The Philippine Statistician, 2007
Vol. 56, Nos. 1-2, pp. 19-31

A Comparison of Four Approaches to Handling Missing Data

Jeffrey Scott Granberg-Rademacker[1]

Received: November, 2006; Revised: May, 2007

ABSTRACT

Missing data problems have typically been under-appreciated in the social sciences. In practicality, few students or researchers go beyond the default settings of most statistical packages when conducting analyses on missing data. Four different approaches for handling missing data will be presented and conducted on a set of simulated missing data where the missing data follows a partial Missing at Random (MAR) pattern.

Keywords: Imputation, Missing Data, Simulation, Linear Regression, Monte Carlo

I. INTRODUCTION TO MISSING DATA

One of the most common problems that empirical social science researchers encounter is missing data. However, despite its widespread occurrence, it is still an underappreciated problem in many disciplines. It is the purpose of this article to clearly illustrate how missing data can make parameter estimates biased, and to compare the performance of several different approaches to handling missing data.

II. PATTERNS OF MISSINGNESS

Little and Rubin (2002), uncovered the importance of missingness patterns, by identifying three distinct patterns: Missing Completely at Random (MCAR), Missing at Random (MAR), and Nonignorable (NI). In the following subsections, suppose that $\mathbf{D}$ is a data matrix incorporating both dependent and independent variables such that $\mathbf{D} = \{\mathbf{Y}, \mathbf{X}\}$. Furthermore, suppose that $\mathbf{D}$ can be partitioned into observed and missing observations $\mathbf{D}: \{\mathbf{D}_{obs}, \mathbf{D}_{mis}\}$. Additionally, let $\mathbf{M}$ be a matrix with the same dimensions as $\mathbf{D}$, and allow $M_{ij} = 1$ when $M_{ij} \in \mathbf{D}_{obs}$ and $M_{ij} = 0$ when $M_{ij} \in \mathbf{D}_{mis}$.

2.1. Missing Completely at Random

Data that is MCAR is considered to be the least serious in practice. Theoretically, this type of data does not introduce bias for parameter estimates if the pattern of missingness is independent of the data itself, $P(M \mid D) = P(M)$. The actual $P(M) \sim \text{BIN}(n, p)$ so that if a random variable q is based on an underlying discrete event (like the roll of dice), $q \sim \text{DISUNIF}(f, g)$ and if q is based on a continuous event (like a random number

[1] Assistant Professor of Political Science; Graduate Director and Pre-Law Advisor, Department of Political Science and Law Enforcement, Minnesota State University, Mankato; Minnesota, USA: email address: granbj@mnsu.edu, jeffrey.granberg-rademacker@mnsu.edu

generator), $q \sim \text{UNIF}(f,g)$. If q follows a discrete distribution, then the probability, p, is based on the probability mass function of q such that:

$$p = pmf(q) = \begin{cases} 1/n, & q = q_1,\ldots,q_n \\ 0, & otherwise \end{cases}$$

and if q follows a continuous distribution, then p is based on the probability density function:

$$p = pdf(q;f,g) = \begin{cases} 1/(g-f), & f < q < g \\ 0 & otherwise \end{cases}$$

Because p is a random probability independent of $\mathbb{D}$, the parameter estimates of a linear model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e} \qquad (1)$$

are unbiased (Greene, 2000): $E(B_k) = \beta_k$ since

$$E(B) = E\left(\left[X_{obs}'X_{obs} + X_{mis}'X_{mis}\right]^{-1}\left[X_{obs}'X_{obs}B_{obs} + X_{mis}'X_{mis}B_{mis}\right]\right)$$
$$= E\left(\left[X'X\right]^{-1}\left[X'XB\right]\right) = \beta$$

Unfortunately, MCAR missingness patterns are seldom encountered in practice. If a respondent were to answer or refuse to answer survey questions based on rolls of dice or on the results of a random number generator, the missingness pattern observed in the data would be MCAR.

2.2. Missing at Random

In practice, data that exhibits an MAR pattern of missingness is much more serious than MCAR missingness. The unbiasedness assumption regarding parameter estimates no longer holds under MAR, because $P(M \mid D) = P(M \mid D_{obs})$. This problem can be mitigated if the pattern of missingness can be explained by at least one or more observed variables in $\mathbb{D}$. In practice, analysts can often correct for MAR missingness by including more variables in an imputation process (King et al, 2001). This is done by treating each individual D_{mis} as a parameter to be estimated. Once estimated, $\hat{D}_{mis}$ is then used in place of D_{mis} and standard estimation procedures can continue like normal. In linear models such as those shown in Equation 1, the parameter estimates of the linear model itself are unbiased:

$$E(B) = E\left(\left[X_{obs}'X_{obs} + \hat{X}_{mis}'\hat{X}_{mis}\right]^{-1}\left[X_{obs}'X_{obs}B_{obs} + \hat{X}_{mis}'\hat{X}_{mis}\hat{B}_{mis}\right]\right)$$
$$= E\left(\left[X'X\right]^{-1}\left[X'X\hat{B}\right]\right) = \beta$$

provided that the imputation procedure allows for $E(\hat{D}_{mis}) = D_{mis}$ to hold true.

MAR missingness can take different forms. One example of how data might exhibit MAR missingness is if low assessment scores are less likely to be reported on a survey than high assessment scores. The pattern of missingness can be explained by assessment scores, and is thus MAR, provided that at least some survey respondents reported assessment scores. Imputation procedures also improve their ability to model the missingness pattern if there are other variables in the model that explain assessment scores.

2.3. Nonignorable

The most serious and difficult pattern of missing data to handle is NI. This occurs when $P(M \mid D)$ does not simplify. In this case the pattern of missingness, **M**, is not independent of **D**. More specifically, the pattern of missingness is dependent in whole or in part on the unobservable values of D_{mis}. Therefore, attempting to estimate D_{mis} will produce biased estimates, $E(\hat{D}_{mis}) \neq D_{mis}$, leading to biased model estimates $E(B) \neq \beta$. A pattern of NI would occur if low assessment scores were less likely to be reported on a survey, *and* no other observed variables can predict which respondents have low assessment scores.

It is worthwhile to note that a significant amount of work has been done in the field of NI nonresponse in experimental designs. Some notable studies include Scharfstein et al (1999, pp. 1096-1120) use of semi-parametric nonresponse models, Tang's (2003, pp.747-764) use of the pseudo-likelihood method to identify multivariate regression parameters with NI missingness in the data, and Qin et al (2002, pp.193-200) use semi-parametric likelihood to handle NI missingness.

2.4. Missing Data in Social Science

In practice, the line between MAR and NI missingness is blurry. Data may exhibit MAR-like patterns of missingness, whereby other observed variables may be useful in explaining the overall pattern of missingness (King et al 2001, Rubin 1996, Schafer 1997). Yet these patterns of missingness are rarely likely to be completely and truely MAR. More likely is that some instances of missingness can not be explained by variables included in the imputation. If the missingness pattern is expressed as explained and unexplained probabilistic portions, $P(D_{mis}) = P(\breve{D}_{mis}) + P(\hat{D}_{mis})$ where $P(\breve{D}_{mis})$ is the probability that can be explained by D_{obs} and $P(\hat{D}_{mis})$ is the probability that cannot be explained by D_{obs}, then any instance where $P(\hat{D}_{mis}) > 0$ is going to introduce bias into the imputed missing values because $E(\hat{D}_{mis}) \neq D_{mis}$, with the bias becoming more pronounced as $P(\hat{D}_{mis})$ increases. The objective is to then estimate $\hat{D}_{mis}$ so that $\max\left[P(\breve{D}_{mis})\right]$ and $\min\left[P(\hat{D}_{mis})\right]$.

III. DIFFERENT APPROACHES TO HANDLING MISSINGNESS

The following subsections will briefly discuss several different approaches to handling missingness in datasets. Overall, the order of these approaches presented will go from simple to complex.

3.1. Doing Nothing: Listwise Deletion

One option to handling missing data is to ignore the problem and do nothing. This option is theoretically viable if and only if the data missingness pattern is truly MCAR. Specifically, all cells in $\mathbf{M}$ must be independent of $[D_{obs}, D_{mis}]$. If this condition is not met for any cell in $\mathbf{M}$, then $P(\hat{D}_{mis}) > 0$, thereby introducing bias into parameter estimates. The impact of this bias can vary, and may result in different parameter estimate magnitudes or signs of causal or descriptive inferences (Anderson et al 1983).

Further compounding the problem of missingness is what happens to some observed data when the problem is ignored in multivariate analyses. Most standard multivariate statistical methods used in practice (linear regression, ANOVA, etc) are incapable of handling missing observations without correction. This leads many social science researchers to simply rely on statistical package default settings to handling missing observations without being fully aware of the assumptions or ramifications of such a choice. In just about every major statistical software package for social science (SPSS, STATA and SAS among others), the default mechanism for handling missing data is listwise deletion.

In many ways, listwise deletion is a cure that is worse than the illness. Instead of now having one missing observation on one variable, listwise deletion eliminates the entire row of data from the analysis if one observation in one variable is missing. This practice essentially *throws away good data.* Table 1 illustrates how listwise actually compounds the problem of missing data by eliminating validly-observed data points from the analysis. Instead of one missing observation in Variable #2, listwise deletion actually throws away observations in Variables #1, #3, and #4 so that the analysis can be conducted.

Table 1: An Illustration of How Listwise Deletion Excludes Data

Variable #1	Variable #2	Variable #3	Variable #4
Observed	Observed	Observed	Observed
Observed	Observed	Observed	Observed
Excluded	*Missing*	*Excluded*	*Excluded*
Observed	Observed	Observed	Observed

This is theoretically valid if the missingness pattern exhibited in the data is truly MCAR. However, researchers should bear in mind that listwise deletion actually *compounds* the problem of missing data, by selectively eliminating observed data points from the analysis. This is clearly illustrated in Table 1, where one missing observation ($1/16 = 0.0625$ or 6.25% missing) handled by listwise deletion increases the missing problem three-fold ($4/16 = 1/4 = 0.25$ or 25% missing). If the missingness is not MCAR, listwise deletion is likely to only magnify the problem of bias since more observations are being systematically excluded from the analysis.

3.2. Mean Imputation

Another method of dealing with missing data is to use mean imputation. This approach is quite simple: replace missing observations with the average observed value, such that $D_{mis_{i,j}} = \overline{D}_{obs_j} \forall i, j$. Mean imputation, much like listwise deletion, only holds valid if the

missingness pattern is truly MCAR. Additionally, mean imputation is unable to incorporate additional variables which may be able to explain the some aspects of the pattern of missingness. So if the missingness pattern in the data is MAR and D_{obs_j} cannot explain the missingness, then mean imputation is inappropriate and the imputed values are going to be biased: $E\left(\hat{D}_{mis_j}\right) \neq D_{mis_j}$. Common sense suggests that it will be difficult if not impossible for D_{obs_j} to say anything meaningful about D_{mis_j} because mean imputation only considers one variable, j, in its imputation—and it is this same variable that is exhibiting the missingness pattern. So while mean imputation does not exacerbate the missingness problem by discarding data in the same manner as listwise deletion, the inadequate (and impractical) handling of MAR missingness suggests limited usage.

3.3. Multiple Regression Imputation with Stochastic Substitution

Missing data points can also be imputed by modeling the patterns of missingness, and then using that information to derive "plausible" values to explain the missingness. This approach works well when the missingness pattern is MAR. This assumption is crucial to obtaining unbiased parameter estimates. The multiple regression imputation with stochastic substitution procedure is quite simple. Suppose that $\mathbf{D} = \{\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2\}$, and $[D_{obs}, D_{mis}] = \{[Y_{obs}, Y_{mis}], [X_{1,obs}, X_{1,mis}], [X_{2,obs}]\}$, where X_2 is fully observed and Y, X_1 are only partially observed. Regression coefficients are then obtained utilizing $[D_{obs}, D_{mis}]$:

$$Y = B_0^* + B_1^* X_1 + B_2^* X_2 + e^*$$
$$X_1 = B_0^\dagger + B_1^\dagger Y + B_2^\dagger X_2 + e^\dagger$$

Each of the i missing values of Y to be imputed are then constructed as a linear combination of each of the i^{th} values of X_1 and X_2 given B_1^*, B_2^* and e^*:

$$\hat{Y}_i = B_0^* + B_1^* X_{1,i} + B_2^* X_{2,i} + e_r^*$$

where e_r^* is a randomly chosen residual. The inclusion of e_r^* is important because it acts as a buffer against "over-correcting" the imputed values. The importance of random noise as a means to buffer against imputed value over-correction has been discussed elsewhere (Rubin 1977, 1987). The procedure follows in the same fashion, with each variable that contains missing data being modeled as a dependent variable for the purposes of imputing the missing data on that variable:

$$\hat{X}_{1,i} = B_0^\dagger + B_1^\dagger Y_i + B_2^\dagger X_{2,i} + e_r^\dagger$$

where $e_r^\dagger$ is a randomly chosen residual. It is important to stress that this procedure depends on the idea that D_{mis} is an MCAR or MAR pattern. In such instances, $E\left(\hat{Y}_i\right) = Y_i$ and $E\left(\hat{X}_{1,i}\right) = X_{1,i}$. If the missingness follows a NI pattern, the imputed values (and the subsequent parameter estimates) will be biased such that $E\left(\hat{Y}_i\right) \neq Y_i$ and $E\left(\hat{X}_{1,i}\right) \neq X_{1,i}$.

3.4 Multiple Imputation: Expectation-Maximization with Importance Resampling (EMis)

The Expectation-Maximization with Importance Resampling (EMis) algorithm is a multiple imputation algorithm that is designed to impute D_{mis} by creating m number of datasets where the missing data is imputed with "plausible" estimates. The choice of m has largely been settled by the work of Rubin (1987) and Wang and Robbins (1998), which demonstrates that parameter estimates with $m=5$ imputed datasets are quite reasonably efficient, and that the relative efficiency of estimates with $m=10$ is nearly the same as that of estimates with $m=\infty$. This result is also echoed by King et al (2001), who recommend $m=5$ to be adequate for most social science research except in extreme cases where missingness is exceptionally high.

To illustrate how the EMis algorithm works, let $\tilde{D}_{ij}$ denote the simulated value of the i^{th} observation in the j^{th} variable. Also, let $\tilde{D}_{i,\neg j}$ denote the vector values of all observed variables except j. It is possible to create an imputation:

$$\tilde{D}_{ij} = \tilde{D}_{i,\neg j}\tilde{\beta} + \tilde{\varepsilon}$$

where $\tilde{\beta}$ can be calculated from μ and Σ utilizing the likelihood function:

$$L(\mu,\Sigma \mid D_{obs}) \propto \prod_{i=1}^{n} N(D_{obs,i} \mid \mu_{obs,i}\Sigma_{obs,i})$$

Imputation draws are then taken from the conditional predictive distribution of the missing data, and each individual draw is replaced with that observation's predictive value:

$$\tilde{D}_{ij}\text{: } P(D_{mis} \mid D_{obs}, \tilde{\mu}, \tilde{\Sigma})$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ can be iteratively estimated by finding the maximum posterior estimates $\tilde{\phi} = (\hat{\mu}, \hat{\Sigma})$ of each based on the posterior distribution $P(\mu, \Sigma \mid D_{obs}, \tilde{D}_{mis})$. The parameters are then put on unbounded scales, using the log of the standard deviations and Fisher's z for the correlations. An acceptance-rejection algorithm is then used to determine whether draws of $\tilde{\phi}$ are kept or rejected based on whether the following expression is true (keep $\tilde{\phi}$) or false (reject $\tilde{\phi}$):

$$P(\tilde{\phi}) \propto \frac{L(\tilde{\phi} \mid D_{obs})}{N(\tilde{\phi} \mid \tilde{\phi}, V(\tilde{\phi}))}$$

Note that the right hand side of this proportionality is the ratio of the actual posterior to the asymptotic normal approximation, both evaluated only at $\tilde{\phi}$.

It is also worth noting that the estimates generated from the EMis algorithm are asymptotically identical to estimates generated by other Bayesian algorithms, such as the Imputation Posterior algorithm (which utilizes Markov Chain Monte Carlo methods), and frequentist algorithms like the Expectation-Maximization (without importance resampling). For more information about the EMis algorithm, and how it compares to other notable algorithms, see King et al (2001) and Honacker et al (2001).

4. MONTE CARLO COMPARISONS

A total of $n = 1000$ observations were generated for $l = 1000$ datasets from a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}$ and variance-covariance matrix:

$$\boldsymbol{\Sigma} = \begin{Bmatrix} 1 & 0.4 & -0.3 & 0.1 & 0.05 & 0.4 \\ 0.4 & 1 & 0.1 & 0.1 & 0.1 & 0.4 \\ -0.3 & 0.1 & 1 & -0.1 & 0.1 & 0.4 \\ 0.1 & 0.1 & -0.1 & 1 & 0.1 & 0.4 \\ 0.05 & 0.1 & 0.1 & 0.1 & 1 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 1 \end{Bmatrix}$$

The values of $\boldsymbol{\Sigma}$ were chosen to mimic conditions typically found in social science data for the respective variables: $\{Y, X_1, X_2, X_3, X_4, X_5\}$. Specifically, the generated data is multivariate in nature, with varying covariance levels between the dependent and independent variables ($\sigma_{2,1} = 0.4, \sigma_{3,1} = -0.3, \sigma_{4,1} = 0.1, \sigma_{5,1} = 0.05$), and a limited amount of independent variable interdependency (values of $\sigma = \{0.1, -0.1\}$). Note that the last column and row of $\boldsymbol{\Sigma}$ are largely comprised of the value 0.4. This was done specifically to model partial MAR conditions, since X_5 (which is the variable that corresponds to the last column and row of $\sigma = 0.4$), is the variable used to generate the patterns of missingness. Allowing for some covariance between X_5 and the data variables Y, X_1, X_2, X_3 and X_4 allows the data to somewhat fit a MAR definition whereby the existing data can (to some extent) predict the pattern of missingness. The lower the covariance between X_5 and the data variables, the more the missingness pattern becomes NI and the less it becomes MAR. A covariance of 0.4 across all data variables means that the missingness pattern is mostly MAR, but not completely. This is mixed-type of missingness (still mostly MAR) is a typical case in social science research.

Three new variables were created in each generated dataset to simulate three levels (10% missingness, 30% missingness, and 50% missingness) of partial MAR-patterned missingness on Y. To construct these three new variables the quantile values of X_5, denoted as κ, were taken at 0.9, 0.7, and 0.5. Missingness patterns were then reflected in the newly created variables:

$$Y^*_{\kappa=0.9,i} = \begin{cases} Y_i & X_{i,5} \le \kappa_{0.9,i} \\ Missing & X_{i,5} > \kappa_{0.9,i} \end{cases}$$

$$Y^*_{\kappa=0.7,i} = \begin{cases} Y_i & X_{i,5} \le \kappa_{0.7,i} \\ Missing & X_{i,5} > \kappa_{0.7,i} \end{cases}$$

$$Y^*_{\kappa=0.5,i} = \begin{cases} Y_i & X_{i,5} \le \kappa_{0.5,i} \\ Missing & X_{i,5} > \kappa_{0.5,i} \end{cases}$$

Regression parameter estimates and standard errors were generated for full datasets as well as the imputed datasets, with X_1, X_2, X_3, and X_4 as the independent variables, and Y and Y^*_κ as the dependent variables.

It is important to note that several other simulations were conducted by the author varying the elements of $\boldsymbol{\Sigma}$ from positive to negative, and also varying the degree and nature of the missingness. Changing the elements of $\boldsymbol{\Sigma}$ from positive to negative yielded no significant changes in the results, since the bias measures used in the following sections have been designed to detect both positive and negative bias. Other varying degrees of missingness were consistent with the results presented here (more missingness meant more bias). Also, bias increased as the missingness pattern became less MAR and more NI. These additional results are not presented here because they do not significantly change the findings, but are available from the author by request.

4.1 Assessing Coefficient Bias

If parameter estimates are unbiased then $E(B) = \beta$. Using the mean square error (MSE), it is thus possible to assess the asymptotic unbiasedness of the regression parameter estimates since:

$$\lim_{l\to\infty}\left(\frac{1}{l}\sum_{m=1}^{l}(B_m - \beta_m)^2\right) = 0$$

if the procedure for handling the missing data values is unbiased. Figures 1, 2, and 3 map out the kernel densities of the parameter squared error averaged across the regression estimates for each of the $l = 1000$ datasets:

$$E\left[(B_p - \beta_p)^2\right] \qquad (2)$$

where $p = 0,\ldots,4$. If B_p is unbiased, then the term in Equation (2) should tend toward zero on the positive side.

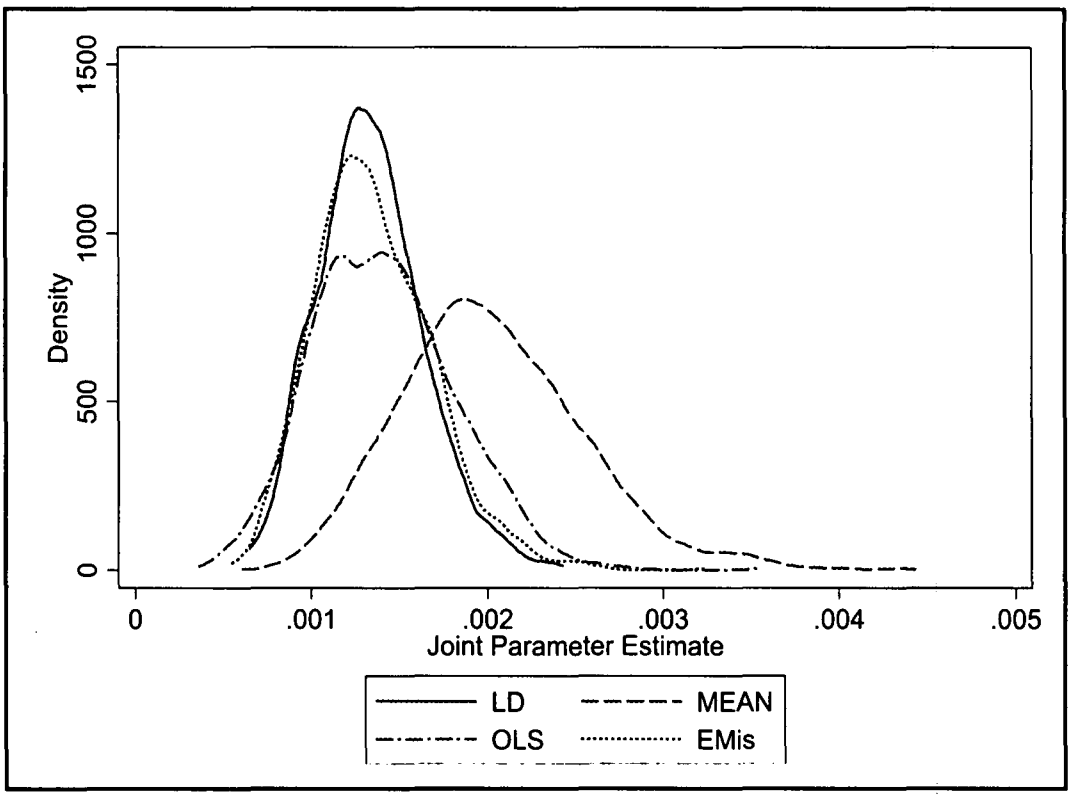

Figure 1: Joint Parameter Estimates with 10% MAR Missingness

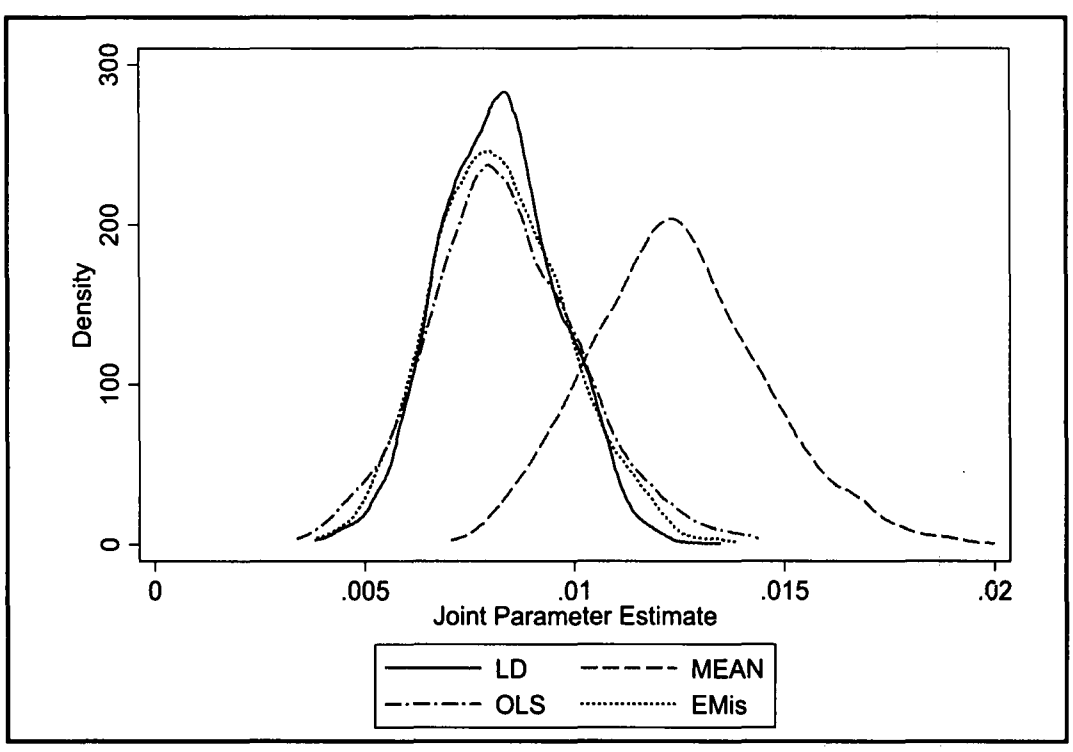

Figure 2: Joint Parameter Estimates with 30% MAR Missingness

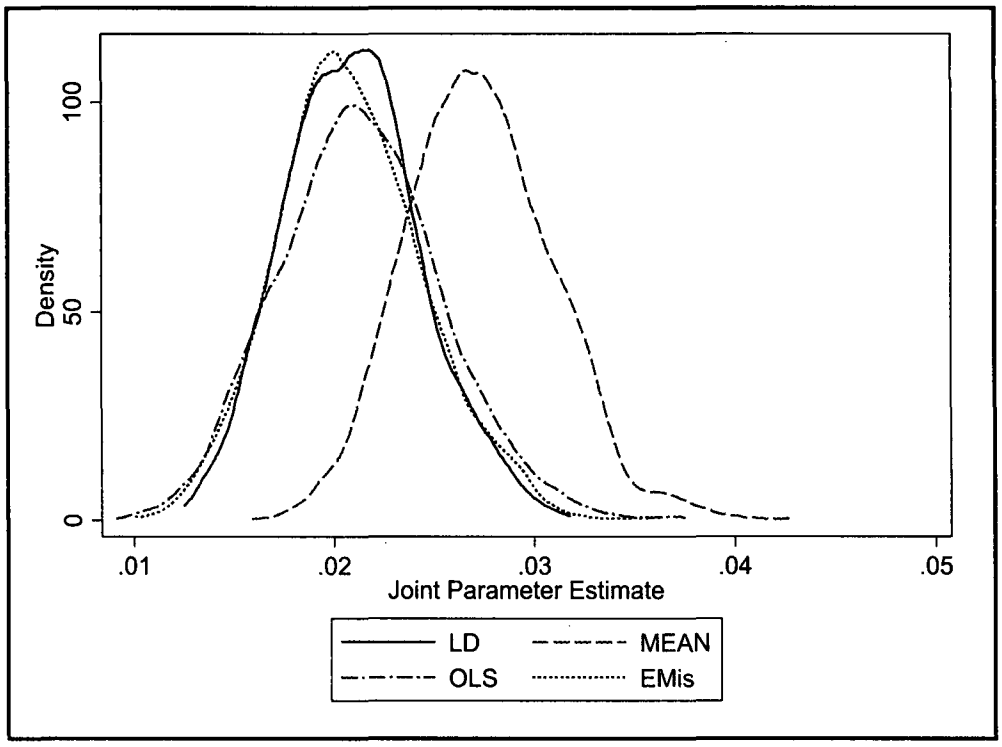

Figure 3: Joint Parameter Estimates with 50% MAR Missingness

One can see that listwise deletion, multiple regression with stochastic substitution and EMis generated the least biased estimates, and mean substition produced the most biased. Also note how the tails of each of the kernel densities become heavier as the missingness increases (note the x-axis range)—as the amount of uncertainty increases, so too does the distribution of the coefficients.

One may be inclined to wonder why listwise deletion seems to have performed so well. One possible explanation is that this a large-sample simulation. The problems of listwise deletion are exacerbated most in small-sample situations (King et al 2001). Also, keep in mind this simulation recreates partial MAR missingness, not total MAR missingness. Partial MAR missingness (with other elements of MCAR and NI patterns also acting on the data) is more likely to reflect social science data, where the variable responsible for the missingness (X_5) is itself unobserved, but other observed variables are partially correlated with it. Table 2 shows the MSE for each procedure across missingness percentages.

Table 2: Mean Square Error by Procedure and Missingness

Procedure	10% Missing	30% Missing	50% Missing
LD	0.00134	0.00818	0.02094
MEAN	0.00203	0.01246	0.02729
OLS	0.00141	0.00834	0.02133
EMis	0.00136	0.00822	0.02086

As can be seen from Table 2, LD performed comparable to Multiple Regression with Stochastic Substitution and the EMis algorithm, only taking a small uptick in coefficient bias when 50% of the rows are observed. Mean substitution consistently produced the most biased

parameter estimates under each condition, but its performance was only moderately worse than the other approaches.

4.2 Assessing Standard Error Bias

Standard errors of parameter estimates are important to researchers, because of their usefulness in hypothesis testing and assessing parameter significance. However, biased standard errors can lead to Type I and Type II errors. Assessing the bias of standard errors can be done by examining if $E[se(B)] = se(\beta)$. The bias can be further isolated so that the inclination toward either a Type I or Type II error is probabilistically known:

$$se(B) - se(\beta) \begin{cases} = 0 & \text{Unbiased} \\ < 0 & \text{Type I Error} \\ > 0 & \text{Type II Error} \end{cases}$$

where $P(se(B) - se(\beta)) : N(0, \sigma^2)$.

Figures 4 shows the joint standard error bias for each procedure as a deviation from the fully observed joint standard error for each level of missingness. Listwise deletion and mean substitution led to the greatest amount of bias in estimated standard errors. The direction of the bias is also important, particularly for the listwise deletion estimates, since a negatively biased standard error increases the likelihood of Type I errors. Additionally, parameter standard error bias becames dramatically more pronounced in estimates that use listwise deletion and mean substitution as the level of missingness increases. The positive bias in the standard errors when mean substitution is used makes Type II errors more likely since the inflated standard errors are going to lead to lower the values of test statistics.

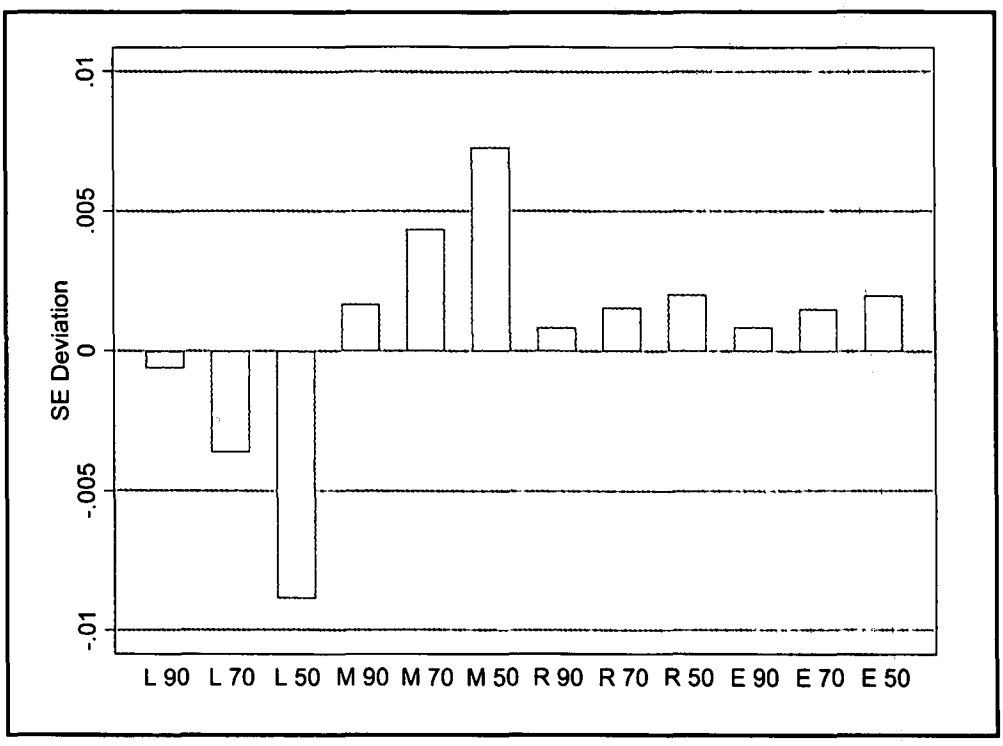

Figure 4: SE Bias by Procedure and Missingness Level

Multiple regression with stochastic substitution and the EMis algorithm imputed observations that lead to less-biased standard errors in the estimated model, especially when missingness is high. The increase in the standard error bias as the missingness level increased was minimal. Table 3 shows the numeric values of the joint standard error biases. EMis imputed values led to the least biased standard errors when missingness was over 10%, though there was little significant difference between EMis and multiple regression with stochastic substitution.

Table 3: Value and Direction of SE Bias Based on Missingness Level

Procedure	10% Missing	30% Missing	50% Missing
LD	-0.00062	-0.00362	-0.00885
MEAN	0.00169	0.00437	0.00731
OLS	0.00085	0.00156	0.00202
EMis	0.00085	0.00151	0.00198

V. CONCLUSION

When deciding how to handle missing data, it is important for the researcher to be aware of the underlying assumptions inherent in missing data procedures. Linear models utilizing listwise deletion and mean imputation would produce unbiased estimates of both parameter estimates and standard errors that are unbiased if the data missingness follows an MCAR pattern. Unfortunately, this seldom happens in practice.

Multiple regression with stochastic substitution and the EMis algorithm impute missing data well for both MCAR and MAR patterns of missingness. The results of this analysis show relative unbiasedness in parameter estimates and standard errors, even when the missingness pattern is only partly MAR. However, a serious limitation of multiple regression with stochastic substitution is that it parametrically assumes a linear relationship which may be inappropriate--thereby leading to biased imputations. EMis imputations may produce biased estimates if the normal approximation is inappropriate, though the heavier-tailed *t*-distribution with a larger variance matrix and additional factors, can correct this problem (King et al 2001).

All GAUSS files and code used in this article are available from the author.

Acknowledgement

The author thanks Dr. Joselito Magadia and the anonymous reviewer for their very helpful suggestions for improving this article. Any errors or omissions lie solely with the author.

References

ANDERSON, A. B., BASILEVSK, A., & HUM, D. P. J. (1983). "Missing Data: A Review of the Literature" in P. H. Rossi, J. D. Wright, A. B. Anderson. (eds). *Handbook of Survey Research*, pp. 415-494. New York: Academic Press.

GREENE, W. H. (2000). *Econometric Analysis*, 4th ed. Upper Saddle River, NJ: Prentice Hall.

HONACKER, J., KATZ, J., & KING, G. (2001). "An Improved Statistical Model for Multiparty Electoral Data." Social Science Working Paper 1111. California Institute of Technology, Division of the Humanities and Social Sciences.

KING, G., HONACKER, J., JOSEPH, A., & SCHEVE, K. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review*, 95, 49-69.

LITTLE, R. J. A., & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. New York, NY: Wiley-Interscience.

QIN, J., LEUNG, D., & SHAO, J. (2002). "Estimation with Survey Data Under Nonignorable Nonresponse or Informative Sampling." *Journal of the American Statistical Association*, 97, 193-200.

RUBIN, D. B. (1977). "Formalizing Subjective Notion about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association*, 72, 538-543.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.

RUBIN, D. (1996). "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association*, 91, 473-489.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

SCHARFSTEIN, D. O., ROTNITZKY, A., & ROBINS, J. M. (1999). "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models." *Journal of the American Statistical Association*, 94, 1096-1120.

TANG, G., (2003). "Analysis of Multivariate Missing Data with Nonignorable Nonresponse." *Biometrika*, 90, 747-764.

WANG, N., & ROBINS, J. (1998). "Large-Sample Theory for Parametric Multiple Imputation Procedures." *Biometrika*, 85, 935-948.