

The Missing Link: Data Analysis with Missing Information

Venita DePuy, Duke Clinical Research Institute, Durham, NC

ABSTRACT

How do you handle missing data? Deletion of those subjects frequently leads to biased outcomes. Mean imputation assumes that non-responders are no different than responders, and can bias variances toward zero. Last observation carried forward methods, while still often used, can cause bias and even induce an apparent treatment effect. Multiple imputation is an improved method to deal with these issues. This paper will focus on the Markov chain – Monte Carlo based method of multiple imputation using SAS®'s PROCs MI and MIANALYZE.

INTRODUCTION

One of the leading concerns in data analysis is how to appropriately incorporate missing information. This paper will discuss different types and causes of missing data, review case deletion and single imputation methods, and discuss the use of multiple imputation, with a focus on the MCMC-based method for arbitrary missing data patterns.

While single imputation may not have a significant effect on subsequent analyses when only small amounts of information are missing, it frequently adds bias and distorts the relationship between variables. In contrast, multiple imputation computes statistics based on several different datasets, which maintain the relationships between variables and provide a measure of the uncertainty in the estimates.

The attached appendix provides notated SAS® code for a variety of relationships between variables utilizing PROCs MI and MIANALYZE.

CAUSES AND CLASSIFICATIONS OF MISSING INFORMATION

Missing information can occur in data for a variety of reasons. In clinical trials, information is typically collected at scheduled visits. A subject in a clinical trial may not complete all items on a questionnaire, or may miss an entire visit, resulting in no data collection at that time point. Subjects may miss visits for reasons unrelated to the study, such as transportation difficulties, or for reasons potentially related to study medications, such as experiencing an adverse event. Similarly, item-level missing information can occur for both study-related and unrelated reasons. It is not uncommon for these types of missing data to be followed by observed data at subsequent time points. This pattern of missing data may be followed by observed data, are considered an arbitrary pattern of missing data.

On the other hand, if a subject withdraws from the study or dies, future scheduled observations will be missing. In other words, if an observation is missing, all subsequent observations are also missing. This is called a monotone pattern of missing data, and allows for more flexibility in analysis choices than arbitrary patterns.

MISSINGNESS MECHANISMS

We shall refer to the matrix of complete data as Y , which is composed of columns of p variables and rows of n subjects. Y can be separated into two parts: Y_{obs} , the observed data, and Y_{mis} , the missing data. We can also create a matrix R of response indicators (elements r_{ij} are 0 if y_{ij} is missing, 1 if observed).

The simplest case, missing completely at random (MCAR), implies a random, arbitrary pattern to the missing data. In other words, subjects with missing data are like a random subsample of the data; there is no difference between responders and non-responders. For instance, if every subject was equally likely to record his or her weight, the missing weight data would be missing completely at random. In technical terms, MCAR indicates that the probability of missingness is independent of the data. In other words,

$$P(R|Y) = P(R|Y_{obs}, Y_{mis}) = P(R)$$

Unfortunately, the MCAR assumption is rarely realized. A less restrictive, and more realistic scenario is that the data are missing at random (MAR). This assumption states that the probability of missingness depends only upon observed variables. For example, if women were less likely to record their weight, and gender was recorded for every subject, the probability of missing information would be MAR. Obviously, this assumption becomes more likely as more variables are recorded. It is possible to test the MCAR assumption against the MAR¹, although multiple imputation methods can be used on MCAR data as well as MAR data. Since the missingness is independent of the unobserved data, the missing at random assumption can also be written as

$$P(R|Y) = P(R|Y_{obs}, Y_{mis}) = P(R|Y_{obs})$$

The third mechanism, missing not at random (MNAR), is the most difficult to analyze appropriately. In this case, missingness is dependent on the unobserved data. For example, subjects with very high incomes are less likely to report income on a survey². The probability that the income level is missing is dependent on the value of the missing data, which violates the MAR assumption.

MNAR data are often analyzed via selection models or pattern mixture models³, which are beyond the scope of this paper. Molenberghs et al⁷ discuss the drawbacks to MNAR analyses, and suggest that their optimal place is within a sensitivity analysis. It also should be noted that it is impossible to test the MAR assumption against the MNAR alternative without additional information⁴.

METHODS OF HANDLING MISSING INFORMATION

Case Deletion

The simplest manner of handling missing information is to delete the subjects with missing values, a method often referred to as case deletion, or the complete-case analysis method. This method has been commonly used in the past, and is still the default in many analysis programs, including SAS. While case deletion is easy to do, it makes the assumption that the missing data have exactly the same distribution as the observed data (in other words, that the data are MCAR). Since this is rarely the case, and subjects with complete information may systematically differ from those with missing data, this method of exclusion has a strong potential bias as well as the potential for underestimating the standard error. Reducing the sample size in this manner can reduce the power of the analyses as well. However, if very few cases are incomplete (<5%), case deletion may be an acceptable method⁶.

Weighting

One way to make the data better reflect the actual population being sampled, when using the case deletion method, is by a system of weighting similar to that employed in survey sampling. Weighting to compensate for non-response involves weighting each respondent by the inverse of the response rate for that combination of covariates.

For example, consider a simplified case of gender (g) and age group (a), simplified to pre-adults and adults. If n_{ga} people are sampled from each of these four combinations of gender and age, and r_{ga} respond from each group, then each respondent should be weighed by $w_{ga} = n_{ga}/r_{ga}$ ⁷. In other words, if 60 of 100 female adults respond, each of those women should be weighted by 100/60 or 1.67. Similarly, if 50 of 55 male adults respond, each of those respondents should be rated by 55/50 or 1.10.

This weighting method may be used to correct for bias, but still discards partial information from subjects with missing values, where single and multiple imputation methods do not.

Single imputation

One common method of dealing with missing information while also retaining the information contained in subjects with only partial data is to replace each missing value with a value derived from the data set. While this has an advantage over case deletion in that it allows the entire data set to be used in analysis, there are several key drawbacks. First, the standard errors of the calculated statistics (such as regression coefficients) are rarely adjusted to reflect the uncertainty inherent in imputed data. Second, these imputations may cause systematic bias. For example, it seems apparent that, as a subject's health declines, at some point he or she will be too ill to attend a scheduled appointment. If the last recorded observation is substituted for the missing information, that subject will appear to maintain a constant state of health, resulting in falsely optimistic results. A variety of different methods of single imputation are used today; some of the most common methods are addressed below.

Mean or Median Substitution

One simple method of imputing data is to substitute the mean (or median) of the non-missing values of that variable. Mean imputation of baseline variables in randomized trials can be a reasonable method if variables are not imputed within arms, which results in lost precision and underestimation of standard errors, and missingness of baseline does not predict the outcome⁸. Arnold and Kronmal⁹ demonstrated that mean imputation of baseline variables in a large observational study produced very consistent results.

Mean or median substitution of covariates and outcome variables is still frequently used. This method is slightly improved by first stratifying the data into subgroups (by gender and age category, for example) and using the subgroup average. The primary drawback to this method is that non-response bias is ignored, and data are assumed to be MCAR, or MCAR within subgroups. Subjects who are too sick to attend an appointment are given the average value of subjects who are well enough to attend, which could lead to an overly optimistic estimate. Mean/median imputation results in the mean or median of the entire data set being the same as it would be with case deletion, but the variability between individuals' responses is decreased, biasing variances and covariances toward zero⁶.

Another facet of mean substitution is the probability imputation technique introduced by Schemper and Smith¹⁰. These researchers demonstrated that this method maintains the correct type I error rate when that covariate is independent of the treatment, and that the power exceeds that of the case deletion method.

In spite of its potential flaws, mean imputation has some significant advantages over the case deletion method. Mean imputation is readily available in SAS's PROC Standard.

Hot Deck Imputation

Hot deck imputation replaces each missing observation with a value randomly drawn from the observed data via sampling without replacement. As with mean/median substitution, this method is more accurate when stratifying prior to imputation. The primary drawbacks are the lack of guidance in creating subgroups, which may result in groups with very few observations to draw from, and the potential to distort relationships among variables. Hot deck imputation also assumes no difference between respondents and non-respondents, i.e. that data are MCAR, and may grossly underestimate variability by treating imputed data as observed data².

Regression

There are two types of regression (single) imputation that use models of the non-missing data to predict values of the missing data. In the simpler case, the predicted value from the regression replaces the missing value. A method more consistent with the true distribution of the data is to impute the predicted value plus a residual, where the residual is normally distributed with mean of zero and variance equal to the mean squared error. This latter method can be done by conducting the regression, creating a normally distributed random variable as specified, and then adding them together before inserting in the data set. Both of these methods inflate the correlation between variables and require model specification. They may also potentially result in biased parameter and standard error estimates.

Neighboring observations

The most common method of this type is the Last Observation Carried Forward (LOCF) method. Engels and Diehr¹¹ found that LOCF performed the best of the single imputation methods based only on data prior to the missing information, although Schafer¹² refers to this model as inferior to regression because it ignores regression to the mean. Molenberghs et al also noted that the bias in the LOCF estimator typically does not vanish under MCAR assumptions. In addition, the bias can be either positive or negative, and can even induce an apparent treatment effect when there is none⁴.

When data were available after the missing period, the Next Observation Carried Backward (NOCB) method performed the best of all single imputation methods in studies where patients tended to decline, but studies suggest that LOCF would be preferable if data over time reflected an increasing trend. Averaging the observations immediately before and after the missing data also performed well¹¹. These two methods are obviously only applicable for parts of arbitrary missing data for which subsequent data exists, and cannot address the problem of missing data due to dropout.

While these three methods have the advantage of only using data from a particular subject, they make the strong assumption that there is no change in the subject between the observed time points and the missing time period, which can lead to biased estimates and reduced variances.

The main drawback to single imputation methods is that subsequent analyses do not reflect the uncertainty inherent in missing data. The sample size is overstated, confidence intervals are too narrow, and Type 1 error rates are generally too high¹². Many single imputation methods are prone to the introduction of bias, although it is very difficult to measure how much bias is introduced. These problems become worse as the rates of missing information and the number of parameters increase. Various authors^{5,11} provide further comparisons.

Multiple Imputation

Multiple imputation (MI) is a method that involves generating multiple complete datasets and combining the individual estimates to obtain parameter estimates with standard errors that reflect the uncertainty inherent in imputation. MI allows all data pertinent to the missing information to be included in the simulation model, leading to more accurate parameter estimates. Missing values are predicted from each participant's observed values, and joint relationships among variables are estimated from all available data³. If the MAR assumption is reasonable, multiple imputation may provide less bias than other approaches, if the imputation model is correctly specified¹³.

Rubin² states that MI has the disadvantages of requiring more effort, more time, and more computer storage space than single imputation methods. In addition, a final unique answer is not produced in MI; instead, the combined datasets are used to produce estimates.

While many single imputation methods assume that data are MCAR, multiple imputation models generally assume that data is MAR. However, it should be noted that the MAR assumption is not a requirement of multiple imputation theory³, although it is assumed in programs such as SAS's PROC MI. The MAR assumption becomes more plausible as Y_{obs} is enriched with more variables.

Model Selection

The first step in conducting multiple imputation is to select a model. In most cases, a multivariate normal model is used⁶ which assumes that all variables are normally distributed, linearly related, and have normal homoscedastic error terms. As with other normality-based analyses, highly skewed variables should be transformed prior to analysis. Although the normality assumptions are rarely satisfied in a strict sense, simulations³ have shown this model to be robust to departures from normality^{3,14}. All variables which may help describe the missing data, and those which will be used in future analyses, should be included in the model. It should be noted that nominal variables are not appropriate for these multiple imputation models, and should be replaced with an appropriate number of dichotomous "dummy" variables⁶. SAS assumes a multivariate normal model when either the regression or MCMC methods of multiple imputation are used.

Generate Initial Estimates

Initial estimates are needed to begin the multiple imputation process. The most common method is to generate maximum likelihood estimates of the means and covariances based on the EM algorithm¹⁵. This iterative process continues until convergence and yields a maximum likelihood estimate¹⁶. Schafer⁶ provides more details on this Process. SAS also provides the option of entering initial estimates from a data set.

Data Augmentation / Imputation

Three main types of multiple imputation are available in PROC MI. Parametric regression models may be used for data with the multivariate normal model that has a monotone pattern of missing data. A regression model is fitted for each variable that has missing values, using the earlier variables as covariates. The Process is then repeated for the next missing time point and subsequent time points until all missing values are imputed.

The non-parametric propensity score method does not require the multivariate normal assumption, but is only applicable for data with monotone missingness. This method of imputation generates propensity scores for each variable with missing values to find the probability of missingness, then groups observations together by these scores. Approximate Bayesian bootstrapping is applied to sample values within each group for imputations. While this method is effective for inferences about the distribution of individually imputed variables, it is inappropriate when the purpose is to assess relationships between variables¹⁷.

The third option, Markov chain Monte Carlo (MCMC) methods, assumes the multivariate normal model but is appropriate for arbitrary missing data patterns. A Markov chain is a sequence of random variables in which each element's distribution depends on the value of the previous element. The MCMC method constructs multiple chains of variables long enough to allow the distributions to stabilize. Specifically, each imputed dataset is created by alternating the imputation of missing values and the simulation of the mean and covariances of the posterior population until the results converge to a stationary distribution. Further details can be found in Schafer⁶.

One of these three methods is used to impute m complete data sets. Schafer⁶ recommends that no more than 10 data sets are generally needed, although more may be needed if there is a great deal of missing information. Generally, five data sets are imputed.

Examination of imputed data

After the imputations are completed, variables should be changed to their appropriate form. For example, any skewed variables that were transformed prior to imputation should be re-transformed. Dichotomous variables with values of 0 or 1 may have imputed values that are fractions, and should be rounded to the correct value. It may also be possible for a subject to have contradictory imputed values, such as dichotomous variables indicating multiple races, which will need resolved. If imputations are calculated without limiting the ranges of imputations, data ranges should be examined to ensure imputed values are acceptable.

Analyzing results

After imputation is complete, the parameter and variance estimates, and any other analyses of interest, should be calculated separately for each data set. Then, overall estimates may be calculated with the following equations as per Rubin², which are also available in software packages such as SAS's PROC MIANALYZE. These statistics account for the variability of the imputations and, assuming the imputation model is correct, provide consistent estimates of the parameters and their standard errors¹³.

From the m different data sets, each with parameter estimate $\hat{Q}^{(t)}$ and variance estimate $\hat{U}^{(t)}$, we can calculate overall estimates as follows.

Point estimate, calculated as the mean of the m point estimates for parameter Q:

$$\bar{Q} = \frac{1}{m} \sum_{t=1}^m \hat{Q}^{(t)}$$

Within-imputation variance, calculated as the mean of the m variances for parameter Q:

$$\bar{U} = \frac{1}{m} \sum_{t=1}^m \hat{U}^{(t)}$$

Between-imputation variance for parameter Q:

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

Total resulting variance associated with $\bar{Q}$

$$T = \bar{U} + (1 + \frac{1}{m})B$$

Rubin² recommended the use of $\frac{\bar{Q} - Q}{\sqrt{T}} \sim t_v$ for confidence intervals and tests, where

$$v = (m-1) \left[1 + \frac{\bar{U}}{(1 + m^{-1})B} \right]^2$$

In cases where only a modest proportion of data is missing and the complete-data degrees of freedom v_0 is small, this computed degrees of freedom can be significantly larger than v_0 . Therefore, Barnard and Rubin¹⁸ recommend the following adjusted degrees of freedom:

$$v^* = \left[\frac{1}{v} + \frac{1}{v_{obs}} \right]^{-1} \quad \text{where} \quad v_{obs} = \frac{v_0 + 1}{v_0 + 3} v_0 \left[1 - \frac{(1 + m^{-1})B}{T} \right]$$

The relative increase in variance due to non-response is defined by Rubin² as

$$r = \frac{(1 + m^{-1})B}{T}$$

which is used to calculate the estimated fraction of missing information about Q:

$$\hat{\lambda} = \frac{r + 2/(v + 3)}{r + 1}$$

$\hat{\lambda}$ denotes how the missing data influence the uncertainty of the estimates, and can be noisy for small m ¹². Schafer^{6,12} notes that even with a large fraction of missing information, a relatively small number of imputations provides adequate standard error estimates.

PROGRAMMING

SAS introduced two procedures to conduct multiple imputation, PROC MI and PROC MIANALYZE in release 8.2. The imputation step is carried out in PROC MI, after which the appropriate analysis (correlation, regression, etc.) is conducted on m separate data sets. Appropriate estimates are then computed by combining results in the PROC MIANALYZE step. SAS also has the ability to limit potential variable ranges to those appropriate for each variable, and the ability to round imputed values within the imputation procedure.

FURTHER INFORMATION

The number of ways of dealing with missing data are greatly beyond the scope of this paper.

Suggestions for further reading include:

- Yuan's paper¹⁷, presented at SUGI 25, provides further information on the PROC MI options, and is very similar to SAS Online Documentation.
- Lanning and Berry¹⁹ suggested a large-sample alternative to PROC MI at SUGI 28.
- Paulin et al²⁰ discuss the limitations of PROC MI and introduce methods for model-based multiple imputation in their SUGI 29 paper.
- Raghunathan⁷ discusses various weighting methods and compares them to multiple imputation and likelihood construction methods, and comments that multiple imputation is perhaps the most practical method of these three.
- Barzi and Woodward²¹ compared the case deletion method to eight single imputation and four multiple imputation methods in 28 cohort studies. They concluded that, when < 10% of the data were missing, all methods gave similar results; when 10-60% were missing, clear differences existed and multiple imputation is the optimal choice, and no method was satisfactory when >60% were missing.
- Barnes et al²² introduced the completion score method of multiple imputation and compared it to other methods for monotone data, such as the predictive mean matching, Bayesian least squares, and modified propensity scores methods.

CONCLUSION

A variety of methods are available to handle missing data. However, care should be taken to prevent bias from being introduced and produce accurate parameter estimates and appropriate standard errors. Case deletion methods discard valuable information and assume that data are missing completely at random, which is seldom true in practice. Single imputation methods preserve the sample size but typically do not reflect the inherent uncertainty of imputation in their standard errors.

Multiple imputation, when used appropriately, provides statistically valid inferences from incomplete datasets. MCMC-based multiple imputation is valid for a wide range of analyses, even when data has an arbitrary pattern of missing data, provided that the covariates upon which missingness depends are included in the data set.

TRADEMARK INFORMATION

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates US registration.

REFERENCES

- [1] Diggle PJ, Liang KY, Zeger SL. (1994). *Analysis of Longitudinal Data*, Oxford: Clarendon Press.
- [2] Rubin, DB (1987). *Multiple imputation for nonresponse in surveys*, New York: Wiley.
- [3] Schafer JL, Graham JW. (2002). "Missing Data: Our View of the State of the Art." *Psychological Methods*, 7(2): 147-177.
- [4] Molenberghs G, Thijs H, Jansen I, Beunckens C.(2004). "Analyzing Incomplete Longitudinal Clinical Trial Data." *Biostatistics*. 5(3):445-64.
- [5] Little RJA, Rubin DB. (1987). *Statistical analysis with missing data*, New York: Wiley.
- [6] Schafer, JL. (1997). *Analysis of incomplete multivariate data*, London: Chapman and Hall.
- [7] Raghunathan, TE. (2004). "What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data." *Annual Review of Public Health*, 25:99-117.
- [8] White IR, Thompson SG. (2005). "Adjusting for partially missing baseline measurements in randomized trials." *Statistics in Medicine*, 24: 993-1007.
- [9] Arndt AM, Kronmal RA. (2003) "Multiple Imputation of Baseline Data in the Cardiovascular Health Study." *American Journal of Epidemiology*, 157(1): 74-84.
- [10] Schemper M, Smith TL. (1990). "Efficient evaluation of treatment effects in the presence of missing covariate values." *Statistics in Medicine*, 9:777-784.
- [11] Engels JM, Diehr P. (2003). "Imputation of missing longitudinal data: a comparison of methods." *Journal of Clinical Epidemiology*, 56: 968-976.

- [12] Schafer, JL. (2000). *Multiple Imputation for Missing-Data Problems*, presented January 24-25, 2000, Durham, NC.
- [13] Horton NJ, Lipsitz SR. (2001). "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables." *The American Statistician*, 55(3): 244-254.
- [14] Garson GD. (2004). Data Imputation for Missing Values. [<http://www2.chass.ncsu.edu/garson/pa765/missing.htm>] *North Carolina State University*.
- [15] Dempster AP, Laird NM, Rubin DB. (1977). "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society, Series B*. 29: 1-38.
- [16] Patrician PA. (2002). "Focus on Research Methods: Multiple Imputation for Missing Data." *Research in Nursing and Health*, 25: 76-84.
- [17] Yuan YC. (2001). "Multiple imputation for missing data: concepts and new development SAS/STAT 8.2." [<http://www.sas.com/statistics>] *SAS Institute Inc. Cary, NC*. 2001.
- [18] Barnard J, Rubin DB. (1999). "Small-Sample Degrees of Freedom with Multiple Imputation". *Biometrika*, 86:948-955.
- [19] Lanning D, Berry D. (2003). *An Alternative to PROC MI for Large Samples*, presented at SAS Users Group International (SUGI) 28, Seattle, Washington.
- [20] Paulin G, Tsai S, Grance M. (2004). *Model-Based Multiple Imputation*, presented at SAS Users Group International (SUGI) 29, Montreal, Canada.
- [21] Barzi F, Woodward M. (2004). "Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies." *American Journal of Epidemiology*, 160(1):34-45.
- [22] Barnes SA, Lindborg SR, Seaman JW. (2004). "Multiple Imputation Techniques in Small Sample Clinical Trials" presented at Joint Statistical Meetings, Toronto, Canada.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Venita DePuy, Duke Clinical Research Institute, PO Box 17969, Durham, NC 27715.
 You may reach her via e-mail at venita.depuy@duke.edu.

APPENDIX

Multiple Imputation Examples using SAS

Code is intended for SAS v9, adjustments for v8.2 given as needed.

Notes regarding data, particularly to explain max/min statements in PROC MI:

Correct, chemo, and rad_ther are 0/1 variables

Risknum is an ordinal variable, with levels 1, 2, 3

Age is rounded to the nearest year, theoretical range 18-100

Adverse_rxn is the patient-estimated probability, scale 0-100

*****;

*******MULTIPLE IMPUTATION: note can NOT put any nominal variables in this! *****;**

```
PROC MI data=file1
  minimum = 0 1 18 0 0 0 /* variable ranges, IN ORDER. */
  maximum = 1 3 100 100 1 1 /* leaving min & max missing = no restrictions */
  round = 1 /* if need output w/o decimals - I use always */
  seed = 55418 /* round statement: can do diff # for each var */
  nimpute=5 /* ALWAYS use (any #), or non-reproducible */
  out = outmi1 /* # of imputations, 5 is default*/
  simple ; /* output dataset name */
  mcmc /* calculates N, mean, etc. & pairwise corr. */
  ACFplot (wlf) /* Markov chain Monte Carlo method */
  chain=multiple /* specify wlf (worst linear function) plots */
  displayinit /* separate chain used for each imputation */
  initial=em(itprint) /* initial parameter values for MCMC process */
  prior=jeffreys; /* specifies EM algorithm for initial estimates */
  var correct risknum AGE adverse_rxn chemo rad_ther;
```

```
DATA outmi1;
  set outmi1;
  if correct < .001 then correct=0; run; /*accounts for error in MI in SAS v9.1.2;
```

******* To do a simple regression on multiple imputation data *****;**

```
PROC REG data=outmi1
  outest=outreg /*outputs estimates to outreg file */
  covout /* outputs covariance matrix */
  noprint;
  model risknum = age chemo;
  by _Imputation_; /* IMPORTANT! Regresses separately on all 5 datasets */
```

```
PROC MIANALYZE data=outreg;
  var Intercept age chemo; /* intercept but not response variable here */
  run;
```

*******To do a regression using PROC GLM on multiple imputation data *****;**

**To do a T test, say model y = /inverse (no explanatory variables) ;

*v8.2: CLASS statements not in MIANALYZE, can manually create dummy vars;

```
PROC GLM data=outmi1;
  model adverse_rxn = risknum age b_sdrace chemo b_female / INVERSE solution;
  by _Imputation_;
  ods output ParameterEstimates=glmparms InvXPX=glmxpxi;run;
```

```
PROC MIANALYZE parms=glmparms xpxi=glmxpxi ;*edf=324; /*edf= #obs-#vars */
  var Intercept risknum age b_female chemo;
  run;
```

*****To do a logistic regression on multiple imputation data *****;
 *****PROC GENMOD also done this way *****;

```
PROC LOGISTIC descending data=outmi1;
  model correct = age adverse_rxn chemo / COVB;
  by _Imputation_;
  ods output ParameterEstimates=lgparms covb=lgcovb;
```

```
PROC MIANALYZE parms=lgparms(rename=(Variable=Parameter)) covb=lgcovb;
  var Intercept age adverse_rxn chemo;
  run;
```

*****To use PROC MIXED on multiple imputation data *****;

```
*note: different MIANALYZE statments for SAS v8 and v9;
PROC MIXED data=outmi1;
  model risknum = age chemo / solution covb;
  by _Imputation_;
  ods output SolutionF=mixparms CovB=mixcovb;
PROC MIANALYZE parms=mixparms covb(effectvar=rowcol)=mixcovb;          *for SAS v9;
*PROC MIANALYZE parms=mixparms covb=mixcovb edf=326; /*edf= #obs-#vars */ * SAS v8;
  var Intercept age chemo;
  run;
```

*****Computing correlation coefficients *****;

```
PROC CORR data=outmi1 out=outcorr;
  var adverse_rxn age ;
  by _Imputation_;
  run;
```

```
DATA ztrans (type=EST);
  set outcorr (drop=adverse_rxn rename= (age=Z));      *drop 1, rename other;
  if (_type_='N' or _name_='adverse_rxn');
  if (_type_='CORR') then do;          *fisher_s Z transformation;
    _type_='PARMS';                    * z=.5*ln((1+r)/(1-r));
    _name_='';
    z = 0.5*log((1+z)/(1-z));
  end;
  else if (_type_='N') then do;
    _type_='COVB';
    _name_='Z';
    Z = 1. / (Z-3);          *variance = 1/(n-3);
  end;
  run;
```

```
PROC MIANALYZE data=ztrans;
  ods output ParameterEstimates=parms; *version 9;
  *ods output ParmEst=parms;          *version 8;
  var z;
  run;
```

```
DATA corr_ci;
  set parms;
  r = (exp(2*Estimate)-1)/(exp(2*Estimate)+1);
  r_low = (exp(2*LCLMean)-1)/(exp(2*LCLMean)+1);
  r_upp = (exp(2*UCLMean)-1)/(exp(2*UCLMean)+1);
```

```
PROC PRINT data=corr_ci;
  Title 'Estimated Correlation Coefficient w/ 95% CI';
  var r r_low r_upp;
  run;
```

```
*optional print statements to verify correlations;
PROC PRINT data=outcorr (obs=10);
  title 'Correlations (First 2 imputations)';run;
```

```
PROC PRINT data=ztrans;
  title 'EST type data set with Fisher_s Z transformation'; run;
```

```
PROC PRINT data=parms;
  title 'Parameter Estimates with 95% Confidence Limits (terms of Z)';
  var Estimate LCLmean UCLmean;run;
```