Handling missing data in Stata: Imputation and likelihood-based approaches

Rose Medeiros

StataCorp LP

2016 Swiss Stata Users Group meeting

Missing Values

- Missing values are ubiquitous in many disciplines
  - Respondents fail to fully complete questionnaires
  - Follow-up points are missing
  - Equiptment malfunctions
- A number of methods of handling missing values have been developed

Traditional Methods

- Complete case analysis—analyze only those cases with complete data on some set of variables
  - Potentially biased unless the complete cases are a random sample of the full sample
- Hot deck—picking a fixed value from another observation with the same covariates
  - Not necessarily deterministic if there were many observations with the same covariate pattern
- Mean imputation—replacing with a mean
- Regression imputation—replacing with a single fitted value
- The last three methods all suffer from too little variation
  - Replace each missing value with a single good estimate

Principled Methods

- Methods that produce
  - Unbiased parameter estimates when assumptions are met
  - Estimates of uncertainty that account for increased variability due to missing values
- This presentation focuses on how to implement two of these methods Stata
  - Multiple Imputation (MI)
  - Full information maximum likelihood (FIML)
- Other principled methods have been developed, for example Bayesian approaches and methods that explicitely model missingness

Missing Data Mechanisms

The classic typology of missing data mechanisms, introduced by Rubin:

- Missing completely at random (MCAR)
  - Missingness on x is unrelated to observed values of other variables and the unobserved values of x
- Missing at random (MAR)
  - Missingness on x uncorrelated with the unobserved value of x, after adjusting for observed variables
- Missing not at random (MNAR)
  - Missingness on x is correlated with the unobserved value of x
- MI and FIML both assume that missing data is either MAR or MCAR

An Example

- The example used throughout this presentation uses data from the National Health and Nutrition Examination Survey II contained in `nhanes2.dta`
- We'll regress diastolic blood pressure (`bpdiast`) on body mass index (`bmi`) and age in years (`age`)
- The starting dataset contains no missing values on the analysis variables
- Missing values were created for `bmi` and `age`
    - The missing values are MAR

Analysis with Complete Data

```
. webuse nhanes2
. regress bpdiast bmi age

      Source |       SS           df       MS      Number of obs   =    10,351
-------------+----------------------------------   F(2, 10348)     =   1224.34
       Model |  330967.862         2  165483.931   Prob > F        =    0.0000
    Residual |   1398651.4    10,348  135.161519   R-squared       =    0.1914
-------------+----------------------------------   Adj R-squared   =    0.1912
       Total |  1729619.26    10,350  167.112972   Root MSE        =    11.626

------------------------------------------------------------------------------
     bpdiast |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9303882    .023599    39.42   0.000     .8841295    .9766469
         age |   .1530495   .0067377    22.72   0.000     .1398423    .1662567
       _cons |   50.67308   .6425594    78.86   0.000     49.41354    51.93262
------------------------------------------------------------------------------
```

Summarizing Missing Values

Switching to the version of the dataset with missing values, we can summarize the missing values

```
. use nh2miss
. misstable summarize
                                                            Obs<.
                                              +------------------------------
               |                              | Unique
      Variable |     Obs=.     Obs>.     Obs<.  | values        Min         Max
  -------------+------------------------------+------------------------------
           age |       976              9,375  |     55         20          74
           bmi |     1,858              8,493  |   >500    12.3856     61.1297
  -----------------------------------------------------------------------------
```

Missing Value Patterns

```
. misstable patterns

   Missing-value patterns
     (1 means complete)

              |   Pattern
  Percent     |  1  2
  ------------+-------------
     76%      |  1  1
              |
     14       |  1  0
      6       |  0  1
      4       |  0  0
  ------------+-------------
    100%      |

  Variables are  (1) age  (2) bmi
```

Estimation Using Complete Case Analysis

By default, `regress` performs complete case analysis

```
. regress bpdiast bmi age

      Source |       SS           df       MS      Number of obs   =     7,915
-------------+----------------------------------   F(2, 7912)      =    689.23
       Model |  143032.35         2  71516.1748   Prob > F        =    0.0000
    Residual |  820969.154     7,912  103.762532   R-squared       =    0.1484
-------------+----------------------------------   Adj R-squared   =    0.1482
       Total |  964001.504     7,914  121.809642   Root MSE        =    10.186

------------------------------------------------------------------------------
     bpdiast |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .7273228   .0255498    28.47   0.000     .6772383    .7774072
         age |   .1215468   .0066455    18.29   0.000     .1085198    .1345738
       _cons |   53.93006   .6638102    81.24   0.000     52.62882     55.2313
------------------------------------------------------------------------------
```

Comparing Complete Data to Listwise Deletion

Coefficients

	Complete	Listwise
bmi	.93	.727
age	.153	.122
intercept	50.7	53.9

Standard errors

	Complete	Listwise
bmi	.023	.025
age	.007	.006
intercept	.643	.663

What is Multiple Imputation?

- Multiple imputation (MI) is a simulation-based approach for analyzing incomplete data
- Multiple imputation:
  - replaces missing values with multiple sets of simulated values to complete the data—*imputation step*
  - applies standard analyses to each completed dataset—*data analysis step*
  - adjusts the obtained parameter estimates for missing-data uncertainty—*pooling step*
- The objective of MI is to analyze missing data in a way that results in in valid statistical inference (Rubin 1996)
- MI does not attempt to produce imputed values that are as close as possible the missing values

Preparing the Data for Imputation

First, we need to tell Stata how to store the imputations. Stata call these `mi` styles.

```
. mi set wide
```

Next we tell Stata what variables we plan to impute

```
. mi register imputed bmi age
```

Optionally, we can also tell Stata what variables we don't plan to impute

```
. mi register regular bpdiast
```

Imputing Missing Values

```
. mi impute mvn bmi age = bpdiast, add(20)

Performing EM optimization:
note: 398 observations omitted from EM estimation because of all imputation
  variables missing observed log likelihood = -47955.552 at iteration 8

Performing MCMC data augmentation ...
Multivariate imputation                       Imputations =       20
Multivariate normal regression                      added =       20
Imputed: m=1 through m=20                         updated =        0

Prior: uniform                                 Iterations =     2000
                                                  burn-in =      100
                                                  between =      100

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               bmi |       8493         1858      1858 |     10351
               age |       9375          976       976 |     10351
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Obtaining MI Estimates

```
. mi estimate: regress bpdiast bmi age

Multiple-imputation estimates                   Imputations       =         20
Linear regression                               Number of obs     =     10,351
                                                Average RVI       =     0.1619
                                                Largest FMI       =     0.2424
                                                Complete DF       =      10348
DF adjustment:   Small sample                   DF:     min       =     322.12
                                                        avg       =     706.73
                                                        max       =     969.86
Model F test:       Equal FMI                   F(   2,   838.8)  =     970.30
Within VCE type:          OLS                   Prob > F          =     0.0000

------------------------------------------------------------------------------
     bpdiast |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .9283816   .0263465    35.24   0.000     .8766788    .9800844
         age |   .1510538   .0076479    19.75   0.000     .1360076    .1660999
       _cons |   50.86274   .7051584    72.13   0.000     49.47863    52.24685
------------------------------------------------------------------------------
```

Comparing MI Estimates

Coefficients

	Complete	Listwise	MI
bmi	.93	.727	.928
age	.153	.122	.151
intercept	50.7	53.9	50.9

Standard errors

	Complete	Listwise	MI
bmi	.023	.025	.026
age	.007	.006	.008
intercept	.643	.663	.705

Adding Categorical Variables

- If the analysis model includes categorical variables, we'll want to include those in the imputation model as well
- To demonstrate we'll add three categorical variables to our analysis model
- The analysis model is now
  `regress bpdiast bmi age i.race i.female i.region`
  - Respondent's race (`race`) takes on 3 values and has missng values
  - Resondent's sex (`female`) is binary and has missing values
  - Region of the U.S. (`region`) takes on 4 values and is complete

Imputing Categorical Variables

- The multivariate normal model implemented in `mi impute mvn` assumes all variables follow a multivariate normal distribution
- However, it turns out to be surprisingly robust to nonnormality (Schafer 1997; Demirtas et al. 2008), even when imputing categorical variables (e.g., Lee and Carlin 2010)
  - To include `race` and `region` in a model using `mi impute mvn` we would need to create $k-1$ dummy variables to use in the imputation model
- An alternative is to use the multivariate imputation by chained equations (MICE) approach to impute the missing values

MICE

- MICE allows us to specify the method used to impute each of the variables in our model
- In Stata, MICE is implemented in `mi impute chained`
- For our example, we will use
  - A linear model (`regress`) to impute `bmi` and `age`
  - A logistic model (`logit`) to impute `female`
  - A multinomial logit model (`mlogit`) to impute `race`
- `mi impute chained` allows the user to specify models for a variety of variable types, including binary, ordinal, nominal, truncated, and count variables

Using mi impute chained

As before, we prepare the data for imputation

```
. mi set wide
. mi register imputed bmi age race female
. mi register regular bpdiast region
```

Then we can run the imputation model

```
. mi impute chained (regress) bmi age (logit) female ///
    (mlogit) race = bpdiast i.region, add(20)

Conditional models:
               age: regress age bmi i.female i.race bpdiast i.region
               bmi: regress bmi age i.female i.race bpdiast i.region
            female: logit female age bmi i.race bpdiast i.region
              race: mlogit race age bmi i.female bpdiast i.region

Performing chained iterations ...

Multivariate imputation                     Imputations =         20
Chained equations                                 added =         20
Imputed: m=1 through m=20                       updated =          0

Initialization: monotone                     Iterations =        200
                                                burn-in =         10
```

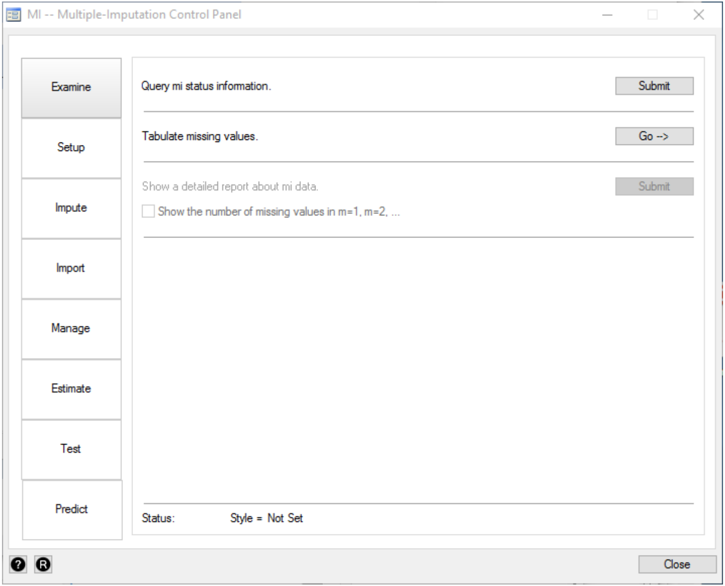
mi impute chained (continued)

```
             bmi: linear regression
             age: linear regression
          female: logistic regression
            race: multinomial logistic regression

------------------------------------------------------------------
                   |               Observations per m
                   |----------------------------------------------
          Variable |   Complete   Incomplete   Imputed |     Total
-------------------+-----------------------------------+----------
               bmi |       8493         1858      1858 |     10351
               age |       9375          976       976 |     10351
            female |       8220         2131      2131 |     10351
              race |       7297         3054      3054 |     10351
------------------------------------------------------------------
 of the number of filled-in observations.)
(complete + incomplete = total; imputed is the minimum across m
```

Additional Features `mi` Suite

- We haven't seen Stata's tools for
  - Data management with `mi` data
  - Use of `mi impute` to impute univariate and monotone missing values
  - Investigating convergence for both `mi impute` and `mi impute chained`
  - Hypothesis tests and predictions after `mi estimate`
  - The use of `mi estimate` with special data types, for example survey or time-series data (see `help mi xxxset`)
- The dialog box for `mi` which guides you through the MI process
  - It can be reached from the menus **Statistics** > **Multiple imputation** or by typing `db mi`


MI -- Multiple-Imputation Control Panel
Examine
Setup
Impute
Import
Manage
Estimate
Test
Predict
Query mi status information.
Submit
Tabulate missing values.
Go -->
Show a detailed report about mi data.
Submit
Show the number of missing values in m=1, m=2, ...
Status:
Style = Not Set
Close

More on the Imputation Step

In practice the imputation process involves a lot of decision making

- Scope of the imputation—Whether to impute for a specific analysis, set of related analyses, or for all analyses on a given dataset
- The type of imputation model to use
- What variables to include in the imputation model
- The number of imputations to create

Selecting an Imputation Model

For the most common missing data pattern the options are

- The multivariate normal model—implemented in (`mi estimate mvn`)
  - Assumes multivariate normality or all variables
  - If the model includes non-normal or categorical variables, you'll have to decide how to include those
- Multivariate imputation by chained equations—implemented in (`mi impute chained`)
  - Offers flexibility in how each variable is modeled

Selecting Variables

The imputation model must maintain the existing characteristics of the data, in order to do so it should include

- All variables in the analysis model
- Any interactions that will be tested in the analysis model
- Transformations of variables
- Auxilary variables–variables that do not appear in the analysis model, but
  - Predict missingness, and
  - Are correlated with the variables with missing values

Full Information Maximum Likelihood Estimation

- Full information maximum likelihood (FIML) estimation adjusts the likelihood function so that each case contributes information on the variables that are observed
- Does not create or impute any data, it just analyzes everything that is there
- FIML is implemented as part of Stata's `sem` command which fits linear structural equation models
- FIML assumes
  - Multivariate normality
  - Missing values are MAR or MCAR

Using `sem`

- The `sem` command uses a form of model specification that is different from other commands
  - Direct paths within variables in a model are specified within sets of parentheses
  - Arrows are used to denote the direction of relationships
- The following all regress `bpdiast` on `bmi` and `age`

```
. regress bpdiast bmi age
. sem (bpdiast <- bmi age)
. sem (bmi age -> bpdiast)
```

- By default `sem` performs maximum likelihood estimation on the complete cases
- To request estimation using FIML use the option `method(mlmv)`

```
. use nh2miss, clear
. sem (bpdiast <- bmi age), method(mlmv)
```

(output omitted)

```
Structural equation model                       Number of obs     =     10,351
Estimation method  = mlmv
Log likelihood     = -105553.76

------------------------------------------------------------------------------
              |                 OIM
              |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Structural    |
  bpdiast <-  |
          bmi |   .9229957   .0276157    33.42   0.000       .86887    .9771214
          age |    .152064   .0076274    19.94   0.000     .1371146    .1670133
        _cons |   50.95577   .7217014    70.61   0.000     49.54126    52.37028
--------------+---------------------------------------------------------------
     mean(bmi)|   25.46282   .0518402   491.18   0.000     25.36121    25.56442
     mean(age)|   47.72442   .1827953   261.08   0.000     47.36615    48.08269
--------------+---------------------------------------------------------------
var(e.bpdiast)|   135.9395   1.985341                      132.1035     139.887
      var(bmi)|   22.67168   .3509293                       21.9942    23.37003
      var(age)|   307.4869   4.563105                      298.6722    316.5618
--------------+---------------------------------------------------------------
  cov(bmi,age)|   16.85967    .965718    17.46   0.000      14.9669    18.75244
------------------------------------------------------------------------------
LR test of model vs. saturated: chi2(0)   =      0.00, Prob > chi2 =      .
```

Comparing FIML Estimates

Coefficients

	Complete	Listwise	MI	FIML
bmi	.93	.727	.928	.923
age	.153	.122	.151	.152
intercept	50.7	53.9	50.9	51

Standard errors

	Complete	Listwise	MI	FIML
bmi	.023	.025	.026	.028
age	.007	.006	.008	.008
intercept	.643	.663	.705	.722

Comparison

Multiple imputation

- If the chained equation approach is used, there is not assumption of multivariate normality
- MI generally makes it easier to include auxilary variables
- Allows for a wide variety of analysis models
- Care is required when constructing the imputation model

Full information maximum likelihood

- Repeated runs of the same model produce the same results
- Easier for others to reproduce, since fewer decisions need to be made and documented

Conclusion

- Stata provides multiple options for analyzing data that contain missing values
- MI and FIML both assume missing values are MAR or MCAR
    - Other solutions are necessary for MNAR data

References

Demirtas, H., S.A. Freels, RM Yucel. 2008. Journal of Statistical Computation and Simulation 78(1): 69-84.

Lee, K. J., and J. B. Carlin. “Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation.” American journal of epidemiology 171.5 (2010): 624-632.

Little, R. J. A., & D. B. Rubin. 2002. Statistical analysis with missing data. Hoboken, N.J: Wiley.

Rubin, D. B. 1996. "Multiple imputation after 18+ years." Journal of the American statistical Association 91(434): 473-489.
Schafer, J. L. 1997. Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC.