# Dealing with Non-ignorable Sampling and Non-response in Statistical Matching

Daniela Marella [1] Danny Pfeffermann [2]

[1]Sapienza Università di Roma

[2]University of Southampton, Hebrew University of Jerusalem and former National Statistician and Director General of Israel's Central Bureau of Statistics

Workshop on methodologies for official statistics

# Statistical Matching Problem

Let A and B be two samples from a population generated from the joint *pdf* $f_p(x, y, z; \theta)$:

1. $(X, Y)$ are observed in sample $A$ of size $n_A$;
2. $(X, Z)$ are observed in sample $B$ of size $n_B$.

The aim is to estimate the joint *pdf* of $(X, Y, Z)$.
No joint observations on $(Y, Z)$ $\implies$ the model of $(X, Y, Z)$ is not identifiable.

# Alternatives approaches to SM

Several alternative approaches to overcome the identification problem:

1. Conditional independence assumption (CIA) between $Y$ and $Z$ given $X$;
2. External information regarding the relationship between $Y$ and $Z$;
3. Analyze the uncertainty regarding the joint *pdf* of $(X, Y, Z)$.

# Aim of the talk

The aim of this talk is:

1. to discuss under the CIA the SM problem when the samples $A$ and $B$ are informative and the nonresponse in NMAR by an approach based on EL.
2. to drop the CIA and define a class of plausible joint *pdf*s for $(X, Y, Z)$.
3. to show the results of an application to SHIW and HBS datasets.

# Empirical likelihood approach

The EL approach approximates the population *pdf* by a multinomial model, which support is given by the empirical observations:

1. it combines the robustness of nonparametric methods with the efficiency of the likelihood approach;
2. it does not require specifying the population model, and is thus more robust and often easier to implement;
3. it facilitates the use of calibration constraints. The population mean $\mu_X$ of $X$ is known.

# Empirical likelihood approach

1. $Y$ and $Z$ are continuous;
2. $X$ is a discrete variable taking $K$ distinct values. $A_k = \{i \in A : x_i = x_k\}$ is the set of sample units in A with $X = x_k$ of size $n^x_{k,A}$ such that for $i \in A_k$

$$p_i^X = P(X = x_k) = p_k^X$$

for $k = 1, \ldots, K$.

Under the CIA, the joint population multinomial probability of unit $i$ is given by:

$$p_i^{XYZ} = P(x_i)P(y_i|x_i)P(z_i|x_i) = p_k^X p_i^{Y|X} p_i^{Z|X}$$

# EL under non-ignorable sampling

Under informative sampling, the observed outcomes are no longer representative of the population outcomes and the sample models are different from the corresponding population models. Pfeffermann et al. (1998) establish general conditions under which for independent observations under the population model, the sample measurements are asymptotically independent under the sample model, when increasing the population size but holding the sample size fixed. Then, the sample likelihood can be approximated by the product of the sample *pdf*s over the corresponding sample observations.

# EL under non-ignorable sampling

The sample EL based on $A \cup B$ is

$$EL_{Obs}^{A \cup B} = \prod_{k=1}^{K} (p_{k,A}^{X})^{n_{k,A}^{X}} \prod_{i \in A_k} p_{i,A}^{Y|X} \prod_{k=1}^{K} (p_{k,B}^{X})^{n_{k,B}^{X}} \prod_{i \in B_{xk}} p_{i,B}^{Z|X}$$

where the sample models

$$\begin{aligned} p_{i,A}^{Y|X} &= P(y_i|x_i, I_i^A = 1) = \frac{P(I_i^A = 1|x_i, y_i)}{P(I_i^A = 1|x_i)} p_i^{Y|X} \\ p_{k,A}^{X} &= P(x_k|I_i^A = 1) = \frac{P(I_i^A = 1|x_k)}{P(I_i^A = 1)} p_k^{X} \end{aligned}$$

for $i \in A_k$ where $I_i^A$ the sample indicator. Analogous expression can be obtained for sample $B$.

# EL under non-ignorable sampling

1. The probabilities $P(I_i^A = 1|x_i, y_i) = 1/E_A(w_{i,A}|x_i, y_i)$ (Pfeffermann and Sverchkov (1999)) can be estimated regressing $w_{i,A}$ against $(x_i, y_i)$.
2. $EL_{Obs}^{A \cup B}$ needs to be maximized with regard to $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ under the constraints:

$$
\begin{aligned}
p_k^X &\geqslant 0, p_i^{Y|X} \geqslant 0, p_i^{Z|X} \geqslant 0, \\
\sum_{k=1}^{K} p_k^X &= 1, \sum_{j \in A_{xk}} p_i^{Y|X} = 1, \sum_{j \in B_{xk}} p_i^{Z|X} = 1, \sum_{k=1}^{K} p_k^X x_k = \mu_X
\end{aligned}
$$

# EL under nonignorable sampling and nonresponse

Additionally to informative sampling, $A$ and $B$ are subject to NMAR nonresponse.
The empirical respondents likelihood (ERL) for the sample $A \cup B$

$$ERL_{Obs}^{A \cup B} = \prod_{k=1}^{K} (p_{k,R_A}^{X})^{r_{k,A}} \prod_{i \in R_{k,A}} p_{i,R_A}^{Y|X} \prod_{k=1}^{K} (p_{k,R_B}^{X})^{r_{k,B}} \prod_{i \in R_{k,B}} p_{i,R_B}^{Z|X}$$

where $R_{k,A} = \{i \in R_A : x_i = x_k\}$ and $R_A$ is the set of responding units of size $r_A$.

# EL under nonignorable sampling and nonresponse

The respondent models are

$$
\begin{aligned}
p_{i,R_A}^{Y|X} &= \frac{P(R_i^A = 1|x_k, y_i, I_i^A = 1)}{P(R_i^A = 1|x_k, I_i^A = 1)} \frac{P(I_i^A = 1|x_k, y_i)}{P(I_i^A = 1|x_k)} p_i^{Y|X} \\
p_{k,R_A}^{X} &= \frac{P(R_i^A = 1|x_k, I_i^A = 1)}{P(R_i^A = 1|I_i^A = 1)} \frac{P(I_i^A = 1|x_k)}{P(I_i^A = 1)} p_k^X
\end{aligned}
$$

where $R_i^A$ is the response indicator. Analogous expressions can be obtained from sample $B$.

# EL under nonignorable sampling and nonresponse

1. the response is independent of the sample selection, the $E_A(w_{i,A}|x_i, y_i) = E_{R_A}(w_{i,A}|x_i, y_i)$, then the probabilities $P(I_i^A = 1|x_i, y_i)$ can be estimated by regressing $w_{i,A}$ against $(x_i, y_i)$, using the observed data in $A$.
2. The response probabilities need to be estimated from the available data by a parametric model.

$$P(R_i^A = 1|x_i, y_i, I_i^A = 1) = g_A(\gamma_{0,A} + \gamma_{x,A}x_i + \gamma_{y,A}y_i)$$

for some functions $g_A$ (logit function), with unknown parameters $\gamma_A$.

Analogous expression for sample $B$ with response parameters $\gamma_B$.

# EL under nonignorable sampling and nonresponse

The ERL needs to be maximized with respect $(\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}, \gamma_A, \gamma_B)$ under the constraints

$$\begin{aligned} p_k^X &\geqslant 0, p_i^{Y|X} \geqslant 0, p_i^{Z|X} \geqslant 0, \\ \sum_{k=1}^{K} p_k^X &= 1, \sum_{j \in R_{A,k}} p_j^{Y|X} = 1, \sum_{j \in R_{B,k}} p_j^{Z|X} = 1, \sum_{k=1}^{K} p_k^X x_k = \mu_X \end{aligned}$$

for all $k$ and $i$.

# EL under non-ignorable sampling and non-response

Estimate the probabilities $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ by the profile likelihood:

Step 1 initial estimates of $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ and maximize the profile likelihood function,

$$G(\gamma_A, \gamma_B) = ERL_{Obs}^{A \cup B}(\gamma_A, \gamma_B | p_k^X, p_i^{Y|X}, p_i^{Z|X})$$

with respect to $(\gamma_A, \gamma_B)$.

Step 2 substitute the estimates $\widehat{\gamma}_A, \widehat{\gamma}_B$ in the likelihood and maximized with respect to $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$.

This completes the first iteration in the estimation process. In the second iteration, the estimates of $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ in Step 2 are considered as known and re-estimate the parameters $(\gamma_A, \gamma_B)$ and then the unknown probabilities.

# EL under non-ignorable sampling and nonresponse

The estimates $\widehat{p}^X_{k,A}$, $\widehat{p}^X_{k,B}$ need to be harmonized into a unique estimate.

1. a linear combination of the two estimates;

$$\widehat{p}^X_k = \lambda \widehat{p}^X_{k,A} + (1-\lambda)\widehat{p}^X_{k,B}$$

   with $\lambda \in [0,1]$. For instance $\lambda = n_A/(n_A + n_B)$.
2. taking the value of $\lambda$ minimizing the variance of $\widehat{p}^X_k$. Variance estimates of $\widehat{p}^X_{k,A}$ , $\widehat{p}^X_{k,B}$ can be computed by resampling methods for finite populations.
3. replace $p^X_{k,R_A}$, $p^X_{k,R_B}$ by $\lambda p^X_{k,R_A} + (1-\lambda)p^X_{k,R_B}$ and maximizing the ERL also with respect to $\lambda$.

# Missing data imputation

Once the parameters $\{p_k^X, p_i^{Y|X}, p_i^{Z|X}\}$ have been estimated a fused datasets with joint observations $(x, y, z)$ can be constructed as follows:

Step 1 generate $\tilde{n}$ observations from the estimated distribution of $X$ taking values $(x_i, \ldots, x_K)$ with probabilities $(\widehat{p}_1^X, \ldots, \widehat{p}_K^X)$;

Step 2 for $i = 1, \ldots, \tilde{n}$ and $k = 1, \ldots K$, if $x_i = x_k$ draw at random a value $\tilde{y}$ from the estimated probability function of $\widehat{p}_i^{Y|X}$ taking the values $\Gamma_{k,A} = (y_1^k, \ldots, y_{r_{k,A}^X}^k)$ with probabilities $(\widehat{p}_1^{Y|X}, \ldots, \widehat{p}_{r_{k,A}^X}^{Y|X})$

Step 3 Applied the procedure in Step 2 for drawing values $\tilde{z}$ from the estimated probability function of $\widehat{p}_i^{Z|X}$.

# Uncertainty in Statistical matching

According to *Conti, Marella and Scanu* (2016), we drop the CIA and we proceed:

1. to define the class of plausible *pdf*s for $(Y, Z)|X$ when no auxiliary information is available and under the constraint $Y \leqslant Z$.
2. to compute an uncertainty measure quantify how broad is the class of plausible models.
3. to choose a plausible *pdf* (a matching distribution) from the class according to the IPF algorithm.

# Uncertainty in Statistical Matching

The estimation of the joint *pdf* of $(X, Y, Z)$ requires the estimation of

1. the marginal *pdf* of $X$;
2. the joint conditional *pdf* of $(Y, Z)|X$.

No auxiliary information is available, the only valid statement is that

$$L_c(F_p(y|x_k), G_p(z|x_k)) \leqslant F_p(y, z|x_k) \leqslant U_c(F_p(y|x_k), G_p(z|x_k))$$

$$\begin{aligned} U_c(F_p(y|x_k), G_p(z|x_k)) &= min(F_p(y|x_k), G_p(z|x_k)) \\ L_c(F_p(y|x_k), G_p(z|x_k)) &= max(0, F_p(y|x_k) + G_p(z|x_k) - 1, ) \end{aligned}$$

The bounds are the Fréchet bounds.

# Uncertainty Measures

For $X = x_k$ a natural uncertainty measure is

$$\Delta_p^k = \int_{R^2} (U_c(\cdot,\cdot) - L_c(\cdot,\cdot)) dF_p(y|x_k) G_p(z|x_k)$$

Weight functions different from $dF_p(y|x_k)G_p(z|x_k)$ can be used instead. In our choice larger weights are assigned to intervals with larger marginal densities.

An overall uncertainty measure is defined by the average of the conditional measures.

$$\Delta_p = \sum_{k=1}^{K} \Delta^k p_k^X$$

# Auxiliary information in Statistical Matching

1. Sample data do not allow to discriminate between alternative models then each distribution in the class can be taken as surrogate of the actual distribution.
2. If auxiliary information is not available the class of plausible distributions is too wide to be usable in practice.
3. We consider external auxiliary information assuming that $Y \leqslant Z$.

# Auxiliary information in Statistical Matching

The class of plausible distributions for $(Y, Z)|X$ is,

$$L_c(F_p(y|x_k), G_p(z|x_k)) \leqslant F_p(y, z|x_k) \leqslant U_c(F_p(y|x_k), G_p(z|x_k))$$

where the Fréchet bounds becomes

$$U_c(F_p(y|x_k), G_p(z|x_k)) = min(F_p(y|x_k), F_p(z|x_k), G_p(z|x_k))$$

$$\begin{aligned} L_c(F_p(y|x_k), G_p(z|x_k)) = & \; max(0, F_p(y|x_k) + G_p(z|x_k) - 1, \\ & min(F_p(y|x_k), F_p(z|x_k)) + G_p(z|x_k) - 1) \end{aligned}$$

# Uncertainty in Statistical matching

The corresponding uncertainty measures under the constraint $\Delta^k_{p,c}, \Delta_{p,c}$ are defined similarly to $\Delta^k_p, \Delta_p$

Final step: to choose a matching distribution from the class by using the IPF algorithm.

The more informative the auxiliary information, the less uncertain the statistical model for the variables of interest becomes. The larger the uncertainty reduction (i.e an uncertainty measure under a given threshold), the more plausible the choice of a matching distribution from the class

# IPF algorithm in the EL context

Step 1 Discretize the variables $Y$ and $Z$ by grouping their ascending values in pre-defined classes. For each $x_k$, the range of the variable $Y$ ($Z$ ) is divided into intervals of equal size $\sqrt{r_{k,A}}(\sqrt{r_{k,B}})$. Denote by $Y_{d,k}(Z_{d,k})$ the discretized variable corresponding to $Y(Z)$ taking $h_k(g_k)$ values defined by the midpoints $y_{d,h}(z_{d,g})$.

Step 2 Estimate the marginal probabilities of $p_h^{Y_{d,k}|x_k}$ as follows

$$\widehat{p}_h^{Y_{d,k}|x_k} = \sum_{i=1}^{r_{k,A}^X} \widehat{p}_i^{Y|x_k}, \widehat{p}_g^{Z_{d,k}|x_k} = \sum_{i=1}^{r_{k,B}^X} \widehat{p}_i^{Z|x_k}$$

where $\widehat{p}_i^{Y|x_k}$, $\widehat{p}_i^{Z|x_k}$ are the estimates obtained by ERL maximization.

# IPF algorithm in the EL context

Step 3 Define the contingency table $C^k$ defined by the $h_k g_k$ values

$$(y_{d,1} z_{d,1}), ..., (y_{d,h} z_{d,g}), ..., (y_{d,h_k} z_{d,g_k})$$

with unknown cell probabilities

$$(p_{11}^{Y_{d,k}, Z_{d,k}|x_k}, ..., p_{hg}^{Y_{d,k}, Z_{d,k}|x_k}, ..., p_{h_k g_k}^{Y_{d,k}, Z_{d,k}|x_k}$$

Step 4 the midpoints $(y_{d,h}, z_{d,g})$ are checked to identify the cells with structural zeroes in $C^k$. The unknown probabilities $p_{hg}^{Y_{d,k}, Z_{d,k}|x_k}$ with $(y_{d,h} > z_{d,g})$ are set equal to zero.

Step 5 Initial values of the cell probabilities $p_{hg}^{Y_{d,k}, Z_{d,k}|x_k}$ are

$$p_{hg}^{0, Y_{d,k}, Z_{d,k}|x_k} = \delta_{hg} \widehat{p}_{h.}^{Y_{d,k}|x_k} \widehat{p}_{.g}^{Z_{d,k}|x_k}$$

where $\delta_{hg} = 1$ for cells in $C^k$ do not contain structural zeroes $\delta_{hg} = 0$ otherwise.

# IPF algorithm in the EL context

The IPF modifies the initial cell probabilities iteratively, until convergence. The final fitted cell probabilities define a matching distribution for $(Y_{d,k}, Z_{d,k})|x_k$. A synthetic datatsets can be reconstructed as follows:

1. Generate $\tilde{n}$ observations $\tilde{x}_i$ from the estimated distribution of $X$ taking the values $(x_1, ..., x_K)$ with probabilities $(\widehat{p}_1^X, ..., \widehat{p}_K^X)$;
2. Let $\tilde{n}_k^X$ be the number of observations with $\tilde{x}_i = x_k$. Draw $\tilde{n}_k^X$ pairs $(\tilde{y}_d, \tilde{z}_d)$ from IPF the matching distribution.

# Application to SHIW and HBS datasets

In Italy, information on households income and expenditure is provided by

1. **SHIW** (Survey on Household Income and Wealth) run by Banca d'Italia;
2. **HBS** (Household Budget Survey) run by ISTAT;

No single data source containing information on both income and expenditure exists. This problem is generally overcome with statistical matching.

# Application to SHIW and HBS datasets

1. SHIW is conducted by Banca d'Italia every two years. Its main goal is to study the economic status of Italian households, focusing on income and wealth. The sample is drawn in two stages, with municipalities as the primary sampling units and households as the secondary sampling units. In the present application, we use the 2010 wave, 7951 households.
2. The HBS collects detailed information on socio-demographic characteristics and expenditures on a disaggregated set of commodities (durable and non-durable). As with SHIW, we use the 2010 wave. The sampling design is similar to SHIW with 22227 houselholds.

# Application to SHIW and HBS datatsets

1. $X$=household size taking the values $x_k = 1, 2, 3, 4+$ with probabilities $p_k^X$.
2. $Y$=household expenditure in HBS;
3. $Z$= household income in SHIW.
4. $\sum_{k=1}^{K} p_k^X x_k = 2.4$ (calibration constraint).
5. $E_A(w_{i,A}|x_i, y_i)$ (regressing $w_{i,A}$ against $(x_i, y_i)$). Same for SHIW (B).
6. the estimates of $p_k^X$ have been harmonized:$\lambda \widehat{p}_{k,A}^X + (1 - \lambda)\widehat{p}_{k,B}^X$, $\lambda = n_A/(n_A + n_B)$.

# Application to SHIW and HBS datasets

The response rates is about of 62% in SHIW and HBS. Nonresponse is explained by,

1. the size of the household: The larger the household, the more possibilities exist to find a contact person for an interview. In addition, households consisting of only one or two elder people, often tend not to participate in surveys.
2. the income (expenditure): as often reported in the literature, the response probability tends to decrease, as the household income or expenditure increase

Response sets $R_A$, $R_B$ are generated by a logistic model setting $\gamma_A = (0.2, -0.002)$ and $\gamma_B = (0.2, -0.003)$.

# Application to SHIW and HBS datasets

First, under the calibration constraint we estimate $p_k^X$

1. $p_k^X$: ISTAT's estimates of the household size distribution in Italy in 2010
2. $\widehat{p}_{k,1C}^X$:estimates obtained ignoring the sampling design effects and assuming that all the units responded;
3. $\widehat{p}_{k,2C}^X$:estimates obtained when accounting for the sampling effects;
4. $\widehat{p}_{k,2CM}^X$ : estimates which account for the sampling designs and nonresponse.

# Application to SHIW and HBS datasets

Tabella: Different estimates of $p_k^X$

| *hsize* | $p_k^X$ | $\widehat{p}_{k,1C}^X$ | $\widehat{p}_{k,2C}^X$ | $\widehat{p}_{k,2CM}^X$ |
|---|---|---|---|---|
| 1 | 0.284 | 0.264 | 0.276 | 0.276 |
| 2 | 0.276 | 0.293 | 0.281 | 0.280 |
| 3 | 0.209 | 0.208 | 0.200 | 0.205 |
| 4+ | 0.232 | 0.233 | 0.243 | 0.239 |

Tabella: Hellinger distance

| $HD(p_k^X, \widehat{p}_{k,1C}^X)$ | $HD(p_k^X, \widehat{p}_{k,2C}^X)$ | $HD(p_k^X, \widehat{p}_{k,2CM}^X)$ |
|---|---|---|
| 1.76% | 1.24% | 0.85% |

# Application to SHIW and HBS datasets

The correlations in HBS and SHIW are: $\rho_{XY} = 0.38, \rho_{XZ} = 0.31$.

Next, we estimate the probabilities $\{p_i^{Y|X}, p_i^{Z|X}\}$ and we generate fused dataset of size 10000:

1. when ignoring the sampling designs and nonresponse: $\rho_{XY} = 0.34$, $\rho_{XZ} = 0.28$, $\rho_{YZ} = 0.08$;
2. when both processes are taken into account: $\rho_{XY} = 0.38$, $\rho_{XZ} = 0.32$, $\rho_{YZ} = 0.13$;

The SHIW questionnaire also contains a section on household expenditures aimed at constructing an approximate measure of total expenditure.

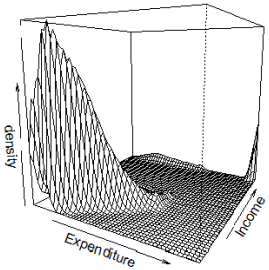
$\rho_{YZ}^{SHIW} = 0.65 \implies \rho_{YZ}^{CIA} = \rho_{XY}^{CIA}\rho_{XZ}^{CIA} = 0.12$ (CIA is inappropriate)

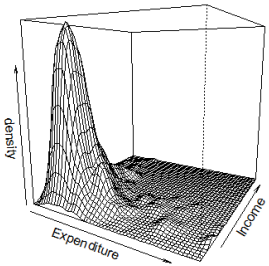
# Uncertainty Measure

1. When accounting for the sampling and nonresponse effects and imposing the constraint $Y \leqslant Z$ , the estimated uncertainty measure $\widehat{\Delta}_p$ decreases from 0.16, (its maximum value when no constraint is used), to 0.11.
2. The uncertainty measure increases to 0.13 when the sampling and nonresponse processes are ignored.

# IPF algorithm in the EL context

Figura: Estimated density of $(Y, Z)$ under the constraint $Y \leqslant Z$ for $h = 3$ . Estimates obtained by IPF (left) and under the CIA (right).



# IPF algorithm in the EL context

The correlation between the imputed values of expenditure and income is 0.55 when applying the IPF.

Our proposed methodology seems to recover pretty well the correlation of 0.65 between income and expenditure in the original SHIW data set.

# References


- Conti, P.L., Marella, D. and Scanu, M. (2016) Statistical matching analysis for complex survey data with applications. JASA, 111, 1715-1725.
- Conti P.L., Marella D., Mecatti F. and Andreis F. (2020). A unified principled framework for resampling based on pseudo-populations: Asymptotic theory. Bernoulli, 26, 2, 1044-1069.
- Marella, D. and Pfeffermann, D. 2019. Matching information from two independent informative sampling. Journal of Statistical Planning and Inference, 203, 70-81.
- Marella, D. and Pfeffermann, D. 2022. Accounting for Non-ignorable Sampling and Non-response in Statistical Matching. International Statistical Review, doi: 10.1111/insr.12524.
- Pfeffermann D. and Sverchkov M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. SankhyÄ, Series B, 61, 166â186.