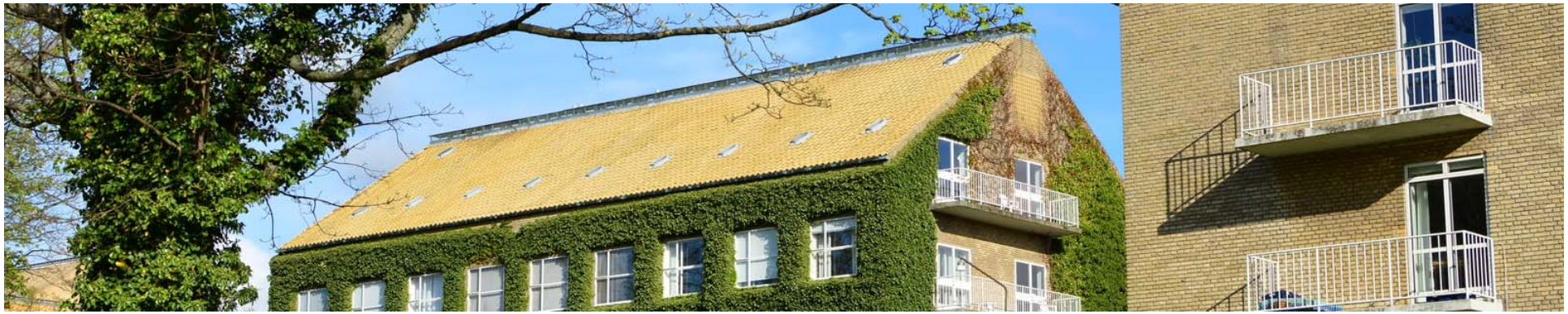


Missing data and Multiple Imputation (II)

Henrik Støvring & Morten Frydenberg

stovring@ph.au.dk – morten@ph.au.dk

December 13, 2017 – 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

?

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 *Wooddata2.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 (-mi set- and -mi register-)
4. Use –mi impute chained– to make a ”black box” imputation model with 10 imputed datasets:

```
mi impute chained ///
    (mlogit) wooddustgrp packryg = fev1aendaar, add(10)
```

5. Use –mi estimate– to obtain estimates of the analysis:

```
mi estimate: regress fev1aendaar i.wooddustgrp i.packryg
```

Towards a first MI model based on MICE (MI Chained Equations)

- Specify the model you want to estimate from a subject matter point of view (Stage specific incidence rates according to period and age category, for example) – the analysis model
- Outline of MICE
  - Make a list of variables to be imputed (variables with missing values which are in the analysis model)
  - Add variables which are predictors of any of these
  - If any added variables have missings they should also be imputed
  - Specify a model for each variable to be imputed – the imputation model, i.e.:
    - Distribution of outcome
    - Which covariates to include

Common regression models used in MICE

- Ordinary linear regression (regress)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

where ϵ follows a Normal distribution with a mean of zero and σ as its SD

- If a person has a missing y but observed x's, then we should impute the missing y with randomly drawn values from a Normal distribution with mean

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots$$

where the x values are those of the specific person. And we use $\hat{\sigma}$ as SD

Common regression models used in MICE

- Logistic regression (logit)

$$\log \text{odds}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

- If a person has a missing y, but observed x's, then we should impute the missing y as 1 with probability

$$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots\right)}$$

  and zero otherwise

Common regression models used in MICE

- Multinomial logistic regression (mlogit)

$$\log \text{odds}(y = 1) = \beta_0^1 + \beta_1^1 x_1 + \beta_2^1 x_2 + \cdots$$
$$\log \text{odds}(y = 2) = \beta_0^2 + \beta_1^2 x_1 + \beta_2^2 x_2 + \cdots$$
$$\ldots$$

where $y = 0$ is reference outcome value

- If a person has a missing y, but observed x's, then we should impute the missing y as 1 with probability

$$\frac{\exp\left(\hat{\beta}_0^1 + \hat{\beta}_1^1 x_1 + \hat{\beta}_2^1 x_2 + \cdots\right)}{1 + \exp\left(\hat{\beta}_0^1 + \hat{\beta}_1^1 x_1 + \hat{\beta}_2^1 x_2 + \cdots\right) + \exp\left(\hat{\beta}_0^2 + \hat{\beta}_1^{22} x_1 + \hat{\beta}_2^2 x_2 + \cdots\right) + \cdots}$$

and so on

Common regression models used in MICE

- Ordered logistic regression (ologit)

- Rather complex formulas, but intuitively:
  If you have ordered categories (packryg, for example) then it is a relevant model to consider
- Main assumption: The effect of a covariate expressed as an Odds Ratio should be the same when considering for example outcome category 2 vs 1 as for outcome category 3 vs 2, and so on
- Requires estimation of fewer parameters than -mlogit-.

The MICE algorithm

- An *iterative* procedure, i.e. we repeat the following an appropriate number of times

1. Estimate the association between the variable with fewest missings (V1) and the other explanatory variables
2. Impute from this model the missing values of V1
3. Estimate the association of the variable with 2nd fewest missings (V2) and the other explanatory variables including the imputed V1
4. Impute from this model the missing values of V2
5. Repeat steps 3 and 4 for V3, V4, …, VK
6. Repeat the above 10 times, say
7. Impute all variables from the K estimated models to create one complete dataset
8. Repeat all of the above *m* times, 100 say, to create *m* complete datasets

MICE in Stata

- Three steps

1. Declare data to be MI data:
   ```
   mi set flong
   ```
2. Declare the variable to be imputed
   ```
   mi register imputed V1 V2 V3
   ```
3. Do the prediction and imputation in a single command:
   ```
   mi impute chained (regress) V1 ///
                     (logit) V2 ///
                     (mlogit) V3 = var1 var2, add(100)
   ```

- Yields 100 complete datasets, where the variables V1, V2, V3 no longer have missing values

Analysis of imputed data

Rubin's rule

- For each of the m imputed datasets we get the estimates $\hat{\theta}_j$ and $SE(\hat{\theta}_j)$
- As overall estimate we use the average estimate:

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

- As uncertainty estimate we use the combined SE:

$$SE(\hat{\theta}) = \sqrt{\overline{SE}^2(\hat{\theta}_j) + \frac{m+1}{m}\left(\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \hat{\theta})^2\right)}$$

- Note: Can be implemented in any spreadsheet
- Is automated in Statas –mi estimate–

Exercise 2 – expanding the model

- In the paper by Jacobsen (2008), age, height and weight change (Table 4) are part of the final model
- Jacobsen (2008) stratify on sex – we will ignore this and instead include sex as a covariate
- Expand the previous model (Exercise 1) to include these variables in the imputation step (hint: Look at the help page for -mi impute chained- towards the bottom to find relevant examples)
- Examine the specified regression equations – make a list of the assumptions which must be checked?
- Consider the variable -vaegtaend- :
  Open the original dataset, and perform the regression of -vaegtaend- on the other variables in the model specified above (use the -predict- command)
  - Check the normality of the residuals
  - Plot residuals against each of the explanatory variables and the predicted values

Exercise 2 – expanding the model (cont'd)

- Consider the variable packryg
- We will now check the assumptions relevant for its imputation model
- Open the dataset Wooddata2.dta
- Create two binary variables:

```
generate packryg1_2 = (packryg == 2) ///
                  if packryg <= 2 & !missing(packryg)
generate packryg1_3 = (packryg == 3) ///
                  if packryg != 2 & !missing(packryg)
```

- Is the association of the two variables linear on the log-odds scale with respect to age?

  - Explore this by categorizing age and use this as explanatory variable together with age as continuous

- Discuss the implications of your findings – how should packryg and vaegtaend be imputed?

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(10)
```

- 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(10)
```

Understanding mi estimate

- Look at the imputed dataset again
- 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: `regress fev1aendaar i.packryg i.wooddustgrp if _mi_m == 1`)
- Now run the `mi estimate` with the same regression, but with the option noisily: `mi estimate, noi: regress ...`
- Compare with what you got, when you ran a regression for each imputed dataset

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 (Hint: Make rectangles of different colors/shading to symbolize subsets of observations and variables)

Exercise 3: Working with different formats (continued)

- Now find two other students with two different data types and try to 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
- Run an mi estimate command on each format to verify that results do not change regardless of format

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

Exercise 4 – Improving on Jakobsen (2008)

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)

Exercise 5: Separate imputations for men and women

- Repeat the previous exercise, 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 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

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
```

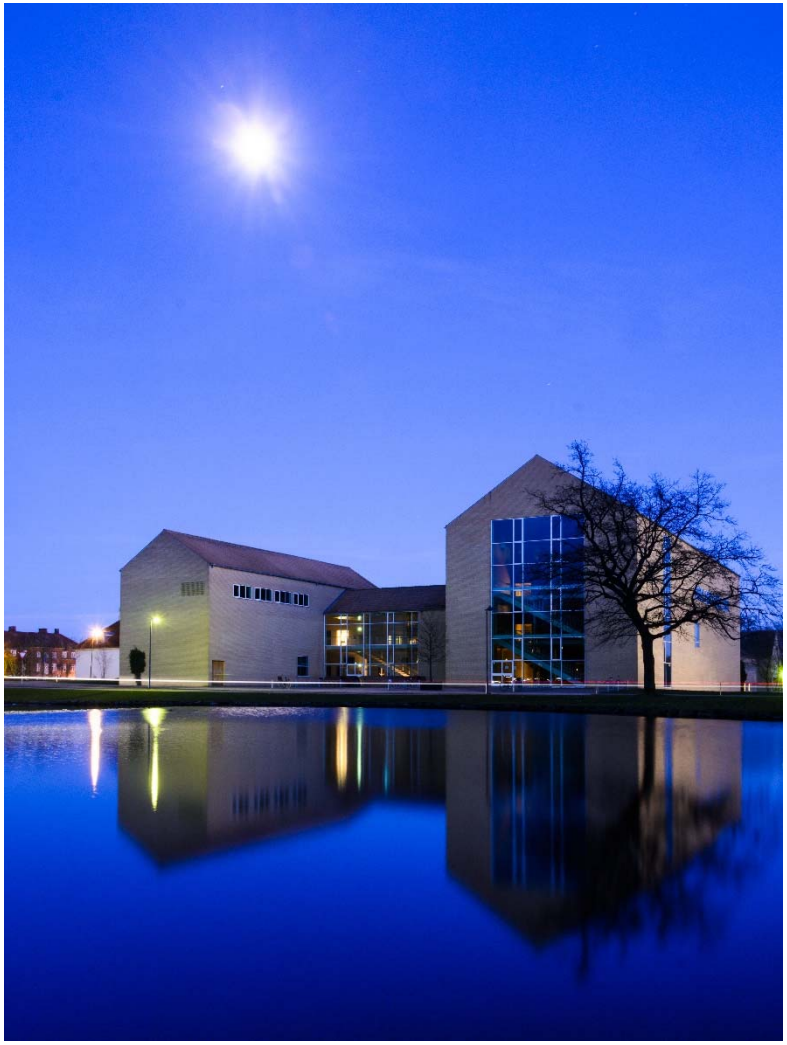
Other useful options

- 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

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)