Chapter 4

Statistical Approaches to the Evaluation Problem: Matching Methods

4.1 Introduction

The problem of selection bias is the central issue in modern microeconometric studies. This bias occurs from two, not necessarily mutually exclusive, reasons; selection on observables and selection on unobservables. The former type of selectivity implies that persons with specific observable characteristics tend to participate in the program. A straight comparison of outcomes between participants and non-participants, then, yields biased estimates unless specific econometric methods are applied. Such an appealing method is the *method of Matching*.

4.2 The idea in Matching Methods

Matching is a widely used method of evaluation in statistics that it was first conducted by Fechner at 1860 in agricultural studies. In the late 1970 and early 1980, it was extensively applied to the evaluation of job training programs.

Rubin (1973a), in the discussion about matching estimators, supposes that analysts have access to a set of conditioning variables T_i that represent attributes of participants and non-participants, decomposed, as shown in Chapter 2, into (X_i, Z_i). In this chapter, let us suppose for reasons of simplicity, and without loss of generality, that $X_i = Z_i$ and thus use only the term X_i on all respective formulas. Matched sampling attempts to compare participants with "similar" non-participants with respect to the conditioning variables

(attributes) measured on all subjects. The contrast of participants' outcomes (Y_{1i}) to "comparable" non – participants' outcomes (Y_{0i}) gives the parameter of interest. Since matching is conducted in terms of the observable characteristics of individuals, matching estimators estimate consistently the causal parameters under selection on observables only. Rosenbaum and Rubin (1985) apply this method to data from a Danish cohort study. In order to evaluate the usefulness of a particular treatment, they compare the outcomes of a number of persons who chose to receive treatment (participants) with similar non-treated (non-participants) in terms of observable characteristics, such as age, sex and socioeconomic status. Intuitively, they assumed that treatment status among persons depended only on these observable characteristics. If another factor affected the treatment status, the analysis would not yield free from selection bias estimates.

Intuitively, this idea is closely related to the idea of randomized experiments to the solution of the evaluation problem. However, treatment is not assigned by randomization but is received by all participants. Matching emulates randomization in the following way: *under some conditions, matched sampling conditions on participants' attributes X_i to generate a "comparison" group of non-participants*. Then, a simple difference of the relative outcomes for each person, conditional on X_i, estimates adequately $\left(\Delta_i \middle| X_i, D_i = 1\right)$. That is, due to the similarity in the characteristics of the matched individuals, we may compare them as if we could compare the outcomes of a single participant *(D_i = 1)* observed at both states (treatment and non-treatment), simultaneously. In this sense X_i can be characterized as a balancing score.

Definition 4.1: Balancing Score

A balancing score, $b(X_i)$, is a function of the observed covariates X_i such that the conditional distribution of X_i given $b(X_i)$ is the same for the participants *(D_i = 1)* and non-participants *(D_i = 0)*.

In this framework $b(X_i) = X_i$ is a balancing score. More interesting balancing scores are the particular functions of X_i, mentioned in paragraph 4.4.

4.3 Conditions of Matching

Dawid (1979) provided the conditions that make the matching estimator a useful tool for an econometric analysis.

1st condition

(C - 1): $(Y_{0i}, Y_{1i}) \perp D_i | X_i$ and $0 \prec P(D_i = 1 | X_i) \prec 1$

More formally, participation variable D_i and response (Y_{1i}, Y_{0i}) are conditionally independent given X_{ii}. Rosenbaum and Rubin (1983a) called condition (C - 1) *strong ignorability*[2] condition for D_i given the vectors of covariates X_i. Condition (C - 1) implies:

$$P(D_i = 1 | Y_{0i}, Y_{1i} X_i) = P(D_i = 1 | X_i)$$

so the probability of selection depends only on observed (by the researcher) characteristics, termed in the literature as observables.

When the strong ignorability condition holds, one can generate the marginal distributions of the counterfactuals:

$$F(Y_{0i} | D_i = 1, X_i) \text{ and } F(Y_{1i} | D_i = 0, X_i)$$

from the matching (comparison group) persons. From the dataset one can also retrieve:

$$F(Y_{0i} | D_i = 0, X_i) \text{ and } F(Y_{1i} | D_i = 1, X_i)$$

However, the joint distribution of $(Y_{0i}, Y_{i1}), F(Y_{0i}, Y_{1i}, D_i | X_i)$ cannot be estimated without further restrictions.

[2] The model is ignorable if the inequalities in the second relationship of condition (C-1) are not strict. If D is strongly ignorable then it is also ignorable, but the converse is not true.

If assumption C-1 is valid we can use non-participants to measure what participants would have earned had they not participated, provided we conditioned on the variables X_i. This is true since the independence of Y_{0i} from the participation state given X_i, implies that:

$$B(X_i) = E\left(Y_{0i} \middle| D_i = 1, X_i\right) - E\left(Y_{0i} \middle| D_i = 0, X_i\right) = 0$$

To ensure that this assumption has empirical content, it is also necessary to bound participation probability in order to make the comparison. More specifically, if for some X_i, $P(X_i) = 0$ *or* $P(X_i) = 1$, one could not use matching conditional on those X_i values to estimate a treatment effect. Persons with such characteristics either never receive treatment (thus are never observed as participants), or always receive treatment and hence they are never observed as non-participants, so matches from both $D_i = 0$ and $D_i = 1$ distributions cannot be performed. This is exactly the idea of the $0 < P(D_i = 1 | X_i) < 1$ assumption. Heckman (1990a) proves the above argument in a mathematical way.

When (C - 1) assumption is invoked, it is possible to construct the "treatment on the treated" parameter, $E\left(\Delta_i \middle| D_i = 1, X_i\right)$, or the effect of "treatment on the non-treated", $E\left(\Delta_i \middle| D_i = 0, X_i\right)$ and estimate them adequately with the method of matching. Nevertheless, the issue of interest is the matching measures that have to be performed to obtain adequate estimates of the above parameters.

4.4 Matching Measures

Let us denote with Y_{1i} the outcomes of participants and with Y_{0i} the outcomes of the comparison group members. Furthermore, assume that I_0 and I_1 denote the set of indices for non-participants and participants, respectively. To estimate the treatment effect for each participant $i \in I_1$, outcome Y_{1i} is compared to Y_{0i} of an appropriate, "similar" person $j \in I_0$ in the non-participation state. Matches can be constructed on the basis of the observed individuals characteristics. The difference in those outcomes is considered as a program evaluation measure.

Although it seems simple in theory, practically it is often difficult to conduct matching on the basis of X_i variables. The practical difficulties vary with the situation. For example, Rosenbaum and Rubin (1985) are referred to a comparison of the economic adjustments of boys who completed high school with boys who dropped out. Starting with reservoirs of 671 and 523 boys relatively, the analyst ended with samples of size 23 after matching on six major variables. This problem is mentioned by Cochran (1965) as the *curse of dimensionality*. That is, even with samples of typical size, when the dimension of X_i is high (that is many X_i*'s* coordinates) matching on X_i is very difficult. Therefore, more easily handled variables have to be defined in the conditioning set.

Rosenbaum and Rubin (1983a) gave an appealing solution to this problem. They demonstrate that matching can be performed on "propensity score", $P\left(D_i = 1 \middle| X_i\right) = p_i$, instead of X_i. When condition (C - 1) is hold, then:

$$\left(Y_{1i}, Y_{0i}\right) \perp D_i \middle| p_i \quad \text{and} \quad 0 \prec p_i \prec 1 \qquad \text{(C′ - 1)}$$

also holds. Conditioning on p_i not only produces conditional independence but also the construction of the desired counterfactual conditional mean $E\left(Y_{0i} \middle| D_i = 1, p_i\right)$ requires only:

$$B\left(p_i\right) = E\left(Y_{0i} \middle| D_i = 1, p_i\right) - E\left(Y_{0i} \middle| D_i = 0, p_i\right) = 0$$

which is implied by the independence property of (C′ - 1).

In this way, exact matching on p_i will tend to balance the X_i distribution in the participants and non-participants groups and make them comparable for further analysis. Therefore p_i can be characterized a balancing score and the following theorem is constituted:

Theorem 4.1

Suppose treatment assignment is strongly ignorable and b(x) is a balancing score. Then the expected difference in observed responses to the two treatments at b(x) is equal to the Average Treatment Effect at b(x), that is:

$$E\{Y_{1i}|b(X_i), D_i = 1\} - E\{Y_{0i}|b(X_i), D_i = 0\} = E\{Y_{1i} - Y_{0i}|b(X_i)\}$$

This theorem is applicable when $b(X_i) = p_i$ (see Appendix 2, Corollary 4.1). From this point of view, the dimensionality problem is reduced to matching in one dimension. Matching on the propensity score seems to be an appealing solution.

Heckman, Itchimura and Todd (1998) examine the above argument from a different perspective and produce controversial results. Since Rosenbaum and Rubin's (1983a) theoretical results are based on strong ignorability conditions and p_i is assumed to be known rather than estimated, several limitations occur in practice where, indeed, p_i have to be estimated. Specifically, when comparing the efficiency of the estimators:

$$E\{Y_{1i}|D_i = 1, X_i\} - E\{Y_{0i}|D_i = 0, X_i\} \quad and \quad E\{Y_{1i}|D_i = 1, p_i\} - E\{Y_{0i}|D_i = 0, p_i\}$$

neither is proved to be more efficient than the other. Three arguments is worth mentioning:

1. *Strong ignorability conditions*: Condition (C - 1) or (C′ - 1) is overly strong for the estimation of the mean effect of treatment on the treated or any other evaluation parameter. A weaker condition of the form:

   $$Y_{0i} \perp D_i|X_i \quad \text{or} \quad Y_{0i} \perp D_i|p_i$$

   which implies that

   $$E(Y_{0i}|D_i = 1, X_i) = E(Y_{0i}|D_i = 0, X_i) \quad and \quad E(Y_{0i}|D_i = 1, p_i) = E(Y_{0i}|D_i = 0, p_i)$$

it suffices to estimate $E\left(Y_{1i} - Y_{0i} \middle| D_i = 1, p_i\right)$, since one can recover the counterfactual mean $E\left(Y_{0i} \middle| D_i = 1, p_i\right)$ from the data of non-participants. Note that condition (C′ - 1) does not rule out the dependence of D_i on Y_{1i} or on Δ_i given p_i.

2. *Comparing estimators when p_i is known*: From the perspective of bias, matching on p_i is better in the sense that it allows √N – consistent estimation of the Treatment on the Treated matching estimator for a wider class of models than in possible if matching is performed directly on X_i. However, from the perspective of variance of these estimators, the asymptotic variance of $\Delta_i^{(p)}$ is not necessarily smaller than that of $\Delta_i^{(X)}$.

3. *Comparing estimators when p_i is estimated*: The propensity score is usually estimated either parametrically or non-parametrically. In the parametric case, under some regularity conditions, the selection bias function for $\hat{p}_i$ is zero. However, when p_i is estimated non-parametrically two problems occur. First, the curse of dimensionality is present. Second, the smaller bias that arose from matching on a known p_i no longer holds if true estimation of p_i is a d-dimensional non-parametric estimation problem where $d > 1$. As far as the asymptotic variance of the estimators concerns, it increases under any method of estimation of p_i.

At this point it is worth referring the paper of Heckman, Itchimura and Todd (1997) who, in an application to JTPA data, estimated the propensity score in a semi-parametric framework. They decomposed the conventional measure of evaluation bias into components. Three components are described. The first components B_1 arose because of non-overlapping support. For some participants there are no comparable non-participants. The second component B_2 arose from different distribution of X_i within the two populations. The third component B_3 reflects the differences in outcomes that remain even after conditioning on observables or on p_i, and is the well-known selection bias

component. Their analysis showed that the first two components are the most significant ones in the sense that they contribute the most to the total bias. After accounting for these sources, B_3 is statistically insignificant different from zero. Based on the hypothesis of selection on observables, this evidence suggests that simply by matching on p_i or a different balancing score does not eliminate total bias but rather selection bias under some conditions. The largest sources of bias are eliminated, provided the evaluation parameter is estimated over a region of common support.

These arguments prove that neither matching on p_i nor matching on X_i is an optimal solution in the sense that they do not produce efficient estimators. However, if the question is what an analyst can do, in practice, in order to solve the evaluation problem through matching, based on the above arguments we would suggest to match on the probability of participation and estimate it parametrically by using an appropriate logit model:

$$p_i = P(D_i = 1|X_i) = \frac{P(D_i = 1) \times P(X_i|D_i = 1)}{P(D_i = 1) \times P(X_i|D_i = 1) + P(D_i = 0) \times P(X_i|D_i = 0)}$$

Alternatively, one may manipulate the above equation to obtain the odds ratio:

$$q(X_i) \equiv log\left[p_i \Big/ 1 - p_i \right]$$

Either estimator can perform matching. Rosenbaum and Rubin (1985) demonstrate that matching in terms of $q(X_i)$ may be preferable in order to "avoid the compression of the p_i scale near 0 and 1 and, moreover, $q(X_i)$ is more nearly normally distributed".

4.5 Matching Estimators

Generally, for each observation *i* in the participant sample, a weighted average of comparison sample observations is formed to estimate the effect of treatment to this observation :

$$Y_{1i} - \sum_{j \in I_0} W_{N_0,N_1}(i,j) \times Y_{0j} \tag{4.1}$$

where $W_{N_0,N_1}(i,j)$ is usually a positive valued weight function, defined so that

$$W_{N_0,N_1}(i,j) = \begin{cases} 1 & \text{if } j \in A_i \\ 0 & \text{otherwise} \end{cases}$$

and N_0 and N_1 denote the number of individuals in I_0 and I_1, respectively. Several matching estimators have been proposed that exploit (C - 1) or (C′ - 1). They differ in the weights attached to members of the comparison group and the metrics that are used to obtain the match. Below we review the most common ones.

Nearest Available Matching on the Propensity Score.

In this method which was studied by Rubin (1973b) a nearest-neighbor algorithm is applied on the propensity score. Participants and non-participants are randomly ordered. Then, the first participant is matched with the non-participant having the nearest $q(X_i)$. In mathematical terms, denoting by i a participant and by j a non-participant define A_i such that only one j is selected according to the rule:

$$A_i^{NA} = \left\{ j \Big| \min_{j \in \{1,\ldots N_c\}} \left[q(X_i) - q(X_j) \right] \right\} \tag{4.2}$$

The effect of treatment on the treated for each person, as this parameter has been defined in paragraph 2.4, is calculated as:

$$E(\Delta_i | q(X_i), D_i = 1) = \frac{1}{N_1} \times \sum_{i=1}^{N_1} \left(Y_{1i} - W_{N_0,N_1}(i,j) \times Y_{0i} \right) \tag{4.3}$$

Matched persons are removed from the list of participant and non-participant persons and the process is repeated for the remaining unmatched individuals.

Mahalanobis Metric Matching Including the Propensity Score.

Mahalanobis metric matching has been described by Cochran and Rubin (1973). Again, in a non-experimental framework, participants and non-participants are randomly ordered. The first participant is matched with the closest non-participant in terms of the Mahalanobis distance:

$$D\left(q(X_i), q(X_j)\right) = \left(q(X_i) - q(X_j)\right)^T \times \Sigma^{-1} \times \left(q(X_i) - q(X_j)\right)$$

where Σ^{-1} is the sample covariance matrix of $\left(q(X_i) - q(X_j)\right)$in the non-participant reservoir. A non-participant *j* is selected using the expression:

$$A_i^{MM} = \left\{ j \middle| \min_{j \in \{1,\ldots N_c\}} D\left(q(X_i), q(X_j)\right) \right\}$$

The mean effect of treatment on the treated is calculated then by (4.3).

Nearest Available Mahalanobis Metric Matching Within Calipers Defined by the Propensity Score.

An extension of Mahalanobis Metric Matching on $q(X_i)$ was developed by Althauser and Rubin (1971) who define that the first participant *i* is matched with a non-participant *j* for whom:

$$A_i^{MMc} = \left\{ j \middle| \min_{j \in \{1,\ldots N_c\}} \left[D\left(q(X_i), q(X_j)\right) \prec c \right] \right\}$$

where *c* is a pre-specified tolerance known as *caliper*. As before, the mean effect of treatment on the treated can be estimated by equation (4.3). Cohran and Rubin (1973) provide a method to determine caliper's width. In the extreme case where no matching results within the caliper, the analyst has to match persons with simply the closest $q(X_i)$ in the sense of the Mahalanobis distance.

Nearest available Mahalanobis metric matching within calipers defined by the propensity score is a combination of the two previous methods. Applications showed that

it is better than the first "in that it yields fewer standardized differences" while it is also better than the second "in controlling the difference along the propensity score" (see Rosenbaum and Rubin, 1985).

Kernel matching

Another appealing way of matching can be derived by using *kernel* estimators. Kernel matching is an alternative method that manipulates the entire comparison sample to match on each participant. More specifically, instead of matching each participant with a specific non-participant (the closest one respectively to a measure), this method uses a weighted average for all non-participants as comparison units. Specifically, unlike the above estimators, kernel matching defines:

$$W_{N_0,N_1}(i,j) = \frac{K_{ij}}{\sum_{k \in \{D_i=0\}} K_{ik}}$$

where $K_{ik} = K\left[\left(X_i - X_k \right) \Big/ \alpha_{N_0} \right]$ is a kernel that downweights distant observations and α_{N_0} is a sequence of smoothing parameters (bandwidth) with the property that $\lim_{N_0 \to \infty} \alpha_{N_0} \to 0$. The impact of treatment on the treated, usually estimated in a particular domain $X_S \in X$, is:

$$E(\Delta_i | D_i = 1, X_i) = \sum_{i \in I_1} w_{N_0,N_1}(i) \left\{ Y_{1i} - \sum_{j \in I_0} W_{N_0,N_1}(i,j) \times Y_{0j} \right\}$$

where different values of $w_{N_0,N_1}(i)$ may be used to select different domains X_S or to account for heteroscedacity in the treated sample. Heckman, Itchimura and Todd (1997) develop an asymptotic distribution theory for kernel-based matching estimators. This theory covers both the cases of a known propensity score and of an estimated one. Furthermore, they demonstrate that conventional functional form restrictions (exclusion restriction and additive separability) invoked in econometrics may improve the impact estimates obtained from kernel matching. Upon this finding they formalize an alternative

approach that is known in the literature as *local-linear matching* and yields smaller variance for the resulted matching estimator. However, its efficiency does not necessarily increases.

Heckman and Smith (1998) and Heckman, Itchimura, Smith and Todd (1998) describe the construction procedure of a kernel estimator of the counterfactual outcome for participant *i*. Using all the comparison group observations, they run a weighted regression with Y_{0j} as the dependent variable. The regression contains only an intercept term and the estimated intercept is the kernel estimate of the counterfactual outcome for participant *i*.

Relatively to the nearest neighbor matching, kernel matching reduces the variance of the matching estimate by making use of information from additional non-participant observations. At the same time though, it increases the bias in small samples because the additional observations are more distant, in terms of participation probabilities, from the observations being matched.

4.6 Other Methods

Apart from matching, other methods have been also suggested for the construction of the desired counterfactual mean, $E\left(Y_{0i}\middle|D_i=1,X_i\right)$, to solve the evaluation problem. These methods are briefly outlined below.

The method of Subclassification.

This method, referred in Cochran (1965), attempts to gain most of the advantages of matching with less expenditure of time. Suppose we have data-lists from a population of participants and from a population of non-participants, including their outcomes Y_{1i} and Y_{0i} respectively and a vector X_i of attributes for each person. Subclassification stratifies each population into subclasses by the values of X_i. Within a given subclass, samples of the same size are drawn from each population, but are not individually matched. Instead

for the calculation of the mean effect of treatment on the treated the usual average estimator is applied:

$$E\left(\Delta_{i(k)}\middle|D_{i(k)}=1,X_{i(k)}\right)=E\left(Y_{1i(k)}\middle|D_{i(k)}=1,X_{i(k)}\right)-E\left(Y_{0i(k)}\middle|D_{i(k)}=1,X_{i(k)}\right)$$

where k refers to a specific subclass. The overall mean effect of treatment on the treated is:

$$E\left(\Delta_{i}\middle|D_{i}=1,X_{i}\right)=\frac{1}{N_k}\times\sum_{k=1}^{N_k}\ E\left(Y_{1i(k)}\middle|D_{i(k)}=1,X_{i(k)}\right)-E\left(Y_{0i(k)}\middle|D_{i(k)}=1,X_{i(k)}\right)$$

where N_k denotes the number of subclasses. In this approach $E\left(\Delta_{i}\middle|D_{i}=1,X_{i}\right)$ is free from bias due to differences between means of X_i in different subclasses. Nevertheless, some bias may remain from variation of X_i within subclasses since subclasses will not be homogeneous in X_i.

Rosenbaum and Rubin (1983a) indicate the major problem of subclassification according to which as the number of confounding variables increases, the number of subclasses grows dramatically. In other words they constitute the dimensionality problem for subclassification. As a solution they suggest subclassification on the propensity score and establish a corollary based on theorem 2, under which subclassification on p_i obtains an unbiased estimator (see Appendix 2, Corollary 4.2).

The method of covariance adjustment

This method, also described in Cochran (1965), consists of conducting a regression of Y_i outcomes on the various X_i attributes for participants and non-participants separately. The samples of participants and non-participants in which the regressions are performed may be matched or random. Then, by adjusting $E\left(\Delta_{i}\middle|D_{i}=1,X_{i}\right)$ to remove the effects of regression, one can compute the mean effect of treatment on the treated. Rosenbaum and

Rubin (1983a) provide a corollary that consecrates the covariance adjustment on a balancing score, like the propensity score p_i (see Appendix 2, Corollary 4.3).

The econometric procedure of Barnow, Cain and Goldberger (1980)

This method produces a regression estimator that exploits condition (C – 1) in a linear regression setting. It assumes that Y_{0i} is linearly related to observables X_i and an unobservable u_{0i}, so that:

$$E\left(Y_{0i}\middle|D_i=1,X_i\right)=X_i\beta+E\left(u_{0i}\middle|D_i=1,X_i\right)$$

so that $E\left(u_{0i}\middle|D_i=0,X_i\right)=E\left(u_{0i}\middle|X_i\right)$ is linear under X_i. Controlling for X_i via linear regression allows one to identify $E\left(Y_{0i}\middle|D_i=0,X_i\right)$ and thus, from the conditional independence assumption, $E\left(Y_{0i}\middle|D_i=1,X_i\right)$. In this way the desired counterfactual mean is obtained.

Heckman, LaLonde and Smith (1999**)** describe the above method. They explain that its advantage lies on the parsimonious usage of the available data and therefore constitutes an alternative to matching on the propensity score. However, its major limitation is that it discards a major advantage of the matching methods because it forces the investigator to make arbitrary assumptions about functional forms of estimating equations. Moreover, in practice, users of this method do not impose a common support condition in generating the estimates obtained from the method. When the distribution of X_i is different in $(D_i = 0)$ and $(D_i = 1)$ samples, the comparability is only achieved by imposing linearity and extrapolating over different regions.

It is obvious, finally, that this last procedure is closely related to the covariance adjustment method in that they both use regression of Y_i*'s* on X_i*'s* to cope with the evaluation problem. Although it seems that they share the major limitation of arbitrary functional assumptions, they differ in an important point. Covariance adjustment method does not impose the linear relationship of Y_i on X_i. As Rubin (1979) comments the response variable Y_i (outcome) may be regressed nonlinearly on X_i or sophisticated

Bayesian and empirical Bayesian methods can be performed over a variety of nonlinear models for the response surfaces.

4.7 Identification of the Impact Distribution

In the context of an evaluation study, estimation of the distribution of outcomes is of great importance for the analyst. However, the inability to observe simultaneously Y_{0i} and Y_{1i} for each person does not allow us to gain insight on several useful features and different interpretations of the program impacts. This limitation is also present in matching methods.

Similarly to the randomized social experiments, matching can easily evaluate the marginal distribution of outcomes $F_1\left(Y_{1i}\middle|D_i=1,X_i\right)$ and $F_0\left(Y_{0i}\middle|D_i=0,X_i\right)$ from the available data. Though, estimation of the joint distribution of outcomes $F\left(Y_{1i},Y_{0i},D_i\middle|X_i\right)$ is not a simple task.

A common feature of experimentation and matching is the set of assumptions imposed for the identification of $F\left(Y_{1i},Y_{0i},D_i\middle|X_i\right)$. Specifically, by assuming the common effect model for the matched persons, one can easily derive the joint distribution of outcomes, as described in paragraph 3.7.1. Yet, in the more general case of heterogeneous preferences, the experimental bounding approach can be adapted to matching. To obtain informative bounds, one has to assume the same dependence across the different quantiles of Y_0 and Y_1 for the matched persons.

In addition, since matching is based on observed information on sampled individuals, it can be stated that participants' statements about ex-post expectations of the program impacts is useful in determining the joint distribution of outcomes. Participants have information not available to external program evaluators on issues such as certain components of the cost of program participation or the value of outcomes by participant relative to their cost. Participants' self-reports, in most cases, lead to more informative data about Y_{0i} and Y_{1i} and as a result to a better estimation of the joint distribution of outcomes.

4.8 Discussion on Matching

Matching is a widely used method of evaluation. It is based on the idea of contrasting the outcomes of program participants, Y_{1i}, with the outcomes of "comparable" non-participants, Y_{0i}. In this way it can be thought as a substitute for experiments. By aligning the distribution of observed characteristics on the $D_i = 0$ population with that in the $D_i = 1$ population, matching mimics the basic feature of the randomized data. The only difference is that, unlike randomization procedures, it uses econometric methods to produce a comparable group of non-participants, termed as comparison group.

Several important authors have studied the method of matching either theoretically or practically. Cochran and Rubin (1973) summarize the work on the efficacy of univariate pair matching procedures. Rubin (1979) applies matching techniques at data from a Monte Carlo simulation. He studies the nearest available pair matching and the nearest available pair matching based on Mahalanobis distance and compares these approaches. He points out that if the distributions of X_i diverge widely, none of the above methods can be trusted to remove all, or nearly all, the bias. Rosenbaum and Rubin (1983a) suggest a solution that removes substantial selection bias. That is matching on p_i instead of X_i. However, as Heckman, Itchimura and Todd (1998) mention this is effective only when p_i is known or at least estimated parametrically. In both situations, though, the variance as a choice criterion between matching on X_i or p_i indicates that "nothing is absolutely true".

Rosenbaum and Rubin (1983a, 1985) develop a theory of matching methods and use real data from a Danish cohort to support it. They mention that in many cases is unprofitable to conduct experiments instead of matching since the costs of experiments are high. Matching can give adequate results free from selection bias in a smaller cost.

On the other hand, Heckman and Smith (1998) argue that the major limitation of non-experimental matching compared to randomized experiments is that the latter guarantees that:

$$Support\left(X_i \middle| D_i = 1, r = 1\right) = Support\left(X_i \middle| D_i = 1, r = 0\right)$$

while in the former this is not always satisfied. The inability to find comparable comparison group members for program participants is a major source of bias for the matching estimator.

Heckman, Itchimura and Todd (1997) outline the main problems an analyst is faced with when applying matching methods:

- Matching on measured characteristics available on a typical non-experimental study is not guaranteed to produce a truly comparable group of non-participants like experiments do. Even if strong assumptions are invoked, these assumptions will be inconsistent with many economic models of program participation in which agents select into the program on the basis of unmeasured components of outcomes unobserved by the researcher.
- Even if a valid comparison group can be found, the distribution theory for the matching estimator remains to be established for continuously distributed matching variables X_i.
- Matching cannot be applied in the situation where the selection occurs on unobservables.
- Matching is a data-hungry method. With a large number of conditioning variables, it is easy to have many cells without matches. This makes the method impractical or dependent on the use of arbitrary sorting schemes to select hierarchies of matching variables.

Despite those pessimistic comments, matching has been applied extensively by statisticians and economists who, in several situations, have produced plausible and interesting results. An illustrative example is the work of Barron, Black and Loewenstein (1989) who analyze data of workers' payments under different amounts of on-the-job training. They find that workers pay part of their on-the-job training costs by accepting a lower starting wage and waiting for a higher future wage after the completion of the training. However, the relationship "higher trainings usually means lower starting wage" that resulted earlier studies of human capital, is not verified. The authors have also found that in more demanding positions, employers spend more time per applicant during the

screening process and, on the average, they see more candidates. Finally, they concluded that wage growth and productivity growth are positively related to on-the-job training.

Barron, Berger and Black (1997) also conduct matching methods in order to compare various measures of on-the-job training. They establish that informal training is more often applied than formal training. Moreover, informal training is measured as accurately as formal one while the effect of training is found to be underestimated in several studies due to measurement errors.