Multiple Imputation of Missing Data Using Stata

Ofira Schwartz-Soicher

oschwart@princeton.edu

October, 2017

http://dss.princeton.edu/training

Multiple imputation (MI) is a statistical technique for dealing with missing data. In MI the distribution of observed data is used to estimate a set of plausible values for missing data. The missing values are replaced by the estimated plausible values to create a “complete” dataset.

The data file mheart5.dta which is available from Stata Corp. will be used for this tutorial:

```
webuse "mheart5.dta"
```

The “`misstable`” command with the “`gen()`” option generates indicators for missingness. These new variables are added to the data file and start with the prefix `miss_`.

To examine the missing data pattern:

```
misstable sum, gen(miss_)
```

```
                                                               Obs<.
                                                +------------------------------
               |                                | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+--------------------------------+------------------------------
           age |        12                 142  |    142   20.73613    83.78423
           bmi |        28                 126  |    126   17.22643    38.24214
  -----------------------------------------------------------------------------
```

This column represents the number of missing values for each variable. If there is no entry for a variable, it has no missing values.

The number of observed values for each variable is listed in this column.

As an additional check you may tabulate the new indicator variables:

```
tab1 miss_age miss_bmi

-> tabulation of miss_age

   (age>=.) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        142       92.21       92.21
          1 |         12        7.79      100.00
------------+-----------------------------------
      Total |        154      100.00

-> tabulation of miss_bmi

   (bmi>=.) |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        126       81.82       81.82
          1 |         28       18.18      100.00
------------+-----------------------------------
      Total |        154      100.00
```

Indicators for missing age and BMI were added to the data file; a value of 1 on these variables indicates the observation is missing information on the specific variable. A value of 0 indicates the observation in not missing. 12 observations are missing information on age, 28 observations are missing on BMI.

MI is appropriate when data are missing completely at random (MCAR) or missing at random (MAR). It would be difficult to perform a legitimate analysis if data are missing not at random (MNAR).

Logistic regression models could be used to examine whether any of the variables in the data file predict missingness. If they do, the data are MAR rather than MCAR.

```
logit miss_bmi attack smoke age female hsgrad

Iteration 0:   log likelihood = -49.994502
Iteration 1:   log likelihood = -47.73123
Iteration 2:   log likelihood = -47.614822
Iteration 3:   log likelihood = -47.614515
Iteration 4:   log likelihood = -47.614515

Logistic regression                               Number of obs   =        142
                                                  LR chi2(5)      =       4.76
                                                  Prob > chi2     =     0.4459
Log likelihood = -47.614515                       Pseudo R2       =     0.0476

------------------------------------------------------------------------------
    miss_bmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attack |   .0101071   .5775173     0.02   0.986    -1.121806     1.14202
      smokes |   .1965135   .5739319     0.34   0.732    -.9283723    1.321399
         age |  -.0485561   .0244407    -1.99   0.047     -.096459   -.0006532
      female |   .0892789   .6256756     0.14   0.887    -1.137023    1.315581
      hsgrad |   .3940007   .6888223     0.57   0.567    -.9560662    1.744068
       _cons |   .1414761   1.423355     0.10   0.921    -2.648249    2.931201
------------------------------------------------------------
```

Age is statistically significantly associated with missingness of BMI, and the cases missing age are also missing BMI suggesting that the data are MAR rather than MCAR.

```
logit miss_age attack smoke female hsgrad

Iteration 0:   log likelihood = -42.144379
Iteration 1:   log likelihood = -40.780233
Iteration 2:   log likelihood = -40.713422
Iteration 3:   log likelihood = -40.713172
Iteration 4:   log likelihood = -40.713172

Logistic regression                               Number of obs   =        154
                                                  LR chi2(4)      =       2.86
                                                  Prob > chi2     =     0.5811
Log likelihood = -40.713172                       Pseudo R2       =     0.0340

------------------------------------------------------------------------------
    miss_age |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attack |  -1.035628   .7108815    -1.46   0.145     -2.42893    .3576738
      smokes |   .2788896   .6369393     0.44   0.661    -.9694886    1.527268
      female |  -.0059384   .7025713    -0.01   0.993    -1.382953    1.371076
      hsgrad |   .5426292   .8029777     0.68   0.499    -1.031178    2.116437
       _cons |  -2.649692   .7993453    -3.31   0.001     -4.21638   -1.083004
------------------------------------------------------------------------------
```

No other variables other than BMI are statistically significantly associated with missingness of age.

T-test may also be informative in evaluating whether the values of other variables vary between the missing and the non-missing groups.

```
foreach var of varlist attack smoke age female hsgrad {
     ttest `var', by(miss_bmi)
}


    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4392         Pr(|T| > |t|) = 0.8785          Pr(T > t) = 0.5608

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     126    57.14571    1.022929    11.48234    55.12121    59.17021
       1 |      16    50.82253    2.810969    11.24388    44.83109    56.81397
---------+--------------------------------------------------------------------
combined |     142    56.43324    .9727211    11.59131    54.51024    58.35624
---------+--------------------------------------------------------------------
    diff |            6.323186    3.040682                .3115936    12.33478
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.0795
Ho: diff = 0                                     degrees of freedom =      140

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9803         Pr(|T| > |t|) = 0.0394          Pr(T > t) = 0.0197
```

T-test suggests a statistically significant relationship between missigness of BMI and age. T-tests between missingness of BMI and the other variables (i.e., attack, smokes, female and hsgrad) were not statistically significant. Results are not presented for brevity.

A decision regarding the variables to be imputed should be made prior to the imputation. The imputation model should always include all the variables in the analysis model, including the dependent variable of the analytic model as well as any other variables that may provide information about the probability of missigness, or about the true value of the missing data. Theory should guide the decision as to which variables to include.

To deal with skewed variables, the imputation model may include transformed variables (such as log and squared transformations - similar to transformation of variables in other regression models). Non-linear terms which are included in the analytic model must be taken into account when creating the imputation model. It is suggested to treat the non-linear terms as just another variable. That is, create a new variable that will represent the non-linear term prior to the imputation and include it as another variable in the imputation model.

Before proceeding with the imputation, a model which includes all the variables in the imputation model should be estimated for each variable separately. This will ensure that the model is specified correctly and that it converges. The addition of interaction terms may be examined at this stage. If the interaction terms are statistically significant a separate imputation for each group (e.g., male and female) should be considered.

```
logit attack smokes age female hsgrad bmi, or

Iteration 0:   log likelihood = -87.082406
Iteration 1:   log likelihood = -75.829436
Iteration 2:   log likelihood = -75.802318
Iteration 3:   log likelihood = -75.802314

Logistic regression                               Number of obs   =        126
                                                  LR chi2(5)      =      22.56
                                                  Prob > chi2     =     0.0004
Log likelihood = -75.802314                       Pseudo R2       =     0.1295

------------------------------------------------------------------------------
      attack | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |   4.545809   1.844206     3.73   0.000     2.052505    10.06788
         age |   1.030523   .0181677     1.71   0.088     .9955232    1.066754
      female |   1.321851   .6152168     0.60   0.549     .5309038    3.291163
      hsgrad |   1.381645   .6161839     0.72   0.469      .576473    3.311418
         bmi |   1.104937   .0553865     1.99   0.047     1.001543    1.219004
       _cons |   .0050166   .0090925    -2.92   0.003     .0001438    .1750652
------------------------------------------------------------------------------

reg bmi attack age female hsgrad smokes

      Source |       SS       df       MS              Number of obs =     126
-------------+------------------------------           F(  5,   120) =    0.90
       Model |  73.1449691     5  14.6289938           Prob > F      =  0.4853
    Residual |  1956.28756   120  16.3023963           R-squared     =  0.0360
-------------+------------------------------           Adj R-squared = -0.0041
       Total |  2029.43253   125  16.2354602           Root MSE      =  4.0376

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
      attack |      1.545   .7775581     1.99   0.049     .0054888    3.084511
         age |  -.0240674   .0318764    -0.76   0.452    -.0871805    .0390457
      female |   -.109108   .8372703    -0.13   0.897    -1.766845    1.548629
      hsgrad |  -.4092541   .8225038    -0.50   0.620    -2.037754    1.219246
      smokes |  -.2470361   .7820658    -0.32   0.753    -1.795472      1.3014
       _cons |   26.31799   1.961478    13.42   0.000      22.4344    30.20158
------------------------------------------------------------------------------


reg age smokes attack female hsgrad bmi

      Source |       SS       df       MS              Number of obs =     126
-------------+------------------------------           F(  5,   120) =    0.77
       Model |  512.408607     5  102.481721           Prob > F      =  0.5731
    Residual |  15968.1205   120   133.06767           R-squared     =  0.0311
-------------+------------------------------           Adj R-squared = -0.0093
       Total |  16480.5291   125  131.844233           Root MSE      =  11.535

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smokes |   .1810474   2.235231     0.08   0.936    -4.244555     4.60665
      attack |   3.711123   2.232172     1.66   0.099    -.7084223    8.130668
      female |   1.116176   2.390081     0.47   0.641    -3.616019     5.84837
      hsgrad |  -.7017856   2.351445    -0.30   0.766    -5.357484    3.953913
         bmi |  -.1964494   .2601901    -0.76   0.452    -.7116077     .318709
       _cons |   60.53406    6.92685     8.74   0.000     46.81938    74.24874
------------------------------------------------------------------------------
```

For this example a number of interactions were examined, however, none was statistically significant, therefore, will not be included in the analytic model.

When considering the number of datasets to impute, it is often suggested that an imputation of `m=5` datasets is adequate. The use of more than 5 to 10 imputations offers little or no practical benefit (Schafer, 1999).

```
mi set wide
```

Prior to imputation, data should be set to wide using the "`mi set`" command. Indicating how the additional imputations should be stored.

```
mi register regular female attack smokes hsgrad
mi register imputed bmi age
```

Variables in the data set have to be registered using the "`mi register`" command. "`mi register imputed`" specifies the variables to be imputed in the procedure. "`mi register regular`" specifies the variables that should not be imputed (either because they have no missing values or because there is no need).

```
mi impute chained (regress) age bmi = attack smokes hsgrad female, add(5) rseed(9478)
replace
```

"`mi impute chained`" (MICE) is an iterative process. The variable with the fewest missing values is imputed first followed by the variable with the next fewest missing values and so on for the rest of the variables.

Both BMI and age, the variables to be imputed, are continuous variables. Therefore, regress is specified as a method.

These are regular variables which should not be imputed.

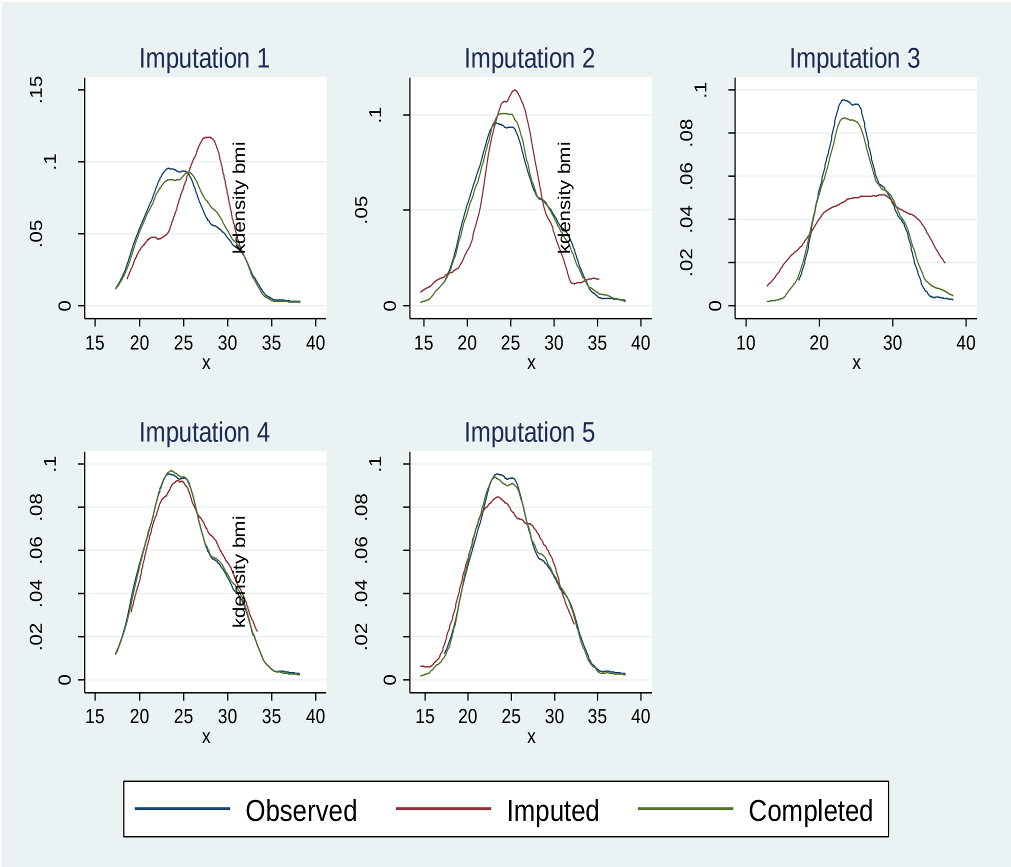
The number of datasets to be imputed (5).

The "`rseed()`" option may be used for results reproducibility.

```
note: missing-value pattern is monotone; no iteration performed

Conditional models (monotone):
               age: regress age attack smokes hsgrad female
               bmi: regress bmi age attack smokes hsgrad female

Performing chained iterations ...

Multivariate imputation                     Imputations =         5
Chained equations                                 added =         5
Imputed: m=1 through m=5                        updated =         0

Initialization: monotone                     Iterations =         0
                                                burn-in =         0

               age: linear regression
               bmi: linear regression

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               age |        142           12        12 |       154
               bmi |        126           28        28 |       154
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Five imputations of variables which have missing values on the observed dataset were added. The new variables are noted with a prefix _x_ where x represent the imputation number (i.e. , 1, 2,…,5).

```
midiagplots bmi, m(1/5) combine

(M = 5 imputations)
(imputed: bmi age)
```



To compare the distribution of imputed variables to that of the observed and the complete (i.e., observed and imputed) data use the "`midiagplots`" command which could be downloaded by typing "`findit midiagplots`" in the command line. The plots below represent the distribution of BMI (this could be done for any of the imputed variables) and suggest a good overlap between observed and completed data.

Once the data have been imputed it is possible to perform the estimation of the analytic model of interest.

In the example below, a logistic regression is used to estimate the probability of a heart attack with the pooled 5 imputed data files:

Using the "or" option will present odds ratios following a logistic regression.

```
mi estimate, or: logit attack bmi age female smokes hsgrad

Multiple-imputation estimates                   Imputations       =          5
Logistic regression                             Number of obs     =        154
                                                Average RVI       =     0.0847
                                                Largest FMI       =     0.2674
DF adjustment:   Large sample                   DF:     min       =      66.30
                                                        avg       =  219859.40
                                                        max       = 1304332.25
Model F test:       Equal FMI                   F(   5, 1620.2)   =       2.99
Within VCE type:          OIM                   Prob > F          =     0.0108

------------------------------------------------------------------------------
      attack | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.092521    .054564     1.77   0.081     .9888445    1.207068
         age |   1.028489    .016718     1.73   0.085     .9961376     1.06189
      female |   .9252642   .3829305    -0.19   0.851     .4110762    2.082616
      smokes |    3.15415   1.121851     3.23   0.001     1.570664    6.334049
      hsgrad |   1.142031   .4547481     0.33   0.739     .5232819    2.492414
       _cons |   .0099802   .0169657    -2.71   0.007     .0003484    .2859078
------------------------------------------------------------------------------
```

5 imputed datasets were used.

The "mi estimate" prefix first runs the estimation command on each of the imputations separately. It then combines the results and displays the combined output.

Suggested Reading

Allison, Paul D. (2001), *Missing Data* (Series: Quantitative Applications in the Social Sciences). *A SAGE University paper*.

Azur, Melissa J., Stuart, Elizabeth A., Frangakis, Constantine & Leaf, Philip J. (2011), Multiple Impuitation by Chained Equations: What is it and how does it work? *International Journal Methods Psychiatric Research*, 20(1), 40-49.

Schafer, Joseph L. (1999), Multiple Imputation: a primer. *Statistical Methods in Medical Research*, 8, 3-15.

van Buuren, Stef. (2012), *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL.

White, Ian R., Royston Patrick & Wood, Angela M. (2011), Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377-399.

Stata Multiple-Imputation Reference manual (release 13):
http://www.stata.com/manuals13/mi.pdf

A Mutiple Imputation in Stata tutorial of UCLA's Institute for Digital Research and Education could be found at: http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm