Combining latent class analysis and multiple imputation to correct for misclassification in combined datasets

Laura Boeschoten

Tilburg University – Statistics Netherlands

18 oktober 2018

Outline

Starting point: Combined dataset

Multiple sources:

- Administrative sources
  - (Large part of) population
- Surveys
  - Sample of population
- Linked on person level

Different sources sometimes contain the same (categorical) variable

Official Statistics output

Predefined large cross-tables that meet a number of requirements

Requirement 1
Account for impossible combinations of scores

Problem:

- Sometimes, a combination of scores is observed that is not possible in practice
- Caused by misclassification in one of the variables (De Waal, Pannekoek & Scholtus, 2011)

Solution:

- Incorporate edit restrictions into the LC model

Example:

$$P(\text{Gender} = \text{Male} | \text{Pregnant} = \text{Yes}) = 0$$

Requirement 2

Numerical consistence over different cross-tables

Problem:

- Sometimes variables are only observed by means of sample surveys
- Weighting leads to inconsistent estimates of the different cross-tables (De Waal, 2014)

Solution:

- Mass imputation

Requirement 3

Incorporate uncertainty into the variance estimates

Problem:

- Missing and conflicting values make us more uncertain about our estimates

Solution:

- Multiple imputation

Summary

Multiple Imputation of Latent Classes (MILC)

Latent Class model

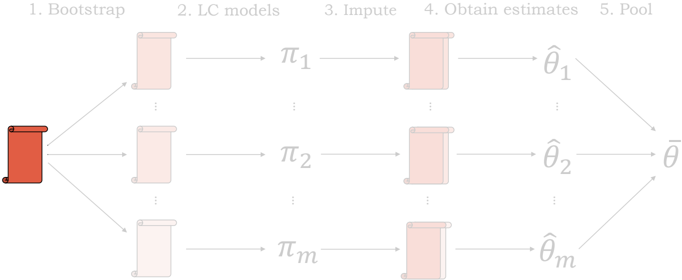
- Variables that measure the same construct as indicators of a latent variable that measures the 'true scores'
- Edit restrictions

Multiple Imputation

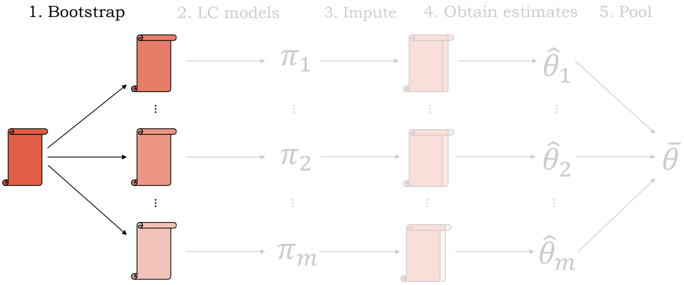
- Consistent estimates
- Uncertainty due to missing and conflicting values

Multipele Imputation of Latent Classes

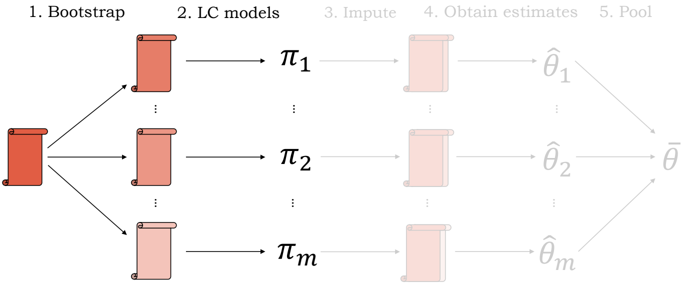
MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

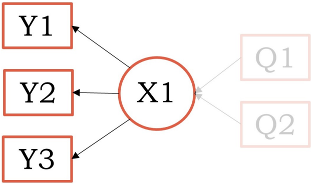
MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

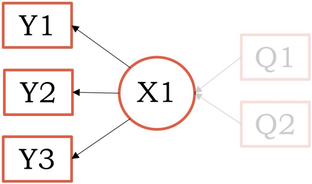
MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

LC model



- Variables measuring the same construct used as indicators of a latent variable
- Number of LC's = number of categories in indicators



- Mixture:

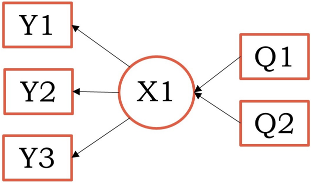
$$\mathrm{P}(\mathbf{Y}) = \sum_x \mathrm{P}(X)\mathrm{P}(\mathbf{Y}|X)$$

- Local independence

$$\mathrm{P}(\mathbf{Y}|X) = \prod_l \mathrm{P}(Y_l|X)$$

LC model

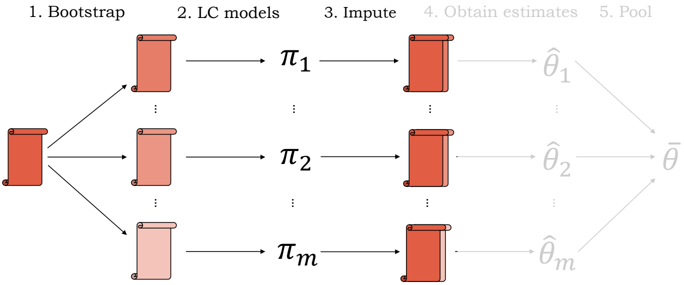
Covariates



- Misclassification independent of covariates
- Covariates are free of error
- Edit restrictions are ‘hard edits’:

$$P(X = 2|Q = 1) = 0$$

MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

Multiple imputation

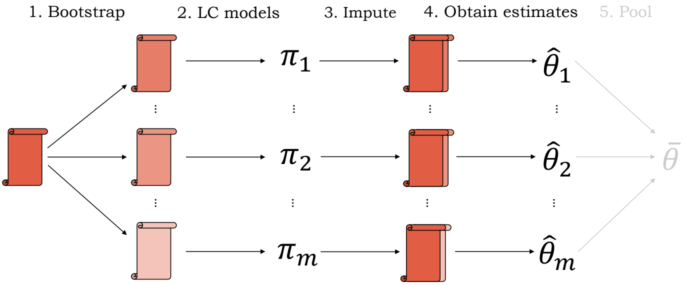
Using posteriors

- For every case, one imputation is created using the posterior belonging to the corresponding profile

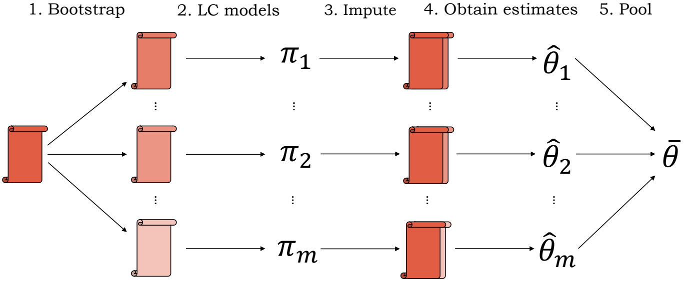
$$P(X|\mathbf{Y}) = \frac{P(X)\prod_l P(Y_l|X)}{\sum_x P(X)\prod_l P(Y_l|X)}$$

- This is done for every LC model (belonging to every bootstrap sample)
- Resulting in m imputations

MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

MILC


1. Bootstrap
2. LC models
3. Impute
4. Obtain estimates
5. Pool
π_1
π_2
π_m
$\hat{\theta}_1$
$\hat{\theta}_2$
$\hat{\theta}_m$
$\bar{\theta}$

Pooling of the results

Using the pooling rules defines by Rubin

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

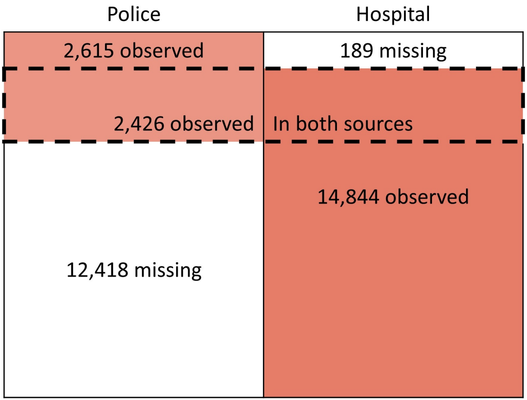
- T = Total variance
- $\bar{U}$ = 'Within' variance (uncertainty about the assigend score)
- B = 'Between' variance (uncertainty about the model)

Simulation conclusions

- Quality of the imputations depends on the quality of the data
    - Entropy R^2: Score between 0 and 1 indicating how well you can classify based on your observed data
- Required quality depends on desired output
- Low number of imputations seems sufficient
- (Boeschoten, Oberski & De Waal, 2017)

Application: Number of serious road injuries per vehicletype

Number of serious road injuries per vehicletype in 2013

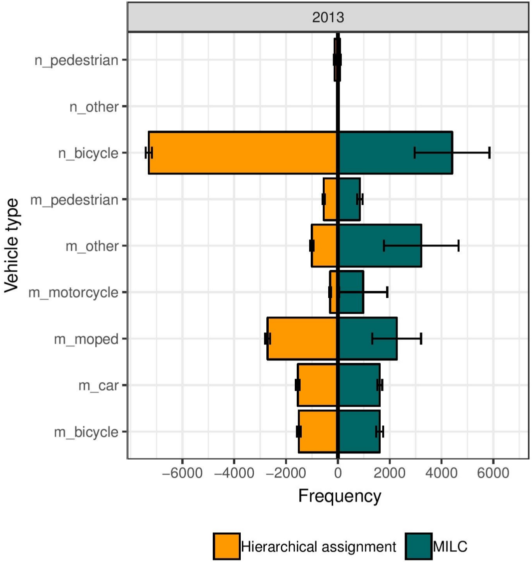


In total 15,033 serious road injuries

Number of serious road injuries per vehicletype in 2013

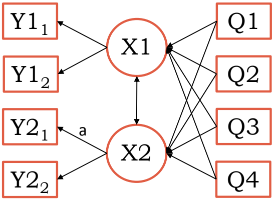
- Nine categories for vehicletype
- Currently, if the police assigned a score, this is used. Otherwise, the score assigned by the hospital is used
- Vehicletype is classified differently for 30% of the cases

Number of serious road injuries per vehicletype in 2013



Simultaneously impute missing values

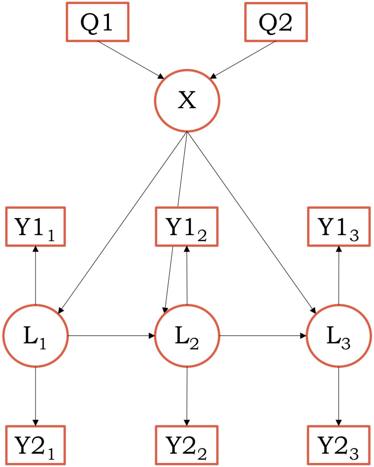
By using a quasi-latent variable



- Simultaneously impute the variable 'region of accident'
- Only measured by police
- restriction 'a': 'region of accident is a perfect indicator of X2.
- If missing: LC model is used, with 'region of hospital' as another indicator

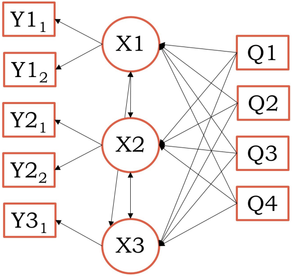
EXTENSIONS OF THE METHOD

Extension 1: Longitudinal



- Model to estimate monthly employment rates (Pavlopoulos & Vermunt, 2015)
- Create monthly imputations of employment status
- Investigated and applied in collaboration with ISTAT

Extension 2: Population census



- Comparable to model discussed in application
- Simulate from finite population
  - Different approach for bootstrap
- Many observations; many cells

Extension 3: External covariates

Using three-step methodology

- Estimate the relationship between an imputed latent variable and external covariates (not included in the LC model)
- Apply a correction procedure (ML or BCH)
- Update the imputations with the newly obtained posteriors
- (Boeschoten, Oberski, De Waal & Vermunt, 2018) & (Boeschoten, Croon & Oberski, 2018)

Summary

- Quality of the output depends on the quality of the data
- Missing values can be imputed simultaneously
- MILC can easily be adjusted to specific situations
- Covariates can be added at a later time-point

Discussion

- Is there a more flexible alternative for the bootstrap?
  - Gibbs sampler?
- How can we use MILC for other types of errors?
  - Or use it in combination with other types of correction methods?

Literature

- A.G. De Waal, J. Pannekoek, S. Scholtus (2011) Handbook of statistical data editing and Imputation *John Wiley & Sons Inc.*
- A.G. De Waal (2014) Consistent estimates for categorical data based on a mix of administrative data sources and surveys *CBS Discussion Paper*
- L. Boeschoten, D.L. Oberski & A.G. De Waal (2017) Estimating classification errors under edit restrictions in composite survey-register data using multiple imputation latent class modelling *Journal of Official Statistics 33-4*
- L. Boeschoten, D.L. Oberski, A.G. De Waal & J.K. Vermunt (2018) Updating latent class imputations with external auxiliary variables *Structural Equation Modeling: A Multidisciplinary Journal*
- L. Boeschoten, M. Croon & D.L. Oberski (2018) A note on applying the BCH method under linear equality and inequality constraints *Journal of Classification*
- D. Pavlopoulos & J.K. Vermunt (2015) Measuring temporary employment. Do survey or register data tell the truth? *Survey Methodology* 41-1, p.197-214

Thank you!!

l.boeschoten@tilburguniversity.edu