Research Report

Department of Statistics

No. 2003:11

Propensity Score Weighting with Double Samples: A Simulation Study

Boris Lorenc

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Propensity Score Weighting with Double Samples: A Simulation Study

Boris Lorenc
Department of Statistics, Stockholm University
SE-106 91 Stockholm, Sweden
E-mail: boris.lorenc@stat.su.se

Abstract

The propensity score adjustment technique for web surveys, introduced by Terhanian, is theoretically known to reduce bias caused by nonrepresentativeness of web panel respondents with respect to the general population, on provision that the assumptions that pertain to the technique hold. In practical applications though, the technique implies making choices whose implications for the weighted point and variance estimates are too complex to express in analytic terms.

The propensity for participation in a web survey is not known and thus needs to be estimated. Choice of the model and estimation of the model parameters introduce uncertainties that do not exist in a purely theoretical analysis. Further, covariance structure of the variables chosen to be analysed cannot easily be estimated as the study variables are not observed in the unrestricted sample. Even characteristics specific to the propensity score technique, like number of strata or ratio of the sample sizes, may play a role when making inference from a finite set of data. Finally, the assumptions required by the technique may or may not actually be fulfilled.

As analytic expressions relating these factors to the point estimates and the estimates of their variance are lacking, a simulation was run to study their effects. A known probability distribution was used to generate artificial populations whose parameters were then estimated while varying the factors under study in a setup of a designed experiment. The results in general confirmed the bias reducing effect of the propensity score weighting. They also indicated that, under a particular covariance structure, the adjustment could actually increase the estimator's bias if a relevant variable was left out of the model by which the propensity score was estimated.

Introduction

To reduce the bias of the estimates of population parameters that arises when web panels are surveyed instead of random samples from the population, Terhanian (e.g. Terhanian, Marcus, Bremer, and Smith, 2001) suggested a new application of the propensity score weighting technique. The technique was originally developed by Rosenbaum and Rubin (1983a) to obtain a proper estimate of a treatment difference between *two populations*. The new in Terhanian's approach was to apply, with minor changes, this method to *two samples* from the same population—the name "double samples procedure" seems thus appropriate.

Situations where the method might be suitable arise when (*i*) access to and data collection from one of the samples is much cheaper than from the other one, and (*ii*) the inclusion probabilities into the former (cheaper to collect data from) sample are not explicitly known. It is often, with respect to *ii*, further suspected that the unknown inclusion probabilities are related to auxiliary[1] information or to variables under study, that is, this sample is presumably gathered from some specific subset of the general population. Additionally, self-selection into this sample might be present. (The researcher, when modelling this situation, may focus on estimating the inclusion probabilities or on estimating the difference in the distribution of the variables in the population and in the subset. The propensity score technique is geared towards the latter goal.)

In Terhanian's practical application of the technique, a web panel takes the role of the *restricted* sample as it consists of web users presumably different from the general population on important properties like age, income, and the like. In order to produce appropriate weights, web panel data are augmented with incomplete data (only the auxiliary variables are collected) from an *unrestricted* sample—a sample from the general population with the element inclusion probabilities known.

Terhanian and the colleagues reported some very accurate predictions of elections outcomes obtained by the propensity score adjustment technique (Terhanian, Marcus, et al., 2001, Terhanian, Taylor, Siegel, Bremer, and Smith, 2001). Their presentations were, unfortunately, given in descriptive and rather vague terms, with formal expressions and technical details omitted. In a study related to this paper (Lorenc, 2003), the present author gave a simple analytic demonstration of the effectiveness of the propensity score weighting in a double samples setting. Akin to the study of Cochran (1968) who investigated the case of sampling from two populations, this new study showed that Terhanian's method performed as indicated by its creator, reducing the bias of the unadjusted estimator by about 90% in the situations studied. The technique gives theoretically an unbiased estimator, but in practical applications the necessity to estimate the parameters from data hinders the accomplishment of this goal.

The technique may be said to consist of these steps:

1. collecting complete data—the auxiliary variables and the variables under study—from a web panel (the restricted sample) and collecting the auxiliary variables from a random sample of the population (the unrestricted sample),

2. given the whole set of auxiliary information (from the unrestricted and the restricted samples) but not the sample membership indicator, estimating for each unit the probability of being a panel member (this magnitude is known as the estimated propensity score); a common way of estimating this probability is by building a logistic regression model,

[1] What in the survey literature are known as *auxiliary* (or sometimes *background*) variables are in regression analysis referred to as *independent* variables, in the biomedical research as *covariates*, and in the econometric literature as *conditioning* variables.

3. estimating the distribution of the propensity score in the population by considering the distribution of the estimated propensity score in the unrestricted sample only; in particular, identifying cutoff points for stratification: usually equidistant cutoff points are chosen and 5 intervals are used, in which case the cutoff points would be the 20^{th}, 40^{th}, 60^{th}, and 80^{th} percentile of the estimated propensity score distribution in the population,

4. classifying the units in the restricted sample (panel) into appropriate strata based on their individual estimated propensity score values,

5. for each stratum, building a mean of the study variable values of the panelists in that stratum; then, weighting the strata means appropriately together to produce the final, adjusted estimate for that study variable; in the case of equidistant intervals the weighting amounts to calculating the arithmetic mean of the strata means.

Justification for the procedure and its details were given in (Lorenc, 2003).

Despite its theoretical clarity, the method nevertheless leaves some questions open. For one, the propensity score is in real applications not known and thus needs to be estimated. This uncertainty comes atop of the usual one, that of the unrestricted sample being a sample rather than a census. The most common method for estimating the propensity score is logistic regression, but discriminant analysis is an alternative suggested in the literature.

An expression for variance of the propensity score adjusted point estimate is lacking. The applied variance estimate (e.g. Rosenbaum and Rubin, 1984) is conditional on the chosen model, not taking into account the uncertainty regarding the model itself. Yet another issue is that of choosing the correct cutoff points for stratification. This is still an open research issue.

In addition, there are other factors that might influence the point estimate and the estimate of its variance. They are: absolute and relative sample sizes, covariance structure of the data, whether all the variables relevant to the study were observed or only a part of them, and so on. The effects of these factors are not clear because explicit expressions for the variance estimates are lacking. In such a situation, an alternative way to address these issues is through a simulation study. This path was chosen for the present study.

The present study is a simulation study that investigates the effect of the above mentioned factors on performance of the propensity score adjusted estimator in the situation of taking double samples from the same population. The study is performed in a manner of a designed experiment. In Section 1, goals of the study and motivation for investigating the chosen factors are presented. Details of the method are provided in Section 2, while Section 3 gives the results. Some general conclusions arising from the study and comments are given in the final Section 4.

1 Goal and studied factors

1.1 Goal

The present study investigates the influence of a number of factors on the effectiveness—here understood as the ability to reduce bias—of the propensity score technique under certain conditions (given in more detail below), with the main aim of preparation for the future analyses of real data. The setup of the study is of an experiment with a number of factors.

1.2 Studied factors

1.2.1 Preliminaries

Let $\mathbf{X}$ denote the auxiliary variables, let Y be the variable under study, and let Z be an indicator of inclusion into a subset (to be specified) of the population. Let two samples be drawn by the simple random sampling mechanism: an unrestricted sample, s, of size n from the population, and a restricted sample, r, of size k from the subset of the population.

The propensity score, denoted $e(\mathbf{x})$, is defined as $e(\mathbf{x}) = \Pr(Z = 1|\mathbf{X} = \mathbf{x})$. Let the following two conditions be fulfilled:

- a positive probability at every level of the propensity score for every unit in the population to be assigned to any of the samples[2], $0 < e(\mathbf{x}) < 1$, and
- independence of the study variable and the subset membership indicator conditional on the auxiliary information[3], $Y \perp Z \mid \mathbf{X}$.

The two assumptions are jointly referred to as *strongly ignorable treatment assignment* (SITA). Then, as proved by Rosenbaum and Rubin (1983a),

- conditioning on the propensity score establishes a (conditional) independence of the auxiliary information and the subset membership indicator, $\mathbf{X} \perp Z \mid e(\mathbf{X})$—which is sometimes referred to as *balancing*—and,
- conditioning on the propensity score removes, in expectation, the bias of the unadjusted mean of the study variable between the population and the subset[4], $E\{Y \mid e(\mathbf{X}),\ Z = 1\} = E\{Y \mid e(\mathbf{X}),\ Z = 0\}$.

The propensity score is in practice not known and must be estimated, all relevant auxiliary information might have been gathered or some of it left out, and the assumptions may hold or may not hold. These, and some further issues, are now discussed in more detail.

[2] This requirement, in effect when treating two populations with mutually exclusive membership, may in the present setting of double samples from the same population be relaxed to $0 < e(x) \leqslant 1$: even if a member of r with certainty, an element has still a positive probability to appear in s, thus satisfying the main requirement of a positive chance for each element to appear in any of the samples.

[3] Or conditional on any balancing score, of which the propenisity score is the coarsest. (The proof is given in Rosenbaum and Rubin,1983a, as the proof of Theorem 3.)

[4] In double samples procedures, for this to hold it is additionally required that there is no effect of data collection method on measurement (see Lorenc, 2003).

1.2.2 Estimation of the propensity score

Methods

The propensity score is in practice most often estimated by the use of a logistic regression with the logit link. Instead of the logit link, other links like the probit link might be considered, but in the present study, with a relatively simple population model used and a simple subpopulation inclusion rule (cf. Section 2) these give the same results, so this comparison was omitted.

Instead, discriminant analysis was used as an alternative for estimating the propensity score. Rosenbaum and Rubin (1983a) point out though that with multivariate normal distributions having common covariance in both treatment groups (which is the case in the present study), the propensity score is a monotone function of the discriminant score, why the eventual difference between the results of applying the two methods would stem from the estimation algorithms rather than from theoretical differences.

Available amount of data

With limited amount of data on which to base the estimation, a problem of insufficient overlap of the distributions of the propensity score in the two samples may arise. Practically, it may happen that, when classified in step 4 of the procedure outlined in the Introduction, no unit in the restricted sample falls into one of the strata. No contribution to the adjusted average response can thus come from this stratum, with the consequence that if an adjusted estimate is produced, it will then be biased. Occurrence of this is clearly related to the sample sizes, why the sample size factor was included in the study.

Theoretically, the larger the number of strata into which the sample is stratified based on the estimated propensity score, the lesser the bias of the adjusted estimate (Section 2 of Lorenc (2003) gives the details). In practical situations with a limited amount of data, though, a finer-grained stratification yields greater risk for empty strata which in turn might lead to a biased estimate. Coarser stratification lowers the risk for empty strata but might preserve a larger intrinsic bias compared to a finer-grained stratification. The number of strata was thus varied in the study.

Additionally, proportions in which the two samples partake in the joint sample might be of significance to the propensity score estimation methods. This was also included in the study, as a factor.

Variance estimation

While conditionally on the regression model variance estimation of the propensity score adjusted estimate is not difficult to obtain (see p. 10 under heading "Procedure", below), the model selection process includes sources of variance that are harder to express analytically. A simulation study provided the opportunity to estimate the variance of the propensity score adjusted estimate without access to analytic expressions, by simply obtaining the adjusted point estimates in each repeated drawing of the population and of the samples, and then calculating the variance of these values and comparing them to the variances calculated conditional on the regression model.

1.2.3 Assumption violations

Treating the same problem as in Cochran (1968) and Cochran and Rubin (1973), namely that of estimating treatment effects, Rosenbaum and Rubin (1983a) showed in a situation with multivariate covariates that conditioning the response on the propensity score (or on any balancing score, of which the propensity score is the coarsest) eliminates bias due to nonrandom treatment assignment provided that the assumptions given in SITA hold.

But, what would the consequences be if these assumptions did not hold? While analytical results might be derived in simpler cases, there was an interest in checking the behaviour of the estimator in more complex situations that in a sense resemble those that might be met in real applications.

SITA violations

SITA assumptions might be violated in two ways: some units may have a zero chance of being included in the restricted sample, $\exists i : e(\mathbf{x}_i) = 0$ where i indexes units in the population, and a dependence between the study variable and the subset membership indicator may remain even after conditioning on the auxiliary information, here symbolically denoted as $(Y \angle Z) \mid \mathbf{X}$. The effects of both violations were investigated in the simulation study.

Unobserved covariates

In addition, violation of the latter of the SITA assumptions occurs when an important auxiliary variable is left unobserved. In practical applications, the difficulty precisely it that there is an uncertainty whether all the covariates causally related to the response, Y, and the subset membership indicator, Z, are observed.

This factor is a variant of the preceding one, the only difference being a conceptual one: here, there was a variable that we ought to have observed but failed to do so while, in the previous case ($(Y \angle Z) \mid \mathbf{X}$), the nature of the phenomenon was such that Y and Z are tangled and could not be untangled by conditioning.

When investigating sensitivity of the propensity score technique to not recording a covariate, Rosenbaum and Rubin (1983b) considered an unobserved binary covariate. In the general multivariate normal setup of the present study (see subsection "The general setup" in next section), it seemed natural to investigate the effect of not including a continuous covariate, while varying the degree of correlation between this covariate and the other variables.

2 Method

Point of departure for the double samples application of the propensity score technique, as noted in Introduction, is that the values of the study variable Y are not observed for the units in the unrestricted sample from the population, s. So, given the information available in the two samples—s with only auxiliary variables and the restricted sample from the subset, r, with the complete information—the goal was to estimate the mean of the study variable Y for the whole population by applying the propensity score weighting technique. For the model presented in (1), p.7 below, this mean of Y in the population is zero.

The factors from the preceding section were included in a simulation experiment which consisted of repeatedly drawing s and r samples from populations with known characteristics defined by levels of the factors pertaining to the population properties and the samples' properties, and producing point and variance estimates based on levels of the factors pertaining to the estimator properties.

A summary of the factors is given after a presentation of the general setup.

2.1 The general setup

The following multivariate normal model was used:

$$(X_1, X_2, Y, V) \sim N(\mathbf{0}, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = \left[\begin{array}{cccc} 1 & \rho_{12} & \rho_{13} & \rho_{\cdot 4} \\ \rho_{12} & 1 & \rho_{23} & \rho_{\cdot 4} \\ \rho_{13} & \rho_{23} & 1 & \rho_{\cdot 4} \\ \rho_{\cdot 4} & \rho_{\cdot 4} & \rho_{\cdot 4} & 1 \end{array} \right]. \tag{1}$$

This model defined a finite population, $U = \{1, 2, , ..., i, ..., N\}$, while inclusion into the subset was defined through either $I_{V<X_2}$ or $I_{\max(V,0)<X_2}$. The variables in the model were given the following meanings, not uncommon in the survey literature:

X_1, an auxiliary variable,
X_2, another auxiliary variable, also involved in defining
the subset—"the participation variable",
Y, the study variable,
V, another variable involved in defining the subset.

The present model differed from the one studied in (Lorenc, 2003) in that here there are two auxiliary variables instead of one, each with its specific correlation coefficients with the other variables, and also that V may here even have a nonzero correlation with the other variables. In order not to inflate the number of factors in the experiment, V's correlation coefficient with the other variables was kept the same, $\rho_{\cdot 4}$, across the variables.

Setting V aside for the moment, the covariance matrix in (1) produces 8 different models when each of ρ_{12}, ρ_{13}, and ρ_{23} is held on one of the two positive levels, "high" and "low". Varying the covariance structure in this way gave the opportunity to investigate the efficiency of the propensity score adjustment under the "high" and "low" levels of correlation between each of the covariates and the response (Table 1).

The values of "low" and "high" for the ρ's in this reduced, 3×3, covariance matrix were set to .22 and .78, respectively—in any combination producing a positive definite matrix, as required by the model. The value of $\rho_{\cdot 4}$ was set to .175, producing with both signs ($-$ and $+$) positive definite matrices in all the 8 models. Three of the matrices, those pertaining to the structures 5, 6, and 7, became thereby almost singular (i.e., the third variable an almost deterministic function of the other two).

Table 1: Denotations for the covariance structures investigated in the study.

Covariance structure	ρ_{12}	ρ_{13}	ρ_{23}
1	low	low	low
2	low	low	high
3	low	high	low
4	high	low	low
5	low	high	high
6	high	low	high
7	high	high	low
8	high	high	high

Table 2: Expected values of the study variable Y in the subset for the chosen values of ρ_{23} and $\rho_{\cdot 4}$, which also are the biases of the corresponding sample means of Y in r if they were to estimate the expected value of Y in the population (the latter being zero).

	$\rho_{23} = .22$		$\rho_{23} = .78$	
	$Z = I_{V<X_2}$	$Z = I_{\max(V,0)<X_2}$	$Z = I_{V<X_2}$	$Z = I_{\max(V,0)<X_2}$
$\rho_{\cdot 4} = -.175$	.206	.244	.497	.719
$\rho_{\cdot 4} = 0$	.124	.201	.440	.708
$\rho_{\cdot 4} = .175$	.028	.141	.376	.692

Assignment to the subset was set either by $Z = I_{V<X_2}$, conforming to the SITA assumption that $\forall i \in U : 0 < e(\mathbf{x}_i) \leqslant 1$, or by $Z = I_{\max(V,0)<X_2}$, violating this assumption through $\exists i \in U : x_{2,i} < 0 \longrightarrow e(\mathbf{x}_i) = 0$. For this latter case the reference “SITA violation #2” is used below.

The coefficient $\rho_{\cdot 4}$—determining the correlation between X_2 and V—may here, in contrast to the setup in (Lorenc, 2003), have taken on values different from 0. When $\rho_{\cdot 4} \neq 0$, assignment to the subset was not strongly ignorable: the assumption $(Y \perp Z) \mid X_2$ was violated. For this condition the reference “SITA violation #1” is used in what follows.

With no SITA assumption violations, the expected value of the study variable Y in the subset was biased with respect to the expected value of Y in the population, with the bias expressible as $\rho_{23} \times \pi^{-\frac{1}{2}} = \rho_{23} \times .564$ (ibid.), giving for the case $\rho_{23} = .22$ the bias of .124, and for the case $\rho_{23} = .78$ the bias of .440. The expected values of Y in the subset, both with and without the SITA violations, are given in Table 2.

Two samples were drawn using simple random sampling, one from the complete population, denoted by s and of size n, and another from the subset, denoted by r and of size k.

The variable X_2, which took part in defining the subset from which the restricted sample r was drawn (“the web users”), is sometimes referred to as “the participation variable” meaning participation in the restricted sample, the only sample providing information on Y.

2.2 Summary of the studied factors

The following factors were included in the study:

1. Covariance structure [denoted COVSTR in the Tables and Figures]: 8 levels (the 8 models presented in Table 1),

2. Sample sizes [SSIZE]: 2 levels ("low", $n_{low} = 1000$, and "high", $n_{high} = 5000$, for the sample s),

3. Ratio of k, the size of the sample r, to n, the size of the sample s [KNRATIO]: 3 levels ($\frac{1}{2}$, $\frac{2}{2}$, and $\frac{3}{2}$, giving the restricted sample's sizes $k_{low} = \{500, 1000, 1500\}$ for the "SSIZE low" condition and $k_{high} = \{2500, 5000, 7500\}$ for the "SSIZE high" condition),

4. Method [METHOD]: 2 levels (logistic regression with logit link and discriminant analysis),

5. Observed covariates [OBSERVED]: 4 levels (an "analytic" level (A), where the known propensity score is used (see the heading "Recorded statistics", below), and three levels where the propensity score is estimated after only X_1 is observed, only X_2 is observed, and both covariates are observed and forced into the model),

6. Number of strata into which the empirical distribution of the propensity score is stratified [NSTRATA]: 2 levels (5 and 7 strata),

7. SITA violation #1 [SITAVIO1]: 3 levels ("yes, negative correlation", $\rho_{\cdot 4,yes-} = -.175$, "no", $\rho_{\cdot 4,no} = 0$, and "yes, positive correlation", $\rho_{\cdot 4,yes+} = .175$),

8. SITA violation #2 [SITAVIO2]: 2 levels ("no", $\forall i : 0 < e(\mathbf{x}_i) < 1$, and "yes", $\exists i : e(x_i) = 0$),

2.3 Procedure

For each of the level combinations of the all the factors except METHOD and OBSERVED, $b = 1000$ independent trials were run, where a trial consisted of generating a simulated population of size $N = 50000$ with the required properties, taking an unrestricted sample s and a restricted sample r, and calculating the required statistics (see next subsection) from them. As comparisons between the two propensity score estimation techniques, and between the effects of observing differing amount of information, were of interest, the required statistics for the levels of the factors METHOD and OBSERVED were calculated on the same sets of data.

For every point estimate, a corresponding estimate of its variance was calculated. For this, the method of Mosteller and Tukey was used. This particular formulation, "the method of Mosteller and Tukey", is the consequence of Rosenbaum and Rubin's (1984) special mention of it; they namely say: "standard errors for the adjusted proportions were calculated following Mosteller and Tukey (1977, Chap. 11c)."

Chapter 11c of Mosteller and Tukey (1977) considers primarily the choice of the standard population when comparing samples from two populations. All the examples in

11c concern proportions, and the usual variance estimators for the estimated proportions, $\hat{p}$, are applied within each stratum l, $\hat{V}_l\left(\hat{p}_l\right) = \frac{1}{n_l}\hat{p}_l\left(1-\hat{p}_l\right)$, the only specific issue being the weight given to each stratum's variance in building the overall variance.

In their example in the reference above, Rosenbaum and Rubin seem to give equal weights to each stratum; this could not be fully confirmed though, as using the data supplied in their Table 1 did not give an exact replicate of their reported standard errors (equal weights gave though the values closest to those in the article, about 0.01 below, of the several considered weightings; the difference seems though to be too large for a result of rounding errors, why this remains an open issue).

In the present application, each stratum's standard deviation was given equal weight as the strata cutoff points were equidistant:

$$\hat{V}\left(\widehat{E\left(\bar{Y}\right)}_{\{\cdot\}}\right) = \sum_{l=1}^{L}\left(\frac{N_l}{N}\right)^2\frac{s_{yl}^2}{k_l} = \sum_{l=1}^{L}\left(\frac{\frac{1}{L}N}{N}\right)^2\frac{s_{yl}^2}{k_l} = \frac{1}{L^2}\sum_{l=1}^{L}\frac{s_{yl}^2}{k_l}, \tag{2}$$

where k_l denotes the number of units in r falling into stratum l of the quantized distribution of the estimated propensity score in the population, and s_{yl}^2 denotes the variance of the observed values on Y for these units.

The simulation was performed in Matlab.

2.4 Recorded statistics

The following statistics were recorded for every simulated population:

1. the propensity score adjusted estimates of the population mean for Y based on the restricted sample r:

   - one estimate based on the true propensity score, $\widehat{\bar{Y}}_{\{r,PS\}}$, (*i.e.*, it was taken as known that $e\left(\mathbf{X}\right) = \Phi\left(X_2\right)$, the cumulative distribution function of X_2, as in the demonstration example of (Lorenc, 2003), subsection "Reducing the bias of $\bar{Y}_r$ as an estimator of $E\left(Y\right)$ by stratification on the propensity score")[5]; the cutoff points were estimated from the empirical distribution of the propensity score in the unrestricted sample s rather than found in the table of $\Phi\left(\cdot\right)$; this estimate corresponds to the level A of the factor OBSERVED,
   - three estimates based on the estimated propensity score obtained by the current method (either logistic regression or discriminant analysis), where the information observed and available for estimation of the propensity score varied: it could be only X_1, which gave the point estimator $\widehat{\bar{Y}}_{\left\{r,\widehat{PS}_{X_1}\right\}}$, only X_2, which gave $\widehat{\bar{Y}}_{\left\{r,\widehat{PS}_{X_2}\right\}}$, and ($X_1$, X_2) forcing both variables in the model, which gave $\widehat{\bar{Y}}_{\left\{r,\widehat{PS}_{X_1X_2}\right\}}$; the cutoff points were estimated from the empirical distribution of $\widehat{PS}_{\mathbf{X}} = \widehat{e\left(\mathbf{X}\right)}$ in the unrestricted sample s, with $\mathbf{X}$ replaced by the currently observed $\mathbf{X}$-values (X_1, X_2, or X_1 and X_2).

[5] In the case SITAVIO2=YES (i.e. $Z = I_{\max(V,0)<x_2}$) the "usual" true propensity score $e\left(\mathbf{X}\right) = \Phi\left(X_2\right)$ was somewhat inappropriately used.

2. the bias of the estimates in item 1 with respect to μ_Y, the mean of the variable Y in the current population,

3. variance estimates for the estimates in item 1 calculated using (2),

4. whether the population mean of the current population, μ_Y, was contained in the estimated nominal 95% confidence interval built around the point estimates in item 1 by subtracting and adding 1.96 times the root of the corresponding variance estimates in item 3.

As $b = 1000$ trials were performed at each combination of the factor levels, some summary statistics could be generated:

I. mean bias across the b trials (*MeanBias* in the reported Figures and Tables),

II. empirical confidence level: proportion of confidence interval "hits"—the mean of the statistics in item 4 above across the b trials (*Clevel*),

III. difference between the mean across the b trials of the root of the variance estimate in item 3 above and the observed simulation standard deviation across the b trials of the corresponding point estimate in item 1 above (*StDiff*).

The statistics I-III are presented as results in the next section. The statistics were analysed using the design-of-experiments module in the statistical package Minitab. When percentage reduction in bias for the compound simulation statistics *MeanBias* is presented, it was calculated using

$$prb\left(\hat{\theta}_{\{\cdot\}}\right) = 100\left(1 - \frac{\left|\frac{1}{b}\sum_{j=1}^{b}\hat{\theta}_{\{\cdot\},j} - \theta\right|}{\left|\hat{\theta}_r - \theta\right|}\right),$$

where $\hat{\theta}_{\{\cdot\}}$, $\hat{\theta}_r$, and θ are the estimator adjusted using the technique $\{\cdot\}$, the unadjusted estimator (based on the r sample only), and the parameter they aim to estimate, respectively. Thus, prb was calculated from the summary data, and not for each generated population separately.

3 Results

The results of the simulation are presented in tabular and graphical form. The main table of results[6] consists of percentages reduction in bias and empirical confidence levels of the propensity score weighted estimator under conformance and the deviations from the assumptions. Second-order interaction plots of the studied factors are added with the aim to give the reader an impression about the individual contributions of the studied factors on the simulation statistics, as well as about the contributions of their interactions. Two additional kinds of tables, containing more detailed information, also exist: ANOVA tables for each of the summary statistics, up to second order effects, and tables of means

[6] Table 4 on p. 17.

Table 3: The eight cases of assumption violations.

Case	Only X_1 observed	$\rho_{.4} \neq 0$	$Z = I_{\max(V,0)<X_2}$
0	no	no	no
1	yes	no	no
2	no	yes	no
3	no	no	yes
4	yes	yes	no
5	yes	no	yes
6	no	yes	yes
7	yes	yes	yes

of the first and second order effects, across all the levels partaking in the current analysis. These two kinds of Tables—too large and detailed to constitute a part of the text—are given in the Appendix.

The factors that proved to have a dominating effect on the observed simulation statistics were those related to violations of the assumptions for the propensity score technique. In order to give a clear picture of the contributions of all the factors investigated, first presented is the case where all the assumptions held, followed by the cases where they were violated in various ways. There are 8 such cases all in all (including the one where all the assumptions held), as the Table 3 illustrates. The results are presented in this order.

Within cases, the results are presented first for the point estimation (i.e., the simulation statistic *MeanBias*), followed by those regarding variance estimation (i.e., the statistics *StDiff* and *Clevel*). But, as variance estimates are of little use for the production of correct confidence intervals if the point estimates are seriously biased, which in general turned out to be the case when the SITA assumptions did not hold, the variance estimates are not always presented for such cases (but can always be found in the Tables in the Appendix).

3.1 Case 0: SITA assumptions hold

Besides the two factors explicitly named so, SITAVIO1 and SITAVIO2, even not observing an important covariate constitutes a violation of a SITA assumption. Specifically, observing only X_1 would amount to observing incomplete information as it does not hod that $(Y \perp Z) \mid X_1$. So, in presenting here the results where the SITA assumptions held, the estimates obtained using the true propensity score, the information in X_2, and the information in X_1 and X_2 together are included. Thus, the effects of the factors COVSTR, SSIZE, KNRATIO, METHOD, NSTRATA, and OBSERVED, excluding OBSERVED=X1, are analysed here.

3.1.1 Case 0: Point estimation

When SITA assumptions held, major factors that influenced the remaining bias in the propensity score adjusted point estimates of the mean of Y in the population were COVSTR and NSTRATA (Figure 1). Inspection of the ANOVA table (Table I in the Appendix)

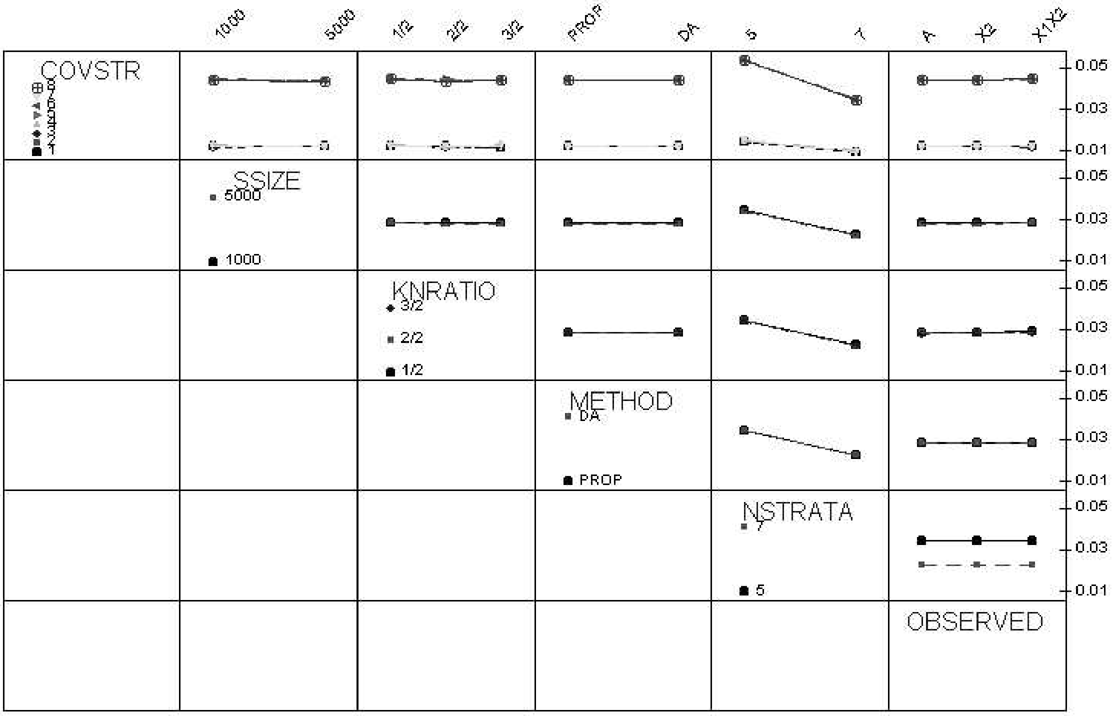

Figure 1: Case 0, interaction plot for *MeanBias*. (Identification of the COVSTR levels given in the text.)

showed that all the main effects except METHOD had a statistically significant contribution ($p \leqslant .001$), but that the dominating ones were those two mentioned. In addition, all the two-way interactions involving COVSTR except that with METHOD plus the interaction KNRATIO×NSTRATA were the significant ones among the two-way interactions ($p \leqslant .005$).

The covariance structures whose point estimates had the higher mean level of the bias in the first row of plates in Figure 1 (about .044 on the average) were those denoted by the numbers 2, 5, 6, and 8, that is, those where the correlation between the participation variable and the study variable was high (i.e. $\rho_{23} = .78$). This confirmed the analytically derived values of the adjusted point estimates for the same situation (Lorenc, 2003): the values of the bias of these point estimates falling under "5" and "7" in the pane for the interaction COVSTR×NSTRATA in Figure 1—the means across the structures 2, 5, 6, and 8—were .0537 and .0352 respectively, deviating only in the last decimal from the theoretically derived ones, .0532 and .0351 (also, see Table IV in the Appendix).

3.1.2 Case 0: Variance estimation

When the point estimates were based on the true propensity scores, that is, when the estimator $\widehat{Y}_{\{r,PS\}}$ was used, the corresponding variance estimates, calculated applying the suggested method (”the method of Mosteller and Tukey”), were approximately correct irrespective of the covariance structure (Figure 2; the estimator $\widehat{Y}_{\{r,PS\}}$ is represented as METHOD=A). But, when the point estimates were based on the estimated propensity

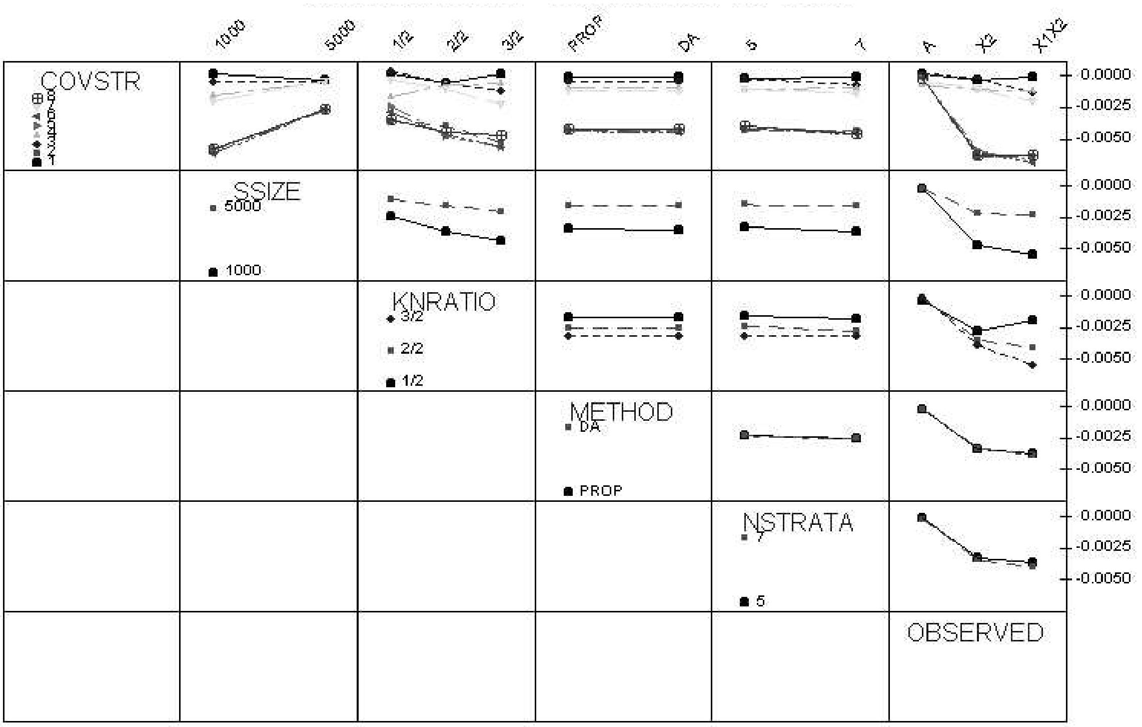

Figure 2: Case 0, interaction plot for *StDiff*. (Identification of the COVSTR levels given in the text.)

scores instead, resulting in the estimators $\widehat{\widehat{Y}}_{\{r,\widehat{PS.}\}}$, the accuracy of the variance estimates depended on the covariance structure: less underestimating or approximately correct were those where the original bias was low (that is, those with the low correlation between X_2 and Y, denoted by 1, 3, 4, and 7), while more underestimating were those where the original bias was high (that is, those where the correlation between X_2 and Y was high, denoted by 2, 5, 6, and 8).

But, that the variance estimates of the point estimates were eventually correct did not necessarily result in correct confidence intervals, those where the parameter of interest would be found in the interval in a prespecified proportion of trials. The reason for this could be found in the biasedness of the point estimators, where intervals of the eventually correct lengths were built around wrong point estimates, thus performing below the required coverage level. Thus, for the estimator $\widehat{\widehat{Y}}_{\{r,PS\}}$, whose variance was estimated approximately correctly irrespective of the covariance structure, the attained confidence level was nevertheless much below the required one when the correlation between the participation variable X_2 and the study variable Y was high (i.e., the covariance structures 2, 5, 6, and 8, that gave rise to the more biased point estimates—Figure 3).

It can be noted that the levels corresponding to the smaller amounts of data (SSIZE=1000 and KNRATIO=$\frac{1}{2}$) led to confidence intervals with somewhat better coverages than those with the larger amounts of data.

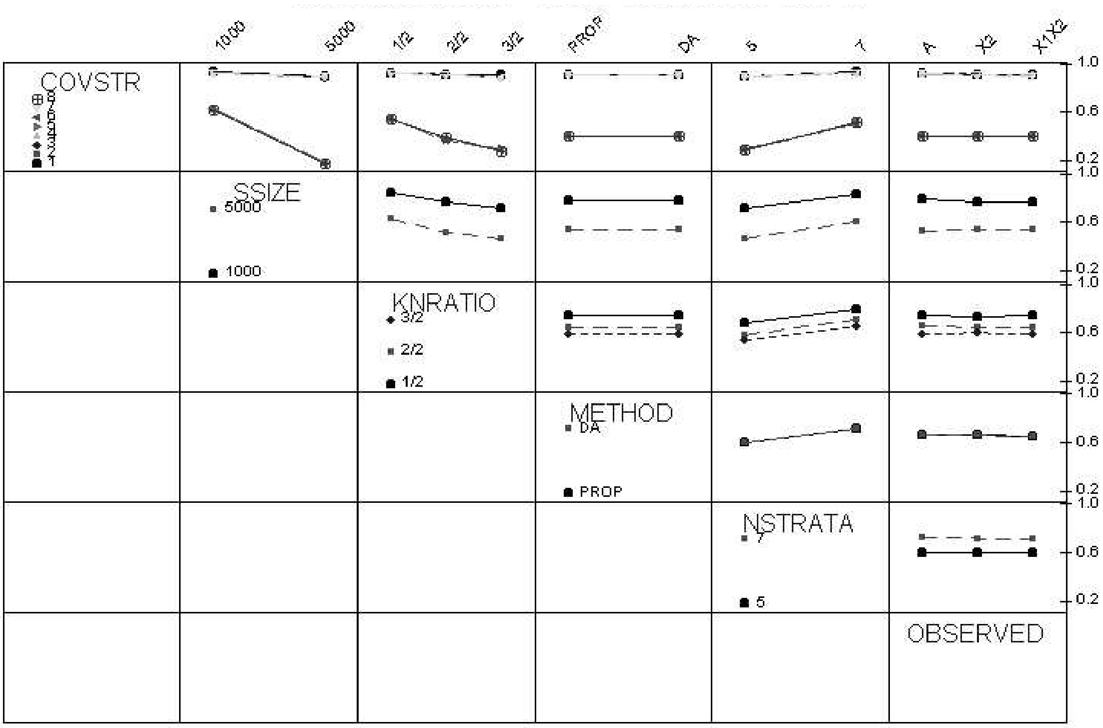

Figure 3: Case 0, interaction plot for *Clevel.* (Identification of the COVSTR levels given in the text.)

3.2 Case 1: SITA violated, not all the relevant information observed

The results presented thus far had not included the estimates that were based on just the auxiliary variable X_1 being observed (i.e., OBSERVED=X1). Observing only X_1 would amount to observing incomplete information: the participation variable X_2 is required instead in order to 'explain' Z. Only insofar as X_1 and X_2 would be correlated would the observation of X_1 help when X_2 ought to have been observed instead. So, here even the condition OBSERVED=X1 is added into the analysis besides the 6 factors and levels explored previously.

3.2.1 Case 1: Point estimation

The estimates of the mean of Y in the population based on only X_1 were considerably more biased than those derived using the true propensity score or when X_2 was amongst the variables observed (Figure 4). And, an interesting pattern arose: a comparison of the unadjusted estimates ($\bar{Y}_r$, the means of Y in the restricted sample r) with those adjusted on the estimated propensity scores after observing only X_1, across all the other factors, is given in Table 4 together with the corresponding correlation coefficients in the covariance matrix.

The bias of the unadjusted estimator $\bar{Y}_r$ was a function of ρ_{23}, which is the correlation between the participation variable X_2 and the study variable Y. When the participation variable X_2 was observed, the bias was reduced to approximately 10% of its original value,

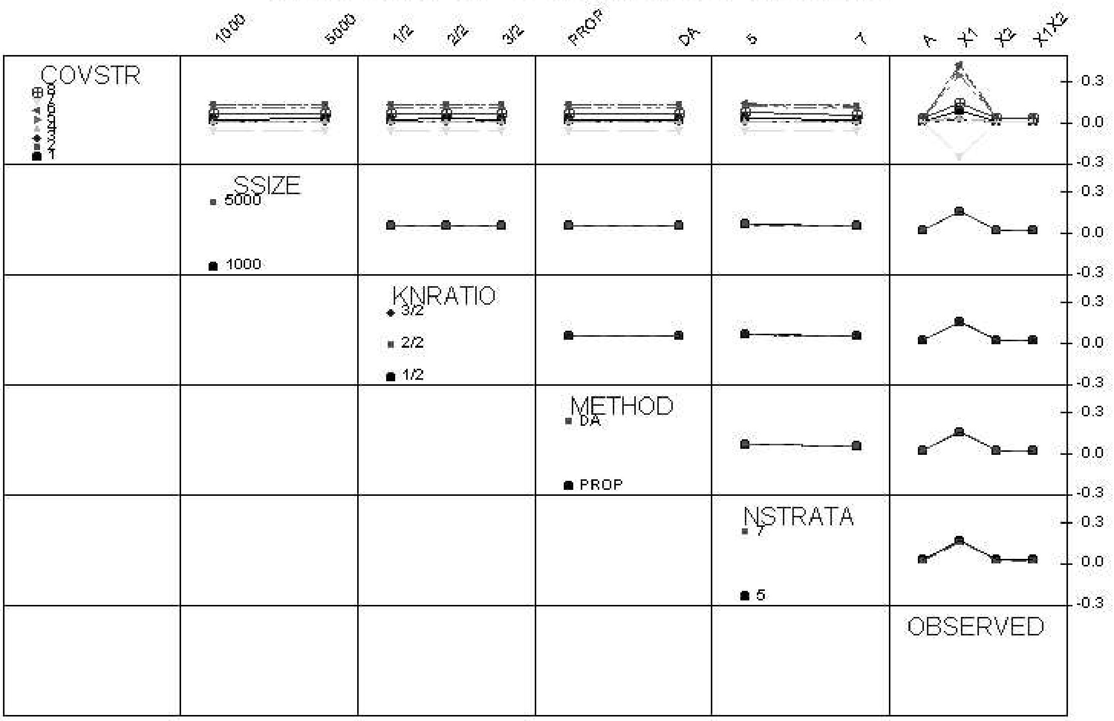

Figure 4: Case 1, interaction plot for *MeanBias*. (Identification of the COVSTR levels given in the text and Table 4.)

as reported in the preceding section of the results.

When X_2 was not observed but X_1 was, adjustment for the bias could be attempted using the information contained in the correlation between X_1 and X_2 (i.e., ρ_{12}) and the correlation between X_1 and Y (i.e., ρ_{13}). When information in both ρ_{12} and ρ_{13} was low, then the adjustment attempts in general resulted in small percentage reduction in bias, though smaller when ρ_{23} was high then when it was low (rows 1 and 2 corresponding to the estimator $\widehat{\widehat{Y}}_{\{r,\widehat{PS}_{X_1}\}}$—denoted for short $\widehat{\widehat{Y}}_1$— in Table 4). When the original bias was low (i.e., ρ_{23} was low), it took one of ρ_{12} and ρ_{13} to be high in order to produce a considerable reduction in bias (rows 3 and 4 corresponding to $\widehat{\widehat{Y}}_1$ in Table 4). But, when the original bias was high (i.e., ρ_{23} was high), it did not suffice that just one of ρ_{12} and ρ_{13} was high to produce a large reduction in bias (rows 5 and 6 corresponding to $\widehat{\widehat{Y}}_1$ in Table 4). Both needed to be high, and yet the percentage reduction in bias could only be moderate (row 8 corresponding to $\widehat{\widehat{Y}}_1$ in Table 4). But, when both the information contained in ρ_{12} and in ρ_{13} was at the high level but when the original bias was low, then the estimator "overadjusted" for the bias, actually almost doubling it (row 7 corresponding to $\widehat{\widehat{Y}}_1$ in Table 4).

The factor OBSERVED (including the level OBSERVED=X1) was by far the strongest among those included (Table V in the Appendix).

Table 4: The unadjusted estimators $\bar{Y}_r$ (the means across all drawn populations), and the adjusted estimators $\widehat{\bar{Y}}_{MI,\cdot}$ with their corresponding percentages reduction in bias (*prb*) and empirical confidence levels of the nominal 95% confidence intervals (*Clevel*) for the 8 treated cases and, within each, for the 8 covariance structures across all the levels not defining a case.

Note: As the factor SITAVIO1 consists of three levels that generate different population and subset properties and have a different effect on the adjusted estimators, the component tables of Table 4 that correspond to negative values of $\rho_{.4}$ have a negative sign before their indices (in the lower part of the table), while the component tables that correspond to positive values of $\rho_{.4}$ do not have this sign (in the upper part of the table). The neutral case (case 0, no assumption violations) is placed in the middle of the table, just below the component tables refering to just SITA violation #2 (i.e., cases 3 and 5).

COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_6$	*prb*	*Clevel*	$\widehat{\bar{Y}}_7$	*prb*	*Clevel*
1	.22	.22	.22	.141	.096	32	.06	.116	18	.05
2	.22	.22	.78	.691	.546	21	0	.675	2	0
3	.22	.78	.22	.141	.114	19	.04	.041	71	.30
4	.78	.22	.22	.141	.092	35	.06	.028	80	.84
5	.22	.78	.78	.691	.563	19	0	.600	13	0
6	.78	.22	.78	.691	.539	22	0	.938	−36	0
7	.78	.78	.22	.141	.101	28	.04	−.594	−321	0
8	.78	.78	.78	.691	.549	21	0	.317	54	0
COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_2$	*prb*	*Clevel*	$\widehat{\bar{Y}}_4$	*prb*	*Clevel*
1	.22	.22	.22	.028	−.100	−258	.17	.022	21	.77
2	.22	.22	.78	.376	.008	98	.83	.371	1	0
3	.22	.78	.22	.028	−.078	−180	.37	.008	72	.79
4	.78	.22	.22	.028	−.105	−275	.15	−.056	−100	.36
5	.22	.78	.78	.376	.029	92	.63	.356	5	0
6	.78	.22	.78	.376	−.004	99	.76	.348	7	0
7	.78	.78	.22	.028	−.091	−224	.23	−.281	−904	0
8	.78	.78	.78	.376	.010	97	.80	.122	68	.01
COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_3$	*prb*	*Clevel*	$\widehat{\bar{Y}}_5$	*prb*	*Clevel*
1	.22	.22	.22	.200	.156	22	0	.164	18	0
2	.22	.22	.78	.708	.553	22	0	.686	3	0
3	.22	.78	.22	.200	.156	22	0	.058	71	.18
4	.78	.22	.22	.200	.156	22	0	.093	54	.34
5	.22	.78	.78	.708	.553	22	0	.581	18	0
6	.78	.22	.78	.708	.553	22	0	.972	−37	0
7	.78	.78	.22	.200	.156	22	0	−.553	−177	0
8	.78	.78	.78	.708	.553	22	0	.327	54	0
COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_0$	*prb*	*Clevel*	$\widehat{\bar{Y}}_1$	*prb*	*Clevel*
1	.22	.22	.22	.124	.013	90	.92	.101	19	.09
2	.22	.22	.78	.440	.045	90	.41	.421	4	0
3	.22	.78	.22	.124	.012	90	.92	.036	71	.37
4	.78	.22	.22	.124	.013	90	.91	.043	65	.52
5	.22	.78	.78	.440	.044	90	.41	.357	19	0
6	.78	.22	.78	.440	.044	90	.40	.435	1	0
7	.78	.78	.22	.124	.013	90	.91	−.240	−93	0
8	.78	.78	.78	.440	.044	90	.40	.152	66	0
COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_{-2}$	*prb*	*Clevel*	$\widehat{\bar{Y}}_{-4}$	*prb*	*Clevel*
1	.22	.22	.22	.206	.123	37	.14	.171	17	0
2	.22	.22	.78	.497	.083	83	.17	.474	5	0
3	.22	.78	.22	.206	.107	48	.28	.061	70	.16
4	.78	.22	.22	.206	.134	35	.13	.139	32	.05
5	.22	.78	.78	.497	.060	88	.35	.364	27	0
6	.78	.22	.78	.497	.096	81	.14	.531	−7	0
7	.78	.78	.22	.206	.120	42	.18	−.207	−1	0
8	.78	.78	.78	.497	.080	84	.18	.184	63	0
COVSTR	ρ_{12}	ρ_{13}	ρ_{23}	$\bar{Y}_r$	$\widehat{\bar{Y}}_{-6}$	*prb*	*Clevel*	$\widehat{\bar{Y}}_{-7}$	*prb*	*Clevel*
1	.22	.22	.22	.245	.207	15	0	.203	17	0
2	.22	.22	.78	.719	.556	23	0	.697	3	0
3	.22	.78	.22	.245	.194	21	0	.072	70	.11
4	.78	.22	.22	.245	.212	14	0	.152	38	.15
5	.22	.78	.78	.719	.543	25	0	.566	21	0
6	.78	.22	.78	.719	.564	22	0	1.005	−40	0
7	.78	.78	.22	.245	.204	17	0	−.516	−110	0
8	.78	.78	.78	.719	.557	23	0	.337	53	0

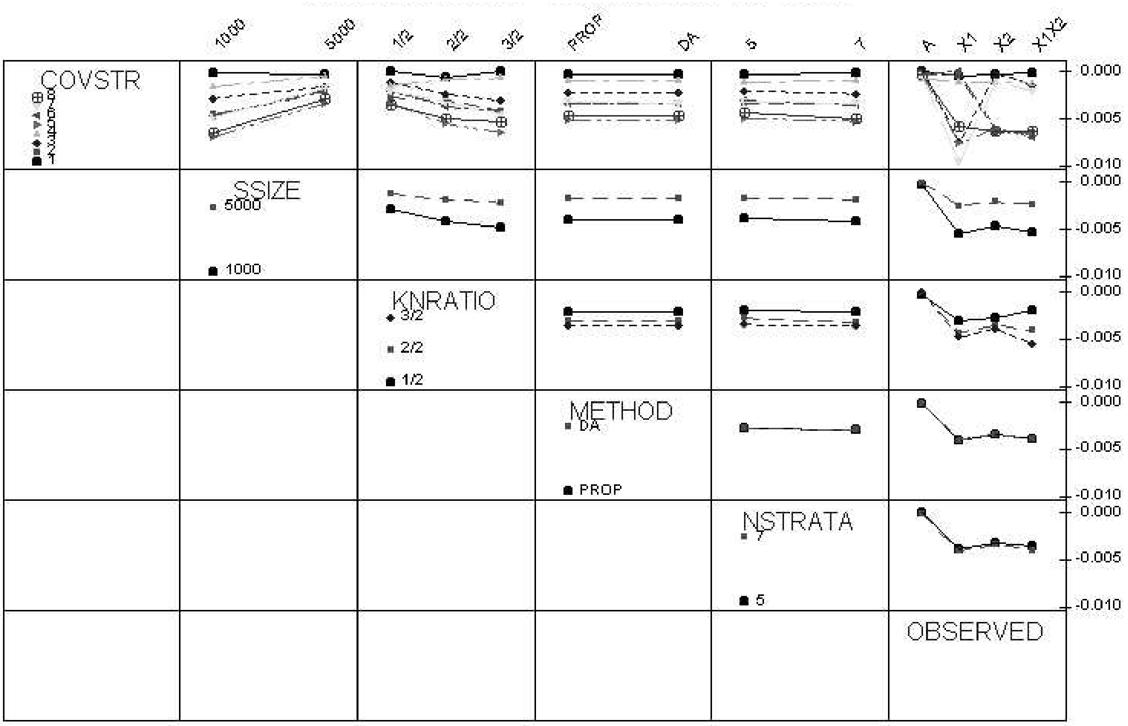

Figure 5: Case 1, interaction plot for *StDiff*. (Identification of the COVSTR levels given in the text.)

3.2.2 Case 1: Variance estimation

The accuracy of the variance estimation for the estimator involving only X_1 followed in a sense the same pattern as for those reported for the corresponding results for Case 0. It was less biased or practically unbiased for some covariance structures, here those where the correlation between X_1 and Y was low ($\rho_{13} = .22$): that is, the structures denoted by the numbers 1, 2, 4, and 6. And it was more biased for the covariance structures where the correlation between X_1 and Y was high ($\rho_{13} = .78$): that is, those denoted by the numbers 3, 5, 7, and 8 (Figure 5, the pane COVSTR × OBSERVED). But, as previously, that the variance estimate was correct did not help to achieve the desired significance level of the confidence intervals: the biased point estimate prevented this (Figure 6).

3.3 Case 2: SITA violated, Y and Z correlated after all the relevant information observed

A dependence between the study variable Y and the indicator of the subset membership Z that remains after conditioning their joint distribution on the auxiliary information, symbolically represented as $(Y \angle Z) \mid \mathbf{X}$, violates one of the assumptions of SITA. In such a situation, in words, there exists information in the subset membership Z about Y that is not available for adjustment by the propensity score technique. This factor, named SITAVIO1, was now added to the 6 previously analysed ones: COVSTR, SSIZE, KNRATIO, METHOD, NSTRATA, and OBSERVED (excluding OBSERVED=X1).

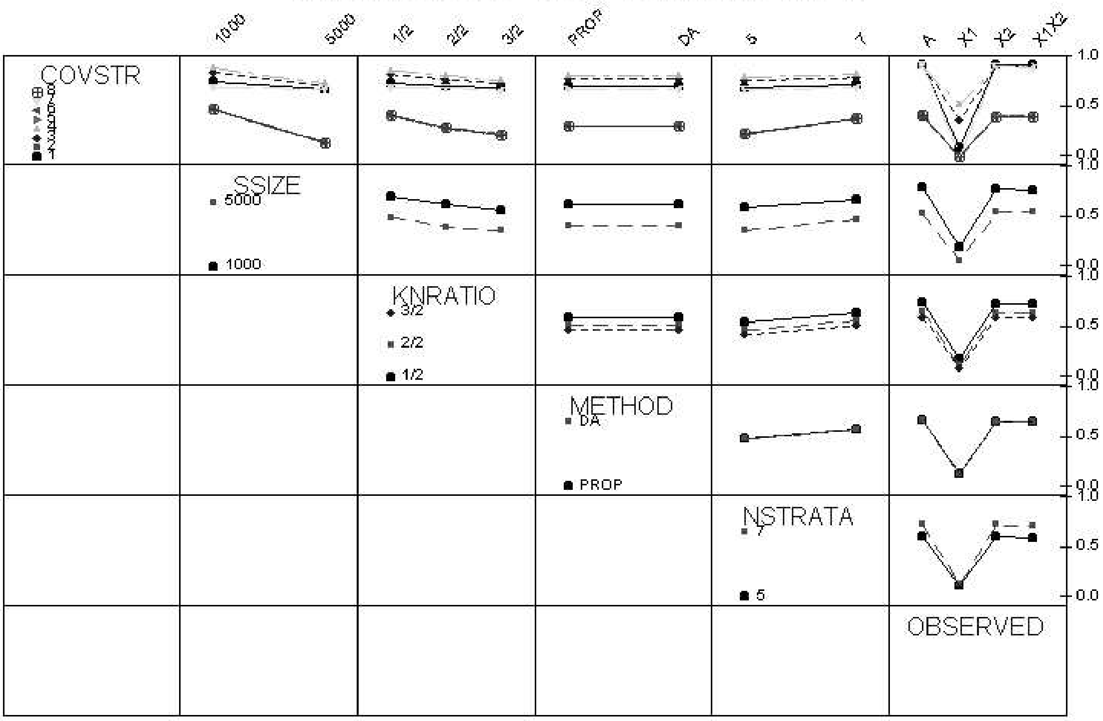

Figure 6: Case 1, interaction plot for *Clevel*. (Identification of the COVSTR levels given in the text.)

3.3.1 Case 2: Point estimation

The direction of influence of a remaining conditional dependence, $(Y \angle Z) \mid \mathbf{X}$, on the estimate of the mean of Y in the population was a function of the sign of the correlation coefficient $\rho_{.4}$ between $\mathbf{X}$ and Y. Negative correlations increased the point estimate and positive ones decreased it (Figure 7). If with $\rho_{.4} = 0$ the point estimate was an overestimate, then a change in the covariance structure to a negative $\rho_{.4}$ would increase the bias, while a change to a positive $\rho_{.4}$, for an appropriate range of small values of $\rho_{.4}$, would decrease the bias but for a large value of $\rho_{.4}$ again increase it. The reverse would hold if, with $\rho_{.4} = 0$, the point estimate was an underestimate.

The factor SITAVIO1 was by far the strongest among those included (Table IX in the Appendix).

The covariance structures differed in how much they were affected by the departures of $\rho_{.4}$ from zero. More resistant were those where the correlation between the participation variable X_2 and the study variable Y was high (the structures 2, 5, 6, and 8), while less resistant were those where the correlation between X_2 and Y was low (the structures 1, 3, 4, and 7).

3.3.2 Case 2: Variance estimation

The simulation statistics related to variance estimation and confidence levels, *StDiff* and *Clevel*, followed in general the patterns detected for the corresponding results for the Cases 0 and 1.

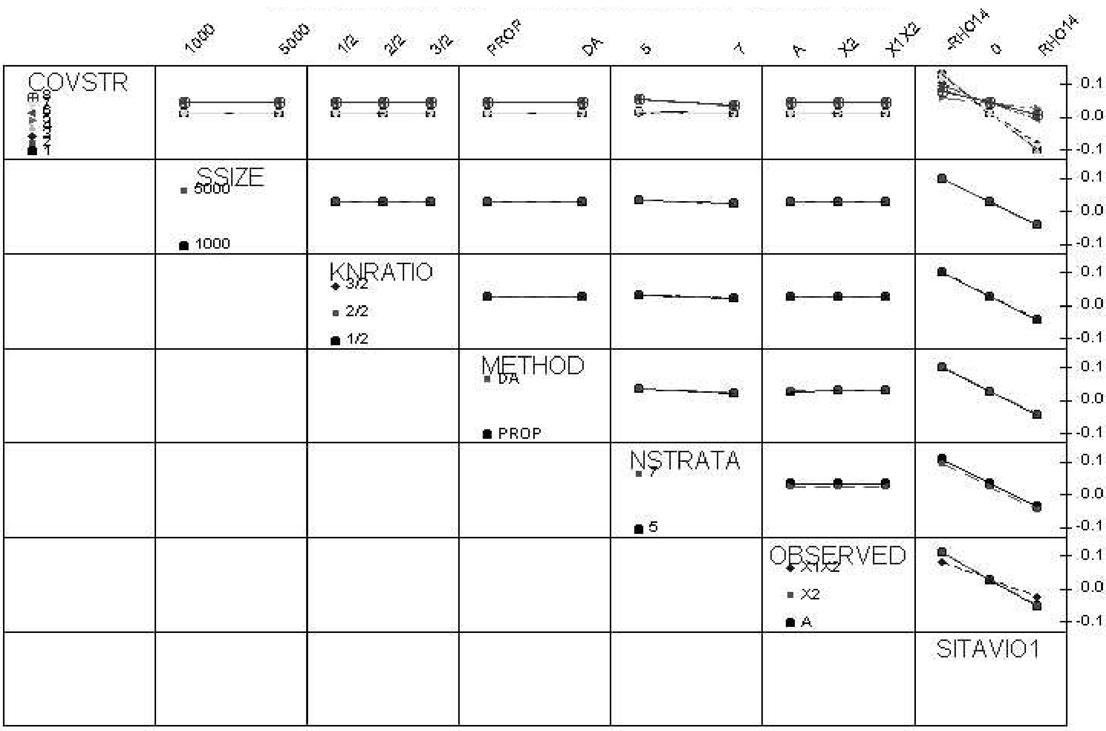

Figure 7: Case 2, interaction plot for *MeanBias*. (Identification of the COVSTR levels given in the text.)

For the level METHOD=A, *StDiff* was almost correct irrespective of the covariance structure and of the $\rho_{.4}$, but the confidence level was approximately correct in only two of the six cases: it was approximately correct when $\rho_{.4} = 0$ for the covariance structures with low correlation between the participation variable X_2 and the study variable Y (the structures 1, 3, 4, and 7), and it was approximately correct when $\rho_{.4} = .175$ for the covariance structures with high correlation of the two variables (the structures 2, 5, 6, and 8). The latter is result due to the point estimate having been "drawn" towards the correct value by the positive $\rho_{.4}$, as mentioned under the preceding heading, "Point estimation". For all the other combinations of ρ_{23} and $\rho_{.4}$ the confidence level was quite low, between .1 and .4.

When the propensity score was estimated (i.e. METHOD$\neq$A), the variance estimates for the covariance structures 1, 3, 4, and 7 (those with low correlation ρ_{23} between the participation variable X_2 and the study variable Y) were approximately correct for all the three levels of $\rho_{.4}$, while the variance estimates for the structures 2, 5, 6, and 8 (those with high ρ_{23}) were considerably underestimating the true variances for all the three levels of $\rho_{.4}$. The effect on confidence level was even here that the positive $\rho_{.4}$ "drew" the point estimate towards the correct value, enabling even underestimated confidence intervals to cover the true parameter in a somewhat higher proportion of outcomes then when $\rho_{.4} = 0$.

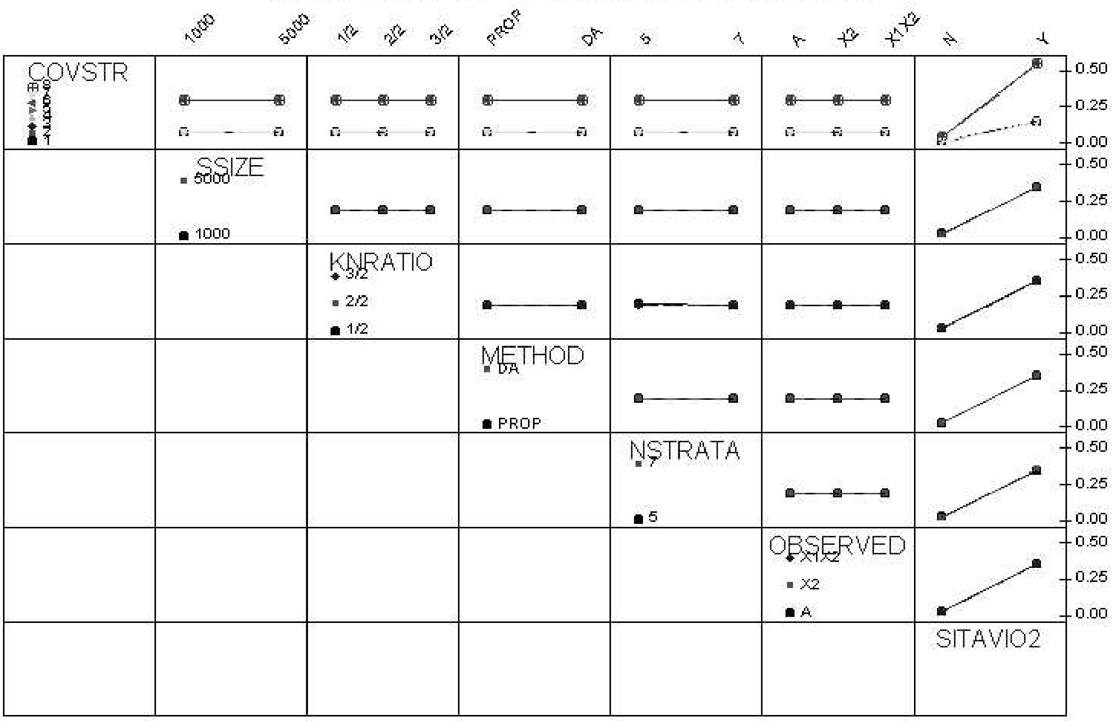

Figure 8: Case 3, interaction plot for *MeanBias*. (Identification of the COVSTR levels given in the text.)

3.4 Case 3: SITA violated, not all units given a positive probability to appear in r

When the determining property of the subset was $Z = I_{V<X_2}$, a unit in the subset may have taken on any value of X_2, it was only more probable that it would have taken on a higher X_2 value then what a unit in the population would, on the average. When, on the other hand, the determining property of the subset was $Z = I_{\max(V,0)<X_2}$—as in the alternative termed "SITA violation #2"—no unit in the subset may have taken on a negative value of X_2. In other words, units with the negative X_2 had no chance to appear in the sample from the subset, r. While far from the only such rule, in what follows the effects of just the rule $I_{\max(V,0)<X_2}$ were investigated.

3.4.1 Case 3: Point estimation

That only the units with positive X_2 values could appear in the restricted sample r had a big effect on the adjusted point estimates of the mean of Y in the population, which was due to the correlation between X_2 and Y. Accordingly, the covariance structures with high ρ_{23} (i.e., the structures 2, 5, 6, and 8) were more biased than those with low ρ_{23}.(Figure 8).

It is interesting to note that not much of the unadjusted bias could in this case of SITA violation be corrected by the propensity score technique. For the covariance structures with high ρ_{23}, the reduction was from .622 to .553, that is, 11%, and for the covariance structures with low ρ_{23}, it was from .175 to .156, again 11%. This held irrespective of

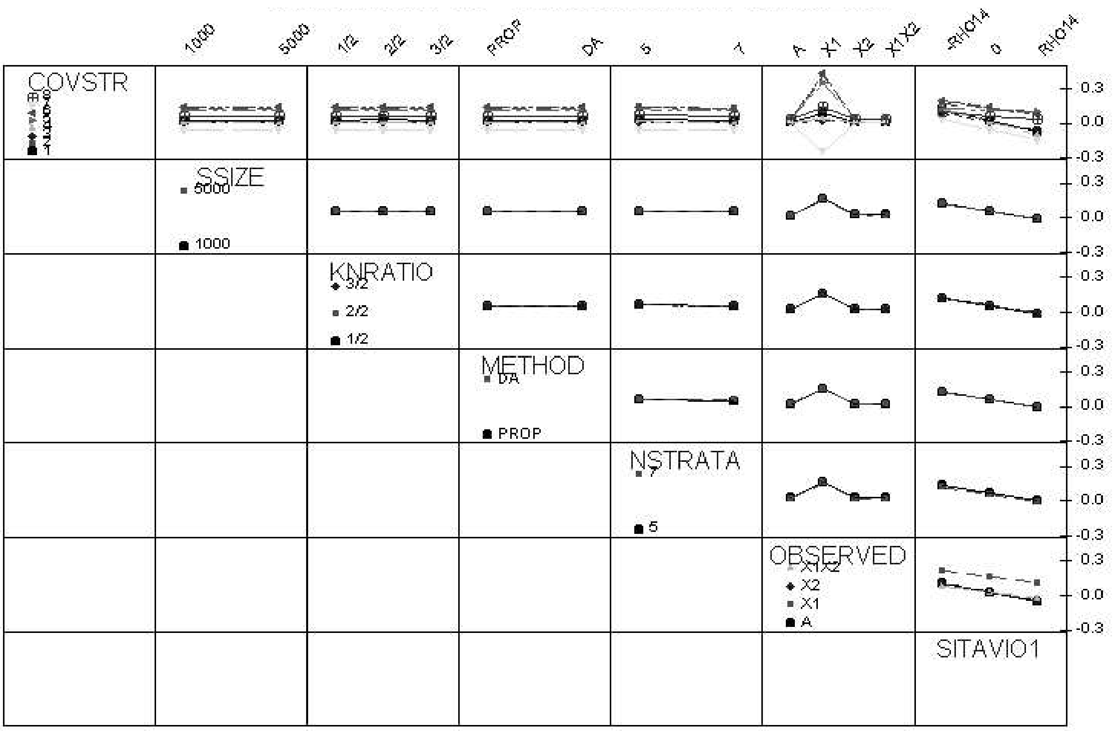

Figure 9: Case 4, interaction plot for *MeanBias*.

whether the true or the estimated propensity score was used.

The factor SITAVIO2 was by far the dominating one in the ANOVA decomposition for *MeanBias* (Table XIII in the Appendix).

3.4.2 Case 3: Variance estimation

With the current violation of SITA, variance estimates underestimated the true variance of the point estimators: the mean of *StDiff* for SITAVIO2=YES, across all the other factors, was $-.0145$ (Table XVI in the Appendix). This, coupled to grossly biased point estimates, yielded empirical confidence levels that were off the mark practically all the time.

3.5 Case 4–Case 6: Pairwise combinations of the SITA violations

The cases 4-6 present pairwise combinations of the SITA violations investigated under Case 1 – Case 3. The results are reported only for *MeanBias*, in Figures 9–11 (but the details concerning all the simulation statistics can be found in the Appendix).

The main conclusion could be drawn that also with combinations of the SITA violations, the adjusted point estimates of the mean of Y in the population were in general badly biased. But, in some cases of interaction of the factor levels, these estimates were less biased than without such an interaction. Two identified sources for this were the level RHO14 of the factor SITAVIO1 and the level OBSERVED=X1. Regarding the former, the impact of RHO14 was to“draw” the adjusted point estimates in the negative direction

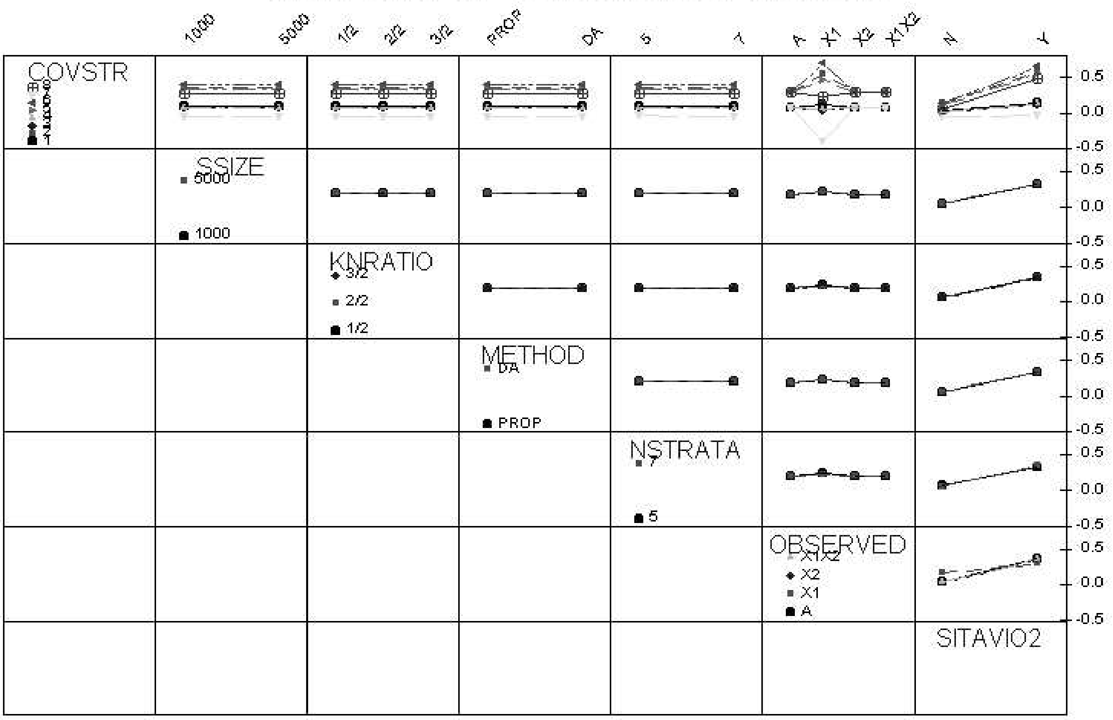

Figure 10: Case 5, interaction plot for *MeanBias*.

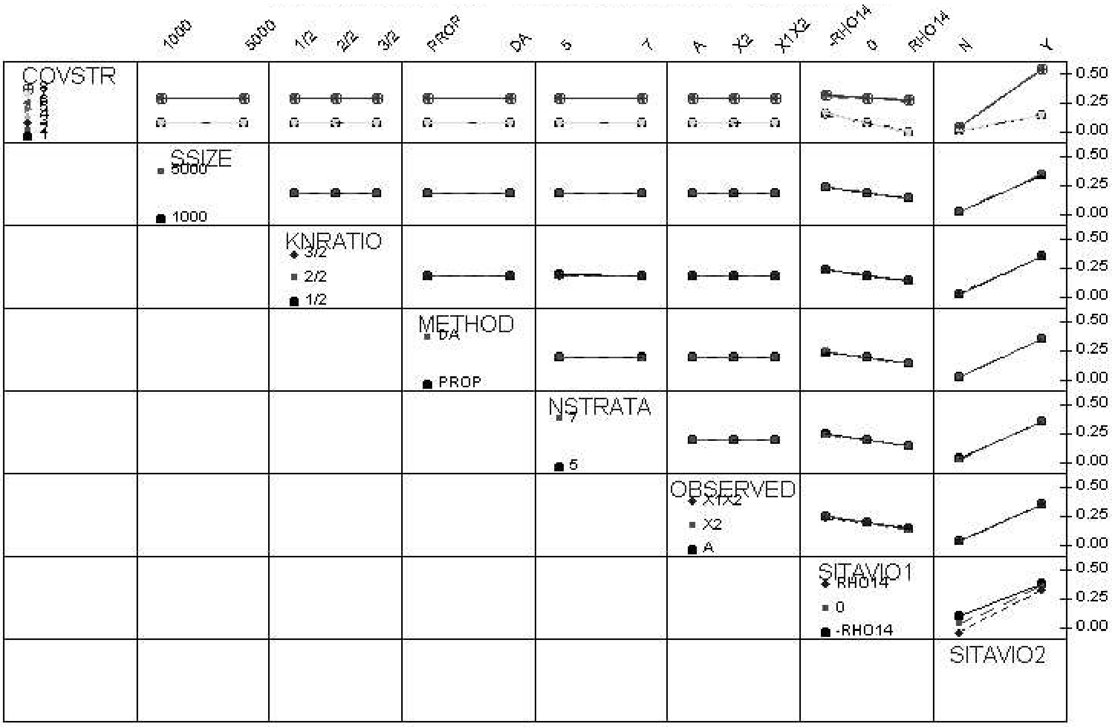

Figure 11: Case 6, interaction plot for *MeanBias*.

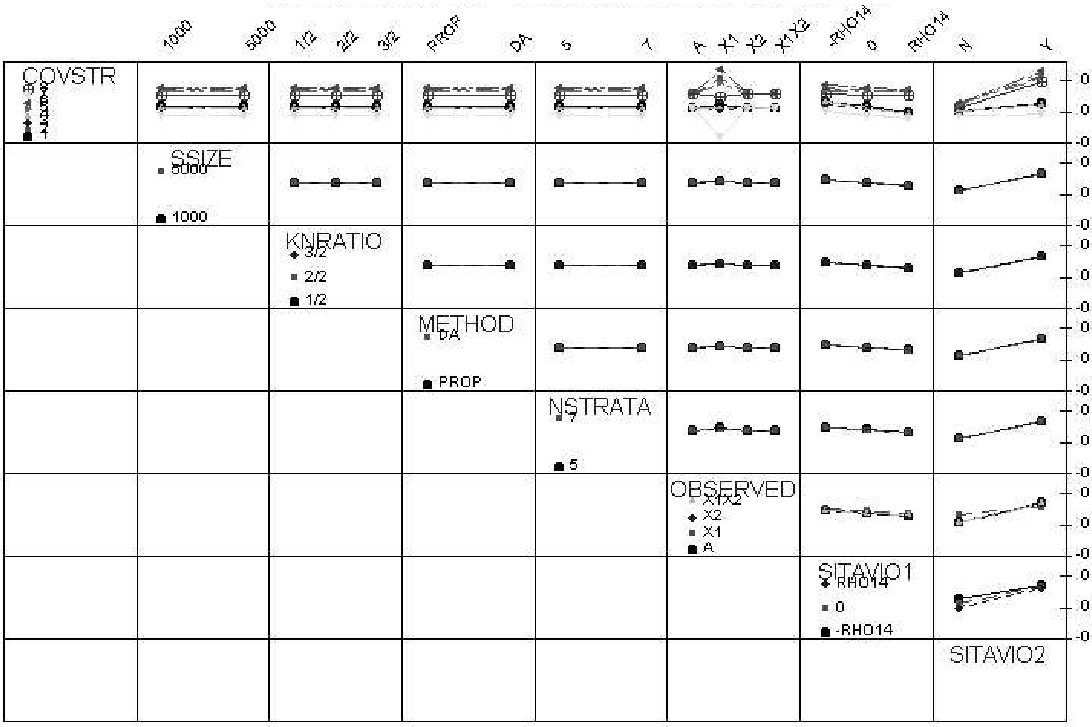

Figure 12: Case 7, interaction plot for *MeanBias*. (Identification of the COVSTR levels given in the text.)

thus having the potential to reduce the remaining bias of any overestimating adjusted point estimate. Regarding the latter, OBSERVED=X1 coupled with covariance structure 7 led to major underestimation and thus it too had the potential to reduce the bias of an overestimating point estimate.

How beneficial would the effect of these sources together be is something predictable only provided all the relevant information on the exact nature of the SITA violations is known in advance—which is difficult to achieve in practice. In other words, while under some fortuitous occasions violation of the SITA assumptions could actually reduce the bias, it is difficult to determine, without the complete knowledge, for some situation at hand whether it is such an occasion or not.

3.6 Case 7: All the violations at the same time

The preceding remarks concerning the influence of the pairwise violations of the SITA assumptions on the point estimates were applicable also when all the violations were introduced into the same experiment (Figure 12). Of interest remained to compare the relative contributions of the studied factors in this case when they all were present.

The comparison was done using the ANOVA decomposition for *MeanBias* (Table XXIX in the Appendix). In Table 5, the 11 most important effects from this table are presented in the descending order of their F-statistics values; together, these effects accounted for the 98% of the observed sums of squares.

The two factors with the greatest impact on the bias of the adjusted point estimate were SITAVIO2 and COVSTR, in this order, while their interaction had the third largest

Table 5: Case 7, partial reproduction of the ANOVA decomposition for *MeanBias* (Table XXIX in the Appendix): the effects with P<.05 presented in an approximate increasing order of magnitude (the first 9 P-values were evaluated as 0 by the *F cdf* function of a mathematical program with precise calculation (Maple), why they are given in the order where main effects come first, followed by their interactions).

Source	DF	Seq SS	Adj SS	Adj MS	F	P
SITAVIO2	1	86.8276	86.8276	86.8276	62000	0
COVSTR	7	110.0035	110.0035	15.7148	11000	0
SITAVIO1	2	6.6733	6.6733	3.3366	2387.23	0
OBSERVED	3	1.0825	1.0825	0.3608	258.15	0
COVSTR*SITAVIO2	7	39.2281	39.2281	5.604	4009.44	0
OBSERVED*SITAVIO2	3	8.3244	8.3244	2.7748	1985.26	0
COVSTR*OBSERVED	21	52.8658	52.8658	2.5174	1801.11	0
SITAVIO1*SITAVIO2	2	1.2206	1.2206	0.6103	436.65	0
COVSTR*SITAVIO1	14	2.5494	2.5494	0.1821	130.28	0
OBSERVED*SITAVIO1	6	0.2637	0.2637	0.044	31.45	0
NSTRATA*SITAVIO2	1	0.0592	0.0592	0.0592	42.36	0
METHOD*SITAVIO1	2	0.0197	0.0197	0.0099	7.06	0.001
NSTRATA	1	0.0078	0.0078	0.0078	5.57	0.018

impact. SITAVIO1 had the next largest impact, followed by some of the interactions of the factor OBSERVED and then this factor by itself together with some other of the pairwise interactions.

It is noteworthy that two of the "regular" factors—not involving violations of the assumptions for the propensity score technique—showed a large influence on the adjusted point estimate even when the assumptions were not fulfilled. These were COVSTR and OBSERVED, emphasizing the two facts: (a) that the true (and not known beforehand) covariance structure of the data at hand may have a profound effect on the adjusted point estimate, and (b) that failure to observe all the relevant information may too have a profound effect on this estimate.

4 Conclusions

This simulation of the propensity score adjustment technique partially demonstrated the practical viability of the approach and partially investigated the effects of certain factors whose influence it was not possible to express in a closed statement. An important aspect concerning the latter goal was the behaviour of the propensity score adjusted point estimate under violations of the assumptions required for the technique to work optimally. The following summary of the results starts with the situation when the assumptions held.

4.1 The factors conforming to the SITA assumptions

4.1.1 Method

The only factor that in the present study had no significant effect on the remaining bias of the adjusted point estimate was the method for estimation of the propensity score: applying logistic regression versus discriminant analysis resulted in only negligible differences between the point estimates. It was already noted that the two methods yield theoretically the same results under the models chosen for this study: only the differences related to the estimation algorithms were eventually expected to show up—these turned out to be practically nonexistent. With other models, the outcome might have been different.

4.1.2 Observed variables

The rest of the factors did have a significant effect. Whether the distribution of the true propensity score was known, or estimated using the variable that "really" determined the participation, or using this variable plus another variable, did have an increasing effect on the remaining bias of the resulting point estimate.

Even interaction of this factor with the sample size was significant: the aforementioned effect was noticeable with the smaller sample size but had practically vanished with the larger sample size.

4.1.3 Balance of the samples

Next, the ratio of the sizes of the samples s and r did have an effect on the remaining bias of the adjusted point estimate. Increasing this factor—from the restricted sample r being half the size of s, to being as large as s, to being one and a half the size of s—led to the reduction of the bias of the point estimate. An increase in this factor always implied access to more data, which resulted in more accurate point estimates. Another candidate explanation, needing further exploration, would be that with more data available, the chances for a stratum being empty in the restricted sample were smaller, leading to a smaller bias of the resulting estimate: the significant interaction between this factor and the sample size, with much smaller effect on the high sample size level, possibly gives support to this interpretation.

4.1.4 Sizes of the samples

The sample size was a factor whose increase led to a smaller bias of the resulting point estimates, with the straightforward interpretation that with more data the estimates became more precise, as well as that the chance of a stratum being empty in the restricted sample was smaller with larger sample sizes.

4.1.5 Number of strata

Finally, the two factors with the largest impact on the remaining bias of the adjusted point estimate were the number of strata into which to classify the observations and the covariance structure of the data at hand. Both were theoretically expected, as it was demonstrated in for instance (Lorenc, 2003). Interestingly, the interaction NSTRATA×SSIZE proved to be insignificant contrary to the expectation that with the small sample size the larger number of strata would more frequently lead to empty strata in the restricted sample and thus to larger bias of the resulting point estimates, compared to the large sample size. Possibly, with a further reduction of the sample size, the effect would eventually show up.

4.1.6 Covariance structure

The remaining bias of the point estimate for the 8 covariance structures was dependent on the correlation between the "participation" variable and the study variable. But, this correlation also determined the original bias: it was higher for the structures with the higher correlation, and lower for those with the lower correlation. So, the propensity score adjustment did in fact have in both cases (i.e., with both high and low ρ_{23}) the same percentage reduction in bias.

4.2 The factors violating the SITA assumptions

The remaining three factors—whether all the relevant auxiliary information was observed, whether the study variable Y could be "untangled" from the subset indicator Z, and whether all the units in the population had a positive chance to appear in the restricted sample r—had substantial effects on the remaining bias of the adjusted point estimator. As the levels of particularly the latter two of the factors (SITAVIO1 and SITAVIO2, respectively) were set rather arbitrarily, their impact in the present study should be taken as an illustration of their potential rather than as something to directly generalize to other situations.

4.2.1 Not all relevant information observed

The effect of not observing all relevant auxiliary information—more precisely, of missing the participation variable—proved particularly interesting in the case when the available auxiliary information was highly correlated with both the participation variable and the study variable but the correlation of the latter two was poor. In this case, a strong overadjustment occurred, resulting in almost doubling the original bias.

4.2.2 Y and Z correlated after all relevant information observed

Peculiar to the case of a remaining correlation between the study variable and the subset indicator after conditioning on the auxiliary information was that, dependent on the sign of the correlation, this could increase or decrease the remaining bias of the adjusted point estimator. If there was a remaining bias in this estimator, as in general is the case with the propensity score adjusted estimates, then under "lucky" occasions the remaining correlation may have reduced the bias of the adjusted estimator, compared to the case with no remaining correlation. Unfortunately, the researcher in real applications has no such deep knowledge of the variables of relevance (and in particular of the study variables) to be able to count on this: had this knowledge existed, no study would have been needed in the first place. Thus, counting on this effect would in general be like “shooting in the dark”.

4.2.3 Not all units had a positive chance to appear in r

That the factor SITAVIO2 had a profound effect in the study depended to a large degree on the rough alternative inclusion function chosen, $Z = I_{\max(V,0)<X_2}$. There are infinitely many other functions that could be conceived, with a varying effect on the remaining bias of the point estimate.

4.3 Summary

This study demonstrated the effectiveness of the propensity score weighting technique in the situation with known properties of the underlying population, and investigated the effect of some of the factors plausibly in effect even in real-life studies, thus providing some impression of what might be expected in practical applications of this same technique. The study proved practically that the propensity score weighting works, reducing most of the bias in the situation when the assumptions that pertain to the technique hold. It further demonstrated the relative sensitivity of the resulting estimates to variation of both the factors internal to the technique (like sample sizes, ratio of the sample sizes, number of strata, etc.) and the factors that determine the nature of data to which the technique is applied (like covariance structure and the violations of the assumptions).

Acknowledgement

Support for this study from the Bank of Sweden Tercentenary Foundation, Grant no. 2000-5063, is gratefully acknowledged.

References

[1] Cochran, W.G. (1968). “The effectiveness of adjustment by subclassification in removing bias in observational studies”. *Biometrics*, 24:205-13.

[2] Cochran, W.G. and Rubin, D.B. (1973). “Controlling bias in observational studies: a review”. *Sankya*, ser. A, 35:417-46.

[3] Lorenc, B. (2003). “Effectiveness of weighting by stratification on the propensity score using double samples”. Research report 2003:10. Department of statistics, Stockholm university.

[4] Mosteller, F. and Tukey, J.W. (1977). *Data Analysis and Regression: A Second Course in Statistics.* Reading, MA: Addison-Wesley.

[5] Rosenbaum, P.R. and Rubin, D.B. (1983a). ”The central role of the propensity score in observational studies for causal effects”. *Biometrika*, 70:41-55.

[6] small Rosenbaum, P.R. and Rubin, D.B. (1983b). ”Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome”. *Journal of the Royal Statistical Society*, ser. B, 45:212-18.

[7] Rosenbaum, P.R. and Rubin, D.B. (1984). ”Reducing bias in observational studies using subclassification on the propensity score.” *Journal of the American Statistical Association*, 79:516-24.

[8] Terhanian, G., Marcus, S., Bremer, J., and Smith, R. (2001). ”Reducing error associated with non-probability sampling through propensity scores: evidence from election 2000”. *Joint Statistical Meeting 2001*, August 5-9, 2001, Atlanta, Georgia, USA.

[9] Terhanian, G., Taylor, H., Siegel, J., Bremer, J., and Smith, R. (2001): “The Accuracy of Harris Interactive's Pre-Election Polls of 2000”. *AAPOR 2001 Annual Conference*, 17-20 May 2001, Montreal, Quebec.