



AgEcon SEARCH
RESEARCH IN AGRICULTURAL & APPLIED ECONOMICS

The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
<http://ageconsearch.umn.edu>
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

2002

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

DRAFT for BROWNBAG SEMINAR 5/6/2002

762
Econometrics

JUN 06 2002
UNIVERSITY OF CALIFORNIA
DAVIS

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.

Copyright 2002 by Yuko Onozaka. All rights reserved. Readers may make verbatim copies of this document for non-commercial purpose by any means, provided that this copyright notice appears on all such copies.

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 smaller sum of squared errors (SSE) and highest power. Case deletion performs particularly poorly, and results in by far the largest mean squared errors (MSE) and lowest power. Regression imputation has smaller MSE when correlations among variables are high but results in larger SSE and lower power in general. Mean imputation has relatively small MSE, but SSE is high and power is low.

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 $Y=(Y_1, Y_2, \dots, Y_n)$ and $X=(X_1, X_2, \dots, X_n)$ 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:

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

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 ignability 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

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

since

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

where $l(\theta | X_{obs})$ is a log-likelihood function of model parameters given observed data, and $P(X_{mis} | X_{obs}, \theta)$ is called the predictive distribution of the missing data given θ . c is an arbitrary constant. However, $P(X_{mis} | X_{obs}, \theta)$ 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 $P(X_{mis} | X_{obs}, \theta^{(t)})$ where $\theta^{(t)}$ is an estimate of θ for the t^{th} iteration. Then, use $P(X_{mis} | X_{obs}, \theta^{(t)})$ 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 $P(X_{mis} | X_{obs}, \theta)$ 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 $P(X_{mis} | X_{obs}, \theta^{(t)})$ 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 $P(\theta | X_{obs}, X_{mis}^{(t+1)})$. This yields a stationary distribution $P(X_{mis} | X_{obs})$, 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:

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

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:

$$V = \frac{1}{M} \sum_{m=1}^M V_m$$

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

The total variance T can be obtained by calculating:

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

The estimator is distributed approximately as:

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

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.

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

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

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

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 work. 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.

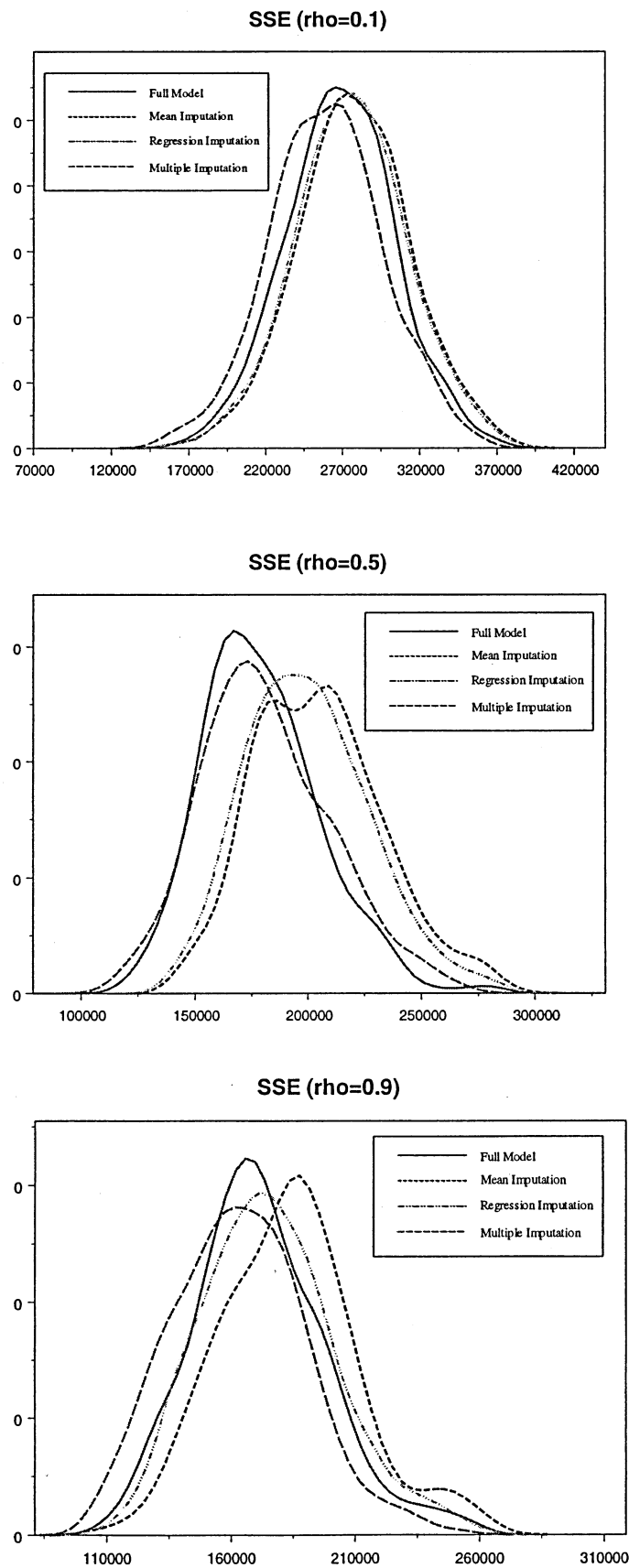
IV. Estimation Results

SSE:

The sum of squared errors (SSE) is the measure of how well the response variable is predicted. Figure.1 shows the density plot of SSE 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.

Smaller SSE indicates better prediction of the response variable. Thus, more distribution mass towards the left is an indication of better predictive power. Several observations can be made from Figure.1. First, the performances of the imputation methods depend on the correlation parameter ρ . For $\rho=0.1$, SSE is large in general, even with the full data. When $\rho=0.5$, SSE decreases, and the densities for mean imputation and regression imputation deviate from the full model, while the density for MI stays close to it. For $\rho=0.9$, regression imputation becomes somewhat closer to the full model while mean imputation is still off from it. MI remains close to full model.

Figure.1 Density of SSE for each method with different correlation parameter rho



MSE:

The mean squared error (MSE) allows overall comparison among estimators since it is a function of both the variances and the bias of the estimators. Figure.2 shows box-plots of the MSE for each method with different ρ . Note that MI incorporates the uncertainty of missingness by adding the variability within estimates where as other methods ignore this uncertainty. Consequently, methods other than MI have lower MSE due to the uncounted uncertainty term. In order to compare fairly, MI calculated without the uncertainty term is also provided in the figure.

Box-plots graphically show the quantiles of the data. The stem shows the minimum and maximum, and the box shows the inner quantile (first and third). The middle dot is the median. More observations to the right indicate larger MSE.

For almost all the cases, case deletion has by far the largest MSE. It is striking that case-deletion, the most common method in practice, works quite poorly. Mean imputation, regression imputation and MI perform relatively better compared to case deletion. It is not unambiguous which performs the best among these three. However, regression imputation seems have small MSE as correlations among variables get larger. It is as expected since larger correlation implies better predictive power of the auxiliary regression.

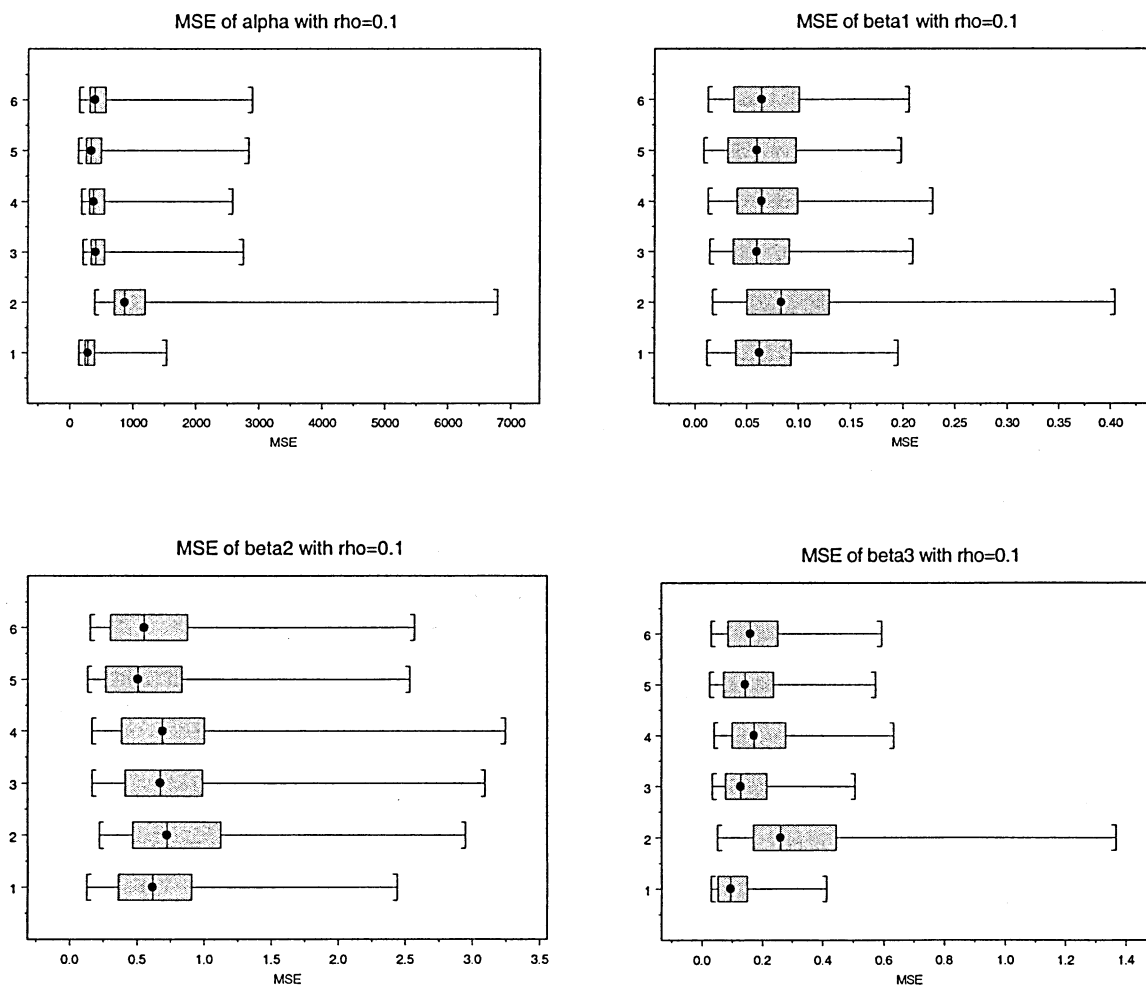
Several things are worth noting. There seems to be some mixed cases where higher correlations affect the MSE positively or negatively. This is probably due to the fact that imputations use the relationships among variables. However, for large correlations, there is an effect of collinearity, where the variance estimates are inflated. Second, the regression method tends to bias the variance downward. This would explain why the regression method has relatively small MSE. It is well known that regressions are in general sensitive to misspecification (e.g. omitted variable bias, violation of linearity). Thus, the small MSE for regression imputation may not be taken at face value. In this sense, MI is more flexible, since it does not require any model assumptions as in regression imputation.

Figure.2 MSE of each method with different correlation parameter rho

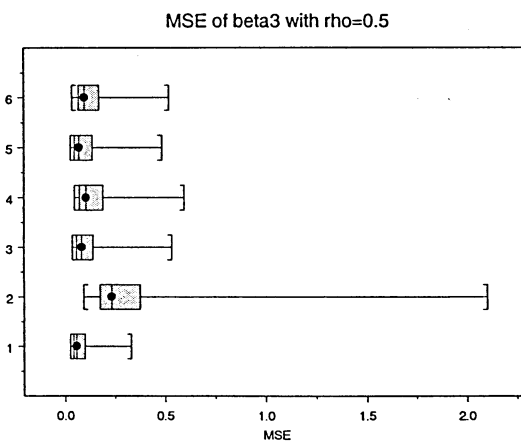
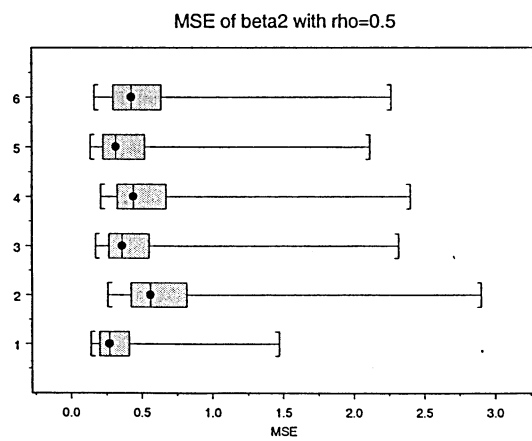
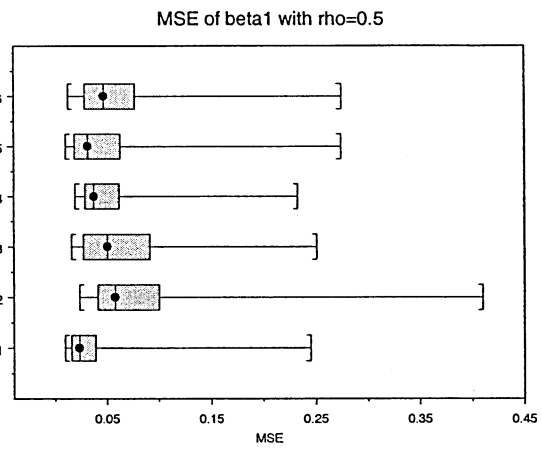
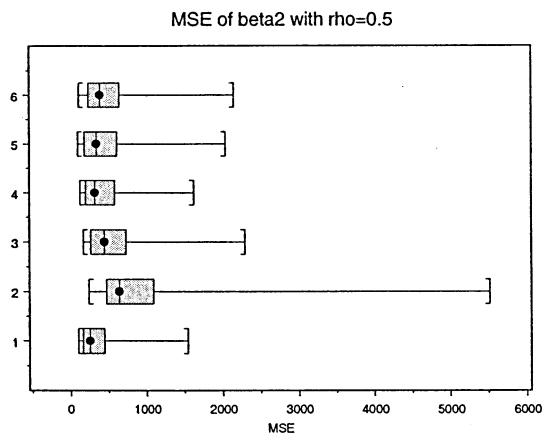
For all the plots:

- 1 : Full Model
- 2 : Case Deletion
- 3 : Mean Imputation
- 4 : Regression Imputation
- 5 : MI
- 6 : MI with uncertainty counted

Rho=0.1

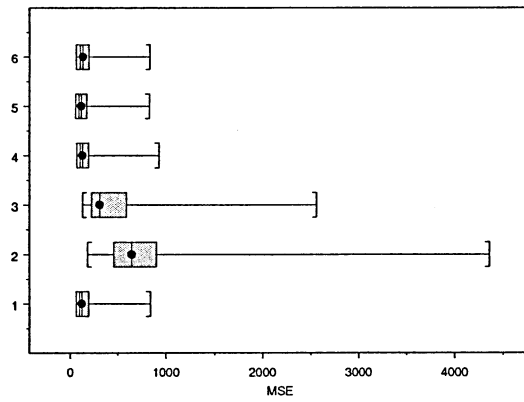


Rho=0.5

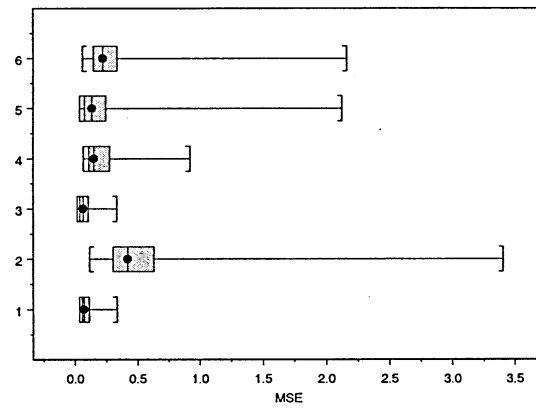


Rho=0.9

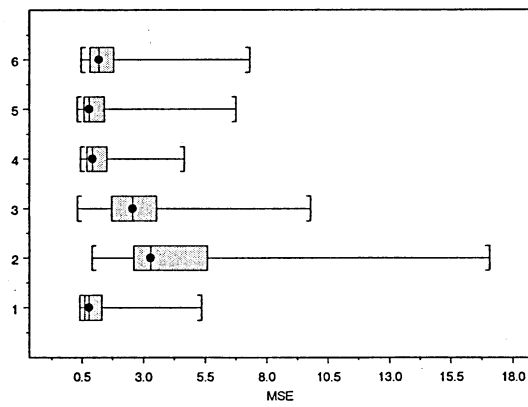
MSE of alpha with rho=0.9



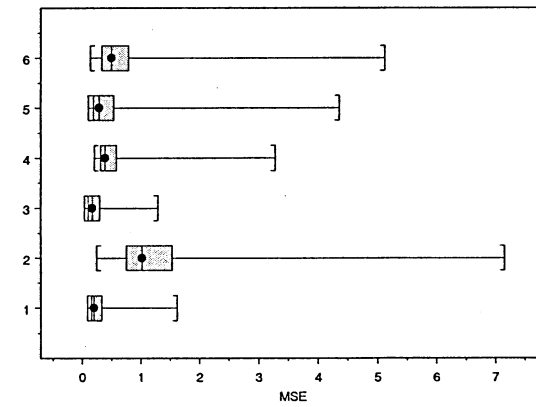
MSE of beta1 with rho=0.9



MSE of beta2 with rho=0.9



MSE of beta3 with rho=0.9



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. Figure.3 shows how many times each model rejects this hypothesis correctly at the 95% confidence level in 300 repetitions.⁶

Table.1 Ratio of correctly rejected null that coefficient is zero

rho=0.1	beta1	beta2	beta3
Full model	0.32	0.96	0.89
Case deletion	0.09	0.67	0.16
Mean imputation	0.24	0.91	0.59
Regression Imputation	0.23	0.90	0.46
Multiple Imputation	0.38	0.97	0.66
Multiple imputation (including b/w variability)	0.26	0.92	0.53
rho=0.5	beta1	beta2	beta3
Full model	0.79	1.00	0.98
Case deletion	0.32	0.91	0.26
Mean imputation	0.36	0.96	0.84
Regression Imputation	0.56	0.99	0.71
Multiple Imputation	0.61	0.99	0.86
Multiple imputation (including b/w variability)	0.45	0.97	0.73
rho=0.9	beta1	beta2	beta3
Full model	0.32	0.67	0.44
Case deletion	0.11	0.29	0.12
Mean imputation	0.22	0.14	0.39
Regression Imputation	0.08	0.57	0.29
Multiple Imputation	0.36	0.74	0.48
Multiple imputation (including b/w variability)	0.19	0.60	0.27

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. Mean imputation and regression imputation performs better than case deletion, but not as well as MI. Neither model perform well when ρ is high or low. 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. On the other hand, when correlation is low, there is not much information in variables to predict missing observations. Thus, imputation does not work as well, which results in the

⁶ The intercept estimates, not reported here, were significant for all the methods for all the correlations.

lower power. MI outperforms all the other methods in terms of the power of the test and shows robustness to multicollinearity.

V. Conclusion

In general, the performance of 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 better estimation in terms of smaller SSE, more powerful test results, and relatively small MSE.

Results from MSE comparison were more ambiguous and difficult to determine which methods performed the best. However, regression imputation performed well when correlation is high. There seems to be a tradeoff between high correlation among variables and multicollinearity when linear regression is fit. Thus, regression imputation performs well in terms of smaller MSE compared to other variables, but it also results in lower power in hypothesis testing.

In this analysis, I used only continuous data for simplicity. 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., Laird, 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: Springer-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 Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

Appendix A: Sample Imputation Result:

The actual imputed values for one sample iteration are shown by different patterns of missingness. There are 6 patterns of missingness, (1) no missing values, (2) only X1 missing, (3) only X2 are missing, (4) only X3 are missing, (5) X1 and X3 are missing, and (6) X2 and X3 are missing. The percentage missing for each variable is such that 33%, 20%, and 23% are missing from X1, X2, and X3 respectively.

The overall fit of this model can be diagnosed by SSE, shown in Table.A1. As one can see, regression with full data has SSE equals to 162,643. Mean imputation and regression imputation both has higher SSE than full data model. MI has smaller SSE on average than any other methods.

Table.A1 SSE

Full Data	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
162643	173617	177152	155382	165002	145691	174241	155250	136725

Table.A2 to A6 shows the comparison between actual and imputed values for each method, by different missing patterns.

Table.A2 Imputed values in cases when only X1 is missing

Actual X1	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
34.13	67.52	82.70	65.45	-3.31	126.54	99.84	54.47	49.69
86.66	67.52	114.97	127.37	220.12	170.01	95.19	114.31	37.23
201.05	67.52	108.20	134.94	111.56	156.61	46.32	157.23	202.97
65.73	67.52	94.37	89.60	133.23	88.35	57.09	129.03	40.29
103.38	67.52	104.07	154.29	265.34	97.52	138.37	153.51	116.72
100.03	67.52	87.49	71.55	-3.83	117.59	45.34	80.77	117.87
72.77	67.52	90.81	69.19	60.10	14.53	114.49	86.64	70.20
46.89	67.52	84.46	99.95	69.58	131.00	124.55	89.04	85.58
122.56	67.52	97.63	93.45	52.25	112.92	95.87	79.30	126.89
131.14	67.52	144.67	152.86	224.99	181.22	60.44	165.80	131.85
147.49	67.52	103.96	77.37	85.58	47.21	97.91	77.37	78.75
90.02	67.52	97.53	122.86	123.95	172.55	65.70	116.92	135.17
78.85	67.52	113.73	104.46	136.00	64.91	149.09	82.01	90.31
8.03	67.52	89.21	66.12	64.95	114.91	4.71	157.27	-11.24
81.89	67.52	128.69	115.20	140.96	141.32	83.94	142.31	67.49
69.35	67.52	73.80	57.05	48.10	81.87	50.46	66.73	38.08
105.86	67.52	138.01	149.49	128.95	121.21	177.19	160.01	160.10
114.10	67.52	125.68	102.44	46.13	126.35	35.74	124.03	179.96
79.93	67.52	96.56	89.25	32.74	103.98	72.03	119.60	117.91
113.88	67.52	116.92	116.06	112.07	131.29	154.40	114.35	68.20
87.42	67.52	96.33	69.47	143.23	28.45	124.29	38.33	13.03
107.60	67.52	84.01	82.88	83.15	43.56	125.15	71.02	91.51
160.67	67.52	86.25	125.41	117.59	116.32	122.24	82.11	188.77
68.23	67.52	95.93	136.29	76.68	181.88	105.10	184.87	132.93
83.05	67.52	69.59	57.12	-0.75	138.77	56.29	63.98	27.29

Table A.3 Imputed values in cases when only X2 is missing

Actual X3	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
20.77	24.43	24.17	17.49	15.90	35.61	6.90	9.08	19.95
36.83	24.43	23.96	23.54	43.79	7.09	9.48	39.49	17.86
21.36	24.43	24.18	21.28	11.07	22.76	33.66	15.37	23.57
16.32	24.43	16.64	24.06	23.40	41.32	28.74	2.45	24.37
12.28	24.43	21.29	19.64	8.68	25.68	14.56	23.28	25.99
25.92	24.43	11.90	-3.14	-12.81	6.63	1.76	-13.09	1.81
7.77	24.43	20.88	6.54	0.32	-3.79	17.57	17.72	0.88
7.70	24.43	16.82	25.69	25.52	36.47	26.34	21.01	19.12
45.94	24.43	32.12	34.79	19.40	40.08	34.01	48.11	32.38
13.63	24.43	21.32	12.10	24.38	7.80	12.42	9.41	6.50
14.67	24.43	15.62	18.96	16.89	27.42	9.73	25.27	15.51
16.47	24.43	12.59	14.15	8.78	18.72	27.29	5.81	10.18
10.76	24.43	9.09	13.41	12.24	16.57	5.28	26.31	6.66
42.72	24.43	24.67	23.38	23.48	16.90	30.05	17.16	29.32

Table.A4 Imputed values in cases when only X3 is missing

Actual X3	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
33.94	47.63	46.49	42.53	15.72	36.78	26.58	73.35	60.20
34.75	47.63	35.32	40.39	50.15	26.29	55.28	32.72	37.53
34.33	47.63	35.34	26.88	25.82	8.92	22.66	42.66	34.34
93.67	47.63	30.69	43.59	23.48	5.78	73.92	76.79	37.98
24.26	47.63	40.73	44.92	41.14	70.09	72.79	12.31	28.26
9.99	47.63	26.88	32.29	32.96	8.16	57.08	13.72	49.52
3.43	47.63	34.70	34.85	31.99	-11.52	55.01	43.94	54.85
59.82	47.63	46.49	25.81	7.53	43.03	23.87	29.61	25.02
18.50	47.63	33.92	47.14	81.99	37.70	63.03	24.96	28.05

Table.A5 Imputed values in cases when X1 and X3 are missing

Table 1: Mean Regression									
	Actual X3	Imputation	Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
X1	70.80	67.52	74.50	67.10	15.75	93.87	58.15	68.19	99.55
	67.52	67.52	72.65	56.35	72.40	76.12	57.70	9.75	65.78
	67.91	67.52	65.37	95.31	70.67	123.67	12.83	109.98	159.40
	6.74	67.52	97.50	66.02	44.03	38.97	131.85	44.36	70.88
	4.10	67.52	104.44	93.81	38.33	61.46	133.35	167.84	68.07
	72.48	67.52	69.36	57.35	138.88	4.37	-0.52	82.38	61.64
	29.07	67.52	96.99	59.42	101.93	96.25	14.76	60.08	24.10
	60.28	67.52	113.93	61.39	76.72	36.10	20.40	118.24	55.49
	Actual X3	Imputation	Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
X3	51.91	47.63	47.36	48.40	7.11	60.57	64.01	44.73	65.58
	83.84	47.63	47.39	33.94	50.22	49.70	26.95	35.08	7.74
	60.97	47.63	47.92	44.82	51.25	67.41	40.28	27.55	37.59
	46.71	47.63	47.31	38.86	26.53	52.24	11.84	48.74	54.94
	71.55	47.63	49.51	56.42	14.11	82.06	68.29	56.94	60.72
	81.32	47.63	47.11	16.60	28.36	3.96	2.49	34.74	13.45
	64.90	47.63	48.11	51.73	49.77	57.70	41.32	56.36	53.52
	67.38	47.63	49.67	43.29	13.19	63.18	40.83	70.20	29.07

Table.A6 Imputed values in cases when X2 and X3 are missing

	Actual X3	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
X2	13.92	24.43	15.57	22.00	27.88	13.90	16.46	21.78	30.01
	18.67	24.43	14.84	18.12	15.19	21.59	17.58	27.55	8.71
	33.53	24.43	18.48	16.21	15.08	18.70	22.85	13.58	10.84
	31.96	24.43	18.87	28.31	42.87	26.33	36.64	19.36	16.37
	17.96	24.43	14.49	10.11	1.20	8.56	11.73	18.81	10.26
	18.62	24.43	14.31	15.28	17.35	35.86	-6.80	12.49	17.53
	Actual X3	Mean Imputation	Regression Imputation	Average MI	MI 1	MI 2	MI 3	MI 4	MI 5
X3	34.03	47.63	39.83	34.19	15.54	50.29	23.05	40.36	41.69
	0.53	47.63	37.61	23.28	-27.87	51.23	35.57	21.27	36.22
	39.61	47.63	45.92	24.79	32.95	37.07	20.28	33.74	-0.09
	42.49	47.63	40.86	57.12	100.62	63.90	46.64	17.61	56.83
	11.78	47.63	32.50	37.82	46.56	14.49	49.90	40.99	37.18
	1.05	47.63	35.62	15.86	14.50	40.45	3.40	3.55	17.40