Alternative approaches in statistical matching when sample surveys are drawn according to complex survey designs

Mauro Scanu
with ideas taken from previous papers written with Pier Luigi Conti and Daniela Marella, and other stuff discussed with Marco Di Zio, Marcello D’Orazio, Hans Kiesl, Susanne Raessler, Nicola Torelli, Li-Chun Zhang,…

ITACOSM- Bologna, 14 June 2017
Statement of the problem and non-identifiability of the model for the data

How to deal with non-identifiability: matching error and uncertainty

A quick overview of traditional approaches (calibration, file concatenation)

A new approach based on the asymptotic properties of the estimators
Statistical matching

- Two independent samples drawn from the same population
- The only common information is in the X variables (*glue*, according to Reiter)
- Example: Y (Y1) expenditure variables, Z (Y2) income variables (an application on EU Statistics on Income and Living Conditions (EU-SILC) and Household Budget Survey (HBS) is being presented in another session)
Two main problems:

- The model for \((X, Y_1, Y_2)\) is not identifiable given the data sets \(A\) and \(B\) (unless specific models are imposed)

- The two samples could be drawn according to complex survey designs, and it is not of immediate solution how to use survey weights in the statistical matching context
If instead of the samples A and B there was complete knowledge on the distribution of \((Y_1, X)\) and \((Y_2, X)\), the joint \((Y_1, Y_2, X)\) distribution is still problematic for \(Y_1\) and \(Y_2\) given \(X\). Generally speaking, it is possible to say that

\[
\max(0, \ F_{1N}(y_1 | x) + F_{2N}(y_2 | x) - 1) \\
\leq H_{N}(y_1, y_2 | x) \leq \min(F_{1N}(y_1 | x), F_{2N}(y_2 | x))
\]

These are the traditional Fréchet bounds for cumulative distribution functions. This set of distribution is named **uncertainty set**

They can be complemented with additional information, so that this space of distributions becomes narrower
For instance, if $Y_1$, $Y_2$ and $X$ were normal, it is possible to work on the fact that the corresponding variance matrix (were $\sigma_{12}$ is not estimable) should be positive definite (Kiesl and Raessler, 2009).

D’Orazio et al (2006) introduce the notion of restriction of the space of equally plausible distributions given knowledge on $(X,Y_1)$ and $(X, Y_2)$ by introducing:

- Inequalities between parameters
- Structural zeros
Conti, Marella and Scanu (2012) formalize the notion of uncertainty for any kind of model and the presence of structural zeros, and in (2016) give appropriate formulas for the definition of the constrained Fréchet bounds when the constraint follows this form

\[ a_x \leq f_x(y_1, y_2) \leq b_x \]

i.e., the formula becomes

\[ K_{N-}^x(y_1, y_2) \leq H_N(y_1, y_2 \mid x) \leq K_{N+}^x(y_1, y_2) \]
Statistical matching: identifiability

where

\[
K_{N-}^x(y_1, y_2) = \max(0, \min(F_{2N}(y_2 | x), F_{2N}(\gamma_{y_1}(a_x) | x)) + \min(F_{1N}(y_1 | x), F_{1N}(\delta_{y_2}(b_x) | x)) - 1, F_{1N}(y_1 | x) + F_{2N}(y_2 | x) - 1) \\
K_{N+}^x(y_1, y_2) = \min(F_{2N}(y_2 | x), F_{2N}(\gamma_{y_1}(a_x) | x), F_{1N}(y_1 | x), F_{1N}(\delta_{y_2}(b_x) | x)),
\]

It can be seen that

\[
K_{N-}^x(y_1, y_2) \geq \max(0, F_{1N}(y_1 | x) + F_{2N}(y_2 | x) - 1), \\
K_{N+}^x(y_1, y_2) \leq \min(F_{1N}(y_1 | x), F_{2N}(y_2 | x)),
\]
1. A matching procedure is a procedure to choose one d.f. from the uncertainty set.

2. Suppose the chosen distribution is $H_N^*(y_1, y_2|x)$ while the true but unknown distribution (in the same set) is $H_N(y_1, y_2|x)$. The discrepancy between the two distribution is the **matching error**.

3. A **matching error measure**, conditionally on X, is

$$ME_x(H_N^*, H_N) = \int_{\mathbb{R}^2} \left| H_N^*(y_1, y_2|x) - H_N(y_1, y_2|x) \right| \times d \left[ F_{1N}(y_1 |x) F_{2N}(y_2 |x) \right].$$

4. An **unconditional measure of the matching error** is

$$ME(H_N^*, H_N) = \int ME_x(H_N^*, H_N) dQ_N(x).$$
Statistical matching: bounding the matching error

Hence

\[ ME_x(H^*_N, H_N) \leq \Delta^x(F_{1N}, F_{2N}) \forall x, \]

where

\[ \Delta^x(F_{1N}, F_{2N}) = \int_{\mathbb{R}^2} \left( K^{-}_N(y_1, y_2) - K^{+}_N(y_1, y_2) \right) \times d \left[ F_{1N}(y_1 | x) F_{2N}(y_2 | x) \right] \]

which is a measure of the size of the plausible set of distributions given knowledge on \((Y_1, X)\) and \((Y_2, X)\) and constraints.

A bound for the unconditional matching error is

\[ ME(H^*_N, H_N) \leq \Delta(F_{1N}, F_{2N}) = \sum_x \Delta^x(F_{1N}, F_{2N}) p_N(x) \]
Any distribution in the uncertainty set is equally plausible, but can we squeeze further information and favour some of them?

Let’s restrict to parameters, and let’s turn to the old case studied in D’Orazio et al (2006), where $X$, $Y_1$ and $Y_2$ consisted of only categorical variables.

Some parameter values are “more frequent” in the uncertainty set than others!
Let’s turn to the case A and B are just samples, so that the \((X,Y_1)\) and \((X,Y_2)\) distributions can only be estimated.

Traditionally, two approaches have been considered

- Rubin (1986) proposes to concatenate the two samples as if they were a unique sample, with a unique system of sampling weights.

\[
P(A \cup B \ni u) = P(A \ni u) + P(B \ni u) - P(A \cap B \ni u)
\]

The concatenated sample is the starting point for the analyses.

- Renssen (1998) harmonizes the statistical information in A and B.

Hence, the survey weights in the two samples are calibrated in order to obtain known totals, distributions.
A comparison between the two approaches (D’Orazio et al, 2009) does not allow to say that one method outperforms the other. A change in the way to proceed was necessary.

Instead of looking on how efficient estimators based on concatenated samples or on calibration are, let’s look at the asymptotic properties of estimators.

These survey designs are those that, together with a set of technical assumptions on the sequence of sample designs for each N, n₁ and n₂, satisfy

\[ d_H(P_h, P_{R,h}) \to 0 \text{ as } N \to \infty, h = 1, 2. \]

where \( P_h \) are the actual sample designs with inclusion probabilities \( \pi_{1,H}, \ldots, \pi_{N,H}, \) h=1,2, for A and B respectively, while \( P_{R,h} \) are the corresponding conditional Poisson sampling design with the same probabilities as \( P_h \).

This class of survey designs includes simple random sampling, successive sampling, Sampford design, Chao design, stratified desing,…
The idea is to consider \( N, n_h, h=1,2 \), large enough so that we can take advantage of the asymptotic properties of the estimators proposed.

Use the Hajék estimators for the estimable distributions:

\[
\hat{F}_h(y \mid x) = \frac{1}{\hat{N}_h(x)} \sum_{i=1}^{N} \frac{D_{i,h}}{\pi_{i,h}} I(y_{hi} \leq y) I(x_{i}=x), \quad h = 1, 2,
\]

where

\[
\hat{N}_h(x) = \sum_{i=1}^{N} \frac{D_{i,h}}{\pi_{i,h}} I(x_{i}=x), \quad h = 1, 2.
\]
Statistical matching: estimators

It is now possible to get estimates of anything else is based on the estimation of the estimable distributions of $Y_1|X$ and $Y_2|X$, e.g.

\[
\Delta^x = \frac{1}{\hat{N}_1(x)\hat{N}_2(x)} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \hat{K}^x_+(y_{1i}, y_{2j}) - \hat{K}^x_-(y_{1i}, y_{2j}) \right) \\
\times \frac{D_{i,1} D_{i,2}}{\pi_{i,1} \pi_{i,2}} I(x_i=x) I(x_j=x),
\]

(40)

For almost all $(x_i, y_{1i}, y_{2i})$'s values, conditionally on $x_N, y_{1N}, y_{2N}$, the following result holds:

\[
\left( \hat{n}_1^{-1} + \hat{n}_2^{-1} \right)^{-1/2} \left( \Delta - \Delta(F_{1N}, F_{2N}) \right) \\
\xrightarrow{d} N(0, V(F_1, F_2)) \text{ as } N \to \infty
\]

(58)

with

\[
V(F_1, F_2) = \sum_{k=1}^{K} p(x^k) V(F_1, F_2; x^k) \\
+ \frac{(\xi_1 - 1)(\xi_2 - 1)}{(\xi_1 + \xi_2 - 2)^2} \Delta^x(F_1, F_2)' \Sigma \Delta^x(F_1, F_2).
\]
the evaluation of the reliability of a matching distribution can be dealt with in terms of testing the hypotheses:

$$\left\{ \begin{array}{l} H_0 : \Delta^x(F_{1N}, F_{2N}) \leq \epsilon^x \\ H_1 : \Delta^x(F_{1N}, F_{2N}) > \epsilon^x \end{array} \right. \quad (60)$$

Given the (asymptotic) significance level $\gamma$, the null hypothesis $H_0$ is accepted if

$$\hat{\Delta}^x \leq \epsilon_N^x + z_\gamma \sqrt{\hat{V}_x \left( \hat{n}_1(x)^{-1} + \hat{n}_2(x)^{-1} \right)^{1/2}},$$

where $z_\gamma$ is the $\gamma$th quantile of the standard normal distribution, and $\hat{V}_x$ is an appropriate estimator of the variance $\hat{V}(F_1, F_2; x)$. 
Statistical matching: selection of one df

At a population level, i.e. knowing exactly \( F_{1N}(y_1|x) \) and \( F_{2N}(y_2|x) \) a simple idea is the following

1. Choose a starting distribution fulfilling the possible constraints (e.g. structural zeros), \( H_{ST;N}(y_1, y_2|x) \)

2. Use the Iterative Proportional Fitting algorithm (IPF) in order to adjust the joint \( H_{ST;N}(y_1, y_2|x) \) to the marginals \( F_{1N}(y_1|x) \) and \( F_{2N}(y_2|x) \)

The same holds also when A and B are sampled.

A starting distribution could be

\[
\int_B dH_{ST;N}(y_1, y_2|x) = C_{Nx} \int_B I(a_x \leq f_x(y_1, y_2) \leq b_x) d\left[F_{1N}(y_1|x)F_{2N}(y_2|x)\right] \quad \forall B \in \mathcal{B}(\mathbb{R}^2),
\]
1. In statistical matching, the treatment of non-identifiable models is recognized: the uncertainty set is the base for measuring the error

2. Hajék estimators are considered, as well as the asymptotic equivalence of some survey designs with the conditional Poisson sampling design

3. Given the asymptotic nature of the solution, for the weights no harmonization issues are considered here (opposite to Rubin’s and Renssen’s approaches)

4. A test for the width of the uncertainty set, which is a bound on the maximal matching error, is proposed

5. A way to build a candidate distribution inside the uncertainty set based on a “constrained conditional independence assumption” is proposed

6. There’s space for building R packages in this area
That's all Folks!