EVALUATING PROGRAMMES: EXPERIMENTS, NON-EXPERIMENTS AND PROPENSITY SCORES

Denis Conniffe, Vanessa Gash
and
Philip J. O'Connell

March 2000

Working Paper No. 126

Evaluating Programmes: Experiments, Non-Experiments and Propensity Scores

DENIS CONNIFFE, VANESSA GASH and PHILIP O'CONNELL
The Economic and Social Research Institute, Dublin

Abstract: Evaluations of programmes - for example, labour market interventions such as employment schemes and training courses – usually involve comparison of the performance of a treatment group (recipients of the programme) with a control group (non-recipients) as regards some response (gaining employment, for example). But the ideal of randomisation of individuals to groups is rarely possible in the social sciences and there may be substantial differences between groups in the distributions of individual characteristics that can affect response. Past practice in economics has been to try to use multiple regression models to adjust away the differences in observed characteristics, while also testing for sample selection bias. The Propensity Score approach, which is widely applied in epidemiology and related fields, focuses on the idea that "matching" individuals in the groups should be compared. The appropriate matching measure is usually taken to be the prior probability of programme participation. This paper describes the key ideas of the Propensity Score method, compares it with the common approach in economics, reviews the arguments in the literature and illustrates application by reanalysis of some Irish data on training courses.

I INTRODUCTION

Application of the direct experimental approach in the economy and society is usually considered unpalatable, or even unethical, even when it would clearly provide the ideal comparison. For example, we would like to assess an active labour market policy – say, training to enhance skills – by drawing a large and fully random sample from the relevant population and then randomly assigning individuals to a training group and a control group. Then, although an individual's subsequent average performance (in terms of employment, earnings, productivity, or whatever) will depend on characteristics like age, education and previous work experience, these factors cancel out of the difference in the averages for the two groups[1]. So the difference can be validly interpreted as the effect of the programme or policy. While there have been a few such assessments in the US (for example, LaLonde, 1986), allocation to a control group can be seen as disadvantageous, so that randomisation is unpopular, to say the least.

Evaluations have sometimes been based on the performances of the programme participants only, without employing any control group. But as it is most unlikely that, in the absence of a programme, individuals would not have tried to improve their own positions, information has to be sought about this, or else the programme benefits could be considerably overestimated. An Irish example is

[1] The ideal comparison, though impossible to make, would use the *same* people to compare the effect of participation in the programme with non-participation. In experimental approaches the *causal* effect of a treatment on an individual is defined as the difference in the potential responses when receiving and not receiving the treatment. The average of these differences over the whole population is the parameter of interest. While theoretically a useful concept, it is usually unmeasurable and has to be assumed estimable by the difference between the means of the treatment and control groups, given prior randomisation of individuals to groups.

provided by Breen and Halpin (1988), who evaluated the FAS *Enterprise* programme by interviewing a sample of participants and, besides ascertaining how well they had got on, also asked what they would have done had the programme not existed. Again, Breen and Halpin (1989), in assessing a job subsidisation scheme, asked employers if they would have hired anyway in the absence of the scheme. In both these studies and, no doubt, in many others, there was simply no other way to proceed. But depending on questions of this nature, with all the possibilities for "wisdom by hindsight", seems less attractive than comparing with a 'control' group, even if the allocations of individuals to groups has been far from random.

Situations when we have observational data on a programme (henceforward called treatment) group and on a control group, but without the deliberate randomisation of individuals to groups that characterises true experimentation, are often called "natural experiments" in the economics literature. Without randomisation, there may well be substantial differences between groups in the distributions of individual characteristics that affect performance (henceforward called response). Sometimes quite simple methods are used to analyse the data. For example, if responses of all individuals are measured at two time points, corresponding to before and after treatment for the treatment group, the "difference of differences" method may be employed. This just compares the mean improvement in response for the treatment group with the mean improvement (if any) for the control group. The idea is that in taking "before" from "after" individual characteristics may cancel out. For example, this type of analysis was used by Eissa and Liebman (1996) in their study of labour supply response to an earned income tax credit.

However, most analyses of "natural experiments" have taken a far more sophisticated form. At simplest, a regression model, perhaps embodying economic theory and previous findings, is estimated relating the response to all the characteristics as well as to a dummy variable defining groups. This is interpreted as adjusting the treatment versus control comparison to what it would have been had there been no variation in characteristics between groups. The added possibility of sample selection biases is addressed by postulating another model expressing the probability that an individual is included, or not included, in the treatment group as a function of the individual's characteristics. When estimated, this model *may* permit adequate adjustment to compensate for the biases. This econometric modelling

approach to the analysis of "natural experiments"[2], which will be reviewed in Section II, owes much to the work commenced by Heckman (1976, 1979) and continued in subsequent papers. However, for reasons to be discussed, the greater sophistication of approach does not always pay off in terms of elimination of bias, precision of estimation or clarity of findings .

The main purpose of this paper is to present an alternative approach based on propensity score matching. This has been developed in the statistical and biometrical, rather than in the econometric, literature. The propensity score for an individual is the probability that he or she belongs to the treatment group and is a function of the individual's characteristics. 'Matching' is a simple device advocated by Cochran (1965, 1968) to help draw causative inferences from survey data. The Propensity Score method, which derives from the papers by Rosenberg and Rubin (1983a, 1984), will be detailed in Section III and contrasted with the modelling and Heckman adjustment approach. The approach already dominates in biomedical fields and many expository papers and reviews have appeared, including those by Drake and Fisher (1995), Rubin (1997), D'Agostino (1998) and Perkins, Tu, Underhill, Zhou and Murray (2000) as well as papers describing variants to methods (for example, Tu, Perkins, Zhou and Murray, 2000) and a host of papers describing applications to specific biomedical observational studies[3].

There has been very little application of the propensity score method in the economic literature as yet, although there has been some debate in theoretical journals about methodological issues and whether variations on the approach, or even outright alternatives to it, might be more appropriate to the social sciences. We review this literature in Section IV and emphasise the degree of agreement that actually exists, in spite of apparently contentious matters. Finally, in Section V we illustrate the propensity score method by reanalysing data used by O'Connell and McGinnity (1997) to evaluate some training programmes.

II THE ECONOMETRIC, OR SPECIFICATION ERROR, APPROACH

If individuals from the population of interest were randomly allocated to treatment and control, we would be happy to accept the difference in mean responsiveness as a consistent estimate of the

[2] Some authors reserve the term "natural experiments" for cases where only simple analyses like the "difference of differences" are employed and call what is being described here the "modelling approach". However, we think the defining feature is the presence of treatment and control groups.
[3] However, the term "Natural Experiment" is not used outside the economic literature.

programme effect[4]. In a somewhat more roundabout account, we are fitting the model

$$y_{ij} = a + bD + e_{ij}, \qquad (1)$$

where y_{ij} is the response (taken as continuous, for the present) of the j th individual in group i (with i=1 for the treatment group and i=0 for the control group), D is a dummy variable equal to 1 for the treatment group and to zero for the control group, a and b are intercept and coefficient respectively, and e_{ij} is the error or disturbance term. This term contains the effects of such characteristics as age, gender, education, etc. The true means of the control and treatment groups are a and a+b respectively and so b is the true difference in means. Standard regression is appropriate because D and e_{ij} are independent (by randomisation), which is the fundamental condition for consistent estimation by Ordinary Least Squares. The OLS estimate of b is, as expected,

$$\bar{y}_1 - \bar{y}_0, \qquad (2)$$

where the over bars denote means. The numbers in the groups do not, of course, need to be equal. With non-randomised observational data D and e_{ij} in (1) are no longer independent and so (2) is no longer a plausible estimate. The approach in the econometric literature has then been to treat model (1) as "misspecified" because it does not contain all the covariates (of which there are p, say) that can affect the response. Instead

$$y_{ij} = a^* + b^* D + \sum_{k=1}^{k=p} c_k x_{kij} + u_{ij}, \qquad (3)$$

where x_{kij} is the value of covariate k as measured on individual j of group i, is estimated. The asterisks on a and b just indicate they are somewhat different from what they were in equation (1), now that the covariates are included. Perhaps the OLS estimate of b^* is now an adequate measure of the treatment effect, having been "adjusted" for differences in covariate values, although there is still the possibility of self-selection bias. But before considering that topic, it is important to note the assumptions already being built into (3). The effects of covariates are being assumed linear and to operate identically in the two groups. The consequences of departures from assumptions on the

[4] It should be said that some literature on programme evaluation questions the acceptability of this difference as the measure of programme effect and might even deny the appropriateness of the experimental paradigm. The points will be returned to, but are ignored for the present to simplify development.

estimate of b^* are far more serious when the distribution of covariate values differs greatly between groups than when it does not. Of course, it could be argued that sometimes economic theory or researchers' prior experience mean they "know" the true model. Or that the researcher may have tried out different functional forms, considered interactions of treatment with covariates etc. and, via goodness of fit measures and specification tests, found the "right" model to replace (3). But very often researchers are unsure about the correct model, a variety of equations may fit almost equally well (or poorly) and data may not be fully informative for various reasons, including multicollinearity. Even more seriously, some data points may be totally uninformative about the comparison of treatment and control and their inclusion just serves to magnify specification biases[5].

Turning to self-selection bias, Heckman's (1979) formulation of the problem and its solution was also in terms of misspecification due to an omitted variable. Initially, Heckman just considered the drawing of one sample, with possible self-selection operational, from a population. So the situation was somewhat simpler than where two samples, corresponding to treatment and control, are involved. So, for the present, we rewrite equation (3) as if there is only one group, so that b, D and the subscript i do not appear. The model is now:

$$y_j = a + \sum_{k=1}^{k=p} c_k x_{kj} + u_j, \qquad (4)$$

where u is assumed to have mean = 0 and variance = σ^2. Given a truly random sample the coefficients could be estimated by OLS. Suppose the selection bias takes the form that sample y values are only observed if an unobserved, or latent, variable w is positive. However, x values are not only observed for the sample, but also for another sample (for which the y's are not), or even for the whole population.(Many selection mechanisms can be reformulated into this framework – for example if y is a working mother's earnings and the x's are age, years of education, etc., w could be the difference between earnings and perceived costs incurred by joining the workforce.) If w is independent of e there is still no difficulty, but it may well not be (in the example, the particularly high earners are more likely to exceed their perceived costs) and it may also depend on the x's. Then u, instead of being random with mean zero, will have a component which is related to the x's and the fundamental condition for OLS is broken. Heckman showed progress is possible if each w_j is assumed normally distributed with unit variance and mean

$$d_0 + \sum_{k=1}^{k=p} d_k x_{kj} = -z_j \quad \text{, say.} \tag{5}$$

Then, although w is only known to the extent of its sign, Heckman demonstrated that the true model is got by adding another variable λ_j, which is a function of z_j, to (4), giving

$$y_j = a^* + \sum_{k=1}^{k=p} c_k^* x_{kj} + g\lambda_j + u_j, \tag{6}$$

with

$$\lambda_j = \frac{\phi(z_j)}{\Phi(-z_j)}, \tag{7}$$

where ϕ is the ordinate (or density) and Φ is the integral (or distribution function) of the Standard Normal. Formula (7) is called the "inverse Mill's ratio", or sometimes "Heckman's lambda". As before, the asterisks in (6) just indicate that these coefficients differ from those in (4) in requiring interpretation as effects of covariates holding the new variable λ constant. Of course, the z's needed to calculate the λ's depend on the unknown d coefficients in (5) and this necessitates a two-stage estimation process. First a binary variable is defined, equalling 1 when y is observed and zero when it is not, and a probit analysis is conducted with the x's as explanatory variables to estimate the d's. Then the λ's are calculated and the coefficients of (6) estimated.

With this formulation of the selection bias problem, the observed covariates are being taken as quite sufficient for adequate model specification and estimation. The x's actually come into equation (6) twice – explicitly in the second term on the right hand side and implicitly in λ. It could be said that Heckman sees the effect of an omitted covariate w as replacing equation (4), which is linear in the x's, by an equation (6), which still involves only the x's, but which has the non-linear component λ. Indeed, estimation of (6) could be approached as a case of non-linear equation estimation. Note, however, the crucial importance of the assumption, embodied in (5), that the unobserved w has a linear regression on the observed x's and is normally distributed. These assumptions permit the x's to "handle" the omitted variable and identify the appropriate non-linear adjustment to the equation.

The situation when comparing treatment and control groups is just a little more complicated. Now individuals may select themselves into either the treatment or control group. For example, individuals of above average ability (the unobserved w could now be individual's ability minus mean ability and is positive for the treatment group) might judge a training programme to be likely to benefit them and opt

[5] This point will be returned to.

for training, while those of below average ability might judge themselves unable to benefit from training and opt for the control group[6]. The Heckman lambda values, λ_{tj}, for the treatment group are obtained from z_{tj} values derived from a probit analysis where the binary variable equals 1 for the treatment group and 0 for the control group and the x's are the explanatory variables. They are given by (7). In the corresponding derivation for the control group, w is negative, the 1's and 0's of the probit analysis are interchanged, but the covariates are the same. So it is intuitively clear that the λ_{cj} values are given by

$$\lambda_{cj} = \frac{-\phi(-z_j)}{\Phi(z_j)} = \frac{-\phi(z_j)}{\Phi(z_j)}.$$

The modification of model (3) to be actually estimated is then

$$y_{ij} = a' + b'D + \sum_{k=1}^{k=p} c_k' x_{kij} + g\lambda_{ij} + u_{ij}, \qquad (8)$$

where the primes indicate coefficients have changed from those of equation (3) because of inclusion of the lambda variable. A consistent estimate of b' will now follow from OLS[7], assuming, of course, that all has been specified correctly in the model.

As in the sample from a population case, the approach is heavily dependent on the postulated selection bias process corresponding reasonably to reality and on the bivariate normality assumption about w and y. It is also sensitive to specification errors. For example, when the response equation should have contained some non-linear components, it is known that the λ_{ij} term in (8) can pick these up, suggesting there is selection bias where none exists. In other cases the λ_{ij} as estimated functions of the x's can be nearly collinear with the covariates in the model, leading to unstable coefficients, high

[6] Postulating such behaviour might seem to be attributing near prophetic powers to the individuals in both groups, since they have neither experienced the programme nor is any objective assessment of its effectiveness yet available. However, such postulated behaviour has also been used (for example in Heckman and Smith, 1996) to make radical criticisms of the standard experimental paradigm. Not only should the measure of treatment effect really include some selection effect, but randomisation could destroy an experiment, because high ability individuals would not accept assignment to a control group and by overt or covert means would attain the treatment group's conditions. However, Angrist, Imbens and Rubin (1996a) have incorporated "compliers" and "refusers" into the experimental framework.

[7] Because some variance heterogeneity has been introduced at the probit analysis stage, OLS, while consistent, may be less efficient than the appropriate GLS or Maximum Likelihood solution. Many econometric computing packages now provide these procedures and we will not detail them here, but rather concentrate on the ideas and assumptions embodied in the econometric approach to the problem.

standard errors and insignificant "t" values. Then it can be difficult to conclude anything at all from the results after "adjusting" for selection bias. This sensitivity is reduced if some of the variables defining the mean of w_j in (5) do not occur (that is, are known to have zero coefficients) in the response equation (4). So some standard textbooks (for example, Johnston and Di Nardo, 1997, p. 450) recommend that Heckman correction should only be performed in these circumstances.[8]

It is worth noting that knowledge of zero coefficients (or "zero exclusions", as they are termed) does permit a possible alternative analysis to the "standard Heckman" via (8). The zero exclusions define instrumental variables and so direct instrumental variable (IV) estimation is also possible. Returning to equation (3), the dummy variable D for treatment v control could now be regarded as endogenous. However, the IV method does not try, or need, to specify the precise source of endogeneity. Given instrumental variables, that is, some c's zero, the endogeniety can be "washed out" by a procedure closely analogous to two stage least squares. No explicit assumptions about a latent variable or its distribution, or about the precise mechanism of the selection need be made. This may sometimes make the approach preferable to the "standard Heckman" analysis. Sample selection biases can easily be visualised as operating in a much more complex way than the scenario of people of high ability opting for training and of low ability opting for the control. There could be selection biases originating with the programme administrators, possibly interacting with ,or countering, selection associated with individuals' abilities. There could be selection effects at various stages of programmes – perhaps at recruitment to the programme before separation into treatment and control, then at the treatment allocation stage and perhaps selective dropout from the programme. Then it might be sensible to just think of an endogenous D as an aggregate effect of an assembly of selection biases and use the IV method.

The IV approach is still, of course, quite valid if the "standard Heckman" scenario holds. The high ability person opting for the training programme can be regarded as affecting both y and D, making the latter endogenous, which can be corrected by IV estimation. However, it could be argued that the IV method neglects information under these circumstances and is less efficient and less informative than the standard method. However, many authors have considered the assumptions required for Heckman correction of model (8) to be very fragile. Criticisms go back in the econometric literature as far as

[8] If the response variable is itself binary - for example, if a programme is being evaluated in terms of employment gain – at least one zero coefficient is essential for any estimation.

Goldberger (1983), but have recurred strongly in recent propensity score and causal modelling literature, for example in Angrist et al (1996a) and Imbens and Rubin (1997). Empirical findings have sometimes differed with the approach to selection bias. The belief, following "standard Heckman" correction, that OLS estimation overestimates returns to education (because of selection on "ability") has not been borne out by some IV analyses of US data (see Angrist and Kruger, 1991; Imbens and Rubin, 1997; and other references contained in the latter.

Using the term "standard Heckman" may actually be unfair to Heckman, whose views on the appropriate treatment of selection bias have evolved over time. Heckman and Robb (1986) and Heckman and MaCurdy (1986) were very positive about IV estimation and also examined other alternatives to the "standard Heckman" approach. Heckman and Hotz (1989), defending non-experimental studies of manpower training from criticisms about conflicting findings, emphasised that several selection bias adjustment procedures exist and that the correct choice depended on the source of the selection bias (see, especially, the reply to Holland, 1989). Heckman (1990) tried to relax the assumptions further[9] and move to non- or semi-parametric approaches and some of his more recent papers will be considered in the next section. But it is the "standard Heckman" procedure that is in the textbooks and econometric software packages and that has been very frequently employed in programme evaluation and other assessments, such as on returns to education. Irish examples include Breen (1986); Breen (1991); Breen, Hannan and O'Leary (1995); O'Connell and Lyons (1995) and O'Connell and McGinnity (1997).

Returning to IV estimation, it is often very difficult to find truly valid and effective instrumental variables. Few economists can be fully confident in relying on economic theory to provide a set of "zero exclusions". Inserting the candidate instrumental variables into the model, along with the treatment dummy variable and the "true" covariates, and finding their coefficients not statistically significant would provide some support. But the lack of statistical significance could be because the model did not fit well anyway, due to incorrect functional form, or important omitted covariates, or whatever. The choice and framing of assumptions justifying IV estimation, as presented in some econometric papers on programme evaluation and selection, have been subject to critical scrutiny by such authors as Little (1985), Angrist (1995) and Angrist et al (1996a). These criticisms have been disputed and there have even been disagreements over precisely what assumption is implied by a zero

[9] So did Newey, Powell and Walker (1990)

exclusion. For example, Angrist et al (1996a) state that a valid instrumental variable z has to be *independent* of the response variable y given the non-zero coefficient variables (that is, independent of the disturbance term). Heckman (1996) maintained the required assumption is much weaker: z has only to be *uncorrelated* (mean independence) with the disturbance term. Probably not everyone would agree the assumption is much weaker, but in any event Angrist et al (1996b)[10] replied that even if, without independence, there was no correlation with y as the response variable, there would be if log y (or another function) replaced y. Overall, recent work on IV stresses emphasis on ensuring the validity of instruments, although those that survive scrutiny are often found less related to the endogenous regressor of interest that might be desired (see, for example, Imbens and Rubin, 1997). These considerations will also be relevant in the next section when we discuss using IV in conjunction with propensity scores.

III THE PROPENSITY SCORE APPROACH

In the statistical analysis of designed experiments, there is rarely any attempt to estimate a "true" model for response in terms of *all* the variates that could affect it. Instead the model usually relates response to the factors being deliberately varied by the researcher and relies on a combination of seeking homogeneity of experimental material and randomisation to balance out all the other factors. For programme evaluation, this implies the treatment v control comparison is of paramount interest and that the covariate effects, if not subjected to randomisation, are just complicating nuisances. It might be argued that the relationship between response and some covariate is of interest in its own right. But that can probably be studied (and perhaps has) without the complication of conducting a comparative social experiment[11]. Most causal findings in science have followed from controlled or randomised experiments and social scientists frequently complain they cannot conduct such trials. From this perspective, the problem with model (1) is not that covariates require specification, but the lack of randomisation to D=1 and D=0. In a sense, the *entire* problem is specification bias in allocation between treatment and control.

The propensity score solution depends on the idea of "matching" individuals from the treatment and control groups. Cochran (1968) gave the example of mortality rates for US smokers being lower, on

---

[10] This argument appears elsewhere in the literature, too, for example in Imbens and Rubin (1997).
[11] Matters may be different if a particular treatment by covariate interaction (for example, what training achieves for females) is considered vital to the evaluation. But this is really a definition (which should have preceded the study) of a key sub-population and a stipulation that there be two treatment groups and two controls.

average, than for non-smokers – the reason being that smokers were younger, on average, than non-smokers. When groups of smokers and non-smokers of equal ages were compared, the mortality rates were always higher for smokers. Cochran advocated seeking causative effects from observational data by matching individuals from the treatment and control using all the covariates. If no important covariate has been omitted, it seems plausible to suppose that the difference between the responses of two such matching individuals, one receiving the treatment and the other the control, is the treatment effect plus a random element. Then averaging over the set of differences estimates the treatment effect. Cochran showed that perfect matching, in terms of exact equality of continuous covariates, is unnecessary and that matching on intervals can work well. Nonetheless the method will meet trouble if there are a lot of covariates, because the number of matching cells increases exponentially with the number of covariates and cells could quickly become empty of treatment individuals, or control cases, or both. That difficulty could be overcome, however, if all covariates could somehow be combined into a single "balancing score". Several ways of constructing such a balancing score have been proposed, but Rosenberg and Rubin's (1983a) "Propensity Score" approach is overwhelmingly the most popular.

The propensity score for an individual is the a priori probability (which is a function of the covariate values) that the individual is in the treatment group. Rosenberg and Rubin show that if we consider two sets of people, one set in the treatment group and one in the control group, with the same value of the propensity score, then the two sets have the same distributions of covariates. Rosenberg and Rubin gave a rigorous version of the intuitive argument that follows. If two individuals, one in the treatment group and one in the control group, have the same propensity score, their subsequent "allocation" to treatment or control can be regarded as random. It may as well have been decided by a coin toss. The difference in their responses is the treatment effect plus a random element and averaging over the set of such differences estimates the treatment effect. So although individuals have not actually been randomly allocated to treatments, the fact that overall distributions of covariates differ between the groups is "ignorable", given matching on the propensity score.

A Propensity Score analysis commences with estimation, by probit or logit, of a treatment assignment equation, where all known covariates affecting assignment and response are included as explanatory variables and the observed "dependent" variable is D=1 for an individual in the treatment group and D=0 for someone in the control group. Then propensity scores are calculated for all individuals and some matching process is implemented. The most commonly employed is stratification of the

propensity score distribution by quintiles or sextiles - the "binning" procedure. Then the distributions of covariates for treatment and control within each subclass are compared and, if they still differ appreciably, the assignment equation is further developed. For example, if a particular covariate still differs between groups within subclasses, the assignment model could be modified by trying powers of the covariate and its interactions with other variables. It is important to search for the best model for participation. When a satisfactory model is arrived at, the treatment v control effect on the response variable within subclass i is just the difference in means $\bar{y}_{1i} - \bar{y}_{0i}$, if the response is continuous, or a difference in proportions $\hat{p}_{1i} - \hat{p}_{0i}$, if the response variable is qualitative. Then the overall measure of treatment effect can be taken as simply

$$\frac{1}{s}\sum(\bar{y}_{1i} - \bar{y}_{0i}), \tag{9}$$

where s is the number of bins, or strata, in which there are both treatment and control units and the summation is over these strata. The standard error is

$$\frac{1}{s}\sqrt{\sum\left\{\frac{\hat{\sigma}_{1i}^2}{n_{1i}} + \frac{\hat{\sigma}_{2i}^2}{n_{2i}}\right\}}, \tag{10}$$

where the $\hat{\sigma}_{1i}^2$ and $\hat{\sigma}_{0i}^2$ are the within group and within stratum i variances among the n_{1i} treated individuals and the n_{0i} controls. For example,

$$\hat{\sigma}_{1i}^2 = \frac{1}{n_{1i} - 1}\sum_j (y_{1ij} - \bar{y}_{1i})^2.$$

If the response variable is qualitative, formulae (9) and (10) become

$$\frac{1}{s}\sum(\hat{p}_{1i} - \hat{p}_{0i}) \qquad \text{and} \qquad \frac{1}{s}\sqrt{\sum\left\{\frac{\hat{p}_{1i}\hat{q}_{1i}}{n_{1i}} + \frac{\hat{p}_{0i}\hat{q}_{0i}}{n_{2i}}\right\}}, \quad \text{with q = 1- p.}$$

The choice of (9) as an estimate needs some discussion. If the treatment versus control comparison could be presumed the *same* within all strata (that is, training has the same consequences for a low propensity individual as for a high one) then (9) is not the *best* estimate, although it is an unbiased one. The best estimate, in a minimum variance sense, would weight the within strata estimates inversely as their variances (so giving most weight to the most precise estimate). But if we want to allow for varying treatment effects across strata, we need to weight each stratum contrast in proportion to the

stratum's fraction of the population, which means equally, given quintiles or sextiles. The constant treatment effect assumption is implicit in the econometric approach of Section 2, but is not essential for the Propensity Score approach. However, it may well be plausible in some cases and can permit more precise estimates and more powerful tests of treatment effects.

It is possible that the Propensity Scores approach could fail to achieve a comparison of treatment with control. The difference within stratum i, $\overline{y}_{1i} - \overline{y}_{0i}$, obviously presupposes that there are some treated and control individuals present. If there are no representatives of one group, that stratum does not contribute to the comparison. If no stratum contains representatives of both groups, the approach fails, because there is no overlap in the propensity scores. The interpretation is that the characteristics (as measured by covariates) of the two groups are so dissimilar that no meaningful comparison is possible. This is not a disadvantage of the Propensity Scores method relative to the econometric modelling approach. The data deficiencies would feed into and undermine the regression analyses, although the cause of the problem might not seem at all obvious. One of the virtues of the Propensity Scores approach is that it reveals just how much of the data truly provide information on the comparison. Dehejia and Wahba (1998), who reanalysed Lalonde's (1986) data, emphasise this point. Rubin (1997), writing in the context of drawing deductions from health care databases, has stressed that the first use of propensity scores should be to decide whether a question of interest can be legitimately addressed to the database at all.

Some other advantages of the method deserve mention at this point. The approach is nonparametric as regards the response variable and is very sparing on assumptions. Nothing has been specified about the actual relationships of the response to the covariates, so avoiding the accumulation of biases due to the combination of model misspecifications and unbalanced covariates, which can have serious consequences for the regression modelling approach (for example, see Rubin, 1997). The estimation and testing of the treatment v control difference will usually be more efficient because the standard error will be smaller. Several factors contribute to this, but the most important is that multicollinearity induced variance inflation, so often associated with multiple regression on a lot of correlated variables, is being avoided. The reduction from multidimensional covariates to a unidimensional propensity score also makes results easier to summarise and understand. This point might seem trivial, but it recurs frequently in the literature comparing (for example, Rubin, 1997; Oberchain and Melfi, 2000; Perkins et al, 2000) the Propensity Scores approach with multiple regression/econometric/Heckman methods.

At this point it should be said that there are other matching procedures besides stratifying into quintiles or sextiles. More subclasses could be employed, for example by stratifying into deciles. Or bins could be based on equal ranges, rather than frequencies, of propensity scores as in Tu, et al (2000). Leaving stratification entirely, each treatment individual could be matched with the control individual with the closest propensity score value, or matched to a group of "close" control cases. Decisions on what constitutes "close" can be based on "callipers" – pre-selected ranges. Although these procedures sometimes have advantages, most applications of the Propensity Score methodology use "binning" because Cochran (1968) showed that stratification into quintiles usually removes 90% of the bias due to differing covariate distributions between the treatment and control groups. It should also be said that multiple regression, or econometric, modelling can have a refining role, if applied within strata following matching on the propensity score (see Drake and Fisher, 1995, or Rubin, 1997). This is because model misspecifications cause minimal biases if the covariate distributions are similar for the treatment and control groups. While it is probably obvious, the y variables in (9) could be replaced by differences of post- and pre-treatment (and control) values if the earlier measurements exist. This would give a matched or Propensity Score "difference of differences analysis", very analogous to common practice in randomised experiments.[12]

When there are more than two groups comparisons are made pairwise, with separate derivation of propensity scores for each pair. The need for this can be appreciated by extending (following Rubin, 1997) the smoking example to three groups – non-smokers, cigarette smokers and pipe smokers – and two covariates – age and social class. As before a mortality measure is the response variable. Suppose non-smokers and cigarette smokers have the same age distributions, but unequal, though overlapping, social class membership. Suppose non-smokers and pipe smokers have the same distributions by social class, but unequal, though overlapping, age distributions. Then for the non-smokers v cigarette smokers comparison the propensity score matching should "balance" social class differences, while for non-smokers v pipe smokers it should "balance" age differences. Clearly separate derivations of propensity scores are appropriate.

[12] For example, in randomised animal growth experiments, the difference between mean final weights of treatment and control groups is a valid estimate of treatment effect, but the difference in mean weight gains (assuming initial weights were recorded prior to treatment application) is the preferred measure, because it reduces random error.

Matching on the propensity score balances over observed covariates. What if an important covariate – one that has a substantial effect on the response *and* whose distribution differs considerably between treatment and control groups – has not even been observed? If it, w say, is uncorrelated with the observed covariates, matching on the propensity scores can no longer be assumed to have adequately simulated a randomisation situation. However, if it is correlated with them, matching on the propensity scores based on the observed covariates will also, at least partially, balance the unobserved covariate effect. This is because such matching produces treatment and comparison groups with the same distribution of x variables. They would balance the unobserved variable most effectively if w had a well fitting regression on the observed variables with a disturbance uncorrelated with y. If the regression explained enough of the variation in w, they would still balance effectively even if the disturbance was not uncorrelated with y. For this reason the Propensity Score approach stresses searching for and examining the maximum number of attainable covariates. The Heckman correction procedure given by (5) and (6) also assumes w has a linear regression on the observed covariates, though with the additional assumption of bivariate normality (and, some critics would say, without the stress on searching). The normality assumption would be invalidated by the existence of a non-random unrecorded covariate. In reality, both the Propensity Score and the econometric modelling/Heckman correction approaches assume the observed covariates are sufficient to work with, but the difference between the approaches can be very clearly seen by imagining that w, "ability" say, becomes observable. Now the econometric procedure just involves estimation of equation (3) with w added as an extra variable. There is no estimation of a probit or logit participation equation, because the selection bias was presumed to operate only through the formerly unobserved variable. But the Propensity Score method will, of course, retain the treatment assignment equation, just adding w to the variables in it, on the supposition that all variables can contribute to selection bias.

However, if an unobserved, covariate is not well explained by observed covariates and has a substantial effect on y, the "strongly ignorable assumption" – that treatment assignment is effectively random within strata of the propensity score – does not hold and progress would depend on other assumptions. It is possible to employ IV methods within a Propensity Score framework (for example, Angist, 1995) provided truly valid instruments can be identified, but this an issue about which there are considerable disagreements in the recent technical literature. Before returning to this topic it is

desirable to review the measure of agreement that has been reached about the Propensity Score approach in this literature as well as the caveats that have been expressed and the unresolved issues.

IV RECENT DEBATES IN THE LITERATURE

A quick reading of recent literature can give a first impression of substantial, even acrimonious, disagreement between authors with the opposing sides (largely) consisting of econometricians on the one hand and "mainline" statisticians on the other. But at least some of this is due to rival claims about who deserves the credit for the formulation of key concepts and to criticisms of each others' originality and presentation of approaches. In fact, there is more or less general agreement about the desirability of "balancing" covariate distributions through matching and acceptance that, in most if not all cases, propensity scores, or closely related statistics, should be employed. It is now generally appreciated that, whatever about selection bias arising from an unobserved characteristic, considerably larger biases can follow from failure to correct adequately for very different covariate distributions between treatment and control.

Heckman, Ichimura, Smith and Todd (1996) said that their data analysis "appeared to provide a strong endorsement for matching on the propensity score". They stressed that incomparable individuals should not be used in contrasting the treatment and control (in a "binning" context, strata containing individuals from only one group should be discarded), or selection bias would result. But their comment that this "essential" requirement was "hitherto unnoticed" is unfair to many previous authors outside econometrics. The data analysis in Heckman, Ichimura and Todd (1997) decomposed bias into three components: that due to non-overlapping support (the comparing of incomparables), that due to different distributions of covariates within groups and finally that due to selection on unobservables. They said that this last component "called selection bias in econometrics" had been emphasised in the previous econometric literature, but was actually smaller in magnitude than the others and they concluded "Simple balancing of observables . . . goes a long way towards . . . effective evaluation . . ."

This is not to say these authors were uncritical of Propensity Score methodology. They believed that bias due to selection on unobservables, even if it was smaller than the other components, required attention, and this topic will be returned to shortly. They also made other criticisms and comments that are perhaps not of great practical import in most situations. Heckman et al (1996) said that because