Imputation of Data Missing due to Truncation ———
A SAS Macro for Maximum Likelihood Estimates Using Symmetry Property

Zhengyu Wang, Sears Roebuck & Co., Hoffman Estates, IL.
Viswanathan Ramakrishnan, University of Illinois, Chicago, IL.
John M. Davis, Illinois State Psychiatric Institute, Chicago, IL.

ABSTRACT

Clinical trials comparing treatments that follow patients over a period often suffer from patient dropouts, either due to clinical improvement or deterioration. These dropouts are usually due to a particular response to a treatment or due to lack of efficacy rather than occurring at random. Most missing data methods operate under the assumptions that the dropouts are "missing at random," which usually does not hold. There are only a few methods try to tackle the problem of data missing due to truncation. It has been shown that these methods can not provide satisfactory results in the truncated distribution. Under the assumption of a symmetric distribution for the random error in a model at any given time point, a SAS macro program to produce symmetric maximum likelihood estimation for truncated data has been developed.

Based on the symmetry assumption, collected sample data at any given time is partitioned into three segments: (1) data beyond a threshold that are missing due to dropout; (2) the lower tail of the distribution, which corresponds to the mirror image of the unobserved portion; and (3) the data in between. Since the lower and middle part of the data are observed, the mean and the variance-covariance matrix for the truncated distribution are estimated under the symmetry property. Then imputation estimates for truncated values are obtained by applying both the group level information and the individual measurements from the data observed before the time point at which dropout occurs. A simulation study with various conditions was used to examine the sensitivity of the method. It has been verified that this method has several major advantages over existing methods. Analytically, the estimate is unbiased under normal distribution, it is less sensitive to the form of the response function as other existing methods are, and it yields more realistic estimates even if the underlying distribution is skewed.

INTRODUCTION

Consider that in a study some data are missing because the response is higher than a certain threshold. Let the threshold value at which a patient drops out be \( M \), and assume that \( M \) is greater than \( \mu \), the marginal mean at time point \( t \). \( M \) is the threshold value beyond which a patient would be considered as MDT. Under the assumption that the observations at each time point are marginally normally distributed, and imposing the truncation at the threshold \( M \), the distribution is divided into three parts. Based on the assumption that the truncation occurs at the upper tail of the distribution, and on Schneider (1986)'s discussion about the truncation of normal distribution, the three parts can be described as follows.

1. The portion above \( M \): The observations from this part of the distribution are missing. The truncated probability density function (pdf) for this part is:

\[
 f_{\text{high}}(y) = \begin{cases} 
 \frac{1}{\sigma} \phi \left( \frac{y - \mu}{\sigma} \right) & \text{if } y \geq M \\
 1 - \Phi \left( \frac{M - \mu}{\sigma} \right) & \text{if } y < M \\
 0 & \text{if } y < M \end{cases}
\]

where \( \phi() \) and \( \Phi() \) are pdf and cdf from the standard normal distribution, respectively, \( \mu \), and \( \sigma \), are the mean
and variance of the normal distribution at the $t$th time point.

2. The portion between $M'\leq y < M$: This corresponds to a distribution left truncated at $M'$, and right truncated at $M$. Here, $M'$ is a mirror image of $M$ about $\mu$, and it is defined as $M' = \mu - (M - \mu)$. This part of the data is observed, and will be treated as a sample from the truncated normal distribution with pdf

$$f_{L,med}(y) = \begin{cases} \frac{1}{\sigma_t} \left( \frac{y - \mu}{\sigma_t} \right) & \text{if } M' \leq y < M \\ \frac{\phi \left( \frac{M' - \mu}{\sigma_t} \right) - \phi \left( \frac{M' - \mu}{\sigma_t} \right)}{\phi \left( \frac{M - \mu}{\sigma_t} \right) - \phi \left( \frac{M - \mu}{\sigma_t} \right)} & \text{otherwise} \end{cases}$$

3. The portion below $M'$: The data from this part are also observed, and are from the truncated normal distribution with pdf similar to (1.1), which is

$$f_{L,low}(y) = \begin{cases} \frac{1}{\sigma_t} \left( \frac{y - \mu}{\sigma_t} \right) & \text{if } y < M' \\ \phi \left( \frac{M' - \mu}{\sigma_t} \right) - \phi \left( \frac{M - \mu}{\sigma_t} \right) & \text{if } y \geq M' \end{cases}$$

In applications, the actual truncation value $M$ is an unknown nuisance parameter. However, the method proposed in this paper provides imputation estimates that are functionally independent of $M$. The theoretical importance of $M$ is that it helps to define a mirror image, $M'$, in the pdf. Suppose at time $t$, there are $r_t$ observations which are MDT, the same number of observations are less than the truncation value $M'$ under the assumption of symmetric distribution.

There are several widely accepted methods for MDT. Among them are the "last observation carried forward" method (Bombardier and Tugwell, 1985), the regression prediction method which extrapolates the missing observations based on a regression fit (using time as the independent variable) specific to the individual whose data are MDT (Buck, 1960), and mixed-effect regression method (Dempster, Laird, Rubin, 1977). But, based on the described truncated distribution, none of the existing methods cab produce unbiased estimate if MDT is the case.

**A MAXIMUM LIKELIHOOD ESTIMATE USING SYMMETRY PROPERTY**

It can be shown analytically that the mean and variance of the truncated normal distribution are:

$$E(y) = \mu + \sigma_t \left\{ \frac{\phi \left( \frac{L - \mu}{\sigma_t} \right)}{\phi \left( \frac{U - \mu}{\sigma_t} \right) - \phi \left( \frac{L - \mu}{\sigma_t} \right)} - \frac{\phi \left( \frac{U - \mu}{\sigma_t} \right)}{\phi \left( \frac{U - \mu}{\sigma_t} \right) - \phi \left( \frac{L - \mu}{\sigma_t} \right)} \right\}$$

and
Posters

\[
\text{Var}(y) = \sigma^2 \left[ 1 + \frac{\phi \left( \frac{L - \mu_{L}}{\sigma_{t}} \right)}{\phi \left( \frac{U - \mu_{U}}{\sigma_{t}} \right) - \phi \left( \frac{L - \mu_{L}}{\sigma_{t}} \right)} L - \mu_{L} - \frac{\phi \left( \frac{U - \mu_{U}}{\sigma_{t}} \right)}{\phi \left( \frac{U - \mu_{U}}{\sigma_{t}} \right) - \phi \left( \frac{L - \mu_{L}}{\sigma_{t}} \right)} U - \mu_{U} \right]
\]

\[
- \sigma^2 \left[ \frac{\phi \left( \frac{L - \mu_{L}}{\sigma_{t}} \right)}{\phi \left( \frac{U - \mu_{U}}{\sigma_{t}} \right) - \phi \left( \frac{L - \mu_{L}}{\sigma_{t}} \right)} \right]^2
\]

respectively.

Without loss of generality, MDT in a repeated measure design can be considered truncated at the upper tail of the distribution, say at \( M \). That is, the unobserved data are from a truncated normal distribution, with \( L = M \), and \( U = \infty \).

Since the original underlying distribution is symmetric, there exists a part of the data which is symmetric to the missing data, and comes from the observed lower tail of the distribution. These observed data are from a truncated distribution with \( L = \infty \), and \( U = M' \), where \( M' \) is the mirror image of \( M \) about \( \mu_{t} \) (\( M' = \mu_{t} - (M - \mu_{t}) \)). The remaining observed part of data can be considered as the middle part, being left truncated at \( L = M' \) and right truncated at \( U = M' \). From these symmetry settings, it can be shown that the mean of the truncated portion can be simplified as

\[
\mu_{\text{middle}} = \mu_{\text{low}} + (\mu_{\text{middle}} - \mu_{\text{low}}),
\]

the corresponding variances can be simplified as

\[
\sigma^2_{\text{middle}} = \sigma^2_{\text{low}}
\]

and the covariances

\[
\sigma_{\text{low}, \text{middle}} = \sigma_{\text{low}, \text{low}}
\]

where \( t \) is the time when a subject dropped out, and \( t' \) is any time point prior to \( t \) (i.e., \( t' = 1, 2, ..., t-1 \)).

Based upon the truncated normal distribution, the conditional estimation procedure can be described as follows. Suppose, at time \( T \), \( r_{T} \) individuals have missing observations. By definition, these are the \( r_{T} \) largest observations that are to be imputed. Consider the following partition for this individual:

\[
y_{IT} = \begin{bmatrix}
y_{1i}
y_{2i}
\vdots
y_{r_{T}i}
y_{(r_{T}-1)i}
y_{r_{T}i}
\end{bmatrix}
\]

with mean

\[
y_{IT} = \begin{bmatrix}
y_{1i}
y_{2i}
\vdots
y_{r_{T}i}
y_{(r_{T}-1)i}
y_{r_{T}i}
\end{bmatrix}
\]
\[
\mu_\tau = \begin{bmatrix}
\mu_{\text{obs}} \\
\mu_\tau
\end{bmatrix},
\]

and variance-covariance matrix

\[
\Sigma_\tau = \begin{bmatrix}
\Sigma_{11(\text{obs})} & \Sigma_{12} \\
\Sigma_{21} & \sigma_\tau^2
\end{bmatrix},
\]

where \(y_{\text{obs}}\) is the \((T-1)\times1\) vector of the observed part of the data for the \(i\)th individual with \(\mu_{\text{obs}}\) denoting the mean, and \(\Sigma_{11(\text{obs})}\) denoting the variance-covariance matrix. In general, since \(y_\tau\) is multivariate normal distributed, the conditional distribution of \(y_\tau\) given \(y_{\text{obs}}\) is (untruncated) univariate normally distributed (conditional on the observed part of the data). The conditional mean is therefore

\[
E(y_\tau \mid y_{\text{obs}}) = \mu_\tau - \Sigma_{21} \Sigma^{-1}_{11(\text{obs})} (y_{\text{obs}} - \mu_{\text{obs}}),
\]

and the corresponding variance is

\[
\text{Var}(y_\tau \mid y_{\text{obs}}) = \sigma_\tau^2 - \Sigma_{21} \Sigma^{-1}_{11(\text{obs})} \Sigma_{22}.
\]

(T.W. Anderson, 1984). Now, apply the fact that \(y_\tau\) is MDT, the expressions for truncated mean and variance-covariance can be substituted into above two equations, the mean and the variance of the truncated distribution for the data in the upper tail reduce to

\[
E(y_{\tau,\text{MPL}} \mid y_{\text{obs}}) = \mu_{\tau,\text{middle}} + (\mu_{\tau,\text{middle}} - \mu_{\tau,\text{low}}) + \Sigma_{23,\text{low}} \Sigma^{-1}_{11(\text{obs})} (y_{\text{obs}} - \mu_{\text{obs}}),
\]

and

\[
\text{Var}(y_{\tau,\text{MPL}} \mid y_{\text{obs}}) = \sigma_{\tau,\text{low}}^2 - \Sigma_{23,\text{low}} \Sigma^{-1}_{11(\text{obs})} \Sigma_{23,\text{low}}.
\]

This is the imputation equation for symmetric maximum likelihood estimate for the truncated missing data (SMLT). After having substituted the SMLT imputed values, a complete sample with total data points \(N \times T\) is obtained. However, the number of independently measured data values remains the same, which is

\[
\sum_{i=1}^{T} N_i = N - \sum_{i=1}^{T} r_i.
\]

Since the \(\sum r_i\) imputed values are actually linear combinations of the \(\sum N_i\) observed values, the degrees of freedom for error sum square has to be adjusted by subtracting \(\sum r_i\). The suggested imputation value for \(Y_\tau\) and its variance are unbiased.

If the estimate of the untruncated population mean is the only parameter of interest, the imputation step in SMLT can be skipped because: \(\mu_i = \mu_{\text{middle}}\). However, imputing the missing data has many advantages. Substituting the imputed value for the missing observation allows one to treat the data as complete and hence allows the use of methods that are well accepted in the current literature.
THE MACRO IMPLEMENTATION

A SAS macro is developed to implement the SMLT estimation. The macro SMLTEST involves SAS/BASE* and SAS/IML*, and it is independent of platforms. Suppose you have a group of observations in a data file MYDATA, each observation identified by PTID has 4 repeated measured variables Y1, Y2, Y3, and Y4. Some of these Y's are missing due to truncation (the values are too high to keep them in the study), you want to have the imputed complete sample coded as EST1, EST2, EST3, and EST4 in the data file, without knowing the technical details of how to get the imputation values. The actual macro calls would look like this:

```sas
%SMLTEST(mydata, mydata, 0, 1, 4, , ptid, Y1-Y4, EST, 0, 0, 0);
```

Here, the first and second parameters are the names of input and output data files, the next three parameters are the specifications of (1) if you want to detailed intermediate results (1 = yes); (2) which tail of the distribution is truncated (1 = right, 2 = left), and (3) the number of time points you are considering. Parameter 6 is for any IF statement you want to put in the data step, which basically will create a subset of data. Parameter 7 is for specifying the value of missing data, the default value is ".". The next 3 parameters are the names of subject’s identification, the actual names of variables which will be included in the SMLT estimation procedure, and the prefix of estimated (imputed) variables. The last three parameters are to specify that if a low or upper bound is needed (1 = yes) for SMLT imputation, and if so what are the minimum and maximum values they are going to be.

If you have two groups of MDT data, one is because of the clinical deterioration (the values are too high), and the other is because of major improvement (the measures are very low), you can call the macro twice as in the case described above as:

```sas
%SMLTEST(mydata, mydata1, 0, 1, 4, if GROUP = 1, , ptid, Y1-Y4, EST, C, 0, 0);
%SMLTEST(mydata, mydata2, 0, 2, 4, if GROUP = 2, , ptid, Y1-Y4, EST, C, 0, 0);
DATA mydata; SET mydata1 mydata2; run;
```

If you are interested in getting the macro SMLTEST, you can send a request e-mail to the e-mail address u13576@uicvm.cc.uic.edu.

CONCLUSION

The advantage of the SMLT estimate is that it includes information from $Y_{j(i)}$ which is specific to the $i$th individual as well as responses from the other individuals in the group. On the contrary, in methods of the last observation carried, and the regression prediction, the imputation for the $i$th patient, whose value at time $T$ is missing, only depends on that individual’s previous data. Mixed-MRM method imputes the MDT values towards the group mean, which is a biased estimate of population mean in the MDT situation. The SMLT method can be used in a study where missing data occur on several time points, and also the MDT occur at both tails of the distribution.

In addition to the analytical study, two sets of simulation studies were performed in terms of the underlying distribution assumption: one using the normal distribution, which is symmetric, and the other using the beta distribution, whose parameters are set so that the distribution is reasonably skewed. The results from the simulations of various MDT conditions show that in terms of the average absolute bias and the MSE, the SMLT method produces better results than all other methods compared. This is because that the method of symmetric estimate for truncated data has several useful features:

1. SMLT estimates the truncated data values by using the symmetry property, and therefore the estimated sample mean is much closer to the expectation of the untruncated distribution.

2. SMLT imputes the MDT data closer to unobserved truncated values than imputations from regression does as demonstrated by simulation study, and it is also restricted by the distribution at the final time point, and therefore it will not produce extreme and unreasonable results as is the case in regression imputations.
3. SMLT can handle the situation where the truncation occurs at both tails of the distribution, even though it does not guarantee unbiasedness, as is the case when the MDT only occurs on one side of the distribution. But the imputed values produced will be located at the desired range.

The SMLT method is useful if there is a known treatment related reason for drop out. The main advantage in using the SMLT over the other methods is that SMLT utilizes both the individual level information as well as the group level information in the imputation process, it also complies with the assumptions of the analysis, because when a large proportion of drop-outs occur resulted from truncation, the distribution resembles that of a truncated normal rather than a regular normal. The imputations under SMLT are constructed by correctly assuming a truncated normal, while the other methods do not assume MDT in a normal distribution. When MDT is assumed, and when the variation of measurements among time points is high, SMLT should be considered.

But, if the missing data mechanism is unknown, or MDT is mixed with MAR, it might not be easy to identify that an incomplete subjects belongs to MDT, or MAR. The method described in the present work suggests SMLT is only applicable to MDT data, and other methods which assume MAR should be considered to impute non-MDT data. The other drawback of the SMLT method is that when the distribution is not symmetric, as presented in the simulation with the beta distribution, SMLT seems to tend to exaggerate (or understate, depending on the shape of the asymmetry) the effect size at the later points, since all the imputed values are large (or small). However, this criticism does not hold in general, because most well-established statistical methods to analyze repeated measure design data require the normality assumption.

After all, the intention of proposing the SMLT is not to sway the outcome of the study in one way or the other but to efficiently use the MDT information to produce accurate results.

REFERENCES


Schneider, H. (1986), Truncated and Censored Samples from Normal Populations, New York, Marcel Dekker, Inc.

* SAS, SAS/BASE, SAS/IML are the registered trademarks of SAS Institute, Inc., Cary, NC, USA

Zhengyu Wang, Ph.D., Sears, Roebuck & Co., 3333 Beverly Road, A4-217A, Hoffman Estates, IL. 60179 Telephone: (708) 286-8655, E-mail: u13576@uicvm.cc.uic.edu