Evaluating Alternative Methods of Dealing with Missing Observations - An Economic Application

By: Yuko Onozaka
Department of Agricultural and Resource Economics
University of California, Davis
e-mail: onozaka@primal.ucdavis.edu

Paper prepared for the annual meeting of the American Agricultural Economics Association, Long Beach, July 28-31, 2002.

Abstract:

This paper compares methods to remedy missing value problems in survey data. The commonly used methods to deal with this issue are to delete observations that have missing values (case-deletion), replace missing values with sample mean (mean imputation), and substitute a fitted value from auxiliary regression (regression imputation). These methods are easy to implement but have potentially serious drawbacks such as bias and inefficiency. In addition, these methods treat imputed values as known so that they ignore the uncertainty due to 'missingness', which can result in underestimating the standard errors. An alternative method is Multiple Imputation (MI). In this paper, Expectation Maximization (EM) and Data Augmentation (DA) are used to create multiple complete datasets, each with different imputed values due to random draws. EM is essentially maximum-likelihood estimation, utilizing the interdependency between missing values and model parameters. DA estimates the distribution of missing values given the observed data and the model parameters through Markov Chain Monte Carlo (MCMC). These multiple datasets are subsequently combined into a single imputation, incorporating the uncertainty due to the missingness. Results from the Monte Carlo experiment using pseudo data show that MI is superior to other methods for the problem posed here.

I. Introduction

This paper compares methods to remedy missing value problems in survey data. The analysis shows that commonly applied methods such as deleting the observations with missing values can result in bias and inefficiency. The method of Multiple Imputation appears to provide more reliable estimates for imputing missing values.

The increasing interest in valuing environmental goods has created an explosion of data collection and estimation techniques for nonmarket valuation. The most widely used methods of nonmarket valuation, such as the Contingent Valuation Method (CVM) and the Travel Cost Method (TCM), commonly involve some form of survey data collection. Frequently, some people leave questions partially unanswered. Although more rigorous methods to deal with non-responded items are available (e.g., Mitchell and Carson, Hanemann and Kanninen), typically researchers apply "ad-hoc" methods, such as deleting observations with missing values, replacing missing values with sample mean, or imputing with regression estimation. These methods are easy to implement but could be inefficient or cause bias. An alternative method is Multiple Imputation (MI), a method developed by Rubin (Rubin, 1987). MI uses some imputation methods, such as Data Augmentation (DA) to create multiple complete datasets, each with different imputed values due to random draws. These datasets are subsequently combined into a single imputation. The relative advantages of this different approach have not yet been explored in the nonmarket valuation setting.

The purpose of this paper is to compare MI to the ad hoc applications. The analysis shows that MI is superior to other methods, with performance closest to the full model. Case deletion results in significant efficiency loss and small power. Mean imputation had lower R^2 compared to the full model and MI, and causes bias in some of the coefficients. Regression imputation exhibits the worst fit in terms of R^2, and downward bias in standard errors with slight bias in coefficients.

II. Background

Common Methods and Their Drawbacks:

The common practices when facing missing values are (1) deleting observations that have missing values (case deletion), (2) substituting a sample mean for missing items (mean imputation), and (3) substituting a fitted value from auxiliary regression (regression imputation). These are easy to implement but have serious drawbacks. I will give a quick overview on the drawbacks in following. For more complete discussion, see Little and Rubin (1987).

Case deletion is particularly common in practice. However, by throwing away the information in incomplete observations, it is inefficient and also can bias the estimation when data are missing in a systematic manner. Mean imputation, also a common method, can distort the marginal densities of the data and the covariance among variables. The regression method will underestimate the variability of the data by substituting fitted values from the regression. An extension of regression imputation is a method called stochastic regression imputation in which an error term is added to the imputed value. It will reduce the bias somewhat but will still not be able to mimic the variability of the full data. Regression methods in general are sensitive to model specification.

The common drawback in all these methods is that they ignore the uncertainty due to the 'missingness' by treating imputed values as known. As a result, standard errors for the estimated coefficients are underestimated. This will increase the probability of Type I error (Schafer and Olsen, 1998).

Development of Multiple Imputation:

Rubin first proposed the paradigm of MI in late 70's (e.g., Rubin, 1977, 1978), but it was used only by experts since it typically required extensive statistical knowledge and computational tools. However, recent improvement in the power and convenience of personal computers along with the

development of the method of Markov Chain Monte Carlo (MCMC) simulation in the late 90's have made MI more accessible.[1]

Assumptions:

Before going into the details of the MI algorithm, the assumptions necessary for its application are discussed. First are the assumptions on the population distribution and parameter distribution. The common assumptions are normal or logistic distribution.[2] Schafer points out that the normal distribution works well in many discrete cases even when the normal assumption is only approximately true (Schafer, 1997).

The other key assumption is on the mechanism of "missingness" (Rubin, 1976, Little and Rubin, 1987). MI assumes that data is *missing at random* (MAR). This is different from saying that the dataset has no systematic way of missing values, which is called *missing completely at random* (MCAR). MCAR is equivalent of random sampling, where the missingness is independent of both observed and missing data. MAR assumes that the missingness depends on observed data, but is independent of the missing data. Another way to put this is that the datum is a random sample of the subset of the dataset. MAR is less restrictive than MCAR, since the missingness can depend on the variable itself through its relationship with other variables (but not directly).

Schafer and Olsen (1998) have an illustrative example. Consider two variables Y and X where $Y=(Y_1, Y_2, \ldots, Y_n)$ and $X=(X_1, X_2, \ldots, X_n)$ and assume that some correlation exists between the two variables. For simplicity, assume further that Y has complete data while X has some missing values. Under MCAR, Y does not provide any information on missing X's since they are missing completely at random. However, under MAR, Y contains some information on the missing X's, for example, X's

[1] For example, Schafer developed PC-based software for computing MI. The software can be downloaded from his webpage (http://wtat.psu.edu/~jls/mysoftwa.html). King modified Schafer's algorithm (King, et. al, 2000) and developed a Gauss-based software available on his web (http://Gking.Harvard.Edu). There is also a built-in S-Plus library called "missing". It was built upon Schafer's code, and it also provides useful commands to analyze the pattern of missingness.

[2] In S-Plus, a researcher can choose Gaussian, log-linear, or mixture of two as an estimation model, where log-linear can be used to impute factor/discrete variables.

corresponding to larger values of *Y* are more likely to be missing. They also discuss that when both variables have missing values, the principle here still applies.

More formally, let R be the matrix of missing pattern, with the same dimension as the dataset. Each element of R takes the value 1 if the datum is observed and 0 otherwise. Let ξ be the unknown parameter(s) of the missing mechanism. Then, MAR is defined as follows:

$$P(R \mid X_{obs}, X_{mis}, \xi) = P(R \mid X_{obs}, \xi)$$

This indicates that the probability of observing or missing the datum depends on the observed data and the missing mechanism, but not on the missing portion of the data.

Another assumption for MI is called "distinctness" (Rubin, 1976, Little and Rubin, 1987). This is not an intuitive assumption, especially from frequentist perspective. It means that the "joint parameter space of (θ, ξ) must be the Cartesian cross-product of the individual parameter spaces for θ and ξ (Schafer, 1997)". In Bayesian sense, this basically says that the joint prior distribution of the model parameter and the parameter of the missingness mechanism can be factored into the independent marginal densities, i.e., $\pi(\theta,\xi) = \pi_\theta(\theta)\pi_\xi(\xi)$, where π's are prior distributions. When MAR and distinctness hold, the missingness mechanism is said to be "ignorable". If ignorability holds, then the likelihood function can be factored into two terms; one only involving model parameters and observed data, and one with the missingness mechanism and missingness parameters. Thus, we can ignore the nuisance term of the missing mechanism.[3] As one can imagine, ignorability makes the estimation a lot easier.

In general, the ignorability assumption holds when missingness is under the control of the researcher. For example, double sampling (concentrating efforts to obtain responses of the random sample of non-respondents from the first phase, for more discussion on double sampling, see for example, Rao 1983) is known to create a MAR situation. In this case, responses from non-respondents who are not chosen for the second phase are missing, but are missing randomly within the subset of non-respondents. Thus, the MAR assumption is satisfied. For item-nonresponse cases in CVM studies, whether MAR

[3] For more detailed discussion on the assumptions, see Rubin (1976), Little and Rubin (1987) and Schafer (1997).

holds or not is more ambiguous. Cases in which ignobility does not hold are still subject to an active discussion among statisticians.[4,5] nonetheless, researchers may use MI as an alternative for other ad-hoc methods since these methods require even stricter assumptions.

The Multiple Imputation Algorithm:

The basic idea of MI is to estimate the missing value with an unbiased estimator using the parameter estimates and observed data, repeated M times. This will create M full datasets with imputed values different from each other due to the random draws. Since these are full datasets, a researcher can conduct analysis in the usual manner. At the end, the M results are combined, incorporating the uncertainty due to the missingness. The imputations can be obtained through a number of methods. One of the most popular approaches is the combination of EM and DA. This approach can be divided into three steps. (1) EM estimation; (2) DA estimation using EM as starting values; and (3) combining results from DAs to obtain the overall estimation.

Step1: Expectation Maximization

EM is essentially maximum-likelihood estimation, utilizing the interdependency between missing values and model parameter θ. Let X be the dataset; then X can be partitioned into $X = (X_{obs}, X_{mis})$ where X_{obs} contains the observed items of the data and X_{mis} contains the missing items of the data. Then, the log-likelihood function can be written as

$$l(\theta \mid X) = l(\theta \mid X_{obs}) + \log P(X_{mis} \mid X_{obs}, \theta) + c$$

since

$$P(X \mid \theta) = P(X_{obs} \mid \theta) P(X_{mis} \mid X_{obs}, \theta)$$

where $l(\theta \mid X_{obs})$ is a log-likelihood function of model parameters given observed data, and $P(X_{mis} \mid X_{obs}, \theta)$ is called the predictive distribution of the missing data given θ. c is an arbitrary constant. However, $P(X_{mis} \mid X_{obs}, \theta)$ is unknown since X_{mis} is not observed. Instead, we take the average

[4] See Schafer 1997, section 2.5.3, for example, for more discussion on the literatures of nonignorable cases.
[5] One of the familiar examples of nonignorable case for economists is Heckman's censored model (1976).

of the likelihood function over the predictive distribution $P(X_{mis} \mid X_{obs}, \theta^{(t)})$ where $\theta^{(t)}$ is an estimate of θ for the t[th] iteration. Then, use $P(X_{mis} \mid X_{obs}, \theta^{(t)})$ to calculate the log-likelihood iteratively until it converges. For more complete discussion of EM, see Dempster, et. al. (1977).

Step2: Data Augmentation:

While EM converges to a single parameter estimate deterministically, DA will estimate the distribution $P(X_{mis} \mid X_{obs}, \theta)$ itself using Markov Chain Monte Carlo (Tanner and Wong, 1987). The idea is to draw a missing value estimate $X_{mis}^{(t+1)}$ from the distribution $P(X_{mis} \mid X_{obs}, \theta^{(t)})$ where $\theta^{(t)}$ is an estimate of θ for the t[th] iteration. Then, draw a new estimate $\theta^{(t+1)}$ from the complete-data posterior $P(\theta \mid X_{obs,} X_{mis}^{(t+1)})$. This yields a stationary distribution $P(X_{mis} \mid X_{obs})$, the true distribution of missing values conditional on observed data from which we can draw an estimate of missing values.

Assessing convergence is an important issue in MCMC. We need to check if the stationary distributions are attained so that draws from these distributions are in fact, from the desired distributions. Schafer suggests to look at the auto-correlation functions and time-series plots (Schafer, 1997). If distributions converge, auto-correlations should die out and time-series plots exhibit white noise. Researchers also should use enough burn-in period so that draws are from the stationary distribution. Burn-in period is pre-convergence iterations not used to the actual analysis. Auto-correlation and time-series plots help researchers to determine the length of burn-in periods that ensures the convergence. For more detailed discussion on MCMC, see Robert and Casella (1999).

Step3: Combining Results

In MI, each missing value is imputed for M times, which yields M complete datasets. M is typically 3 to 5, since more than 5 iterations does not gain much more efficiency (Rubin, 1987). The

point-estimate is simply the mean of the M imputations. The variance estimate is calculated by incorporating the uncertainty of substituting missing values (Rubin, 1987, also see Rubin, 1996).

The point estimate is the mean of M imputations, thus calculated as:

$$\overline{Q} = \frac{1}{M}\sum_{m=1}^{M} Q_m$$

Let the estimated variance for each imputation be V_m, then the within-imputation variance V and between-imputation variance B can be calculated as follows:

$$V = \frac{1}{M}\sum_{m=1}^{M} V_m$$
$$B = \frac{1}{M-1}\sum_{m=1}^{M} (Q_m - \overline{Q})^2$$

The total variance T can be obtained by calculating:

$$T = V + (1+\frac{1}{M})B$$

The estimator is distributed approximately as:

$$\frac{Q-\overline{Q}}{T^{1/2}} \sim t_v, \quad where\ v\ is \quad v = (m-1)\left[1+\frac{V}{(1+M^{-1})B}\right]$$

Thus, use this distribution for inferences such as hypothesis testing and confidence intervals.

III. An Application

Data Generating Process and Estimation Models

There are three variables, X_1, X_2, and X_3, generated as multivariate normal random variable with different values for correlations ρ, where ρ takes the values 0.1, 0.5, and 0.9.

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\mu = \begin{pmatrix} 65 \\ 20 \\ 40 \end{pmatrix}, \Sigma = \begin{pmatrix} 50^2 & \rho(50)(15) & \rho(50)(30) \\ \rho(50)(15) & 15^2 & \rho(15)(30) \\ \rho(50)(30) & \rho(15)(30) & 30^2 \end{pmatrix}\right)$$

The chosen sample size is *n = 100*. These variables are to mimic explanatory variables to calculate willingness to pay (WTP) for some good. In particular, X_1 can be thought of as an income variable, and

X_2 and X_3 are some other socioeconomic variables or taste variables that affect WTP. The mean and variance of X_1 were taken from the income variable in a CVM study by Larson and Lew (2000). Mean and variance for the other two variables are arbitrarily chosen such that sensible values of WTP are generated. WTP is calculated as a linear function of these variables plus an error term. Parameters are set to be $\beta^T = (100, 0.4, -2, 0.8)$. For example, WTP for the i[th] individual is calculated as

$$WTP_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad where\ \varepsilon_i \sim N(0, 40)$$

Here, WTP is directly observable and is a linear function of the explanatory variables. It is also assumed that this is a true data generating process. Surely, this is not what we face in reality. However, we would like to compare how well different methods works. For such comparisons, a simple model is more desirable because it allows for comparisons of differences purely due to the imputation methods.

After the complete data is generated, some data points are deleted according to a mechanism such that some observations on X_1 are missing for higher values (above average) of X_3, some X_2 are missing for lower values (below average) of X_3, and some X_3 are missing for higher values of X_1 (above average) with probability 0.6, 0.5, and 0.5 for X_1, X_2, and X_3 respectively. This process deletes between 20 to 40 percent of each variable. The model applied for the estimation is a simple linear model. The methods of case deletion, mean imputation, stochastic regression imputation, and MI are applied to impute missing values. The imputed values for a sample iteration is shown in Appendix A. For comparison, estimation using the full data before the data deletion is also calculated. After models are fit, sum of squared errors, mean squared errors, and numerical power of tests are calculated. This process is repeated 300 times as a Monte Carlo experiment.

IV. Estimation Results

R^2:

R^2 measures the goodness of fit of the model. In particular, it shows the proportion of the total variation of data explained by the explanatory variables. Thus, R^2 is an indicator of the overall performance of the model. One of the advantages of using R^2 is that it makes the comparison among datasets with different correlation coefficient ρ easier since it has a range from 0 to 1. Figure.1 shows the

density plot of R^2 for each method, with different correlation parameter ρ. Note that case-deletion is not shown here because it is not directly comparable due to the discarded observations.

Larger values for R^2 indicate a better fit. Thus, more distribution mass towards the right is an indication of better performance. Also, Table.1 shows the mean of R^2.

Table.1 Mean of R2 for each imputation method

rho	Full Model	Mean Imputation	Regression Imputation	Multiple Imputation
0.1	0.39	0.25	0.18	0.41
0.5	0.37	0.23	0.22	0.38
0.9	0.20	0.11	0.13	0.23

The performances of imputation methods depend on the correlation parameter ρ. For ρ=0.1, R^2 is larger, and it decreases as the correlation becomes higher, in general. This is because of the multicollinearity among variables. The full model has R^2 of 0.39 on average even when ρ=0.1. This is relatively low due to the large variability of error terms in the data generating process. MI performs closest to the full model, while both mean imputation and regression imputation diverges from the full model. Regression imputation performs poorly overall, although it did slightly better with ρ=0.5. It is counterintuitive that regression imputation did not do well when correlation is high. It is probably because of the multicollinearity. Mean imputation did better than regression imputation for ρ=0.1, but it becomes the worst method for larger correlations. Overall, MI did as well as the full model, while mean imputation and regression imputation performed poorly.

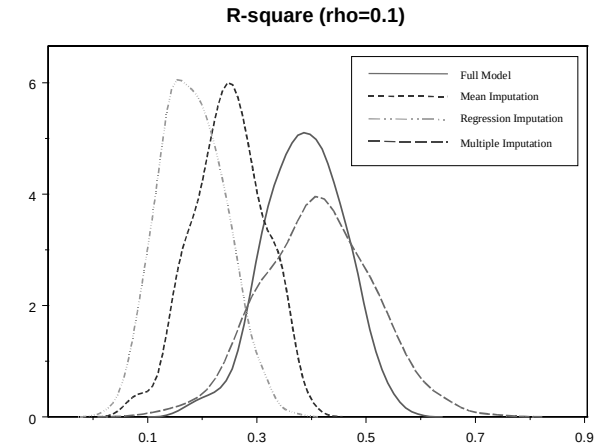
Coefficient Estimates:

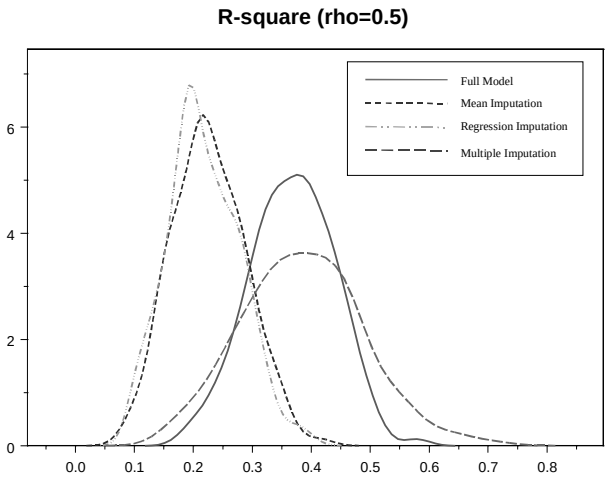
The density plots of coefficient estimates are the convenient way to compare the quality of estimated coefficients, in terms of bias and standard errors. Figure.2 shows density plots of each coefficient for 300 Monte Carlo iterations for each method with different correlation among variables. From the peak of each density, we can assess the bias since these distributions are roughly symmetric. The dispersion of the density shows the variability of estimates. Thus, efficiency can be assessed by the dispersion of the density.

The density plots show that the full model is always unbiased. There seems to be some dispersion from the full model for all the methods, but most of them are minor. However, mean imputation shows strong bias for the estimate of β_3, especially when correlation is 0.1 and 0.9. They are quite dispersed from the full model. For all the cases, case deletion seems to be unbiased, but has particularly large dispersion. This shows the significant loss of efficiency in estimation due to the thrown away data.

In many cases, the density plots of case deletion contain zero, which suggests that it may wrongly fail to reject the null that coefficients are zero. The shape of the density plots of regression imputation is "slim" compared to the full model. This confirms that regression imputation tends to underestimate standard errors. Regression imputation uses auxiliary regression to predict the missing values. It overstates the correlations, which results in smaller standard errors. In the density plots, regression imputation has slight bias in some cases. This is potentially a serious problem especially when estimates are biased, since it could result in completely wrong inferences. Overall, it seems that MI stays closest to the full model, with no major bias or loss of efficiency.

Figure.1 Density of R^2 for each method with different correlation parameter rho





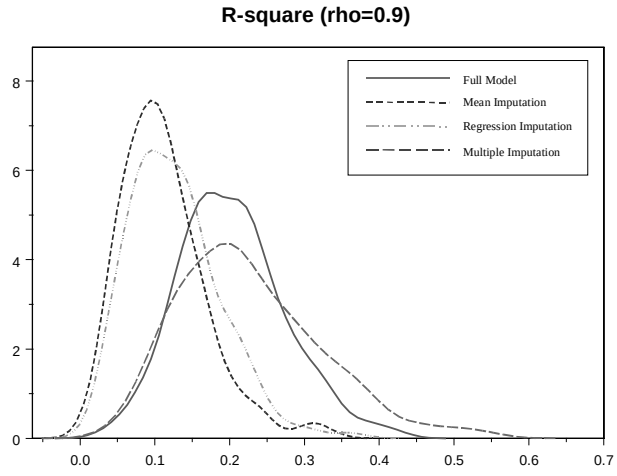
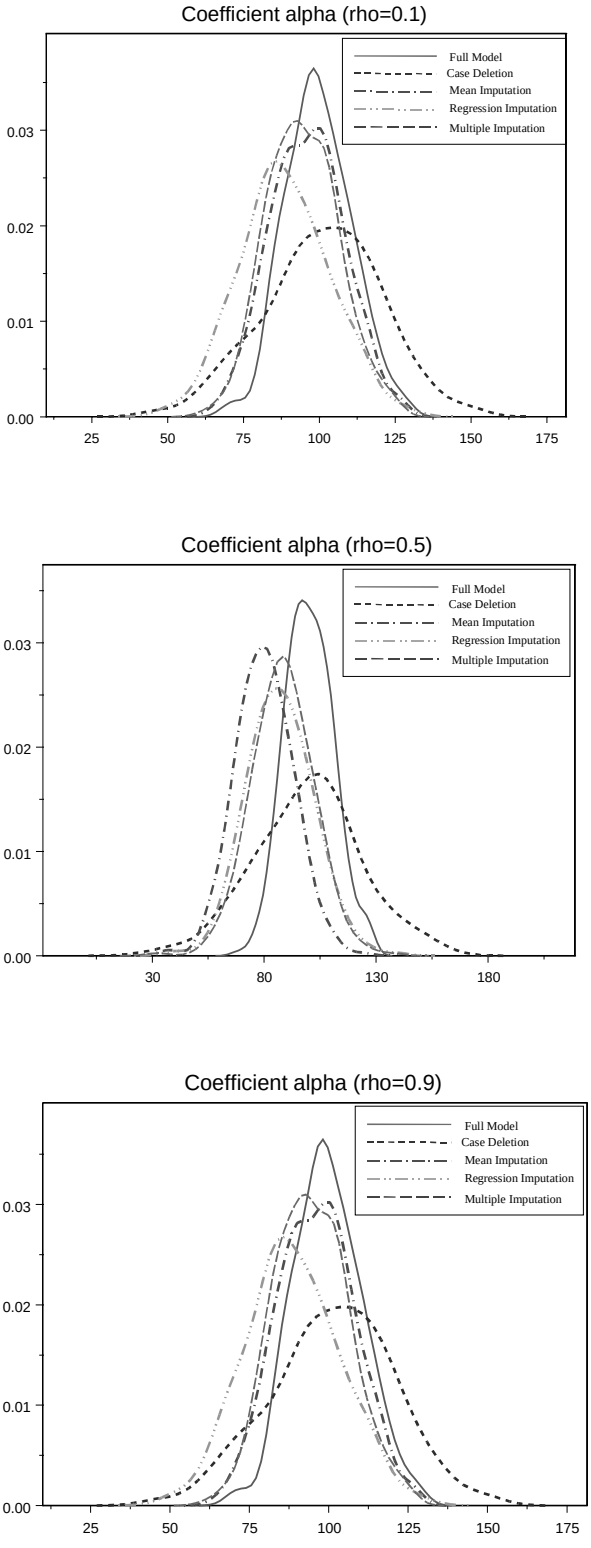
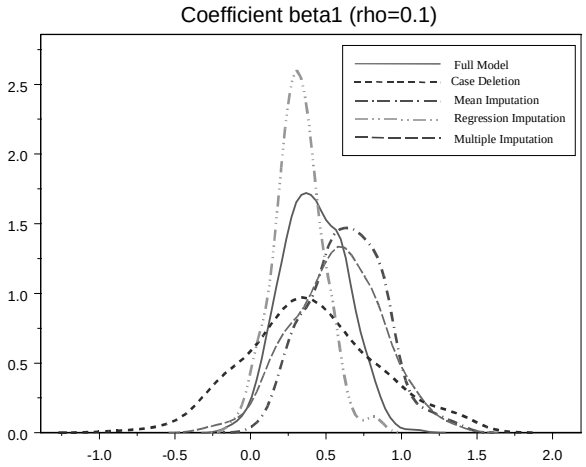


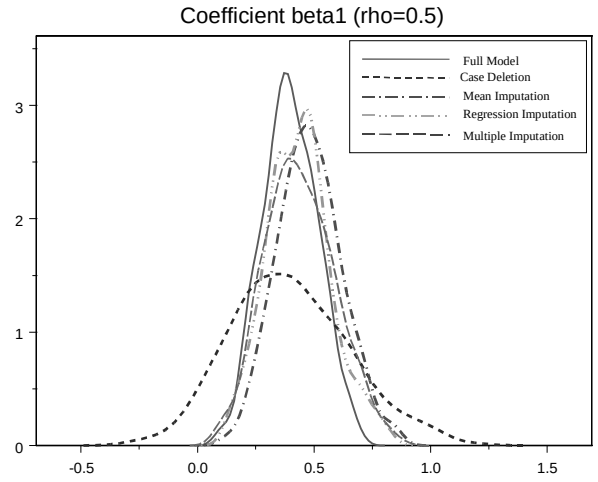
Figure.2 Coefficient estimate of each method with different correlation parameter rho

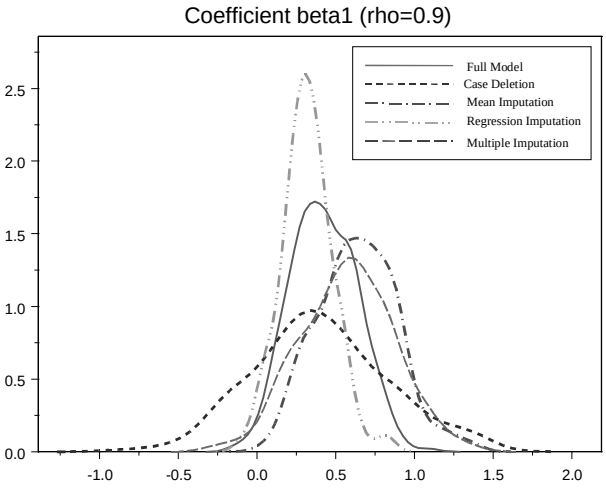
Alpha (True alpha = 100)



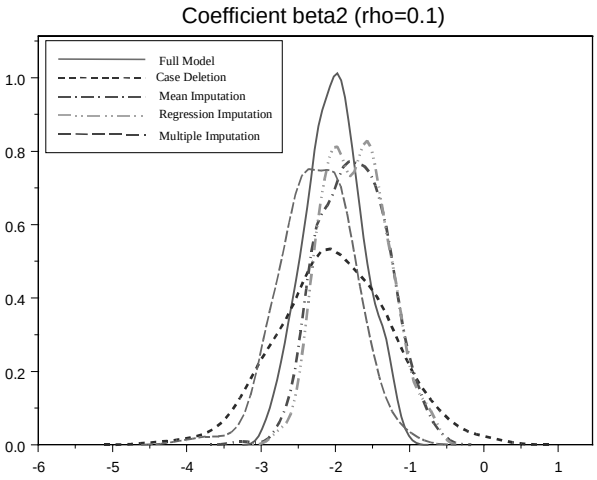
Beta1 (True beta1 = 0.4)

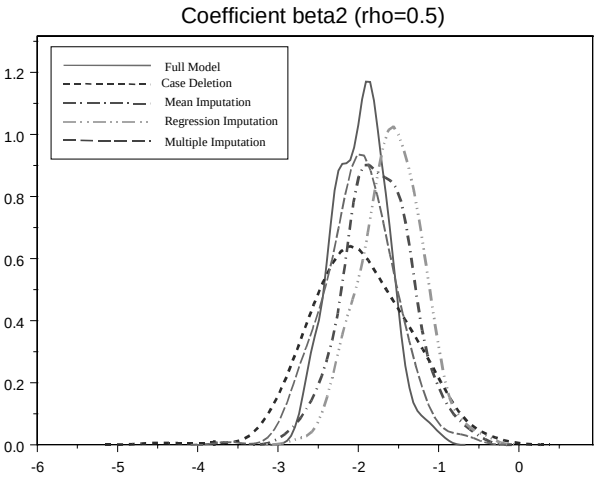


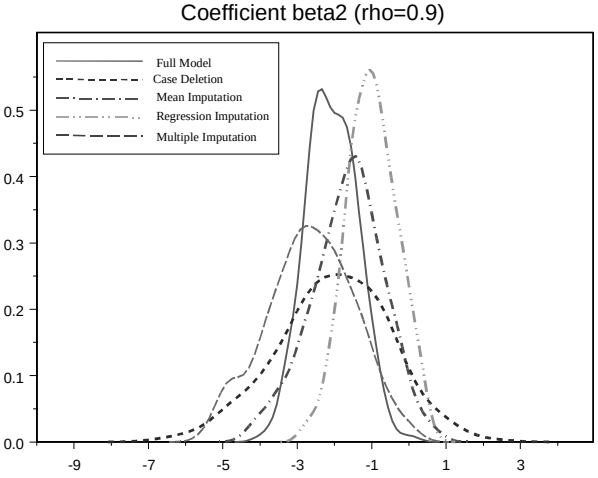




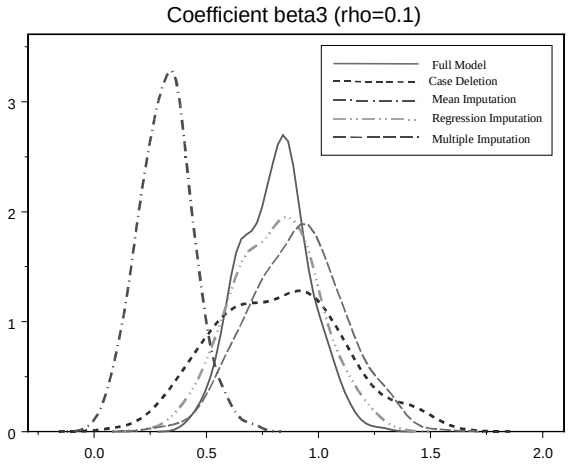
Beta2 (True beta2 = -2)

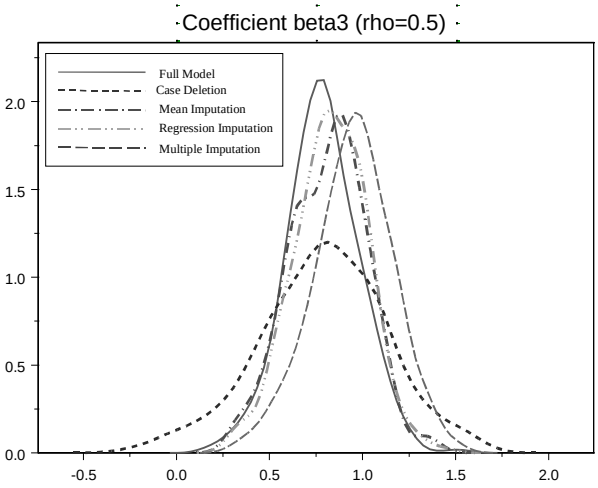


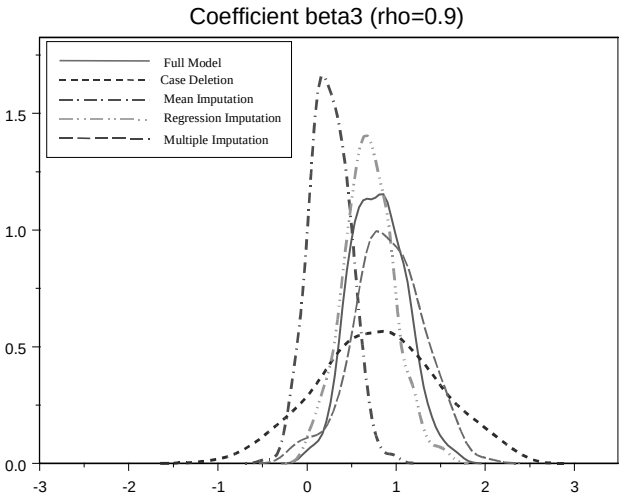




Beta3 (True beta3 = 0.8)







Power of test:

The third important criterion is whether the model correctly rejects the null that coefficients are zero when they are significant in truth. Table.2 shows how the ratio that each model rejects this hypothesis correctly at the 95% confidence level in 300 repetitions.

Table.2 Ratio of correctly rejected null that coefficient is zero

rho=0.1	**alpha**	**beta1**	**beta2**	**beta3**
Full model	1.00	1.00	1.00	1.00
Case deletion	0.99	0.99	0.76	0.79
Mean imputation	1.00	0.93	0.97	0.97
Regression Imputation	1.00	0.95	0.96	1.00
Multiple Imputation	1.00	0.99	1.00	0.99
Multiple imputation (including b/w variability)	1.00	0.95	1.00	0.99
rho=0.5	**alpha**	**beta1**	**beta2**	**beta3**
Full model	1.00	0.91	1.00	0.98
Case deletion	0.98	0.30	0.87	0.67
Mean imputation	1.00	0.80	0.97	0.97
Regression Imputation	1.00	0.87	0.97	0.95
Multiple Imputation	1.00	0.92	0.99	1.00
Multiple imputation (including b/w variability)	1.00	0.71	0.98	0.98
rho=0.9	**alpha**	**beta1**	**beta2**	**beta3**
Full model	1.00	0.53	0.85	0.67
Case deletion	1.00	0.15	0.26	0.23
Mean imputation	1.00	0.45	0.24	0.57
Regression Imputation	1.00	0.71	0.34	0.20
Multiple Imputation	1.00	0.72	0.78	0.78
Multiple imputation (including b/w variability)	1.00	0.54	0.64	0.60

For all the cases, MI correctly rejects the null more frequently than case deletion, mean imputation, and regression imputation. Case deletion performs quite poorly, rejecting null very few times. This is because of the efficiency loss we observed in Figure.1. Mean imputation and regression imputation performs better than case deletion, but not as well as MI. Neither model perform well when ρ is high. When correlation is high among variables, multicollinearity will result, which inflates the

variance. The poor performance when correlation is high is due to the multicollinearity. MI outperforms all the other methods in terms of the power of the test.

V. Conclusion

In general, the performance of the imputation methods depends on the quality of the data and how well the full model describes the data. Given that the data and full model are good, MI seems to outperform other methods, at least for the problems analyzed here. The analysis shows strong evidence that MI provides the closest estimation results to the full model. Case deletion seems to be unbiased, but has a significant efficiency loss, which results in low power. Mean imputation had lower R^2, bias in some estimates, and lower power than MI. Regression imputation had the lowest R^2, downward bias in standard errors, and lower power.

In this analysis, I used a simple linear continuous model to illustrate the comparison. However in the CV literature, it is more common to observe categorical variables. A natural extension of this research is to see how well MI works for categorical data. The effect of misspecification would be another topic to explore. In this paper, the model was "correct". However in any real situations, we never know what the true model is. Also, the validity of the ignorability assumption in the survey data should be explored more extensively.

In conclusion, MI is computationally more intensive than other methods, but it appears well worth implementing for better estimation results.

Reference:

1. Dempster, A.P., Laired, N.M., and Rubin, D.B. (1997). Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, 39, 1-38.

2. Hanamann, Michael W. and Kanninen, Barbara (1998). The Statistical Analysis of Discrete-Response CV Data. Working Paper No.798.

3. King, Gary; Honaker, James; Joseph, Anne and Scheve, Kenneth (2000). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. Working Paper.

4. Larson, Douglas M. and Lew, Daniel K. (2000). *The Public's Willingness to Pay for Improving California's Water Quality*. Presented at the 2001 W-133 Western Regional Research Conference, Miami Florida.

5. Little, Roderick J.A and Rubin, Donald B. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons.

6. Little, Roderick; Raghunathan, Trivellore (1993). Should Imputation of Missing Data Condition on All Observed Variables? *Proceeding of the Annual Research Conference*, 617-622, Bureau of the Census, Washington DC.

7. Mitchell, Robert C. and Carson, Richard T. (1989). *Using Surveys to Value Public Goods: The Contingent Valuation Method*, Washington D.C. : Resource for the Future.

8. Rao, P.S.R.S, Randomization Approach (1983). *Incomplete Data in Sample Surveys*, *Volume 2, Theory and Bibliographies*, 97-105, Panel on Incomplete Data: New York, Academic Press.

9. Robert, Christian P., and Casella, George (1999). *Monte Carlo Statistical Methods*, New York: Sprinter-Verlag New York, Inc.

10. Rubin, Donald B. (1976). Inference and Missing Data. Biometrika, Volume 63, Issue 3, 581-592.

11. Rubin, Donald B. (1977). Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72, 538-543.

12. Rubin, Donald B. (1978). Multiple Imputations in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse. *Proceeding of the Survey Research Methods Section, American Statistical Association*, pp.20-34.

13. Rubin, Donald.B. (1996). Multiple Imputation after 18+ Years (with discussion). *Journal of the American Statistical Association*, 91, 473-489.

14. Schafer, Joseph.L. (1997). *Analysis of Incomplete Multivariate Data*, London: Champan & Hall.

15. Schafer, Joseph.L. and Olsen, Maren K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Working Paper.

16. Tanner, Martin.A. and Wong, Wing Hung. (1987). The Calulation of Posterior Distibutions by Data Augmentation (with discussion*). Journal of the American Statistical Association*, 82, 528-550.