Istat | Istituto Nazionale di Statistica

12th International Conference the use of R in Official Statistics, uRos2024
Athens, 27-29 November 2024

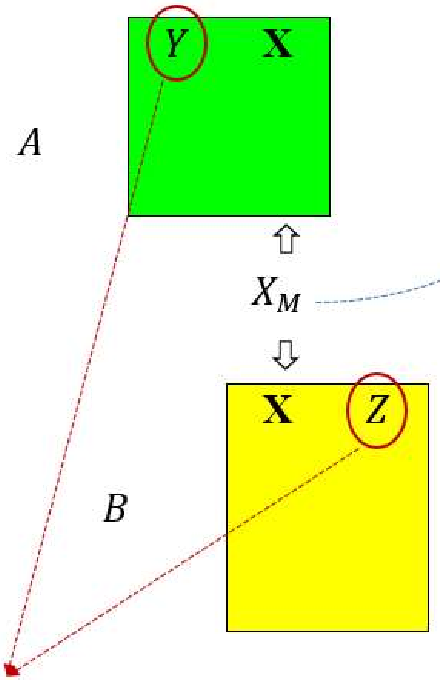
Is Statistical Matching Feasible?

Marcello D’Orazio

Italian National Institute of Statistics – Istat, Rome, Italy | Directorate for Methodology and Statistical Process Design

What is Statistical Matching?

Statistical Matching (SM) 'basic' case:

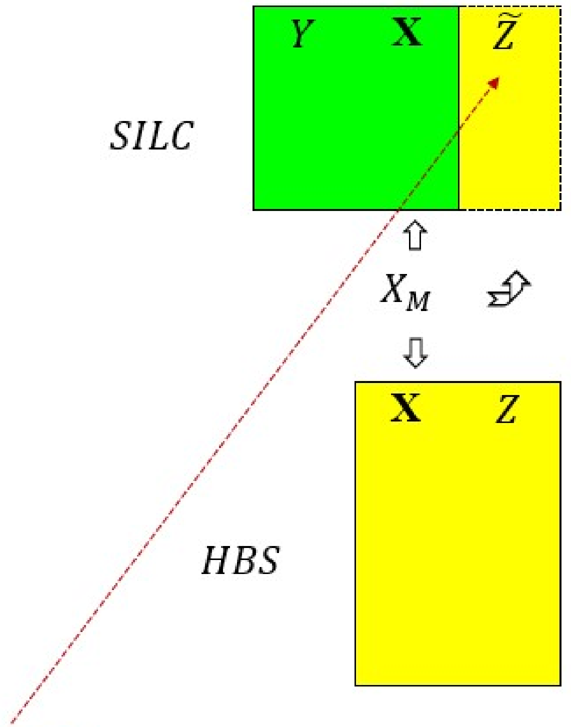


1. A and B are representative samples of the same population
2. $\boldsymbol{X}$ are common variables
   X_M ($X_M \subseteq \boldsymbol{X}$) are the **matching variables** (with same definition and support)
3. Y and Z are NOT jointly observed
4. probability=0 of finding the same unit in both A and B

GOAL: investigate **relationship between Y and Z** (ρ_{YZ} or $\beta_{Y|Z}$; contingency table $Y \times Z$, …)

Example of Statistical Matching

Example SM at Istat



1. *SILC* and *HBS* are representative samples of the Italian HHs:
   Y = HH income
   Z = HH total expenditures

2. ***X*** (many) common variables X_M ($X_M \subseteq \boldsymbol{X}$) the **matching variables**:
   - Macro-regions (3 cat)
   - No. of owned durable goods (5-9)
   - Ownership of the house (Yes/No)
   - HH Income from tax register
   - Rough approximate HH expenditures

3. SM method: **Nearest Neighbour hotdeck** and ***k*-NN**

Feed the Eurostat's experimental statistics on the joint distribution of income, consumption, and wealth (ICW)

Centralized SM exercise for many EU countries

https://ec.europa.eu/eurostat/web/experimental-statistics/income-consumption-wealth

GOAL: **impute HH total expenditures in SILC** and use this "fused" dataset to investigate relationship between HH income and HH consumption (e.g. propensity to consume)

Assumptions Underlying Statistical Matching

Major limiting assumption:

The relationship between Y and Z is fully explained by the matching variables X_M

In other words, Y and Z are independent conditional on X_M:

$$Y \perp Z |\, X_M$$

$$\rho_{YZ|X} = 0 \qquad \text{and} \qquad \rho_{YZ} = \rho_{YX}\,\rho_{XZ}$$

It's a very **strong assumption** seldom valid, and cannot be tested with available data

BUT…

Uncertainty in the Basic SM Setting

In the simple case of three continuous (X,Y,Z) variables following the multivariate Gaussian distribution:

Lower bound   Upper bound

$$\rho_{xy}\rho_{xz} - \sqrt{(1-\rho_{xy}^2)(1-\rho_{xz}^2)} \leq \rho_{yz} \leq \rho_{xy}\rho_{xz} + \sqrt{(1-\rho_{xy}^2)(1-\rho_{xz}^2)}$$

Midpoint

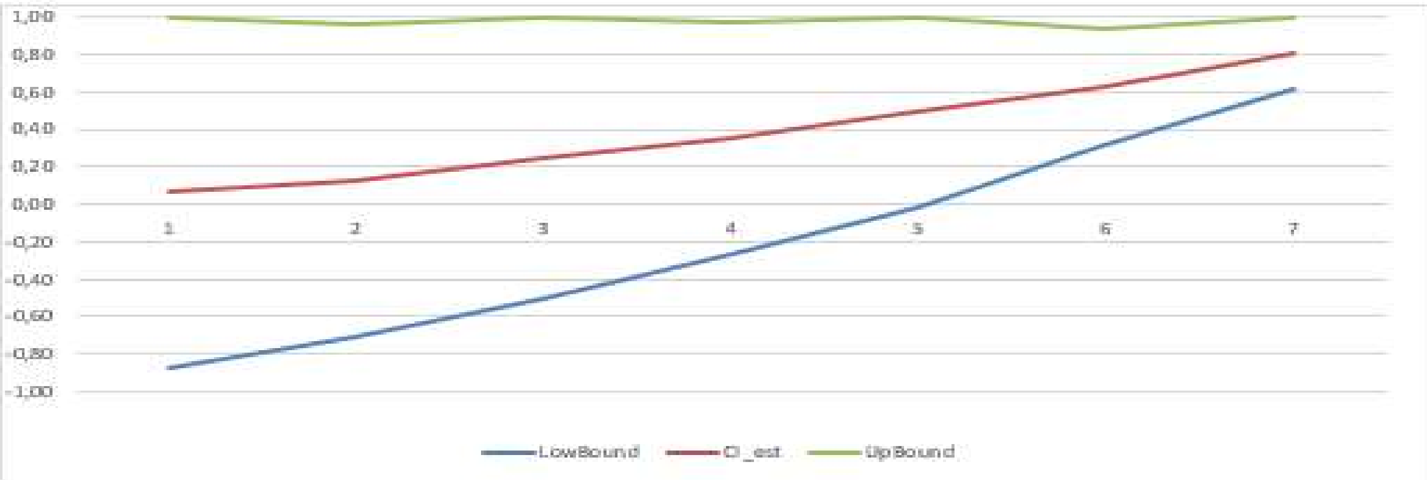
$$\rho_{YZ}^{(CI)} = \rho_{YX}\,\rho_{XZ}$$

(estimate under Conditional Independence)

Wide interval → High uncertainty → CI poor assumption → NOT worth doing matching

Uncertainty Bounds Width

$\hat{\rho}_{xy}$	$\hat{\rho}_{xz}$	LowBound	CI_est	UpBound
0.25	0.25	-0.88	0.06	1.00
0.25	0.50	-0.71	0.13	0.96
0.50	0.50	-0.50	0.25	1.00
0.50	0.70	-0.27	0.35	0.97
0.70	0.70	-0.02	0.49	1.00
0.70	0.90	0.32	0.63	0.94
0.90	0.90	0.62	0.81	1.00



Istat

Assess Uncertainty for Decision on Feacibility of SM

Work strategy: assess uncertainty before carrying out SM, i.e. obtain estimates of the bounds:

$$\left[\tilde{\rho}_{yz}^{(low)}, \tilde{\rho}_{yz}^{(up)}\right]$$

- If the interval is **wide**: give up doing SM
- If the interval is **narrow**: go on with SM

Main difficulties:

a) Many X_M variables; if all continuous and multivariate Gaussian distribution holds → Rodgers and de Vol (1982) give the expression to estimate bounds

b) Some X_M variables are categorical → replace with dummies → Gaussian?; Problem of bi-serial correlation

c) There are many X variables and X_M not identified

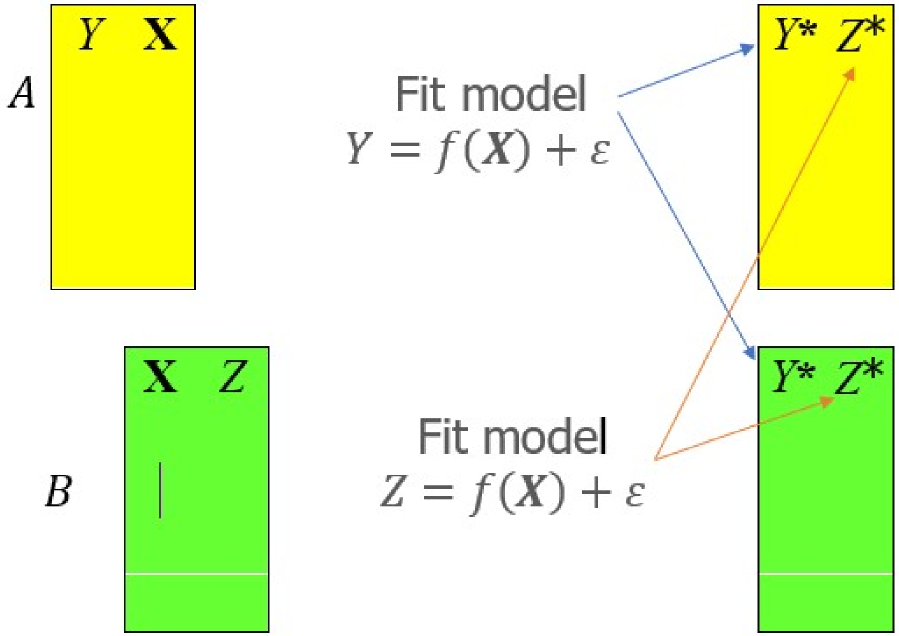
Approximate Estimation of Uncertainty Bounds

1.a) On A fit a «model» having Y as response;
And use fitted model to get predictions in both A and B

1.b) On B fit a «model» having Z as response;
And use fitted model to get predictions in both A and B

Implemented in **R**; available models:

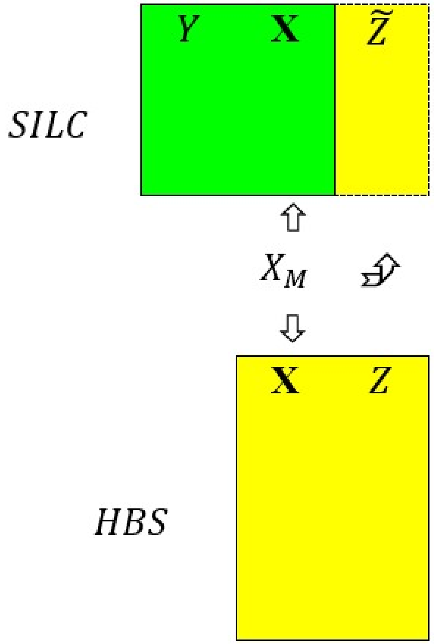
- Linear regression
- Robust linear regression
- Linear regression with *lasso* feature selection
- Random Forest



2) Use predictions of Y and predictions Z as Xs to assess the uncertainty

Example in Statistical Matching of SILC-HBS

data year 2016



Y = HH income (*log transf.*)
Z = HH total expenditures (*log transf.*)

matching variables:

- Macro-regions (3 cat)
- No. of owned durable goods (5-9)
- Ownership of the house (Yes/No)
- Income from tax register (*log transf.*)
- Rough approx. HH expenditures (*log transf.*)

	low	midp	up
w dummies	0.0764	0.3899	0.7035
lm pred	0.0297	0.3595	0.6893

Starting with a **larger** set of (potential) matching variables:

	low	midp	up
lm	0.0552	0.3665	0.6778
rob lm	0.0602	0.3733	0.6865
lasso	0.0427	0.3620	0.6812
rnd forest	0.3160	0.4217	0.5273

Istat

Categorical tearget variables

Y and Z are categorical → GOAL: contingency table crossing Y and Z

Same way of working but

uncertainty assessed using Frechet-Hoeffding property → estimation of the **expected values of the conditional bounds** (conditional to a categorical matching variable X) for each cell in the contingency table crossing Y and Z

$$\bar{p}^{(low)}_{Y=j,Z=k} \leq p_{Y=j,Z=k} \leq \bar{p}^{(up)}_{Y=j,Z=k}, \qquad j = 1, \dots, J; k = 1, \dots, K$$

Implemented in **R**; available models to predict Y and Z:

- Multinomial model
- Multinomial model with *lasso* feature selection
- Random Forest

What's Next

Add new "models"

Developed R code → new functions of the **StatMatch** package (D'Orazio, 2024)

Thank You

Marcello D’Orazio | marcello.dorazio@istat.it

Some References

Balestra, C. and F. Oehler (2023) "Measuring the joint distribution of household income, consumption and wealth at the micro level. Methodological issues and experimental results. Edition 2023", *Statistical Working Papers*, Eurostat/OECD, Luxembourg. https://ec.europa.eu/eurostat/web/products-statistical-working-papers/w/ks-tc-22-003

D'Orazio, M. (2019) "Statistical learning in official statistics: The case of statistical matching*". Statistical Journal of the IAOS*, 35(3), pp. 435-441. DOI: 10.3233/SJI-190518

D'Orazio, M. (2024) "StatMatch: Statistical Matching or Data Fusion". R package version 1.4.2. https://CRAN.R-project.org/package=StatMatch

D'Orazio, M and Di Zio, M and Scanu, M (2006) *Statistical Matching: Theory and Practice*. Wiley, Chichester

Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2014) "Statistical Matching of Income and Consumption expenditures". *International Journal of Economic Science*, Vol. III (No. 3), pp. 50-65.

Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016b) "The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics". DGINS - Conference of the Directors General of the National Statistical Institutes, 26-27 September 2016, Vienna.

Donatiello G., D'Orazio M., Frattarola D., Spaziani M. (2022) "The joint distribution of income and consumption in Italy: an in-depth analysis on statistical matching". *Rivista di Statistica Ufficiale - Review of Official Statistics*, N. 3/2022, pp. 77-109

Rodgers, W.L. and DeVol E.B. (1982) "An evaluation of statistical matching". *Report Submitted to the Income Survey Development Program*, Dept. of Health and Human Services, Institute for Social Research, University of Michigan