Strategies for dealing with Missing Data

Maarten L. Buis

Institut für Soziologie
Eberhard Karls Universität Tübingen
http://www.maartenbuis.nl

What do we want from an analysis strategy?

- Simple example
  - We have a theory that working for cash is mainly “men’s work” and collecting stuff from the forest is mainly “women’s work”.

What do we want from an analysis strategy?

- Simple example
  - We have a theory that working for cash is mainly “men’s work” and collecting stuff from the forest is mainly “women’s work”.
  - We go to PEN and make a table

% women	% forest income
0–25	
25–50	
50–75	
75–100	

What do we want from an analysis strategy?

- Simple example
  - We have a theory that working for cash is mainly “men’s work” and collecting stuff from the forest is mainly “women’s work”.
  - We go to PEN and make a table

% women	% forest income
0–25	
25–50	
50–75	
75–100	

- Question → Observe stuff → Answer

What do we want from an analysis strategy?

- Simple example
  - We have a theory that working for cash is mainly “men’s work” and collecting stuff from the forest is mainly “women’s work”.
  - We go to PEN and make a table

% women	% forest income
0–25	
25–50	
50–75	
75–100	

- Question → Observe stuff → Answer
- analysis strategy is just there to summarize the observed stuff so we can see the answer

What are missing data?

person or company or village	var1	var2	var3
1	2	3	5
2	3	7	3

What are missing data?

person or company or village	var1	var2	var3
1	2	3	5
2	3	7	3
3	4	?	5

What is the problem?

- you can quickly loose a frightening proportion of your data

What is the problem?

- you can quickly loose a frightening proportion of your data
  - composite variable, like income

What is the problem?

- you can quickly loose a frightening proportion of your data
  - composite variable, like income
  - regression with multiple variables

What is the problem?

- you can quickly loose a frightening proportion of your data
    - composite variable, like income
    - regression with multiple variables
- you may not measure what you want to measure (bias)
    - prices more often forgotten remembered when collected small amounts

Imputation

- We want to be able to use the observed part of a case without “adding” information on the missing data.

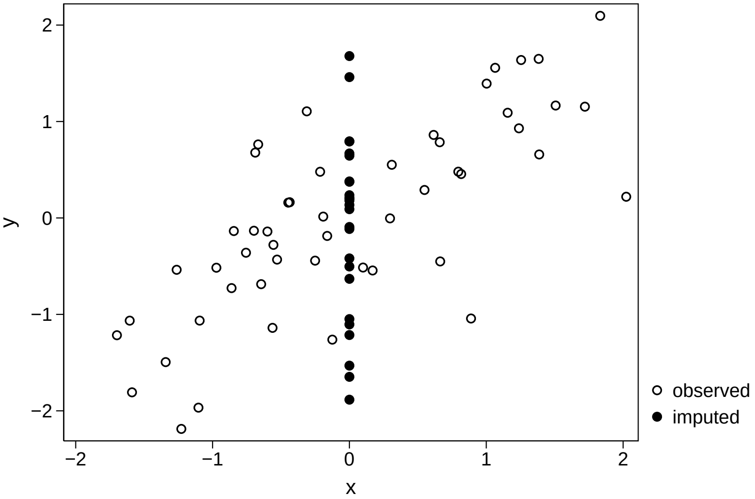
Imputation

- We want to be able to use the observed part of a case without “adding” information on the missing data.
- Reproducing patterns in observed data on the missing data, so we can use the observed part of a case

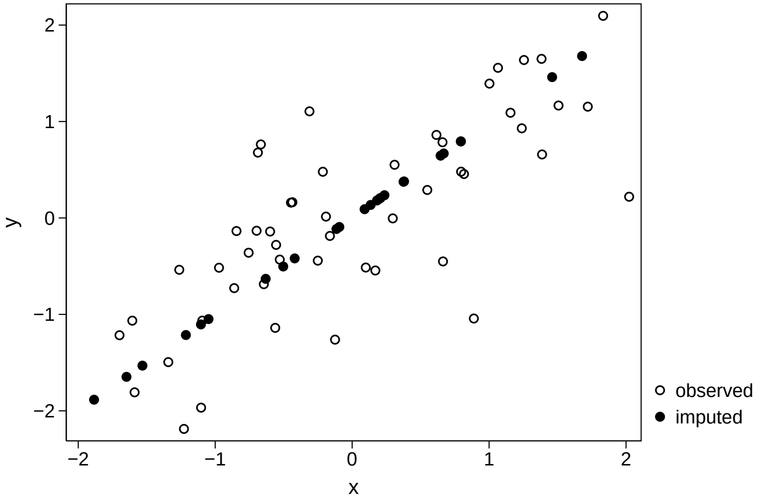
Imputation

- We want to be able to use the observed part of a case without “adding” information on the missing data.
- Reproducing patterns in observed data on the missing data, so we can use the observed part of a case
- Not recovering the lost observation

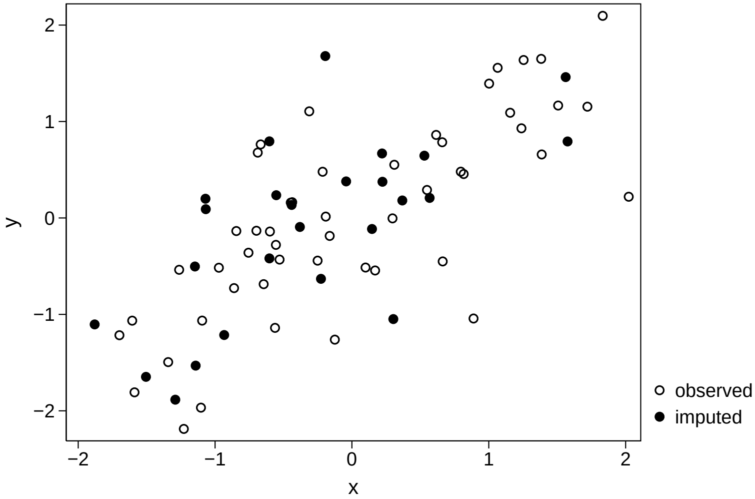
mean median mode imputation



regression imputation



regression imputation + uncertainty



missingness dummy (1)

- Mean imputation is simple, can’t we safe it?

missingness dummy (1)

- Mean imputation is simple, can't we safe it?
- What about adding a control variable that says that it is imputed?

$$y = b_0 + b_1 x + b_2 D$$

missingness dummy (1)

- Mean imputation is simple, can't we safe it?
- What about adding a control variable that says that it is imputed?

$$y = b_0 + b_1 x + b_2 D$$

- What happens when x is observed:

$$y = b_0 + b_1 x + b_2 0$$

$$y = b_0 + b_1 x$$

missingness dummy (1)

- Mean imputation is simple, can't we safe it?
- What about adding a control variable that says that it is imputed?

$$y = b_0 + b_1 x + b_2 D$$

- What happens when x is observed:

$$y = b_0 + b_1 x + b_2 0$$

$$y = b_0 + b_1 x$$

- What happens when x is missing:

$$y = b_0 + b_1 \bar{x} + b_2 1$$

$$y = b_0^*$$

missingness dummy (2)

- What if the missingness happens in a control variable (z)

$$y = b_0 + b_1 x + b_2 z + b_3 D$$

missingness dummy (2)

- What if the missingness happens in a control variable (z)

$$y = b_0 + b_1 x + b_2 z + b_3 D$$

- What happens when z is observed:

$$y = b_0 + b_1 x + b_2 z + b_3 0$$

$$y = b_0 + b_1 x + b_2 z$$

missingness dummy (2)

- What if the missingness happens in a control variable (z)

$$y = b_0 + b_1 x + b_2 z + b_3 D$$

- What happens when z is observed:

$$y = b_0 + b_1 x + b_2 z + b_3 0$$

$$y = b_0 + b_1 x + b_2 z$$

- What happens when z is missing:

$$y = b_0 + b_1 x + b_2 \bar{z} + b_3 1$$

$$y = b_0^* + b_1 x$$

missingness dummy (2)

- What if the missingness happens in a control variable (z)

$$y = b_0 + b_1 x + b_2 z + b_3 D$$

- What happens when z is observed:

$$y = b_0 + b_1 x + b_2 z + b_3 0$$

$$y = b_0 + b_1 x + b_2 z$$

- What happens when z is missing:

$$y = b_0 + b_1 x + b_2 \bar{z} + b_3 1$$

$$y = b_0^* + b_1 x$$

- b_1 is now a mixture of the effect of x while controlling for z, and the effect of x while not controlling for z.

multiple imputation

- OK, so regression + uncertainty seems to be the way to go, but...

multiple imputation

- OK, so regression + uncertainty seems to be the way to go, but...
- Shouldn't the imputed observations count less than observed observations?

multiple imputation

- OK, so regression + uncertainty seems to be the way to go, but...
- Shouldn't the imputed observations count less than observed observations?
- Aren't we artificially increasing our sample size if we don't do that?

multiple imputation

- OK, so regression + uncertainty seems to be the way to go, but...
- Shouldn't the imputed observations count less than observed observations?
- Aren't we artificially increasing our sample size if we don't do that?
- Yes

multiple imputation

- OK, so regression + uncertainty seems to be the way to go, but...
- Shouldn’t the imputed observations count less than observed observations?
- Aren’t we artificially increasing our sample size if we don’t do that?
- Yes
- This is where multiple imputation comes in

multiple imputation (2)

- Imputed values are random draws.

multiple imputation (2)

- Imputed values are random draws.
- Give each missing value multiple imputed values.

multiple imputation (2)

- Imputed values are random draws.
- Give each missing value multiple imputed values.
- Result: multiple imputed datasets.

multiple imputation (2)

- Imputed values are random draws.
- Give each missing value multiple imputed values.
- Result: multiple imputed datasets.
- Do your analysis separately in each dataset.

multiple imputation (2)

- Imputed values are random draws.
- Give each missing value multiple imputed values.
- Result: multiple imputed datasets.
- Do your analysis separately in each dataset.
- Summarize the results:
    - point estimate is the average of the point estimates

multiple imputation (2)

- Imputed values are random draws.
- Give each missing value multiple imputed values.
- Result: multiple imputed datasets.
- Do your analysis separately in each dataset.
- Summarize the results:
    - point estimate is the average of the point estimates
    - Uncertainty about the point estimate (standard error) is a combination of:
        1. the average uncertainty and
        2. the degree to which the results in the different datasets differ from one another

multiple imputation (3)

- what variables to include in imputation model?

multiple imputation (3)

- what variables to include in imputation model?
- All variables included in your model of interest.

multiple imputation (3)

- what variables to include in imputation model?
- All variables included in your model of interest.
- Including your dependent variable.

Where we are going with PEN

- Challenges
  - we don’t know how the dataset is going to be used
  - not all users have access to the way of analyzing multiply imputed data

Where we are going with PEN

- Challenges
  - we don’t know how the dataset is going to be used
  - not all users have access to the way of analyzing multiply imputed data
- Criteria
  - document what is real and what is imputed
  - useable for users
  - conservative (don’t impute everything)

Where we (think we) are going with PEN

- By default users will get no imputed values.

Where we (think we) are going with PEN

- By default users will get no imputed values.
- Optionally, they can get single imputed data

Where we (think we) are going with PEN

- By default users will get no imputed values.
- Optionally, they can get single imputed data
- Only impute prices and quantities

Where we (think we) are going with PEN

- By default users will get no imputed values.
- Optionally, they can get single imputed data
- Only impute prices and quantities
- We expect most people to use income components from different sectors

Where we (think we) are going with PEN

- By default users will get no imputed values.
- Optionally, they can get single imputed data
- Only impute prices and quantities
- We expect most people to use income components from different sectors
- This would consist of e.g. price$_{\text{fire wood}} \times$ quantity$_{\text{fire wood}} +$ price$_{\text{brazil nuts}} \times$ quantity$_{\text{brazil nuts}}$

Where we (think we) are going with PEN

- By default users will get no imputed values.
- Optionally, they can get single imputed data
- Only impute prices and quantities
- We expect most people to use income components from different sectors
- This would consist of e.g. price$_{\text{fire wood}}\times$ quantity$_{\text{fire wood}}+$ price$_{\text{brazil nuts}}\times$ quantity$_{\text{brazil nuts}}$
- So this is where there are 4 opportunities of getting a missing value, and each imputed value adds relatively little information.

Recommendations

- While in the field collect as much information as possible.

Recommendations

- While in the field collect as much information as possible.
- If you do your own imputation go for multiple imputation.
  - Use all explanatory variables *and* the explained variable.
  - Look at the imputed and observed values (graphs).

Recommendations

- While in the field collect as much information as possible.
- If you do your own imputation go for multiple imputation.
  - Use all explanatory variables *and* the explained variable.
  - Look at the imputed and observed values (graphs).
- If you impute for general use data
  - Document the imputed values.
  - Get an idea of how the data is going to be used.
  - Make a trade-off between ‘correct’ statistical procedure and what the users are comfortable with using.