

# Using auxiliary information and non parametric methods in weighting adjustments

Emilia Rocco

**Abstract** Weighting adjustments are commonly used in survey estimation to compensate for unequal selection probabilities, nonresponse, non coverage, and sampling fluctuations from known population values. Over time many weighting methods have been proposed, mainly in the nonresponse framework. These methods generally make use of auxiliary variables to reduce the bias of the estimators and improve their efficiency. Frequently, a substantial amount of auxiliary information is available and the choice of the auxiliary variables and of the way in which they are employed may be significant. Moreover, the efficacy of weighting adjustments is often seen as a bias-variance trade off. In this paper, we analyze this aspects of the nonresponse weighting adjustments and investigate the properties of mean estimators adjusted by individual response probabilities estimated through non parametric methods in situations where multiple covariates are both categorical and continuous.

**Key words:** Bias-variance trade-off, Nonresponse adjustment, Quasi-randomization inference, Survey nonresponse

## 1 Introduction

Estimation in sample surveys is carried on mainly by attaching weights to observed data and then computing weighting summation. The weights are generally developed in a series of stages to compensate for unequal selection probabilities, nonresponse, non coverage, and sampling fluctuations from known population values. At the first stage, a base weight equal to the inverse of the element's selection probability is assigned to each sampled element (respondent or not respondent). The second step of weight

---

Emilia Rocco

Department of Statistics "G. Parenti" - University of Florence - viale Moragni 59, 50134 Firenze,  
e-mail: rocco@ds.unifi.it

development is usually to compensate for unit nonresponse. The base weights of the respondents are increased using estimates of the probabilities that they respond to the survey. The third stage involves a further adjustment to the weights to make the resultant weighted estimates from the sample conform to known population parameters for some auxiliary variables. This type of adjustment aims to reduce variances of the sample estimates and to compensate for noncoverage, but can also be used to compensate for nonresponse. The base weights are usually readily determined from selection probabilities. On the contrary, all their adjustments involve models or assumption of some kind, explicit or implicit, and make use of auxiliary information. As a consequence their efficacy in terms of bias and/or variance reduction depends on the validity of the assumed model and on the choice of auxiliary variables.

In the remainder of the paper we just consider the adjustments for nonresponse. Their essence is to increase the weights of the respondents so that they represent the nonrespondents. The respondent base weights are increased by a multiplying factor given by the reciprocal of the estimated response probabilities. Here the nonresponse is viewed as an additional phase of sampling whose probability mechanism is unobserved and that follows the first phase of sampling, which is determined by the original sample design. A critical step of the weighting adjustment procedure is the estimation of the response probabilities. This step is usually carried on under the assumption that the response mechanism is missing at random (MAR), after conditioning on a set of covariates. Therefore, the response probabilities are estimated under explicit or implicit models relating response occurrences and auxiliary variables. The most common method used to estimate the response probabilities is the weighting within cell method, where respondents and nonrespondents are classified into adjustment cells based on covariate information known for all the sample units, and for all cases in a cell the response weight is calculated as the reciprocal of the response rate in the same cell. Here the unspecified assumption is that sample units in “weighting cells” are homogeneous with respect to the response mechanism. Another popular way to estimate the individual response probabilities is by fitting a parametric model, such as logistic, probit or exponential. An alternative, less known approach to estimate the response probabilities is through kernel regression methods. The kernel type-smoothing methods in the nonresponse context, was first proposed by Giommi (1984) as an alternative to weighting within cell method when no clearly identifiable “weighting cells” exist. This may occur when many auxiliary variables are available or also in the case of a unique continuous auxiliary variable. Giommi (1984) proposed to turn the idea of homogeneity within “weighting cells” into the idea that couples of units with similar or identical value of an auxiliary variable connected with the response mechanism have similar response probabilities. This hypothesis does not fit well the usual cell structure where the weighting cells are non overlapping. The “new” weighting cells are defined in a way allowing overlapping. For each unit in the sample a group of nearest neighbours is identified and the response probability is estimated as proportion of response unit in the group. The resulting estimator of the response function is a running mean of the response indicators but can also be viewed as a simple kernel smoother that uses a kernel uniform function. Its expression corresponds to that of the Nadaraya Watson-estimator. Giommi (1987) extends this estimator using more general Kernel functions. The author used the estimated response probabilities to construct estimators for the population mean according to the Cassel, Särndal & Wretmann (1983) adjusted regression estimator. The response probability estimators suggested by Giommi were also further considered by Niyonsenga (1994, 1997). More

recently Da Silva and Opsomer (2006) extended the results of the previous authors: They used the response probabilities estimated through the Nadaraya-Watson estimator to adjust the Horvitz Thompson mean estimator and the ratio-adjusted mean estimator (Hájek estimator); they derived theoretically the asymptotic properties of these two weighted mean estimators and evaluated empirically their small sample properties. In a subsequent paper (Da Silva and Opsomer, 2008) the same authors extended these results to the estimation of the response probabilities by local polynomial regression.

In this paper, we first discuss for all the three response probability methods quoted above (weighting with cells, parametric estimation, kernel regression estimation) the characteristics that a covariate used in the weighing adjustment process must have so that the weighting adjustments are effective to reduce both bias and variance of the unweighted estimator. Then, the three weighting methods are empirically contrasted for different scenarios, some using a unique continuous covariate and other using two covariates of different type. Most of all, we are interested in evaluating the nonparametric approach in the multiple covariate case since it has not been investigated so far.

The paper is organized as follows: in Section 2, the notation and framework for the sampling design and the nonresponse model are introduced and the estimation of the response probabilities through kernel regression is described. In Section 3, the characteristics that a covariate must have to reduce both bias and variance of the unweighted estimators are discussed. The performances of two kernel-type weighted mean estimators are empirically investigated for different scenarios in Section 4. Section 5 concludes with final comments and ongoing questions.

## 2 Notation and framework for the sampling design and response model

Let  $U$  be a finite population of  $N$  units labelled  $i$  ( $i = 1, \dots, N$ ) and let the population mean  $\bar{y} = \sum_{i \in U} y_i / N$  the parameter of interest, where  $y_i$  is the value of the study variable for unit  $i$ . A sample  $s$  of  $n$  units is drawn from  $U$  according to a sampling design  $p(s)$ ; let  $\pi_i$  be the inclusion probability of unit  $i$  for all  $i \in U$ . When nonresponse occurs the  $y_i$ 's values are observed only on a subset  $r \subseteq s$  of  $n_r$  ( $n_r \leq n$ ) respondents, and to account for the information lost in the estimation of the parameters of interest, it becomes necessary to model the response process. To this end, a response indicator,  $R_i$ , assuming value one if the unit  $i$  responds and zero otherwise, is defined. The distribution of the vector  $(R_i : i \in s)$  is called response mechanism and need to be modelled. We will keep the response mechanism quite general. Precisely we assume that the response mechanism is MAR given a vector of  $q$  auxiliary variables  $\mathbf{x}$  correlated both to nonresponse and to  $y$ , fully observed throughout the sample. Moreover, we assume that, given the sample, the response indicators are independent random variables with:

$$\Pr(R_i = 1 | i \in s, y, \mathbf{x}) = \phi(\mathbf{x}_i) \equiv \phi_i, \quad \text{for all } i \in U \quad (1)$$

where the exact form of  $\phi(\cdot)$  is unspecified, but it is assumed to be a smooth function of  $\mathbf{x}_i$ , with  $\phi(\cdot) \in [0,1]$ . If all the response probabilities were conceptually known, the theory of two phase sampling lead us to use as possible mean estimator:

$$\hat{\bar{y}}_{\pi} = \sum_{i \in S} y_i \pi_i^{-1} \phi_i^{-1} R_i / N \quad (2)$$

or

$$\hat{\bar{y}} = \sum_{i \in S} y_i \pi_i^{-1} \phi_i^{-1} R_i / \sum_{i \in S} \pi_i^{-1} \phi_i^{-1} R_i \quad (3)$$

that are the Horvitz-Thompson and the Hájek estimators adjusted to compensate for nonresponse (Little and Rubin 2002). The second, even if not unbiased, is usually preferred as it is location-scale invariant and generally more efficient. In practice, both formulas in (2) and (3) are unfeasible as the response probabilities are unknown. Then we need to replace  $\phi_i$  with their estimates  $\hat{\phi}_i$ , satisfying  $0 < \hat{\phi}_i \leq 1$ . The resulting weighting adjusted mean estimators are:

$$\hat{\bar{y}}_{\pi} = \sum_{i \in S} y_i \pi_i^{-1} \hat{\phi}_i^{-1} R_i / N \quad (4)$$

or

$$\hat{\bar{y}} = \sum_{i \in S} y_i \pi_i^{-1} \hat{\phi}_i^{-1} R_i / \sum_{i \in S} \pi_i^{-1} \hat{\phi}_i^{-1} R_i \quad (5)$$

In order to implement (4) and (5), it is necessary to estimate the response probabilities  $\phi_i$ . Since the weighting within cells method and the parametric method are well known and extensively present in literature (among others Groves, Dillman, Eltinge and Little (2002), Little and Vartivarian (2005), Kim and Kim (2007)), here we describe only the kernel regression method to estimate the response probability. When auxiliary information consists in a scalar continuous variable  $x$  the Kernel regression estimator (or local constant regression estimator) of response probability assumes the following expression:

$$\hat{\phi}_i = \frac{\sum_s K\left(\frac{X_j - X_i}{h}\right) R_j}{\sum_s K\left(\frac{X_j - X_i}{h}\right)} \quad (6)$$

where  $K(\cdot)$  denotes the kernel function and  $h$  is referred to as the bandwidth. In the originally version of Gijbels (1984)  $K(\cdot)$  is an uniform kernel function and more in detail a function that assigns weight 1 to all the units  $j \in S$  for which  $|x_j - x_i|/h \leq 1$ .

Subsequently, this idea is generalized by using a more general kernel function, but only the case of one continuous auxiliary variable is considered.

A possibility to generalize this approach to the case of more auxiliary variables lies in using a “generalized product kernel function”:

$$K\left(\frac{X_j - X_i}{h}\right) = k_1\left(\frac{x_j - x_i}{h_1}\right) \dots k_q\left(\frac{x_j - x_i}{h_q}\right) \quad (7)$$

i.e. the product of  $q$  univariate kernel function, each one may be a continuous data kernel function or a categorical ordered data kernel function or a categorical unordered data kernel function.

Moreover Da Silva and Opsomer (2008) generalized the kernel regression estimator with the estimation of the response probability by local polynomial regression (general  $p$ th order). In other words, as the estimator in (6) can be obtained as the solution of the following minimization problem :

$$\min_a \sum_{j=1}^n (R_j - a)^2 K\left(\frac{X_j - X_i}{h}\right), \quad (7)$$

they suggested a more general estimator based on minimizing the following expression:

$$\min_{\{b_0, b_1, \dots, b_p\}} \sum_{j=1}^n \left( R_j - b_0 - b_1(X_j - X_i) - \dots - b_p(X_j - X_i)^p \right)^2 K\left(\frac{X_j - X_i}{h}\right) \quad (8)$$

We refer to Da Silva and Opsomer (2008) for the detailed description of this estimator. Its expression is just cumbersome in the case of only one continuous variable. For polynomial of order 1, that is for local linear regression, Li and Racine (2004), in a more general context, suggested the extension to the case of more mixed continuous and discrete variables. The extension is based on the generalized product kernel function and it actually treats the continuous regressors in a local linear fashion and the discrete regressors in a local constant one. We refer to Li and Racine (2004) and Li and Racine (2006, cap.4) for a detailed description. Here, their idea is applied to estimate nonresponse probability in presence of mixed discrete and continuous covariates.

### 3 Weigthing nonresponse adjustments and bias-variance trade-off

A widespread view in the survey sampling literature is that the nonresponse weighting methods aim at reducing nonresponse bias at the expense of an increase in variance. Kalton e Kasprzyk (1986) and Kish (1992) among others, define the effectiveness of nonresponse weighting adjustments as a trade-off between bias reduction and variance increase. Kish also gives a measure of this loss of precision that is  $F = 1 + CV^2(w_i)$ ,

where  $CV^2(w_i)$  is the coefficient of variation of the weights  $w_i = \phi_i^{-1}$ . The measure  $F$  represents the multiplying factor that is applied to the variance of a survey estimate due to the variability in the weights in the situation where equal weights are optimal. According to Little and Vartivarian (2005) this view is an oversimplification, “*nonresponse weighing can in fact lead to a reduction in variance as well as in bias*”. The fact that nonresponse weighting can reduce variance is also implicit in the paper of Oh and Scheuren (1983), and noted in Little (1986) and in Holt and Smith (1979). Some other authors, including Rosebaun (1987), Robins, Rotnitzky and Zhao (1994), Beamount (2005) noted that the weighted adjusted estimator using the estimated response probability can be more efficient than the estimator using the true response probability. This last result is known as “super efficiency”.

Actually the question is related to the characteristics of the auxiliary information used in the estimation of the response probability, that is to its relation with both the nonresponse and the outcome of interest.

With respect to the parametric estimation of response probability, Kim and Kim (2007) theoretically show that adjustments based on response probability estimated using a parametric maximum likelihood method, improve efficiency of point estimation, in addition to reduce bias because they incorporate additional information contained in the auxiliary variables used in the response model and it is generally more efficient than an estimator which uses the true response probability.

For the weighting within cell methods Little and Vartivarian (2005) propose an equation that captures both bias and variance components whether or not the adjustment cell variable is associated with the study variable. From such equation it is evident that:

1. a substantial bias reduction requires adjustment cell variables that are related both to nonresponse and to outcome variables;
2. cell variables related only to the nonresponse increase the variance by the Kish multiplying factor without any reduction in bias;
3. cell variables related only to the study variable tend to have no impact on the bias but reduce variance as far as these variables are good predictors of the outcome variable;
4. If the adjustment cell variables are good predictors of both nonresponse mechanism and outcome variables, weighting tends to reduce both bias and variance.

The negative effect on the variance of the variability of weights is present also in situation described at points 3) and 4) but it is generally small respect to the positive effect of a good predictor of the study variable.

The leading idea is that some of the considerations listed in point 1-4 can be extended to the case of overlapping classes that are the starting point of kernel regression methods. As regards point 1, we can argument as follows: suppose that for each respondent a group of neighbour units is identified on the base of a covariate  $x$ . In this case, if  $x_i$  is the value of  $x$  for the respondent  $i$ , the group of units centred on  $i$  contains all units  $j \in s$  for which  $|x_j - x_i| \leq h$  with  $h$  fixed. Let  $n_i$  and  $n_{ir}$  be the number of sampled and respondent units in that group respectively. Then, assuming that units are selected by simple random sampling, the weighted mean estimator is:

$$\bar{y}_w = \sum_{i \in r} \frac{y_i}{n} \frac{n_i}{n_{ir}} = \sum_{i \in r} \frac{y_i}{n_r} \frac{n_i}{n_{ir}} \frac{n_r}{n} \quad (9)$$

whereas the unweighted (or naive) mean estimator is :

$$\bar{y}_r = \sum_{i \in r} \frac{y_i}{n_r} = \sum_{i \in r} \frac{y_i}{n} \frac{n}{n_r} . \quad (10)$$

The construction of the groups of neighbours respect to variable  $x$  is equivalent to sort the units respect to  $x$  and allows us to consider the terms  $n_{ir}/n_i$  as a running mean of the response indicators. If we assume that the conditional distribution of  $y$  given  $x = x_i$  and  $R_i = 1$  has mean  $\mu_{ir}$ , then, conditional to  $n$ ,  $n_r$  and to the observed distribution of sampled and respondent units among the cells, the mean of  $\bar{y}_w$  and  $\bar{y}_r$  are:

$$E[\bar{y}_w | D] = \sum_{i \in r} \frac{\mu_{ir}}{n} \frac{n_i}{n_{ir}} \text{ and } E[\bar{y}_r | D] = \sum_{i \in r} \frac{\mu_{ir}}{n_r} = \mu_r .$$

where  $D = \{n, n_r, (n_i, n_{ir}) i \in r\}$ . Moreover, neglecting the difference between the weights and their expectation the bias reduction may be expressed as:

$$b(\bar{y}_r) - b(\bar{y}_w) = \sum_{i \in r} \frac{\mu_{ir}}{n_r} \left( 1 - \frac{n_i}{n_{ir}} \frac{n_r}{n} \right)$$

and this difference is zero if either nonresponse is unrelated to the  $x$  (in this case  $n_i/n_{ir} \approx (n_r/n)^{-1}$ ) or  $y$  is unrelated to  $x$  (in this case  $\mu_{ir} \approx \mu_r \forall i$  and as  $\sum_{i \in r} \frac{n_i}{n_{ir}} \approx n$  we have  $b(\bar{y}_r) - b(\bar{y}_w) = \mu_r - \frac{\mu_r}{n} \sum_{i \in r} \frac{n_i}{n_{ir}} \approx 0$ ). Thus, also in the case

of overlapping classes, a substantial bias reduction requires that the variable used to identify the groups of neighbor units is related both to nonresponse mechanism and to outcome variable.

The other points (2-4) refer to the effect of weighting on the variance. Even if the analytical proof is still in progress, it is reasonable to think that also in the case of overlapping classes the effect of weighting on the variance may be measured as the difference between the variance of the unweighted estimator and the variance of the weighted estimator and may be expressed as the sum of two components. The first one is negative deriving from the difference among the expected value of the conditional variance of the two estimators. This therefore, under the assumption of constant variability in the population, is function only of the variability of the weights. The other one is positive deriving from the difference of the variances of the conditional expectations of the two estimators and it increases when the differences between the  $\mu_{ir}$  increase. This is likely to happen when the variable used to identify the groups of neighbor units is a good predictor of the outcome variables. Empirical results, in next section, already show this property: kernel regression estimation method, as well as the

logistic method and weighting within cell method, tends to reduce both bias and variance of the mean if it uses covariates that are good predictors of both nonresponse mechanism and outcome variables.

## 4 Simulation study

Some Monte Carlo experiments have been performed in order to compare the local constant and the local linear estimators of response probabilities with the two more common response probability estimation methods: weighting within cells and logistic regression method. In the first setting a unique continuous covariate  $x_1 \sim \text{Uniform}(0,1)$  is considered: two outcome variables were generate,  $y_1 = 20 + 60x_1 + \varepsilon_1$ , where  $\varepsilon_1 \sim N(0,16)$ , and  $y_2 = 25 + 135(x_1 - 0.5)^3 + \varepsilon_2$ , where  $\varepsilon_2 \sim N(0,6.25)$ . For both population unit nonresponse is generated according to two response functions of the auxiliary variable  $x_1$ , a quadratic function and a logistic function. The two response functions are:

$$\begin{aligned}\phi_1(x_1) &= -\beta_0 x_1^2 + \beta_1 x_1 + \beta_2 \\ \phi_2(x_1) &= \exp(\beta_3 + \beta_4 x_1) / (1 + \exp(\beta_3 + \beta_4 x_1))\end{aligned}\tag{9}$$

where the coefficients  $\beta_0, \dots, \beta_4$ , are chosen so that the response rate is approximately equal to 0.7. In the second setting we consider three different couples of covariates:

1.  $x_1 \sim \text{Uniform}(0,1)$  plus an indicator variable,  $x_2$ ,
2.  $x_1 \sim \text{Uniform}(0,1)$  plus a categorical variable,  $x_3$  with 4 categories,
3.  $x_1 \sim \text{Uniform}(0,1)$  plus a discrete variable,  $x_4$  assuming 4 values.

In case 1 two outcome variables are generated:  $y_3$  as linear function of  $x_1$  with different parameters in the two subpopulation identified by  $x_2$ , and  $y_4$  as quadratic function of  $x_1$  with different parameters in the two subpopulation identified by  $x_2$ . For case 2 an outcome variable  $y_5$  is generated as linear function of  $x_1$  with 4 different sets of parameters corresponding to the 4 categories of  $x_3$ . For case 3 an outcome variable  $y_5$  is generated as linear function of  $x_1 \times x_4$ . For all four populations nonresponse is generated according to two response function  $\phi_3$  and  $\phi_4$  equivalent respectively to  $\phi_1$  and  $\phi_2$  but with different parameters corresponding to the two categories of  $x_2$  or to two groups of categories of  $x_3$  or of  $x_4$ . In all cases the response rate for the whole population is approximately equal to 0.7 and varies at most from 0.65 to 0.75 for the two subpopulation corresponding to different parameters of the response function.

For each combination of population model and response mechanism, the simulation procedure consists in the following steps:

1. select a simple random sample of 300 units (populations' size is  $N = 3000$ );
2. perform a Bernoulli trial for each unit  $i \in s$  with probability  $\phi_i$  ( $i=1,2,3,4$ ) for "success"(response) and  $(1 - \phi_i)$  for "failure" (nonresponse);
3. compute expression (5) on the set of respondents to obtain the following weighted estimators of the mean:
  - $T_0$  : adjusted with true response probabilities;
  - $T_1$  : adjusted with response probabilities estimated through local constant regression;
  - $T_2$  : adjusted with response probabilities estimated through local linear regression;
  - $T_3$  : adjusted with response probabilities estimated through a logistic model;
  - $T_4$  : adjusted with response probabilities estimated by the weighting within cell method;
  - $T_5$  : adjusted with  $\hat{\phi}_i = 1, i \in s$  (naive estimator);
4. repeat steps 1-3 10000 times.

For the first two populations, all the 4 response estimation methods use the unique covariate  $x_1$ . For the other three populations  $y_3$ ,  $y_4$  and  $y_5$  the couple  $(x_1, x_2)$ ,  $(x_1, x_3)$ ,  $(x_1, x_4)$  are respectively used. In the local regression methods  $x_2$  and  $x_3$  are considered as factors, whereas, for  $x_5$  an ordered kernel function is used. For the weighting within cell method the cells are identified in the univariate case by the deciles of the distribution of  $x_1$  and in the bivariate case crossing  $x_1$  (firstly categorized using the quartiles) respectively with  $x_2, x_3$  and  $x_4$ . In the Kernel regression methods other choices in estimator settings concern the kernel functions and the typology and the selection criterion of the bandwidth. Epanechnikov kernel function is used for  $x_1$ , the Aitchison and Aitken kernel function (Aitchison and Aitken, 1976) is used for  $x_2$  and  $x_3$  and Wang and van Ryzin kernel function (Wang and van Ryzin, 1981) is used for  $x_4$ . With respect to the selection criterion of the bandwidth the least-squares cross-validation selection method is.

The experimental results are reported in Tables 1 and 2: Table 1 gives for each combination of population and response mechanism the relative bias that is bias divided by the population mean percent. Table 2 shows the corresponding empirical mean square errors (MSE). It is immediately evident that in all the scenarios each weighting adjustment method reduce successfully the bias and MSE of the unweighted (naïve) estimator.  $T_0$  is not compared in bias with the other weighted estimators, since it is conditionally unbiased for the full sample estimates. In the single covariate scenarios: a)  $T_1$  shows a higher bias than all the other weighted estimators but its MSE is lower than that of  $T_0$ . b) The same MSE behaviour is observed for the other estimators using estimated response probabilities. c) The estimators  $T_2$  and  $T_4$  are almost equivalent both in the bias and in the MSE. d)  $T_3$  shows better bias performance than  $T_2$  and  $T_4$  for response mechanism  $\phi_2$  while it is worse for response mechanism

$\phi_1$  and for both response mechanism it is almost equivalent to  $T_2$  and  $T_4$  in the MSE. For scenarios with two covariates: a) For all populations, when the response mechanism is  $\phi_2$ ,  $T_2$  appears the best either in bias and in MSE. b) When the response mechanism is  $\phi_1$ , in two scenarios  $T_2$  still appears the best either in bias and in MSE, in one scenario the best is  $T_3$  and in another is  $T_4$ . c) For all scenarios the MSE of  $T_2$  and  $T_3$  is lower than that of  $T_0$ , whereas this behaviour is not always valid for  $T_1$  and  $T_4$ .

Another simulation experiment, not reported here for lack of space, considers a situation with two covariate  $z_1$  and  $z_2$  and two outcomes, one function only of  $z_1$  and the other function of  $z_2$  and a response mechanism function of both  $z_1$  and  $z_2$ . If the response probability is estimated using only a covariate, all the weighting estimators works “well” both in bias and in MSE for the outcome variable function of this covariate, but they are more biased and more variable than the unweighted estimator for the other outcome variable.

## 5 Conclusions

Both the theoretical considerations in section 3 and the empirical results in section 4 show that the most important aspect in weighting methods is the choice of the auxiliary information. This is in some way an expected result, but here we can give light on some features than can be sometimes hidden in the shadow of the real life applications.

Two main results stem out from our simulation study. Firstly, when the set of auxiliary information is related both to the nonresponse mechanism and to the outcome variable, all examined weighting estimation methods are effective to reduce both bias and variance of the basic and unweighted estimator of the mean. Moreover they may also be more effective to reduce MSE than the method based on the true response probability. Therefore we must admit that there is no evidence of a bias-variance trade-off, due to using or not using a weighting method. Secondly, given the set of auxiliary variables, the bias-variance trade-off may regard the choice among different weighting methods because an inverse relationship between the bias reduction and the variance reduction may be sometimes observed.

When there are more outcome variables and the same weights are applied to all, the problem may become more intricate. Some variables may be unrelated to the response mechanisms and for this variables the unweighted estimator should be better. Moreover to include in a weighting estimation method all the relevant information may be “difficult” and may produce more unstable response probability estimations.

With respect to the comparison among the non parametric estimation methods and the others, the empirical results show that the local linear regression method, may be an appealing alternative mainly in the multivariate case. Anyway, more investigations are needed as we have to note that when the number of auxiliary variables increases the application of local linear regression may be very complex. Finally we do not exclude that other nonparametric models such, as generalized additive models – still not explored here – may lead to more effective results.

**Table 1:** Percentage biases of weighting adjusted mean estimator (one continuous covariate)

| Response<br>Funtions | Outcomes | $T_0$  | $T_1$  | $T_2$         | $T_3$         | $T_4$        | $T_5$  |
|----------------------|----------|--------|--------|---------------|---------------|--------------|--------|
| $\phi_1$             | $y_1$    | -0.015 | 0.458  | <b>-0.042</b> | 0.272         | 0.050        | 2.664  |
| $\phi_1$             | $y_2$    | -0.005 | 0.374  | <b>-0.040</b> | 0.249         | -0.051       | 1.822  |
| $\phi_2$             | $y_1$    | -0.009 | -0.981 | -0.073        | <b>-0.017</b> | -0.076       | -7.432 |
| $\phi_2$             | $y_2$    | -0.005 | -0.795 | -0.041        | <b>-0.011</b> | -0.093       | -5.076 |
| $\phi_3$             | $y_3$    | 0.093  | 0.994  | <b>0.022</b>  | 0.166         | 0.673        | 7.717  |
| $\phi_3$             | $y_4$    | 0.021  | 0.658  | 0.203         | -0.237        | <b>0.167</b> | 4.405  |
| $\phi_3$             | $y_5$    | 0.042  | 0.793  | 0.191         | <b>-0.078</b> | 1.316        | 5.255  |
| $\phi_3$             | $y_6$    | 0.074  | 1.908  | 1.060         | <b>0.565</b>  | 2.063        | 8.734  |
| $\phi_4$             | $y_3$    | -0.010 | -1.102 | <b>-0.218</b> | 0.729         | -0.458       | -7.297 |
| $\phi_4$             | $y_4$    | -0.028 | -0.452 | <b>0.026</b>  | 0.073         | -0.320       | -3.122 |
| $\phi_4$             | $y_5$    | -0.019 | -0.735 | <b>-0.135</b> | 0.313         | -1.378       | -4.519 |
| $\phi_4$             | $y_6$    | -0.011 | -0.731 | <b>0.113</b>  | -0.965        | -1.723       | -5.011 |

**Table 2:** Empirical Mean square Error of weighting adjusted mean estimator (one continuous covariate)

| Response<br>Funtions | Outcomes | $T_0$ | $T_1$ | $T_2$        | $T_3$        | $T_4$        | $T_5$  |
|----------------------|----------|-------|-------|--------------|--------------|--------------|--------|
| $\phi_1$             | $y_1$    | 1.596 | 1.022 | 0.946        | <b>0.937</b> | 0.944        | 3.031  |
| $\phi_1$             | $y_2$    | 0.262 | 0.173 | <b>0.154</b> | 0.157        | 0.161        | 0.396  |
| $\phi_2$             | $y_1$    | 1.582 | 1.227 | 0.952        | 0.955        | <b>0.943</b> | 14.903 |
| $\phi_2$             | $y_2$    | 0.254 | 0.208 | 0.162        | 0.178        | <b>0.160</b> | 1.802  |
| $\phi_3$             | $y_3$    | 1.488 | 1.112 | <b>0.856</b> | 0.880        | 0.964        | 13.596 |
| $\phi_3$             | $y_4$    | 0.588 | 0.522 | 0.387        | 0.387        | <b>0.377</b> | 6.525  |
| $\phi_3$             | $y_5$    | 1.016 | 0.848 | 0.645        | <b>0.621</b> | 1.193        | 9.410  |
| $\phi_3$             | $y_6$    | 1.040 | 1.282 | 0.877        | <b>0.739</b> | 1.374        | 12.346 |
| $\phi_4$             | $y_3$    | 1.392 | 1.137 | <b>0.851</b> | 0.977        | 0.894        | 12.278 |
| $\phi_4$             | $y_4$    | 0.637 | 0.456 | <b>0.378</b> | 0.394        | 0.413        | 3.456  |
| $\phi_4$             | $y_5$    | 0.975 | 0.814 | <b>0.629</b> | 0.678        | 1.282        | 7.117  |
| $\phi_4$             | $y_6$    | 1.192 | 0.848 | <b>0.746</b> | 0.864        | 1.167        | 4.580  |

## References

1. Aitchison, J., Aitken C.G.G.: Multivariate binary discrimination by the kernel method. *Biometrika*. 68, 301-309 (1976)
2. Beaumont, J.F.: Calibrated imputations in survey under a model-assisted approach. *Journal of the Royal Statistical Society B*. 67, 445-458 (2005)
3. Cassel, C.M., Särndal, C.E., Wretman J.H. : Some uses of statistical models in connection with the non-response problem. In *incomplete Data in Sample Surveys, 3, Theory and Bibliographies*, (Eds. W.G.Madow,I. Olkin and D.B.Rubin) Accademic Press. New York. 143-160 (1983)
4. Da Silva, D.N., Opsomer, J.D.: Properties of the weighting cell estimator under a nonparametric response mechanism. *Survey Methodology*. 30, 45-55 (2004)
5. Da Silva, D.N., Opsomer, J.D.: A kernel smoothing method of adjusting for unit nonresponse in sample survey. *The Canadian Journal of Statistics*. 34, 563-579 (2006)
6. Da Silva, D.N., Opsomer, J.D.: Nonparametric propensity weighting for survey nonresponse through local polynomial regression. To appear in *Survey Methodology* (2008)
7. Giommi, A.: A simple method for estimating individual response probabilities in sampling from finite populations. *Metron* 42, 185-200 (1984)
8. Giommi, A.: Nonparametric methods for estimating individual response probabilities. *Survey Methodology*. 13, 127-134 (1987)
9. Groves R., Dillman, D., Eltinge, J.L., Little, R.J.A.: *Survey Nonresponse*. Wiley, New York (2002)
10. Holt, D., Smith, T.M.F.: Post stratification. *Journal of the Royal Statistical Society, A*. 142, 33-46 (1979)
11. Kalton , G., Kasprzyk, D.: The treatment of missing survey data. *Survey Methodology*. 12, 1-16 (1986)
12. Kim, J.K., Kim, J.J.: Nonresponse weighting adjustment using estimated response probabilities *The Canadian Journal of Statistics*. 35, 501-514 (2007)
13. Kish, L.: Weighting for unequal  $P_i$ . *Journal of Official Statistics*. 8, 183-200 (1992)
14. Li, Q., Racine, J.S.: Cross validated local linear nonparametric regression. *Statistica Sinica*. 14, 485-512 (2004)
15. Li, Q., Racine, J.S.: *Nonparametric econometrics: Theory and Practice*, Princeton University Press, Princeton, NJ. (2006)
16. Little, R.J.A. :Survey nonresponse adjustment. *International Statistical Review*. 54, 139-157 (1986)
17. Little, R.J.A., Rubin, D.B.: *Statistical Analysis with Missing Data*, Wiley, New York. (2002)
18. Little, R.J.A., Vartivarian, S.: Does weighting for nonresponse increase the variance of survey mean. *Survey Methodology*. 31,161-168 (2005)
19. Oh, H.L., SCheuren, F.S. : Weighting adjustments for unit nonresponse. In *incomplete Data in Sample Surveys, 2, Theory and Bibliographies*, (Eds. W.G.Madow,I. Olkin and D.B.Rubin) Accademic Press. New York. 143-184 (1983)
20. Niyonsenga, T.: Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology*. 20, 177-184 (1994)
21. Niyonsenga ,T.: Response probability estimation. *Journal of Statistical Planning and Inference*. 59, 111-126 (1997)
22. Robins, J.M., Rotnitzky, A., Zhao, L.P.: Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 89, 846-866 (1994)
23. Rosenbaum, P.R.: Model-based direct adjustment. *Journal of the American Statistical Association*. 82, 387-394 (1987)
24. Wang, M.C., van Ryzin, J.: A class of smooth estimators for discrete distributions. *Biometrika*. 68:301-309 (1981)