An empirical evaluation of latent class models for multisource statistics

Di Consiglio L., Di Zio M., Filipponi D.

Istat, Italian National Institute of Statistics

ITACOSM 2019 - Florence 5-7 June 2019

Outline

1. Background - Motivational case
2. Accuracy for LCM Register-based statistics
3. Resampling algorithms for accuracy estimation
4. Empirical evaluation
5. Final remarks and next steps

Integrated System of Registers in Istat

- Istat is currently in the middle of a strong modernization effort aimed at overcoming traditional stovepipe production model
- The backbone of the new production system will be the 'Integrated System of Statistical Registers' (ISSR)
- A system of connected registers that will be used as reference for all the statistical programs carried out by Istat
- Multisource context. ISSR will integrate as much as possible administrative data and survey data concerning related topics

Variables

- Registers contain some ('core') variables
- In register of population: Place and date of birth, gender, citizenship, attained level of education, employment status

Variables

- Variables will be used as reference for all the statistics produced in Istat
- Estimates on those variables will be 'register-based' statistics.
- Register-based statistics. Computation of the target parameter directly on register data: Mean, median, ...

Variable prediction in a multisource informative context

- Some core variables are easily obtained by using admin data, see for instance sex, age, (high reliability of admin data).
- For other core variables, although admin data are strongly related to the target variable, a model should be used for the prediction
- a sample can be used to improve the prediction

Mass imputation - latent model

- Two strategies can be envisaged: *Supervised* and *unsupervised* learning
- Supervised approach. A source is taken as reference, i.e., the variable observed in the source is considered as target variable (gold standard).
- Supervised approach with a sample survey: Mass imputation
- Unsupervised approach. All sources contain information close to the target variable, but none of them can be directly assumed as target variable.
- In this case, a latent variable model can be adopted to predict target values

LC Model

Goal: Estimation of a latent variable,
e.g., employment status two categories: 0= *not employed* , 1= *employed*,

- Latent variable Y^*: (true employment status $Y^* \in \{0,1\}$)
- Observed measures Y_i, for $i = 1, \ldots, k$: (employment status according to the i-th source $Y_i \in \{0,1\}$)
- covariates
  - X: *e.g., retirement status, student, income, age, sex*
- Target. Prediction of Y^* for all the units int the register using the estimated conditional probabilities $Pr(Y^*|Y, X)$

Example in ISSR

- Mass imputation of *attained level of education*.
    - Supervised approach: Admin data on course attendance, sample survey.
- Unsupervised approach: Hidden Markov Models (HMM) for the estimation of *monthly employment status.*[1]

[1] *Filipponi et al., (2018).*

ACCURACY EVALUATION FOR REGISTER-BASED STATISTICS

- Mass imputation for level of education: *Scholtus* (2018) proposes analytical and resampling techniques
- We study a bootstrap approach to evaluate accuracy of a LC model w.r.t. two frameworks
  - design based
  - model-design based
- other random mechanisms affecting accuracy are neglected (linkage, nonresponse,...).

Pseudo-population bootstrap - Design based

(Chauvet 2007, Shao Sitter 1996)

1. generation of ONE pseudo-population U^* from observed data (sample S integrated with admin data).
2. Take a bootstrap sample S^* from U^* using the same sampling design that led to S.
3. estimate the latent model for imputation, predict the values of the latent variable Y^* over the register, compute the bootstrap statistics $\hat{\theta}^*$
4. Repeat Steps 2 and 3 a large number of times, B, to get $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$
5. Define $v\hat{a}r^* = \frac{\sum_{b=1}^{B}(\hat{\theta}^*_b - \hat{\theta}^*_{(\cdot)})^2}{(B-1)}$, where $\hat{\theta}^*_{(\cdot)} = \frac{\sum_{b=1}^{B}\hat{\theta}^*_b}{B}$

Pseudo-population bootstrap - Model-Design based - (Chen, Haziza, Leger, Mashreghi, 2019)

1. estimate the LC model $\hat{M}$ on observed data (sample S integrated with admin data)
2. parametric generation of a pseudo-population U^* (including latent variable) w.r.t. $\hat{M}$
3. Draw a bootstrap sample S^* from U^* using the same sampling design that led to S.
4. estimate the latent model for imputation, predict the values of latent variable Y^* over the register, compute the bootstrap statistics $\hat{\theta}^*$
5. Repeat Steps 1 and 4 a large number of times, B, to get $\hat{\theta}^*_1, \dots, \hat{\theta}^*_B$
6. Define $v\hat{a}r^* = \frac{\sum_{b=1}^{B}(\hat{\theta}^*_b - \hat{\theta}^*_{(.)})^2}{(B-1)}$, where $\hat{\theta}^*_{(\cdot)} = \frac{\sum_{b=1}^{B}\hat{\theta}^*_b}{B}$.
7. Alternative to step 6. $v\hat{a}r^* = \frac{\sum_{b=1}^{B}(\hat{\theta}^*_b - \hat{\theta}^*_{U^*})^2}{(B-1)}$ where $\hat{\theta}^*_{U^*}$ is the statistic computed on U^*.

Empirical evaluation based on simulations: LCM (1)

- Standard LCM. 4 dichotomous manifest variables $Y_i \in \{0, 1\}$, one dichotomous X (known without error in the whole population)
- Latent variable $Y^* \in \{0, 1\}$ depends on X
- Target parameter $\theta = \sum_{i=1}^{N} Y_i^*$

Empirical evaluation based on simulations: LCM (2)

- Misclassification errors

	Y_1		Y_2		Y_3		Y_4	
Y^*	0	1	0	1	0	1	0	1
0	0.9	0.1	0.8	0.2	0.9	0.1	0.9	0.1
1	0.1	0.9	0.1	0.9	0.2	0.8	0.05	0.95

- mixing prop $P(Y^* = 1|X = 0) = 0.7$, $P(Y^* = 1|X = 1) = 0.3$

Empirical evaluation based on simulations: LCM (3)

- Large population ($N = 50,000$)
- Observed data. X, Y_1, Y_2, Y_3 observed in the whole pop. Missing on Y_4 with sampling prob depending on X, i.e., sample gathering info on Y_4.

Observed data - Integration of admin and survey data

	Admin			Survey	Lat. Var
X	Y_1	Y_2	Y_3	Y_4	Y^*
$x_{1,1}$	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	?
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	?
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	?
$x_{n,1}$	$y_{n,1}$	$y_{n,2}$	$y_{n,3}$	$y_{n,4}$	?
$x_{n+1,1}$	$y_{n+1,1}$	$y_{n+1,2}$	$y_{n+1,3}$	?	?
$\dots$	$\dots$	$\dots$	$\dots$	?	?
$x_{N,1}$	$y_{N,1}$	$y_{N,2}$	$y_{N,3}$	?	?

Empirical evaluation based on simulations: LCM (4)

- Sampling rate: 2%, 5%, 10%
- Estimate LCM and predict the latent variable Y^* on the register with two methods:
  - expected value of the LCM (EX)
  - random draw from conditional prob. of LCM (RD)
- evaluate the case when X is considered in the mixing proportions of LCM (LCM.X), and when X is not taken into account in the LCM.

Empirical evaluation - Monte Carlo results

Design based

	LCM-EX		LCM-RD		LCM.X-EX		LCM.X-RD	
	rmse	bias	rmse	bias	rmse	bias	rmse	bias
2%	171	171	175	169	147	147	151	144
5%	231	230	234	230	136	134	143	135
10%	323	322	326	322	121	119	130	120

Model-Design based

	LCM-EX		LCM-RD		LCM.X-EX		LCM.X-RD	
	rmse	bias	rmse	bias	rmse	bias	rmse	bias
2%	644	591	645	592	240	4	241	5
5%	612	590	614	591	157	0	160	0
10%	601	591	603	591	104	2	107	0

Empirical evaluation results - Bootstrap

Design based - se estimation - LCM.X

	EX			RD		
	se 2%	se 5%	se 10%	se 2%	se 5%	se 10%
Target MC	16	23	26	47	46	51
Boot	25	28	31	50	52	52

Model-Design based - se estimation - LCM.X

	EX			RD		
	se 2%	se 5%	se 10%	se 2%	se 5%	se 10%
Target MC	240	157	104	241	160	107
BootRD	236	149	108	236	152	112
BootMean	256	182	150	257	185	153

Final remarks and further steps

- Register-based LCM estimates
  - bias in the design context
  - model-design unbiased
- Pseudo-population bootstrap estimates
  - pseudo-population bootstrap gives good results for LCM
  - in model-design, bootstrap with random generation of pseudo-population is preferable
- Next steps
  - apply the pseudo-population bootstrap method to the occupation estimation by means of HMM
  - develop analytical methods for LCM

References

- Filipponi D., Guarnera U., Varriale R. (2019), Hidden Markov Models to Estimate Italian Employment Status. *NTTS 2019, Bruxelles 11-13 March 2019*.
- Scholtus S. (2018). Variances of Census Tables after Mass Imputation, CBS.
- Chauvet G. (2007), Méthodes de bootstrap en population finie. PhD Dissertation, Laboratoire de statistique d'enquetes, CREST-ENSAI, Université de Rennes 2. Available at http://tel.archives-ouvertes.fr/docs/00/26/76/89/PDF/thesechauvet.pdf
- Chen S., Haziza D., Léger C., Mashreghi Z., (2019) Pseudo-population bootstrap methods for imputed survey data, *Biometrika*

Thank you

The motivational case. The informative context of HMM for Employment status

- Admin data
  - Social Security data
  - Chamber of Commerce data
- Sample survey.
  - The labour force survey (LFS)

COMPARING LABOUR FORCE AND ADMIN DATA

TABLE: Cross-classification of the employment status measured by LFS and AS. LFS data, Year 2014

$LFS \setminus AS$	Out	In	Total
Not Employed	52.9	7.3	60.2
Employed	2.5	37.3	39.8
Total	55.4	44.6	100.0

About 10% of units are differently classified

Modeling employment data: HMM.

Filipponi et al., (2018)

Goal: Estimation of the *monthly employment status*

three categories: $1 =$ *employed*, $2 =$ *unemployed*, $3 =$ *others*
two categories: $1 =$ *employed*, $0 =$ *not employed*

- S_t: true employment status (latent)
  $S_t \in (1, 0) \qquad t \in (1, \ldots, 12)$
- Y^L: employment status according to the LFS
  $Y_t^L \in (1, 0)$
- Y^A: employment status according to the AS
  $Y_t^A \in (1, 0)$
- covariates
  - X: *retirement status, student, income, age, sex*
  - Z: *type of administrative sources, admin measure at previous time.*

EMPIRICAL EVALUATION RESULTS - HIGHER ERRORS

Model-Design based - MC

	LCM-EX		LCM.X-EX		LCM-RD		LCM.X-RD	
	rmse	bias	rmse	bias	rmse	bias	rmse	bias
2%	1384	1206	596	16	1383.	1205	599	15
5%	1251	1184	342	8	1250	1182	343	8
10%	1209	1174	247	-5	1210	1174	249	-6

Model-Design based - bootstrap se estimation - LCM.X

	EX			RD		
	se 2%	se 5%	se 10%	se 2%	se 5%	se 10%
Target MC	596	342	247	599	342	249
BootRD	610	378	268	611	380	271
BootMean	607	391	286	608	393	289