数理解析研究所講究録
第1703巻 2010年 138-150

Rubin's Model for Causal Inference: a review

千葉大学・大学院理学研究科　汪　金芳 (Jinfang Wang)
Graduate School of Science
Chiba University

1 Introduction

Rubin's model for causal inference, or simply Rubin causal model (RCM), sometimes referred to as the Neyman-Rubin causal model or Neyman-Rubin-Holland model for causal inference, is developed in a series of papers by Rubin ([31, 32, 33]), though RCM may be traced back to the work of Neyman ([24]), while Holland ([18]) and Holland and Rubin ([19]) provide penetrating reviews of this model. A more complete picture of RCM may be found in a collection of papers by Rubin ([36]). This model has found applications in diversity of areas including statistics, medicine, economics, political science, sociology and law, among others; see [39] for references on some of the recent applications. Recently some rigorous results on RCM have been established in the econometric literature, see, e.g., [13], [17], [1]. In this paper we give a review of RCM with emphases on the more resent developments.

2 Rubin Causal Model

Borrowing from the language of design of experiments, suppose that we have a population from which we draw a random sample of n units. Each unit is able to be exposed to either a *treatment* or a *control*. Let Z_i represent a random variable of *treatment assignment* so that $Z_i = 1$ if the ith unit is assigned to the treatment group and $Z_i = 0$ if the ith unit is assigned to the control group. Thus, the ith unit has two potential outcomes, $Y_i(1)$ if it is exposed to the treatment when $Z_i = 1$, or $Y_i(0)$ if it is exposed to the control when $Z_i = 0$. The observed data on the ith unit consist of the pair (Z_i, Y_i), where

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) .$$

The effect *caused* by the teatment for the ith unit (relative to the control), or simply the *treatment effect* for the ith unit, is defined as the difference $Y_i(1) - Y_i(0)$. This quantity measures the gain in the outcome variable under the assignment to the treatment relative to the control. We suppose that each unit can be exposed to only the treatment or the control, therefore we can observe either $Y_i(1)$ or $Y_i(0)$, but never both. That is, either $Y_i(1)$ or $Y_i(0)$ is missing for the ith unit, implying that the treatment effect for the ith unit is not *observable*. This fact is called by Holland ([18]) the *fundamental problem of causal inference*.

To statistically overcome the fundamental problem of causal inference, the first thing we do is to replace the inferential goal of estimating the treatment effect for an individual unit by considering the problem of estimating the *average treatment effect*:

$$\theta = \mathbb{E}\{Y_i(1)\} - \mathbb{E}\{Y_i(0)\} \tag{1}$$

where the expectation is assumed to be independent of i. We note that since the operational meanings of the two random variables $Y_i(0)$ and $Y_i(1)$ involve the random variable Z_i, the expectations $\mathbb{E}\{Y_i(1)\}$ and $\mathbb{E}\{Y_i(0)\}$ therefore almost always depend on the distribution of Z_i, that is, the mechanism of the treatment assignment. More explicitly, we can write

$$\mathbb{E}\{Y_i(1)\} = \mathbb{E}[\mathbb{E}\{Y_i(1)\}|Z_i]$$

and similarly for $\mathbb{E}\{Y_i(0)\}$. The average treatment effect θ has the potential to be estimated because potential outcomes $Y_i(1)$ and $Y_i(0)$ on different units may now be used to estimate the expectations $\mathbb{E}\{Y_i(1)\}$ and $\mathbb{E}\{Y_i(0)\}$. To achieve this goal, further fundamental assumptions on the treatment assignment mechanism are however required since the observed data (Z_i, Y_i) only provide information on the expectations

$$\mathbb{E}\{Y_i|Z_i = 1\} = \mathbb{E}\{Y_i(1)|Z_i = 1\} \quad \text{and}$$
$$\mathbb{E}\{Y_i|Z_i = 0\} = \mathbb{E}\{Y_i(0)|Z_i = 0\}.$$

The fundamental problem of causal inference can be overcome by considering two such assumptions, namely the *independence assumption* ([18]) and the assumption of *strong ignorability* ([29]). Both conditions are natural in the sense that they can be derived when one considers the relations between the expectations $\mathbb{E}\{Y_i(1)\}, \mathbb{E}\{Y_i(0)\}$ and the conditional expectations $\mathbb{E}\{Y_i(1)|Z_i = 1\}, \mathbb{E}\{Y_i(0)|Z_i = 0\}$. The independence assumption concerns the classical case of *randomized experiment*, where we assume that the treatment assignment Z_i is independent of the potential outcomes $(Y_i(1), Y_i(0))$ and all other potential confounding variables. Causal inference for randomized experiment is straightforward because under this independence assumption we have the basic identities

$$\mathbb{E}\{Y_i(1)\} = \mathbb{E}\{Y_i(1)|Z = 1\}$$
$$\mathbb{E}\{Y_i(0)\} = \mathbb{E}\{Y_i(0)|Z = 0\}.$$

Thus the independence assumption ensures that

$$\theta = \mathbb{E}\{Y_i(1)|Z = 1\} - \mathbb{E}\{Y_i(0)|Z = 0\} \tag{2}$$

So the sample difference in the two groups in this case will give an unbiased estimate for θ.

The large body of the literature on causal inference however concerns the second case when the experiment is not randomized, that is, the independence assumption does not hold true. These cases are known as nonrandomized experiments or *observational studies* ([28]), and will be the topic in the rest of this paper.

In order to estimate the average treatment effect in observational studies, we assume, as is usually the case, that in addition to (Z_i, Y_i), we also observe for each unit i the value on a pretreatment variable X_i, a vector of length p. The value of the pretreatment variable X_i usually measures the characteristics of the ith unit (e.g., gender, paraent's educational level, etc.) before the treatment assignment, and thus is not affected by the treatment. We now relax the independence assumption in a randomized experiment by the following assumption of *strong ignorability* ([29]):

ASSUMPTION 2.1 (strong ignorability). *The following hold for each i.*

(i) (Unconfoundedness)

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i \mid X_i$$

(ii) (Overlap)

$$0 < \Pr(Z_i = 1 | X_i = x) < 1$$

where $A \perp\!\!\!\perp B|C$ is the Dawid's ([8]) notation denoting the conditional inpdenepdence of A and B given C. The conditional probability of assignment to treatment given the pretreatment variable is known as the *propensity score* ([29]):

$$e(x) = \Pr(Z_i = 1 | X_i = x) = \mathbb{E}\{Z_i | X_i = x\} \tag{3}$$

To see why the strongly ignorable treatment assignment should lead to an estimation procedure for the average treatment effect, we note the following basic identity under the unconfoundedness assumption:

$$\begin{aligned} \mathbb{E}\{Y_i(z) | X_i = x\} &= \mathbb{E}\{Y_i(z) | Z_i = z, X_i = x\} \\ &= \mathbb{E}\{Y_i | Z_i = z, X_i = x\} \end{aligned} \tag{4}$$

where z takes values 0 or 1. Thus, estimation of the average treatment effect θ can be done by first estimating the average treatment effect for a subpopulation at $X = x$:

$$\theta(x) = \mathbb{E}\{Y_i | Z_i = 1, X_i = x\} - \mathbb{E}\{Y_i | Z_i = 0, X_i = x\},$$

by using the averaged sample treatment-control difference within the subpopulation at $X = x$. We then average this difference over all possible values of x to give an unbiased estimator for θ because we have

$$\theta = \mathbb{E}[\mathbb{E}\{Y_i(1) | X_i\} - \mathbb{E}\{Y_i(0) | X_i\}] = \mathbb{E}\{\theta(X_i)\}. \tag{5}$$

Thus, in observational studies the fundamental problem of causal inference is now overcome by the additional knowledge on pretreatment varibles and the unconfoundedness assumption. Note that the overlap assumption is crucial in estimating $\theta(x)$, for violation of this assumption at x will mean that there are only treated or control units at x thus making the estimation of either $\mathbb{E}\{Y_i(1) | X_i = x\}$ or $\mathbb{E}\{Y_i(0) | X_i = x\}$ an impossibility. It is also worthy of noting that the basic equation (4) itself may be used as a weaker assumption instead of the unconfoundedness in order to estimate the average treatment effect ([15]). The assumption (4) is however almost as difficult to verify as with the unconfoundedness condition in practice.

To conclude this section we note that although we shall focus on estimation of the average treatment effect θ, there is also considerable interest in the literature on estimation of the *treatment effect for the treated* (e.g., [14], [15], [16]):

$$\theta_T = \mathbb{E}\{Y_i(1) - Y_i(0) | Z_i = 1\}$$

We need the strong ignorability assumption here as well for estimating θ_T because $Y_i(0)$ are unobserved for the treated units.

3 Estimating the Average Treatment Effect

3.1 Regression Estimators

Regression adjustment for estimating the average treatment effect in observational studies has a long history (e.g., [25], [3], [6], [7], [32], etc.). The idea is to use regression techniques to find estimates $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ of the two regression functions in (4), namely, $\mu_1(x) = \mathbb{E}\{Y_i(1)|X_i = x\}$ and $\mu_0(x) = \mathbb{E}\{Y_i(0)|X_i = x\}$. By (5), we then average the difference $\hat{\mu}_1(x) - \hat{\mu}_0(x)$ over the empirical distribution of x to get an unbiased estimate of θ:

$$\hat{\theta} = \frac{1}{n}\sum_i \{\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\} \tag{6}$$

where $\hat{\mu}_1(x_i)$ and $\hat{\mu}_0(x_i)$ are estimated, due to unconfoundedness, using the treatment group samples and the control group samples respectively.

For instance, suppose that we may assume, as in [32], the regression functions are linear in x

$$\begin{aligned} \mu_1(x) &= \alpha_1 + \beta_1 x \\ \mu_0(x) &= \alpha_0 + \beta_0 x \end{aligned}$$

where x is a univariate continuous covariate. Let $\hat{\beta}_1, \hat{\beta}_0$ denote the respective within sample least squares estimators and $\bar{y}_1, \bar{x}_1, \bar{y}_0, \bar{x}_0$ the respective within sample means. The predicted values of the regression functions are then given by

$$\begin{aligned} \hat{\mu}_1(x) &= \bar{y}_1 + \hat{\beta}_1(x - \bar{x}_1) \\ \hat{\mu}_0(x) &= \bar{y}_0 + \hat{\beta}_0(x - \bar{x}_0) \end{aligned}$$

Note that the predictors $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ rely on extrapolation and thus the estimator (6) under this linear model can have poor properties when the distributions of the covariate differ significantly in the treated and control groups.

Recently attempts have been made which focus on nonparametric estimation of the regression functions $\mu_1(x)$ and $\mu_0(x)$. One such method is to use the method of sieves ([11]), see, e.g. [20] and [4]. The resulting estimator is efficient in the sense defined in the next section. Another type of estimator for estimating $\mu_1(x)$ and $\mu_0(x)$ is to use kernel methods (see, e.g., [14, 15], [16]).

3.2 Matching Estimators

Estimators constructed by matching the covariates X are among the most popular estimators due to their algorithmic simplicity. These estimators closely resemble the nonparametric kernel regression estimators, where the number of matched samples plays the role of the bandwidth in kernel regression. Large sample properties of simple matching estimators are however established only recently ([1]). When the covariate X is used in matching, the Mahalanobis distance between two

multivariate observations are usually employed (e.g., [7], [34], [35]). The unidimensional propensity score can also be used in matching. We will discuss propensity score matching in the next subsection.

Matching estimators are usually used to estimate the average treatment effect for the treated θ_T, this is because in many observational studies there are more controls than the treated so that it is easier to impute missing values $Y_i(0)$ for units with $Z_i = 1$. To find an imputed value for $Y_i(0)$ we compute the distance $d(X_j, X_i)$ for all X_j in the control group, then retain the k units with the closest distance with X_i. Then we use the average value

$$\hat{Y}_i(0) = \frac{1}{k}\sum_j Y_j(0)$$

as a predicted value for $Y_i(0)$. Here k is an arbitrary integer, usually small such as two or one. The estimator takes the form

$$\hat{\theta}_T = \frac{\sum_i \{Z_i Y_i(1) - Z_i \hat{Y}_i(0)\}}{\sum_i Z_i} \tag{7}$$

If both groups are of relatively large size then we can impute either $Y_i(1)$ or $Y_i(0)$ for all $i = 1, \ldots, n$. Let $\hat{Y}_i(1)$ or $\hat{Y}_i(0)$ be the imputed values, then the resulting matching estimator for the average treatment effect θ is simply taken as the averaged difference

$$\hat{\theta} = \frac{1}{n}\sum_i (\hat{Y}_i(1) - \hat{Y}_i(0)) \tag{8}$$

In [1] it is shown that the estimator (8) has a bias of order $O(n^{-1/k})$, where k is the number of the continuous components of X. So if $k \geq 2$, when enlarged by a factor $\sqrt{n}$, the bias of this estimator will not vanish as $n \to \infty$, although this bias may not be so large in practice as to concern the practitioner.

In the above discussion one usually use the Mahalanobis metric to measure the distance between X_i and X_j,

$$d(X_i, X_j) = \sqrt{(X_i - X_j)' V^{-1} (X_i - X_j)} \tag{9}$$

where V is the estimated covariance matrix of X. In [34], V is taken to be the pooled within sample covariance matrix

$$V = \frac{(\mathbf{X}_1'\mathbf{X}_1 - n_1 \overline{\mathbf{X}}_1' \overline{\mathbf{X}}_1) + (\mathbf{X}_2'\mathbf{X}_2 - n_2 \overline{\mathbf{X}}_2' \overline{\mathbf{X}}_2)}{n - 2}$$

where $\mathbf{X}_i$ is the $n_i \times p$ data matrix for the ith group.

To achieve even better balance in the covariate between the treated and control group, in [10] and [38] the Mahalanobis distance (9) is generalized to

$$d_G(X_i, X_j) = \sqrt{(X_i - X_j)' (V^{-1/2})' W V^{-1/2} (X_i - X_j)} \tag{10}$$

where $V^{-1/2}$ is the Cholesky decomposition of V and W is a $p \times p$ positive definite weight matrix to be estimated. In (10) the covariate may be enlarged to include the propensity score $e(X_i)$ if one

has a reliable model for $e(X_i)$. This method is called genetic matching because a genetic algorithm ([22], [40]) is used to estimate the components of the weight matrix W.

By giving specific weights to W in (10), the genetic matching estimator using metric (10) reduces to the Mahalanobis matching estimator or the propensity score matching estimator, The genetic matching method may be of merits relative to the Mahalanobis matching especially when the covariate has a large dimension and is nonellipsoidally distributed ([36, p. 462]). For some applications of this method see [26], [23] and [12].

When matching is applied to the covariate X, the metric used plays an important role. See also [41] for an alternative metric which takes into account the consideration of the correlation of X_i, Z_i and $(Y_i(1), Y_i(0))$.

3.3 Propensity Score Methods

Significant progress has been made on estimating the avergae treatment effect under RCM by the discovery of a property for the propensity score ([29]). This property says that if treatment assignment Z_i is unconfounded given the pretreatment variable X_i, then Z_i is also unconfounded given the one-dimensional propensity score $e(X_i)$. That is, under unconfoundedness, it holds that

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp Z_i \mid e(X_i) \tag{11}$$

This property may be proved by showing that

$$\Pr\{Z_1 = 1 | Y_i(1), Y_i(0), e(Z_i)\} = \Pr\{Z_1 = 1 | e(Z_i)\}$$

which is equal to $e(X_i)$. To show this, we express the probabilities as expectations and by conditioning on the covariate X_i. Thus due to (11), the fundamental problem of causal inference can now be overcome by conditioning on the propensity score because of ([29])

$$\theta = \mathbb{E}[\mathbb{E}\{Y_i(1) | Z_1 = 1, e(X_i)\} - \mathbb{E}\{Y_i(0) | Z_1 = 0, e(X_i)\}] \tag{12}$$

This is an important result because bias due to the imbalance of the covariate can now be corrected by conditioning on the univariate propensity score, not the covariate vector X_i. Now we discuss several methods for estimating θ which use the propensity score.

3.3.1 Matching

In Section 3.2 we discussed how to construct an estimator of θ by matching the covariate X. Due to (12) we can alternatively match on the propensity score $e(X)$ instead of the full covariate X. When the propensity score $e(X)$ is unknown we have to first estimate it, usually using a logistic regression model:

$$e(X_i) = \frac{e^{\beta' X_i}}{1 + e^{\beta' X_i}}$$

To avoid side effect near zero and one, it is preferable to match on the linear predictor $\hat{\beta}' X_i$ instead of the propensity score directly ([39]). When the propensity score is known, the asymptotic result

of Abadie and Imbens ([1]) then shows that matching estimator using the scalar propensity score produces a $\sqrt{n}$ consistent estimator.

3.3.2 Blocking

Blocking, or subclassification ([29]) is a method which divides the unidimensional propensity score into B blocks, usually equally lengthed. Within each block we treat the data as if they come from a randomized experiment, and theorefore use the averaged treatment-control difference $\hat{\theta}_b$ to estimate the avergae treatment effect for the bth block. The blocking estimator for the average treatment effect is taken as the weighted mean

$$\hat{\theta} = \sum_{b=1}^{B} \frac{n_{1b} + n_{0b}}{n} \hat{\theta}_b \tag{13}$$

where n_{1b} and n_{0b} are the respective numbers of treated and controls in the bth blok. Estimator of variance for $\hat{\theta}$ of (13) is discussed in [21].

For a one-dimensional covariate, with equal-sized block and assuming normality, it is shown ([5]) that $B = 5$ is adequate for removing more than 95% of the bias associated with the simple treatment-control difference. This is the reason that $B = 5$ is usually employed in defining the block estimator ([30], [9], [2]).

3.3.3 Regression

In Section 3.1 we discussed the idea of estimating θ by using regression techniques to estimate the two conditional means $\mu_1(x) = \mathbb{E}\{Y_i(1)|X_i = x\}$ and $\mu_0(x) = \mathbb{E}\{Y_i(0)|X_i = x\}$. Due to (12), under unconfoundedness, we can alternatively estimate the regression functions

$$\eta_1(p) = \mathbb{E}\{Y_i(1)|e(X_i) = p\} \quad \text{and}$$
$$\eta_0(p) = \mathbb{E}\{Y_i(0)|e(X_i) = p\}$$

Using estimates $\hat{\eta}_1(x)$ and $\hat{\eta}_0(x)$, we can estimate θ by

$$\hat{\theta} = \frac{1}{n} \sum_i \{\hat{\eta}_1(e(x_i)) - \hat{\eta}_0(e(x_i))\} \tag{14}$$

Note that to use this estimator we have to specify a model for the propensity score $e(X_i)$ in order to estimate $\hat{\eta}_1(x)$ and $\hat{\eta}_0(x)$. It is of interest to investigate conditions under which the estimator using $\hat{\eta}_z(x)$ may perform better than that using $\hat{\mu}_z(x)$. For estimator (14) to have a chance of success one needs a reasonably good model for the regression functions $\eta_z(p)$ ([21]).

3.3.4 Weighting

A weighting estimator for the average treatment effect takes the form

$$\frac{1}{n} \sum_i \left(\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1 - Z_i) Y_i}{1 - \hat{e}(X_i)} \right)$$

where $\hat{e}(X_i)$ is a nonparametric sieve estimator of the propensity score. This estimator is semiparametrically efficient. We will discuss this estimator in more detail in Section 4.2.

4 Semiparametric Efficiency Bounds and Efficient Estimation

4.1 Efficiency Bounds

In estimating the average treatment effect, it is the bias of the estimators rather than the variance of these estimators that should be of primary concern to the researcher ([36]). However, when an estimator is known to be unbiased or asymptotically unbiased, it is then of interest to consider the variance of such estimators. For instance, for a randomized experiment, it is known that the unbiased simple averaged treatment-control difference is not an efficient estimator for the average treatment effect ([17]). To construct an efficient estimator in this case with known constant propensity score, one can inversely weight the observations using the nonparametrically estimated propensity scores. We will discuss this estimator in detail in next subsection.

Under unconfoundedness and other regularity conditions, Hahn ([13]) established the efficiency bound of a regular estimator $\hat{\theta}$ for the average treatment effect θ. He showed that $\hat{\theta}$ is asymptotically normally distributed

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$$

with variance bounded by

$$V \geq \mathbb{E}\left\{ \frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\theta(X_i) - \theta)^2 \right\} \tag{15}$$

In (15), $\theta(X_i)$ is the average treatment effect for the subpopulation at X_i, and $\sigma_1^2(X_i) = \text{var}(Y_i(1)|X_i)$, $\sigma_0^2(X_i) = \text{var}(Y_i(0)|X_i)$ are the conditional variances. The r.h.s. of (15) gives the semiparametric efficiency bound for a regular estimator for the average treatment effect θ. This efficiency bound plays an analogous role as the Cramér-Rao lower bound for parametric estimation. Hahn showed that the efficiency bound in (15) remains unchanged even though the propensity score is known in advance. In the special case when the propensity score equals an unknown constant, $e(X_i) = p$, that is, the treatment assignment is randomized, we have $\theta = \theta_T$ and the efficiency bound for the commom parameter becomes

$$\mathbb{E}\left\{ \frac{\sigma_1^2(X_i)}{p} + \frac{\sigma_0^2(X_i)}{1 - p} + (\theta(X_i) - \theta)^2 \right\}.$$

When the propensity score is not known, a similar bound exists for the average treatment effect on the treated θ_T:

$$\mathbb{E}\left\{ \frac{e(X_i)\sigma_1^2(X_i)}{p^2} + \frac{e(X_i)^2\sigma_0^2(X_i)}{p^2(1 - e(X_i))} + \frac{(\theta(X_i) - \theta_T)^2 e(X_i)}{p^2} \right\}$$

where $p = \mathbb{E}\{e(X_i)\}$. When the propensity score is known, the corresponding efficiency bound decreases by an amount

$$\mathbb{E}\left\{\frac{(\theta(X_i) - \theta_T)^2 e(X_i)(1 - e(X_i))}{p^2}\right\}.$$

which may be considered as the gain in efficiency by the knowledge of the propensity score.

4.2 Efficient Estimators

Hahan also proposed estimators for both the average treatment effect θ and the average treatment effect on the treated θ_T, which achieve the respective efficiency bound described above. To motivate these estimators, observe that, under unconfoundedness, we have

$$\mathbb{E}\{Z_i Y_i | X_i\} = \mathbb{E}\{Z_i Y_i(1) | X_i\} = \mathbb{E}\{Z_i | X_i\}\mathbb{E}\{Y_i(1) | X_i\}$$

implying

$$\mathbb{E}\{Y_i(1) | X_i\} = \frac{\mathbb{E}\{Z_i Y_i | X_i\}}{\mathbb{E}\{Z_i | X_i\}} = \frac{\mathbb{E}\{Z_i Y_i | X_i\}}{e(X_i)} \tag{16}$$

Similarly we also have

$$\mathbb{E}\{Y_i(0) | X_i\} = \frac{\mathbb{E}\{(1 - Z_i) Y_i | X_i\}}{1 - e(X_i)} \tag{17}$$

These two expressions relate the conditional expectations $\mu_1(X_i) = \mathbb{E}\{Y_i(1)|X_i\}$ and $\mu_0(X_i) = \mathbb{E}\{Y_i(0)|X_i\}$ to the conditional expectations $\mathbb{E}\{Z_i Y_i | X_i\}, \mathbb{E}\{(1-Z_i)Y_i|X_i\}$ and $e(X_i) = \mathbb{E}\{Z_i|X_i\}$. The idea is to use nonparametric regression techniques to estimate the quantities $\mathbb{E}\{Z_i Y_i|X_i\}$, $\mathbb{E}\{(1-Z_i)Y_i|X_i\}$ and $e(X_i)$ to give estimates $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ for $\mathbb{E}\{Y_i(1)|X_i\}$ and $\mathbb{E}\{Y_i(0)|X_i\}$ respectively. These estimates $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ may be used as imputed values for $Y_i(1)$ and $Y_i(0)$ when they are missing. With the imputed values we now have a 'complete' data situation:

$$\begin{aligned} \hat{Y}_i(1) &= Z_i Y_i(1) + (1 - Z_i)\hat{\mu}_1(X_i) \quad \text{under 'treatment'} \\ \hat{Y}_i(0) &= (1 - Z_i) Y_i(0) + Z_i \hat{\mu}_0(X_i) \quad \text{under 'control'} \end{aligned}$$

Hahn proved that the efficient estimator for θ and θ_T are given respectively by

$$\hat{\theta} = \frac{1}{n}\sum_i (\hat{Y}_i(1) - \hat{Y}_i(0)) \tag{18}$$

and

$$\hat{\theta}_T = \frac{\sum_i Z_i(\hat{Y}_i(1) - \hat{Y}_i(0))}{\sum_i Z_i} \tag{19}$$

Alternatively, we note that

$$\begin{aligned} \theta &= \mathbb{E}\{\theta(X_i)\} \\ &= \mathbb{E}\{\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]\} \\ &= \mathbb{E}\{\mu_1(X_i) - \mu_0(X_i)\} \end{aligned}$$

This motivates the following estimator

$$\tilde{\theta} = \frac{1}{n}\sum_i (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)) \tag{20}$$

which is again shown by Hahn to be efficient for estimating θ. Similarly, the efficient estimator for θ_T is

$$\tilde{\theta}_T = \frac{\sum_i Z_i(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))}{\sum_i Z_i} \tag{21}$$

So far we have left unspecified the estimates for $\mathbb{E}\{Z_iY_i|X_i\}$, $\mathbb{E}\{(1-Z_i)Y_i|X_i\}$ and $e(X_i) = \mathbb{E}\{Z_i|X_i\}$, which are used to form the estimates $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$. When X_i has finite support, we can use the following estimates

$$\begin{aligned}
\hat{\mathbb{E}}\{Z_iY_i|X_i = x\} &= \frac{\sum_j Z_jY_j \cdot \mathbf{1}(X_j = x)}{\sum_j \mathbf{1}(X_j = x)}, \\
\hat{\mathbb{E}}\{(1-Z_i)Y_i|X_i = x\} &= \frac{\sum_j (1-Z_j)Y_j \cdot \mathbf{1}(X_j = x)}{\sum_j \mathbf{1}(X_j = x)}, \\
\hat{\mathbb{E}}\{Z_i|X_i = x\} &= \frac{\sum_j Z_j \cdot \mathbf{1}(X_j = x)}{\sum_j \mathbf{1}(X_j = x)},
\end{aligned}$$

where $\mathbf{1}(X_j = x)$ is the indicator function.

When X_i has a continuous distribution, Hahn suggests to use the series estimators for these conditiona expectations. One difficulty with the series estimators is that one has to choose a somewhat arbitrary number of terms in the series. Hirano *et al.* ([17]) considered another type of efficient estimator for θ so that the series estimators are required only for estimating the propensity score. The merits of using estimated propensity score in gaining efficiency even when the propensity score is known has been poined by a number of researches (e.g., [27], [37], [13], [15]). To motivate their estimator, we notice that, by (16) and (17), the average treatment effect θ can also be expressed as

$$\begin{aligned}
\theta &= \mathbb{E}\{\mathbb{E}[Y_i(1)|X_i] - \mathbb{E}[Y_i(0)|X_i]\} \\
&= \mathbb{E}\left\{\frac{\mathbb{E}[Z_iY_i|X_i]}{e(X_i)} - \frac{\mathbb{E}[(1-Z_i)Y_i|X_i]}{1-e(X_i)}\right\} \\
&= \mathbb{E}\left\{\mathbb{E}\left[\frac{Z_iY_i}{e(X_i)}\middle|X_i\right] - \mathbb{E}\left[\frac{(1-Z_i)Y_i}{1-e(X_i)}\middle|X_i\right]\right\} \\
&= \mathbb{E}\left\{\frac{Z_iY_i}{e(X_i)} - \frac{(1-Z_i)Y_i}{1-e(X_i)}\right\}
\end{aligned}$$

The sample version of the last expectation, with the propensity score estimated, gives an estimator for θ:

$$\hat{\theta} = \frac{1}{n}\sum_i \left(\frac{Z_iY_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}\right) \tag{22}$$

where $\hat{e}(X_i)$ in (22) is the nonparametric sieve estimator for the propensity score. Hirano, *et al.* ([17]) showed that $\hat{\hat{\theta}}$ attains the semiparametric efficiency bound (15), thus is an efficient estimator for θ. The advantage of $\hat{\hat{\theta}}$ over $\hat{\theta}$ or $\tilde{\theta}$ is that to compute $\hat{\hat{\theta}}$ we only need estimattion for the propensity score.

References

[1] Abadie, A. and Imbens, G.: Large sample properties of matching estimators for average treatment effects, *Econometrica*, Vol. 74, pp. 235–267 (2006).

[2] Becker, S., and Ichino, A.: Estimation of average treatment effects based on propensity scores, *The Stata Journal*, Vol. 2, No. 4, pp. 358–377 (2002).

[3] Belson, W.A.: A technique for studying the effects of a television broadcast, *Applied Statistics*, Vol. 5, pp. 195–202 (1956).

[4] Chen, X., Hong, H. and Tarozzi, A. Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors *Annals of Statistics*, Vol. 36, No. 2, pp. 808–843 (2008).

[5] Cochran, W.G: The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics*, Vol. 24, pp. 295–314 (1968).

[6] Cochran, W.G: The use of covarance in observational studies, *Applied Statistics*, Vol. 18, pp. 270–275 (1969).

[7] Cochran, W.G. and Rubin, D.: Controlling bias in observational studies, *Sankhya A*, Vol. 35, pp. 417–446 (1973).

[8] Dawid, A. P.: Conditional independence in statistical theory, *Journal of the Royal Statistical Society, Series B*, Vol. 41, No. 1, pp. 1–31 (1979).

[9] Dehejia, R., and Wahba, S.: Causal effects in nonexperimental studies: reevaluating the evaluation of training programs, *Journal of the American Statistical Association*, Vol. 94, pp. 1053–1062 (1999).

[10] Diamond, A. and Sekhon, J.: Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies, `http://sekhon.berkeley.edu/papers/GenMatch.pdf` (2005).

[11] Geman, S. and Hwang, C.: Nonparametric maximum likelihood estimation by the method of sieves, *Annals of Statistics*, Vol. 10, 401–414 (1982):

[12] Gordon, S. and Huber, G.: The Effect of Electoral Competitiveness on Incumbent Behavior, *Quarterly Journal of Political Science*, Vol. 2, No. 2, pp. 107–138 (2007).

[13] Hahn, J.: On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica*, Vol 66, No. 2, pp. 315–331 (1998).

[14] Heckman, J., Ichimura, H. and Todd, P.: Matching as an econometric evaluation estimator: evidence from evaluating a job training program, *Review of Economic Studies*, Vol. 64, pp. 605–654 (1997).

[15] Heckman, J., Ichimura, H. and Todd, P.: Matching as an econometric evaluation estimator, *Review of Economic Studies*, Vol. 65, 261–294 (1998)

[16] Heckman, J., Ichimura, H., Smith, J. and Todd, P.: Characterizing selection bias using experimental data, *Econometrica*, Vol. 66, No. 5, 1017–1098 (1998).

[17] Hirano, K., Imbens, G. and Ridder, G.: Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, Vol. 71, No. 4, pp. 1161–1189 (2003). July

[18] Holland, P. W.: Statistics and causal inference, *Journal of the American Statistical Association*, Vol. 81, No. 396, pp. 945–960 (1986).

[19] Holland, P. W. and Rubin, D.: Causal Inference in Retrospective Studies, *Evaluation Review*, Vol. 12, No. 3, pp. 203–231 (1988).

[20] Imbens, G., Newey, W. and Ridder, G.: Mean-squared-error calculations for average treatment effects, Working Paper, Department of Economics, UC Berkeley.

[21] Imbens, G. and Wooldridge, J.: Recent developments in the econometrics of program evaluation, NBER Working Paper No. 14251 (2008).

[22] Mebane, W. and Sekhon, J.: GENetic optimization using derivatives (GENOUD), `http://sekhon.berkeley.edu/rgenoud/` (1998).

[23] Morgan, S. and Harding, D.: Matching estimators of causal effects: prospects and pitfalls in theory and practice, *Sociological Methods & Research*, Vol. 35, No.1, pp. 3–60 (2006).

[24] Neyman, J.: On the application of probability theory to agricultural experiments. essay on principles. Section 9, translated in *Statistical Science* (with discussion), Vol. 5, No. 4, pp. 465–480 [1990](1923)

[25] Peters, C.C.: A method of matching groups for experiment with no loss of population, *Journal of Educational Research*, Vol. 34, pp. 606–612 (1941).

[26] Raessler, S. and Rubin, D.: Complications when using nonrandomized job training data to draw causal inferences, *Proceedings of the International Statistical Institute*, (2005).

[27] Rosenbaum, P.: Model-based direct adjustment, *Journal of the American Statistical Association*, Vol. 82, pp.387–394 (1987).

[28] Rosenbaum, P. R.: *Observational Studies*, New York: Springer-Verlag, 2nd edition (2002).

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

[30] Rosenbaum, P., and Rubin, D.: Assessing the sensitivity to an unobserved binary covariate in an observational study with binary outcome, *Journal of the Royal Statistical Society, Ser. B*, Vol. 45, No. 212–218 (1983b).

[31] Rubin, D.: estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, Vol. 66, pp.688–701 (1974).

[32] Rubin, D.: Assignment to treatment group on the Basis of a covariate, *Journal of Educational Statistics*, Vol. 2, No. 1, pp. 1–26 (1977).

[33] Rubin, D.: Bayesian inference for causal effects: the role of randomization, *Annals of Statistics*, Vol, 6, pp. 34–58 (1978).

[34] Rubin, D.: Using multivariate sampling and regression adjustment to control bias in observational studies, *Journal of the American Statistical Association*, Vol. 74, pp. 318–328 (1979).

[35] Rubin, D.: Bias reduction using Mahalanobis-metric matching, *Biometrics*, Vol. 36, No. 2, pp. 293–298 (1980).

[36] Rubin, D.: *Matched Sampling for Causal Effects*, Cambridge, England: Cambridge University Press (2006).

[37] Rubin, D. and Thomas, N.: Matching using estimated propensity scores: relating theory to practice, *Biometrics*, Vol. 52, pp. 249–264 (1996).

[38] Sekhon, J.: Matching: algorithms and software for multivariate and propensity score matching with balance optimization via genetic search, *Journal of Statistical Software*, (2007).

[39] Sekhon, J.: Causal inference in quantitative and qualitative research, In *The Oxford Handbook of Political Methodology* (eds. Box-Steffensmeier, J., Brady, H. and Collier, D.), Oxford University Press (2008).

[40] Sekhon, J. and Mebane, W.: Genetic optimization using derivatives: theory and application to nonlinear models, *Political Analysis*, Vol. 7, pp. 189–203 (1998).

[41] Zhao, Z.: Using matching to estimate treatment effects: data requirements, matching metrics and an application, *Review of Economics and Statistics*, Vol. 86, No. 1, pp.91–107 (2004).