
Statistical Matching:

Methodological issues and practice with R-StatMatch

Vitoria-Gasteiz, 21-22 November 2013

INTRODUCTION TO STATISTICAL MATCHING

Marcello D'Orazio*

[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*



Techniques to integrate two or more data sources:

- 1) **record linkage**
- 2) **statistical matching**

Record linkage

Aims at identifying pairs of records, coming from different data files, which belong to the same entity on the base of the agreement between common indicators (Fortini, 2009)

the data sources are assumed to have observations in common!!!

- **exact record linkage**: perfect agreement between indicators which are assumed to be free of errors; typically used when it is available a Personal Identification Number (PIN)
- **probabilistic record linkage**: a PIN is available but it is not free of errors or, the PIN is not available and there are some common indicators which may present errors (name, surname, address, ...).
Uses probabilities for deciding when a given pair of records refers to the same unit (is a "match") or not

Statistical Matching (SM or **data fusion** or **synthetic matching**)

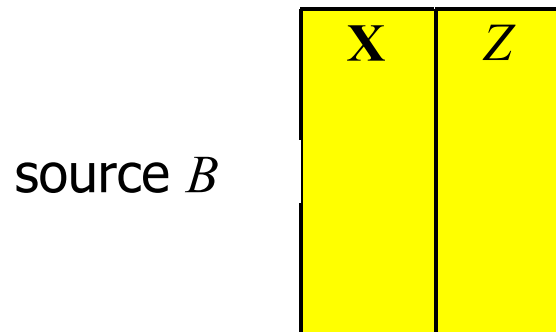
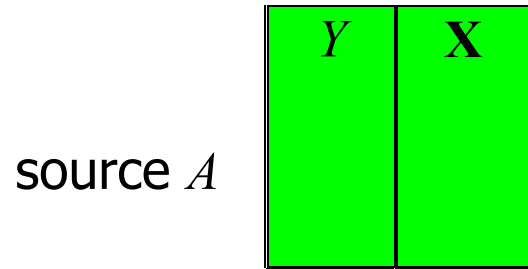
A series of statistical methods whose objective is the integration of two (or more) data sources (usually samples) referred to the same target population.

The objective is to study relationship among variable not jointly observed in a single data source

The data sources share a subset of variables (common variables) and, at the same time, each source observes distinctly other sub-sets of variables.

There is a negligible chance that data in different sources observe the same units (disjoint sets of units).

Traditional SM framework



1. X variables are in common
2. Y and Z are NOT jointly observed
3. The chance of observing the same unit in A and B is close to zero (usually A and B are samples)

The objective of SM is to study relationship between Y and Z or X , Y and Z

Objectives of SM:

- ✓ **micro**: derive a “synthetic” data-set with **X**, **Y** and **Z**

For instance A filled-in with Z :

Y	X	Z

or $A \cup B$ with Z filled in A and Y filled in B (**file concatenation**):

	Y	X	Z
A			
B			

“synthetic”: the records in the integrated source are not really observed

- ✓ **macro**: estimation of parameters; for instance:
 - correlation coef. (ρ_{YZ})
 - regression coefficient (β_{YZ})
 - a contingency table $Y \times Z$
 - ...

These methods do not require to integrate A and B at micro level

Objectives SM	Approaches		
	Parametric	Nonparametric	Mixed
Macro	Yes	Yes	No
Micro	Yes	Yes	Yes

Mixed has 2 steps:

- 1) Parametric: a model (e.g. regression) is considered and its parameters are estimated. It is used to impute the values of the missing variables
- 2) Nonparametric: imputed values provided by the model are used as input of a nonparametric SM method

Advantages of both parametric and nonparametric approaches are maintained:

- the model is parsimonious
- nonparametric techniques offer protection against model misspecification

A further complexity element: which type of inference?

- i) "**classic**": A and B are samples of i.i.d. obs., i.e. independent outcomes of a set of random variables (X, Y, Z) whose joint distribution follows a given (known or unknown) model (*model based inference*)
- ii) "**design based**": A and B are the results of complex sample surveys carried out on the same finite population. The values assumed by X , Y and Z for each unit in the finite population are viewed as fixed values (not outcomes of random variables).
The randomness is introduced by the probability criterion (sampling design) used to select the sample from the population (*design based inference*) (cf. Särndal et al. 1992).

Most of the SM methods developed under (i)

Relatively few methods available in case of (ii). In some case the methods developed for (i) are applied in situation (ii) ignoring the sampling design

Some References

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

Rässler S (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag.

Särndal, C. E., Swensson, B, and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer, New York.

One of the major problems in the application of SM techniques consisted in the lack of software

Problem tackled in different manners:

- a) Use/adapt existing software, typically programs developed for imputation of missing data
- b) Writing ad hoc code:
 - SAS codes by Moririaty (2001)
 - S-Plus code by Rassler (2002)
 - R code by D'Orazio et al. (2006)
 - ...

These reasons led develop two ad hoc software:

- SAMWIN (Sacco, 2008): standalone program for MS Windows; limited functionalities (mainly nonparametric micro approach)
- StatMatch (D'Orazio, 2013; first release in 2008), a free package for the R environment (R Core Team, 2013; <http://www.r-project.org/>) which implements different SM techniques and provides some other additional functions <http://CRAN.R-project.org/package=StatMatch>

StatMatch does not come as a standalone software: it is an open source additional package for the R environment.

The first release of StatMatch on the repositories of the Comprehensive R Archive Network (CRAN) dates back to 2008. This release was based on R codes provided in the Appendix of D'Orazio et al. (2006).

Since then a number of updates have been released.

Latest version is 1.2.0 released 2012-12-03.

A valuable contribution to the improvement of StatMatch is the work done within the Eurostat's ESSnet project on "Data Integration" (D'Orazio, 2011)

http://www.cros-portal.eu/sites/default/files//Statistical_Matching_with_StatMatch.pdf

Statmatch 1.2.0 provides a series of functions to perform:

- SM macro when dealing with variables distributed according to a multivariate normal distribution, function `mixed.mtc`
- SM micro methods:
 - random hot-deck: function `RANDwNND.hotdeck`
 - distance hot deck: function `NND.hotdeck`
 - rank hot deck: function `rankNND.hotdeck`
 - mixed method (based on predictive mean matching, starting from a multivariate normal distribution): function `mixed.mtc`

- SM of data from complex sample surveys (micro or macro) (mainly categorical variables). The Renssen's calibration based approach is considered:
 - function `harmonize.x`
 - function `comb.samples`
- Exploring uncertainty in SM when dealing with categorical variables:
 - function `Frechet.bounds.cat`
 - function `Fbwidht.by.x`

- Additional useful functions:
 - **comp.prop** comparison of the marginal distribution of the same variable(s)
 - **pw.assoc** association and PRE measures for categorical variables
 - **gower.dist**, **mahalanobis.dist**, **maximum.dist** to compute distances
 - **fact2dummy** substitutes a categorical variables with the corresponding dummies
 - **create.fused** physically creates the synthetic data source after the application of hot deck SM methods

Some References on StatMatch

D'Orazio, M (2011b) “Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment” R package vignette

http://www.cros-portal.eu/sites/default/files//Statistical_Matching_with_StatMatch.pdf

D'Orazio, M (2012) “StatMatch: Statistical Matching”, R package version 1.2.0

<http://CRAN.R-project.org/package=StatMatch>

Statistical Matching:

Methodological issues and practice with R-StatMatch

Vitoria-Gasteiz, 21-22 November 2013

STATISTICAL MATCHING UNDER THE CONDITIONAL INDEPENDENCE ASSUMPTION

Marcello D'Orazio*

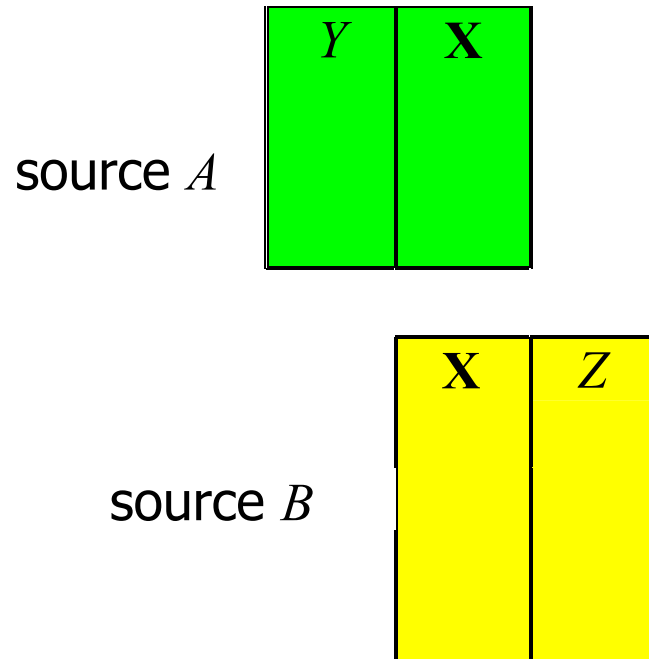
[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*



The Conditional Independence Assumption

Traditional SM framework



All the SM methods that use just the X variable to perform SM, implicitly assume that the relationship between Y and Z is completely explained by X

This means that Y and Z are independent once conditioning on the X variables:

$$f(x, y, z) = f(y|x)f(z|x)f(x)$$

i.e. **conditional independence assumption** (CIA)

When X , Y and Z are categorical variables the Conditional Independence Assumption means that:

$$\Pr(X = i, Y = j, Z = k) = \Pr(Y = j | X = i) \times \Pr(Z = k | X = i) \times \Pr(X = i)$$

$$i = 1, \dots, I; \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

Under this assumption the probabilities for the cells in the table $Y \times Z$ are obtained by summing over X categories:

$$\Pr(Y = j, Z = k) = \sum_{i=1}^I \Pr(Y = j | X = i) \times \Pr(Z = k | X = i) \times \Pr(X = i)$$

An example: X =Gender; Y =have a cat; Z =buying a specific cat's product

$X \times Y$ table estimated from A

	Cat	No cat	Tot.
M	10	38	48
F	32	20	52
Tot.	42	58	100

$X \times Z$ table estimated from B

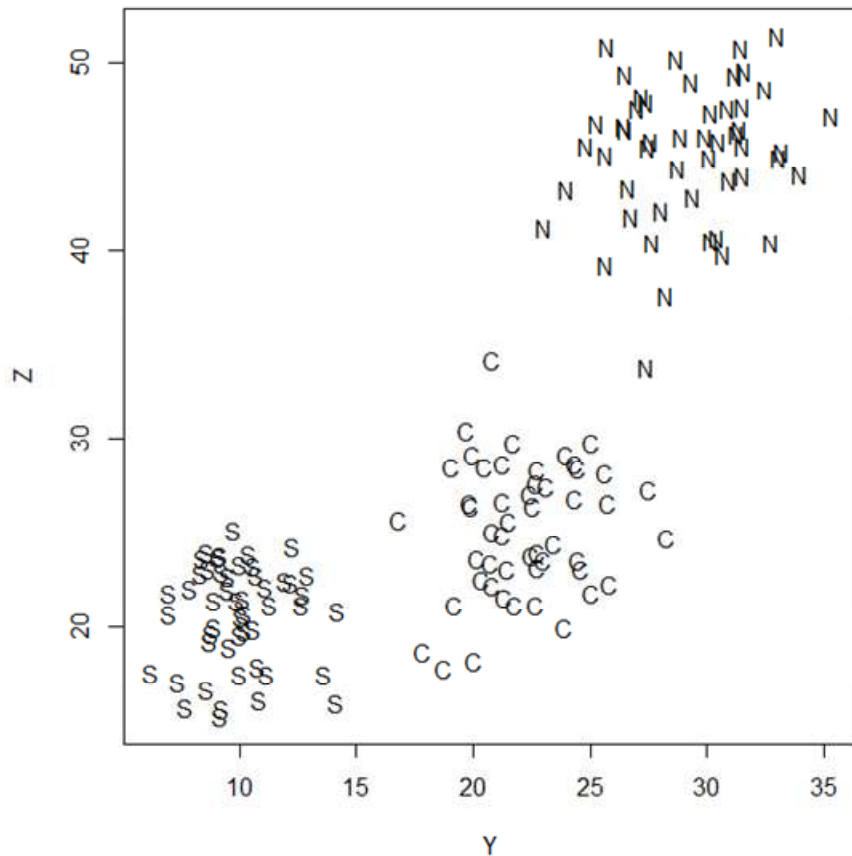
	Buy	Not buy	Tot
M	4	44	48
F	16	36	52
Tot.	20	80	100

Under the CIA:

$$\begin{aligned}\Pr(Y = 'no cat', Z = 'buy') &= \Pr(Y = 'no cat' | X = 'M') \times \Pr(Z = 'buy' | X = 'M') \times \Pr(X = 'M') + \\ &\quad + \Pr(Y = 'no cat' | X = 'F') \times \Pr(Z = 'buy' | X = 'F') \times \Pr(X = 'F') \\ &= 20/52 \times 16/52 \times 52/100 + 20/52 \times 16/52 \times 52/100 \\ &= 0.0615 + 0.0317 = 0.0932\end{aligned}$$

Example: X is region ('North', 'Center', 'South'); Y and Z are continuous

By matching A and B by considering the Region (X) it would result



$$\rho_{YZ} = 0.82$$

But:

$$\rho_{YZ|X=South} = 0.04$$

$$\rho_{YZ|X=Center} = 0.14$$

$$\rho_{YZ|X=North} = 0.18$$

Most of the method developed to perform SM assume the CIA

The CI assumption is very strong!

In most of the real world phenomena it does not hold

In such case results of SM derived under it are not valid.

Remark: in the traditional SM framework, when just A and B , are available
it is not possible to test explicitly whether the CIA holds or not

Parametric macro under the CIA: continuous variables

y_1	x_1
...	...
x_{n_A}	x_{n_A}

x_1	z_1
...	...
x_{n_B}	z_{n_B}

X , Y and Z are all continuous

Joint distr. of (X, Y, Z) is the trivariate normal distribution

μ_X and σ_X^2 could be estimated on A or on B or on $A \cup B$

μ_Y and σ_{XY} can be estimated on A

μ_Z and σ_{XZ} can be estimated on B

There are no data to estimate σ_{YZ}

The unique possibility to estimate σ_{YZ} is to assume the **CI of Y and Z given X**

$$\sigma_{YZ} = \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X^2}$$

$$\text{i.e. } \rho_{YZ} = \rho_{XY}\rho_{XZ} \quad (\rho_{YZ|X} = 0)$$

Assuming the CIA, how do I estimate in practice the various parameters?

Option 1) use the “sample counterparts” by exploiting all the available information (Kadane, 1978; Moriarity & Scheuren, 2001):

$$\hat{\mu}_X = \bar{x}_{A \cup B} = \frac{n_A \bar{x}_A + n_B \bar{x}_B}{n_A + n_B}; \quad \hat{\sigma}_X^2 = s_{X, A \cup B}^2 = \frac{(n_A - 1)s_{X, A}^2 + (n_B - 1)s_{X, B}^2}{n_A + n_B - 1}$$

$$\hat{\mu}_Y = \bar{y}_A = \frac{1}{n_A} \sum_{a=1}^{n_A} y_a; \quad \hat{\sigma}_Y^2 = s_{Y, A}^2 = \frac{1}{n_A - 1} \sum_{a=1}^{n_A} (y_a - \bar{y}_A)^2$$

$$\hat{\mu}_Z = \bar{z}_B = \frac{1}{n_B} \sum_{b=1}^{n_B} z_b; \quad \hat{\sigma}_Z^2 = s_{Z, B}^2 = \frac{1}{n_B - 1} \sum_{b=1}^{n_B} (z_b - \bar{z}_B)^2$$

$$\hat{\sigma}_{XY} = s_{XY, A} = \frac{1}{n_A - 1} \sum_{a=1}^{n_A} (x_a - \bar{x}_A)(y_a - \bar{y}_A)$$

$$\hat{\sigma}_{XZ} = s_{XZ, B} = \frac{1}{n_B - 1} \sum_{b=1}^{n_B} (x_b - \bar{x}_B)(z_b - \bar{z}_B),$$

And finally:

$$\hat{\sigma}_{YZ} = \frac{S_{XY,A} S_{XZ,B}}{S_{X,A \cup B}^2}$$

To sum up:

$$\hat{\Sigma} = \begin{pmatrix} S_{X,A \cup B}^2 & S_{XY,A} & S_{XZ,B} \\ S_{XY,A} & S_{Y,A}^2 & \hat{\sigma}_{YZ} \\ S_{XZ,B} & \hat{\sigma}_{YZ} & S_{Z,B}^2 \end{pmatrix}$$

This way of working is not wrong but my pose problems!!!

The matrix $\hat{\Sigma}$ May not be positive semidefinite!!!

Parametric macro under the CI assumption: continuous variables

	y	x	z	
1	0.2439127	0.91480717	NA	$\bar{x}_A = -0.5333$
2	-0.7085639	1.27087791	NA	
3	-0.4648018	-0.04236420	NA	
4	1.1225702	0.35148475	NA	$\bar{x}_B = -0.0510$
5	0.2617050	-1.17026594	NA	
6	0.0493663	-0.05189664	NA	$s^2_{X,A} = 0.7883$
7	0.2629912	-1.56363633	NA	
8	-0.6125185	0.15057822	NA	
9	NA	-0.19028070	0.01697295	
10	NA	0.26155390	-0.42745199	$s^2_{X,B} = 1.1537$
11	NA	-0.18158490	0.34473947	
12	NA	-2.14735610	-1.77710613	$s^2_{Y,A} = 0.3643$
13	NA	1.31195750	0.26088178	
14	NA	-1.39430890	-0.43961454	
15	NA	1.31225100	1.28938131	$s^2_{Z,B} = 0.6896$
16	NA	-0.17518800	-0.15255273	
17	NA	0.38074790	-0.15720819	$s_{XY,A} = 0.1111$
18	NA	0.31178000	0.88172958	$s_{XZ,B} = 0.7244$

$$\bar{x}_{A \cup B} = -0.1774$$

$$s^2_{X,A \cup B} = 0.9565$$

$$\hat{\sigma}_{YZ} = \frac{s_{XY,A} s_{XZ,B}}{s^2_{X,A \cup B}} = \frac{0.1111 \times 0.7244}{0.9565} = 0.0841$$

$$\hat{\Sigma} = \begin{pmatrix} s_{X,A \cup B}^2 & s_{XY,A} & s_{XZ,B} \\ s_{XY,A} & s_{Y,A}^2 & \hat{\sigma}_{YZ} \\ s_{XZ,B} & \hat{\sigma}_{YZ} & s_{Z,B}^2 \end{pmatrix} = \begin{pmatrix} 0.9565 & 0.1111 & 0.7244 \\ 0.1111 & 0.3643 & 0.0841 \\ 0.7244 & 0.0841 & 0.6896 \end{pmatrix}$$

By considering the determinant it comes out:

$$|\hat{\Sigma}| = \begin{vmatrix} 0.9565 & 0.1111 & 0.7244 \\ 0.1111 & 0.3643 & 0.0841 \\ 0.7244 & 0.0841 & 0.6896 \end{vmatrix} = 0.0474$$

i.e. the matrix $\hat{\Sigma}$ is positive semidefinite.

Option 2) use the ML estimation methods for partially observed data (Anderson, 1976):

$$\hat{\mu}_X = \bar{x}_{A \cup B}, \quad \hat{\sigma}_X^2 = s_{X, A \cup B}^2,$$

(Var. and cov. estimated considering sample size at denominator)

The parameters involving Y are estimated by considering the regression eq.

$$Y = \alpha_Y + \beta_{YX}X + \varepsilon_{Y|X}.$$

Provided that

$$\hat{\beta}_{YX} = s_{XY;A} / s_{X;A}^2, \quad \hat{\alpha}_Y = \bar{y}_A - \hat{\beta}_{YX} \bar{x}_A$$

It comes out:

$$\hat{\mu}_Y = \hat{\alpha}_Y + \hat{\beta}_{YX} \mu_X$$

and

$$\hat{\sigma}_Y^2 = s_{Y,A}^2 + \hat{\beta}_{YX} (s_{X, A \cup B}^2 - s_{X,A}^2), \quad \hat{\sigma}_{XY} = \hat{\beta}_{YX} s_{X, A \cup B}^2$$

The same happens as far Z is concerned.

Parametric macro under the CI assumption: continuous variables

	y	x	z
1	0.2439127	0.91480717	NA
2	-0.7085639	1.27087791	NA
3	-0.4648018	0.04236420	NA
4	1.1225702	0.35148475	NA
5	0.2617050	1.17026594	NA
6	0.0493663	0.05189664	NA
7	0.2629912	1.56363633	NA
8	-0.6125185	0.15057822	NA
9	NA	-0.19028070	0.01697295
10	NA	0.26155390	-0.42745199
11	NA	-0.18158490	0.34473947
12	NA	-2.14735610	-1.77710613
13	NA	1.31195750	0.26088178
14	NA	-1.39430890	-0.43961454
15	NA	1.31225100	1.28938131
16	NA	-0.17518800	-0.15255273
17	NA	0.38074790	-0.15720819
18	NA	0.31178000	0.88172958

```
> mx <- mean(AuB$x)
```

```
> mx
```

```
[1] -0.1773666
```

```
> s2.Ax <- var(A$x) * (nA-1) / nA
```

```
> s2.Ax
```

```
[1] 0.6897415
```

```
> s2.Bx <- var(B$x) * (nB-1) / nB
```

```
> s2.Bx
```

```
[1] 1.038287
```

```
> s2.x <- var(AuB$x) *
```

```
+ (nA+nB-1) / (nA+nB)
```

```
> s2.x
```

```
[1] 0.9033252
```

```
> reg.yx <- lm(y~x, data=A)
> cA <- coefficients(reg.yx)
> cA
(Intercept)          Byx
 0.06657144  0.14089719

> my <- cA[1]+cA[2]*mx
> my
 0.04158098

> s2.Ay <- var(A$y)*(nA-1)/nA
> s2.Ay
[1] 0.3187551

> s2.y <- s2.Ay+cA[2]*
+ (s2.x-s2.Ax)
> s2.y
0.3488484
```

```
> reg.zx <- lm(z~x, data=B)
> cB <- coefficients(reg.zx)
> cB
(Intercept)          Bzx
 0.01602623  0.62788598

> mz <- cB[1]+cB[2]*mx
> mz
-0.09533979

> s2.Bz <- var(B$z)*(nB-1)/nB
> s2.Bz
[1] 0.6206646

> s2.z <- s2.Bz+cB[2]*
+ (s2.x-s2.Bx)
> s2.z
0.5359237
```



```
> s.yx <- var(A$x,A$y) * (nA-1) / nA
> s.yx
[1] 0.09718264
> s.zx <- var(B$x,B$z) * (nB-1) / nB
> s.zx
[1] 0.6519261
>
> s.yx*s.zx/s2.x
[1] 0.07013631
```

$$|\hat{\Sigma}| = \begin{vmatrix} 0.9033 & 0.0972 & 0.6519 \\ 0.0972 & 0.3488 & 0.0701 \\ 0.6519 & 0.0701 & 0.5359 \end{vmatrix} = 0.0200$$

Function `mixed.mtc` in **StatMatch**

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
          method="ML", rho.yz=0, micro=FALSE,  
          constr.alg="Hungarian")
```

data.rec: data set A

data.don: data set B

match.vars: vector with the names of the matching variables

y.rec: the name of the Y variable (in A)

z.don: the name of the Z variable (in B)

method: method for estimating the parameters; `method="ML"` denotes Maximum Likelihood; `method="MS"` the one proposed by Moriarity and Scheuren (2001 e 2003).

micro: when FALSE (default) just parameters estimates are returned

Parametric macro under the CIA: categorical variables

A is a sample of n_A iid observations

B is a sample of n_B iid observations

The categorical variables X and Y are available in A :

$$X = [1, \dots, I], \quad Y = [1, \dots, J]$$

The categorical variables X and Z are available in B

$$X = [1, \dots, I], \quad Z = [1, \dots, K]$$

Parameters of interest:

$$\theta_{ijk} = \Pr(X = i, Y = j, Z = k), \quad 0 \leq \theta_{ijk} \leq 1, \quad \sum_{i,j,k} \theta_{ijk} = 1$$

$$i = 1, \dots, I; \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

Under the CI assumption:

$$\Pr(X = i, Y = j, Z = k) = \Pr(Y = j|X = i)\Pr(Z = k|X = i)\Pr(X = i)$$

$$\theta_{ijk} = \theta_{j|i} \theta_{k|i} \theta_{i++} = \frac{\theta_{ij+}}{\theta_{i++}} \frac{\theta_{i+k}}{\theta_{i++}} \theta_{i++} = \frac{\theta_{ij+} \theta_{i+k}}{\theta_{i++}}$$

$$i = 1, \dots, I; \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

Obviously the marginal table of $Y \times Z$ is obtained as:

$$\sum_i \theta_{ijk} = \sum_{i=1}^I \frac{\theta_{ij+} \theta_{i+k}}{\theta_{i++}}, \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

If

$n_{A,ij+}$ are the counts in the table $X \times Y$ derived from A

$n_{B,i+k}$ are the counts in the table $X \times Z$ derived from B

Using the ML estimation it comes out:

$$\hat{\theta}_{i++} = \frac{n_{A,i++} + n_{B,i++}}{n_A + n_B}, \quad i = 1, \dots, I$$

$$\hat{\theta}_{j|i} = \frac{n_{A,ij+}}{n_{A,i++}}, \quad i = 1, \dots, I, \quad j = 1, \dots, J$$

$$\hat{\theta}_{k|i} = \frac{n_{B,i+k}}{n_{B,i++}}, \quad i = 1, \dots, I, \quad k = 1, \dots, K$$

Function `Frechet.bounds.cat` in **StatMatch**

```
Frechet.bounds.cat(tab.x, tab.xy, tab.xz,  
                  print.f="tables", tol=0.0001)
```

tab.x: contingency table with the estimated joint distribution of the X variables

tab.xy: estimated contingency table $X \times Y$

tab.xz: estimated contingency table $X \times Z$

Comments on the parametric macro

- 1) specification of a model
- 2) estimation of the parameters: the method depends on the SM framework

Commonly encountered difficulties:

- too many variables -> discard less relevant ones
- variables of mixed type (categorical and continuous). Some transformations may help:
 - i) substitute a categorical variable with dummies and treat them as continuous; or
 - ii) categorize the continuous variables

Wrong results if the model is misspecified!

Parametric micro under the CIA: continuous variables

The final synthetic data source is obtained by:

- a) **concatenating** the two data set

$$S = A \cup B$$

a.1) the variable Z is filled in A

a.2) the variable Y is filled in B

- b) selecting one of the available data set as **recipient** (the other is the **donor**)

then, the missing variable is imputed in the recipient by using the data available in the donor.

E.g.: if A is the recipient, the synthetic data set will be A filled in with the variable originally missing in it, i.e. Z .

Usually the recipient is the smaller one, $n_{rec} < n_{don}$

Some parametric methods:

- conditional mean matching
- draws from the predicted distribution

Conditional mean matching: regression imputation

In the case of three continuous variables X , Y , and Z , **regression imputation** is considered (cf. D'Orazio, 2006, Sec. 2.2.1):

a) A is filled in with the predicted values:

$$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a, \quad a = 1, 2, \dots, n_A$$
$$\hat{\alpha}_Z = \bar{z}_B - \hat{\beta}_{ZX} \bar{x}_B, \quad \hat{\beta}_{ZX} = s_{XZ;B} / s_{X;B}^2$$

b) B is filled in with the predicted values:

$$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b, \quad b = 1, 2, \dots, n_B$$
$$\hat{\alpha}_Y = \bar{y}_A - \hat{\beta}_{YX} \bar{x}_A, \quad \hat{\beta}_{YX} = s_{XY;A} / s_{X;A}^2$$

c) file concatenation: $S = A \cup B$

file B

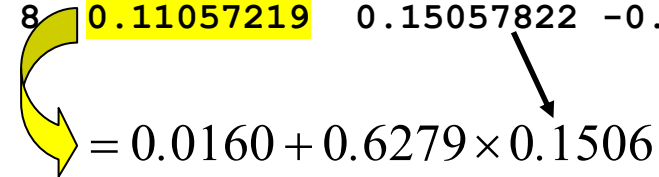
	x	z
1	-0.1902807	0.01697295
2	0.2615539	-0.42745199
3	-0.1815849	0.34473947
4	-2.1473561	-1.77710613
5	1.3119575	0.26088178
6	-1.3943089	-0.43961454
7	1.3122510	1.28938131
8	-0.1751880	-0.15255273
9	0.3807479	-0.15720819
10	0.3117800	0.88172958

$$\hat{z}_b = 0.0160 + 0.6279 \times x_b$$

$$s_Z^2 = 0.6896$$

file A imputed with Z

	z.imp	x	y
1	0.59042084	0.91480717	0.2439127
2	-0.78194019	-1.27087791	-0.7085639
3	-0.01057365	-0.04236420	-0.4648018
4	0.23671858	0.35148475	1.1225702
5	-0.71876735	-1.17026594	0.2617050
6	-0.01655894	-0.05189664	0.0493663
7	-0.96575909	-1.56363633	0.2629912
8	0.11057219	0.15057822	-0.6125185



$$= 0.0160 + 0.6279 \times 0.1506$$

$$s_{Z.imp}^2 = 0.3108$$

file A

	x	y
1	0.91480717	0.2439127
2	-1.27087791	-0.7085639
3	-0.04236420	-0.4648018
4	0.35148475	1.1225702
5	-1.17026594	0.2617050
6	-0.05189664	0.0493663
7	-1.56363633	0.2629912
8	0.15057822	-0.6125185

$$\hat{y}_a = 0.0666 + 0.1409 \times x_a$$

$$s_Y^2 = 0.3643$$

file B imputed with Y

	x	z	y.imp
1	-0.1902807	0.01697295	0.03976140
2	0.2615539	-0.42745199	0.10342362
3	-0.1815849	0.34473947	0.04098662
4	-2.1473561	-1.77710613	-0.23598499
5	1.3119575	0.26088178	0.25142253
6	-1.3943089	-0.43961454	-0.12988277
7	1.3122510	1.28938131	0.25146388
8	-0.1751880	-0.15255273	0.04188793
9	0.3807479	-0.15720819	0.12021772
10	0.3117800	0.88172958	0.11050034

$$= 0.0666 + 0.1409 \times 0.3118$$

$$s_{Y.imp}^2 = 0.0229$$

	x	y	z
A1	0.91480717	0.24391268	0.59042084
A2	-1.27087791	-0.70856393	-0.78194019
A3	-0.04236420	-0.46480182	-0.01057365
A4	0.35148475	1.12257016	0.23671858
A5	-1.17026594	0.26170497	-0.71876735
A6	-0.05189664	0.04936630	-0.01655894
A7	-1.56363633	0.26299123	-0.96575909
A8	0.15057822	-0.61251852	0.11057219
B1	-0.19028074	0.03976140	0.01697295
B2	0.26155388	0.10342362	-0.42745199
B3	-0.18158491	0.04098662	0.34473947
B4	-2.14735609	-0.23598499	-1.77710613
B5	1.31195749	0.25142253	0.26088178
B6	-1.39430891	-0.12988277	-0.43961454
B7	1.31225096	0.25146388	1.28938131
B8	-0.17518799	0.04188793	-0.15255273
B9	0.38074788	0.12021772	-0.15720819
B10	0.31178000	0.11050034	0.88172958

$$s_Y^2 = 0.3643 \quad \text{in } A$$

$$s_{Y.imp}^2 = 0.0229 \quad \text{imputed in } B$$

$$s_{Y,A \cup B}^2 = 0.1625$$

$$s_Z^2 = 0.6896 \quad \text{in } B$$

$$s_{Z.imp}^2 = 0.3108 \quad \text{imputed in } A$$

$$s_{Z,A \cup B}^2 = 0.5014$$

$$\rho_{XY,A} = 0.2073$$

$$\rho_{XZ,B} = 0.8121$$

Draws from the predicted distribution: stochastic regression imputation

Regression imputation provides values lying on the regression line and there is no variability around it.

To preserve variability it is suggested (cf. Little and Rubin, 2002) to add a random residual to each predicted value.

a) A is filled in with the values:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX}x_a + e_a, \quad a = 1, 2, \dots, n_A$$

with e_a a residual generated randomly from $N(0, \hat{\sigma}_{Z|X})$ being

$$\hat{\sigma}_{Z|X}^2 = s_{Z,B}^2 - \hat{\beta}_{ZX}^2 s_{X,B}^2,$$

b) B is filled in with the values:

$$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX}x_b + e_b, \quad b = 1, 2, \dots, n_B$$

with e_b a residual generated randomly from $N(0, \hat{\sigma}_{Y|X})$ being

$$\hat{\sigma}_{Y|X}^2 = s_{Y,A}^2 - \hat{\beta}_{YX}^2 s_{X,A}^2,$$

c) file concatenation: $S = A \cup B$

file B

	x	z
1	-0.1902807	0.01697295
2	0.2615539	-0.42745199
3	-0.1815849	0.34473947
4	-2.1473561	-1.77710613
5	1.3119575	0.26088178
6	-1.3943089	-0.43961454
7	1.3122510	1.28938131
8	-0.1751880	-0.15255273
9	0.3807479	-0.15720819
10	0.3117800	0.88172958

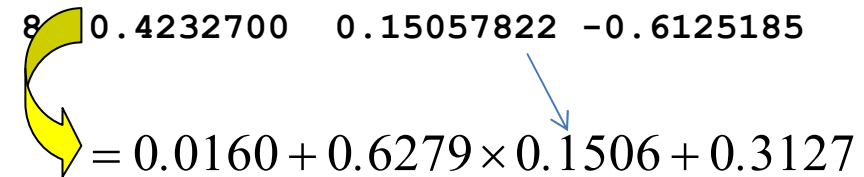
$$\hat{z}_b = 0.0160 + 0.6279 \times x_b$$

$$s_Z^2 = 0.6896$$

$$\hat{\sigma}_{Z|X}^2 = 0.5140$$

file A imputed with Z

	z.imp	x	y
1	1.5656758	0.91480717	0.2439127
2	-1.2210559	-1.27087791	-0.7085639
3	-0.4672171	-0.04236420	-0.4648018
4	0.8758671	0.35148475	1.1225702
5	-0.9275315	-1.17026594	0.2617050
6	0.2351697	-0.05189664	0.0493663
7	-0.3758059	-1.56363633	0.2629912
8	0.4232700	0.15057822	-0.6125185



$$= 0.0160 + 0.6279 \times 0.1506 + 0.3127$$

$$s_{Z.imp}^2 = 0.8803$$

file A

	x	y
1	0.91480717	0.2439127
2	-1.27087791	-0.7085639
3	-0.04236420	-0.4648018
4	0.35148475	1.1225702
5	-1.17026594	0.2617050
6	-0.05189664	0.0493663
7	-1.56363633	0.2629912
8	0.15057822	-0.6125185

$$\hat{y}_a = 0.0666 + 0.1409 \times x_a$$

$$s_Y^2 = 0.3643$$

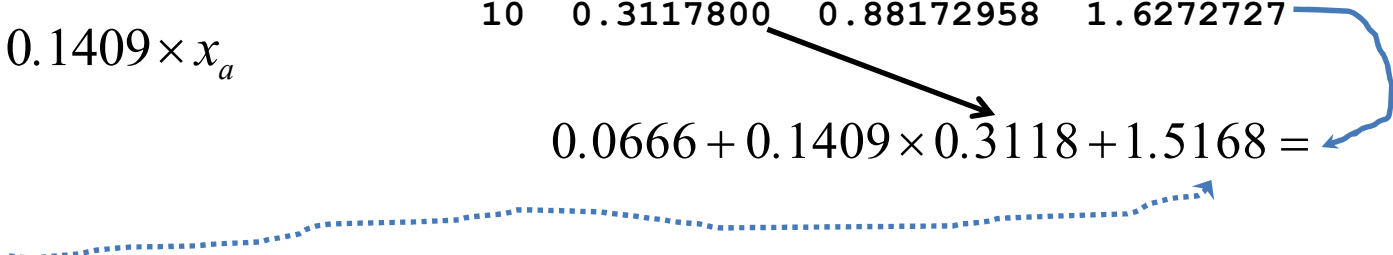
$$\hat{\sigma}_{Y|X}^2 = 0.6378$$

file B imputed with Y

	x	z	y.imp
1	-0.1902807	0.01697295	0.3551362
2	0.2615539	-0.42745199	0.3461535
3	-0.1815849	0.34473947	1.1919399
4	-2.1473561	-1.77710613	-0.9706588
5	1.3119575	0.26088178	-0.2781156
6	-1.3943089	-0.43961454	-0.7478982
7	1.3122510	1.28938131	0.3439470
8	-0.1751880	-0.15255273	-0.6048051
9	0.3807479	-0.15720819	0.9686518
10	0.3117800	0.88172958	1.6272727

$$0.0666 + 0.1409 \times 0.3118 + 1.5168 =$$

$$s_{Y.imp}^2 = 0.7575$$



Interesting findings in this context are those from Kadane (1978) and Moriarity & Scheuren (2001, 2003); D'Orazio et al. (2005)

In both the examples the imputed values are “artificial” values not really observed

Nonparametric micro under the CIA

Frequently encountered in SM applications

Do not require specifying in advance a model.

Strictly linked to the nonparametric imputation procedures applied in sample surveys to fill in missing values.

Typically the synthetic data set is obtained as a **recipient** data set filling in with the values of the missing variable selected from **donor** the data set

Usually the recipient is the smaller one ($n_{rec} < n_{don}$)

In the case of more recipients than donors ($n_{rec} > n_{don}$)

↳ some donor records will be selected more than once

↳ risk of altering the marginal distribution of the imputed variable

Let's assume:

A is the recipient dataset containing variables X and Y

B is the donor dataset containing variables X and Z

$$n_A \ll n_B$$

The synthetic data set (S) is obtained by filling in Z in A , the imputed values are values of Z really observed on units in B

Some commonly used methods (Singh et al., 1993):

- **random hot deck**
- **distance hot deck**
- **rank hot deck**

Nonparametric micro under the CIA: **random hot deck**

For each record in A , a donor record is randomly selected in B and the value of Z observed on this donor is imputed in A

Usually:

- 1) before the random selection the units in both the dataset are grouped into homogeneous strata (**donation classes**) according to the values of one or more categorical variables, X_G , chosen among the available common variables ones ($X_G \subseteq X$). For instance gender, region etc.
- 2) for a record in A in a given group (e.g. males), it is randomly chosen a donor in B in the same group (males in B)

Remark: a unit in B can be selected as a donor more than once

Such a way of working is equivalent to estimating the conditional distribution of Z given X_G (assumed categorical) and drawing and observation from it

Z continuous and X_G categorical, conditional distribution estimated via:

$$\text{Empirical cum. distr. } \hat{F}_{Z|X} = \frac{\sum_{b=1}^{n_B} I(z_b < z) I(x_{G,b} = i)}{\sum_{b=1}^{n_B} I(x_{G,b} = i)}$$

Z categorical and X_G categorical, conditional distribution estimated via:

$$\hat{\theta}_{k|i} = \frac{\sum_{b=1}^{n_B} I(z_b = k) I(x_{G,b} = i)}{\sum_{b=1}^{n_B} I(x_{G,b} = i)}$$

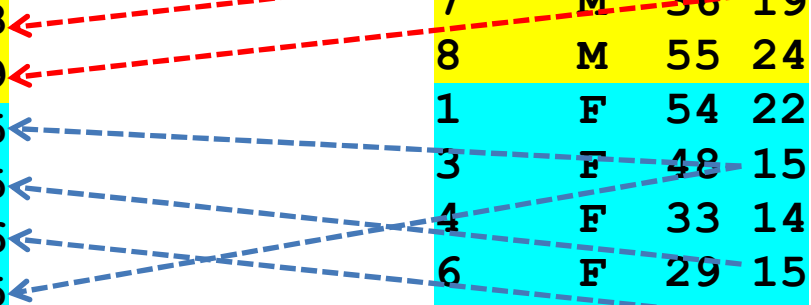
Donation classes formed using
"sex"

A

	sex	age	y	z	imp
2	M	35	19		13
3	M	41	47		19
1	F	27	22		15
4	F	61	41		15
5	F	52	17		26
6	F	39	26		15

B

	sex	age	z
2	M	21	17
5	M	63	13
7	M	36	19
8	M	55	24
1	F	54	22
3	F	48	15
4	F	33	14
6	F	29	15
9	F	50	26
10	F	27	18



When X_G is continuous, a donor can be chosen within the subset of donor units which are closest to the recipient unit according to a distance computed on X_G (cf. Andridge & Little, 2010).

For instance:

$$d_{ab}(x_{G,a}, x_{G,b}) \leq \delta, \quad b = 1, 2, \dots, n_B, \quad \delta > 0$$

Otherwise a donor is chosen within the subset of q donors units which are closest to the recipient units according to a distance on X_G

Donation classes formed using "sex" and $d_{ab}(x_{age,A}, x_{age,B}) \leq 5$

A

	sex	age	y	z.imp			sex	age	z
					[35-5, 35+5]	2	M	21	17
2	M	35	19	19		7	M	36	19
3	M	41	47	NA		8	M	55	24
1	F	27	22	15		5	M	63	13
6	F	39	26	NA	[27-5, 27+5]	10	F	27	18
5	F	52	17	26		6	F	29	15
4	F	61	41	NA	[52-5, 52+5]	4	F	33	14
						3	F	48	15
						9	F	50	26
						1	F	54	22

In this case the use of donation classes formed using "sex" and the random selection of the donor within the subset of the ones with an age of 5 years less or more than the recipient returns some empty subsets of donors

Function `RANDwNND.hotdeck` in **StatMatch**

```
RANDwNND.hotdeck(data.rec, data.don, match.vars=NULL,  
                  don.class=NULL, dist.fun="Manhattan",  
                  cut.don="rot", k=NULL, weight.don=NULL,  
                  ...)
```

data.rec: *recipient* data set (file *A*)

data.don: *donor* data set (file *B*)

match.vars: names of the matching variables (optional)

don.class: names of the variables (categorical) to create donation classes

dist.fun: how to compute distance on the matching variables ("Euclidean", "Manhattan", "Gower", etc.)

cut.don="k.dist": donors with distance from recipient $\leq k$

cut.don="exact": the k closest donors

Nonparametric micro under the CIA: **distance hot deck**

For each record in A , it is selected the closest donor record in B according to a distance computed on a suitable subset of the common variables X_M ($X_M \subseteq X$); the value of Z observed on the donor unit it is imputed in A

$$d_{ab}(\mathbf{x}_{M,a}, \mathbf{x}_{M,b}) = \min, \quad b = 1, 2, \dots, n_B$$

Many distances can be used to compute proximity between the units (see Appendix C in D'Orazio et al., 2006)

Before searching for donors it may be convenient and more efficient to divide units in A and B into donation classes according to the values of some categorical variables X_G ($X_G \subseteq X$).

For instance, for male recipient in A it will be searched the closest male donor in B

Donation classes formed using "sex", selection of the closest donor in terms of age

A

	sex	age	y	z.imp
2	M	35	19	19
3	M	41	47	19
1	F	27	22	18
6	F	39	26	14
5	F	52	17	22
4	F	61	41	22

	sex	age	z
2	M	21	17
7	M	36	19
8	M	55	24
5	M	63	13
10	F	27	18
6	F	29	15
4	F	33	14
3	F	48	15
9	F	50	26
1	F	54	22

Connections shown:

- Row 2 (M, 35) connects to Row 7 (M, 36)
- Row 3 (M, 41) connects to Row 7 (M, 36)
- Row 1 (F, 27) connects to Row 10 (F, 27)
- Row 6 (F, 39) connects to Row 6 (F, 29)
- Row 5 (F, 52) connects to Row 9 (F, 50)
- Row 4 (F, 61) connects to Row 1 (F, 54)

A crucial step is the choice of the **matching variables** X_M that have to be used in computing distances.

Usually, it corresponds to the subset of the X variables that at the same time are connected with Z and with Y :

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

X_Y : subset of the common variables ($X_Y \subseteq X$) that better explains Y

X_Z : subset of the common variables ($X_Z \subseteq X$) that better explains Z

Choosing too many matching variables may affect negatively the matching results: the marginal distribution of Z imputed in A may not reflect the one observed in B

Function `NND.hotdeck` in **StatMatch**

```
NND.hotdeck(data.rec, data.don, match.vars,  
            don.class=NULL, dist.fun="Manhattan",  
            constrained=FALSE, constr.alg="Hungarian",  
            ...)
```

data.rec: *recipient* data set (file *A*)

data.don: *donor* data set (file *B*)

match.vars: names of the matching variables (optional)

don.class: names of the variables (categorical) to create donation classes

dist.fun: how to compute distance on the matching variables ("Euclidean", "Manhattan", "Gower", etc.)

Remark: in distance hot deck a unit in B can be chosen more than once as a donor.

In order to avoid this, a **constrained distance hot deck** can be used: a donor can be used just once and the subset of the donors is selected in order to minimize the overall matching distance

Set `constrained=TRUE` in `NND.hotdeck`

In this case the selection of the donors requires the solution of an optimization problem (transportation problem).

Nonparametric micro under the CIA: rank hot deck

For each record in A , it is selected a closest donor record in B according to a distance computed on the percentage points of the empirical cumulative distribution function of the unique (continuous) common variable X_M being considered (Singh et al., 1993):

$$d_{ab}(\hat{F}(x_{M,a}), \hat{F}(x_{M,b})) = |\hat{F}(x_{M,a}) - \hat{F}(x_{M,b})|, \quad b = 1, 2, \dots, n_B$$

$$\hat{F}(x_{M,a}) = \frac{1}{n_A} \sum_{t=1}^{n_A} I(x_{M,t} \leq x_{M,a}), \quad a = 1, 2, \dots, n_A$$

$$\hat{F}(x_{M,b}) = \frac{1}{n_B} \sum_{t=1}^{n_B} I(x_{M,t} \leq x_{M,b}), \quad b = 1, 2, \dots, n_B$$

This transformation of the origin values produces values uniformly distributed in the interval $[0,1]$

Useful when the values of X_M can not be directly compared because of measurement errors which however do not affect the “position” of a unit in the whole distribution (cf. D’Orazio et al., 2006a, pp. 199-200)

Function `rankNND.hotdeck` in **StatMatch**

```
rankNND.hotdeck(data.rec, data.don, var.rec,  
                var.don=var.rec, don.class=NULL,  
                weight.rec=NULL, weight.don=NULL,  
                constrained=FALSE, constr.alg="Hungarian")
```

data.rec: *recipient* data set (file *A*)

data.don: *donor* data set (file *B*)

var.rec: name of the unique matching variables in *A*

var.don: name of the unique matching variables in *B*

don.class: names of the variables (categorical) to create donation classes

Micro objective: parametric vs. nonparametric

Parametric

- requires the specification of a model
- the imputed values are "artificial", not really observed
- unreliable results if the model is misspecified
- + the model is parsimonious

Nonparametric

- + does not require specifying a model
- handles easily complex situations (mixed type variables)
- requires a selection of a subset of the common variables (grouping, matching):
 - variables with low predictive power on the target variable may influence negatively the distances
- + the imputed values are really observed values

Micro objective: the mixed approach

Combines parametric and nonparametric:

- 1) a parametric model is specified and its parameters are estimated
- 2) nonparametric technique is used to derive a synthetic data source

Joins advantages of both the approaches; the nonparametric 2nd step offers protection against model misspecification in the 1st step.

When dealing with continuous X , Y and Z variables, various techniques based on predictive mean matching are available

An example:

Let X , Y and Z be all continuous

A is recipient and B is the donor

Step 1) A regression model is assumed as far as Z is concerned e.g.

$$Z = g(X; \theta)$$

Its parameters θ are estimated. The estimated model is used to derive “artificial” values, \tilde{z}_a , of Z in A (predicted values, or predicted plus a random error term)

Step 2) For each record in A its is selected the closest record in B according to a distance among artificial and truly observed values of Z , $d_{ab}(\tilde{z}_a, z_b)$.

The closest record in B donates to A the value of Z observed on it

There are variants of such a procedure (Rubin, 1986; Singh et al., 1993; Moriarity & Scheuren, 2001 and 2003)

It presents the following advantages:

- + offers protection against model misspecification
- + avoids the problem related to computing the distances by considering several common variables: variables with low predictive power on the target variable may influence negatively the distances

The procedures MM5 (see D'Orazio et al. 2006) for continuous variables:

Step 1) Two regression models are considered:

$$Y = \alpha_Y + \beta_{YX}X + \varepsilon_Y; \quad Z = \alpha_Z + \beta_{ZX}X + \varepsilon_Z$$

Their parameters are estimated and then

1.a) A is filled in with the values:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX}x_a + e_a, \quad a = 1, 2, \dots, n_A$$

being e_a a residual generated randomly from $N(0, \hat{\sigma}_{Z|X})$

1.b) B is filled in with the values:

$$\tilde{y}_b^{(B)} = \hat{y}_b^{(B)} + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX}x_b + e_b, \quad b = 1, 2, \dots, n_B$$

being e_b a residual generated randomly from $N(0, \hat{\sigma}_{Y|X})$

Step 2) For each record in A its is selected the closest record in B according to a distance

$$d_{ab}((y_a, \tilde{z}_a), (\tilde{y}_b, z_b)).$$

The Mahalanobis distance is considered and the matching is constrained.

In yellow the values provided by stochastic regression imputation in step 1)

	x	y	z
A1	0.91480717	0.24391268	0.59042084
A2	-1.27087791	-0.70856393	-0.78194019
A3	-0.04236420	-0.46480182	-0.01057365
A4	0.35148475	1.12257016	0.23671858
A5	-1.17026594	0.26170497	-0.71876735
A6	-0.05189664	0.04936630	-0.01655894
A7	-1.56363633	0.26299123	-0.96575909
A8	0.15057822	-0.61251852	0.11057219
B1	-0.19028074	0.03976140	0.01697295
B2	0.26155388	0.10342362	-0.42745199
B3	-0.18158491	0.04098662	0.34473947
B4	-2.14735609	-0.23598499	-1.77710613
B5	1.31195749	0.25142253	0.26088178
B6	-1.39430891	-0.12988277	-0.43961454
B7	1.31225096	0.25146388	1.28938131
B8	-0.17518799	0.04188793	-0.15255273
B9	0.38074788	0.12021772	-0.15720819
B10	0.31178000	0.11050034	0.88172958

$$d_{ab}((y_a, \tilde{z}_a), (\tilde{y}_b, z_b))$$

Function `mixed.mtc` in **StatMatch** with `micro=TRUE`

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
          method="ML", rho.yz=0, micro=FALSE,  
          constr.alg="Hungarian")
```

D'Orazio (2011) introduced a two step procedure for SM similar to the mixed one which is characterized by using a nonparametric approach also in the step 1):

Regression trees are used instead of linear regression models

Such a procedure is more flexible, and can handle both categorical and continuous predictors

Moreover it is possible to have also categorical response variables (classification trees)

Some References

- Andridge R.R., Little R.J.A. (2010) “A Review of Hot Deck Imputation for Survey Nonresponse”. *International Statistical Review*, **78**, 40–64.
- D’Orazio M. (2011) “Statistical matching through regression trees”. Paper presented at the SCo 2011 - 7th Conference on Statistical Computation and Complex Systems. Univ. Padova, September 19-21, 2011.
- D’Orazio M., Di Zio M., Scanu, M. (2005) “A comparison among different estimators of regression parameters on statistically matched files through an extensive simulation study”. *Technical Report Contributi 2005/10*, Istat, Roma.
- D’Orazio M., Di Zio M., and Scanu M. (2006) *Statistical Matching, Theory and Practice*. Wiley, New York.
- Kadane, J.B. (1978) “Some statistical problems in merging data files”, Reprinted in 2001 in *Journal of Official Statistics*, **17**, 423-433.
- Little R.J.A., Rubin D.B. (2002) *Statistical Analysis with Missing Data*, 2nd Edition. Wiley, New York.
- Moriarity C., Scheuren F. (2001) “Statistical matching: a paradigm for assessing the uncertainty in the procedure”. *Journal of Official Statistics*, **17**, 407–422.
- Moriarity C., Scheuren F. (2003). “A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation”, *Jour. of Business and Economic Statistics*, **21**, 65–73.
- Paass G. (1986) “Statistical match: evaluation of existing procedures and improvements by using additional information”. In *Microanalytic Simulation Models to Support Social and Financial Policy* (ed. Orcutt GH and Quinke H) Elsevier Science, pp. 401–422
- Rässler S (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag.
- Rubin D.B. (1986) “Statistical matching using file concatenation with adjusted weights and multiple imputations”. *Journal of Business and Economic Statistics*, **4**, 87-94
- Singh A.C., Mantel H., Kinack M., Rowe G. (1993). “Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption”. *Survey Methodology*, **19**, 59–79.

Statistical Matching:

Methodological issues and practice with R-StatMatch

Vitoria-Gasteiz, 21-22 November 2013

STATISTICAL MATCHING WITH AUXILIARY INFORMATION

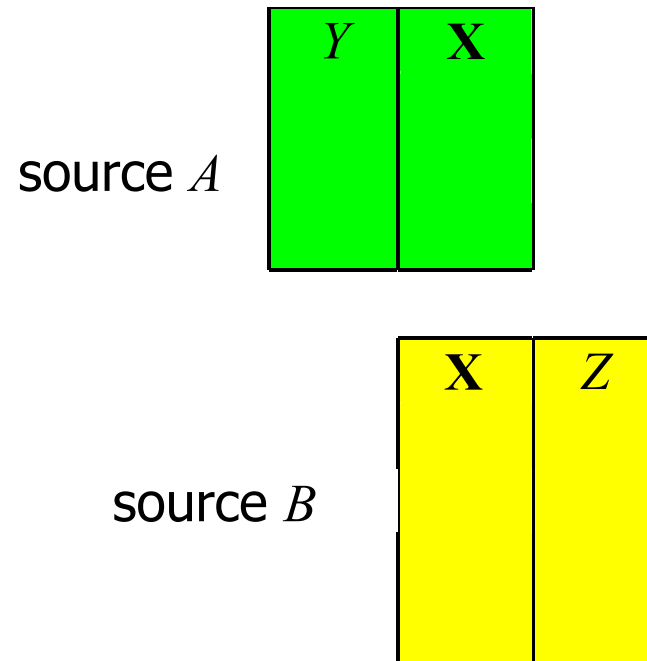
Marcello D'Orazio*

[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*



In the basic SM framework



It is not possible to test whether the CI assumption holds or not

When CI is not valid, the SM results obtained by assuming it are biased.

When the CI assumption does not hold, the SM results will be reliable if the SM application is based on some auxiliary information, e.g.:

- a) a third data source C where X , Y and Z or just Y and Z are jointly observed (e.g. small survey, past survey or census data, admin register, ...)
- b) estimates related to the parameters of $(Y, Z)|X$ or simply Y and Z (e.g. covariance σ_{YZ} or corr. coefficient ρ_{YZ} or partial corr. coef. $\rho_{YZ|X}$; a contingency table of $(Y \times Z)$; etc.)
- c) a priori knowledge of the investigated phenomenon which allow to identify some **logical constraints** on the parameters' values. For instance:

$$\Pr(\text{Age}=14 \text{ AND } \text{Edu.Level}=\text{"Univ. Degree"}) = 0$$

Parametric macro: use of an additional data source

As in the CI case the estimation is carried out by concatenating the available data sources:

$$A \cup B \cup C$$

and using this concatenated file to carry out estimation by exploiting all the available information.

An important issue to account for:

- X is available in C : some parameters are estimable in closed form while iterative methods (e.g. EM) are necessary to estimate parameters related to $(Y, Z)|X$ (information available just in C)
- X is NOT available in C : iterative methods (e.g. EM) are necessary to estimate all the parameters

For major details see D'Orazio et al. (2006, pp. 68-71).

Parametric macro: use of an external estimate

Is not simple to handle an estimation process with an estimate of a parameter which comes from external sources

The external estimate may be not compatible with the available data.

Example: Let consider the trivariate std normal distribution. The available data sources are A and B and it is available an estimate of the correlation coefficient ρ_{YZ}^* .

The correlation matrix must be positive semidefinite =>

$$\rho_{XY}\rho_{XZ} - \left[(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2) \right]^{1/2} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \left[(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2) \right]^{1/2}$$

Hence ρ_{YZ}^* is compatible with the available data if:

$$\hat{\rho}_{XY}\hat{\rho}_{XZ} - \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2} \leq \rho_{YZ}^* \leq \hat{\rho}_{XY}\hat{\rho}_{XZ} + \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2}$$

Example

$$\rho = \begin{pmatrix} 1 & 0.66 & 0.63 \\ 0.66 & 1 & ? \\ 0.63 & ? & 1 \end{pmatrix}$$

It comes out

$$\rho_{YZ}^{(low)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} - \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2} = -0.1676$$

$$\rho_{YZ}^{(up)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} + \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2} = +0.9992$$

$$\rho_{YZ}^{(CIA)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} = 0.66 \times 0.63 = 0.4158$$

$$\rho_{YZ}^{(up)} - \rho_{YZ}^{(CIA)} = 0.9992 - 0.4158 = 0.5834$$

$$\rho_{YZ}^{(CIA)} - \rho_{YZ}^{(low)} = 0.4158 - (-0.1676) = 0.5834$$

Function `mixed.mtc` in **StatMatch**

with

`micro=FALSE` and `method="MS"`

`rho.yz` is set equal to ρ_{YZ}^* , the guess for the correlation coefficient

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
          method="MS", rho.yz, micro=FALSE)
```

Multivariate normal distribution: known partial correlation coefficient

This type of information can easily be used in the tradition framework of SM, when the estimation is carried out using ML

In fact:

$$\tilde{\sigma}_{YZ|X} = \rho_{YZ|X}^* \sqrt{\hat{\sigma}_{Y|X}^2 \hat{\sigma}_{Z|X}^2}$$

$\rho_{YZ|X}^*$ will always be compatible

See Kadane (1978); Moriarity & Scheuren (2001, 2003) and D'Orazio (2005) for further details.

Function `mixed.mtc` in **StatMatch**

with

`micro=FALSE` and `method="ML"`

`rho.yz` is set equal to $\rho_{YZ|X}^*$, the guess for the partial correlation coefficient

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
          method="ML", rho.yz, micro=FALSE)
```

Parametric micro: use of an external information

Parameters are estimated accorded to the available data and taking into account for auxiliary information.

Then the data sources are concatenated and the missing values are imputed via:

- conditional mean matching

or

- draws from the predicted distribution

Remark: when the additional information is represented by an additional data source C (iid sample), then the synthetic dataset is derived as:

$$S = A \cup B \cup C$$

Nonparametric micro: use of an external information

These type of methods are applied when the auxiliary information is represented by an additional data source C

Singh et al. (1993) suggest various techniques based on hot deck imputation

Distance hot deck

Let

A be the recipient with X and Y

B be the donor with X and Z

C the additional data source of n_C observations

If C is a sample large enough and with reliable information then:

Impute Z in A using C as donor (B is NOT used!). Distance is

$d_{ac}((x_a, y_a), (x_c, y_c)),$ if C contains X, Y and Z .

$d_{ac}(y_a, y_c),$ if C contains just Y and Z

If Y values cannot be directly compared rank hot deck can be used

File A

	x1	x2	Y
	age	sex	work
1	32	1	1
2	17	1	1
3	41	1	1
4	24	2	2
5	67	2	2
6	48	2	1

File B

	x1	x2	z
	age	sex	netIncome
1	38	2	20257.67
2	56	1	23771.08
3	39	1	27189.86
4	27	1	17965.50
5	55	2	6515.51
6	32	1	14665.87
7	53	1	4445.99
8	20	1	1142.47
9	36	2	15800.87
10	30	1	20949.13

File C

	x1	x2	Y	Z
	age	sex	work	netIncome
1	21	1	2	0.00
2	41	1	1	19026.54
3	41	1	1	27596.34
4	40	2	1	8537.53
5	58	1	2	38794.79
6	35	1	1	25901.53
7	75	1	2	20870.23
8	35	2	2	22652.40
9	35	2	1	19389.03
10	74	2	2	22009.38

$$d_{ac}((x_{1a}, x_{2a}, y_a), (x_{1c}, x_{2c}, y_c))$$

Z imputed in A using donors selected in C

Distance computed using Xs and Y

File A

	x1	x2	Y
	age	sex	work
1	32	1	1
2	17	1	1
3	41	1	1
4	24	2	2
5	67	2	2
6	48	2	1

$d_{ac}(y_a, y_c)$

File C

	Y	Z
	work	netIncome
1	2	0.00
2	1	19026.54
3	1	27596.34
4	1	8537.53
5	2	38794.79
6	1	25901.53
7	2	20870.23
8	2	22652.40
9	1	19389.03
10	2	22009.38

File B

	x1	x2	Z
	age	sex	netIncome
1	38	2	20257.67
2	56	1	23771.08
3	39	1	27189.86
4	27	1	17965.50
5	55	2	6515.51
6	32	1	14665.87
7	53	1	4445.99
8	20	1	1142.47
9	36	2	15800.87
10	30	1	20949.13

Z imputed in A using donors selected in C

Distance computed using just Y

If C is a small sample with not fully reliable information then:

Step 1) impute Z in A using C as donor. Distance:

$$\begin{aligned} d_{ac}((x_a, y_a), (x_c, y_c)), & \quad \text{if } C \text{ contains } X, Y \text{ and } Z. \\ d_{ac}(y_a, y_c), & \quad \text{if } C \text{ contains just } Y \text{ and } Z \end{aligned}$$

Step 2) impute Z in A using B as a donor with distance computed using

$$d_{ab}((x_a, \tilde{z}_a), (x_b, z_b))$$

being \tilde{z}_a imputed in A after the step 1)

These two steps “robustify” the procedure taking into account that the information in C is not fully reliable (provides an idea of the relationship existing among the variables but the values of the variables are considered to be affected by measurement errors or outdated)

File A

	x1	x2	Y	Z.imp
	age	sex	work	netInc
1	32	1	1	
2	17	1	1	
3	41	1	1	
4	24	2	2	
5	67	2	2	
6	48	2	1	

File B

	x1	x2	Z
	age	sex	netIncome
1	38	2	20257.67
2	56	1	23771.08
3	39	1	27189.86
4	27	1	17965.50
5	55	2	6515.51
6	32	1	14665.87
7	53	1	4445.99
8	20	1	1142.47
9	36	2	15800.87
10	30	1	20949.13

File C

	x1	x2	Y	Z
	age	sex	work	netIncome
1	21	1	2	0.00
2	41	1	1	19026.54
3	41	1	1	27596.34
4	40	2	1	8537.53
5	58	1	2	38794.79
6	35	1	1	25901.53
7	75	1	2	20870.23
8	35	2	2	22652.40
9	35	2	1	19389.03
10	74	2	2	22009.38

$$d_{ac}((x_{1a}, x_{2a}, y_a), (x_{1c}, x_{2c}, y_c))$$

$$d_{ac}((x_{1a}, x_{2a}, \tilde{z}_a), (x_{1b}, x_{2b}, z_b))$$

Step 1) Z imputed in A using donors selected in C
Distance computed using Xs and Y

Step 2) Z imputed again in A using B as donor. Distance computed on Xs and Z

Mixed micro: use of an external estimate

Let consider the case of the trivariate normal distribution.
The procedure MM6 proposed and suggested in D'Orazio et al (2005) permits to plug in an external estimate $\rho_{YZ|X}^*$ in the step (1) when estimating the regression parameters.

$$\hat{\sigma}_{YZ|X} = \rho_{YZ|X}^* \sqrt{\hat{\sigma}_{Y|X}^2 \hat{\sigma}_{Z|X}^2} \quad (\rho_{YZ|X}^* = 0 \text{ in case of CIA}).$$

$$\hat{\sigma}_{YZ} = \hat{\sigma}_{YZ|X} + (\hat{\sigma}_{XY} \hat{\sigma}_{XZ}) / \hat{\sigma}_X^2$$

The step 2 remains unchanged.

Procedure MM6 (D'Orazio et al., 2006, pp. 87-90):

1) regression step. The following intermediate values are imputed

$$\text{in file } A: \quad \tilde{z}_a = \hat{\mu}_Z + \frac{\hat{\sigma}_{ZX|Y}}{\hat{\sigma}_{X|Y}^2} (x_a - \hat{\mu}_X) + \frac{\hat{\sigma}_{ZY|X}}{\hat{\sigma}_{Y|X}^2} (y_a - \hat{\mu}_Y) + e_a, \quad a = 1, \dots, n_A$$

$$\text{in file } B: \quad \tilde{y}_b = \hat{\mu}_Y + \frac{\hat{\sigma}_{YX|Z}}{\hat{\sigma}_{X|Z}^2} (x_b - \hat{\mu}_X) + \frac{\hat{\sigma}_{YZ|X}}{\hat{\sigma}_{Z|X}^2} (z_b - \hat{\mu}_Z) + e_b, \quad b = 1, \dots, n_B$$

e_a is a random residual drawn from $N(0, \hat{\sigma}_{Z|XY}^2)$;

e_b is a random residual drawn from $N(0, \hat{\sigma}_{Y|XZ}^2)$.

2) matching step. For each record a in file A , a live value z_{b^*} is imputed. Constrained distance hot deck (Mahalanobis distance)

$$d[(y_a, \tilde{z}_a), (\tilde{y}_b, z_b)]$$

A similar procedure has been proposed by Moriarity and Scheuren (2001, 2003) which is based on the:

- estimation of the parameters using their sample counterpart
- usage of auxiliary information in terms of known correlation coefficient ρ_{YZ}^*

Crucial to test whether ρ_{YZ}^* is compatible with the other estimates of the correlation coefficients derived from the available data.

In practice, given

$$\rho_{YZ}^{(low)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} - \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2} \quad \rho_{YZ}^{(up)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} + \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2}$$

If $\rho_{YZ}^{(low)} \leq \rho_{YZ}^* \leq \rho_{YZ}^{(up)}$, ρ_{YZ}^* is compatible and can be used in estimation step;

If $\rho_{YZ}^* > \rho_{YZ}^{(up)}$ or $\rho_{YZ}^* \leq \rho_{YZ}^{(low)}$ it is considered a new value ρ_{YZ}^{**} , close to ρ_{YZ}^* , such that $\rho_{YZ}^{(low)} \leq \rho_{YZ}^{**} \leq \rho_{YZ}^{(up)}$.

Function `mixed.mtc` in **StatMatch** with `micro=FALSE`

when `method="MS"` then `rho.yz` is set equal to ρ_{YZ}^* , the guess for the correlation coefficient

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
           method="ML", rho.yz, micro=TRUE,  
           constr.alg="Hungarian")
```

when `method="ML"` then `rho.yz` is set equal to $\rho_{YZ|X}^*$, the guess for the partial correlation coefficient

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
           method="ML", rho.yz, micro=TRUE,  
           constr.alg="Hungarian")
```

Singh et al. introduced various mixed methods to deal with the case of categorical variables (Cf. D'Orazio et al. 2006, Section 3.6.3 and 3.7)

Comment: importance of auxiliary information in SM

Approach to SM applications up to now:

Attempt of integrating surveys but SM was not planned in advance:

Does the CI hold?

If CI does not hold, is auxiliary information available?

If CI does not hold and auxiliary information is not available, the **SM should not be applied** (unless the **uncertainty** approach is considered)

New approach:

Start planning the survey with the idea of integrating them via SM:

- ☞ The X variables share the same definition/classification
- ☞ The auxiliary information is collected (e.g. some of the Z variables are collected in A or all the relevant Z s are collected in a random subsample of A)

Some References

- D'Orazio M., Di Zio M., Scanu, M. (2005) "A comparison among different estimators of regression parameters on statistically matched files trough an extensive simulation study". *Technical Report Contributi 2005/10*, Istat, Roma.
- D'Orazio M., Di Zio M., and Scanu M. (2006) *Statistical Matching, Theory and Practice*. Wiley, New York.
- Kadane, J.B. (1978) "Some statistical problems in merging data files", Reprinted in 2001 in *Journal of Official Statistics*, **17**, 423-433.
- Little R.J.A., Rubin D.B. (2002) *Statistical Analysis with Missing Data, 2nd Edition*. Wiley, New York.
- Moriarity C., Scheuren F. (2001) "Statistical matching: a paradigm for assessing the uncertainty in the procedure". *Journal of Official Statistics*, **17**, 407–422.
- Moriarity C., Scheuren F. (2003). "A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation", *Jour. of Business and Economic Statistics*, **21**, 65–73.
- Rässler S (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag.
- Rubin D.B. (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations". *Journal of Business and Economic Statistics*, **4**, 87-94
- Singh A.C., Mantel H., Kinack M., Rowe G. (1993). "Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption". *Survey Methodology*, **19**, 59–79.

Statistical Matching:

Methodological issues and practice with R-StatMatch

Vitoria-Gasteiz, 21-22 November 2013

THE UNCERTAINTY IN STATISTICAL MATCHING

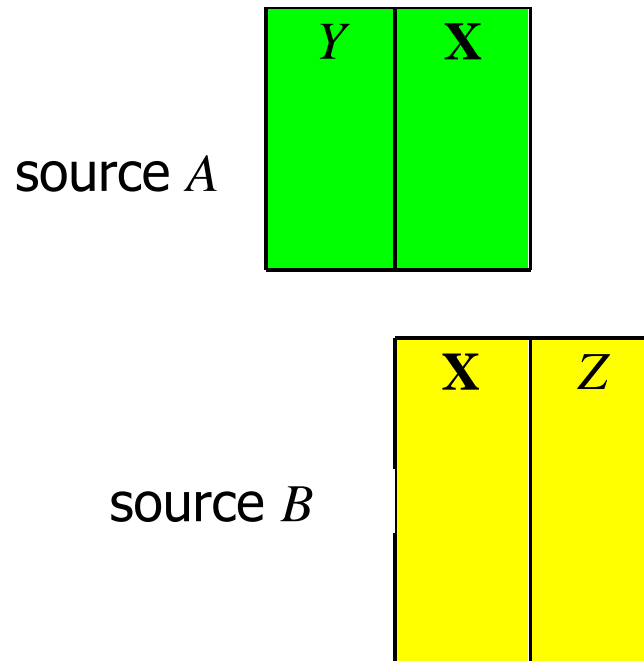
Marcello D'Orazio*

[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*



In the basic SM framework



It is possible to perform SM if:

- the CI of Y and Z given X is assumed
- or
- it is available some additional auxiliary information

In a parametric approach to SM with a macro objective it is possible to reason in different terms

In the case of categorical variables

The parameters of interest are:

$$\theta_{ijk} = \Pr(X = i, Y = j, Z = k), \quad i = 1, \dots, I; \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

These probabilities cannot be directly estimated because the three variables are not jointly observed

In practice there is a problem of identification of θ_{ijk} .

The data in A allow to estimate the contingency table $X \times Y$

$$\hat{\theta}_{ij+} = \hat{P}_{X=i,Y=j} = \hat{P}_{Y=j|X=i} \times \hat{P}_{X=i} \quad i = 1, \dots, I; \quad j = 1, \dots, J$$

The data in B allow to estimate the contingency table $X \times Z$

$$\hat{\theta}_{i+k} = \hat{P}_{X=i,Z=k} = \hat{P}_{Z=k|X=i} \times \hat{P}_{X=i} \quad i = 1, \dots, I; \quad k = 1, \dots, K$$

Therefore among all the distributions which satisfy the basic constraints:

$$0 \leq \theta_{ijk} \leq 1, \quad i = 1, \dots, I; \quad j = 1, \dots, J, \quad k = 1, \dots, K$$

$$\sum_{i,j,k} \theta_{ijk} = 1$$

Just those satisfying these further constraints are admissible:

$$\sum_k \theta_{ijk} = \hat{\theta}_{ij+}, \quad i = 1, \dots, I; \quad j = 1, \dots, J,$$

$$\sum_j \theta_{ijk} = \hat{\theta}_{i+k}, \quad i = 1, \dots, I; \quad k = 1, \dots, K$$

The probabilities can be estimated via the following formula

$$\hat{P}_{X=i} = \frac{\sum_{k=1}^n w_k I(x_k = i)}{\sum_{k=1}^n w_k}$$

$I(\bullet) = 1$ if condition within parenthesis is satisfied and 0 otherwise.

w_k survey weight ($w_k = 1$ in Simple Random Sampling)

The theory permits to derive some interesting conclusions for the probabilities θ_{+jk} of the marginal distribution $Y \times Z$

Frechét bounds for joint distribution $H(y, z)$:

$$\max\{0; F(y) + G(z) - 1\} \leq H(y, z) \leq \min\{F(y); G(z)\}$$

Hence in the case of categorical variables:

$$\max\{0; P_{Y=j} + P_{Z=k} - 1\} \leq P_{Y=j, Z=k} \leq \min\{P_{Y=j}; P_{Z=k}\}$$

$$j = 1, \dots, J, \quad k = 1, \dots, K$$

Identification problem in SM: categorical variables

Table 3. Distribution of Professional Status vs Age in file A

Age	Professional Status			Total
	M	E	W	
1	–	–	9	9
2	–	5	17	22
3	179	443	486	1108
4	6	1	2	9
Tot.	185	449	514	1148

Table 4. Distribution of Educational Level vs Age in file B

Age	Educational Level				Total
	C	V	S	D	
1	6	0	–	–	6
2	14	6	13	–	33
3	387	102	464	158	1111
4	10	0	3	2	15
Tot.	417	108	480	160	1165

	Y	Z	low.u	CIA	up.u
1	M	C	0	0.05859895	0.16114983
2	E	C	0	0.13667241	0.35793991
3	W	C	0	0.16239762	0.35793991
4	M	V	0	0.01422922	0.09270386
5	E	V	0	0.03619792	0.09270386
6	W	V	0	0.04197432	0.09270386
7	M	S	0	0.06611248	0.16114983
8	E	S	0	0.16255474	0.39111498
9	W	S	0	0.18344421	0.41201717
10	M	D	0	0.02296366	0.13733906
11	E	D	0	0.05470296	0.13733906
12	W	D	0	0.06015152	0.13733906

Remark: with categorical variables the estimate under the CI assumption

$$P_{Y=j,Z=k}^{(CI)} = \sum_{i=1}^I P_{Y=j|X_D=i} \times P_{Z=k|X_D=i} P_{X_D=i}$$

is always included in the uncertainty bounds

$$P_{j,k}^{(low)} \leq P_{Y=j,Z=k}^{(CI)} \leq P_{j,k}^{(up)}$$

but it is NOT the central point of the bound

Let assume that X_D is the variable obtained by the crossproduct of the chosen X variables

by conditioning on X_D it comes out

$$P_{j,k}^{(low)} \leq P_{Y=j,Z=k} \leq P_{j,k}^{(up)}$$

$$P_{j,k}^{(low)} = \sum_{i=1}^I P_{X_D=i} \max \left\{ 0; P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1 \right\}$$

$$P_{j,k}^{(up)} = \sum_{i=1}^I P_{X_D=i} \min \left\{ P_{Y=j|X_D=i}; P_{Z=k|X_D=i} \right\}$$

with $j = 1, \dots, J, \quad k = 1, \dots, K$

Example (previous data)

	Y	Z	low.u	low.cx	CIA	up.cx	up.u
1	M	C	0	0.003458712	0.05859895	0.16190430	0.16114983
2	E	C	0	0.000000000	0.13667241	0.34073569	0.35793991
3	W	C	0	0.011168756	0.16239762	0.35305736	0.35793991
4	M	V	0	0.000000000	0.01422922	0.08807807	0.09270386
5	E	V	0	0.000000000	0.03619792	0.09240146	0.09270386
6	W	V	0	0.000000000	0.04197432	0.09240146	0.09270386
7	M	S	0	0.000000000	0.06611248	0.15706211	0.16114983
8	E	S	0	0.000000000	0.16255474	0.39012803	0.39111498
9	W	S	0	0.003963107	0.18344421	0.41211143	0.41201717
10	M	D	0	0.000000000	0.02296366	0.13781814	0.13733906
11	E	D	0	0.000000000	0.05470296	0.13758756	0.13733906
12	W	D	0	0.000000000	0.06015152	0.13781814	0.13733906

The estimation of the bounds permits to estimate the **overall uncertainty**

Two alternatives:

a) the **average width of the bounds**:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} \left[\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right]$$

b) the **weighted average width** suggested by Conti *et al.* (2012)

$$\hat{\Delta} = \sum_{i,j,k} \left(\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right) \times \hat{P}_{Y=j|X_D=i} \times \hat{P}_{Z=k|X_D=i} \times \hat{P}_{X_D=i}$$

The probabilities can be estimated via the following formula

$$\hat{P}_{X=i} = \frac{\sum_{k=1}^n w_k I(x_k = i)}{\sum_{k=1}^n w_k}$$

w_k survey weight ($w_k = 1$ in Simple random Sampling)

Example: estimated uncertainty with previous data

	Av. Width (\bar{d})	Weighted av. Width ($\hat{\Delta}$)
Not conditioning on X	0.2109534	-
conditioning on X	0.2068761	0.2894636

Conti et al. (2013) studied the uncertainty when dealing with categorical ordered variables

Function `Frechet.bounds.cat` in **StatMatch**

```
Frechet.bounds.cat(tab.x, tab.xy, tab.xz,  
                   print.f="tables", tol=0.0001)
```

`tab.x`: contingency table with the distribution of the X variables

`tab.xy`: contingency table with the joint distribution $X \times Y$

`tab.xz`: contingency table with the joint distribution $X \times Z$

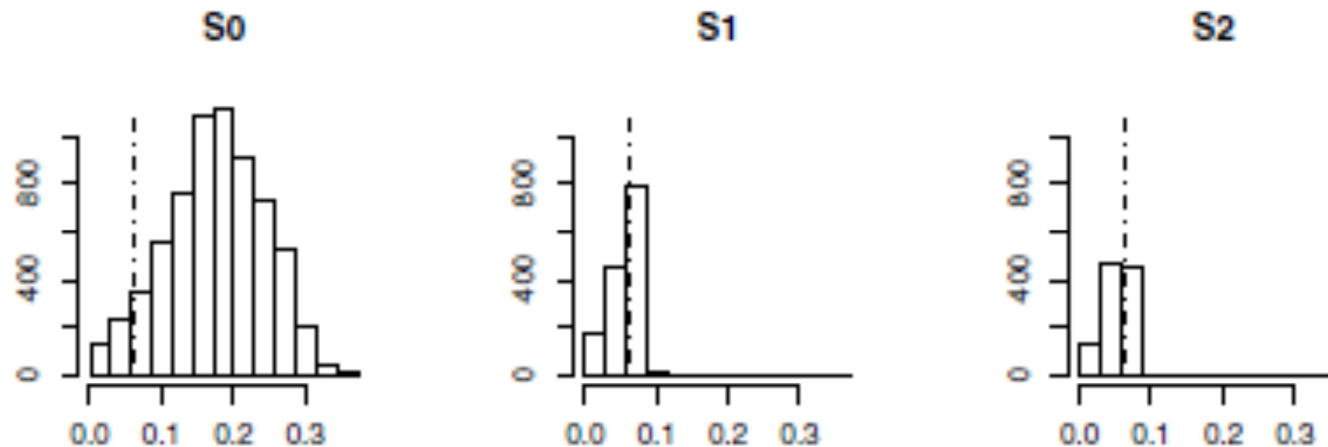
D'Orazio et al. (2006b) investigated the problem of uncertainty in the presence of categorical variables and additional information in terms of

- structural zeros (events that can not happen)

$$\theta_{ijk} = 0 \text{ for some } i, j, k$$

- inequality constraints concerning couples of probabilities: $\theta_{ijk} \leq \theta_{i'j'k'}$

Example: joint distr. of X =Age (in classes), Y =Edu. Level, Z =Prof.Status



Uncertainty bounds concerning
 $X=3$ (23-64 years); Y ="Secondary school"; Z ="Worker"

S0: unrestricted bounds

S1: 15-17 years old people cannot have a unv. Degree;
"Managers" should have at least a secondary school degree

S2: S1 AND $\theta_{X=3,Y='D',Z='M'} \geq \theta_{X=3,Y='D',Z='E'}$

Let consider the simple case of continuous X , Y and Z distributed according to the trivariate normal

The interest parameter is the covariance matrix or the correlation matrix:

$$\rho = \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix}$$

The unique parameter which cannot be directly estimated ρ_{YZ}
(CIA $\Rightarrow \rho_{YZ} = \rho_{XY}\rho_{XZ}$)

In practice there is a problem of identification of ρ_{YZ} .

In absence of external estimates of ρ_{YZ} (or $\rho_{YZ|X}$) and without assuming the CI, the unique conclusion is:

$$\rho_{XY}\rho_{XZ} - \left[(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2) \right]^{1/2} \leq \rho_{YZ} \leq \rho_{XY}\rho_{XZ} + \left[(1 - \rho_{XY}^2)(1 - \rho_{XZ}^2) \right]^{1/2}$$

due to the fact that the correlation matrix must be positive semidefinite (determinant greater or equal to zero: $|\boldsymbol{\rho}| \geq 0$)

By considering the estimates of the others corr. coefficients it comes out:

$$\hat{\rho}_{XY}\hat{\rho}_{XZ} - \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2} \leq \rho_{YZ} \leq \hat{\rho}_{XY}\hat{\rho}_{XZ} + \left[(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2) \right]^{1/2}$$

Remark: the estimate of ρ_{YZ} under the CIA ($\rho_{YZ} = \rho_{XY}\rho_{XZ}$) is the central point of the interval

In practice, this approach permits to explore the **uncertainty** on ρ due to SM framework (no CIA and no auxiliary information)

For further details see:

Kadane (1978), Rubin (1986), Moriarity and Scheuren (2001, 2003)

Rassler (2002), D'Orazio Di Zio and Scanu (2006a, b)

Remark: Manski's (1995) monograph deals with the identification problem for missing data

Example

$$\rho = \begin{pmatrix} 1 & 0.66 & 0.63 \\ 0.66 & 1 & ? \\ 0.63 & ? & 1 \end{pmatrix}$$

It comes out

$$\rho_{YZ}^{(low)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} - [(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2)]^{1/2} = -0.1676$$

$$\rho_{YZ}^{(up)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} + [(1 - \hat{\rho}_{XY}^2)(1 - \hat{\rho}_{XZ}^2)]^{1/2} = +0.9992$$

$$\rho_{YZ}^{(CIA)} = \hat{\rho}_{XY}\hat{\rho}_{XZ} = 0.66 \times 0.63 = 0.4158$$

$$\rho_{YZ}^{(up)} - \rho_{YZ}^{(CIA)} = 0.9992 - 0.4158 = 0.5834$$

$$\rho_{YZ}^{(CIA)} - \rho_{YZ}^{(low)} = 0.4158 - (-0.1676) = 0.5834$$

Function `mixed.mtc` in **StatMatch**

with `method="MS"` and `micro=FALSE`

```
mixed.mtc(data.rec, data.don, match.vars, y.rec, z.don,  
          method="MS", rho.yz=0, micro=FALSE)
```

provides uncertainty bounds of in the presence of one or more X variables

Remark

The CIA estimate of the interest parameter is always included in the uncertainty bounds.

When uncertainty decreases (shorter intervals) increases the trust on the CI assumption

In this sense, the exploration of the uncertainty can be viewed as kind of test on the CI assumption.

Some References

- D'Orazio, M., Di Zio, M. e Scanu, M. (2006a), “Statistical Matching for Categorical Data: displaying uncertainty and using logical constraints”, *Journal of Official Statistics*, vol. 22, n. 1, pp. 1-12.
- D'Orazio M., Di Zio M., and Scanu M. (2006b) *Statistical Matching, Theory and Practice*. Wiley, New York.
- Kadane, J.B. (1978) “Some statistical problems in merging data files”, Reprinted in 2001 in *Journal of Official Statistics*, **17**, 423-433.
- Manski, C. F. (1995), *Identification Problems in the Social Sciences*. Cambridge, Massachusetts: Harvard University Press.
- Moriarity C., Scheuren F. (2001) “Statistical matching: a paradigm for assessing the uncertainty in the procedure”. *Journal of Official Statistics*, **17**, 407–422.
- Moriarity C., Scheuren F. (2003). “A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputation”, *Jour. of Business and Economic Statistics*, **21**, 65–73.
- Rässler S (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag.
- Rubin D.B. (1986) “Statistical matching using file concatenation with adjusted weights and multiple imputations”. *Journal of Business and Economic Statistics*, **4**, 87-94
- Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall

Statistical Matching:

Methodological issues and practice with R-StatMatch

Vitoria-Gasteiz, 21-22 November 2013

STATISTICAL MATCHING OF DATA FROM COMPLEX SAMPLE SURVEYS

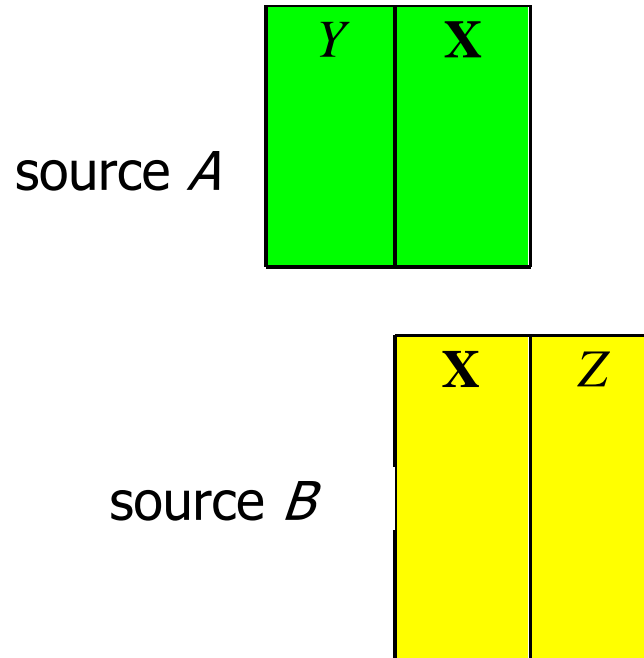
Marcello D'Orazio*

[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*



In the basic SM framework



Most of the SM methods introduced in literature assume that A and B are iid samples from infinite population.

The only source of variation is that from of the model generating the data,
(**model based inference**)

In sample surveys the inference is carried out in a different framework

The target population U is finite and consists of N units ($N < \infty$)

(y_1, y_2, \dots, y_N) is the set of the unknown values assumed by Y for each unit in the population. Such values are **fixed values** and NOT outcomes of a random variable.

The only randomization mechanism is the probability criterion used to select the sample s from U .

The selection criterion assigns to each unit U of a non-null probability π_a ($0 < \pi_a \leq 1$) of being included into the sample.

The inclusion probabilities π_a are the base of the inference.

If the objective of the inference is the total amount of Y in U

$$t_y = \sum_{a=1}^N y_a$$

It can be estimated by the Horvitz-Thompson estimator (1952) (HT):

$$\hat{t}_{\pi y} = \sum_{a=1}^n \frac{y_a}{\pi_a} = \sum_{a=1}^n d_a y_a \quad (n \text{ is the sample size})$$

$d_a = 1/\pi_a$: [design weight](#) or [base weight](#) or [direct weight](#) of the unit a

The HT estimator is an unbiased estimator of t_y wrt to the sampling design (design unbiased) whatever sampling design has been chosen.

Remark:

$$\hat{t}_{\pi 1} = \sum_{a=1}^n \frac{1}{\pi_a} = \sum_{a=1}^n d_a = \hat{N}$$

It is possible to ignore the probability distribution induced by the sampling design and treat the data as an iid sample (inferences are carried out in the model based framework) if:

- i) **sampling design is noninformative**: the sampling design does not depend on the (y_1, y_2, \dots, y_N) values, conditioned to the design variables (strata, etc.)
- ii) the sampling design and the design variables are known and are not related to the parameters which are objective of the inference

Ignoring the sampling design when these conditions do not hold may lead to unreliable results.

In general, when dealing with data from complex sample surveys it is common that the sampling design is not fully known or just a limited number of the design variable are available.

In practice when dealing with micro data from complex sample surveys it is common that:

- the observations in a dataset are less than the planned sample size mainly due to:
 - unit nonresponse (non contacts, refusals, etc.)
 - discarding of ineligible units, maybe because of errors in the sampling frame (units no more belonging to the population)
- some units may present missing values or some of the values are imputed values
- some of the observed values are affected by measurement errors (not detected by checks)
- the final weights w_a associated to the available units are the direct weights corrected to compensate for unit nonresponse, frame undercoverage, to satisfy known given population totals concerning some auxiliary variables
- the design variables are partially available due to the risk of disclosure.

In such a case, in the SM framework the available data set are characterized by two sets of weights (final survey weights)

y_1	x_{11}	\dots	x_{p1}	w_{A1}
y_2	x_{12}	\dots	x_{p2}	w_{A2}
\dots	\dots	\dots	\dots	\dots
y_{m_A}	x_{1m_A}	\dots	x_{pm_A}	w_{Am_A}

A containing m_A units

B containing m_B units

x_{11}	\dots	x_{p1}	z_1	w_{B1}
x_{12}	\dots	x_{p2}	z_2	w_{B2}
\dots	\dots	\dots	\dots	\dots
x_{1m_B}	\dots	x_{pm_B}	z_{m_B}	w_{Bm_B}

And the objective of inference typically concerns finite population parameters concerning the relationship between Y and Z ($\rho_{U,YZ}$, $B_{U,YZ}$, N_{jk} , ecc.) or among \mathbf{X} , Y and Z

Approaches to perform SM when dealing with data from complex sample surveys (cf. D'Orazio, 2010):

"old" approaches:

- Naïve approach (micro objective)
- File concatenation (Rubin, 1986) (micro objective)
- Case weights calibrations (Renssen, 1998) (macro and micro objective)

"new" approach (not developed for SM):

- Empirical Likelihood (Wu, 2004)
 - Separate EL approach
 - Combined EL approach

Naïve approach (cf. D'Orazio, 2012)

It is used a nonparametric micro SM technique to create a synthetic data file by filling in A (the recipient) with the values of the missing variable.

Then the inference is carried out on A in a design based framework by considering the origin survey weights associated to the units in A

Weights can be used or not in the matching. In particular:

- random hot deck: the donors in B can be selected with probability proportional to their survey weights (**weighted random hot deck**)
- rank hot deck: the empirical cumulative distribution of X can be estimated in A and B by considering the corresponding survey weights:

$$\hat{F}^{(A)}(x) = \frac{\sum_{i=1}^{m_A} w_i^{(A)} I(x_{A,i} \leq x)}{\sum_{i=1}^{m_A} w_i^{(A)}}, \quad \hat{F}^{(B)}(x) = \frac{\sum_{i=1}^{m_B} w_i^{(B)} I(x_{B,i} \leq x)}{\sum_{i=1}^{m_B} w_i^{(B)}}$$

Function `RANDwNND.hotdeck` in **StatMatch**

```
RANDwNND.hotdeck(data.rec, data.don, match.vars=NULL,  
                  don.class=NULL, dist.fun="Manhattan",  
                  cut.don="rot", k=NULL, weight.don=NULL,  
                  ...)
```

`data.rec`: *recipient* data set (file *A*)

`data.don`: *donor* data set (file *B*)

`match.vars`: names of the matching variables (optional)

`don.class`: names of the variables (categorical) to create donation classes

...

`weight.don`: name of the variable in *B* to be used in weighted selection of donors

Function `rankNND.hotdeck` in **StatMatch**

```
rankNND.hotdeck(data.rec, data.don, var.rec,  
                var.don=var.rec, don.class=NULL,  
                weight.rec=NULL, weight.don=NULL,  
                constrained=FALSE, constr.alg="Hungarian")
```

`data.rec`: *recipient* data set (file *A*)

`data.don`: *donor* data set (file *B*)

`var.rec`: name of the unique matching variables in *A*

`var.don`: name of the unique matching variables in *B*

`weight.rec`: name of the variable containing weights in *A*

`weight.don`: name of the variable containing weights in *B*

These approaches not fully investigated

The major risk: the marginal distribution of imputed variable Z in A is not coherent with the reference one (that of Z in B)

D'Orazio et al. (2012) compared several naive procedures in a simulation study:

- rank and random hot deck using the weights tend to perform quite well in terms of preservation in the synthetic data set of the marginal distribution of X and of the joint distribution $X \times Z$.
- distance hot deck, provides good results only when constrained matching is used and a design variable (e.g. a stratification variable) is considered in forming donation classes.

File concatenation (Rubin, 1986)

The samples are concatenated $S = A \cup B$ and a new inclusion probability is computed for each unit in the **concatenated sample** $A \cup B$:

$$\begin{aligned}\pi_{A \cup B, c} &= \pi_{A, c} + \pi_{B, c} - \pi_{A \cap B, c} \\ &\cong \pi_{A, c} + \pi_{B, c}\end{aligned}\quad c \in A \cup B$$

- ☹ handles theoretic samples and does not account for nonresponse, etc.
- ☹ difficulties in deriving $\pi_{A \cup B, c}$; required:
 - knowledge of sampling designs used to select A and B respectively
 - the A design variables to be available in A and in B
 - the B design variables to be available in A and in B(for details see Ballin *et al.*, 2008)
- 😊 $\theta_{\mathbf{X}}$ can be estimated directly on $A \cup B$ (higher accuracy)
- ☹ Methods to deal with missing values have to be chosen for estimating $(\theta_{Y|\mathbf{X}}, \theta_{Z|\mathbf{X}})$

Case weights calibrations (Renssen, 1998)

Macro objective: the procedure aims at estimating the two-way contingency table $Y \times Z$ starting from the data of two independent complex sample surveys carried out on the same finite population



It deals mainly with categorical variables

It permits to exploit eventual auxiliary information represented by a third data source C in which (X, Y, Z) or simply (Y, Z) are available

It is based on a series of steps of **calibration** of the survey weights in A and in B .

The two data sources are kept separate.

How calibration works:

Assume

d_k : “starting” weights

w_k : final calibrated weights

the final weights w_k are derived as the solution of:

$$\min \left[\sum_{k \in r} D(d_k, w_k) \right] \quad D(d, w) \text{ is a distance measure}$$

Subject to (for instance):

$$\sum_{k=1}^m w_k x_k = \sum_{k=1}^N x_k ; \quad \sum_{k=1}^m w_k = N$$

Renssen's procedure consists of two phases (calibrations)

it will be assumed that all the variables (X_D, Y, Z) are categorical, being X_D a complete or an incomplete crossing of the matching variables X_M

1st phase: harmonisation of two surveys wrt to the totals of X_D :

Case 1.a): pop. totals of X_D are **known**

1.a.2.A) weights w_a in A are calibrated so that the new calibrated weights $w_a^{(1)}$ reproduce the known totals of X_D

1.a.2.B) weights w_b in B are calibrated so that the new calibrated weights $w_b^{(1)}$ reproduce the known totals of X_D

case 1.b): pop. totals of X_D are **unknown**

Compute the **pooled estimate** for the totals:

$$\tilde{t}_{X_D} = \lambda \hat{t}_{X_D}^{(A)} + (1 - \lambda) \hat{t}_{X_D}^{(B)} \quad \lambda \in [0,1]$$

Usually $\lambda = m_A / (m_A + m_B)$

1.b.2.A) Weights $w_a^{(1)}$ in A are calibrated so that the new calibrated weights $w_a^{(2)}$ reproduce the pooled totals

1.b.2.B) Weights $w_b^{(1)}$ in B are calibrated so that the new calibrated weights $w_b^{(2)}$ reproduce the pooled totals

Harmonization of distributions in **StatMatch** via `harmonize.x()`

```
harmonize.x(svy.A, svy.B, form.x, x.tot=NULL,  
            cal.method="linear", ...)
```

svy.A: data frame *A* and the corresponding sampling design (see package **survey** by Lumley, 2012)

svy.B: data frame *B* and the corresponding sampling design (see package **survey** by Lumley, 2012)

form.x: formula specifying the *X* variables whose distribution has to be harmonized

x.tot: eventual external known totals for the *X* variables

cal.method: calibration method (see package **survey** by Lumley, 2012)

2nd phase: use of an additional data source C in which \mathbf{X} , Y and Z are jointly observed in order to provide estimates concerning the relationship among Y and Z .

Two approaches (still based on calibrations):

- **Incomplete two-way stratification**
 - 2.1.1) Estimate total of Y on A and total of Z on B
 - 2.1.2) calibrate weights of C to reflect these totals.
 - 2.1.3) estimates $Y \times Z$ from file C considering the new calibrated
- **Synthetic two-way stratification**
 - 2.2.1) Estimate the relationship among $Y \times Z$ under the CIA.
 - 2.2.2) "corrects" the CIA according to the "distance" between the CIA and the estimate of $Y \times Z$ in C

See Renssen (1998) for major details.

Matching samples with the Renssen's approach in **StatMatch** Under CI assumption (C is NOT available)

```
comb.samples(svy.A, svy.B, svy.C=NULL, y.lab, z.lab, form.x,  
             estimation=NULL, micro=FALSE, ...)
```

svy.A: data frame A and the corresponding sampling design

svy.B: data frame B and the corresponding sampling design

y.lab: name of the variable Y

z.lab: name of the variable Z

form.x: formula specifying the matching variables X

Matching samples with the Renssen's approach in **StatMatch**
with incomplete two-way stratification (C is available)

```
comb.samples(svy.A, svy.B, svy.C=svy.C, y.lab, z.lab, form.x,  
             estimation="incomplete", micro=FALSE, ...)
```

svy.A: data frame A and the corresponding sampling design

svy.B: data frame B and the corresponding sampling design

svy.C: data frame C and the corresponding sampling design containing
at least Y and Z

y.lab: name of the variable Y

z.lab: name of the variable Z

form.x: formula specifying the matching variables X

estimation: how to estimate $Y \times Z$ ("incomplete" or "synthetic")

Matching samples with the Renssen's approach in **StatMatch**
with synthetic two-way stratification (C is available)

```
comb.samples(svy.A, svy.B, svy.C=svy.C, y.lab, z.lab, form.x,  
             estimation="synthetic", micro=FALSE, ...)
```

svy.A: data frame A and the corresponding sampling design

svy.B: data frame B and the corresponding sampling design

svy.C: data frame C and the corresponding sampling design containing
 X, Y and Z

y.lab: name of the variable Y

z.lab: name of the variable Z

form.x: formula specifying the matching variables X

estimation: how to estimate $Y \times Z$ ("incomplete" or "synthetic")

Renssen's calibration procedure performs well when dealing with categorical variables. It is developed to estimate the contingency table $Y \times Z$

It works with the available survey data, there is no need of having the theoretic sample.

The harmonization procedures permit to harmonize the marginal or the joint distribution of the \mathbf{X} variables among the two data sources A and B .

Calibration may not be successful

This is likely to happen in the presence of too many categories, or when dealing with mixed type variables.

Tips to solve the problem: reduce categories for categorical variables; categorize continuous variables.

The Renssen's approach has a macro objective (estimation of the contingency table $Y \times Z$)

For these purposes the **linear probability models** are fitted at unit level by taking into account the survey weights in the estimation of the regression parameters.

The usage of the models enables to perform a regression imputation of the missing variables at micro level.

Matching samples with the Renssen's approach in **StatMatch**
micro imputation (all the methods, *C* available or not)

```
comb.samples(svy.A, svy.B, svy.C=NULL, y.lab, z.lab, form.x,  
             estimation=NULL, micro=TRUE, ...)
```

the micro imputation with the Renssen's procedure:

- 😊 provides a synthetic data set that when all the variables X , Y and Z are categorical preserves the marginal distribution of the imputed variable and its joint distribution with X .
- 😞 the imputed value for each recipient is a set of estimated probabilities: i.e. the probabilities that the given unit assumes one of the categories of the imputed variable
- 😞 the linear probability models can provide negative or greater than one estimates for the probabilities (other well-known drawbacks of these models are heteroskedasticity and residuals not normally distributed).

For these reasons, such predictions should be used carefully

Empirical Likelihood to combine data from multiple surveys (Wu, 2004)

Harmonization phase similar to the calibration one:

- derive new weights for units in A
- derive new weights for units in B

in order to satisfy some constraints concerning the totals of \mathbf{X} variables.

The target totals of the \mathbf{X} variables can be:

- known from external sources
- estimated by combining estimates obtained separately from A and from B (**Wu's separate approach**) (similar to Renssen's approach)
- unknown and not estimated: they are set to be equal as a further constraint in the optimization problem (**Wu's combined approach**)

Wu's approach:

- is more flexible if compared to calibration (no negative weights)
- in the combined case, it does not require to estimate the totals of the **X** variables
- major complexity in the presence of a stratified sampling with allocation that is NOT proportional

For major details on the EL see Chen and Sitter (1999)

To sum up:

Naïve approach

- 😊 very simple and flexible to be performed
- 😊 starts from the available data sources no matter whether they refer to respondents to a survey or to the theoretic sample
- 😞 it is required the choice of the matching variables
- 😞 the practical and theoretical implications have not been investigated

File concatenation (Rubin, 1986)

- 😊 quite simple in practice; it is used just the concatenated file
- 😊 better estimates for **X** variables
- 😊 it is not required an harmonization step among the two files, the target totals of the **X** variables are the ones estimated on the concatenated file
- 😞 it is very difficult to determine the $\pi_{A \cup B, c}$
- 😞 Methods to deal with missing values have to be chosen for estimating characteristic related to *Y*, *Z* and their relationship
- 😞 the theory is developed to concatenate the theoretic samples but in practice the available data refer to the **respondents** to a survey (final weights incorporate corrections for unit nonresponse, coverage, ...)

Case weights calibration (Renssen, 1998)

- 😊 harmonizes marginal (joint) distributions of the \mathbf{X} variables in both the samples
- 😊 after the harmonization allows the direct estimation of the parameters related to Y and to (\mathbf{X}, Y) on A , and those for Z and (\mathbf{X}, Z) on B
- 😊 allows to easily introduce eventual auxiliary data sources (file C) into the estimation process
- 😊 starts from the available data sources and the weights associated to the units (no matter whether they refer to respondents to a survey or to the theoretic sample)
- 😊 provide micro imputations whose distribution is usually coherent wrt the one in the donor

- ☹ subjective choice (λ) on how to estimate the totals of the **X** variables when they are not known
- ☹ calibration may fail
- ☹ it is difficult to handle continuous and both continuous and categorical variables
- ☹ imputed micro values for categorical variables consist in estimated probabilities, these probabilities can be negative or greater than one

Empirical Likelihood to combine samples (Wu, 2004)

- 😊 is more flexible if compared to calibration (no negative weights; handles mixed type variables)
- 😊 it does not require to estimate the totals of the \mathbf{X} when they are not known ("combined" approach)
- 😊 introduces a comprehensive framework (EL) to make inference in the presence of data from complex sample surveys
- 😞 theory is more complex. A major complexity is required to deal with nonproportional stratified sampling
- 😞 Methods allow combining theoretic samples, but it is difficult to handle unit nonresponse

A further investigation of the EL approach for Statistical Matching is required.

Some References

- Ballin, M., Di Zio, M., D'Orazio, M., Scanu, M., Torelli, N. (2008) "File Concatenation of Survey Data: a Computer Intensive Approach to Sampling Weights Estimation". *Rivista di Statistica Ufficiale*, 2-3, 5-12
- Chen, J., Sitter, R.R. (1999) "A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys". *Statistica Sinica*, 9, 385–406
- D'Orazio, M. (2012) "Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment". R Package Vignette.
- D'Orazio, M., Di Zio, M., Scanu, M. (2006) *Statistical Matching: Theory and Practice*. Wiley, Chichester.
- D'Orazio, M., Di Zio, M., Scanu, M. (2009) "Uncertainty intervals for nonidentifiable parameters in statistical matching" *Proceedings 54th ISI Session*, 16-22 August 2009, Durban, South Africa
- D'Orazio, M., Di Zio, M., Scanu, M. (2009) "Old and New Approaches in Statistical Matching when Samples are Drawn with Complex Survey Designs". 45th SIS Scientific Meeting, Padua, June 16-18, 2010
- Renssen, R.H. (1998) "Use of statistical matching techniques in calibration estimation". *Survey Methodology*, 24, 171–183
- Rubin, D.B. (1986) "Statistical matching using file concatenation with adjusted weights and multiple imputations". *J. Bus. Econ. Stat.* 4, 87–94
- Särndal C.E, Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- Wu, C. (2004) "Combining information from multiple surveys through the empirical likelihood method". *Can. J. Stat.* 32, 1-12 (2004)

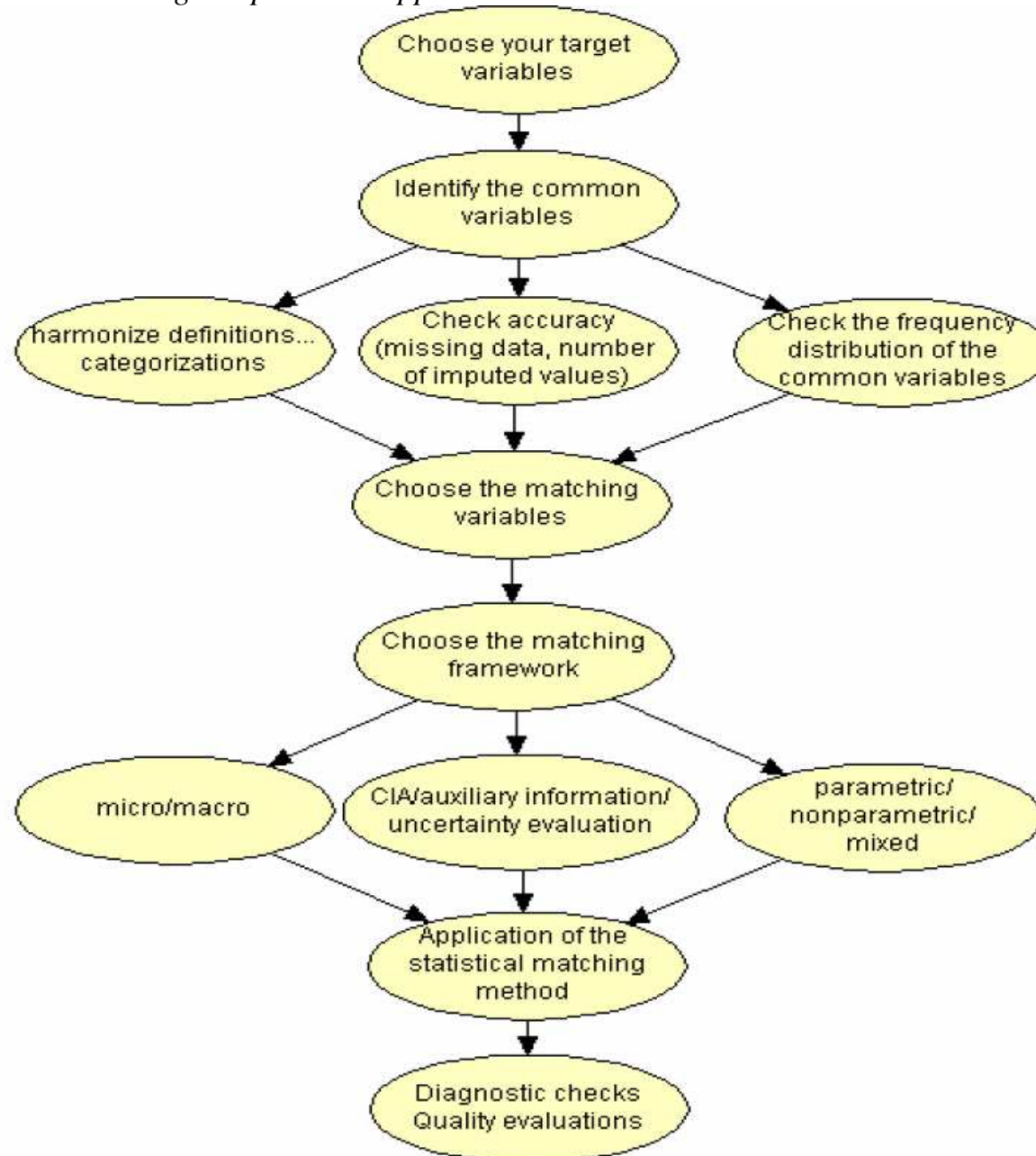
Statistical Matching:
Methodological issues and practice with R-StatMatch
Vitoria-Gasteiz, 21-22 November 2013

PRACTICAL ISSUES IN STATISTICAL MATCHING

Marcello D'Orazio*
[madorazi\(at\)istat.it](mailto:madorazi@istat.it)

**Italian National Institute of Statistics,*





Cf .Chapter 3 in ESSnet Report of WP2 (2009)

- 1) Choice of the target variables, i.e. of the variables observed distinctly in two sample surveys.
- 2) Identification of all the common variables in the two sources. Not all can be used due to lack of harmonization, different definitions, etc.
- 3) Choice of the matching variables: just those that are able to predict the target variables
- 4) Application of the chosen SM technique
- 5) Evaluation of the results of SM

Step 2: Identification of the common variables

A key role is played by the common variables \mathbf{X} . The matching variables \mathbf{X}_M are a subset of the common variables ($\mathbf{X}_M \subseteq \mathbf{X}$).

Before this selection, it is necessary to modify and discard those variables that present some problems:

- i) definitions and the corresponding classifications
- ii) accuracy (missing values, imputed values, measurement errors)
- iii) statistical content: the two samples A and B should estimate (almost) the same distribution for each common variable

If all the common variables are characterized by estimated distributions with many differences there is the possibility that the target populations of the two surveys are different

2.iii) Comparison of the marginal distributions of the common variables

Formally requires statistical tests (Chi-Square, Kolomogorov-Smirnov)

Problem of using tests modified to account for complex sampling design
(for modified Chi-Square test cf. Sarndal et al., 1992, pp. 500-513)

An empirical approach is often used which consists in comparing the marginal distributions estimated from the two surveys using [similarity/dissimilarity](#) measures

Total variation distance:

$$\Delta(\hat{p}_A, \hat{p}_B) = \frac{1}{2} \sum_{j=1}^J |\hat{p}_{A,j} - \hat{p}_{B,j}|, \quad p_{s,j} = \frac{\hat{N}_{s,j}}{\hat{N}_s}$$

$$\Delta(\hat{p}_A, \hat{p}_B) = 1 - OV(\hat{p}_A, \hat{p}_B) = 1 - \sum_{j=1}^J \min(\hat{p}_{A,j}, \hat{p}_{B,j})$$

The smallest fraction of units that would need to be re-classified in order to make the two distributions equal

OV: overlap among the two distributions

When Δ is used to assess the closeness of an observed distribution to the expected values under a given model then, following Agresti (2002, pp. 329-330), $\Delta < 0.02$ or 0.03 , denotes that "*the sample data follow the model pattern quite closely, even though the model is not perfect*".

when comparing estimated marginal distributions of the same variable but from independent data sources, the expected distribution of the variable is

$$\hat{P}_{+c} = \frac{n_A \hat{P}_{Ac} + n_B \hat{P}_{Bc}}{n_A + n_B} = \lambda_A \hat{P}_{Ac} + (1 - \lambda_A) \hat{P}_{Bc} \quad \lambda_A = n_A / (n_A + n_B)$$

(weighted average of the two estimated distribution)

Thus, when comparing both the distributions with the expected one, following the Agresti's rule of thumb, the starting distributions can be considered coherent if $\Delta_{AB} \leq 0.06$.

Bhattacharyya coefficient:

$$BC(\hat{p}_A, \hat{p}_B) = \sum_{j=1}^J \sqrt{\hat{p}_{A,j} \hat{p}_{B,j}} \quad 0 \leq BC \leq 1$$

Hellinger Distance

$$d_H(\hat{p}_A, \hat{p}_B) = \sqrt{1 - BC(\hat{p}_A, \hat{p}_B)}$$

Satisfies the properties of distance ($0 \leq d_H \leq 1$, symmetry, triangle inequality).

It is not possible to determine a threshold of acceptable values of d_H for say two distribution to be close. However, it is possible to show that:

$$d_H^2 \leq \Delta \leq d_H \sqrt{2}$$

admitting values such that $\Delta \leq 0.06$ means admitting $d_H \leq 0.042$ (a rule of thumb often recurring in literature considers two distributions close if the Hellinger's distance is not greater than 0.05)

Comparing marginal distribution of categorical variables in **StatMatch** (No reference distribution)

```
comp.prop(p1, p2, n1, n2=NULL, ref=FALSE)
```

p1 is the first distribution

p2 is the second distribution

n1 size of first sample

n2 size of second sample

ref when TRUE then the second distribution is the reference one

Comparing the marginal distributions of continuous variables poses major problems.

Descriptive statistics (minimum, maximum, average, standard deviation, coefficient of variation, percentiles) and graphical analysis can be of help.

In alternative, the continuous variables can be categorized and the tools previously introduced can be used.

Step 3: Choice of the matching variables

In practical situations the two data sources A and B may share several variables in common

The selection of the matching variables should be performed by:

- using statistical methods (descriptive, inferential, ...)
- consulting subject matter experts

Keeping in mind the principle of parsimony: choosing too many matching variables may:

- increase complexity of the matching application
- affect negatively the accuracy of the matching results

Cohen (1991): the choice should be carried out in a “multivariate sense” in order to identify the subset X_M ($X_M \subseteq X$) connected at the same time with Y and Z

This would require the availability of a data source in which all the variables X , Y and Z are observed.

In the standard SM framework only A and B are available:

file A: permits to explore the relationship between X_S and Y , leading to identify X_Y ($X_Y \subseteq X$), the subset of the common variables that better explains Y

file B: permits to explore the relationship among X_S and Z , leading to identify X_Z ($X_Z \subseteq X$), the subset of the common variables that better explains Z

The results of the separate analysis are evaluated jointly

Usually the subset of the matching variables is obtained as:

$$X_M = X_Y \cup X_Z$$

Such a way of working may provide too many matching variables.

For this reason in most of the cases the set of the matching variables is obtained as a compromise:

$$X_Y \cap X_Z \subseteq X_M \subseteq X_Y \cup X_Z$$

Subject matter experts can be of help in finding a good balance between the two extrema

Let consider the data source A , containing X and Y variables.

The simplest procedure to identify X_Y consists in computing **pairwise correlation/association measures** between the Y and each of the available X variables

The measures to consider depends on the nature of the variables

- a) Y continuous or categorical ordered, X continuous or cat. ordered
Spearman correlation coefficient (correlation coefficient computed on the ranks of $s = \text{rank}(Y)$ and $r = \text{rank}(X)$):

$$\rho_s = \frac{\sum_{i=1}^{n_A} (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n_A} (r_i - \bar{r})^2 \sum_{i=1}^{n_A} (s_i - \bar{s})^2}}$$

Can detect nonlinear monotonic relationship.

$\rho_s^2 = R_{sr}^2$ i.e. the coefficient of determination of the linear regression:

$\text{rank}(Y)$ vs. $\text{rank}(X)$

Harrell (2001) suggests considering the adjusted version of R^2

When non-monotonic relationship is supposed to exist (e.g. "U" shaped relationship) Harrell suggests considering R_{adj}^2 :

$\text{rank}(Y)$ vs. $\text{rank}(X) + [\text{rank}(X)]^2$

b) Y continuous or categorical ordered, X categorical nominal

The ranks of Y values are considered, while X is substituted by the corresponding $I - 1$ dummy variables (I are the categories of X).

It can be considered the coefficient of determination of the regression

rank(Y) vs. dummies(X)

R^2 in this case corresponds to the **Eta-squared** (η^2) measure of effect size used in ANOVA but computed on $s = \text{rank}(Y)$

$$\eta_s^2 = \frac{\sum_{i=1}^I n_i (\bar{s}_i - \bar{s})^2}{\sum_{i=1}^I \sum_{a=1}^{n_i} (s_{ia} - \bar{s})^2}$$

such a measure is strictly related to the Kruskal-Wallis test statistic

c) Y categorical nominal, X categorical nominal (or categorical ordered).

Chi-squared based association measures can be considered. E.g. Cramer's V ($0 \leq V \leq 1$; association between two variables as a percentage of their maximum possible variation). V^2 is the mean square canonical correlation between the variables. The more unequal the margins, the more V will be less than 1

It is better to reason in terms of variance (as with R^2 in linear regression analysis) by considering the **proportional reduction of the variance** of Y when passing from the marginal distribution to its conditional distribution given X (cf. Agresti, 2002, p. 56):

$$\frac{V(Y) - E[V(Y|X)]}{V(Y)}$$

$$E[V(Y|X)] = \sum_{i=1}^I p_{i+} V(Y|i)$$

c) Y cat. nominal, X cat. nominal (or cat. ordered) (*cont.*)

Unfortunately when dealing with categorical variables there is not a general accepted definition of variance.

- If variance is measured in terms of **entropy**

$$V(Y) = -\sum_{j=1}^J p_{+j} \log p_{+j}$$

and the proportional reduction of variance formula gives the Theil's **uncertainty coefficient** (cf. Agresti, 2002, p. 56):

It provides the relative reduction of uncertainty when predicting Y using the information provided by X . $U_{YX} = 0$ denotes that X is not of help in predicting Y

- When **concentration** is considered:

$$V(Y) = 1 - \sum_{j=1}^J p_{+j}^2$$

and the Goodman & Kruskal **concentration coefficient** comes out:

- If one looks at **classification errors**:

$$V(Y) = 1 - \max_j (p_{+j})$$

and the Goodman & Kruskal λ association measure results. It can be interpreted as the proportional reduction in error when predicting Y based on X

Other methods for the selection of the best predictors.

Fitting a **linear regression model**: it is possible to use automatic procedures (backward, forward, stepwise) to reduce the predictors. It would be preferable to use procedures based on the residual Chi-square or the Akaike's information criterion (AIC) (cf. Harrell, 2001).

When all the predictors are continuous **Least Angle Regression** procedures can be used (Efron et al., 2004)

Procedures similar to Least Angle Regression have been developed for **linear regression model with ordinal predictors** (Gertheiss 2011) and for the **Generalised Linear Mixed Models** (Groll & Tutz, 2011).

In complex cases (categorical response and/or mixed type predictors), using nonparametric regression procedures can be of help.

Classification and regression trees (CART; Breiman et al. 1984) can detect nonlinear relationship among response and the predictors.

CART suffer of some well-known drawbacks: selection bias (tend to prefer predictors with many possible splits) and collinearity.

For this reason it would be better to resort to **Random Forests** procedures (Breiman, 2001).

Random forests procedure, present the further advantage of providing measures of predictors' importance (to be used carefully).

Some problems:

Too many common variables: before searching the best predictors it would be preferable to discard redundant predictors (**redundancy analysis**, cf. Harrell 2012; or **variable clustering**, cf. Sarle, 1990; Harrell 2012)

Units' weights (complex sample surveys data): sometimes they can be used directly in the analyses (it is suggested to compare weighted and un-weighted results)

E.g.: in regression analysis one should use ad hoc methods, i.e. **design weighted least squares**

or

it is possible to use weight but introduce the design variables into the model as explanatory variables.

Selecting matching variables using uncertainty approach

Relies on the evaluation of the uncertainty due to the matching framework (Y and Z not jointly observed)

The idea consists in searching for the **subset of common variables more effective in reducing this uncertainty**

In the case of categorical variables, assuming that X_D corresponds to the complete crossing of some of the X variables

$$P_{j,k}^{(low)} \leq P_{Y=j,Z=k} \leq P_{j,k}^{(up)},$$

with

$$P_{j,k}^{(low)} = \sum_{i=1}^I P_{X_D=i} \times \max\{0; P_{Y=j|X_D=i} + P_{Z=k|X_D=i} - 1\}$$
$$P_{j,k}^{(up)} = \sum_{i=1}^I P_{X_D=i} \times \min\{P_{Y=j|X_D=i}; P_{Z=k|X_D=i}\}$$

This result is used in searching the subset of the X variables more effective in reduction of uncertainty

The function **Fbwidths.by.x** in **StatMatch**

estimates $(P_{j,k}^{(low)}, P_{j,k}^{(up)})$ for each cell in the contingency table $Y \times Z$ in correspondence of all the possible combinations of the subsets of the X s

The reduction of uncertainty is measured according to the proposal of Conti *et al.* (2012):

$$\hat{\Delta} = \sum_{i,j,k} \left(\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)} \right) \times \hat{P}_{Y=j|X_D=i} \times \hat{P}_{Z=k|X_D=i} \times \hat{P}_{X_D=i}$$

or, naively, by considering the average widths of the intervals:

$$\bar{d} = \frac{1}{J \times K} \sum_{j,k} (\hat{P}_{j,k}^{(up)} - \hat{P}_{j,k}^{(low)})$$

Fbwidths.by.x(tab.x, tab.xy, tab.xz)

tab.x joint distribution (contingency table) of all the X variables
candidate as matching variables

tab.xy joint distribution (contingency table) $X \times Y$

tab.xz joint distribution (contingency table) $X \times Z$

Step 5: quality evaluations of the SM results

Common problems:

- (i) The general objective of SM is to study the relationship of **phenomena not jointly observed**, unless an additional auxiliary data source is available
- (ii) The SM can provide **different outputs**:
 - a synthetic data set in the micro case
 - one or more estimates (e.g. a correlation coefficient, a regression coefficient, probabilities in a contingency table, etc.) in the macro case
- (iii) There are two or more **input data sources with different quality "levels"** (sampling design, sample size, data treatment and processing)

The absence of observations where to study the relationship between the interest variables

It is the major source of uncertainty concerning the matching results

This lack of information has to be filled in the by:

- making some assumptions (e.g. the conditional independence of the target variables given the matching variables)
- using additional auxiliary information (an external estimate of the interest parameters or an additional data source, etc.)

Unless the approach based on the evaluation of uncertainty it is considered

The results of the SM will necessarily reflect the assumptions/information being used:

- results of a matching application based on the CI assumption will reflect it; they will be unreliable if CI is not holding
- when auxiliary information is used (CIA avoided), the result of the SM are expected to reflect such input. If the input information is not reliable, the results of SM will be unreliable.

In this setting the researcher “knows” what to expect he has check just whether the chosen matching method has been applied correctly, avoiding the introduction of additional noise or bias.

Checks in the macro case

The outputs are estimates of parameters

It may be easy to check whether there is some undesired additional variation or bias

Under the CI assumption in some cases it is possible derive analytic estimation formulas for the parameters of interest

Correlation coefficient: $\tilde{\rho}_{YZ}^{(CI)} = \hat{\rho}_{YX} \times \hat{\rho}_{ZX}$

cell probabilities: $\tilde{P}_{Y=j,Z=k}^{(CI)} = \sum_{i=1}^I \hat{P}_{X=i} \times \hat{P}_{Y=j|X=i} \times \hat{P}_{Z=k|X=i}$

Simulation based studies can be of help

Check in the micro case

The output of SM is a synthetic file with all the needed variables

It is a common practice to investigate the representativeness of the synthetic file (in a wide sense considering the relationship between variables too)

Rässler's (2002) suggests to look at the "validity" of the SM procedure by analyzing how the synthetic data set:

- preserves the **marginal distribution of the imputed variable** (reference is the one in the donor data set);
- preserves the **joint distribution of the imputed variable with the matching variables** (reference is the one in the donor data set).

In order to compare marginal or joint distributions of the variables in the synthetic data set with respect to the one in the donor it is possible to use statistical tests and/or descriptive measures

In the case of categorical variables the descriptive measures (dissimilarity index, Hellinger's distance, etc.) are the ones introduced previously

Note: the distribution in the donor is the reference

Comparing marginal distribution of categorical variables in **StatMatch**
(p2 is the reference distribution)

```
comp.prop(p1, p2, n1, n2, ref=TRUE)
```

In the case of continuous variables descriptive statistics (mean, sd, percentiles, etc.) graphical analysis, etc.

Example: matching of the Survey of Income and Wealth (SHIW) of Italian households (Italian Central Bank) and the Household Budget Survey (HBS) (Istat)

Results of the SM nearest neighbour hot-deck under CI

Estimates from fused dataset compared with HBS estimates (HBS=100)

	(d1)	(d2)	(d3)	(d4)	(d5)	(d6)
Average C_{tot}	100.00	99.70	99.49	99.17	99.08	97.93
sqm C_{tot}	100.01	99.15	99.88	99.69	101.72	120.75

Estimates of correlations from the fused data set

	(d1)	(d2)	(d3)	(d4)	(d5)	(d6)
$\text{Cor}(M_{tot}, \tilde{C}_{tot})$	0.3037	0.3345	0.3084	0.3009	0.2779	0.2531

- Donation classes AREA5 & NCOMP
- Variables used to compute distance
 - (d1) HSURF
 - (d2) HSURF, NOCC4
 - (d3) HSURF, NOCC4, NUNID3
 - (d4) HSURF, NOCC4, NUNID3, NSSD4
 - (d5) HSURF, NOCC4, NUNID3, NSSD4, NRET3
 - (d6) HSURF, NOCC4, NUNID3, NSSD4, NRET3, HRENT

Some References

- Breiman, L. (2001) “Random Forests”, *Machine Learning*, **45**, 5-32.
- Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees*. Wadsworth.
- Cohen, M.L. (1991) “Statistical matching and microsimulation models. Improving Information for Social Policy Decisions”, *The Use of Microsimulation Modeling*, vol. II. National Academy Press.
- Conti, PL and Marella, D and Scanu, M (2012) “Uncertainty Analysis in Statistical Matching”, *Journal of Official Statistics*, **28**, pp. 69-88.
- D’Orazio, M. (2012) “Statistical Matching and Imputation of Survey Data with the Package StatMatch for the R Environment”. R Package Vignette.
- D’Orazio, M., Di Zio, M., Scanu, M. (2006) *Statistical Matching: Theory and Practice*. Wiley, Chichester.
- Efron B., Hastie T., Johnstone I., Tibshirani R. (2004), “Least Angle Regression” (with discussion), *Annals of Statistics*, **32**, 407-499
- Eurostat (2009) “Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data”. ESSnet Statistical Methodology Project on Integration of Survey and Administrative Data
- Harrell, F.E. (2001) *Regression Modeling Strategies, with Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer-Verlag, New York.
- Hothorn T., Hornik, K., Zeileis, A. (2006) “Unbiased Recursive Partitioning: A Conditional Inference Framework”. *Journal of Computational and Graphical Statistics*, **15**, pp. 651–674
- Lee, E.S., and Forthofer, R. N. (2005) *Analyzing Complex Survey Data. Second Edition*. Sage Publications
- McCullagh P. and Nelder, J. A. (1989) *Generalized Linear Models*. London: Chapman and Hall.
- Paass G. (1986) “Statistical match: evaluation of existing procedures and improvements by using additional information”. In *Microanalytic Simulation Models to Support Social and Financial Policy* (ed. Orcutt GH and Quinke H) Elsevier Science, pp. 401–422
- Rässler S (2002) *Statistical Matching: a Frequentist Theory, Practical Applications and Alternative Bayesian Approaches*. Springer Verlag.
- Särndal, C. E., Swensson, B, and Wretman, J. (1992) *Model Assisted Survey Sampling*. Springer, New York.