WP. 35
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (v): Software & tools for data editing and imputation.

MULTIPLE IMPUTATION OF TURNOVER IN EDINET DATA: TOWARD THE IMPROVEMENT OF IMPUTATION FOR THE ECONOMIC CENSUS

Invited Paper

Prepared by Masayoshi Takahashi and Takayuki Ito, National Statistics Center[1], Japan

I. Introduction

1. For the first time in Japanese history, the Economic Census for Business Activity was conducted February 2012, covering all enterprises and establishments in Japan. The Economic Census will be an important data source for a variety of economic statistics; however, due to various types of response units, missing values and error will be frequently produced. Therefore, we are engaging in research on data editing strategies, in order to improve the quality of the next Economic Census. In this paper, we describe standard single imputation techniques and their limitations, illustrate the mechanism and advantages of multiple imputation, and introduce R package Amelia (a general-purpose multiple imputation tool). As of writing, the information from the 2012 Economic Census is not yet available; thus, our analysis is based on EDINET[2] data. Our research shows that the fit of multiple imputation is generally better than that of single imputation, and that Amelia will be a useful tool for multiple imputation.

II. Single Imputation Techniques and Their Limitations

2. When missing values exist in a dataset, available data size shrinks and efficiency decreases; furthermore, if there is a systematic difference between respondents and non-respondents, bias is likely to exist (Rubin, 1987, p.1). Therefore, we always need to deal with missing values in one way or another. As a method to deal with missing data, single imputation is often utilized because it is intuitively attractive. In single imputation, we fill in missing values by some type of "predicted" values, such as mean imputation, cold deck imputation, hot deck imputation, and regression imputation (Little and Rubin, 2002, pp.60-61; Ton de Waal *et al.*, 2011, p.230, pp.246-247, p.249). The common problem in single imputation is to replace an unknown missing value by a single value and then treat it as if it were a true value (Rubin, 1987, pp.12-13). As a result, single imputation ignores uncertainty and almost always underestimates the variance. Multiple imputation overcomes this problem, by taking into account both within-imputation uncertainty and between-imputation uncertainty.

III. Multiple Imputation

3. As early as the 1970's, Rubin (1978) proposed the theory of multiple imputation. Let us first introduce the basics of Rubin's multiple imputation (Rubin, 1987, pp.15-22, pp.75-81; Little and Rubin,

[1] The views and opinions expressed in this paper are the authors' own, not necessarily those of the institution.
[2] EDINET stands for **E**lectronic **D**isclosure for **I**nvestors' **NET**work, which is maintained by the Financial Services Agency of the Japanese Government. The data we used cover 3,587 companies listed on the Tokyo Exchange, whose end of term is March 2011. Since there are no missing values in the EDINET turnover data, we can artificially create missing values in this dataset. It is beneficial that we can compare the true and the imputed values.

2002, pp.85-89). In multiple imputation, missing values are replaced by *M* simulated values, where *M* > 1. Conditional on observed data, we construct a posterior distribution of missing data, draw a random sample from this distribution, and create several imputed datasets. In these *M* multiply-imputed datasets, all of the observed values are the same, but the imputed values are different, reflecting the uncertainty about imputation (Schafer, 1999; King *et al.*, 2001, p.53; Gill, 2008, p.324). Then, we conduct standard statistical analysis, separately using each of the *M* multiply-imputed datasets, and combine the results of the *M* statistical analyses in the following manner to calculate a point estimate.[3] Let $\hat{\theta}_m$ an estimate based on the *m*-th multiply-imputed dataset. The combined point estimate $\bar{\theta}_M$ is equation (1).

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

4. The variance of the combined point estimate consists of two parts. Let v_m the estimate of the variance of $\hat{\theta}_m$, $\mathrm{var}(\hat{\theta}_m)$, let $\bar{v}_M$ the average of within-imputation variance, let $\tilde{v}_M$ the average of between-imputation variance, and let T_M the total variance of $\bar{\theta}_M$. Then, the total variance of $\bar{\theta}_M$ is equation (2), where $(1 + 1/M)$ is an adjustment factor because *M* is not infinite.[4] In short, the variance of $\bar{\theta}_M$ takes into account within-imputation variance and between-imputation variance.

$$T_M = \bar{v}_M + \left(1+\frac{1}{M}\right)\tilde{v}_M = \frac{1}{M}\sum_{m=1}^{M} v_m + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}_M\right)^2\right] \tag{2}$$

5. One way to take uncertainty into account is stochastic regression imputation, which adds random noise to single imputation. While stochastic imputation can reflect within-imputation uncertainty, it cannot reflect between-imputation uncertainty. On the other hand, by constructing several imputation models, multiple imputation can reflect not only within-imputation uncertainty, but also between-imputation uncertainty (Gelman and Hill, 2006, p.542).

6. If we simply repeat single imputation *M* times, however, we only obtain the same imputed value *M* times. In order to conduct multiple imputation, we need to construct an appropriate posterior distribution and calculate *M* different imputed values. For this purpose, we need to use some kind of statistical model. The model that is considered most general is a multivariate normal distribution. Also, the assumption of missing at random (MAR) [5] is not required in multiple imputation (Schafer, 1999, p.8), but many algorithms including the one introduced in this paper assume that the missing mechanism is MAR. In the following, we show the multiple imputation model used in R package Amelia (King *et al.*, 2001, pp.53-54; Honaker and King, 2010, pp.576-578).

7. Let D an $n \times p$ dataset (n = sample size, p = number of variables). If no data are missing, D is normally distributed with mean vector μ and variance-covariance matrix Σ, i.e., $D \sim N_p(\mu, \Sigma)$. Assuming a multivariate normal distribution, missing values are linearly imputed.[6] Suppose that D_{ij} is missing (i = observation index and j = variable index). Let $\widetilde{D}_{ij}$ a simulated, imputed value of observation *i* in variable *j*. Let $D_{i,-j}$ all of the observations in row *i*, except variable *j*. An imputed value, $\widetilde{D}_{ij}$, is calculated using

[3] In simple words, we simulate values for missing data based on observed values adding random noise to them, repeat this process several times, and use the average of these multiply-imputed values as the final product (Shadish, Cook, and Campbell, 2002, p.337). By taking the average over the *M* imputed values, we can increase the efficiency of the estimator, compared with that of single imputation (Little and Rubin, 2002, p.86).

[4] If *M* is infinite, $\lim_{M\to\infty}\left(1+\frac{1}{M}\right)\tilde{v}_M = \tilde{v}_M$.

[5] Let $D = \{Y, X\}$, where *Y* is the dependent variable and *X* is explanatory variables. Let *K* a missingness indicator matrix. The dimensions of *K* and *D* are the same, and whenever *D* is observed, *K* takes the value of 1; otherwise, *K* = 0. Also, let D_o observed data and D_k missing data: $D = \{D_o, D_k\}$. The first assumption is Missing Completely At Random (MCAR), where $P(K|D) = P(K)$: *K* is independent of *D*. The second assumption is Missing At Random (MAR), where $P(K|D) = P(K|D_o)$: *K* is independent of D_k. The third assumption is NonIgnorable (NI), where $P(K|D)$ cannot be simplified: *K* is not independent of *D* (Little and Rubin, 2002, pp.11-12, pp.312-313; King *et al.*, 2001, pp.50-51).

[6] However, just as in a regression analysis, by transforming variables, multiple imputation can be generally applied to non-linearly distributed data as well.

equation (3), where ~ means random sampling from an appropriate posterior distribution. $\tilde{\varepsilon}_i$ represents fundamental uncertainty (i.e., within-imputation uncertainty).

$$\widetilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \tag{3}$$

8. Recall that the slope of regression lines is the square root of the covariance of *X* and *Y* divided by the variance of *X* and that the intercept is the difference between the mean of *Y* and the mean of *X* multiplied by the slope. Thus, the information we need to calculate regression coefficients is the mean, variance, and covariance, all of which are included in μ and Σ. If we fully know μ and Σ, we can deterministically calculate the true regression coefficient β based on D_j, and we can deterministically impute missing values, where the likelihood function of complete data is equation (4).

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \tag{4}$$

9. Nevertheless, in reality, missing values exist in a dataset. Let us assume MAR when forming the likelihood of observed data D_o, i.e., $P(K|D) = P(K|D_o)$. Let $D_{i,o}$ an observed value of row *i* in D, let $\mu_{i,o}$ a subvector of μ, and let $\Sigma_{i,o}$ a submatrix of Σ, where $\mu_{i,o}$ and $\Sigma_{i,o}$ do not change over *i*. Since the marginal densities are normal, the likelihood function of observed data D_o is equation (5).

$$L(\mu, \Sigma|D_o) \propto \prod_{i=1}^{n} N\left(D_{i,o}|\mu_{i,o}, \Sigma_{i,o}\right) \tag{5}$$

10. Since μ and Σ are not fully known, we cannot know β with certainty. $\tilde{\beta}$ shows this estimation uncertainty, which represents between-imputation uncertainty. By way of traditional methods, it is not easy to compute equation (5) and to randomly draw μ and Σ from this posterior distribution. In order to solve this problem, we use the EMB algorithm, which we will explain in the next section.

IV. Expectation Maximization with Bootstrapping Algorithm

11. In this paper, we introduce Amelia, which is a general-purpose multiple imputation package in R. Amelia utilizes the Expectation-Maximization (EM) algorithm combined with bootstrapping. In this section, we review the EM algorithm and bootstrapping, and explain the EMB algorithm.

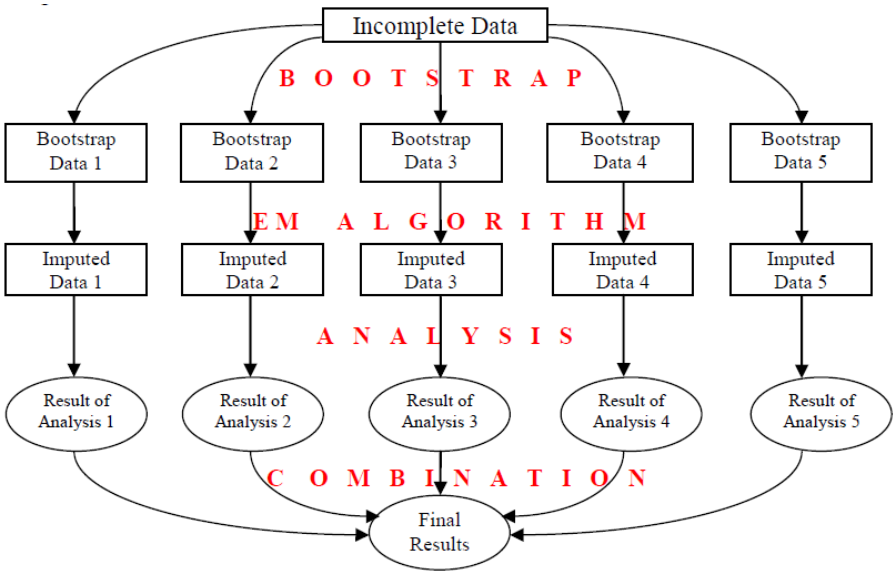
12. When all the information in survey data is not obtained, the entire dataset including the missing part of the data is called incomplete data. To make incomplete data complete, we need information about the distribution of the data, such as the mean and the variance; however, we need to use this incomplete data to estimate the mean and the variance, which is a chicken and egg problem. Therefore, it is not straightforward to analytically solve this problem. As a method to tackle this problem, iterative methods were proposed to estimate such quantities of interest. In the usual EM algorithm application, we temporarily assume a certain distribution and set temporary starting values of the mean and the variance. Based on these temporary values, we calculate an expected value of model likelihood, maximize the likelihood, estimate parameters that maximize the obtained expected values, and update the distribution. After repeating these expectation and maximization steps several times, the value that converged is known to be a maximum likelihood estimate (Watanabe and Yamaguchi, 2000, pp.32-35; Gill, 2008, p.309).

13. Bootstrapping is a resampling method, which is an alternative way to asymptotic approximations. Its objective is to estimate the distribution of parameters without resorting to the first-order asymptotic theory (Wooldridge, 2002, p.379). There are many variations in bootstrapping, but in nonparametric bootstrapping, the observed sample is used as the pseudo-population (Shao, 2002, pp.309-310; Shao and Tu, 1995, pp.9-15; DeGroot and Schervish, 2002, pp.753-763). In other words, a subsample of size *n* is

randomly drawn from this observed sample of size n with replacement, and we repeat this process M times.[7]

14. Figure 4.1 schematically shows multiple imputation, where $M = 5$, using the EMB algorithm. First, there is incomplete data (sample size = n), where q values are observed and $n - q$ values are missing. Using the nonparametric bootstrapping method, a bootstrap subsample of size n is drawn from this incomplete data M times (here, five times). We, then, apply the EM algorithm to each of these M bootstrap subsamples, calculate M point estimates of μ and Σ, impute missing values by M equations of equation (3), and construct M multiply-imputed datasets. Using these M multiply-imputed datasets separately, we conduct statistical analysis and combine the results by equation (1) into the final results (Honaker and King, 2010, p.565; Congdon, 2006, p.504).

Figure 4.1: Multiple Imputation via the EMB Algorithm



Source: Honaker, King, and Blackwell (2011, p.4)

V. R Package Amelia II

15. In the late 1970's, Rubin (1978) proposed the theory of multiple imputation. Despite its theoretical beauty, multiple imputation was computationally challenging, and it had not been used for so long in practice. In a social science field during the late 1990's, about 94% of the journal articles used list-wise deletion to deal with missing data. In light of such reality in the social sciences, a team led by Harvard University Professor Gary King developed general-purpose multiple imputation software, Amelia[8] (King *et al*., 2001). Since then, ten years have passed. Amelia has been used in many social science fields, but in order to accommodate the needs to apply multiple imputation to a gigantic dataset, it was reborn as Amelia II, implemented with the new EMB Algorithm (Honaker and King, 2010).[9] As reported in Honaker and King (2010, p.565), Amelia II can handle 240 variables and 32,000 observations, i.e., 7.68 million variable-observations.

16. There are two assumptions in Amelia II (Honaker, King, and Blackwell, 2011, p.3). First, the theoretical true complete data are a multivariate normal distribution, which is often used as an approximation to the true data distribution. By way of transformation, the distributions of many variables

[7] Generally, the estimate based on bootstrap subsamples is less biased than the estimate based on the original sample. Also, as n and M go to infinity, the variance of bootstrap subsamples becomes a consistent estimate (Little and Rubin, 2002, p.80).

[8] Amelia is named after an American female aviator, Amelia Earhart, who successfully flew over the Atlantic Ocean for the first time in history as a female aviator. In July 1937, during the around-the-world flight, she became missing, and her whereabouts are still considered a mystery. With her fabulous career, she is still a legendary female aviator in the United States of America.

[9] Amelia II can be downloaded from the following websites and can be implemented in R for free: http://gking.harvard.edu/amelia/ or http://cran.r-project.org/web/packages/Amelia/
For software issues on multiple imputation, see Yucel (2011), Schmidt (2009), and Drechsler (2009).

can be made realistically similar to a normal distribution. Second, missingness is either MAR or MCAR. Therefore, if missingness is NI, a specific method is called for. About a specific way to deal with missing data, please refer to King *et al.* (2001, pp.65-66), which is out of the scope in this research.

17. Assuming that Amelia II has already been downloaded and installed on the computer, let us briefly explain Amelia functions.[10] First, start Amelia II, using the `library` function.

```
library(Amelia)
```

18. Next, we should set the starting value, using the `set.seed` function. Amelia II utilizes bootstrapping; thus, if we do not set the seed, imputed values will become always different.[11]

```
set.seed(1223)
```

19. The `amelia` function is used to perform multiple imputation as follows. Here, `a.out` is an arbitrary variable name to store the results of multiple imputation, `data` is the name of the data we are using, and the right hand side of `m =` refers to the number of multiply-imputed datasets.

```
a.out <- amelia(data, m = 5)
```

20. Using the `write.amelia` function, the multiply-imputed datasets can be saved as a csv file.

```
write.amelia(obj = a.out, file.stem = "outdata")
```

21. While it is possible to conduct statistical analysis using the *M* separate datasets by hand, it is fortunate that R has another package called Zelig to perform this task more easily.[12]

```
require("Zelig")
z.out<-zelig(Y~X,data=a.out$imputations,model="ls",cite=F)
summary(z.out)
print(summary(z.out), subset = 1:3)
```

VI. Multiple Imputation of Turnover in EDINET Data

A. Descriptions of Dataset

Table 6.1: Summary Statistics (Raw Data)

Variable	Observation	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Turnover (E)	1222	67	10060	23690	119300	66000	8243000	413242
Worker (E)	1222	3	81	169	419	386	20950	1072
Turnover (I)	571	47	12500	31250	144300	88830	8981000	577050
Worker (I)	571	7	63	133	273	256	7683	557
Turnover (D)	158	230	18420	44800	112200	110200	1154000	202486
Worker (D)	158	6	100	183	394	349	5874	733
Turnover (G)	276	20	2340	6908	55450	17010	3373000	309069
Worker (G)	276	7	76	168	454	433	9783	929
Turnover (L)	191	9	960	4482	26520	12420	1397000	110531
Worker (L)	191	1	25	59	164	133	6284	508

22. The data we used are based on EDINET data (Sectors E = manufacturing, I = retailing, D = construction, G = communication, and L = service). Our dataset has two variables. One is Turnover (unit = million yen), which is our dependent variable. The other is Worker (unit = person), which is our independent variable. Our intuition suggests that as the number of workers increases, the amount of turnover also increases. In the original dataset, there are no missing values in both variables, which means that we know the true values in these variables. For the purpose of experiment, we will artificially create

[10] For more information on Amelia II, see Honaker, King, and Blackwell (2012a) and Honaker, King, and Blackwell (2012b).

[11] Philosophically, random numbers should become different, every time they are generated. However, the retaining of reproducibility is important in scientific analyses; thus, we often retain reproducibility of random numbers by setting the seed. Nonetheless, if we repeatedly use the same "random" number, it is no longer random, so that we have to be careful about how a specific seed affects the results we obtain. If a specific seed generates a skewed dataset, using this seed cannot be recommended. Since Amelia assumes a normal distribution, it can be said that a seed that can generate an exact normal distribution is a good seed. In R, we generated random samples of size 1000, using `set.seed` (1) to `set.seed` (5000), and checked skewness, kurtosis, the Jarque-Bera test, and kernel densities. Among the 5000 seeds, we found that the following can be recommended: 43, 393, 864, 1223, 1403, 1712, 1992, 2725, 2748, and 2902. Our analysis is based on `set.seed` (1223).

[12] For more information on Zelig, see Honaker, King, and Blackwell (2011, pp.35-36) and Imai, King, and Lau (2008).

missing values in the Turnover variable. The models we used are first-order polynomial and natural logarithm transformation. Summary statistics in raw data are presented in Table 6.1.

23. Figures 6.1 and 6.2 show the histograms of Turnover and Worker in raw data for Sector E, respectively. Figure 6.3 shows the scatterplot between Turnover and Worker in raw data for Sector E.

Figure 6.1: Turnover (Raw)

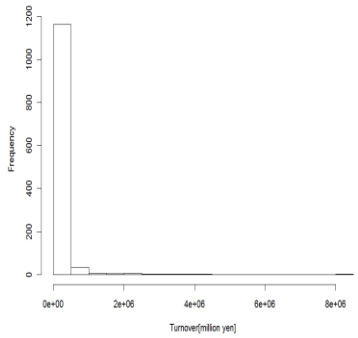


Figure 6.2: Worker (Raw)

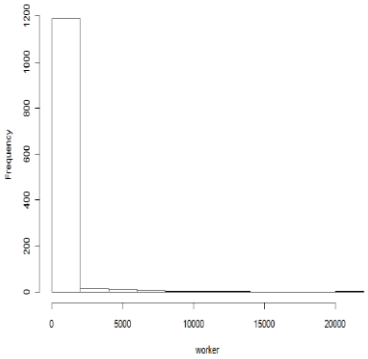
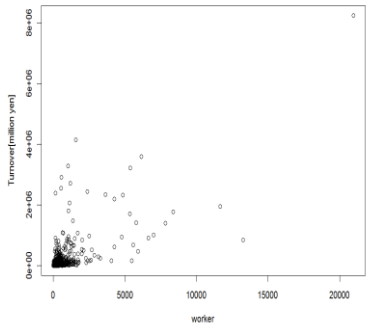


Figure 6.3 (Raw, r = 0.682)



24. Summary statistics in log are presented in Table 6.2.

Table 6.2: Summary Statistics (Natural Log)

Variable	Observation	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Standard Deviation
Turnover (E)	1222	4.204	9.216	10.070	10.220	11.100	15.920	1.553
Worker (E)	1222	1.099	4.394	5.127	5.195	5.955	9.950	1.195
Turnover (I)	571	3.850	9.433	10.350	10.400	11.390	16.010	1.582
Worker (I)	571	1.946	4.139	4.887	4.903	5.545	8.947	1.100
Turnover (D)	158	5.439	9.821	10.710	10.690	11.610	13.960	1.413
Worker (D)	158	1.792	4.600	5.207	5.254	5.856	8.678	1.151
Turnover (G)	276	3.008	7.758	8.840	8.850	9.741	15.030	1.677
Worker (G)	276	1.946	4.327	5.124	5.206	6.071	9.188	1.309
Turnover (L)	191	2.178	6.867	8.407	8.245	9.427	14.150	2.023
Worker (L)	191	0.000	3.219	4.078	4.089	4.887	8.746	1.342

25. Figures 6.4 and 6.5 show the histograms of Turnover and Worker in log for Sector E, respectively. Figure 6.6 shows the scatterplot between Turnover and Worker in log for Sector E.

Figure 6.4: Turnover (Log)

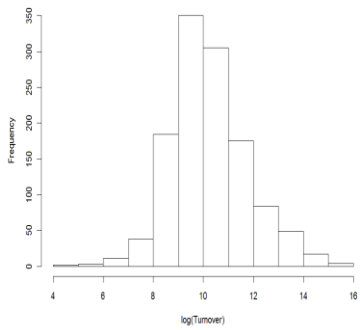


Figure 6.5: Worker (Log)

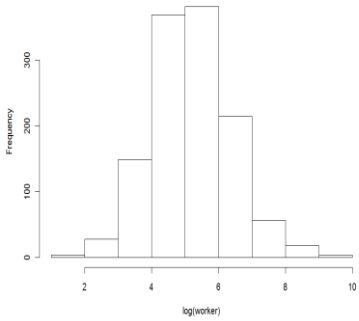
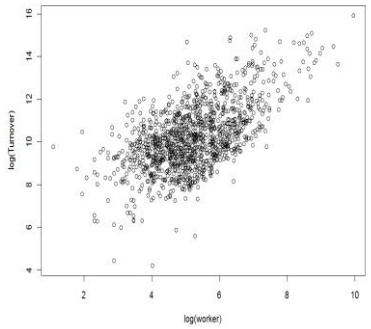


Figure 6.6 (Log, r = 0.593)



B. Assumptions of Missing Mechanism

26. We created the following six patterns of missing mechanisms to cover both MCAR and MAR, but not NI: (1) Completely Random sampling (MCAR); (2) Turnover is missing if worker size is small (MAR); (3) Turnover is missing if worker size is medium (MAR); (4) Turnover is missing if worker size is large (MAR); (5) Turnover is missing if worker size is small or large (MAR); (6) Systematic sampling (MAR). The percentage of missing values in the dataset is 30%, 40%, and 50%. Therefore, there are 18 patterns of missingness in our experiment.

Figure 6.7: MCAR (50%)

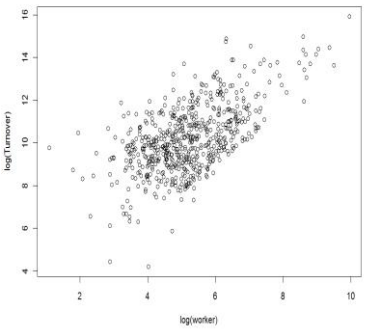


Figure 6.8: MAR (50%, Worker = Small)

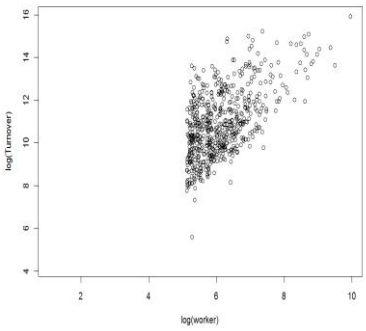
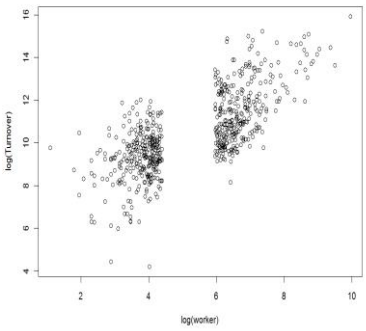


Figure 6.9: MAR (50%, Worker = Medium)



C. Results of Multiple Imputation and Single Imputation

27. We compared the true values of Turnover with the imputed values from multiple imputation and single imputation, respectively. Table 6.3 shows the results of these comparisons. For instance, in sector E (manufacturing), the difference between the true values and the multiply-imputed values is generally (66.7%) less than the difference between the true values and the singly-imputed values. In Sector I (retailing), the difference between the true values and the multiply-imputed values is the same (50%) as the difference between the true values and the singly-imputed values, etc. Overall, in the EDINET data, multiple imputation is closer to the true values than single imputation, 62.5 times out of 100.

Table 6.3

Sector	Multiple Imputation	Single Imputation	Total
E (n = 1222, manufacturing)	66.7%	33.3%	100 %
I (n = 571, retailing)	50.0%	50.0%	100 %
D (n = 158, construction)	10.0%	90.0%	100 %
G (n = 276, communication)	85.7%	14.3%	100 %
L (n = 191, service)	92.3%	7.7%	100 %
Total	62.5%	37.5%	100 %

28. Table 6.4 shows the comparisons across missing patterns. For instance, when the missing pattern is completely random, the difference between the true values and the multiply-imputed values is generally (75%) less than the difference between the true values and the singly-imputed values, etc.

Table 6.4

Missing Pattern	Multiple Imputation	Single Imputation	Total
Completely Random	75.0%	25.0%	100 %
Worker Size Small	75.0%	25.0%	100 %
Worker Size Medium	90.0%	10.0%	100 %
Worker Size Large	66.7%	33.3%	100 %
Worker Size Large & Small	50.0%	50.0%	100 %
Systematic Sampling	20.0%	80.0%	100 %

29. Table 6.5 shows that, in both models (1st order polynomial and logarithm), multiple imputation outperforms single imputation (64.7% and 60.9%, respectively).

Table 6.5

Models	Multiple Imputation	Single Imputation	Total
1st Order Polynomial	64.7%	35.3%	100 %
Natural Log	60.9%	39.1%	100 %

30. Table 6.6 shows that, the higher the rate of missingness, the more multiple imputation outperforms single imputation (55.0%, 69.2%, and 71.4%, respectively).

Table 6.6

Missing Rate	Multiple Imputation	Single Imputation	Total
30%	55.0%	45.0%	100 %
40%	69.2%	30.8%	100 %
50%	71.4%	28.6%	100 %

31. In the following, due to the limited space, we only show the results of **missing patterns (1) and (2)** in the **50% missingness** setting, using **log** transformed data for **Sector E**. The number of multiply-imputed datasets, *M*, is set to 20. For the choice of *M*, see Rubin (1987, p.114) and Schafer (1999, p.7). Our results for multiple imputation are based on the average of the 20 multiply-imputed datasets.

32. In the case of MCAR, the standard deviation of the true values (log) is 1.532. The standard deviation of the multiply-imputed values (log) is 1.525, while the standard deviation of the singly-imputed values (log) is 0.889. Therefore, the true standard deviation is better estimated by multiple imputation than by single imputation. Figure 6.10 shows the scatterplot between Turnover and Worker, including both observed values and singly-imputed values (red circles). Figures 6.11a and 6.11b show the scatterplots between Turnover and Worker, including both observed values and multiply-imputed values (red circles) for m = 1 and m = 10. As visible, the singly-imputed values form a single, straight line, which means that it underestimates the variance, while the multiply-imputed values are scattered just as the true values shown in Figure 6.6.

Figure 6.6:
True Scatterplot in Log

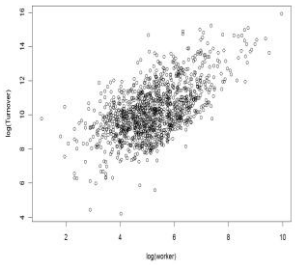


Figure 6.10:
Single Imputation (MCAR)

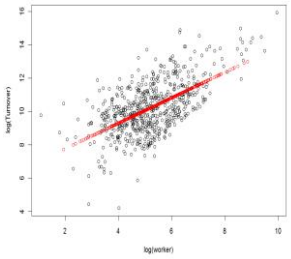


Figure 6.11a:
Multiple Imputation (m = 1)

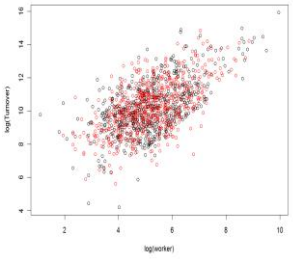
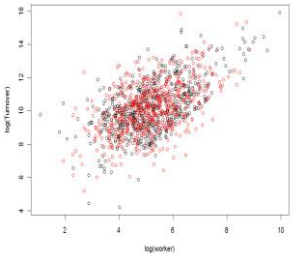


Figure 6.11b:
Multiple Imputation (m = 10)



33. In the above, we compared the true values and the imputed values. For the purpose of experiment, this is the most accurate way of evaluating the performance of imputation models; however, in reality, true values are always unknown. Therefore, the fit of imputation models can never be directly tested in reality. As a result, the diagnostic methods in imputation had long been ignored. Nonetheless, Abayomi, Gelman, and Levy (2008) show that the fit of imputation models and the assumptions of missing mechanisms can be indirectly tested. Amelia II supplies the "comparing densities" function, the "overimpute" function, the "overdispersed starting values" function, and the "missingness map" function (Honaker, King, and Blackwell, 2011, pp.25-35). In the "comparing densities" function, we compare the densities of observed values and imputed values. If the two densities are almost the same, it is likely that the imputation model is free of problems. In the "overimpute" function, we sequentially impute each observed value as if they were missing and then construct 90% confidence intervals. In the "overdispersed starting values" function, we set various starting values for the EM algorithm to see if all of the starting values converge to the same value. In the "missingness" map function, we can visualize the pattern of missingness in the dataset.

34. Suppose that we do not know the true values. We can still diagnose the fit of multiple imputation, using the above mentioned diagnostic techniques. Figure 6.12 compares the densities of observed values and imputed values in the case of missing mechanism (1), i.e., completely random. We can see that the two densities are reasonably similar; thus, we can conclude that the imputation model fits very well. Figure 6.13 shows the missingness map, where Worker is in an increasing order. We see no discernible patterns in Turnover; therefore, we can infer that the missing mechanism is MCAR. Since we created these missing values ourselves, we know that these diagnostics are actually correct. In Figure 6.14, we see that the imputation model is generally captured by the 90% confidence interval. Figure 6.15 shows that all of the starting values for the EM algorithm converged to the same value.

Figure 6.12:
Densities (MCAR)

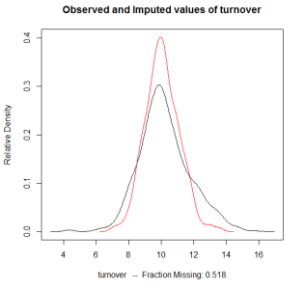


Figure 6.13:
Missingness Map (MCAR)

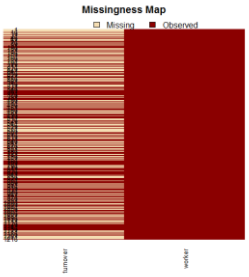


Figure 6.14:
Overimputation (MCAR)

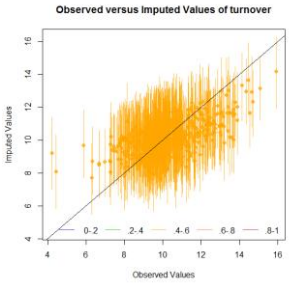
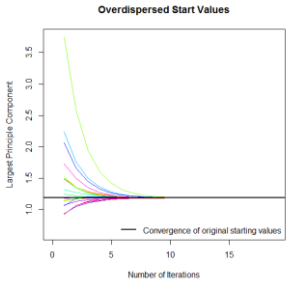


Figure 6.15: Overdispersed Starting Values (MCAR)



35. In the case of missing mechanism (2), i.e., MAR (Worker = Small), the standard deviation of the true values (log) is 1.240. The standard deviation of the multiply-imputed values (log) is 1.426, while the standard deviation of the singly-imputed values (log) is 0.736. Therefore, the true standard deviation is, again, better estimated by multiple imputation than by single imputation. Figure 6.16 shows the scatterplot between Turnover and Worker, including both observed values and singly-imputed values (red circles). Figures 6.17a and 6.17b show the scatterplots between Turnover and Worker, including both observed values and multiply-imputed values (red circles) for m = 1 and m = 10. As visible, the singly-imputed values, once again, form a single, straight line, which means that it underestimates the variance, while the multiply-imputed values are scattered just as the true values shown in Figure 6.6.

Figure 6.6:
True Scatterplot in Log

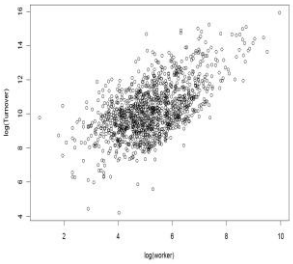


Figure 6.16:
Single Imputation (MAR)

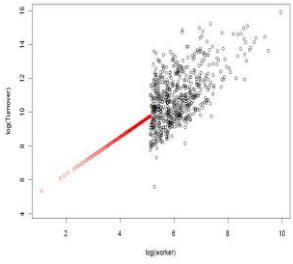


Figure 6.17a:
Multiple Imputation (m = 1)

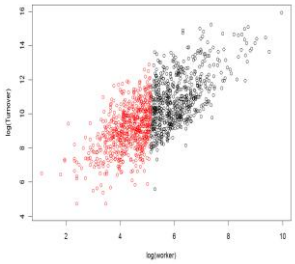
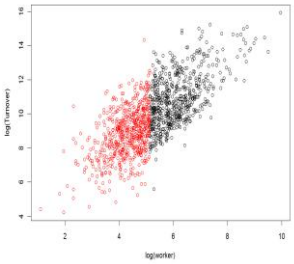


Figure 6.17b:
Multiple Imputation (m = 10)



36. Figure 6.18 compares the densities of observed values and imputed values in the case of MAR. We can see that the two densities are drastically different. When the two densities are different, this does not automatically mean that the imputation model is wrong. This simply implies that we should look into the details to find out why the two densities are different. Figure 6.19 shows the missingness map, where Worker is in an increasing order. We see discernible patterns in Turnover; therefore, we can infer that the missing mechanism is MAR. Thus, we now infer that the density of missing values should be different from the density of observed values, so that Figure 6.18 should not be considered problematic. Again, we know that these diagnostics are, in fact, correct. In Figure 6.20, we see that the imputation model is generally captured by the 90% confidence interval. Figure 6.21 shows that all of the starting values for the EM algorithm converged to the same value.

Figure 6.18:
Densities (MAR)

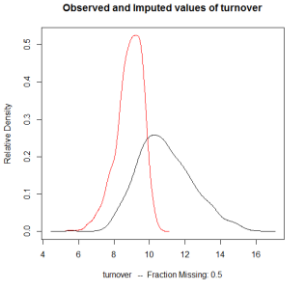


Figure 6.19:
Missingness Map (MAR)

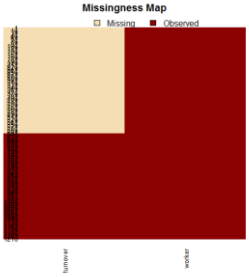


Figure 6.20:
Overimputation (MAR)

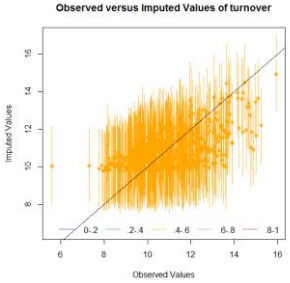
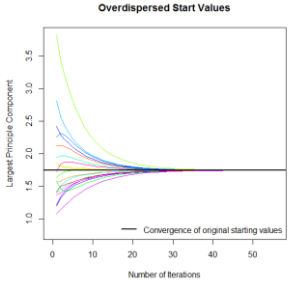


Figure 6.21: Overdispersed Starting Values (MAR)



VII. Conclusions and Future Research

37. This research shows that the overall fit of multiple imputation is excellent, and that it is an efficient method of imputation. Also, we found that R package Amelia II is a useful program for multiple imputation. However, there is no single perfect method to impute missing values, and each imputation method has its own advantages and disadvantages. Even multiple imputation is no exceptions, and if the assumptions are drastically wrong, then the accuracy of imputation cannot be guaranteed. The diagnostic methods in multiple imputation that are shown in this paper are still under development. We diagnosed the *M* combined results, but there is no way of diagnosing each of the *M* multiply imputed datasets, as of writing. In order to guarantee the accuracy of multiply-imputed datasets in practice, for future research, the diagnostic methods should be further developed. Also, in this research, we did not take the existence of outliers in EDINET data into account. However, the accuracy of regression models is largely dependent on the influence of outliers. Thus, in future research, we will consider the impact of outliers on imputation. R package VIM allows us to investigate the missingness structure in a dataset and is expected to be useful for the purpose of diagnostics (Templ, Kowarik, and Filzmoser, 2011).

References

1. Abayomi, Kobi, Andrew Gelman, and Marc Levy. (2008). "Diagnostics for Multivariate Imputations," *Applied Statistics* vol.57, no.3: 273-291.
2. Congdon, Peter. (2006). *Bayesian Statistical Modelling*, Second Edition. West Sussex: John Wiley & Sons Ltd.
3. DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.
4. Drechsler, Jörg. (2009). "Far From Normal - Multiple Imputation of Missing Values in a German Establishment Survey," *Work Session on Statistical Data Editing, Conference of European Statisticians*, Neuchâtel, Switzerland, 5-7 October 2009.
5. Financial Services Agency, the Japanese Government. (2011). *EDINET-Electronic Disclosure for Investors' NETwork*, (Accessed on April 13, 2012), http://info.edinet-fsa.go.jp.
6. Gelman, Andrew, and Jennifer Hill. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
7. Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.
8. Honaker, James and Gary King. (2010). "What to do About Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* vol.54, no.2: 561–581.
9. Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.
10. Honaker, James, Gary King, and Matthew Blackwell. (2012a). *Amelia II: A Program for Missing Data* Version 1.6.1. (Accessed on April 9, 2012), http://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf.
11. Honaker, James, Gary King, and Matthew Blackwell. (2012b). *Package 'Amelia'* Version 1.6.1. (Accessed on April 4, 2012), http://cran.r-project.org/web/packages/Amelia/Amelia.pdf.
12. Imai, Kosuke, Gary King, and Olivia Lau. (2008). "Toward A Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics* vol.17, no.4: 1-22.
13. King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* vol.95, no.1: 49-69.
14. Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.
15. Rubin, Donald 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.
16. Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
17. Schafer, Joseph L. (1999). "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* vol.8: 3-15.
18. Schmidt, Katrin. (2009). "Multiple Imputation with Standard Software: First Application Experiences," *Work Session on Statistical Data Editing, Conference of European Statisticians*, Neuchâtel, Switzerland, 5-7 October 2009.
19. Shadish, William R., Thomas D. Cook, and Donald T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
20. Shao, Jun. (2002). "Replication Methods for Variance Estimation in Complex Surveys with Imputed Data," in *Survey Nonresponse* edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little. New York: John Wiley & Sons, pp.303-314.
21. Shao, Jun and Dongsheng Tu. (1995). *The Jackknife and Bootstrap*. New York: Springer.
22. Templ, Matthias, Alexander Kowarik, and Peter Filzmoser. (2011). "Imputation of Complex Data With R-Package VIM: Traditional and New Methods Based on Robust Estimation," *Work Session on Statistical Data Editing, Conference of European Statisticians*, Ljubljana, Slovenia, 9-11 May 2011.
23. Waal, Ton de, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.
24. Watanabe, Michiko, and Kazunori Yamaguchi. (2000). *EM Algorithm to Fukanzen Data no Shomondai (EM Algorithm and the Problems of Incomplete Data)*. Tokyo: Taga Shuppan.
25. Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
26. Yucel, Recai M. (2011). "State of the Multiple Imputation Software," *Journal of Statistical Software* vol.45, no.1.