Research Report

Department of Statistics

No. 2003:10

Effectiveness of Weighting by Stratification on the Propensity Score Using Double Samples

Boris Lorenc

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

Effectiveness of Weighting by Stratification on the Propensity Score Using Double Samples

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

Abstract

This study demonstrates the validity of Terhanian's suggestion to apply the propensity score weighting technique to the situation of drawing two samples from the same population. First, a simple multivariate population model with a specified covariance structure is introduced. A subset of the population, conforming to the requirements of *strongly ignorable treatment assignment*, is then defined through a relation between two of the model's variables. That is, inclusion into the subset is made stochastically contingent on an auxiliary variable. Two samples from the population are conceived, making it a case of a double samples procedure for surveys: both simple random samples, one from the general population (also, *unrestricted* sample) and the other from the subset (also, *restricted* sample). If a population point estimate concerning the study variable is given based only on values of the variable in the restricted sample, without an attempt to correct for the skewed inclusion probabilities into the subset, the estimate is biased. This is exhibited by deriving expectations for the population and the subset. Then balancing of the auxiliary variable and of the study variable are demonstrated using first the "trivial" balancing score, namely the auxiliary information itself, and then the propensity score (the two turning out to be the same in the univariate case). Graphical illustration of the procedure is given using the regression approach, by which the differences between the propensity score technique in the double samples setting and in the usual setting of two different populations are highlighted.

The study is performed in a model-based framework.

Introduction

The propensity score weighting technique was introduced by Rosenbaum and Rubin (1983) to enable proper inference about treatment effects with data from observational studies rather than from experiments. Typically, a treatment (say, wearing seat belts) is confounded by an auxiliary[1] variable (say, speed at which one drives: people not wearing

[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

seat belts were apparently also prone to driving faster). Thus, to properly evaluate the effect of wearing seat belts on the degree of injury in the case of accidents (here, the study variable), the observed difference in the study variable needs to be adjusted so as to take into account the differing level of the auxiliary variable. The original work for one auxiliary variable was carried out by Cochran (1968). Rosenbaum and Rubin extended the idea to the multivariate case. To do so they introduced and defined the propensity score,

$$e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{X} = \mathbf{x}).$$

Here, $\mathbf{x}$ denotes the observed vector of values of a multivariate auxiliary variable $\mathbf{X}$ for a unit, and Z indicates unit's inclusion into the treatment group, given that it is included in the study. Descriptively, the propensity score is "a scalar function of the auxiliary variables that appropriately summarizes the information required to balance the distributions of the auxiliary variables"[2]. In order for the propensity score technique to balance the distributions and to correctly adjust the treatment effect, two assumptions jointly termed *strongly ignorable treatment assignment* (given in detail below, p. 13) are needed.

As originally conceived, the propensity score technique balances for differences in auxiliary variables between *two populations*. Terhanian (e.g. Terhanian, Marcus, Bremer, Smith, 2001) saw that, with minor adjustments, the technique can be applied to balancing differences in auxiliary variables when having two samples drawn from the *same population*, but with different sampling procedures. One is an ordinary probability sampling from the population using known, positive inclusion probabilities, which gives the *unrestricted* sample. The other sample, the *restricted* sample, comes from a subset of the population. No explicit sampling into this sample is performed; the inclusion propensities, while unknown, are presupposed to be contingent on auxiliary information or on the study variables, or on both.

Primary motivation for using the approach that Terhanian suggested lies presumably in greater ease, higher speed and lower cost of conducting web surveys compared to any other known mode of data collection. While it is relatively easy to build a panel of web respondents, these are not a random sample in the usual meaning of the word from the target population. They often differ from the target population (usually, the general population) in demographic variables like age, income, etc.; in addition, self-selection is present to a considerable degree, which may introduce an additional bias. The road Terhanian took was simply to adjust for these differences, using a random sample from the target population as an indication of how the auxiliary variables really ought to be distributed in the population.

The study variables *need not* be collected from the unrestricted sample: this sample is used to derive adjustment weights. Once produced, the weights may be reapplied to new surveys of the web panel as long as it is believed that the panel or the target population have not changed enough to justify a new derivation of the weights. In what follows, it is assumed that no data exist for the values of the study variables in the unrestricted sample.

---

the econometric literature as *conditioning* variables.

[2] The term *balance* is given a formal meaning below, p. 10.

A comment is in place. Modelling in the present study the situation considered by Terhanian relies on two tenets: *(a)* there is a single population with a (stochastically defined) subset, and *(b)* sampling designs for samples both from the population and the subset are known (i.e., the simple random sampling) but not the defining properties of the subset (i.e., the relations of the auxiliary variables that characterize the subset). In reality, neither are the inclusion probabilities for units in the restricted sample known, nor is, probably, the important assumption of the subset being a stochastic one fulfilled. Nevertheless, in order to model and run simulations, some assumptions need to be made (robustness to assumption misspecification may later be tested). Other choices than those above exist, for instance in a to view web users and non-users as two mutually exclusive populations and in b to view the units in the restricted sample as sampled by unknown inclusion probabilities. There seems nothing to preclude these choices, provided appropriate changes in the modelling procedure and in the estimation goal are made. The decisions reached here were driven by a plausible view on the physical reality of the phenomenon (the decision in a) and by the goal to estimate the properties of the subset rather than to estimate the unknown inclusion probabilities (the decision in b). Namely, a preliminary study has shown that estimation of inclusion probabilities yields an estimator (irrespective of whether HT or the regression estimator is used) that, even if practically unbiased, has quite a large variance.

While the work of Terhanian and his colleagues was made public in descriptive form (e.g. Terhanian, Marcus, et al., 2001; Terhanian, Taylor, Siegel, Bremer, and Smith, 2001), there are to the best of my knowledge no published formal presentations of the technique and of the details of its procedure. It is to the goal of giving a formal exposition of the technique, together with a simple demonstration of its working, that the present text is dedicated. Section 1 introduces a multivariate model of the population and defines, through a relation of two of the variables, a subspace of the sample space. Some population and subspace expectations are derived here for the use in later sections. Section 2 presents a class of scores known as balancing scores and demonstrates balancing of the distribution of an *auxiliary* variable between the population and the subspace by conditioning on a balancing score, namely the variable itself. In Section 3, balancing between the population and the subspace of a *study* variable is performed by conditioning on the propensity score. Section 4 presents the difference between the previous work and Terhanian's approach in terms of regression functions and gives a graphical illustration of the differences. Some final thoughts and suggestions for further work are given in Section 5.

1 The model

The model chosen for the demonstration is a multivariate normal model. The reasons for using this particular kind of model are two: its analytical tractability and its previous use in a similar case (Cochran, 1968). Given that the study is of a demonstrative character, as well as that a normal model may in at least a number of practical situations be applicable, this need not be a serious limitation.

We have the model

$$(X, Y, V) \sim N_3\left(0, 0, 0, \mathbf{\Sigma}\right), \tag{1}$$

where $\mathbf{\Sigma}$ is the covariance matrix with the structure

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The variables are given the following meanings, common in the survey literature:

X, an auxiliary variable,
Y, the study variable,
V, a standard normal variable independent of the previous two.

Let a large number ("population") of independent and identically distributed observations be generated by the model, a single observation pertaining to unit i—realised from the stochastic vector $[X_i, Y_i, V_i]$—being $[x_i, y_i, v_i]$. Next, we define a subspace of the model's sample space by using the relation $V < X$, and introduce an indicator variable Z such that $Z = I_{V<X}$. There is nothing special about the relation $V < X$, but it—as many other relations with the similar composition would—defines a *stochastic* subspace. That is, units satisfying this relation are not confined to only a subspace of X; rather, they occur throughout the whole sample space of X, but with the probability of occurrence related to X. In contrast, the condition e.g. $0 < X$ would restrict the sample space of units with $Z = 1$ to only the positive values of X. In terms of the introductory example, the relation $V < X$ ensures that at least some seat belt wearing drivers drive very fast and at least some drivers not wearing seat belts drive very slow. With the current definition, Z complies to a requirement for strongly ignorable treatment assignment (Rosenbaum and Rubin, 1983; see also p. 13 below).

1.1 The task

Let R denote the set of units in the restricted sample space, so that a unit $i \in R$ iff $z_i = 1$ (but, see the discussion that follow immediately). Let r denote a simple random sample from R, size of this sample denoted by k. And, let s denote a simple random sample from the population, size of this sample denoted by n. It ought to be observed that, with final populations, neither $s \cap r$ nor $s \cap R \cap r^C$ need be empty: that i is an element of s (sampled from the population) does not preclude that the same element is in the subset R or that it is in the sample r (sampled from the subset). Practically, for the case $i \in s$ and $i \in r$ for the same i, such a unit appears twice in the data material: once as a member of s without a y value and with a $z_i = 0$ attached, and once as a member of r with complete observation and with a $z_i = 1$ attached. Also, practically, for the case $i \in s$ and $i \in R \cap r^C$ for the same i, such a unit appears in s without a y value and with a $z_i = 0$ attached. (To take care of this situation formally, a variable Z^* would have needed to be defined, $z_i^* = 0$ iff $i \in s$, $i = 1, ..., n$, and $z_i^* = 1$ iff $i \in r$, $i = n + 1, ..., n + k$. Z^* would indicate the sample through which a vector of observations came to be collected: unrestricted or restricted sample. But, as $Z \to Z^*$ when population size tends to infinity

Table 1: The observed data: a sample s of n units drawn from the population, with data missing on Y, and a sample r of k units drawn from a subset of the population, having complete information. $Z = 0$ indicates inclusion in s and $Z = 1$ indicates inclusion in r.

Obs.	X	Y	Z
1	x_1	$-$	0
2	x_2	$-$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	$-$	0
$n+1$	x_{n+1}	y_{n+1}	1
$n+2$	x_{n+2}	y_{n+2}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n+k$	x_{n+k}	y_{n+k}	1

and the sample sizes are kept constant, this distinction was not explicitly made in the present study.)

For all the sampled units (both those in s and in r), values on the auxiliary variable X are recorded. For units in the restricted sample r even values on the study variable Y are recorded (Table 1). The task is to estimate, based on the recorded data, the expected value, $E(Y)$, of the study variable Y in the population. As is shown under the next heading, for any $\rho \neq 0$, the naive approach of taking $\bar{Y}_r = \frac{1}{k}\sum_{i\in r} Y_i$ as the estimator of $E(Y)$ is biased, as r is not a simple random sample from the population but from the subset. It is for the same reasons biased to take $\bar{X}_r = \frac{1}{k}\sum_{i\in r} X_i$ to estimate $E(X)$, the task we will meet first.

1.2 The joint, conditional and marginal distributions and expectations

Material in this section is mainly a recapitulation of some properties of the multivariate normal distribution, and can be found in sources like Johnson and Kotz (1972). Some particular distributions and their expectations, used in the present and the following sections, are derived here, showing biasedness of the naive estimators, those taken directly from the r sample.

The model in (1) allows for the following decomposition:

$$\begin{aligned}(X, Y) &\sim N_2(0, 0, 1, 1, \rho),\\ V &\sim N(0, 1),\end{aligned}$$

with V and (X, Y) independent, $V \perp (X, Y)$. With $\varphi(\cdot)$ denoting the standard normal density, we can write the joint density of (X, Y) as

$$\begin{aligned}
\varphi_\rho(x,y) &= \varphi(x)\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right) \\
&= \varphi(y)\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right) \\
&= (2\pi)^{-1}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right)
\end{aligned} \tag{2}$$

why we can write the trivariate density of the model in (1) as

$$\begin{aligned}
f(x,y,v) &= \varphi_\rho(x,y)\,\varphi(v) \\
&= (2\pi)^{-\frac{3}{2}}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right)\exp\left(-\frac{1}{2}v^2\right) \\
&\qquad \begin{cases} -\infty < x < \infty \\ -\infty < y < \infty \\ -\infty < v < \infty\ . \end{cases}
\end{aligned} \tag{3}$$

As the multivariate normal distribution in (3) has zero means and unit variances, its component variables (X, Y, V) are marginally standard normal. Thus for Y, the study variable which we are especially interested in, $E(Y) = \mu_y = 0$ in the population. Also, for the auxiliary variable, $E(X) = \mu_x = 0$.

The above applies to the population and, in expectation, to s, the simple random sample from this population. Properties related to the subset of the population defined by $V < X$ and, in expectation, to r—the simple random sample from that subset—are now derived by conditioning the trivariate density $f(x,y,v)$ on the event $Z = 1$, that is, on the event $V < X$.

Due to independence of (X,Y) and V, the joint distribution of (X, Y, V) becomes in this restricted case:

$$\begin{aligned}
g(x,y,v) &= 2f(x,y,v) \\
&= 2\varphi_\rho(x,y)\,\varphi(v) \\
&= 2(2\pi)^{-\frac{3}{2}}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}\right)\exp\left(-\frac{1}{2}v^2\right) \\
&\qquad \begin{cases} -\infty < x < \infty \\ -\infty < y < \infty \\ -\infty < v < x\ . \end{cases}
\end{aligned}$$

The conditional joint and marginal distributions and expectations of X and Y follow. First, the joint distribution of (X,Y) in the subset:

$$\begin{aligned}
g(x,y) &= \int_{-\infty}^{x} 2\varphi_\rho(x,y)\,\varphi(v)\,dv \\
&= 2\varphi_\rho(x,y)\,\Phi(x)\,,
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The distribution of X in the subset:

$$\begin{aligned} g(x) &= \int_{-\infty}^{\infty} g(x,y)\,dy \\ &= 2\int_{-\infty}^{\infty} \varphi_\rho(x,y)\,\Phi(x)\,dy \\ &= 2\varphi(x)\,\Phi(x) \end{aligned} \tag{4}$$

Using the relation $t\varphi(t) = -\varphi'(t)$, the expectation of X in the subset, $E_g(X)$, is then obtained:

$$\begin{aligned} E_g(X) &= 2\int_{-\infty}^{\infty} x\varphi(x)\,\Phi(x)\,dx \\ &= -2\int_{-\infty}^{\infty} \Phi(x)\,d\varphi(x) \\ &= 2\int_{-\infty}^{\infty} \varphi(x)\,d\Phi(x) \\ &= 2\int_{-\infty}^{\infty} \varphi(x)^2\,dx \\ &= \sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty} \varphi\left(\sqrt{2}x\right)dx \\ &= \pi^{-\frac{1}{2}}. \end{aligned} \tag{5}$$

The expectation $E_g(X)$ is equal to $\pi^{-\frac{1}{2}}$, approximately .564, and this holds also for the expectation of the mean of r, $E\left(\bar{X}_r\right)$, because r is a simple random sample from the subset. The unadjusted r sample mean is thus biased in expectation as an estimator of $E(X) = E\left(\bar{X}_s\right) = 0$, with a bias of approximately .564.

To obtain the variance of X in the subset,

$$
\begin{aligned}
E_g\left(X^2\right) &= 2\int_{-\infty}^{\infty} x^2\varphi\left(x\right)\Phi\left(x\right)dx \qquad (6)\\
&= -2\int_{-\infty}^{\infty} x\Phi\left(x\right)d\varphi\left(x\right)\\
&= -\left[2x\Phi\left(x\right)\varphi\left(x\right)\right]_{-\infty}^{\infty} + 2\int_{-\infty}^{\infty}\varphi\left(x\right)d\left[x\Phi\left(x\right)\right]\\
&= 2\int_{-\infty}^{\infty}\varphi\left(x\right)\left[x\varphi\left(x\right)+\Phi\left(x\right)\right]dx\\
&= 2\int_{-\infty}^{\infty} x\varphi\left(x\right)^2 dx + \int_{-\infty}^{\infty} d\Phi\left(x\right)^2\\
&= \sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty} x\varphi\left(\sqrt{2}x\right)dx + 1\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} t\varphi\left(t\right)dt + 1\\
&= 1,
\end{aligned}
$$

from which,

$$
\begin{aligned}
Var_g\left(X\right) &= E_g\left(X^2\right) - \left[E_g\left(X\right)\right]^2 \qquad (7)\\
&= 1-\pi^{-1}.
\end{aligned}
$$

As the study variable Y, through its correlation with X, is not independent of $Z = I_{V<X}$ when $\rho \neq 0$, the distribution of Y in the subset is of interest. It is obtained as:

$$
\begin{aligned}
g\left(y\right) &= \int_{-\infty}^{\infty} g\left(x,y\right)dx \qquad (8)\\
&= 2\int_{-\infty}^{\infty}\varphi_\rho\left(x,y\right)\Phi\left(x\right)dx\\
&= 2\varphi\left(y\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)\Phi\left(x\right)dx\\
&= 2\varphi\left(y\right)\int_{-\infty}^{\infty}\Phi\left(x\right)d\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)\\
&= 2\varphi\left(y\right)\left[1-\int_{-\infty}^{\infty}\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right)d\Phi\left(x\right)\right]\\
&= \int_{-\infty}^{\infty}\pi^{-1}\frac{1}{\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}\frac{x^2-2xy\rho+y^2}{1-\rho^2}\right)\Phi\left(x\right)\,dx.
\end{aligned}
$$

The expected value of Y in the subset:

$$\begin{aligned}
E_g(Y) &= \int_{-\infty}^{\infty} y g(y)\, dy \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\varphi_\rho(x,y)\,\Phi(x)\, dxdy \\
&= 2\int_{-\infty}^{\infty}\varphi(x)\,\Phi(x)\int_{-\infty}^{\infty} y\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right) dydx \\
&= 2\int_{-\infty}^{\infty}\rho x\varphi(x)\,\Phi(x)\, dx \\
&= \rho\int_{-\infty}^{\infty} x d\Phi(x)^2 \\
&= \rho E_g(X) \\
&= \rho\pi^{-\frac{1}{2}}
\end{aligned}$$

and, using the derivation in (6), the variance of Y in the subset:

$$\begin{aligned}
Var_g(Y) &= E_g\left[(Y)^2\right] - \left[E_g(Y)\right]^2 \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2\varphi_\rho(x,y)\,\Phi(x)\, dxdy - \rho^2\pi^{-1} \\
&= 2\int_{-\infty}^{\infty}\varphi(x)\,\Phi(x)\int_{-\infty}^{\infty} y^2\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right) dydx - \rho^2\pi^{-1} \\
&= 2\int_{-\infty}^{\infty}\varphi(x)\,\Phi(x)\left[1-\rho^2+(\rho x)^2\right] dx - \rho^2\pi^{-1} \\
&= 1-\rho^2+\rho^2-\rho^2\pi^{-1} \\
&= 1-\rho^2\pi^{-1}.
\end{aligned}$$

Finally, also using (6), we obtain the conditional covariance of X and Y:

$$\begin{aligned}
Cov_g(X,Y) &= E_g(XY) - E_g(X)\,E_g(Y) \qquad (9) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, dydx - \pi^{-\frac{1}{2}}\times\rho\pi^{-\frac{1}{2}} \\
&= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\varphi_\rho(x,y)\,\Phi(x)\, dydx - \rho\pi^{-1} \\
&= 2\int_{-\infty}^{\infty} x\varphi(x)\,\Phi(x)\int_{-\infty}^{\infty} y\frac{1}{\sqrt{1-\rho^2}}\varphi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right) dydx - \rho\pi^{-1} \\
&= 2\int_{-\infty}^{\infty}\rho x^2\varphi(x)\,\Phi(x)\, dx - \rho\pi^{-1} \\
&= \rho\left(1-\pi^{-1}\right)
\end{aligned}$$

The expectation of Y in the subset, $E_g(Y)$—equal to the expectation of the mean of the simple random sample r from the subset, $E\left(\bar{Y}_r\right)$—is $\rho\pi^{-\frac{1}{2}} \approx \rho\times.564$. Thus, the

unadjusted mean of Y in the restricted sample r is biased as an estimator of $E(Y) = E(\bar{Y}_s) = 0$ whenever $\rho \neq 0$. Its absolute bias is directly proportional to the magnitude of the correlation coefficient, attaining its maximum at .564 when $\rho = \pm 1$.

The remainder of this text gives a demonstration of the reduction of the bias of the two estimators by stratification: of $\bar{X}_r$ as an estimator of $E(X)$ and of $\bar{Y}_r$ as an estimator of $E(Y)$.

2 Reducing the bias of $\bar{X}_r$ as an estimator of $E(X)$ by stratification on X

Based solely on the sample r, both the unadjusted $\bar{X}_r$ and $\bar{Y}_r$ are biased as the estimates of $E(X)$ and $E(Y)$ respectively. Rosenbaum and Rubin (1983) introduced a broad class of scores termed "balancing scores", whose defining property is:

$$(X \perp Z) \,|\, b(X). \tag{10}$$

That is, a function of X is a balancing score, $b(X)$, if conditioning on the function results in a (conditional) independence of the auxiliary information and the treatment assignment.

There are, in general, many functions of X that are balancing scores, the finest, "trivial" (Rosenbaum and Rubin, 1983) balancing score being X itself and the coarsest being the propensity score, $e(x) = \Pr(Z = 1 \mid X = x)$. Here, *coarsest* denotes that there exists no function of the propensity score (other than the identity function) that also produces a balancing score.

As the simplest case of bias reduction by subclassification (stratification), adjustment of $\bar{X}_r$ using X itself—the "trivial" balancing score—is demonstrated first. The same principle is later applied to adjusting $\bar{Y}_r$ using the propensity score.

From the exposition thus far, $E_g(X) = E(\bar{X}_r) \approx .564$ while $E(X) = E(\bar{X}_s) = 0$. In this section, we adjust $\bar{X}_r$ in order to obtain a less biased estimator of $\bar{X}_s$. An intuitive description of the procedure is given first, followed by a more formal account.

The distribution of X in the restricted sample r is stratified into a number of subclasses, with cutoff points determined by the quantiles of the cumulative distribution function (*cdf*) of X in the population. (It is the use of the *cdf* of X *in the population* that, implicitly, performs the role of the balancing score in the present example.) Usually, cutoff points equidistant in terms of quantiles are chosen, as this allows for easier weighting into the final estimate. For each stratum l, the mean of X in the stratum is calculated, $\bar{X}_{r,l} = \left(\sum_{i \in l} X_i\right) / k_l$. An adjusted estimate of $E(X)$ is then built by suitably weighting the stratum means; with equidistant quantile points, this amounts to taking the arithmetic average of the stratum means.

With L strata, the lowest and the highest quantile points would be $q_0 = 0$ and $q_L = 1$, and, aiming at cutoff points equidistant in terms of quantiles, the remaining $L - 1$ points

$$q_l = l \times L^{-1},\ l = 1, 2, ..., L - 1.$$

To each of the quantile points corresponds a cutoff point, c_l, determined by

$$c_l = \Phi^{-1}(q_l), \tag{11}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the *cdf* of the standard normal distribution.

Expectation of X in the restricted sample r at each level of X is determined using the conditional density $g(x)$ given in (4):

$$\begin{aligned} E_{g,l}(X) &= E\left(\bar{X}_{r,l}\right) = \frac{\int_{c_l}^{c_{l+1}} x g(x)\, dx}{\int_{c_l}^{c_{l+1}} g(x)\, dx} \\ &= \frac{\int_{c_l}^{c_{l+1}} x\varphi(x)\,\Phi(x)\, dx}{\int_{c_l}^{c_{l+1}} \varphi(x)\,\Phi(x)\, dx} \\ &= \frac{\int_{c_l}^{c_{l+1}} x\varphi(x)\,\Phi(x)\, dx}{\frac{1}{2}\left[\Phi^2(c_{l+1}) - \Phi^2(c_l)\right]},\ l = 0, 1, ..., L-1. \end{aligned}$$

With equidistant quantile points, the expression in the denominator may be simplified to $(2l+1)/L^2$. It is not known how to further simplify the expression in the numerator—integration must be performed instead. A numerical example with $L = 5$ is given later on in this section. An estimate of $E(X)$, based on the restricted sample r and adjusted by stratification on X, is then obtained by

$$\widehat{E_f(X)}_{\{r,X\}} = \frac{\sum_{l=1}^{L} E_{g,l}(X)}{L} = \frac{\sum_{l=1}^{L} E\left(\bar{X}_{r,l}\right)}{L}.$$

The same weight is given here to each stratum because the quantiles of the *cdf* are equidistant. So, the weights do not appear explicitly in the expression.

The above procedure relies on the assumption that with sufficiently fine stratification on the balancing score, the conditional stratum expectations of X in the population, $E_{f,l}(X)$, and in the subset, $E_{g,l}(X)$, will within each stratum get close enough to each other; that is, that $E_{g,l}(X)$ will become a sufficiently good proxy for $E_{f,l}(X)$. This is expressed in the following derivation.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\, dx \\ &= \sum_{l=0}^{L-1} \int_{c_l}^{c_{l+1}} x f(x)\, dx \\ &= \sum_{l=0}^{L-1} \int_{c_l}^{c_{l+1}} \frac{\int_{c_l}^{c_{l+1}} x f(x)\, dx}{\int_{c_l}^{c_{l+1}} f(x)\, dx} f(x)\, dx \\ &\approx \sum_{l=0}^{L-1} \int_{c_l}^{c_{l+1}} \frac{\int_{c_l}^{c_{l+1}} x g(x)\, dx}{\int_{c_l}^{c_{l+1}} g(x)\, dx} f(x)\, dx \\ &= \frac{1}{L} \sum_{l=0}^{L-1} \frac{\int_{c_l}^{c_{l+1}} x g(x)\, dx}{\int_{c_l}^{c_{l+1}} g(x)\, dx} \end{aligned} \tag{12}$$

Table 2: Adjustment of $\bar{x}_t$ by stratification on X: the strata means of X in the subset, $\bar{\mathbf{x}}_{tl}$, $l = 1, 2, ...5$, are given equal weights.

Stratum	$E\left(\bar{X}_{r,l}\right)$
1	-1.1496
2	-0.4995
3	0.0168
4	0.5458
5	1.427 6
Mean	0.0682

Approximation of $E_{f,l}(X)$ by $E_{g,l}(X)$ is performed on the fourth row. With known $f(x)$ and $g(x)$, such as in the present demonstration, the quality of the approximation is proportional to the number of subclasses into which the distribution $G(x) = \int^x g(x)\,dx$ is stratified. Justification for the last row in (12) exists when the cutoff points are quantile-equidistant.

An example with $L = 5$ follows.

The strata means are given in Table 2, as well as their mean. This mean of the strata means is an adjusted estimate of $E(X)$ based on X values in the sample r only. The estimate, $\widehat{E_f(X)}_{\{r,X\}} = .068$, is closer to the true value, $E(\bar{x}) = 0$, than the unadjusted estimate $\bar{x}_r \approx .564$. With $L = 7$ the adjusted estimate is 0.045, and with $L = 10$ it is 0.029. Using the measure percentage reduction in bias, θ, (Cochran and Rubin, 1973), namely in this case

$$\theta = 100\left(1 - \frac{\widehat{E_f(X)}_{\{r,X\}} - E(X)}{E\left(\bar{X}_r\right) - E(X)}\right),$$

stratifying X into $L = 5, 7$ and 10 strata reduces the bias by 87.9, 92.0, and 94.9 percent, respectively.

A note on the use of the unrestricted sample

In the above example, no use was made of the unrestricted sample, s. Coupling to the population was done on the theoretical level in (11) in the form of using $\Phi^{-1}(x)$. In many cases, this will not be possible as more complicated functions than X itself will be used as balancing scores. In practice, cutoff points in the population will be estimated from the empirical distribution of the balancing score in the two samples taken together, for which the availability of the explicitly drawn sample s from the population, in addition to the sample r of elements from the subset, is essential.

3 Reducing the bias of $\bar{X}_r$ and $\bar{Y}_r$ as estimators of $E(X)$ and $E(Y)$, respectively, by stratification on the propensity score

3.1 Reducing the bias of $\bar{X}_r$ as an estimator of $E(X)$ by stratification on the propensity score

Now, for the chosen model, stratifying on X as demonstrated in the preceding section coincides both in procedure and numerically with stratifying on the propensity score $e(x)$, where

$$e(x) \underset{def.}{=} \Pr(Z = 1|X = x) = \Pr(V < X|X = x) = \Phi(x). \tag{13}$$

In general, when the joint distribution of X and V is a bivariate normal and the indicator for inclusion into the sample r is a linear expression of the form $bV < cX$, the inclusion probabilities are determined by $\Phi(b^{-1}cX)$.

The reason that here balancing on $e(x)$ is the same as balancing on the "trivial" balancing score X is that partitioning the distribution of X into quantiles and partitioning the distribution of $e(x) = \Phi(x)$ into quantiles give the same cutoff points on the X axis. That is, the reason is the simplified structure of the chosen example. With more complex variable structures this would in general change—in particular when X is multivariate.

3.2 Reducing the bias of $\bar{Y}_r$ as an estimator of $E(Y)$ by stratification on the propensity score

When the goal, in addition to balancing for the differences on the auxiliary variable X between the two samples, s and r, is to adjust an estimator of the study variable Y observed only in r, a pair or requirements jointly known as *strongly ignorable treatment assignment* (Rosenbaum and Rubin, 1983) need to hold for the estimate to be—in expectation and with an infinite number of strata—unbiased:

$$\begin{aligned} &(a) \quad 0 < e(X) < 1 \\ &(b) \quad (Y \perp Z) \mid X, \end{aligned} \tag{14}$$

where $\perp$ denotes independence[3].

The first assumption, that each unit's propensity score need be positive and strictly less than 1, mimics the requirement in experimental studies that each unit has a positive chance to be placed into any of the experimental conditions. Intuitively, the requirement says that on any level of $e(X)$ there need be 'comparable' units, 'similar' to each other on X. The second assumption expresses the need that *all* the information relevant for treatment assignment be present amongst the observed auxiliary information—no further

[3]The requirement in (a), 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.

information should there exist in Z, once X is observed; if it did, there might remain a correlation between Y and Z that then could not be adjusted for by stratifying on $e(X)$.

Verification of the above two assumptions with respect to the artificial population and the two samples from it gives that both are fulfilled. For,

$$(a) \qquad 0 < e(x) = \Phi(x) < 1$$

$$(b) \qquad \left.\begin{array}{l} (X,Y) \perp V \\ Z = h(X,V) \end{array}\right\} \rightarrow (Y \perp Z) \mid X$$

Now, to perform the propensity score weighting: the procedure follows exactly the steps of balancing $G(x) = \int g(x)\,dx$, given above. $E_g(Y) = E\left(\bar{Y}_r\right) = \pi^{-\frac{1}{2}}\rho \approx .564 \times \rho$ is biased with respect to $E(Y) = E\left(\bar{Y}_s\right) = 0$. So, $G(y)$ is stratified into L classes using the quantile points

$$q_l = l \times L^{-1}$$

for $l = 0, 1, ..., L$. Stratifying on $e(x)$, the propensity score, where in (13) it was noted that in the present model $e(x) = \Phi(x)$, we again find cutoff points corresponding to the quantiles using $\Phi^{-1}(\cdot)$, the inverse of the standard normal *cdf*, which in expectation is the distribution of X in the sample s:

$$c_l = \Phi^{-1}(q_l),\ l = 0, 1, ..., L.$$

Expectation of Y in the restricted sample r at each level of $e(X)$ is thus determined, using the conditional density $g(y)$ in (8), by

$$E_{g,l}(Y) = E\left(\bar{Y}_{r,l}\right) = \frac{\int_{c_l}^{c_{l+1}} y \int_{-\infty}^{\infty} \varphi_\rho(x,y)\,\Phi(x)\,dxdy}{\int_{c_l}^{c_{l+1}} \int_{-\infty}^{\infty} \varphi_\rho(x,y)\,\Phi(x)\,dxdy},\ l = 0, 1, ..., L-1.$$

Results of a numerical integration with $L = 5$ taken as an example are given immediately. An estimate of $E(Y)$, based on the restricted sample r and adjusted by stratification on the propensity score $e(X)$, is then obtained by

$$\widehat{E_f(Y)}_{\{r,PS\}} = \frac{\sum_{l=1}^{L} E_{g,l}(Y)}{L} = \frac{\sum_{l=1}^{L} E_{g,l}\left(\bar{Y}_{r,l}\right)}{L}.$$

The same weight is given here to each stratum, as the quantiles of the *cdf* $\Phi(x)$ are equidistant.

The strata means, with $L = 5$ and $\rho = .78$, are given in Table 3, together with their mean. This mean of the strata means, .0532, is an adjusted estimate of $\bar{y}$ based on the values of Y in the sample r only. This estimate, $\widehat{E_f(Y)}_{\{r,PS\}}$ [read: an estimate of the expected value of Y in the population, based on the restricted sample r and adjusted using the propensity score as the balancing score], is closer to the true value than the unadjusted estimate from (8), $E\left(\bar{Y}_r\right) = \rho\pi^{-\frac{1}{2}} \approx .78 \times .564 \approx .440$, based on the same sample. It is shown in the following section that regression of Y on X is linear in the restricted subset R, with ρ as the regression coefficient, so we note that values in the table above could have more easily been obtained by $\bar{Y}_{r,l} = \rho \times \bar{X}_{r,l}$ as well as the final adjusted value by $\widehat{E_f(Y)}_{\{r,PS\}} = \rho \times \widehat{E_f(X)}_{\{r,PS\}} = \rho \times \widehat{E_f(X)}_{\{r,X\}}$. Using this shortcut, with

Table 3: Adjustment of $\bar{y}_t$ by stratification on X: the strata means of Y in the subset, $\bar{\mathbf{x}}_{tl},\ l = 1, 2, ...5$, are given equal weights.

Stratum	$E\left(\bar{Y}_{r,l}\right)$
1	−0.8967
2	−0.3896
3	0.0131
4	0.4256
5	1.1135
Mean	0.0532

$L = 7$ the adjusted estimate is .0351, and with $L = 10$ it is .0226. Percentage reduction in bias, θ, stratifying $G(Y)$ into $L = 5, 7$ and 10 strata, using

$$\theta = 100\left(1 - \frac{\widehat{E_f(Y)}_{\{r,PS\}} - E(Y)}{E\left(\bar{Y}_r\right) - E(Y)}\right),$$

is the same as for $\widehat{E_f(X)}_{\{r,X\}}$ with respect to $E(X)$: 87.9, 92.0, and 94.9 percent, respectively (which also follows from the linear relation of X and Y).

4 The regression approach: a graphical illustration

In Cochran (1968) and Cochran and Rubin (1973), the regression function representation is used in a situation that originally motivated the propensity score weighting, namely that of estimating the average difference between *two* populations, conditional on a confounding variable X. As that situation bears resemblance to the present one, and as the regression approach allows for a nice graphical illustration which hopefully will enhance understanding, in the present section the regression approach is illustrated first for the Cochran and Rubin's case and then for the case of double samples from the same population with one of the samples drawn from a subset of the population.

Cochran and Rubin (1973) consider estimating the treatment effect, $\tau_1 - \tau_2$, which is the difference between the average effects of two treatments on the same level of a variable X. Both X and Z (the treatment assignment) and X and Y are known to be correlated (cf. the seat belts example in the introduction). A simple random sample is drawn from each of the two populations. The simple case where the regressions of Y on X are linear and parallel in both populations, gives the model

$$Y_{ij} = \mu_{Y_j} + \beta\left(X_{ij} - \mu_{X_j}\right) + e_{ij},\ j = 1, 2,$$

where for a jointly normal *pdf* as in (2), $\beta_j = \rho\frac{\sigma_{Y_j}}{\sigma_{X_j}} = \rho$, and e_j are here random residuals with zero means and constant variance. Taking expectation over the residuals conditional on X,

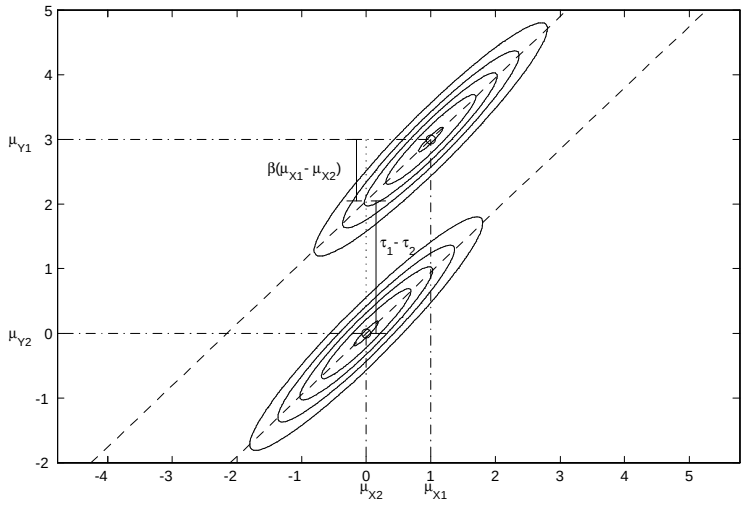

Figure 1: Unadjusted difference $\mu_{X1} - \mu_{X2}$ is biased as an estimator of the treatment effect $\tau_1 - \tau_2$, with the bias equal to $\beta(\mu_{X1} - \mu_{X2})$.

$$\begin{aligned} & E_{mo\,del}(y_1 - y_2 | x_1 = x_2) \\ = \; & \mu_{Y1} + \beta(x - \mu_{X1}) - (\mu_{Y2} + \beta(x - \mu_{X2})) \\ = \; & \mu_{Y1} - \mu_{Y2} - \beta(\mu_{X1} - \mu_{X2}) = \tau_1 - \tau_2, \end{aligned}$$

the magnitude of interest. But, not taking X into account, that is, taking unconditional expectation over the sample gives

$$\begin{aligned} E_{sample}(\bar{y}_1 - \bar{y}_2) &= \mu_{Y1} + \beta(\bar{x}_1 - \mu_{X1}) - (\mu_{Y2} + \beta(\bar{x}_2 - \mu_{X2})) \\ &= \tau_1 - \tau_2 + \beta(\bar{x}_1 - \bar{x}_2), \end{aligned}$$

which differs from $\tau_1 - \tau_2$ by $\beta(\bar{x}_1 - \bar{x}_2)$, and is the bias of the unconditional expectation.

Cochran (1968) considers, amongst others, the normal density function, with $\varphi_1(x) \in N(0, \sigma^2)$ and $\varphi_2(x) \in N(\theta, \sigma^2)$, whose regressions of Y on X are linear and parallel. This is illustrated in Figure 1.

By subclassification (stratification) on X, the bias $\beta(\mu_{X1} - \mu_{X2})$ of $E(\bar{y}_1 - \bar{y}_2)$ with respect to $\tau_1 - \tau_2$ can be reduced as follows. A subclass indexed with l is formed by selecting boundaries x_{l-1} and x_l. With $\bar{y}_{jl}$ denoting sample mean of Y in the l^{th} stratum of the j^{th} population, $E(\bar{y}_{jl}) = \mu_{Y1} + \beta\bar{x}_{jl}$, where

$$\bar{x}_{jl} = \frac{\int_{x_{l-1}}^{x_l} x\varphi_j(x)\,dx}{\int_{x_{l-1}}^{x_l} \varphi_j(x)\,dx}.$$

The bias due to x after adjustment (Figure 2) is

$$\sum w_l \beta(\bar{x}_{2l} - \bar{x}_{1l}),$$

where w_l denotes weight assigned to class l.

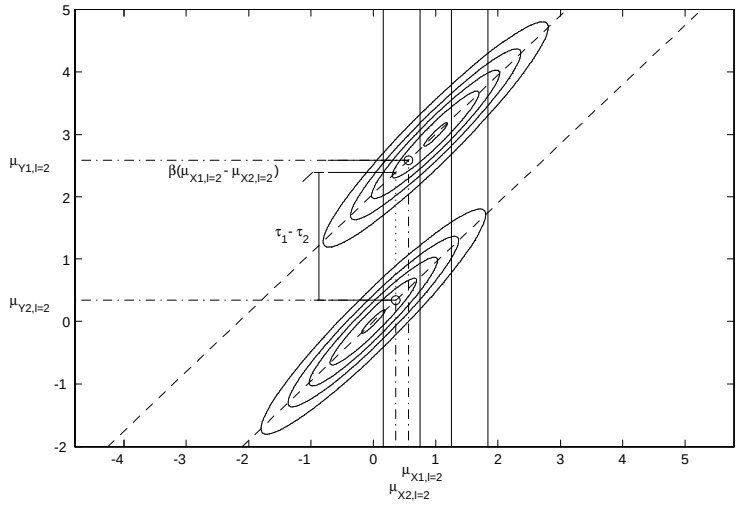

Figure 2: Stratum difference of the means, $\mu_{X1,l} - \mu_{X2,l}$ is less biased as an estimator of $\tau_1 - \tau_2$ than the overall difference of the means. The adjusted estimator is built by averaging over the weighted strata means. Illustrated for $l = 2$ in the figure.

When subclassifying on X in the above mentioned case of two normals, Cochran (ibid.) applies three levels of bias, $\theta/\sigma = 1, \frac{1}{2}, \frac{1}{4}$ and for 5 strata with equal weights assigned to each stratum reports 89.1, 89.6, and 89.7 percent reduction in bias of $\bar{x}_2 - \bar{x}_1$ for the three θ/σ ratios, respectively.

Now, Cochran's (1968) and Cochran and Rubin's (1973) situation is different in two respects from the one considered in the present text, where the differences balance each other: there, responses, Y, are available for both samples, here, responses are available for the r sample only; and on the other hand, there, a treatment effect—conditional expectation of the difference in response between the two groups—is supposed to exist, while here (for the present) no systematic difference in the response caused by the measurement method or otherwise between the samples is assumed to exist.

For a jointly normal *pdf* as in (2)—the case studied by Cochran (1968)— the regression function model, with

$$\begin{aligned} \beta_j &= \rho \frac{\sigma_{Y_j}}{\sigma_{X_j}} = \rho, \\ \alpha_j &= E(Y_j) - \beta_j E(X_j) \end{aligned}$$

is

$$\begin{aligned} Y_j &= \alpha_j + \beta_j X_j \\ &= E(Y_j) - \beta_j E(X_j) + \beta_j X_j \\ &= E(Y_j) + \beta_j (X_j - E(X_j)), \end{aligned} \tag{15}$$

For the double samples procedure and the model in the present study, the regression lines of the two samples, s and r, coincide:

– for the sample s:

$$\begin{aligned}\beta_s &= \rho\frac{\sigma_Y}{\sigma_X} \\ &= \rho \times \frac{1}{1} \\ &= \rho,\end{aligned}$$

$$\begin{aligned}\alpha_s &= E(Y) - \beta_s E(X) \\ &= 0 - \rho \times 0 \\ &= 0,\end{aligned}$$

$$\begin{aligned}Y &= E(Y) - \beta_s (E(X) - X) \\ &= \rho X;\end{aligned}$$

– for the sample r, using the derivations in (7) and (9):

$$\begin{aligned}\beta_r &= \rho_g\frac{\sigma_{Y_g}}{\sigma_{X_g}} \\ &= \frac{Cov_g(X,Y)}{\sigma^2_{X_g}} \\ &= \frac{\rho(1-\pi^{-1})}{1-\pi^{-1}} \\ &= \rho,\end{aligned}$$

$$\begin{aligned}\alpha_r &= E_g(Y) - \beta_r E_g(X) \\ &= \pi^{-\frac{1}{2}}\rho - \rho\pi^{-\frac{1}{2}} \\ &= 0,\end{aligned}$$

$$\begin{aligned}Y_g &= E_g(Y) - \beta_g (E_g(X) - X_g) \\ &= \pi^{-\frac{1}{2}}\rho - \rho\left(\pi^{-\frac{1}{2}} - X|Z\right) \\ &= \rho \times X_g.\end{aligned}$$

With the same intercepts and slopes, the unconditional and conditional regression lines of Y on X are the same. Thus, situation the researcher is facing is the one depicted in Figure 3, but where there actually is no access to the response for the standard group, $j = 1$, only its, possibly estimated, cutoff points on the stratifying variable, here X; so, the situation the researcher really is facing is that in Figure 4.

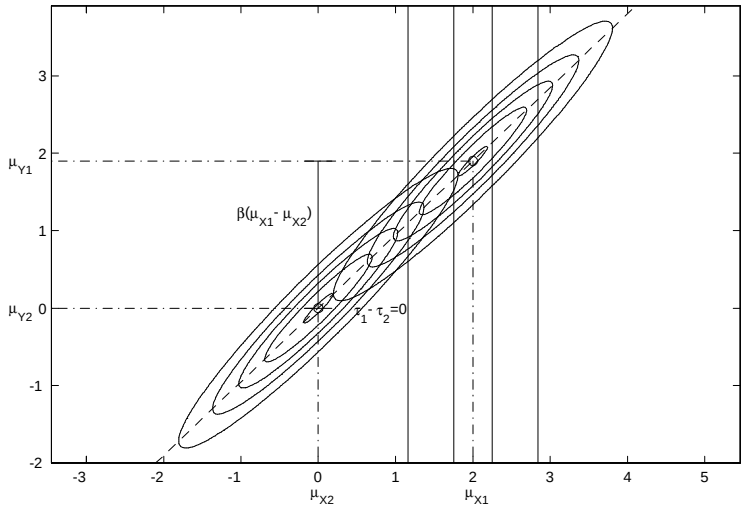

Figure 3: In a single population there is only one regression line. With the assumption of no mode effect, any expected difference between the means indicates bias.

With a limited number of strata, the stratification technique results in a biased estimator of the treatment effect. But, provided that strongly ignorable treatment assignment holds (see above, p. 13), the estimator becomes unbiased when the number of strata approaches infinity. The argument given here for this is intuitive, based on Figure 2; a formal proof is given in Rosenbaum and Rubin (1983).

The bias arises because the lines connecting (μ_{X1}, μ_{Y1}) and (μ_{X2}, μ_{Y2}) in Figure 2 are not vertical. They approach verticality with the increase in the number of strata: with infinitely many strata, the strata mean differences coincide with the treatment effects at these levels of X, and an unbiased overall estimate of the treatment effect is obtained by averaging/integrating over strata means. With limited amount of data, many strata would have none observations or observations for just one of the two samples, why reducing the number of strata is a necessity. But, as shown, with even 5 to 7 strata, an elimination of about 90% of the bias may already be achieved.

While stratification—with a limited number of strata— is inferior to regression adjustment in correcting bias in case of a linear relation between X and Y, there is little reason to expect that a linear relation holds in real applications with multivariate auxiliary information or on a propensity score estimate based on this information. This latter situation is illustrated in Figure 5, where a hypothetical population, indexed by 1, has a peculiar distribution. The part completely observed, corresponding to our sample r, is indexed by 2. The quantile cutoff points of the distribution of X in the population are, as before, represented by solid vertical lines. Plugging the estimated overall expected value of X into a linear model based on the data in r would result in an estimate $\hat{\mu}_{Y1,L}$, which is considerably biased. But, using the means of Y in each of the propensity score defined strata (diamonds in the figure), and giving them weights proportional to the quantile ranges of the strata, would result in an estimate with smaller bias.

Thus, stratification on the propensity score provides a way of adjusting the estimates when regression of Y on the propensity score $e(x)$ is not linear in the standard population.

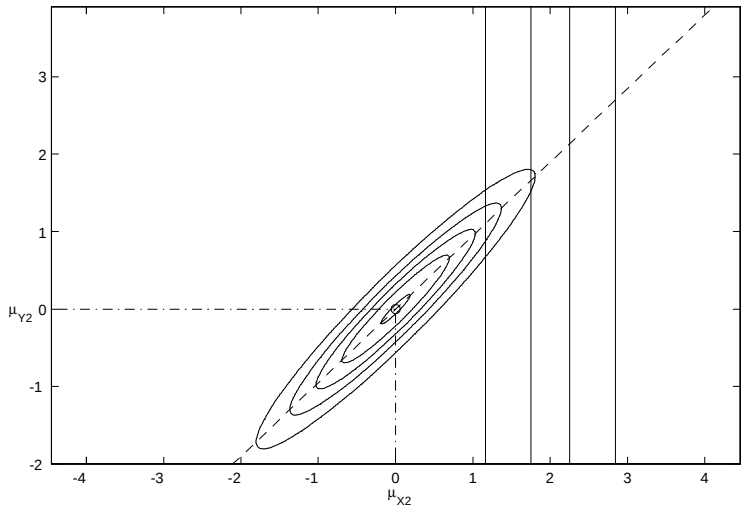

Figure 4: Situation treated in the present study: single population, auxiliary information known for both samples, study variable known only for t sample.

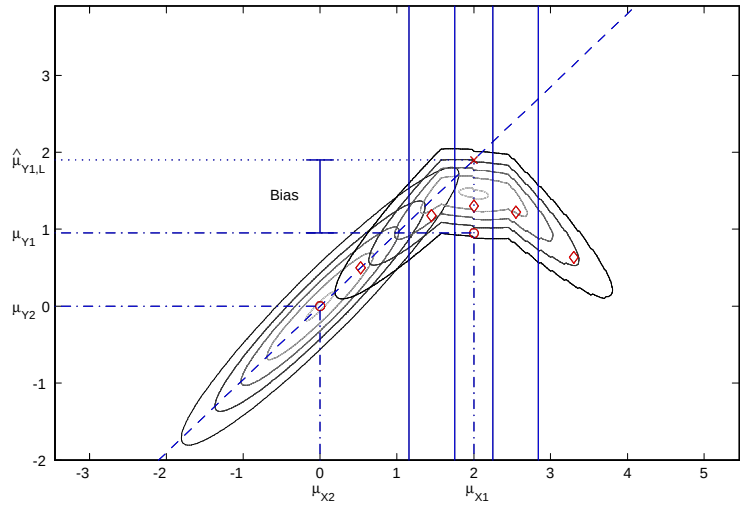

Figure 5: Adjusting the estimate of $E(Y)$ for a hypothetical population with the location parameter (μ_{X1}, μ_{Y1}) based on a sample from the subspace with the location parameter (μ_{X2}, μ_{Y2}): linear regression produces a biased estimate $\hat{\mu}_{Y1,L}$; the propensity score adjusted estimate, which is the mean of the five strata estimates (diamonds in the figure–the mean itself is not depicted in the figure) can be seen to be far less biased.

5 Conclusions

Like in estimating treatment effects when sampling from *two populations*, the propensity score weighting technique efficiently reduces bias when taking two samples from the *same population*—one of the samples drawn from a subset of the population. In fact, in theory the approach can give unbiased estimates provided that the strongly ignorable treatment assignment assumptions hold. In practice, though, a limited amount of available data leads to a limited number of propensity score strata, because observations from both samples need to exist at each stratum level. As a consequence, the method can only reduce bias rather than eliminate it; but, with even 5 to 7 strata, an elimination of about 90% of the bias may be achieved.

Concerning the present study, the following two remarks on its limitations seem to be in place:

Model simplicity. The model introduced here is extraordinarily simple, the sole purpose of its use was to enable simple and clear demonstration of the propensity score weighting. Of course, in this univariate auxiliary variable case, balancing on X and balancing on the propensity score result in the same cutoff points, that is, in the same final estimates. With a multivariate auxiliary information, where the propensity score is a function of more than a single variable, this is not longer the case (but is more difficult to graph). The general principle, though, is the same in both the univariate and the multivariate case.

No mode effect. In this demonstration, no account was taken of an eventual existence of the mode effect, that is, of the effect of the data collection method on values observed. In the survey context, it is known that a "warm", socially intensive, method like personal or telephone interview tends to have a systematic influence in direction of social desirability and conformance, compared to "cold", less socially intensive methods that rely on self-administration, like web surveys. If a mode effect is suspected to exist with respect to some of the variables in the study, it needs to be estimated outside of the propensity score weighting technique and added atop of the weighting results to produce the final estimate.

In practice, the propensity score is estimated rather than known, in situations where the underlying distributions of the variables included in the study are at best assumed. (Usually, the logistic regression model is used.) Uncertainty regarding the resulting, propensity score weighted, point estimate draws from two sources: propensity score was estimated rather than known, and only samples from the population and its subset were measured instead of the whole population. These, and the other topics like influence of the covariance structure of the variables in a multivariate auxiliary information situation on the estimate, choice of the estimation technique, effects of violation of strongly ignorable treatment assignment assumptions, etc., need to be illuminated, preferably through a simulation study.

Acknowledgment

The author is indebted to Professor Ove Frank, whose comments and suggestions have considerably improved the formal expression in the paper, and to Professor Daniel Thorburn for reading and commenting several earlier versions of the manuscript. 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] Johnson, N.L. and Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley: New York, NY.

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

[5] 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.

[6] 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.

[7] 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.