

## **Potentiality of propensity scores matching in inference from Web surveys: a simulation study**

**Silvia Biffignandi**  
*DMSIA, University of Bergamo*

**Monica Pratesi**  
*DSMAE, University of Pisa*

**QUADERNI DIPARTIMENTO MATEMATICA, STATISTICA,  
INFORMATICA E APPLICAZIONI**

**N. 1    2003**

This paper is financially supported by the Cofin Miur 40% 2001 research grant: "The use of new construction strategies of the data and the statistical information for studying and planning economic systems ", local unit University of Bergamo,

The authors thank professor Donald Rubin for its comments and suggestions.

## **INDEX**

### **1. Introduction**

### **2. Propensity score matching**

- 1. Original context**
- 2. Underlying assumptions**
- 3. The assumptions in the Web survey context**

### **3. The proposed estimation method**

- 1. The Web sample**
- 2. The estimation approaches**
  - Horvitz-Thompson approach**
  - Post-stratification approach**

### **4. The simulation study**

- 1. The original population**
- 2. The propensity scores in the original population**
- 3. matching and sampling distribution**
- 4. The results**
- 5. sensitivity of the results to different population models**

### **APPENDIX: Complete tables of the results**

### **References**

## 1. Introduction

Self-selected Web surveys are very common in many fields of application and despite the large number of respondents they often not resemble the intended target population. Couper (2000) gives an extensive review of Web surveys with their benefits and drawbacks.

In traditional quality terms, the main problems in Web surveys are leading to coverage errors, since the target population mostly is broader than the frame population (active Web users), and it has been difficult - up to now - to construct a correct frame of Web users. In the lack of a good frame, much of the collected data are the result of the responses of a self-selected sample of Web users and inference from such samples are reasonably affected by selection bias.

Various non experimental approaches have been developed in order to correct the self-selection bias. One of these techniques, known as propensity score matching has been recently applied to Web survey data. The subject of this paper is to explore the potentiality of this approach in drawing inference from Web surveys.

The format for the paper is as follows. Section 2 summarizes the original context where the propensity matching was proposed, the assumptions underlying the validity of the approach and the circumstances in which, in the Web survey context, those assumptions can be accepted. In section 3 the proposed estimation method is outlined, describing the generation process of the self-selected sample (Web sample), the intended target population and the proposed estimators. In the section there are also identified the questions that the weighting method can answer and those they cannot answer. The statistical properties of the proposed estimators are studied through a Monte Carlo experiment. Section 4 describes the simulation procedures and its results. Section 5 is devoted to some concluding remarks and to the discussion of the obtained results.

## 2. Propensity score matching

Propensity score matching aims to mimic random assignment of cases and controls, matching cases with post hoc controls on a single index (called propensity) reflecting the probability of being exposed to the treatment. The propensity score matching techniques was developed in the 1980s in the context of biomedical studies (Rosenbaum and Rubin, 1983) and has its roots in a conceptual framework outlined further (Rubin, 1974).

The method rests on several assumptions and has some practical constraints. Nevertheless, it is widely applied in biomedical analyses and in many other research fields. The use of propensity matching is now very common in the labour market policy evaluation. In that context became established in 1990s (Dehija and Wahba, 1999). Recent contributions extent the theory to a multiple ordinal treatment (Imbens, 2000).

In the last three years, the techniques have attracted the attention also as a useful tool for weighting self-selected samples of Web respondents (Taylor, 2000; Terhanian et al, 2001). In other words, the technique has been used to construct a control group for the self-selected respondents in order to reduce the severe bias which affects self-selected samples of Web users.

The results are promising but there is the need of further theoretical and empirical work. There are studies which criticize the application of propensity matching and show that it depends critically on assumptions about the nature of the process by which participants select into the factual group, and the data available to the analyst (Smith and Todd, 2000; Heckman et al, 1998).

## **2.1 The original context**

To illustrate the idea behind the propensity matching, imagine we are interested in measuring the effect of a voluntary medical treatment on the health. To know the effect of the treatment on the health we must compare the observed outcome with the outcome that would have resulted had the person not be exposed to the treatment. However, in non experimental studies only one outcome is actually observed, this is called the factual outcome (case). The counterfactual outcome (control) is that which would have resulted had the treated individual not been treated (or the non treated individual treated). This counterfactual outcome cannot be observed and this is the difficulty of treatment evaluation. The outcome of the factual group should be compared with an appropriate control group, simultaneously controlling for all variables that are thought to affect the comparison. Matching on the propensity score could achieve consistent estimates of the treatment effect in the same way as matching on all covariates.

## **2.2 The underlying assumptions and practical constraints**

The matching method relies on several theoretical assumptions and its use is limited by some practical constraints.

Among the underlying theoretical assumptions the first hypothesis is that the selection in the treatment group can be explained purely in terms of observable characteristics. The assumption is that if one can control for observable differences in characteristics between the treated and non treated group, the outcome that would result in the absence of treatment is the same in both cases (Conditional Independence Assumption or Strong Ignorability).

However, if the Conditional Independence Assumption holds, the matching process is analogous to creating an experimental data set in that, conditional on observed characteristics, the selection process is random.

In addition, the method rest on the assumption that the decision to accept the treatment of an individual does not depend on the decision of others. Further the observed outcome of an individual depends only on the individual and not on the mechanism assigning treatment to individuals.

The main practical constraint of the method is that it requires the availability of data sources on the original population of cases. All the variables affecting the comparison between cases and controls that merit inclusion in the matching process should have been known. The analyst should have time and resources to collect those data, and to do so with minimal measurement error. A chance is

prioritising those variables that according to theory and evidence are most likely to influence the exposition to the treatment.

An additional problem is that is possible that there will be nobody in the non-treatment group with a propensity score that is "similar" to that of a particular treatment group individual (support problem). The solution of the problem is achieved refining the matching criterion or omitting the poorly-matched individuals from the estimation of the treatment effect.

### **2.3 The matching in the Web survey context**

The application of the method in the Web survey context relies on the answers given to the questions that one may ask on the extent to which the previous assumptions can be accepted by survey methodologists.

Indeed, the use of propensity scores to adjust for the different populations has been already tested in other fields where it has been combined also with imputation techniques (Czajka et al. 1987; 1992).

In the Web survey context, the first question to be answered concerns the identification of the treatment and the definition of cases and controls in the target population. Then the concern is about the choice of the observable variables to be included in the matching (the so-called pre-intervention variables) and the value of the conditional independence assumption. In more general terms, the questions refers to the relation between the study variable and the assignment to the treatment.

Further, even in the Web survey context, the application suffers from the same data constraints depicted in the previous subsection.

#### *Identification of treatment and definition of cases and controls*

In our perspective the factual outcome is the study variable values of individuals who participated in the Web survey. The counterfactual outcome are those values which would have resulted had the participating individual not participated in the survey. In other words the treatment is the participation in the Web survey. The not treated units are those who belong to the target population but do not participate into the survey.

#### *The target population*

Actually, in the context of Web surveys the assignment to the treatment is the result of being a Web user in the general population. Not all the potential Web users participate in the survey. Given the low Internet penetration, the aim is to try to correct the bias and to make inference on the general population. The goal is ambitious as the collected data are referred to Web users who participated in the survey.

#### *Pre-intervention variables and conditional independence assumption*

The results of Rubin and Rosenbaum hold when conditionally on the pre-intervention variables there is randomization in the assignment to the categories of cases and controls.

Empirical studies have shown that useful pre-intervention variables in our context are the so-called “Webographic” variables. Sex, age, occupation, frequency of Internet usage, place of connection to Internet have predicted the probability of participating in a Web survey in many applications carried out on general population in different countries (Schonlau et al., 2002; Biffignandi and Pratesi, 2002; Varedian and Forsman, 2002; Pratesi et al, 2003).

#### *Relation between the study variable and the assignment to the treatment*

The method rests on the assumption that the decision to participate of an individual does not depend on the decision of others. This assumption is commonly accepted in many traditional surveys and it can be accepted also in the Web survey context. Further the observed outcome of an individual depends only on the individual and not on the Web data collection mode. This hypothesis corresponds to a specific assumption on the nature of the random self-selection process that originates the Web sample of respondents. In other words, the process is such that the Web respondents are not self-selected on the basis of the  $y$  variable. In the random sample design context, the same concept would be expressed saying that the sampling design is not informative.

#### *Data constraints*

The data needed to estimate propensity scores are many. This is true in terms of the number of variables required to estimate the propensity score (probability of participation) but also in the number of participants and non-participants entering the matching process.

The propensity model for the target population can be estimated with the support of a sample survey or of a statistical register (Bryson A., 2001; Green et al., 2001; Varedian and Forsman, 2002). The main concern is that the same pre-intervention variable could be collected for individuals who participate in the Web survey and for controls who did not participated. The implicit assumption is that pre-intervention variables are collected without error and missing values both for cases and controls, in addition to the study variables.

A common warning is to provide that there is a sufficient number of observations on participants and non-participants across the values of the pre-intervention variables. The possible mismatches on propensity values requires to allow for the possibility of neighbourhood matching where exact matching are unlikely.

### **3. The proposed estimation method**

#### **3.1 The Web sample**

In our perspective the set of Web respondents is the result of a self-selection process that can be described at individual level by the indicator  $R_i$  (1 if the individual participates in the survey, 0 otherwise). The expected value of the indicator is the individual probability of self-selection (or of response)  $E(R_i) = p_i$ .

The size of the set of respondent is a random variable  $n = \sum_i^N R_i$ . In other words, the Web sample is the result of a selection with unknown individual probability  $p_i$  of being included in the set. The inference problem is strictly connected with the

estimation of the probability of being a respondent in the Web survey  $E(R_i) = p_i$ . If this probability would be known, the inference problem could be solved using the traditional tools of estimation from random samples (Särndal et al., 1992). The operational assumption is that this probability is estimated by the propensity score, that reflects the probability of being exposed to the "treatment", the Web survey in our context.

### 3.2 The estimation approaches

The propensity model for the target population can be estimated on data collected by a random sample from the target population. There are examples of applications done using survey data obtained by a telephone survey (Varelian and Forsman, 2002). In this paper we imagine to estimate the propensity model for the target population with the support of a statistical register and not using survey data. In fact, for reasons of survey non-response, or over-sampling of particular subgroups, survey data can not be representative of the whole population. It is usual to apply weights to restore the profile of the survey data to that of the population on a number of key characteristics. The role of these weights is not yet completely explored both in the pre-matching operations, estimating the participation model which generates the propensity scores, and in the post-matching ones, weighting each treatment group member and the corresponding comparator(s) (Bryson A., 2001; Green et al., 2001). For these reasons we prefer to avoid the previous problems imagining to dispose of a complete updated statistical register of the target population. This assumption can be easily accepted in many field of application (Biffignandi and Pratesi, 2002).

The idea is to estimate the individual probability  $p_i$  (propensity score) using a multivariate logistic regression, with the pre-intervention variables as regressors, on the data of the statistical register. The estimated model is then used to estimate the propensity on the Web sample, given the same pre-intervention variables collected for the Web sample as well. After the matching of the Web respondents with not respondent individuals of the target population on the estimated propensity scores, the Web sample data are weighted to estimate the mean, total and standard deviation of the analysis variable  $y$ .

We propose for completeness of our analysis two different propensity based weighting approaches: the first is called "Horvitz and Thompson" approach because it produces an estimator that resembles the traditional Horvitz and Thompson weighted estimator, the second is presented as a "Pseudo Post-stratification" approach to distinguish it from the traditional post stratification where the size of the strata in the population and in the sample are not the result of an estimation process.

In order to provide a term of comparison, the Web sample data have been used also to calculate the unweighted sample moments, and the traditional poststratified estimates of the study variable moments. We used the pre-intervention variables to form post-strata.

#### 3.3.1 The Horvitz-Thompson approach

The probability of being a Web respondent is not known and has to be estimated. It is assimilated to the propensity estimated in the target population,

$\hat{p}_i = P(R_i = 1 | X_i)$ , given the vector  $X_i$  of pre-intervention variables. The  $p_i$  probability can be estimated by the fitting of a generalized linear model for the  $R$  variable on the target population. The model estimated at the population level is used to predict the individual  $p_i$  for the units of the Web sample, given the values of the  $X_i$  collected by the Web survey. The predicted  $\hat{p}_i$  are used to find control units in the target population matched on the same probability  $1 - \hat{p}_i = P(R_i = 0 | X_i)$ .

After the matching, the evidence from the Web sample is directly weighted by the reciprocal of the propensity scores, the weights are  $w_i = 1/\hat{p}_i$ , where  $\hat{p}_i$  estimates the probability of participation. The estimation strategy is that of the Poisson sampling design where the probability of inclusion are not known but estimated by  $\hat{p}_i$ .

The matching on the response probability/propensity score reduces the set of Web respondents to the set of respondents that match with the controls in the target population. This ensures that conditionally on the pre-intervention variables (the predictors in the model for propensity scores) the assumption of conditional independence is verified on the set of respondents used in the estimation process. The proposed estimators of the total, mean and standard deviation of the target population are summarized in table 3.1. The performance of these estimators when the self-selection process originates different samples would be evaluated through a simulation study.

### 3.3.2 The Pseudo Post-stratification approach

At the basis of this approach there is the idea of stratifying the population and the sample according to the frequency distribution of the propensity scores.

Here, the original population is classified into strata according to the frequency distribution of the estimated propensity scores, both for cases and controls.  $N_h$  is the size of a class in the population. After the matching, also the sample is classified in strata according to the frequency distribution of the propensities.  $n_h$  is the size of a class in the sample. The estimation strategy is similar to that of simple random sampling with post-stratification according to one criterion.

After the propensity score matching between cases (Web respondents) and controls (non Web users) in the target population, the same distribution can be constructed on the sample. The size of the target population is indicated by  $N = \sum_h N_h$ .

As in the post-stratification approach, the belonging of the sample units to the strata is determined after the selection of the sample itself.

The proposed estimators of the total, mean and standard deviation of the target population are summarized in table 3.1. The performance of these estimators would be evaluated through a simulation study where the self-selection process originates different samples.

| Table 3.1 The proposed estimators |  |  |
|-----------------------------------|--|--|
| Parameter                         | Estimator  |  |
|                                   | Horvitz-Thompson approach  | Post-stratification approach   |
| mean                              | $\bar{y}_p = \frac{\sum_i y_i w_i}{\sum_i w_i}$                          | $\bar{y}_{ps} = \sum_h \frac{N_h}{N} \sum_i \frac{y_{ih}}{n_h}$                          |
| total                             | $t_p = N \cdot \bar{y}_p$  | $t_{ps} = N \cdot \bar{y}_{ps}$  |
| Standard deviation                | $\sigma(y)_p = \sqrt{\frac{\sum_i (y_i - \bar{y}_p)^2 w_i}{\sum_i w_i}}$ | $\sigma(y)_{ps} = \sqrt{\frac{\sum_h \frac{N_h}{n_h} \sum_i (y_{ih} - \bar{y}_h)^2}{N}}$ |

#### 4. The simulation study

##### 4.1 The original population

The distribution of the pre-intervention variables and of the study variable in the population of Web users and Non Web users was generated as it follows.

| Table 4.1: Original Population |                    |                           |         |
|--------------------------------|--------------------|---------------------------|---------|
|                                | Web users<br>N (%) | Non Web<br>users<br>N (%) | p-value |
| <b>Total population</b>        | <b>260</b>         | <b>1500</b>               |         |
| $x_1$ (Mean±sd)                | 5.87 ± 3.95        | 9.97 ± 5.00               | <.005   |
| $x_2$ (Mean±sd)                | 5.13 ± 2.05        | 3.01 ± 1.73               | <.005   |
| $x_3$ (Mean±sd)                | 4.07 ± 1.12        | 1.01 ± 0.99               | <.005   |

Web users:

$$x_1 = N(6,4), x_2 = Gamma(5),$$

$$x_3 = \exp(-t) + 3$$

Non Web users:

$$x_1 = N(10,5), x_2 = Gamma(3),$$

$$x_3 = \exp(-t)$$

The study variable is a linear function of the observed variables:

$$y = 0.2x_1 + 0.3x_2 + 0.5x_3 + 10u$$

where  $u$  has a uniform distribution

in the interval  $[0,1]$

Table 4.1 describes the original population and contains all covariates (pre-intervention variables) that were used in the multivariate logistic regression model to create the propensity score. Differences between groups were evaluated using

the t- test for continuous data. For every covariate, there was a significant difference between the Web users and non Web users ( $p < 0.05$ ).

#### 4.2 The propensity scores in the original population

The model for the probability of being a Web user in the original population is a logistic regression model

$$p_i = P(R_i = 1 | X_i) = \exp(X_i\beta) / (1 + \exp(X_i\beta))$$

where  $X_i$  is the individual vector of pre-intervention variables. The results of the fitting are shown in Table 4.2.1, 4.2.2, 4.2.3.

| Table 4.2.1: Analysis of maximum likelihood estimates |    |          |                |            |           |
|---|----|----------|----------------|------------|-----------|
| Parameter   | DF | Estimate | Standard Error | Chi-Square | Pr> ChiSq |
| Intercept   | 1  | -7.4672  | 0.5711         | 170.9650   | <.0001    |
| $x_1$   | 1  | -0.1653  | 0.0295         | 31.3925    | <.0001    |
| $x_2$   | 1  | 0.5050   | 0.0661         | 58.2998    | <.0001    |
| $x_3$   | 1  | 1.9512   | 0.1329         | 215.6515   | <.0001    |

| Table 4.2.2: Odds Ratios estimates |                |                            |       |
|------------------------------------|----------------|----------------------------|-------|
| Effect                             | Point Estimate | 95% Wald Confidence Limits |       |
| $x_1$                              | 0.848          | 0.800                      | 0.898 |
| $x_2$                              | 1.657          | 1.456                      | 1.886 |
| $x_3$                              | 7.037          | 5.424                      | 9.130 |

| Table 4.2.3: Association of Predicted Probabilities and Observed Responses |        |          |       |
|--|--------|----------|-------|
| Percent Concordant   | 98.3   | Somers'D | 0.967 |
| Percent Discordant   | 1.7    | Gamma    | 0.967 |
| Percent Tied   | 0.0    | Tau-a    | 0.244 |
| Pairs  | 390000 | c        | 0.983 |

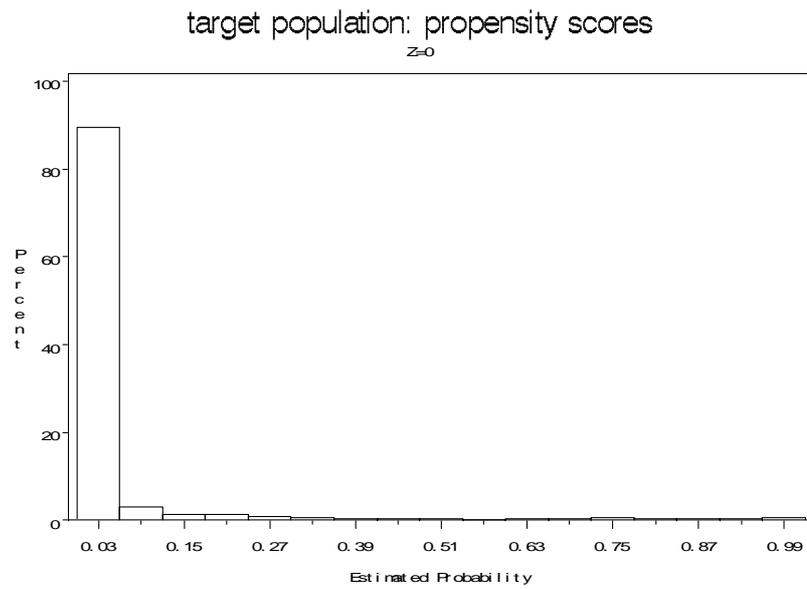
The model for the probability of being a Web user in the original population was used to simulate the population of Web navigators exposed to the risk of compiling

a Web questionnaire. The underlying hypotheses is that, in our target population, the probability of being a Web user is a proxy of the probability of participating in a Web survey after a random contact during a navigation in Internet.

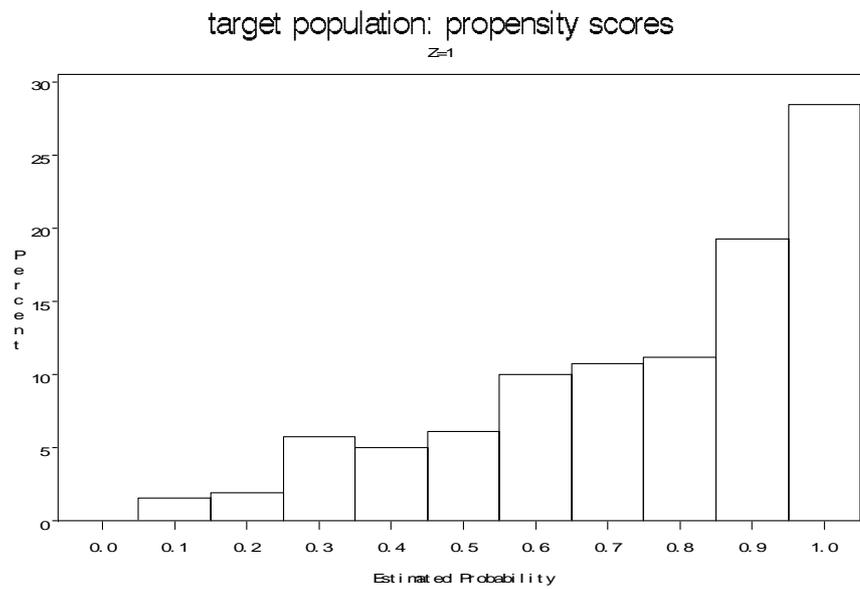
As displayed in Figure 4.2.1, in Figure 4.2.2 and in Figure 4.2.3 (see also Table 4.2.4), user and non users groups may contain a disjoint range of propensity scores in the original population. In the example data, the minimum and maximum propensity score for users was 0.0884221836 and 0.9999973058. For non users, the minimum and maximum propensity score was 0.00002052 and 0.9972057. For the derived population of Web navigators the minimum and maximum propensity score was 0.135204572 and 0.999997306.

| <b>Table 4.2.4: Propensity scores distributions</b> |               |                                   |  |                                    |
|---|---------------|-----------------------------------|--|------------------------------------|
| <b>Quantile</b>                                     |               | <b>Estimate for cases (users)</b> | <b>Estimate for controls (non users)</b> | <b>Estimate for Web navigators</b> |
| 100%  | <b>Max</b>    | 0.9999973058                      | 9.97205732E-01                           | 0.999997306                        |
| 99%   |               | 0.9999575948                      | 8.67063697E-01                           | 0.999957595                        |
| 95%   |               | 0.9993335547                      | 2.37533440E-01                           | 0.999362133                        |
| 90%   |               | 0.9957095590                      | 7.12136611E-02                           | 0.997085980                        |
| 75%   | <b>Q3</b>     | 0.9601288431                      | 1.15225162E-02                           | 0.977122233                        |
| 50%   | <b>Median</b> | 0.8255577743                      | 2.35339270E-03                           | 0.904642828                        |
| 25%   | <b>Q1</b>     | 0.5913303611                      | 7.04062477E-04                           | 0.733164775                        |
| 10%   |               | 0.3709484250                      | 2.96648265E-04                           | 0.535257958                        |
| 5%  |               | 0.2802714501                      | 1.86986852E-04                           | 0.432110438                        |
| 1%  |               | 0.1239299145                      | 6.47313779E-05                           | 0.250451853                        |
| 0%  | <b>Min</b>    | 0.0884221836                      | 2.05253916E-05                           | 0.135204572                        |

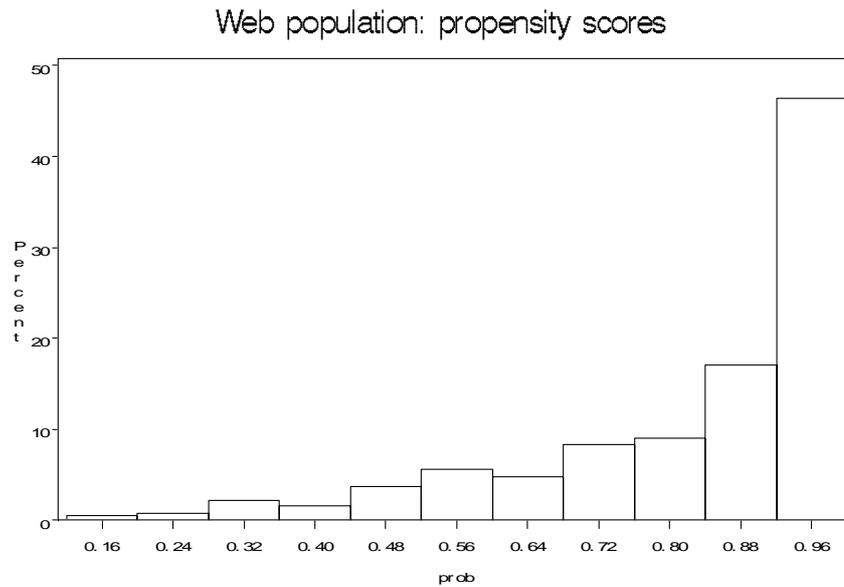
The propensity scores distributions for the original population and the Web navigators (Web population) are displayed in Table 4.2.4. Incomplete matching will result when the Web population members (users or navigators) are considered as cases: the navigators with the highest propensity score (> 0.999 in the example data) and the controls with the lowest propensity score (<0.1352 in the example data) will be excluded.



**Figure 4.2.1 : Estimated propensity scores for Non Web users**



**Figure 4.2.2 : Estimated propensity scores for Web users**



**Figure 4.2.3 : Estimated propensity scores for Web navigators**

### 4.3. Matching and sampling distribution

The self-selection process that generated the sample of Web respondents was simulated selecting from the Web population a sample of Web navigators by a Poisson random sampling design with individual probability equal to the individual propensity score.

The Web navigators of the sample were first matched to controls in the target population on 5 digits of the propensity score. For those that did not match, cases were then matched to controls on 4 digits of the propensity score. This continued down to a 1-digit match on propensity score for those that remained unmatched (Parsons, 2001).

In order to explore the behaviour of the respondents in a set of the possible self-selected samples, the selection process was repeated 100 times: for each sample of size  $n=20$  the matching procedure was repeated as well.

Several variations of the same matching algorithm were performed and are described in Table 4.3.1. The completeness of match, goodness of matched sample and goodness of matched pairs are described in the table summarizing the results of the 100 samples ( $n=20$ ). For the goodness of matched sample, only the mean of  $x_1$  of the cases and controls are shown in Table 3. The reason for this is to show that the goodness of the matched pairs decreases as the completeness of the match improves. This can be noted from the differences in the cases and control means. Imposing a matching on an higher number of digits the mean absolute difference in propensity score of matched pairs decreases.

The sampling distribution of the moments of the propensity scores associated to the sampling units are summarized in table 4.3.4 where the total  $t = \sum_{i=1}^{20} p_i$ , the mean

$\bar{p} = \sum_{i=1}^{20} p_i / 20$  and the standard deviation  $\sigma(p)$  are all referred to the sample of  $n = 20$  units.

The results of section 4.4 are based on the units matched by the 1-Digit Match Algorithm.

| <b>Table 4.3.1: Summary of Matching</b> |                       |                            |                               |  |
|---|-----------------------|----------------------------|-------------------------------|--|
|   | Completeness of Match | Goodness of Matched Sample |                               | Goodness of Matched Pairs                                |
| Algorithm                               | Cases n (% Matched)   | Cases Mean $x_1$ (Mean±sd) | Controls Mean $x_1$ (Mean±sd) | Absolute Difference in Propensity Score of Matched Pairs |
| 3-Digit Match                           | (40%)                 | 5.54± 2.70                 | 6.95 ± 3.05                   | .000813  |
| 2-Digit Match                           | (80%)                 | 5.56± 1.53                 | 7.34 ± 1.48                   | .0014  |
| 1-Digit Match                           | (90%)                 | 5.22 ± 1.45                | 7.42 ± 1.28                   | .0232  |
| <b>population</b>                       |                       | <b>5.87 ± 3.95</b>         | <b>9.97 ± 5.00</b>            |  |

| <b>Table 4.3.2: Moments of the propensity score distribution</b> |            |           |
|--|------------|-----------|
| Variable   | Mean       | Std dev   |
| $t$  | 16.2501990 | 0.8775973 |
| $\bar{p}$  | 0.8320568  | 0.0342385 |
| $\sigma(p)$  | 0.1833420  | 0.0382348 |
| $n_m$  | 19.5400000 | 0.9036402 |

#### 4.4 The results

To illustrate the performance of the methodology relative to other traditional adjustment methods (no adjustment and post-stratification of sample) we summarize a few simulation results. We focus on the estimation of the population total, mean and standard deviation using estimates based on the estimators presented in Table 3.1 under the self-selection process described in section 4.3.

The results are discussed on the basis of the empirical relative percentage bias of the estimators  $(T - \theta) / \sqrt{MSE(T)}$ , where  $MSE(T)$  is the empirical mean squared error of the estimator  $T$  and  $\theta$  is the parameter of interest. The complete set of the results obtained for the original population is presented in Table A1-A4 in the Appendix.

The results are presented in Table 4.4.2 for the two possible target populations: the whole set of individuals, Web users and non Web users and the population of Web navigators (see Table 4.4.1).

| <b>Table 4.4.1: Parameters of the target population</b> |                             |             |                |
|---|-----------------------------|-------------|----------------|
| <i>Target</i>   | <b>Analysis variable: y</b> |             |                |
|   | <i>Sum</i>                  | <i>Mean</i> | <i>Std Dev</i> |
| <i>Whole population</i>                                 | 15001.03                    | 8.52        | 3.20           |
| <i>Web population</i>                                   | 2625.68                     | 10.10       | 3.05           |

To illustrate our results we focus firstly on the performances achieved for the parameter of the whole population relative to the Web population. The relative bias is clearly lower when the target is the Web population but the performance of the proposed estimator is not poor also for the whole population especially for the estimation of the total (R.B.=0,02 versus R.B.=-0.03 for the H.T. estimator and R.B.=0.01 versus R.B.=-0.02 for the Pseudo PS estimator). The estimation of the standard error has the same level of relative bias both for the whole and the Web population.

The second perspective used to illustrate the results is the comparison of the performance of our estimators with that obtained on the same samples without using weights or post-stratifying according to the pre-intervention variables. The strata (two for each variable, stratum 1: individuals above the average, stratum 2: individuals beneath the average) are obtained cross-tabulating the variables both on population and sample. The main result is the good behaviour of the Pseudo post estimator that has a lower level of the relative percentage bias in comparison with the traditional adjustment. The H.T. estimator has not the same performance mainly because of the average high level of the propensity score in the self-selected sample: this produce a low inflation of the H.T estimator and a consequent relative bias that is on line with that of the no weight solution.

| <b>Table 4.4.2: Relative percentage bias of the proposed estimators</b> |                                 |                  |              |                    |                               |                  |              |                    |
|---|---------------------------------|------------------|--------------|--------------------|-------------------------------|------------------|--------------|--------------------|
| <i>Web sample Estimators</i>  | <b>Target: whole population</b> |                  |              |                    | <b>Target: Web population</b> |                  |              |                    |
|   | <i>No weights</i>               | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> | <i>No weights</i>             | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> |
| $t_{ps}$  | 0,03                            | 0,02             | 0,02         | 0,01               | 0,03                          | -26,85           | -0,03        | -0,02              |
| $\bar{y}_{ps}$  | 43,57                           | 41,53            | 40,34        | 14,54              | -6,67                         | -20,78           | -8,61        | -4,37              |
| $\sigma(y)_{ps}$  | -38,09                          | -83,33           | -79,31       | -62,24             | 33,33                         | -23,81           | -32,00       | -62,18             |

#### 4.4 Sensitivity of the results to different population models

The relative performance of the adjustment methods studied here seems to depend on the model that generates the target population: population described in section 4.1 was chosen to represent a situation where the relation between the study variable and the pre-intervention variables is linear and the distribution of the pre-intervention variables is the same (with different moments) for Web users and non Web users.

Here we remove first the assumption of linear relation considering the study variable as a non linear function of the observed variables:

$$y = \frac{0.2}{x_1} + 0.3x_2^2 + 0.5x_3 + 10u$$

where  $u$  has a uniform distribution in the interval [0,1] and the pre-intervention variables are distributed as in the population of section 4.1. We call this model Population 1.

Then we chose different distribution models for the pre-intervention variables for Web users and non Web users, considering again the study variable as a linear function of the observed variables. We call this model Population 2:

$$y = 0.2x_1 + 0.3x_2 + 0.5x_3 + 10u$$

Web users:

$$x_1 = N(6,4), x_2 = \text{Gamma}(5), x_3 = \exp(-t) + 3$$

Non Web users:

$$x_1 = \log(N(10,5)), x_2 = \text{Gamma}(3), x_3 = \exp(-t)$$

The results are presented in Tables 4.4.1 and 4.4.2 considering the whole population as the target.

The pseudo post estimator obtains better results also changing the population model for the study variable: the relative bias of the pseudo post estimator is always lower than the relative bias of the HT estimator for all the descriptive parameters of the population (see Table 4.4.1).

| <b>Table 4.4.1: Relative percentage bias of the proposed estimators</b> |                            |                    |                     |                    |
|---|----------------------------|--------------------|---------------------|--------------------|
| <b>Target: whole population</b>   |                            |                    |                     |                    |
| <i>Web sample Estimators</i>  | <b>Original population</b> |                    | <b>Population 1</b> |                    |
|   | <i>H. T.</i>               | <i>Pseudo post</i> | <i>H. T.</i>        | <i>Pseudo post</i> |
| $t_{ps}$  | 0,02                       | 0,01               | 0,02                | 0,01               |
| $\bar{y}_{ps}$  | 40,34                      | 14,54              | 38,46               | 22,83              |
| $\sigma(y)_{ps}$  | -79,31                     | -62,24             | -26,32              | -21,28             |

The performance of the Pseudo post estimator is not completely confirmed when, in the linear model for the study variable, the pre-intervention variables have different distributions in the sub-populations of Web users and non Web users. In this case the differences in the performance of the pseudo post estimator and the HT estimator are not so relevant (see Table 4.4.2).

| <b>Table 4.4.2: Relative percentage bias of the proposed estimators</b> |                            |                    |                     |                    |
|---|----------------------------|--------------------|---------------------|--------------------|
| <b>Target: whole population</b>   |                            |                    |                     |                    |
| <i>Web sample Estimators</i>  | <b>Original population</b> |                    | <b>Population 2</b> |                    |
|   | <i>H. T.</i>               | <i>Pseudo post</i> | <i>H. T.</i>        | <i>Pseudo post</i> |
| $t_{ps}$  | 0,02                       | 0,01               | 0,008               | 0,008              |
| $\bar{y}_{ps}$  | 40,34                      | 14,54              | 14,44               | 13,97              |
| $\sigma(y)_{ps}$  | -79,31                     | -62,24             | 21,17               | 22,75              |

#### 4. Concluding remarks and future work

In the paper the problem of drawing inference from self-selected samples of Web users has been approached through propensity matching. The goal of the paper has been to explore the potentiality of this technique in order to correct the self-selection bias that often affects Web surveys results.

The focus has been on weighting the Web survey data with adjustment weights obtained after the matching of the Web respondents with the adequate target population members. The matching has been done by the propensity scores estimated by a logistic regression. The propensity matching has been combined with two estimation strategies of several descriptive parameters of the target population: we focused on the "Horvitz-Thompson" strategy and on the "Pseudo Post-stratification" strategy. This last strategy has shown a better performance than the other one both in the estimation of every descriptive parameter and under the three different simulated target populations.

The relative performance of the adjustment methods studied here seems to be sensitive to the model that generates the target population, but the performance of the pseudo post estimator gives always satisfactory results when the target is the whole population.

The results suggests that when using propensity scores in this manner, estimated by logistic regression, the preference is to use sub classification of propensity scores (or post-stratification) rather than Horvitz-Thompson weighting.

In the presented results the Horvitz-Thompson weighting is not enough to compensate for the self-selection bias mainly because of the average high level of the propensity score in the self-selected sample: this produce a low inflation of the H.T estimator and a consequent relative bias that is on line with that of the no weight solution. In addition, another problem can arise: if no further adjustments are to be made, the inverse estimated propensity scores can be unduly huge because of the method of their estimation, which is essentially iteratively reweighted least squares, and can lead to very small and unstable estimated probabilities; post-stratification instead creates discrete categories of weights, resulting in stable adjustments, unaffected by very small estimated propensity scores.

In conclusion, our results show that the propensity matching combined with the pseudo post stratification strategy is a promising solution to the adjustment of the self-selection bias of Web surveys. Future research has to be devoted to implement further adjustments to the propensity weighting. These could be made based on imputation and these adjustments are typically better when applied within the post-strata rather than as a supplement using the estimated propensity scores themselves.

#### References

- Biffignandi S., Pratesi M., (2002), "Internet surveys: the role of time in italian firms response behaviour", *Journal of Research in Official Statistics*, Volume 5, number 2.
- Bryson A. (2001), "The Union Membership Wage Premium: An Analysis Using Propensity Score Matching, Centre for Economic Performance Working Paper n. 1160, London School of Economics.

Couper M. P. (2000), "Web Surveys, A review of Issues and Approaches", *Public Opinion Quarterly*, vol. 64, 464-494.

Czajka J. L., Hirabayashi S. M., Little R. J. A., Rubin D. (1987), "Evaluation of a New Procedure for Estimating Income and Tax Aggregates from Advance Data", *Statistics of Income and Related Administrative Record Research: 1986-1987*. Dept of Treasury, IRS-SOI. Pp. 109-136.

Czajka J. L., Hirabayashi S. M., Little R. J. A., Rubin D. (1992), "Projecting from Advance Data Using Propensity Modelling", *The Journal of Business and Economics Statistics*, vol. 10, 2, pp.117-131.

Dehejia R. H., Wahba S. (1999), "Causal Effects in Nonexperimental Studies: reevaluating the evaluation of training programs", *Journal of the American Statistical Association*, vol. 94, n.448, pp.1053-1062

Green H., Connolly H., Marsh A., Bryson A. (2001), "The Longer-term Effects of Voluntary Participation in ONE", Department of Work and Pensions, Research report Number 149.

Heckman J., Ichimura H., Smith J., Todd P. (1998), "Characterizing selection bias using experimental data", *Econometrica*, 66(5), pp. 1017-1098.

Imbens G. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions", *Biometrika*, vol. 87, pp.706-710.

Parsons LS. (2001) "Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques". Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference, Cary, NC: SAS Institute Inc..

Pratesi M., Lozar Manfreda K., Biffignandi S., Vehovar V. (2003), List based Web Surveys: qualità, timeliness and non-response in the steps of the participation flow, *Journal of Official Statistics*, *forthcoming*.

Rosenbaum, P.R, and D.B. Rubin, (1983), "The Central Role of the Propensity Score in Observational studies for Causal Effect", *Biometrika*, vol 70, pp. 41-55.

Rubin, D. and E. Zanutto (2001), Using matched substitutes to adjust for nonignorable nonresponse through multiple imputation, in *Survey Nonresponse* (R.Groves, D. Dillman, J. Eltinge and R. Little, eds). New York John Wiley, Chapter 26, pp.389-402.

Särndal C. E., Swensson B., Wretman J. (1992), *Model assisted survey sampling*, Springer Verlag.

Smith J., Todd P. (2000), "Does matching overcome LaLonde critique of nonexperimental estimators?", mimeo, downloadable from <http://www.bsos.umd.edu/econ/jsmith/Papers.html>.

Schonlau, M., Fricker R., Elliott, M. (2002), "Evaluation of Web survey Methodology", RAND, Santa Monica, CA. 2002.

Taylor H., (2000), "Does Internet research work? Comparing Online survey Result with telephone Survey" *International Journal of Market Research*, 42 (1) 58-63.

Terhanian G., R. Smith, J. Bremer, R.K. Thomas, (2001), "Exploiting Analytical Advances: Minimizing the biases Associated with Internet-Based Surveys of Non – Random Samples", ARF/ESOMAR: Worldwide Online Measurement, ESOMAR Publication Services, vol. 248, pp. 247-272.

Varedian M., Forsman G. (2002) "Comparing propensity score weighting with other weighting methods: A case study on Web data", *unpublished report*

**APPENDIX: Complete tables of the results**

| <b>Table A1: Horvitz- Thompson approach</b> |                              |                |                                 |             |                               |            |
|---|------------------------------|----------------|---------------------------------|-------------|-------------------------------|------------|
| <i>Propensity weighted Estimators</i>       | <b>Sampling distribution</b> |                | <b>Target: whole population</b> |             | <b>Target: Web population</b> |            |
|   | <i>Mean</i>                  | <i>Std Dev</i> | <i>Bias</i>                     | <i>MSE</i>  | <i>Bias</i>                   | <i>MSE</i> |
| $t_p$                                       | 17540.01                     | 2163.45        | 2538.98                         | 11080118.91 |                               |            |
|   | 2591.14                      | 319.60         |                                 |             | -34.54                        | 102316.37  |
| $\bar{y}_p$                                 | 9.96                         | 1.23           | 1.44                            | 3.57        | -0.13                         | 1.51       |
| $\sigma(y)_p$                               | 2.97                         | 0.49           | -0.23                           | 0.29        | -0.08                         | 0.25       |

| <b>Table A2: Pseudo Post-stratification approach</b> |                              |                |                                 |             |                               |            |
|--|------------------------------|----------------|---------------------------------|-------------|-------------------------------|------------|
| <i>Propensity weighted Estimators</i>                | <b>Sampling distribution</b> |                | <b>Target: whole population</b> |             | <b>Target: Web population</b> |            |
|  | <i>Mean</i>                  | <i>Std Dev</i> | <i>Bias</i>                     | <i>MSE</i>  | <i>Bias</i>                   | <i>MSE</i> |
| $t_{ps}$   | 17977.92                     | 3676.39        | 2514.81                         | 19909483.69 |                               |            |
|  | 2573.09                      | 556.89         |                                 |             | -52.59                        | 309787.84  |
| $\bar{y}_{ps}$                                       | 9.90                         | 2.14           | 1.37                            | 9.42        | -0.20                         | 4.58       |
| $\sigma(y)_{ps}$                                     | 2.31                         | 0.71           | -0.89                           | 1.43        | -0.74                         | 1.19       |

| <b>Table A3: Web sample estimates without weighting</b> |                              |                |                                 |             |                               |            |
|---|------------------------------|----------------|---------------------------------|-------------|-------------------------------|------------|
| <i>Web sample Estimators</i>                            | <b>Sampling distribution</b> |                | <b>Target: whole population</b> |             | <b>Target: Web population</b> |            |
|   | <i>Mean</i>                  | <i>Std Dev</i> | <i>Bias</i>                     | <i>MSE</i>  | <i>Bias</i>                   | <i>MSE</i> |
| $t_{ps}$  | 17626.41                     | 1930.39        | 2625.37                         | 10581743.30 |                               |            |
|   | 2603.90                      | 285.17         |                                 |             | -21.78                        | 80984.19   |
| $\bar{y}_{ps}$  | 10.02                        | 1.10           | 1.49                            | 3.42        | -0.08                         | 1.20       |
| $\sigma(y)_{ps}$  | 3.12                         | 0.45           | -0.08                           | 0.21        | 0.07                          | 0.21       |

| <b>Table A4: Post stratified Web sample estimates</b> |                              |                |                                 |            |                               |            |
|---|------------------------------|----------------|---------------------------------|------------|-------------------------------|------------|
| <i>Web sample Estimators</i>                          | <b>Sampling distribution</b> |                | <b>Target: whole population</b> |            | <b>Target: Web population</b> |            |
|   | <i>Mean</i>                  | <i>Std Dev</i> | <i>Bias</i>                     | <i>MSE</i> | <i>Bias</i>                   | <i>MSE</i> |
| $t_{ps}$  | 17204.14                     | 2121.86        | 2203.10                         | 9310956.19 |                               |            |
|   | 2541.52                      | 313.46         |                                 |            | -84.16                        | 313.46     |
| $\bar{y}_{ps}$  | 9.78                         | 1.21           | 1.25                            | 3.01       | -0.32                         | 1.54       |
| $\sigma(y)_{ps}$                                      | 3.00                         | 0.45           | -0.20                           | 0.24       | -0.05                         | 0.21       |

| <b>Table A5: Relative percentage bias of the proposed estimators<br/>Target: whole population</b> |                            |                  |              |                    |                     |                  |              |                    |
|---|----------------------------|------------------|--------------|--------------------|---------------------|------------------|--------------|--------------------|
| <i>Web sample Estimators</i>  | <b>Original population</b> |                  |              |                    | <b>Population 1</b> |                  |              |                    |
|   | <i>No weights</i>          | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> | <i>No weights</i>   | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> |
| $t_{ps}$  | 0,03                       | 0,02             | 0,02         | 0,01               | 0,02                | 0,02             | 0,02         | 0,01               |
| $\bar{y}_{ps}$  | 43,57                      | 41,53            | 40,34        | 14,54              | 40,72               | 39,11            | 38,46        | 22,83              |
| $\sigma(y)_{ps}$  | -38,09                     | -83,33           | -79,31       | -62,24             | -26,32              | -25,44           | -26,32       | -21,28             |

| <b>Table A6: Relative percentage bias of the proposed estimators<br/>Target: whole population</b> |                            |                  |              |                    |                     |                  |              |                    |
|---|----------------------------|------------------|--------------|--------------------|---------------------|------------------|--------------|--------------------|
| <i>Web sample Estimators</i>  | <b>Original population</b> |                  |              |                    | <b>Population 2</b> |                  |              |                    |
|   | <i>No weights</i>          | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> | <i>No weights</i>   | <i>Trad post</i> | <i>H. T.</i> | <i>Pseudo post</i> |
| $t_{ps}$  | 0,03                       | 0,02             | 0,02         | 0,01               | 0,05                | 0,05             | 0,008        | 0,008              |
| $\bar{y}_{ps}$  | 43,57                      | 41,53            | 40,34        | 14,54              | 9,03                | 9,23             | 14,44        | 13,97              |
| $\sigma(y)_{ps}$  | -38,09                     | -83,33           | -79,31       | -62,24             | 11,76               | 12,92            | 21,17        | 22,75              |