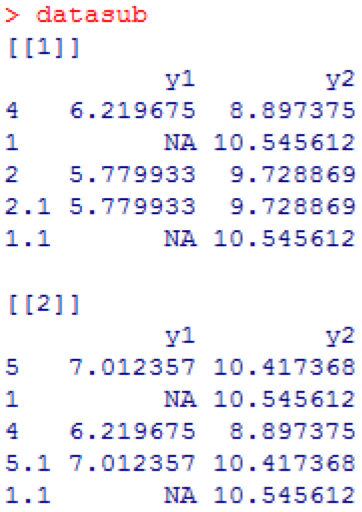
```
> sampleframe
     [,1] [,2]
[1,]    4    5
[2,]    1    1
[3,]    2    4
[4,]    2    5
[5,]    1    1
```

Figure 6.2: Randomly-Drawn Row Numbers

Based on `sampleframe`, our code makes a random draw of the values of y1 and y2 from the original data *M* times. First, let us create a list named `datasub` with the elements of NA and then replace these NAs by appropriate values in the original data, so that `datasub[[i]]` obtains `data[sampleframe[,i],]`, and the `for` loop repeats this process *M* times. In order to use this `datasub` in the EM algorithm below, `datasub` is transformed to a matrix.

```
datasub<-as.list(rep(NA,M))
for(i in 1:M){
  datasub[[i]]<-as.matrix(data[sampleframe[,i],])
}
```

The resulting bootstrap resamples are shown in Figure 6.3, where `datasub[[1]]` and `datasub[[2]]` represent the *m*-th bootstrap resample, respectively.



```
> datasub
[[1]]
          y1        y2
4   6.219675  8.897375
1         NA 10.545612
2   5.779933  9.728869
2.1 5.779933  9.728869
1.1       NA 10.545612

[[2]]
          y1        y2
5   7.012357 10.417368
1         NA 10.545612
4   6.219675  8.897375
5.1 7.012357 10.417368
1.1       NA 10.545612
```

Figure 6.3: Example of Bootstrap Resamples ($M = 2$)

6.3.2 EM Algorithm

Each bootstrap resample created above is likely to be incomplete. Estimates using these resamples are expected to be biased and inefficient. In order to avoid this problem, the EM algorithm is used to refine the estimates in bootstrap resamples.

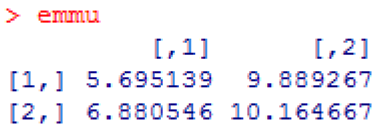
The EM algorithm calculates an expected value of model likelihood, maximizes the likelihood, estimates parameters that maximize the obtained expected values, and updates the distribution. After repeating these expectation and maximization steps several times, the value that converged is known to be an MLE (Little and Rubin, 2002, pp.166-169; Do and Batzoglou, 2008). *R*-package NORM originally created by Schafer (1997) is a multiple imputation program based on Markov chain Monte Carlo (MCMC). The process of multiple imputation by NORM begins with an initial estimate of the parameters by the EM algorithm, which is performed by function `em.norm` (Fox, 2015). Our code does not use NORM for the sake of generating multiple imputation, but function em.norm is useful for the computational purpose of the EM algorithm. First, use the require function to load NORM in *R*. In the code below, p is the number of columns (variables) in the data, para is the number of parameters to be estimated, thetahat is an empty matrix with the dimension of M by para, and emmu is an empty matrix with the dimension of M by p. These are housekeeping issues to perform the EM algorithm by way of function em.norm.

```
require(norm)
p<-ncol(data)
para<-p*(p+3)/2+1
thetahat<-matrix(NA,M,para)
emmu<-matrix(NA,M,p)
```

Function prelim.norm takes care of the preliminary manipulations for a matrix of incomplete data, which is a necessary step for using em.norm, whose results are stored in thetahat. Option showits=FALSE quietly runs em.norm. If the imputer wants to monitor the iteration process of EM, then this option should be set to TRUE. Option maxits=1000 sets the maximum number of iterations to 1,000. Function getparam.norm produces the estimated values of the MLEs, which is stored in emmu. Option corr=FALSE computes the means and variance-covariance matrix. The for loop repeats the em.norm function to be applied to datasub *M* times. This process is the essential part of the EMB algorithm, meaning that the EM algorithm is applied to each of the *M* bootstrap resamples.

```
for(i in 1:M){
  thetahat[i,]<-em.norm(prelim.norm(datasub[[i]]),
  showits=FALSE,maxits=1000)
  emmu[i,]<-getparam.norm(prelim.norm(datasub[[i]]),
  thetahat[i,],corr=FALSE)$mu
}
```

Now, all of the estimates of the means by the EM algorithm are stored in emmu. Thus, typing emmu returns the following matrix in Figure 6.4, where the first column refers to the means for the first variable in the data, and the second column refers to the means for the second variable in the data. Also, the first row refers to the means in $m = 1$ and the second row refers to the means in $m = 2$. Note that these are the MLEs of the means.



```
> emmu
         [,1]      [,2]
[1,] 5.695139  9.889267
[2,] 6.880546 10.164667
```

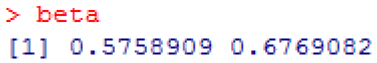
Figure 6.4: MLEs for the Means of y1 and y2

6.3.3 Implementation of Multiple Ratio Imputation

Using matrix `emmu` allows us to estimate multiple ratios of two variables as follows. The estimated ratios are stored in `beta`, which is an empty matrix with the dimension of `M` by `ncol(data)-1`. Ratio imputation has only two variables; thus, the number of columns in the data, i.e., `ncol(data)`, is 2, which means that `beta` is essentially an `M` by 1 column vector.

```
beta<-matrix(NA,M,ncol(data)-1)
beta<-emmu[,1]/emmu[,2]
```

Typing `beta` returns a vector of *M* values, where the first value is the ratio in the first model, the second value in the second model, and so on. This is $\tilde{\beta}$ in equation (5.6).



```
> beta
[1] 0.5758909 0.6769082
```

Figure 6.5: The Values of the Slopes in the Multiple Ratio Imputation Model

As a preparation for multiple ratio imputation, let us define the following matrices. These are housekeeping issues to perform multiple ratio imputation. All of the matrices are empty matrices with the dimensions of `nrow(data)` by `M`.

```
imp<-matrix(NA,nrow(data),M)
resid<-matrix(NA,nrow(data),M)
e<-matrix(NA,nrow(data),M)
imp1<-matrix(NA,nrow(data),M)
imp2<-matrix(NA,nrow(data),M)
```

Now, everything is ready to perform multiple ratio imputation. The values of `beta` are multiplied by `data[,2]` which is the values of the second variable in the data. Specifically, `data[,2]` is y2 in our example. Thus, the following code is $\tilde{\beta}Y_{i2}$ in equation (5.6). The `for` loop repeats this process *M* times. The imputed values are stored in `imp`, where `imp[,1]` is the imputed data from *m* = 1 and `imp[,2]` is the imputed data from *m* = 2.

```
for(i in 1:M){
  imp[,i]<-beta[i]*data[,2]
}
```

To complete the process, a small disturbance term needs to be added to the imputed values, which is $\tilde{\varepsilon}_i$ in equation (5.6). In the following code, `resid` is the differences (residuals) between observed values and predicted values. Also, $\tilde{\varepsilon}_i$ is `e[,i]`, which is normally distributed with the mean of 0 and the standard deviation of the residuals, `resid[,i]`. In the last line, `e[,i]` is added to `imp[,i]`. The `for` loop repeats this whole process *M* times.

```
for(i in 1:M){
  resid[,i]<-data[,1]-imp[,i]
  e[,i]<-rnorm(nrow(data),0,sd(resid[,i],na.rm=TRUE))
  imp1[,i]<-imp[,i]+e[,i]
}
```

Up to this point, all of the values were imputed, both observed and missing. What actually needs to be imputed is the missing part of the data only. Therefore, the final step is to replace `NA` with `imp1` and to keep the observed value as is. In the following code, `imp2` is essentially $\tilde{Y}_{i1}$ in equation (5.6). If `data[j,1]` is missing, then `imp2[j,i]` obtains the imputed value `imp1[j,i]`; otherwise, `imp2[j,i]` obtains `data[j,1]`. In the following loop, `i` refers to the number of imputations and `j` refers to the row number in the data.

```
for(i in 1:M){
  for(j in 1:nrow(data)){
   if (is.na(data[j,1])=="TRUE"){
   imp2[j,i]<-imp1[j,i]
   }else{
   imp2[j,i]<-data[j,1]}
}}
```

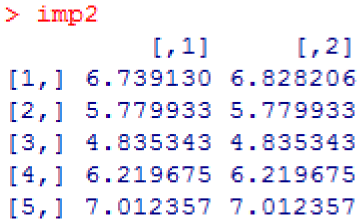
Remember that log-normal data were log-transformed above. Imputed values must be put back to the original scale of incomplete data. The following code unlogs the log-transformed variables.

```
if(log){
  imp2<-exp(imp2)
  data<-exp(data)
}
```

Some variables have logical bounds. For instance, economic variables such as turnover cannot be negative. If this is the case, `zero=TRUE` can be specified in the complete code in Appendix. This option forces negative imputed values to be zero. Warning is that this option may suppress the correct uncertainty in the imputation model (Honaker et al., 2011, pp.23-25); thus, this option should be used cautiously. The default setting is `zero=FALSE`.

```
if(zero){
  imp2[which(imp2<0)]<-0
}
```

Finally, `imp2` returns the following two sets of imputed data, because *M* = 2. The values in row [1,] change over columns [,1] to [,2], because these values are imputed values. The values in the other rows do no change over columns, because these are observed values.



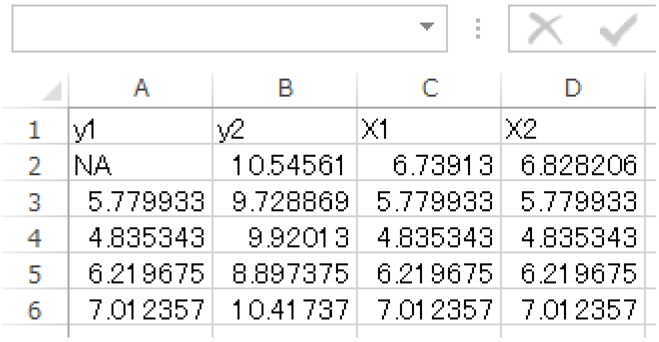
```
> imp2
         [,1]     [,2]
[1,] 6.739130 6.828206
[2,] 5.779933 5.779933
[3,] 4.835343 4.835343
[4,] 6.219675 6.219675
[5,] 7.012357 7.012357
```

Figure 6.6: Example of Multiply-Imputed Data

The `write.csv` function saves the imputed data along with the original data as follows, where `y1` is the original incomplete variable, `y2` is the original auxiliary variable, and `imp2` is a matrix of *M* imputed data created above.

```
y1<-data[,1]; y2<-data[,2]
impdata<-data.frame(y1,y2,imp2)
write.csv(impdata,"mridata.csv",row.names=FALSE)
```

Figure 6.7 is the output data named `mridata` in the csv format, which can be reloaded in *R* or any statistical software of an analyst's choice for subsequent statistical analyses. In this output dataset, Column A (y1) is the original incomplete data, Column B (y2) is the original auxiliary variable, and Columns C to D (X1, X2) are the multiply imputed data.



	A	B	C	D
1	y1	y2	X1	X2
2	NA	10.54561	6.73913	6.828206
3	5.779933	9.728869	5.779933	5.779933
4	4.835343	9.92013	4.835343	4.835343
5	6.219675	8.897375	6.219675	6.219675
6	7.012357	10.41737	7.012357	7.012357

Figure 6.7: Example of Output Data (csv file)

6.4 Analysis Stage

6.4.1 Mean and Standard Deviation

After reading `mridata.csv`, various statistical analyses can be performed. To calculate the mean and the standard deviation of an imputed variable (y1), the analyst first creates two empty vectors of `means` and `sds`, and repeats the calculations *M* times by the `for` loop. Typing `means` and `sds` returns *M* values of the means and the standard deviations.

```
means<-c(NA); sds<-c(NA)
for(k in 1:M){
  means[k]<-mean(imp2[,k])
  sds[k]<-sd(imp2[,k])
}
```

To calculate a combined point estimate, the analyst simply takes the average by equation (5.7). Furthermore, by calculating the standard deviation of `means`, i.e. `sd(means)`, the analyst can estimate the amount of estimation uncertainty due to imputation as a confidence interval.

```
mean(means)            #Combined Point Estimate of Mean
mean(sds)              #Combined Point Estimate of Std. Dev.
sd(means)              #Estimation Uncertainty
mean(means)+2*sd(means) #Confidence Interval Upper Limit
mean(means)-2*sd(means) #Confidence Interval Lower Limit
```

Let us again use the example data in Figure 6.1. In our specific case, the combined point estimate of the means is 6.126, with the combined point estimate of standard deviation 0.868. Estimation uncertainty is measured by `sd(means)`, which is the standard deviation of the *M* means, or the standard error of the estimated *M* means. In our case, it is 0.013. Therefore, the analyst can be approximately 95% confident that the true mean of complete data is somewhere between 6.101 and 6.151, after taking the error due to missingness into account.

6.4.2 Regression of y2 on y1

Suppose that y2 is the dependent variable and y1 is the explanatory variable in regression. To estimate the regression coefficients and the associated standard errors, the analyst first creates four empty vectors, `reg1`, `reg2`, `reg3`, and `reg4`. The `for` loop repeats the estimation of regression models *M* times. The results are stored in `summary(model)$coefficients[i]`, where i= 1 and 3 are regression coefficients and i = 2 and 4 are standard errors.

```
reg1<-c(NA); reg2<-c(NA); reg3<-c(NA); reg4<-c(NA)
for(k in 1:M){
  model<-lm(data[,2]~data[,k+2])
  reg1[k]<-summary(model)$coefficients[1]
  reg2[k]<-summary(model)$coefficients[2]
  reg3[k]<-summary(model)$coefficients[3]
  reg4[k]<-summary(model)$coefficients[4]
}
```

After the analysis stage is complete, there are *M* values of outputs. Using equations (5.7) and (5.8), the results are combined as follows.

```
intercept<-mean(reg1)               #Combined Intercept
  WV1<-mean(reg3^2)                 #Within-Imputation Variance
  BV1<-sum((reg1-intercept)^2)/(M-1) #Between-Imputation Variance
  TV1<-WV1+(1+1/(M))*BV1            #Total Variance
  TSE1<-sqrt(TV1)                   #Total Std. Error
  tstat1<-intercept/TSE1            #t-statistics for Intercept
slope<-mean(reg2)                   #Combined Slope
  WV2<-mean(reg4^2)                 #Within-Imputation Variance
  BV2<-sum((reg2-slope)^2)/(M-1)    #Between-Imputation Variance
  TV2<-WV2+(1+1/(M))*BV2            #Total Variance
  TSE2<-sqrt(TV2)                   #Total Std. Error
  tstat2<-slope/TSE2                #t-statistics for Slope
```

Let us again use the example data in Figure 6.1. In our specific case, the combined point estimate of the regression intercept is 8.231, with the total standard error of 2.512. Thus, the *t*-statistic for the intercept is 3.277. The combined point estimate of the regression slopes is 0.273 with the total standard error of 0.407. Thus, the *t*-statistic for the slope is 0.671.

6.5 Conclusion

This chapter outlined how to implement multiple ratio imputation in *R*, which can be easily copied and pasted into *R* for use (See Appendix 6.1). These codes allow us not only to estimate multiple ratio imputation, but also to statistically analyze imputed data by multiple ratio imputation. Therefore, this will be a valuable addition to the choice for imputation techniques.

However, the code described in this chapter is only a first step toward implementing multiple ratio imputation; thus, the code is expected to be updated so as to maximize computational efficiency and to expand the scope of data that can be handled. Furthermore, the EMB algorithm is a general approach composed of the EM algorithm and nonparametric bootstrapping. Therefore, multiple ratio imputation can be implemented not only in *R*, but also in other statistical environments. Also, multiple ratio imputation is not limited to the EMB algorithm. Depending on the nature of imputation, multiple ratio imputation may be implemented by way of other multiple imputation algorithms, such as MCMC and Fully Conditional Specification (FCS) (van Buuren, 2012).

Appendix 6.1: Software MrImputation

Software *MrImputation* (version 1.0.0), which stands for *m*ultiple *r*atio *imputation*, is a collection of *R*-functions this chapter explained step by step. This appendix combines each of the steps as a set of *R*-functions `mrimpute` and `mranalyze`.

Appendix 6.1.1: User Manual

Copy the following codes into the *R* script and save them as `mrimpute.R` and `mranalyze.R` on the computer. After reading an appropriate data file in *R*, use function `source` to read these functions as follows.

```
source("mrimpute.R")
source("mranalyze.R")
```

Description of `mrimpute`: This function performs the imputation stage of multiple ratio imputation and produces multiply-imputed data named mridata.csv.

Usage: `mrimpute(data = data, M = 100, seed = 1223, log = FALSE, zero = FALSE, outdata = TRUE)`

Arguments:

`data`	A data frame that contains the incomplete variable targeted for imputation. The imputer can specify any name of the data to be used.
`M`	The number of multiply-imputed datasets. The imputer can set any number.
`seed`	Random number seed value. Any number can be specified.
`log`	An option to log-transform the data. The default is FALSE. If log-transformation is optimal, then this option should be set to TRUE.
`zero`	An option to suppress negative values to zero. The default is FALSE. If negative imputed values are unacceptable, this option should be set to TRUE.
`outdata`	An option to save the imputed data as a csv file. The default is TRUE.

Description of `mranalyze`: This function performs the analysis stage. It returns the mean and the standard deviation of the imputed variable. It can also return the result of regression analysis of y2 on y1 if `reg=TRUE`.

Usage: `mranalyze(data, reg = FALSE)`

Arguments:

`data`	The mridata.csv created by mrimpute.
`reg`	An option to perform regression analysis. The default is FALSE. If the analyst wants to see the result of regression analysis, this option should be set to TRUE.

Appendix 6.1.2: *R*-Function mrimpute: Imputation Stage

```
mrimpute<-function(data,M,seed,outdata=TRUE,log=FALSE,zero=FALSE){
data<-data; M<-M; seed<-seed; set.seed(seed)
if(log){data<-log(data)}
sampleframe<-matrix(sample(nrow(data),nrow(data)*M,
             replace=TRUE),nrow=nrow(data),ncol=M)
datasub<-as.list(rep(NA,M))
for(i in 1:M){datasub[[i]]<-as.matrix(data[sampleframe[,i],])}
suppressMessages(suppressWarnings(require(norm)))
p<-ncol(data); para<-p*(p+3)/2+1; thetahat<-matrix(NA,M,para)
emmu<-matrix(NA,M,p)
for(i in 1:M){thetahat[i,]<-em.norm(prelim.norm(datasub[[i]]),
                             showits=FALSE,maxits=1000)
              emmu[i,]<-getparam.norm(prelim.norm(datasub[[i]]),
                             thetahat[i,],corr=FALSE)$mu}
imp0<-as.list(rep(NA,M)); imp<-matrix(NA,nrow(data),M)
resid<-matrix(NA,nrow(data),M); e<-matrix(NA,nrow(data),M)
imp1<-matrix(NA,nrow(data),M); beta<-matrix(NA,M,ncol(data)-1)
beta<-emmu[,1]/emmu[,2]
for(i in 1:M){imp[,i]<-beta[i]*data[,2]}
for(i in 1:M){resid[,i]<-data[,1]-imp[,i]
              e[,i]<-rnorm(nrow(data),0,sd(resid[,i],na.rm=TRUE))
              imp1[,i]<-imp[,i]+e[,i]}
imp2<-matrix(NA,nrow(data),M)
for(i in 1:M){imp2[,i]<-data[,1]}
for(i in 1:M){
       for(j in 1:nrow(data)){
              if (is.na(data[j,1])=="TRUE"){
              imp2[j,i]<-imp1[j,i]
              }else{
              imp2[j,i]<-data[j,1]}
  }}
if(log){imp2<-exp(imp2);data<-exp(data)}
if(zero){imp2[which(imp2<0)]<-0}
impdata<-data.frame(data, imp2)
  if (outdata){
   write.csv(impdata,"mridata.csv",row.names=FALSE)
    }
}
```

Appendix 6.1.3: *R*-Function mranalyze: Analysis Stage

```
mranalyze<-function(data,reg=FALSE){
data<-data; M<-ncol(data)-2; means<-c(NA); sds<-c(NA)

for(k in 1:M){
  means[k]<-mean(data[,k+2])
  sds[k]<-sd(data[,k+2])
}
meanimp<-mean(means);BISD<-sd(means);UL<-mean(means)+2*sd(means);LL<-
mean(means)-2*sd(means);sd<-mean(sds)
outmatrix1<-matrix(c(meanimp, sd, BISD, UL, LL))
colnames(outmatrix1)<-"Summary"
rownames(outmatrix1)<-c("mean","sd","BISD","95%CIUL","95%CILL")

if(reg){
reg1<-c(NA); reg2<-c(NA); reg3<-c(NA); reg4<-c(NA)
for(k in 1:M){
  model<-lm(data[,2]~data[,k+2])
  reg1[k]<-summary(model)$coefficients[1]
  reg2[k]<-summary(model)$coefficients[2]
  reg3[k]<-summary(model)$coefficients[3]
  reg4[k]<-summary(model)$coefficients[4]
}

intercept<-mean(reg1)
WV1<-mean(reg3^2)
BV1<-sum((reg1-intercept)^2)/(M-1)
TV1<-WV1+(1+1/(M))*BV1
TSE1<-sqrt(TV1)
tstat1<-intercept/TSE1

slope<-mean(reg2)
WV2<-mean(reg4^2)
BV2<-sum((reg2-slope)^2)/(M-1)
TV2<-WV2+(1+1/(M))*BV2
TSE2<-sqrt(TV2)
tstat2<-slope/TSE2

outmatrix2<-matrix(c(intercept, TSE1, tstat1, slope, TSE2, tstat2))
colnames(outmatrix2)<-"Regression"
rownames(outmatrix2)<-c("intercept","TSE(intercept)" ,"t-
Stat(intercept)","slope","TSE(slope)" ,"t-Stat(slope)")
}

if(reg){
result<-list(outmatrix1, outmatrix2)
  return(result)
}else{
result<-list(outmatrix1)
  return(result)
}
}
```

7 Conclusion

This dissertation was about how to deal with missing data in official economic statistics. Chapter 2 unveiled the current practice among the UNECE member states and found that ratio imputation was often used in official economic statistics. Furthermore, it proposed multiple imputation as a suitable imputation method for public-use microdata. Chapter 3 gave a unifying approach to ratio imputation with a novel way of identifying an appropriate ratio imputation model based on the magnitude of heteroskedasticity. Chapter 4 compared the existing three multiple imputation algorithms and found that the EMB algorithm would be more useful than the MCMC-based methods. Chapter 5 presented a novel application of the EMB algorithm to create multiple ratio imputation and demonstrated its usefulness by testing it against traditional methods using a variety of simulation data. Chapter 6 provided brand-new software for multiple ratio imputation. The author believes that these findings will be important additions to the literature of missing data in particular and official statistics in general.

Future research may deal with the following issues. The method proposed in Chapter 3 is still a starting point to determine the value of θ. Following the idea of Tukey's boxplot, the method in Chapter 3 divided the data into four groups based on the five number summaries. Preliminary research showed that if the data were divided into ten groups (instead of four groups), the results were not as good as those of the proposed methods. However, the appropriate number of groups may be a function of the number of observations. This issue should be further investigated in future research. Also, an analytical method may be possible by taking the logarithm of residuals. Future research should develop this analytical method, and should test it against the proposed method of this dissertation. Furthermore, ratio imputation in this dissertation is bivariate by definition. Even when many auxiliary variables are available, the model can only use one auxiliary variable. Following Olkin (1958), future research should develop multivariate ratio imputation.

References

[1] Abayomi, K., Gelman, A., and Levy, M. (2008). Diagnostics for multivariate imputations. *Applied Statistics*, 57(3), 273-291.

[2] Abe, T. and Iwasaki, M. (2007). Evaluation of statistical methods for analysis of small-sample longitudinal clinical trials with dropouts. *Journal of the Japanese Society of Computational Statistics*, 20(1), 1-18.

[3] Abe, T. (2016). *Kessoku Data no Toukei Kaiseki* (*Statistical Analysis of Missing Data*). Tokyo: Asakura Shoten.

[4] Acemoglu, D., Johnson, S., and Robinson, J. A. (2005). Institutions as the fundamental cause of long-run growth, in *Handbook of Economic Growth* edited by P. Aghion and S. Durlauf, North Holland: Elsevier.

[5] Allison, P. D. (2002). *Missing Data*. Thousand Oaks, CA: Sage Publications.

[6] Baraldi, A. N., and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.

[7] Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948-955.

[8] Barro, R. J. (1997). *Determinants of Economic Growth: A Cross-Country Empirical Study*. Cambridge, MA: MIT Press.

[9] Bechtel, L., Gonzalez, Y., Nelson, M., and Gibson, R. (2011). Assessing several hot deck imputation methods using simulated data from several economic programs. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 5022-5036.

[10] Blackwell, M., Honaker, J, and King, G. (2015). A unified approach to measurement error and missing data: Details and extensions. *Sociological Methods and Research*, in press.

[11] Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15, 651-675.

[12] Carpenter, J. R., and Kenward, M. G. (2013). *Multiple Imputation and its Application*. Chichester, West Sussex: A John Wiley and Sons Publication.

[13] Carsey, T. M. and Harden, J. J. (2014). *Monte Carlo Simulation and Resampling Methods for Social Science*. Thousand Oaks, CA: Sage Publications.

[14] Central Intelligence Agency. (2016). *The World Factbook*. Available at https://www.cia.gov/library/publications/the-world-factbook/index.html [Last accessed November 27, 2016].

[15] Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, 13(2), 53-75.

[16] Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York, NY: John Wiley and Sons.

[17] Cranmer, S. J. and Gill, J. (2013). We have to be discrete about this: A non-parametric imputation technique for missing categorical data. *British Journal of Political Science*, 43(2),

425-449.

[18] de Waal, T., Pannekoek, J., and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley and Sons.

[19] DeGroot, M. H., and Schervish, M. J. (2002). *Probability and Statistics*, 3rd edition. Boston, MA: Addison-Wesley.

[20] Deng, Y., Chang, C., Ido, M. S., and Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific Reports*, 6(21689), 1-10.

[21] Di Zio, M., and Guarnera, U. (2013). Contamination model for selective editing. *Journal of Official Statistics*, 29(4), 539-555.

[22] Do, C. B., and Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, 26(8), 897-899.

[23] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, 1087-1091.

[24] Egghe, L. (2012). Averages of ratios compared to ratios of averages: Mathematical results. *Journal of Informetrics*, 6(2), 307-317.

[25] Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, 25(3), 76-80.

[26] Enders, C. K. (2010). *Applied Missing Data Analysis*. New York, NY: The Guilford Press.

[27] Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2016). *Penn World Table 9.0*. Available at: http://www.rug.nl/research/ggdc/data/pwt/pwt-9.0 [Last accessed November 3, 2016].

[28] Feng, Y. (2003). *Democracy, Governance, and Economic Performance: Theory and Evidence*. Cambridge, MA: The MIT Press.

[29] Fox, J. (2015). *Package 'Norm'*. Available at: http://cran.r-project.org/web/packages/norm/norm.pdf [Last accessed May 31, 2017].

[30] Freedom House. (2016). *Freedom in the World 2016*. Available at: https://freedomhouse.org/report/freedom-world/freedom-world-2016 [Last accessed November 30, 2016].

[31] Gill, J. (2008). *Bayesian Methods: A Social and Behavioral Sciences Approach*, Second Edition. London: Chapman and Hall/CRC.

[32] Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.

[33] Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3), 206-213.

[34] Greene, W. A. (2003). *Econometric Analysis*, 5th edition. Upper Saddle River, NJ: Prentice Hall.

[35] Gujarati, D. N. (2003). *Basic econometrics*, 4th edition. New York, NY: McGraw-Hill.

[36] Gupta, A. K. and Kabe, D. G. (2011). *Theory of Sample Surveys*. Singapore: World Scientific.

[37] Hardt, J., Herke, M., and Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology,* 12(184), 1-13.

[38] Hoenig, J. M., Jones, C. M., Pollock, K. H., Robson, D. S., and Wade, D. L. (1997). Calculation of Catch Rate and Total Catch in Roving Surveys of Anglers. *Biometrics* 53(1), 306-317.

[39] Honaker, J., King, G., and Blackwell, M. (2016). *Package 'Amelia'*. Available at: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf [Last accessed November 30, 2016].

[40] 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.

[41] Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7), 1-47.

[42] Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman and E. Leamer (Eds), *Handbook of Econometrics* (pp.3160-3228), vol.5. Amsterdam: Elsevier.

[43] Horton, N. J. and Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, 61(1), 79-90.

[44] Horton, N. J. and Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55(3), 244-254.

[45] Hothorn, T., Zeileis, A., Farebrother, R.W., Cummins, C., Millo, G., and Mitchell D. (2015). *Package 'lmtest'*. Available at: https://cran.r-project.org/web/packages/lmtest/lmtest.pdf [Last accessed July 6, 2016].

[46] Hu, M., Salvucci, S., and Lee, R. (2001). *A Study of Imputation Algorithms.* Working Paper No. 2001–17. U.S. Department of Education. National Center for Education Statistics. Available at: http://nces.ed.gov/pubs2001/200117.pdf [Last accessed May 31, 2017].

[47] Hughes, R. A., Sterne, J. A. C., and Tilling, K. (2016). Comparison of imputation variance estimators. *Statistical Methods in Medical Research*, 25(6), 2541-2557.

[48] Imai, K., King, G. and Lau, O. (2008). Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics*, 17(4), 892-913.

[49] Ito, S. and Hoshino, N. (2014). Effectiveness of data swapping based on the microdata from population census. *Statistics*, (107), 1-16.

[50] Iwasaki, M. (2002). *Fukanzen Data no Toukei Kaiseki* (*Foundations of Incomplete Data Analysis*). Tokyo: EconomistSha Publications, Inc.

[51] Jacoby, W. G. (1991). *Data Theory and Dimensional Analysis*. Thousand Oaks, CA: Sage Publications.

[52] Jacoby, W. G. (1999). Levels of measurement and political research: An optimistic view. *American Journal of Political Science*, 43(1), 271-301.

[53] Joenssen, D. W. (2015). *HotDeckImputation: Hot Deck Imputation Methods for Missing Data*, Version 1.1.0. Available at: https://cran.r-project.org/web/packages/HotDeckImputation/index.html [Last accessed May 31, 2017].

[54] 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.

[55] Kropko, J., Goodrich, B., Gelman, A., and Hill, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis*, 22(4), 497-519.

[56] Kurihara, Y. (2015). Estimation precision of statistical matching and selection effects of common variables. *Statistics*, (108), 1-15.

[57] Kyrillidou, M., Morris, S., and Roebuck, G. (2015). *ARL Statistics 2013-2014*. Washington, D.C.: Association of Research Libraries.

[58] Larivière, V. and Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, 5(3), 392-399.

[59] Lee, H., Rancourt, E., and Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10(3), 231-243.

[60] Lee, K. J. and Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171(5), 624-632.

[61] Lee, K. J. and Carlin, J. B. (2012). Recovery of information from multiple imputation: A simulation study. *Emerging Themes in Epidemiology*, 9(3), 1-10.

[62] Leite, W., and Beretvas, S. (2010). The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods*, 9(1), 64-74.

[63] Leon, S. J. (2006). *Linear Algebra with Applications*, 7th edition. Upper Saddle River, NJ: Pearson/Prentice Hall.

[64] Li, F., Yu, Y., and Rubin, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. *Duke University Department of Statistical Science Discussion Paper*, 11(14), 1-35.

[65] Liang, H., Su, H., and Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in AIDS study. *Computational Statistics and Data Analysis*, 53(2), 546-553.

[66] Little, R. J. A. (1992). With missing X's: A review. *Journal of the American Statistical Association*, 87(420), 1227-1237.

[67] Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, second edition. Hoboken, NJ: John Wiley and Sons.

[68] Liu, L., Yujuan, T., Yingfu, L. and Zou, G. (2005). Imputation for missing data and variance estimation when auxiliary information is incomplete. *Model Assisted Statistics and Applications* 1(2), 83-94.

[69] Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*.

Thousand Oaks, CA: Sage Publications.

[70] McNeish, D. (2017). Missing data methods for arbitrary missingness with small samples. *Journal of Applied Statistics*, 44(1), 24-39.

[71] Mooney, C. Z. (1997). *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications.

[72] Nakamura, H. and Hirasawa, K. (2016). Kouteki Toukei no Nijiteki Riyou no Sokushin ni Kansuru Waga Kuni no Torikumi Joukyou. *Proceedings of the 60th (2016) Conference of Japan Economic Society of Statistics*, 36-37.

[73] Office for National Statistics. (2014). Change to imputation method used for the turnover question in monthly business surveys. *Guidance and methodology: retail sales*. Available at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/retail-sales/index.html [Last accessed May 31, 2017].

[74] Olkin, I. (1958). Multivariate ratio estimation for finite populations. *Biometrika*, 45(1/2), 154-165.

[75] Ono, K. and Ikawa, T. (2015). *Monte Carlo hou Nyuumon* (*Introduction to Monte Carlo Methods*). Tokyo: Kinyuu Zaisei Jijou Kenkyuukai.

[76] Poston, D., and Conde, E. (2014). Missing data and the statistical modeling of adolescent pregnancy. *Journal of Modern Applied Statistical Methods*, 13(2), 464-478.

[77] Raghunathan, T. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press.

[78] Rao, T. J., (2002). Mean of ratios or ratio of means or both? *Journal of Statistical Planning and Inference*, 102(1), 129-138.

[79] Ross, S. (2006). *A First Course in Probability*, 7th edition. Upper Saddle River, NJ: Pearson/Prentice Hall.

[80] Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2), 377-387.

[81] Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

[82] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons.

[83] Sakata, S. (2006). Kohyou data to toukei riyou (Individual raw data and its application in statistical analyses). *Statistics*, (90), 31-42.

[84] Schafer, J. L. (2016). *Package 'norm2'*. Available at: https://cran.r-project.org/web/packages/norm2/norm2.pdf [Last accessed November 30, 2016].

[85] Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.

[86] Schafer, J. L. and Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571.

[87] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman and Hall/CRC.

[88] Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association*, 101(475), 924-933.

[89] Scheuren, F. (2005). Multiple imputation: How it began and continues. *The American Statistician*, 59(4), 315-319.

[90] Seaman, S., Galati, J., Jackson, D., and Carlin, J. (2013). What is meant by "Missing at Random"? *Statistical Science*, 28(2), 257-268.

[91] Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology*, 26(1), 79-85.

[92] Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY: Springer.

[93] Shara, N., Yassin, S. A., Valaitis, E., Wang, H., Howard, B. V., Wang, W., Lee, E. T., and Umans, J. G. (2015). Randomly and non-randomly missing renal function data in the strong heart study: A comparison of imputation methods. *PLOS ONE*, 10(9), 1-11.

[94] Snowdon, P. (1992). Ratio methods for estimating forest biomass. *New Zealand Journal of Forestry Science*, 22(1), 54-62.

[95] Statistics Bureau of Japan. (2012). *Economic Census for Business Activity*. Available at: http://www.stat.go.jp/english/data/e-census/2012/index.htm [Last accessed May 31, 2017].

[96] Stuart, E. A., Azur, M., Frangakis, C., and Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, 169(9), 1133-1139.

[97] Takahashi, M. (2017a). Missing data treatments in official statistics: Imputation methods for aggregate values and public-use microdata. *Statistics*, (112), 65-83.

[98] Takahashi, M. (2017b). Multiple ratio imputation by the EMB algorithm: Theory and simulation. *Journal of Modern Applied Statistical Methods*, 16(1), 630-656.

[99] Takahashi, M. (2017c). Implementing multiple ratio imputation by the EMB algorithm (R). *Journal of Modern Applied Statistical Methods*, 16(1), 657-673.

[100] Takahashi, M. (2017d). Statistical inference in missing data by MCMC and non-MCMC multiple imputation algorithms: Assessing the effects of between-imputation iterations. *Data Science Journal*, in press.

[101] Takahashi, M. and Ito, T. (2013a). Imputing missing values of turnover in economic surveys: Assessment of multiple imputation. *Research Memoir of Official Statistics*, (70), 19-86.

[102] Takahashi, M., and Ito, T. (2013b). Multiple imputation of missing values in economic surveys: Comparison of competing algorithms. *Proceedings of the 59th World Statistics Congress of the International Statistical Institute (ISI)*, 3240-3245.

[103] Takahashi, M. and Ito, T. (2014). Comparison of competing algorithms of multiple imputation: Analysis using large-scale economic data. *Research Memoir of Official Statistics*, (71), 39-82.

[104] Takahashi, M., Abe, Y., and Noro, T. (2015). Kouteki toukei ni okeru kessokuchi hotei no kenkyuu: Tajuu dainyuu hou to tanitsu dainyuu hou (Research on the imputation of missing values in official statistics: Multiple imputation and single imputation). *Seihyou Gijutsu*

Sankou Shiryou (*NSTAC Working Paper*), (30), 1-95.

[105] Takahashi, M., Iwasaki, M. and Tsubaki, H. (2017). Imputing the mean of a heteroskedastic log-normal missing variable: A unified approach to ratio imputation. *Statistical Journal of the IAOS*, 33(3), in press.

[106] Takai, K., Hoshino, T., and Noma, H. (2016). *Kessoku Data no Toukei Kagaku: Igaku to Shakai Kagaku heno Ouyou* (*Statistical Science in Missing Data: Application to Medical and Social Sciences*). Tokyo: Iwanami Shoten.

[107] Thompson, K. J., and Washington, K. T. (2012). A response propensity based evaluation of the treatment of unit nonresponse for selected business surveys. *Federal Committee on Statistical Methodology 2012 Research Conference*. Available at: https://fcsm.sites.usa.gov/files/2014/05/Thompson_2012FCSM_III-B.pdf [Last accessed May 31, 2017].

[108] U.S. Bureau of the Census. (1957). *U.S. Census of Manufactures 1954*, Vol.II, Industry Statistics, Part 1, General Summary and Major Groups 20 to 28. Washington, D. C.: U.S. Government Printing Office.

[109] van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC.

[110] van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.

[111] van Buuren, S., and Groothuis-Oudshoorn, K. (2015). *Package 'mice'*. Available at: http://cran.r-project.org/web/packages/mice/mice.pdf [Last accessed May 31, 2017].

[112] von Hippel, P. T. (2016). New confidence intervals and bias comparisons show that maximum likelihood can beat multiple imputation in small samples. *Structural Equation Modeling*, 23(3), 422-437.

[113] Weiss, N. A. (2005). *Introductory Statistics*, 7th edition. Boston, MA: Pearson/Additson Wesley.

[114] Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*, 4th edition. Mason, OH: South-Western.

[115] Zarnoch, S. J. and Bechtold, W. A. (2000). Estimating mapped-plot forest attributes with ratios of means. *Canadian Journal of Forest Research*, 30 (5), 688-697.

[116] Zhu, J. and Raghunathan, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the American Statistical Association*, 110(511), 1112-1124.

[117] Zou, G. H., Li, Y. F., Zhu, R., and Guan, Z. (2010). Imputation of mean of ratios for missing data and its application to PPSWR sampling. *Acta Mathematica Sinica, English Series*, 26(5), 863-874.

Curriculum Vitae

Education

Ph.D., Science and Technology, Seikei University, Expected 2017

A.B.D. (Ph.D. Candidate), Political Science, Michigan State University, 2009

Examination Fields: Political Methodology, Comparative Politics

M.A., Political Science, Michigan State University, 2008

M.A., Political Science, California State University, Los Angeles, 2004

B.A., Philosophy, Keio University, 2001

Employment

Assistant Professor, IR Office, Tokyo University of Foreign Studies, 2016-present

Senior Researcher, Research Division, National Statistics Center, 2011-2016

Professional Service

Committee Member, Open Data Chosa Kenkyuukai, Transdisciplinary Federation of Science and Technology, 2016-present

Visiting Research Fellow, Institute of Economic Research, Chuo University, 2016-present

Committee Member, CBT for Japan Statistical Society Certificate, Center for Japan Statistical Society Certificate, 2015-present

Part-time Lecturer, Department of Economics, Toyo University, 2015-2016

Part-time Lecturer, College of Business, Rikkyo University, 2014-2015

Teaching Assistant, Department of Political Science, Michigan State University, 2006-2010

Teaching Assistant, ICPSR Summer Program in Quantitative Methods of Social Research, 2008-2009

Award

Graduate Student Assistantship, Michigan State University, 2006-2010

Heiwa Nakajima Foundation Scholarship, 2001-2003