Q 2016
European Conference on Quality in Official Statistics

Latent Class Multiple Imputation for multiply observed variables in a combined dataset

Session 15 – June 2, 2016

L. Boeschoten MSc, dr. D.L. Oberski, prof. Dr. A. G. de Waal

Tilburg University & Statistics Netherlands

l.boeschoten@tilburguniversity.edu

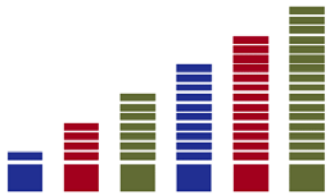
INE Instituto Nacional de Estadística eurostat

Madrid, May 31 - June 3

Q 2016
European Conference on Quality in Official Statistics

Overview

- Introduction
- MILC method
- Simulation
- Application
- Conclusion and discussion



Introduction

Combined datasets

- Registers and surveys
- Linked on unit level
- Examples: Dutch SSD or 2011 census
- Categorical variables
- Used to produce large tables (hypercubes)

Measurement errors in a combined dataset

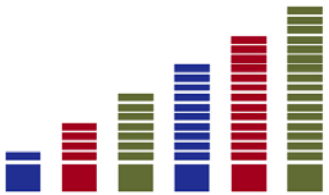
Invisibly and visibly present

Invisibly present errors

- Compare responses on indicators measuring the same latent "true" variable within a combined dataset
- Latent variable models

Visibly present errors

- Logical relations between variables make errors visibly present
- Edit rules



Measurement errors in a combined dataset
Solutions

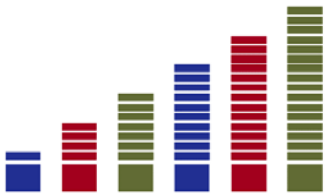
Goal: To estimate cross tables between variables, taking measurement error and edit restrictions into account, and the extra uncertainty this creates.

Invisibly present errors: Multiple indicators from combined dataset

Visibly present errors: Restriction covariates

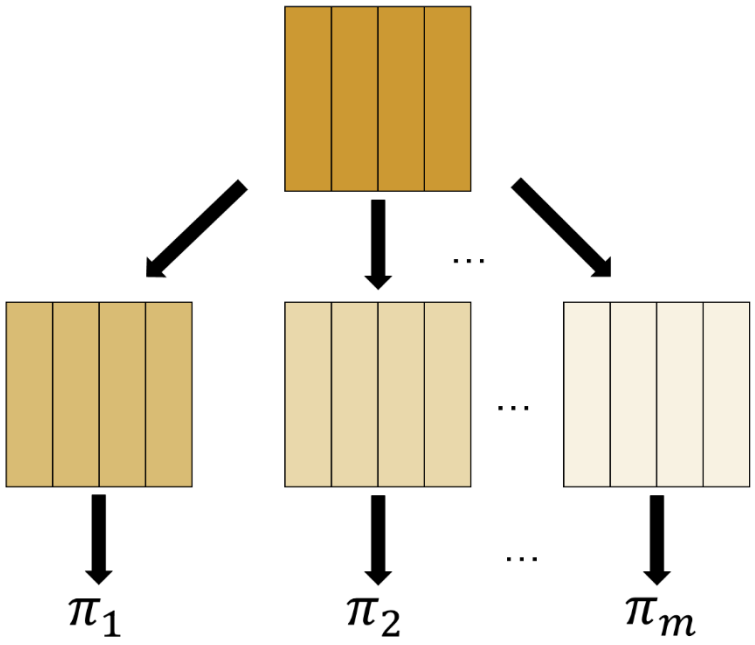
Estimation: Latent Class analysis

Further analyses: Multiple Imputation



MILC method

Step by step



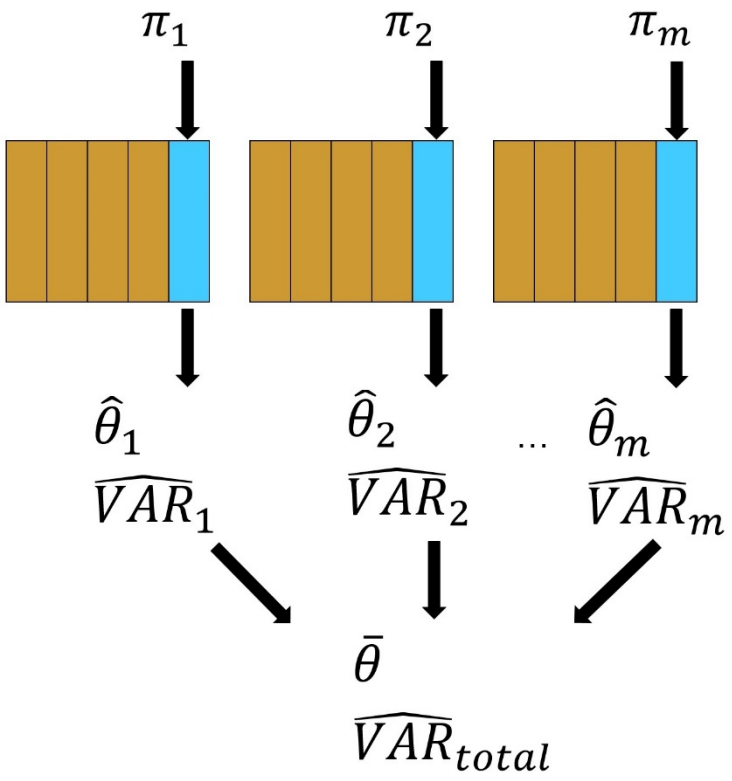
1. Original combined dataset

2. *M* bootstrap samples of the combined dataset

3. M Latent Class models

MILC method

Step by step



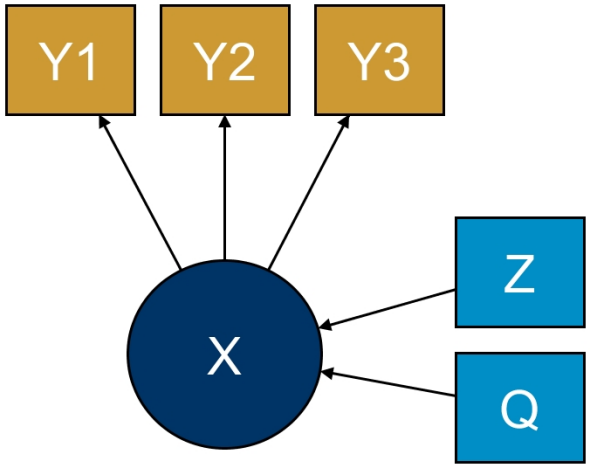
3. *M* Latent Class models

4. *M* new imputed variables in the original combined dataset

5. Estimates of the new imputed variables

6. Pool the estimates using Rubin's rules

Simulation approach



Data generation

- Three dichotomous indicators (*Y1, Y2, Y3*) of latent variable *X*
- Dichotomous covariate *Q* and restriction covariate *Z*

Reference values

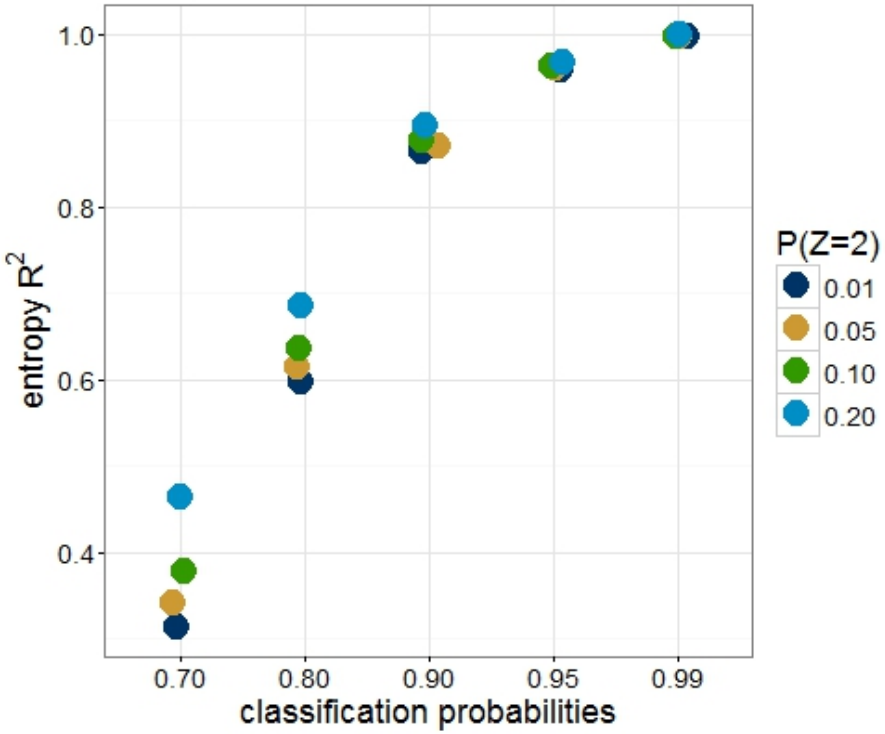
- *2x2* table of imputed latent variable *W* and *Z*
- Logistic regression of *W* on *Q*

Simulation conditions

- Different classification probabilities
- Different *P(Z)* & *P(Q)*

Simulation approach
Entropy R^2

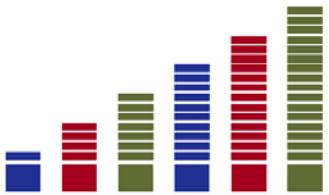
- How well can you predict class membership based on the observed variables?
- Score between *0* and *1*
- *1* means perfect prediction



Q 2016
European Conference on Quality in Official Statistics

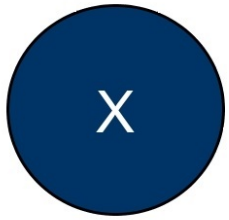
Simulation conclusions

- Quality of the results is very dependent on entropy R^2 of the LC model
- "True" logistic regression estimates can be obtained when the entropy R^2 is at least *0.60*
- "True" cross table counts under edit restrictions can be obtained when the entropy R^2 is at least *0.90*



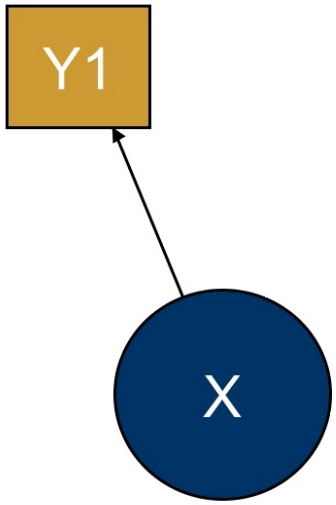
Application on a combined dataset

Latent dichotomous variable *X* measuring *home ownership* (1 =“own”, 2 =“rent”)



Application on a combined dataset

Latent dichotomous variable *X* measuring *home ownership* (1 =“own”, 2 =“rent”)

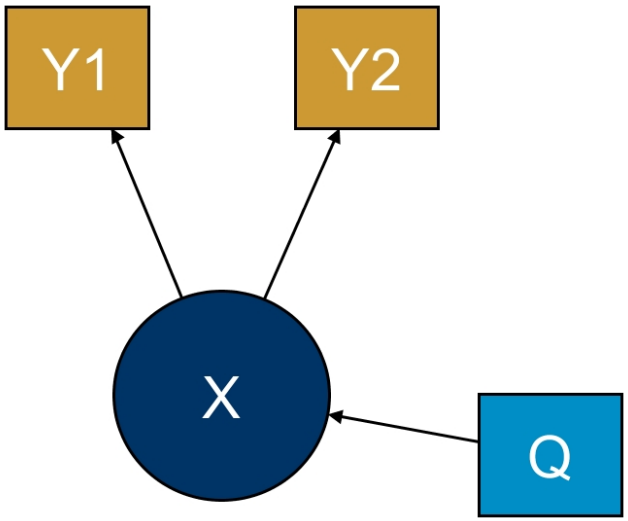


BAG register:

- 1 indicator

Application on a combined dataset

Latent dichotomous variable *X* measuring *home ownership* (1 =“own”, 2 =“rent”)



BAG register:

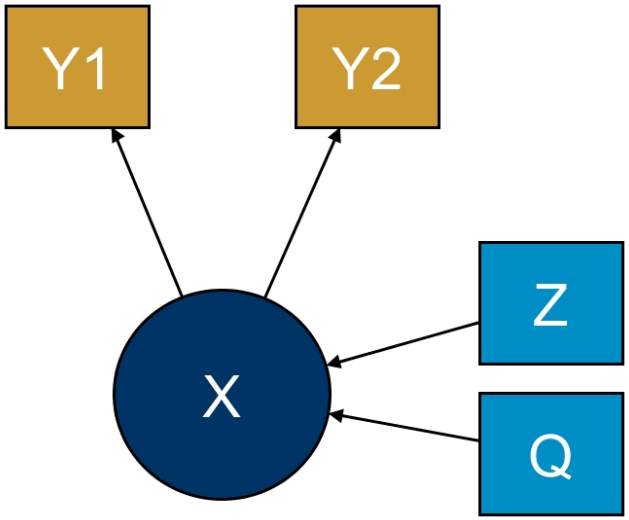
- 1 indicator

LISS background study

- 1 indicator
- 1 covariate: *marriage* (1=“married”, 2=“not married”)

Application on a combined dataset

Latent dichotomous variable *X* measuring *home ownership* (1 =“own”, 2 =“rent”)



BAG register:

- 1 indicator

LISS background study

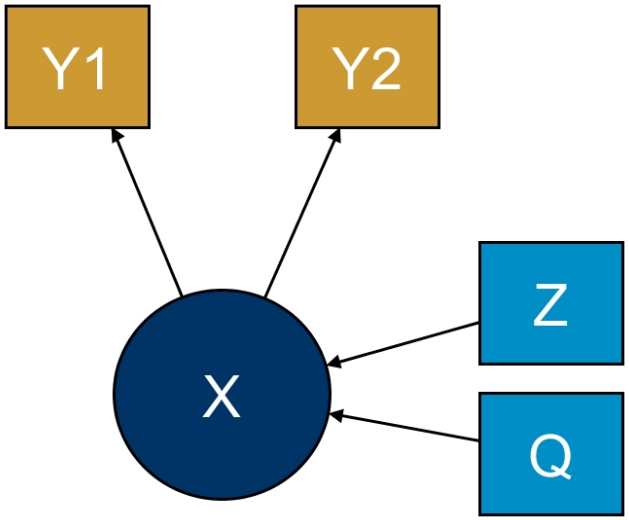
- 1 indicator
- 1 covariate: *marriage* (1=“married”, 2=“not married”)

LISS core study on housing

- 1 restriction covariate: *rent benefit* (1=“yes”, 2=“no”)

Application on a combined dataset

Latent dichotomous variable *X* measuring *home ownership* (1 =“own”, 2 =“rent”)



BAG register:

- 1 indicator

LISS background study

- 1 indicator
- 1 covariate: *marriage* (1=“married”, 2=“not married”)

LC model has an entropy R^2 of *0.93*

LISS core study on housing

- 1 restriction covariate: *rent benefit* (1=“yes”, 2=“no”)

Application on a combined dataset

2x2 table of imputed latent variable *home ownership* and restriction covariate *rent benefit*

	P(own x r.b.)	P(rent x r.b.)	P(own x no)	P(rent x no)
BAG register	0.005	0.295	0.055	0.644
LISS background	0.010	0.289	0.029	0.672
MILC	0.000	0.295	0.021	0.679

Application on a combined dataset

Logistic regression of covariate *marriage* on imputed latent variable *home ownership*

	Intercept		**Not married**	
	estimate	95% CI	estimate	95 % CI
BAG register	2.466	[2.209; 2.723]	-1.233	[-1.390; -1.076]
LISS background	2.762	[2.490; 3.034]	-1.304	[-1.468; -1.141]
MILC	2.822	[2.553; 3.091]	-1.416	[-1.685; -1.147]

Conclusion and discussion

Conclusion

- Quality of the results is very dependent on entropy R^2 of the LC model
- Different entropy R^2 values are required for different types of estimates
- MILC appeared to be useful in practice

Discussion

- Covariates
- Missing values