Analyzing Missing Data

Hsueh-Sheng Wu

CFDR Workshop Series

Summer 2011

BGSU

Center for Family and Demographic Research

Outline

- Importance of analyzing missing data
- Causes of missing data
- Three mechanisms underlying missing data
- Strategies of handling missing data
- What is multiple imputation?
- How to obtain parameters from imputed data?
- The -ice- command in Stata
- The -mim- command in Stata
- Further topics
- Conclusions

Importance of Analyzing Missing Data

- Missing data are common in survey data
  - Unit nonresponse: Respondents did not complete any part of the survey.
  - Item nonresponse: Respondents did not complete one or more questions in the survey.

- Missing data post a dilemma in analyzing data: On the one hand, we do not want to throw out the information contained in the data; on the other hand, most statistical methods are not designed to analyze data with missing values.

- The failure to adequately analyzing missing data results in:
  - upward or downward biases in parameter estimates
  - under- or over-estimated standard errors of the parameters
  - inaccurate findings

Causes of Missing Data

- Question design
  - Skip patterns
  - Sensitive questions
- Data entry or coding errors
- Characteristics of respondents
  - Command factors among respondents missing many items
- Respondents do not want to answer questions

Three Mechanisms Underlying Missing Data

- Missing completely at random (MCAR): No other variables in the data sets can predict whether the values in a variable (e.g., Y) will be missing. Also, the variable, Y, has missing value not because of the unobserved value of Y itself.

- Missing at random (MAR): Other variables in the data sets can predict whether the values in a variable (e.g., Y) will be missing.

- Missing not at random (MNAR): If the value of the variable, Y, determines whether the value of Y will be missing

Strategies of Handling Missing Data

- Delete cases
  - Pairwise deletion
  - Listwise deletion

- Substitution
- Hot deck imputation
- Mean substitution
- Regression substitution
- Multiple imputation

What Is Multiple Imputation?

- A procedure that replaces missing values with multiple sets of plausible values.

- Three steps:

  (1) Create multiple imputed data

  (2) Conduct a analysis on each imputed data set

  (3) Combine the results of analyzing each imputed data set to obtain an averaged estimate of the parameter

Obtain Estimates from Imputed Data

- Mean of the estimate obtained from m imputed data sets

$$\bar{Q} = \frac{1}{m}\sum_{j=1}^{m} \hat{Q}_j$$

- Standard error of the estimate obtained from m imputed data sets
  - – within-imputation variance

$$\bar{U} = \frac{1}{m}\sum_{j=1}^{m} U_j$$

Obtain Estimates from Imputed Data (Con.)

- Between-imputation variance

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

- Total variance

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

- Standard error of the estimate

$$\sqrt{T}$$

-ice- and -mim- Commands in Stata

- Example: We would like to know whether feeling happy is associated with sex, race, being smart, weight, and height among adolescents.

- The data set has six variables:

Happy: a binary variable, with1 indicating being happy, and 0 indicating being unhappy

Sex: a dichotomous variable, 1 indicating male, and 2 indicating female

Race: a categorical variable, with 1 indicating White, 2 indicating Black, 3 indicating Hispanics, and 4 indicating others.

Smart: an ordinal variable, with a higher score indicating being smarter

Weight: a continuous variable

Height : a continuous variable

The -ice- Command

```
ice happy sex smart weight height race black hisp other, /*
*/ saving(c:\temp\mim_impute.dta, replace) m(3) genmiss(m_) /*
*/ eq(happy: sex smart weight height black hisp other, /*
*/       smart: sex happy, /*
*/       weight: sex height,  /*
*/       height: sex weight,  /*
*/       race: sex height weight) /*
*/ cmd(smart:  ologit, race: mlogit) /*
*/ passive (black: (race==2) \ hisp: (race==3) \ other: (race==4) )  /*
*/ substitute(race: black hisp other) /*
*/ match(weight height) /*
*/ seed(123)
```

The -ice- Command (Con.)

ice happy sex smart weight height race black hisp other, /*

This -ice happy sex smart weight height race black hisp other- statement specifies a list of variables that the –ice- command uses for imputation. Notice that if you are using categorical variables (more than 2 categories), you put both the variable with all the values, in the case race, and each dummy variable, black (race=2), hisp (race=3), and other (race=4), whereas white (race=1) is the reference group, assuming that the variable "race" has missing values on it.

The -ice- Command (Con.)

saving(c:\temp\mim_impute.dta, replace) m(3) genmiss(m_)

- This -saving- statement creates a new data set that includes the original data and imputed data and saves this data set with the name of “mim_impute.dta” in a specific directory (i.e., c:\temp\). The -replace- option allows the Stata to overwrite this data file if you re-run this command.
- The -m(3)- statement indicates that three data sets will be imputed. You can create 100 imputed data sets if you use the –m(100).
- The -genmiss(m_)- option generates indicators for all variables included in the -ice- command. For variables without missing values, their values all equal 0. For variable with missing values, the value 1 indicates that missing values have been imputed and 0 otherwise.
- It should be noted that the -ice- command will automatically generate two indicator variables, _mi and _mj. The _mi variable represents the id variable for each observation within the original or the imputed data sets. The _mj variable has values of 0, 1, 2, and 3 because 3 replicates will be created by the m(3) option.

The -ice- Command (Con.)

```
*/ eq(happy: sex smart weight height black hisp other, /*
*/       smart: sex happy, /*
*/       weight: sex height,  /*
*/       height: sex weight,  /*
*/       race: sex height weight) /*
```

- This -eq - statement tells Stata what variables are used for imputing the missing value for particular variables. In this example, Stata uses seven variables (i.e., sex, smart, weight, height, black, hisp, and other) for imputing happy; sex and happy for imputing smart; sex and height for imputing weight; sex and weight for imputing height; and sex, height, and weights for imputing race.
- This -eq- statement allows researchers flexibility in deciding what variables should be used to impute missing values for particular variables, based on the theories or empirical findings. If this -eq- statement is not specified, Stata will use all the other variables named in the -ice- command to impute missing data for each variable.

The -ice- Command (Con.)

```
*/ cmd(smart:  ologit, race: mlogit) /*
```

- This -cmd- statement tells Stata to use ordered logistic regression to impute the values for smart and multinomial logistic regression to impute values for race. If this -cmd- statement is not specified, Stata uses the measurement levels of the imputed variables to decide what regression models should be used. If the imputed variable is a binary variable, logistic regression model will be used. If the imputed variable is a continuous variable, ordinary least square regression will be used. It should be noted that if imputed variables are nominal or ordinal variables, Stata treats them as continuous variables and subsequently OLS regression is used for imputation. Thus, it is important to specify multinomial logistic regression and ordered logistic regression for imputing nominal variables and ordinal variables, respectively.

The -ice- Command (Con.)

```
*/passive (black: (race==2) \ hisp: (race==3) \ other: (race==4) )  /*
*/substitute(race: black hisp other) /*
```

- The -passive- and -substitute- statements deal with the complications of using categorical variable with more than two categories in multiple imputation. When categorical variables are imputed by other variables, there are not complications. However, complications occur when the categorical variables are used to impute other variables. This is because the response categories of these variables do not follow the interval measurement. Thus, these categorical variables, when being used to impute other variables, are better represented by a series of dummy variables.
- The -passive (black: (race==2) \ hisp: (race==3) \ other: (race==4)- statement asks Stata to imputes values (i.e, race) for subjects who did not provide valid answers for this variable. If the imputed value is 2, 3, or 4, this value will be automatically transferred to the three dummy variables representing the different categories of race (i.e., black. hisp, or other),
- the -substitute(race: black hisp other)- statement tells Stata to use black, hisp, and other instead of race to impute other variables with missing values.

The -ice- Command (Con.)

/ match(weight height) /

- This –match- statement allows the imputed values for variables (i.e., weight and height) to stay within the observe range of that variable.

*/ seed(123)

- This –seed- statement tells Stat to start the imputation every time using this seed number (i.e., 123). You can specify any seed number you want. Without specifying a seed number, Stata randomly selects a seed number when imputing missing values. Subsequently, you are likely to get different imputed data each time you execute the –ice- command.

The -mim- Command

Analyzing the imputed data

- mim: logit happy sex smart weight height black hisp other

Analyzing the imputed data with the svy option

- mim: svy: logit happy sex smart weight height black hisp other

Analyzing the imputed data with the -svy- and -subpop- option

- mim: svy, subpop(marker): logit happy sex smart weight height black hisp other

The result

- The original data

```
sum  ID- personwgt

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          ID |         0
       happy |       194    .3505155    .4783659          0          1
         sex |       200       1.435    .4970011          1          2
        race |       190    1.678947    1.006371          1          4
       smart |       190    3.863158    1.104206          1          6
-------------+--------------------------------------------------------
      height |       191    66.73822    4.246925         58         79
      weight |       191    145.0733    35.58866         77        275
```

The result (con.)

- The constructed data for –ice- command

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          ID |         0
       happy |       194    .3505155    .4783659          0          1
         sex |       200       1.435    .4970011          1          2
        race |       190    1.678947    1.006371          1          4
       smart |       190    3.863158    1.104206          1          6
-------------+--------------------------------------------------------
      height |       191    66.73822    4.246925         58         79
      weight |       191    145.0733    35.58866         77         27
       black |       200        .195    .3971949          0          1
        hisp |       200        .075    .2640523          0          1
       other |       200          .1    .3007528          0          1
```

The result (con.)

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          ID |         0
       happy |       794    .3488665    .4769121          0          1
         sex |       800       1.435    .4960671          1          2
        race |       790     1.66962    .9973164          1          4
       smart |       790    3.860759    1.101148          1          6
-------------+--------------------------------------------------------
      height |       791    66.82807     4.30787         58         79
      weight |       791    145.1631     35.4497         77        275
       black |       800       .2025    .4021142          0          1
        hisp |       800       .0775    .2675504          0          1
       other |       800      .10125    .3018482          0          1
-------------+--------------------------------------------------------
         _mi |       800       100.5    57.77042          1        200
         _mj |       800         1.5    1.118733          0          3
     m_happy |       600         .03    .1707296          0          1
       m_sex |       600           0           0          0          0
     m_smart |       600         .05    .2181268          0          1
-------------+--------------------------------------------------------
    m_weight |       600        .045    .2074771          0          1
    m_height |       600        .045    .2074771          0          1
      m_race |       600         .05    .2181268          0          1
     m_black |       600           0           0          0          0
      m_hisp |       600           0           0          0          0
-------------+--------------------------------------------------------
     m_other |       600           0           0          0          0
```

The result

- The logistic regression with original data

- . logit happy sex black hisp other smart height weight

```
Logistic regression                               Number of obs   =        168
                                                  LR chi2(7)      =      15.37
                                                  Prob > chi2     =     0.0315
Log likelihood = -104.42668                       Pseudo R2       =     0.0686

------------------------------------------------------------------------------
       happy |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.6196679   .3934908    -1.57   0.115    -1.390896    .1515599
       black |   -.914049    .467733    -1.95   0.051    -1.830789    .0026909
        hisp |  -.3350371   .6101414    -0.55   0.583    -1.530892    .8608181
       other |   .6513283    .550888     1.18   0.237    -.4283923    1.731049
       smart |   .2784297   .1512894     1.84   0.066     -.018092    .5749514
      height |  -.1068612   .0581648    -1.84   0.066    -.2208622    .0071397
      weight |   .0007083   .0064936     0.11   0.913    -.0120188    .0134354
       _cons |   6.458745    3.71806     1.74   0.082    -.8285187    13.74601
```

The result

```
•   Multiple-imputation estimates (logit)                  Imputations =          3
•   Logistic regression                                    Minimum obs =        200
•                                                          Minimum dof =       56.7

•   ------------------------------------------------------------------------------
•          happy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Int.]     FMI
•   -------------+----------------------------------------------------------------
•            sex |  -.643209    .364004    -1.77    0.078    -1.35751   .071095   0.002
•          black |  -.569596    .476744    -1.19    0.237    -1.52436   .385169   0.207
•           hisp |  -.077812    .595535    -0.13    0.896    -1.24781   1.09219   0.044
•          other |   .634908     .50618     1.25    0.210    -.358988    1.6288   0.031
•          smart |   .293301    .150675     1.95    0.053     -.00385   .590453   0.097
•         height |  -.128749    .058412    -2.20    0.031    -.245451  -.012047   0.193
•         weight |   .000413    .006454     0.06    0.949    -.012482   .013308   0.194
•          _cons |   7.70154    3.59759     2.14    0.034     .592315   14.8108   0.116
•   ------------------------------------------------------------------------------
```

Further Topics

- The use of the –mi- command in Stata
  - The –mi imputation – command
  - The –mi estimate- command

- Switch between the mi and ice data format
  - The -mi import ice – command allows Stata to read in the data created by the -ice- command and then analyze them using the -mi estimate- command
  - The -mi export ice- command exports the mi data set to the format that ice/mim can recognize.

- Full Information Maximum Likelihood (FIML) methods in Mplus, LISREL and EQS
- Imputing data that are missing not at random
- Imputing data collected with complex survey design

Conclusions

- Missing data could create biased results.

- The seriousness of missing data problem depends on how much data are missing and how they are missing.

- To overcome the missing data problem, multiple imputation replace the missing data with multiple data sets and then obtain the averaged parameters from the data sets.

- The methods of handing missing data are still developing. Before analyzing missing data, you may want to search for what the newest methods or techniques for analyzing missing data are.

- For further question, feel free to contact me at wuh@bgsu.edu or stop by my office (5D Williams Hall).