Missing data and Multiple Imputation (II)

Henrik Støvring & Morten Frydenberg

stovring@ph.au.dk – morten@ph.au.dk

December 14, 2016 – Aarhus University

Overview

- A black box imputation
- How Stata organizes imputations
- Looking into the regression equations used in MICE
- Passive variables
- Data formats for imputed datasets
- Checking assumptions for imputation models
- Separate imputations for subgroups
- Testing after mi estimate

- NOTE: Today we will work with *Wooddata2.dta*

What we learned Monday

?

Where we ended Monday: Exercise 6

We want to replicate the analysis reported by Jakobsen (2008) in Table 4, but using multiple imputation. Do this in pairs by completing the following steps:

1. Identify all relevant variables. Use –mi misspattern– to investigate the missingness pattern and amount
2. Choose a relevant regression model for each variable to be imputed (regress, logit, mlogit, etc...)
3. Declare the data to be of mi-type and register the variables to be imputed
4. Use –mi impute chained– to make a ”black box” imputation model (see slide 23) with 100 imputed datasets
5. Use –mi estimate– to obtain the final estimates of the analysis
6. Compare your estimates with those of Jakobsen (2008) and write a short conclusion

Modified Exercise 6 (read: Exercise 1)

We want to run a very first and simple MI model using –mi impute chained– and investigate its effect

1. Open the dataset *Wooddata1.dta*
2. Replace all types of missing values with ‘.’:
   ```
   foreach var of varlist * {
       replace `var’ = . if missing(`var’)
   }
   ```
3. Declare the data to be of mi-type flong and register wooddustgrp and packryg to be imputed
4. Use –mi impute chained– to make a ”black box” imputation model (see slide 23) with 5 imputed datasets:
   ```
   mi impute chained ///
       (mlogit) wooddustgrp packryg = fev1aendaar, add(5)
   ```
5. Use –mi estimate– to obtain estimates of the analysis:
   ```
   mi estimate: regress fev1aendaar i.wooddustgrp i.packryg
   ```

Understanding how Stata organizes imputations

- Stata has created three new variables: _mi_id _mi_miss _mi_m
- Try to understand what the variables record by looking at the two subjects with lbnr 3 and 5:
  ```
  list lbnr _mi_id _mi_miss _mi_m packryg if lbnr == 3
  list lbnr _mi_id _mi_miss _mi_m packryg if lbnr == 5
  ```
- _mi_m indicates which imputed dataset a row belongs to
- Run a regression for each imputed dataset with change in lung function as outcome and packryg and wood dust exposure as categorical explanatory variables
  (Hint: use `regress ... if _mi_m == 1`, etc)
- Now run the `mi estimate:` from the previous slide, but now with the option noisily: `mi estimate, noi: regress ...`
- Compare with what you got, when you ran a regression for each imputed dataset

Understanding the imputation algorithm

- Add the options `noisily showiter(5)` to the `mi impute chained` statement
- What are the regression equations used in the imputation?
- Remember that exposure status varied substantially with sex – we will now add sex as predictor, but only to the equation for wood dust:

```
mi impute chained ///
    (mlogit, include(i.sex)) wooddustgrp ///
    (mlogit) packryg = fev1aendaar, add(5)
```

- Try to work out why the following command yields the same equations:

```
mi impute chained ///
    (mlogit) wooddustgrp ///
    (mlogit, omit(i.sex)) packryg ///
          = fev1aendaar i.sex, add(5)
```

Exercise 6 from Monday

We want to replicate the analysis reported by Jakobsen (2008) in Table 4, but using multiple imputation. Do this in pairs by completing the following steps:

1. Identify all relevant variables. Use –mi misspattern– to investigate the missingness pattern and amount
2. Choose a relevant regression model for each variable to be imputed (regress, logit, mlogit, etc...)
3. Declare the data to be of mi-type and register the variables to be imputed
4. Use –mi impute chained– to make a dedicated imputation model with 100 imputed datasets – start from a simple model and add variables
5. Use –mi estimate– to obtain the final estimates of the analysis
6. Compare your estimates with those of Jakobsen (2008) and write a short conclusion

Imputing a passive variable

- Definition:

  A variable which depends on other variables, whose values are imputed
- Consider for example BMI = Weight (kg) / Height^2 (m)
- Assume that you want to include BMI in the final regression of change in lung function
- We impute missing values of weight (height is completely observed) – the missing values of BMI should be updated accordingly
- Recipe: Create imputed datasets including filled-in values for weight
- Then:

```
    mi passive: gen BMI = vaegt / ((hojde/ 100)^2)
```

- Now BMI can be used in the final analysis

Exercise 2: Impute a passive variable

- Extend the imputation model from slide to include weight
- Create the passive variable BMI:

```
    mi passive: gen BMI = vaegt / ((hojde/ 100)^2)
```

- Estimate the same model as on slide 8, but now include BMI as explanatory variable

Exercise 3: Working with different formats

- The aim here is that you make a drawing of the mi data format you have been assigned to
- Table 1: wide
- Table 2: mlong
- Table 3: flongsep

- First create an mi-dataset with 5 imputed datasets using mi set, mi register, mi impute chained
  (Just use a simple imputation model: wooddustgrp, packryg, fev1aendaar)
- Examine the dataset (Hint: look at subjects 3 and 5 as before) and draw the structure of the dataset on a piece of paper

Exercise 3: Working with different formats (continued)

- Now find two other students with two different data types and try figure out how you would visually move from one drawing to the other – first you may need to explain your data structure to the two others.
- Use the command `mi convert` to transform the data from one format to the other.
- Verify how the information of persons 3 and 5 have moved around

- Finally, consider which dataset type is smallest? Which is easiest to modify by other commands? Which is conceptually simpler?

Exercise 4: What assumptions should be checked?

- Make a list of all regression equations for one version of the solution used
- For every linear regression:
  - Check residuals: normality, linearity vs predicted, linearity vs continuous covariates
- For every logistic regression:
  - Check linearity of log-odds vs. continuous covariates (for example: Categorize the continuous covariate and estimate association – look for constant increase in log-odds)
- For every multiple logistic regression:
  - As for logistic regression for every outcome pairs of categories.
- In parallel: Assign groups to examine assumptions for one imputation equation
- Write a short summary regarding assumptions (Google Docs – see homepage)

Exercise 5: Separate imputations for men and women

- Repeat “Exercise 6 from Monday” on slide 8, but now with separate imputations for men and women, i.e. you add the option –by(sex)– to the –mi impute– statement and you remove the variable sex from any regression equations it may appear in
- Compare your results with those you obtained before

Testing several parameters jointly

- Suppose we want to test for an overall effect of wood dust exposure (4 categories), i.e. report a single p-value
- Likelihood ratio tests are not available after –mi estimate–
- We can instead obtain tests with –mi test– :

```
mi test 1.wooddustgrp 2.wooddustgrp 3.wooddustgrp
```

Other useful options

- mi estimate:
  - errorok: If estimation fails in an imputed dataset, drop it and proceed with the remaining datasets
  - esampvaryok: It is OK that the number of observations included in the estimation vary between imputed datasets
  - cmdok: It is OK to use a “non-authorized regression-like” command in the MI-analysis

Should the outcome always be included in the imputation model?

- (Mentimeter exercise)

Agenda for a Multiple Imputation based analysis

1. Which variables should be imputed – are there passive variables?
2. What is the structure and amount of missing data?
3. Can we identify variables that help explain why other variables have missing values?
4. Formulate a regression model for each variable to be imputed – and do model diagnostics for each
5. Run the imputation model – if it fails, understand why, modify your model and retry.
6. Examine imputed values
7. Estimate the final model
8. Make sensitivity analyses, where you modify the imputation model so as to detect how much your results depends on the assumptions underlying the imputations

Thanks for your attention – questions welcome!

(Aarhus University, March 2016 – H Støvring)

AARHUS
UNIVERSITY
DEPARTMENT OF PUBLIC HEALTH