MODULE FOUR, PART ONE:

SAMPLE SELECTION IN ECONOMIC EDUCATION RESEARCH

William E. Becker and William H. Greene *

Modules One and Two addressed an economic education empirical study involved with the assessment of student learning that occurs between the start of a program (as measured, for example, by a pretest) and the end of the program (posttest). At least implicitly, there is an assumption that all the students who start the program finish the program. There is also an assumption that those who start the program are representative of, or at least are a random sample of, those for whom an inference is to be made about the outcome of the program. This module addresses how these assumptions might be wrong and how problems of sample selection might occur. The consequences of and remedies for sample selection are presented here in Part One. As in the earlier three modules, contemporary estimation procedures to adjust for sample selection are demonstrated in Parts Two, Three and Four using LIMDEP (NLOGIT), STATA and SAS.

Before addressing the technical issues associated with sample selection problems in an assessment of one or another instructional method, one type of student or teacher versus another, or similar educational comparisons, it might be helpful to consider an analogy involving a contest of skill between two types of contestants: Type A and Type B. There are 8 of each type who compete against each other in the first round of matches. The 8 winners of the first set of matches compete against each other in a second round, and the 4 winners of that round compete in a third. Type A and Type B may compete against their own type in any match after the first round, but one Type A and one Type B manage to make it to the final round. In the final match they tie. Should we conclude, on probabilistic grounds, that Type A and Type B contestants are equally skilled?

---

*William Becker is Professor Emeritus of Economics, Indiana University, Adjunct Professor of Commerce, University of South Australia, Research Fellow, Institute for the Study of Labor (IZA) and Fellow, Center for Economic Studies and Institute for Economic Research (CESifo). William Greene is Toyota Motor Corp. Professor of Economics, Stern School of Business, New York University, Distinguished Adjunct Professor, American University and External Affiliate of the Health Econometrics and Data Group, York University.

How is your answer to the above questions affected if we tell you that on the first round 5 Type As and only 3 Types Bs won their matches and only one Type B was successful in the second and third round? The additional information should make clear that we have to consider how the individual matches are connected and not just look at the last match. But before you conclude that Type As had a superior attribute only in the early contests and not in the finals, consider another analogy provided by Thomas Kane (Becker 2004).

Kane's hypothetical series of races is contested by 8 greyhounds and 8 dachshunds. In the first race, the greyhounds enjoy a clear advantage with 5 greyhounds and only 3 dachshunds finishing among the front-runners. These 8 dogs then move to the second race, when only one dachshund wins. This dachshund survives to the final race when it ties with a greyhound. Kane asks: "Should I conclude that leg length was a disadvantage in the first two races but not in the third?" And answers: "That would be absurd. The little dachshund that made it into the third race and eventually tied for the win most probably had an advantage on other traits—such as a strong heart, or an extraordinary competitive spirit—which were sufficient to overcome the disadvantage created by its short stature."

These analogies demonstrate all three sources of bias associated with attempts to assess performance from the start of a program to its finish: sample selection bias, endogeneity, and omitted variables. The length of the dogs' legs not appearing to be a problem in the final race reflects the sample selection issues resulting if the researcher only looks at that last race. In education research this corresponds to only looking at the performance of those who take the final exam, fill out the end-of-term student evaluations, and similar terminal program measurements. Looking only at the last race (corresponding to those who take the final exam) would be legitimate if the races were independent (previous exam performance had no effect on final exam taking, students could not self select into the treatment group versus control group), but the races (like test scores) are sequentially dependent; thus, there is an endogeneity problem (as introduced in Module Two). As Kane points out, concluding that leg length was important in the first two races and not in the third reveals the omitted-variable problem: a trait such as heart strength or competitive motivation might be overriding short legs and thus should be included as a relevant explanatory variable in the analyses. These problems of sample selection in educational assessment are the focus of this module.

SAMPLE SELECTION FROM PRETEST TO POSTTEST AND HECKMAN CORRECTION

The statistical inference problems associated with sample selection in the typical change-score model used in economic education research can be demonstrated using a modified version of the presentation in Becker and Powers (2001), where the data generating process for the change score (difference between post and pre TUCE scores) for the i^{th} student (Δy_i) is modeled as

$$\Delta y_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \varepsilon_i \qquad (1)$$

The data set of explanatory variables is matrix **X**, where $\mathbf{X}_i$ is the row of x_{ji} values for the relevant variables believed to explain the i^{th} student's pretest and posttest scores, the β_j's are the associated slope coefficients in the vector $\boldsymbol{\beta}$, and ε_i is the individual random shock (caused, for example, by unobservable attributes, events or environmental factors) that affect the i^{th} student's test scores. In empirical work, the exact nature of Δy_i is critical. For instance, to model the truncation issues that might be relevant for extremely able students' being better than the maximum TUCE score, a Tobit model can be specified for Δy_i.[i] Also critical is the assumed starting point on which all subsequent estimation is conditioned.[ii]

As discussed in Module One, to explicitly model the decision to complete a course, as reflected by the existence of a posttest for the i^{th} student, a "yes" or "no" choice probit model can be specified. Let $T_i = 1$, if the i^{th} student takes the posttest and let $T_i = 0$, if not. Assume that there is an unobservable continuous dependent variable, T_i^*, representing the i^{th} student's desire or propensity to complete a course by taking the posttest.

For an initial population of N students, let $\mathbf{T}^*$ be the vector of all students' propensities to take a posttest. Let $\mathbf{H}$ be the matrix of explanatory variables that are believed to drive these propensities, which includes directly observable things (e.g., time of class, instructor's native language). Let $\boldsymbol{\alpha}$ be the vector of slope coefficients corresponding to these observable variables. The individual unobservable random shocks that affect each student's propensity to take the posttest are contained in the error term vector $\boldsymbol{\omega}$. The data generating process for the i^{th} student's propensity to take the posttest can now be written:

$$T_i^* = \mathbf{H}_i\boldsymbol{\alpha} + \omega_i \qquad (2)$$

where

$T_i = 1$, if $T_i^* > 0$, and student i has a posttest score, and

$T_i = 0$, if $T_i^* \leq 0$, and student i does not have a posttest score.

For estimation purposes, the error term ω_i is assumed to be a standard normal random variable that is independently and identically distributed with the other students' error terms in the $\boldsymbol{\omega}$ vector. As shown in Module Four (Parts Two, Three and Four) this probit model for the propensity to take the posttest can be estimated using the maximum-likelihood routines in programs such as LIMDEP, STATA or SAS.

The effect of attrition between the pretest and posttest, as reflected in the absence of a posttest score for the i^{th} student $(T_i = 0)$ and an adjustment for the resulting bias caused by excluding those students from the Δy_i regression can be illustrated with a two-equation model formed by the selection equation (2) and the i^{th} student's change score equation (1).[iii] Each of the disturbances in vector $\boldsymbol{\varepsilon}$, equation (1), is assumed to be distributed bivariate normal with the corresponding disturbance term in the $\boldsymbol{\omega}$ vector of the selection equation (2). Thus, for the i^{th} student we have:

$$(\varepsilon_i, \omega_i) \sim \text{ bivariate normal}\,(0,0,\sigma_\varepsilon,1,\rho) \qquad (3)$$

and for all perturbations in the two-equation system we have:

$$E(\boldsymbol{\varepsilon}) = E(\boldsymbol{\omega}) = 0,\ E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma_\varepsilon^2\mathbf{I},\ E(\boldsymbol{\omega}\boldsymbol{\omega}') = \mathbf{I}, \text{ and } E(\boldsymbol{\varepsilon}\boldsymbol{\omega}') = \rho\sigma_\varepsilon\mathbf{I}\ . \qquad (4)$$

That is, the disturbances have zero means, unit variance, and no covariance among students, but there is covariance between selection in getting a posttest score and the measurement of the change score.

The difference in the functional forms of the posttest selection equation (2) and the change score equation (1) ensures the identification of equation (1) but ideally other restrictions would lend support to identification. Estimates of the parameters in equation (1) are desired, but the i^{th} student's change score Δy_i is observed in the TUCE data for only the subset of students for whom $T_i = 1$. The regression for this censored sample of $n_{T=1}$ students is:

$$E(\Delta y_i \mid \mathbf{X}_i, T_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + E(\varepsilon_i \mid T_i^* > 0);\ i = 1,2,\ldots n_{T=1}\ , \text{ for } n_{T=1} < N\ . \qquad (5)$$

Similar to omitting a relevant variable from a regression (as discussed in Module Two), selection bias is a problem because the magnitude of $E(\varepsilon_i \mid T_i^* > 0)$ varies across individuals and yet is not included in the estimation of equation (1). To the extent that ε_i and ω_i (and thus T_i^*) are related, the estimators are biased.

The change score regression (1) can be adjusted for those who elected not to take a posttest in several ways. An early Heckman-type solution to the sample selection problem is to rewrite the omitted variable component of the regression so that the equation to be estimated is:

$$E(\Delta y_i \mid \mathbf{X}_i, T_i = 1) = \mathbf{X}_i\boldsymbol{\beta} + (\rho\sigma_\varepsilon)\lambda_i;\ i = 1,2,\ldots n_{T=1} \qquad (6)$$

where $\lambda_i = f(-T_i^*)/[1 - F(-T_i^*)]$, and $f(.)$ and $F(.)$ are the normal density and distribution functions. The inverse Mill's ratio (or hazard) λ_i is the standardized mean of the disturbance term ω_i, for the i^{th} student who took the posttest; it is close to zero only for those well above the

$T = 1$ threshold. The values of λ are generated from the estimated probit selection equation (2) for all students. Each student in the change score regression (Δy_i) gets a calculated value λ_i, with the vector of these values serving as a shift variable in the persistence regression.

The single coefficient represented by the product of ρ and σ_ε (ie., $\rho\sigma_\varepsilon$) is estimated in a two-step procedure in which the probit selection equation (2) is first estimated by maximum likelihood and then the change-score equation (1) is estimated by least squares with the inverse mills ratio used as an additional regressor to adjust for the selection bias. The estimates of ρ, σ_ε, and all the other coefficients in equations (1) and (2) can also be obtained simultaneously and more efficiently using the maximum-likelihood routines in LIMDEP, STATA or SAS, as will be demonstrated in Parts Two, Three and Four of this module using the Becker and Powers data set.

The Heckman-type selection model represented by equations (1) and (2) highlights the nature of the sample selection problem inherent in estimating a change-score model by itself. Selection results in population error term and regressor correlation that biases and makes the coefficient estimators in the change score model inconsistent. The early Heckman (1979) type two-equation estimation of the parameters in a selection model and change-score model, however, requires cross-equation exclusion restrictions (variables that affect selection but not the change score), differences in functional forms, and/or distributional assumptions for the error terms. Parameter estimates are typically sensitive to these model specifications.

ALTERNATIVE METHODS FOR ADDRESSING SELECTION

As reviewed in Imbens and Wooldridge (2009), alternative nonparametric and semiparametric methods are being explored for assessing treatment effects in nonrandomized experiments but these methods have been slow to catch on in education research in general and economic education in particular. Exceptions, in the case of financial aid and the enrollment decision, are the works of Wilbert van der Klaauw and Thomas Kane. Van der Klaauw (2002) estimates the effect of financial aid on the enrollment decision of students admitted to a specific East Coast college, recognizing that this college's financial aid is endogenous because competing offers are unknown and thus by definition are omitted relevant explanatory variables in the enrollment decision of students considering this college.

The college investigated by van der Klaauw created a single continuous index of each student's initial financial aid potential (based on a SAT score and high school GPA) and then classified students into one of four aid level categories based on discrete cut points. The aid assignment rule depends at least in part on the value of a continuous variable relative to a given threshold in such a way that the corresponding probability of receiving aid (and the mean amount offered) is a discontinuous function of this continuous variable at the threshold cut point. A sample of

individual students close to a cut point on either side can be treated as a random sample at the cut point because on average there really should be little difference between them (in terms of financial aid offers received from other colleges and other unknown variables). In the absence of the financial aid level under consideration, we should expect little difference in the college-going decision of those just above and just below the cut point. Similarly, if they were all given the financial aid, we should see little difference in outcomes, on average. To the extent that some actually get it and others do not, we have an interpretable treatment effect. (Intuitively, this can be thought of as running a regression of enrollment on financial aid for those close to the cut point, with an adjustment for being in that position.) In his empirical work, van der Klaauw obtained credible estimates of the importance of the financial aid effect without having to rely on arbitrary cross-equation exclusion restrictions and functional form assumptions.

Kane (2003) uses an identification strategy similar to van der Klaauw but does so for all those who applied for the Cal Grant Program to attend any college in California. Eligibility for the Cal Grant Program is subject to a minimum GPA and maximum family income and asset level. Like van der Klaauw, Kane exploits discontinuities on one dimension of eligibility for those who satisfy the other dimensions of eligibility.

Although some education researchers are trying to fit their selection problems into this regression discontinuity framework, legitimate applications are few because the technique has very stringent data requirement (an actual but unknown or conceptual defendable continuous index with thresholds for rank-ordered classifications) and limited ability to generalize away from the classification cut points. Much of economic education research, on the other hand, deals with the assessment of one type of program or environment versus another, in which the source of selection bias is entry and exit from the control or experimental groups. An alternative to Heckman's parametric (rigid equation form) manner of comparing outcome measures adjusted for selection based on unobservables is propensity score matching.

PROPENSITY SCORE MATCHING

Propensity score matching as a body of methods is based on the following logic: We are interested in evaluating a change score after a treatment. Let O now denote the outcome variable or interest (e.g., posttest score, change score, persistence, or whatever) and T denote the treatment dummy variable (e.g., took the enhanced course), such that $T = 1$ for an individual who has experienced the "treatment," and $T = 0$ for one who has not. If we are interested in the change-score effect of treatment on the treated, the conceptual experiment would amount to observing the treated individual (1) after he or she experienced the treatment and the same individual in the same situation but (2) after he/she did not experience the treatment (but presumably, others did). The treatment effect would be the difference between the two post-test scores (because the pretest would be the one achieved by this individual). The problem, of

course, is that ex post, we don't observe the outcome variable, *O*, for the treated individual, in the absence of the treatment. We observe some individuals who were treated and other individuals who were not. Propensity score matching is a largely nonparametric approach to evaluating treatment effects with this consideration in mind.[iv]

If individuals who experienced the treatment were exactly like those who did not in all other respects, we could proceed by comparing random samples of treated and nontreated individuals, confident that any observed differences could be attributed to the treatment. The first section of this module focused on the problem that treated individuals might differ from untreated individuals systematically, but in ways that are not directly observable by the econometrician. To consider an example, if the decision to take an economics course (the treatment) were motivated by characteristics of individuals (curiosity, ambition, etc.) that were also influential in their performance on the outcome (test), then our analysis might attribute the change in the score to the treatment rather than to these characteristics. Models of sample selection considered previously are directed at this possibility. The development in this section is focused on the possibility that the same kinds of issues might arise, but the underlying features that differentiate the treated from the untreated can be observed, at least in part.

If assignment to the treatment were perfectly random, as discussed in the introduction to this module, solving this problem would be straightforward. A large enough sample of individuals would allow us to average away the differences between treated and untreated individuals, both in terms of observable characteristics and unobservable attributes. Regression methods, such as those discussed in the previous sections of this module, are designed to deal with the difficult case in which the assignment is nonrandom with respect to the unobservable characteristics of individuals (such as ability, motivation, etc.) that can be related to the "treatment assignment," that is, whether or not they receive the treatment. Those methods do not address another question, that is, whether there are systematic, observable differences between treated and nontreated individuals. Propensity score methods are used to address this problem.

To take a simple example, suppose it is known with certainty that the underlying, unobservable characteristics that are affecting the change score are perfectly randomly distributed across individuals, treated and untreated. Assume, as well, that it is known for certain that the only systematic, observable difference between treated and untreated individuals is that women are more likely to undertake the treatment than men. It would make sense, then, that if we want to compare treated to untreated individuals, we would not want to compare a randomly selected group of treated individuals to a randomly selected group of untreated individuals – the former would surely contain more women than the latter. Rather, we would try to balance the samples so that we compared a group of women to another group of women and a group of men to another group of men, thereby controlling for the impact of gender on the likelihood of receiving the treatment. We might then want to develop an overall average by averaging, once again, this time the two differences, one for men, the other for women.

In the main, and as already made clear in our consideration of the Heckman adjustment, if assignment to the treatment is nonrandom, then estimation of treatment effects will be biased by the effect of the variables that effect the treatment assignment. The strategy is, essentially, to locate an untreated individual who looks like the treated one in every respect except the treatment, then compare the outcomes. We then average this across individual pairs to estimate the "average treatment effect on the treated." The practical difficulty is that individuals differ in many characteristics, and it is not feasible, in a realistic application, to compare each treated observation to an untreated one that "looks like it." There are too many dimensions on which individuals can differ. The technique of propensity score matching is intended to deal with this complication. Keep in mind, however, if unmeasured or unobserved attributes are important, and they are not randomly distributed across treatment and control groups, matching techniques may not work. That is for what the methods in the previous sections were designed.

THE PROPENSITY SCORE MATCHING METHOD

We now provide some technical details on propensity score matching. Let $\mathbf{x}$ denote a vector of observable characteristics of the individual, before the treatment. Let the probability of treatment be denoted $P(T=1|\mathbf{x}) = P(\mathbf{x})$. Because T is binary, $P(\mathbf{x}) = E[T|\mathbf{x}]$, as in a linear probability model. If treatment is random *given* $\mathbf{x}$, then treatment is random given $P(\mathbf{x})$, which in this context is called the *propensity score*. It will generally not be possible to match individuals based on all the characteristics individually – with continuously measured characteristics, such as income. There are too many cells. The matching is done via the propensity score. Individuals with similar propensity scores are expected (on average) to be individuals with similar characteristics.

Overall, for a 'treated' individual with propensity $P(\mathbf{x}_i)$ and outcome O_i, the strategy is to locate a control observation with similar propensity $P(\mathbf{x}_c)$ and with outcome O_c. The effect of treatment on the treated for this individual is estimated by $O_i - O_c$. This is averaged across individuals to estimate the average treatment effect on the treated. The underlying theory asserts that the estimates of treatment effects across treated and controls are unbiased if the treatment assignment is random among individuals with the same propensity score; the propensity score, itself, captures the drivers of the treatment assignment. (Relevant papers that establish this methodology are too numerous to list here. Useful references are four canonical papers, Heckman et al. [1997, 1998a, 1998b, 1999] and a study by Becker and Ichino [2002].)

The steps in the propensity score matching analysis consist of the following:

> Step 1. Estimate the propensity score function, $P(\mathbf{x})$, for each individual by fitting a probit or logit model, and using the fitted probabilities.

Step 2. Establish that the average propensity scores of treated and control observations are the same within particular ranges of the propensity scores. (This is a test of the "balancing hypothesis.")

Step 3. Establish that the averages of the characteristics for treatment and controls are the same for observations in specific ranges of the propensity score. This is a check on whether the propensity score approach appears to be succeeding at matching individuals with similar characteristics by matching them on their propensity scores.

Step 4. For each treated observation in the sample, locate a similar control observation(s) based on the propensity scores. Compute the treatment effect, $O_i - O_c$. Average this across observations to get the average treatment effect.

Step 5. In order to estimate a standard error for this estimate, Step 4 is repeated with a set of bootstrapped samples.

THE PROPENSITY SCORE

We use a binary choice model to predict "participation" in the treatment. Thus,

$$\text{Prob}(T = 1|\mathbf{x}) = F\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K\right) = F(\boldsymbol{\beta}'\mathbf{x}) .$$

The choice of F is up to the analyst. The logit model is a common choice;

$$\text{Prob}(T=1|\mathbf{x}) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x})}{1+\exp(\boldsymbol{\beta}'\mathbf{x})} .$$

The probit model, $F(\boldsymbol{\beta}'\mathbf{x}) = \Phi(\boldsymbol{\beta}'\mathbf{x})$, where Φ(t) is the normal distribution function, is an alternative. The propensity score is the fitted probability from the probit or logit model,

$$\text{Propensity Score for individual } i = F\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right) = P_i .$$

The central feature of this step is to find similar individuals by finding individuals who have similar propensity scores. Before proceeding, we note, the original objective is to find groups of individuals who have the same **x**. This is easy to do in our simple example, where the only variable in **x** is gender, so we can simply distinguish people by their gender. When the **x** vector has many variables, it is impossible to partition the data set into groups of individuals with the same, or even similar explanatory variables. In the example we will develop below, **x** includes age (and age squared), education, marital status, race, income and unemployment status. The working principle in this procedure is that individuals who have similar propensity scores will, if

we average enough of them, have largely similar characteristics. (The reverse must be true, of course.) Thus, although we cannot group people by their characteristics, **x**s, we can (we hope) achieve the same end by grouping people by their propensity scores. That leads to step 2 of the matching procedure.

GROUPING INDIVIDUALS BY PROPENSITY SCORES

Grouping those with similar propensity scores should result in similar predicted probabilities for treatment and control groups. For instance, suppose we take a range of propensity scores (probabilities of participating in the treatement), say from 0.4 to 0.6. Then, the part of the sample that contains propensity scores in this range should contain a mix of treated individuals (individuals with $T = 1$) and controls (individuals with $T = 0$). If the theory we are relying on is correct, then the average propensity score for treated and controls should be the same, at least approximately. That is,

$$\textit{Average } F\left(\hat{\boldsymbol{\beta}}'\mathbf{x}\right)\Big|\left(T=1 \text{ and } \hat{F} \text{ in the range}\right) \approx \textit{Average } F\left(\hat{\boldsymbol{\beta}}'\mathbf{x}\right)\Big|\left(T=0 \text{ and } \hat{F} \text{ in the range}\right).$$

We will look for a partitioning of the range of propensity scores for which this is the case in each range.

A first step is to decide if it is necessary to restrict the sample to the range of values of propensity scores that is shared by the treated and control observations. That range is called the common support. Thus, if the propensity scores of the treated individuals range from 0.1 to 0.7 and the scores of the control observations range from 0.2 to 0.9, then the common support is from 0.2 to 0.7. Observations that have scores outside this range would not be used in the analysis.

Once the sample to be used is determined, we will partition the range of propensity scores into K cells. For each partitioning of the range of propensity scores considered, we will use a standard F test for equality of means of the propensity scores of the treatment and control observations:

$$F_k[1,d] = \frac{\left(\overline{P}_C^k - \overline{P}_T^k\right)^2}{\left(S_{C,k}^2 / N_C^k + S_{T,k}^2 / N_T^k\right)}, k = 1,...,K.$$

The denominator degrees of freedom for F are approximated using a technique invented by Satterthwaite (1946):

$$d = w\frac{(N_C - 1)}{S^2_{C,k} / N^k_C} + (1-w)\frac{(N_T - 1)}{S^2_{T,k} / N^k_T}$$

$$w = \frac{(N_T - 1)\left(S^2_{C,k} / N^k_C\right)^2}{(N_T - 1)\left(S^2_{C,k} / N^k_C\right)^2 + (N_C - 1)\left(S^2_{T,k} / N^k_T\right)^2}.$$

If any of the cells (ranges of scores) fails this test, the next step is to increase the number of cells. There are various strategies by which this can be done. The natural approach would be to leave cells that pass the test as they are, and partition more finely the ones that do not. This may take several attempts. In our example, we started by separating the range into 5 parts. With 5 segments, however, the data do not appear to satisfy the balancing requirement. We then try 6 and, finally, 7 segments of the range of propensity scores. With the range divided into 7 segments, it appears that the balance requirement is met.

Analysis can proceed even if the partitioning of the range of scores does not pass this test. However, the test at this step will help to give an indication of whether the model used to calculate the propensity scores is sufficiently specified. A persistent failure of the balancing test might signal problems with the model that is being used to create the propensity scores. The result of this step is a partitioning of the range of propensity scores into K cells with the $K + 1$ values,

$$[P^*] = [P_1, P_2, ..., P_{K+1}]$$

which is used in the succeeding steps.

EXAMINING THE CHARACTERISTICS IN THE SAMPLE GROUPS

Step 3 returns to the original motivation of the methodology. At step 3, we examine the characteristics (**x** vectors) of the individuals in the treatement and control groups within the subsamples defined by the groupings made by Step 2. If our theory of propensity scores is working, it should be the case that within a group, for example, for the individuals whose propensity scores are in the range 0.4 to 0.6, the **x** vectors should be similar in that at least the means should be very close. This aspect of the data is examined statistically. Analysis can proceed if this property is not met but the result(s) of these tests might signal to the analyst that their results are a bit fragile. In our example below, there are seven cells in the grid of propensity scores and 12 variables in the model. We find that for four of the 12 variables in one of the 7 cells (i.e., in four cases out of 84), the means of the treated and control observations appear to be significantly different. Overall, this difference does not appear to be too severe, so we proceed in spite of it.

MATCHING

Assuming that the data have passed the scrutiny in step 3, we now match the observations. For each treated observation (individual's outcome measure such as a test score) in the sample, we find a control observation that is similar to it. The intricate complication at this step is to define "similar." It will generally not be possible to find a treated observation and a control observation with exactly the same propensity score. So, at this stage it is necessary to decide what rule to use for "close." The obvious choice would be the nearest neighbor in the set of observations that is in the propensity score group. The nearest neighbor for observation O_i would be the O_c* for which $|P_i - P_c|$ is minimized. We note, by this strategy, a particular control observation might be the nearest neighbor for more than one treatment observation and some control observations might not be the nearest neighbor to any treated observation.

Another strategy is to use the average of several nearby observations. The counterpart observation is constructed by averaging all control observations whose propensity scores fall in a given range in the neighborhood of P_i. Thus, we first locate the set [C_t*] = the set of control observations for which $|P_t - P_c| < r$, for a chosen value of r called the caliper. We then average O_c for these observations. By this construction, the neighbor may be an average of several control observations. It may also not exist, if no observations are close enough. In this case, r must be increased. As in the single nearest neighbor computation, control observations may be used more than once, or they might not be used at all (e.g., if the caliper is $r = .01$, and a control observation has propensity .5 and the nearest treated observations have propensities of .45 and .55, then this control will never be used).

A third strategy for finding the counterpart observations is to use kernel methods to average all of the observations in the range of scores that contains the O_i that we are trying to match. The averaging function is computed as follows:

$$\bar{O}_c = \sum_{control\ observations\ in\ the\ cell} w_c O_c$$

$$w_c = \frac{\frac{1}{h} K\left[\frac{P_i - P_c}{h}\right]}{\sum_{control\ observations\ in\ the\ cell} \frac{1}{h} K\left[\frac{P_i - P_c}{h}\right]}$$

The function K[.] is a weighting function that takes its largest value when P_i equals P_c and tapers off to zero as P_c is farther from P_i. Typical choices for the kernel function are the normal or logistic density functions. A common choice that cuts off the computation at a specific point is the Epanechnikov (1969) weighting function,

$K[t] = 0.75(1 - .2t^2)/5^{1/2}$ for $|t| < 5$, and 0 otherwise.

The parameter *h* is the bandwidth that controls the weights given to points that lie relatively far from P_i. A larger bandwidth gives more distant points relatively greater weight. Choice of the bandwidth is a bit of an (arcane) art. The value 0.06 is a reasonable choice for the types of data we are using in our analysis here.

Once treatment observations, O_i and control observations, O_c are matched, the treatment effect for this pair is computed as $O_i - O_c$. The average treatment effect (ATE) is then estimated by the mean,

$$\hat{ATE} = \frac{1}{N_{match}} \sum_{i=1}^{N_{match}} (O_i - O_c)$$

STATISTICAL INFERENCE

In order to form a confidence interval around the estimated average treatment effect, it is necessary to obtain an estimated standard error. This is done by reconstructing the entire sample used in Steps 2 through 4 *R* times, using bootstrapping. By this method, we sample *N* observations from the sample of *N* observations *with replacement.* Then *ATE* is computed *R* times and the estimated standard error is the empirical standard deviation of the *R* observations. This can be used to form a confidence interval for the *ATE*.

The end result of the computations will be a confidence interval for the expected treatment effect on the treated individuals in the sample. For example, in the application that we will present in Part 2 of this module, in which the outcome variable is the log of earnings and the treatment is the *National Supported Work Demonstration* – see LaLonde (1986) – the following is the set of final results:

```
+------------------------------------------------------------------------+
| Number of Treated observations =    185  Number of controls =   1157 |
| Estimated Average Treatment Effect   =        .156255                |
| Estimated Asymptotic Standard Error  =        .104204                |
| t statistic (ATT/Est.S.E.)           =       1.499510                |
| Confidence Interval for ATT = (     -.047985  to        .360496) 95% |
| Average Bootstrap estimate of ATT    =        .144897                |
| ATT - Average bootstrap estimate     =        .011358                |
+------------------------------------------------------------------------+
```

The overall estimate from the analysis is ATE = 0.156255, which suggests that the effect on earnings that can be attributed to participation in the program is 15.6%. Based on the (25) bootstrap replications, we obtained an estimated standard error of 0.104204. By forming a confidence interval using this standard error, we obtain our interval estimate of the impact of the

program of (-4.80% to +36.05%). We would attribute the negative range to an unconstrained estimate of the sampling variability of the estimator, not actually to a negative impact of the program.[v]

CONCLUDING COMMENTS

The genius in James Heckman was recognizing that sample selection problems are not necessarily removed by bigger samples because unobservables will continue to bias estimators. His parametric solution to the sample selection problem has not been lessened by newer semi-parametric techniques. It is true that results obtained from the two equation system advanced by Heckman over 30 years ago are sensitive to the correctness of the equations and their identification. Newer methods such as regression discontinuity, however, are extremely limited in their applications. As we will see in Module Four, Parts Two, Three and Four, methods such as the propensity score matching depend on the validity of the logit or probit functions estimated along with the methods of getting smoothness in the kernel density estimator. One of the beauties of Heckman's original selection adjustment method is that its results can be easily replicated in LIMDEP, STATA and SAS. Such is not the case with the more recent nonparametric and semi-parametric methods for addressing sample selection problems.

REFERENCES

Becker, William E. "Omitted Variables and Sample Selection Problems in Studies of College-Going Decisions," *Public Policy and College Access: Investigating the Federal and State Role in Equalizing Postsecondary Opportunity*, Edward St. John (ed), 19. NY: AMS Press. 2004: 65-86.

_____ "Economics for a Higher Education," *International Review of Economics Education*, 3, 1, 2004: 52-62.

_____ "Quit Lying and Address the Controversies: There Are No Dogmata, Laws, Rules or Standards in the Science of Economics," *American Economist*, 50, Spring 2008: 3-14.

_____ and William Walstad. "Data Loss from Pretest to Posttest as a Sample Selection Problem," *Review of Economics and Statistics*, 72, February 1990: 184-188.

_____ and John Powers. "Student Performance, Attrition, and Class Size Given Missing Student Data," *Economics of Education Review*, 20, August 2001: 377-388.

Becker, S. and A. Ichino. "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, 2, 2002: 358-377.

Deheija, R. and S. Wahba "Causal Effects in Nonexperimental Studies: Reevaluation of the Evaluation of Training Programs," Journal of the American Statistical Association, 94, 1999, pp. 1052-1062.

Epanechnikov, V. "Nonparametric Estimates of a Multivariate Probability Density," *Theory of Probability and its Applications*, 14, 1969: 153-158.

Greene, William. H. "A statistical model for credit scoring." Department of Economics, Stern School of Business, New York University, (September 29, 1992).

Heckman, James. "Sample Bias as a Specification Error," *Econometrica*, 47, 1979: 153-162.

Heckman, J., H. Ichimura, J. Smith and P. Todd. "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 5, 1998a: 1017-1098.

Heckman, J., H. Ichimura and P. Todd. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies*, 64, 4,1997: 605-654.

Heckman, J., H. Ichimura and P. Todd. "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 2, 1998b: 261-294.

Heckman, J., R. LaLonde, and J. Smith. 'The Economics and Econometrics of Active Labour Market Programmes,' in Ashenfelter, O. and D. Card (eds.) The Handbook of Labor Economics, Vol. 3, North Holland, Amsterdam, 1999.

Krueger, Alan B. and Molly F. McIntosh. "Using a Web-Based Questionnaire as an Aide for High School Economics Instruction," *Journal of Economic Education*, 39, Spring, 2008: 174-197.

Huynh, Kim, David Jacho-Chavez, and James K. Self. "The Efficacy of Collaborative Learning Recitation Sessions on Student Outcomes?" *American Economic Review*, (Forthcoming May 2010).

Imbens, Guido W. and Jeffrey M. Wooldridge. "Recent Developments in Econometrics of Program Evaluation," *Journal of Economic Literature*, March, 2009: 5-86.

Kane, Thomas. "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going." NBER Working Paper No. W9703, May, 2003.

LaLonde, R., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76, 4, 1986: 604-620.

Satterthwaite, F. E. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin*, 2: 1946: 110–114.

van der Klaauw, W. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discounting Approach." *International Economic Review*, November, 2002: 1249-1288.

ENDNOTES

[i]. The opportunistic samples employed in the older versions of the TUCE as well as the new TUCE 4 have few observations from highly selective schools. The TUCE 4 is especially noteworthy because it has only one such prestige school: Stanford University, where the class was taught by a non-tenure track teacher. Thus, the TUCE 4 might reflect what those in the sample are taught and are able to do, but it does not reflect what those in the know are teaching or what highly able students are able to do. For example Alan Krueger (Princeton University) is listed as a member of the TUCE 4 "national panel of distinguished economists;" yet, in a 2008 *Journal of Economic Education* article he writes: "a long standing complaint of mine, as well as others, for example Becker 2007 and Becker 2004, is that introductory economics courses have not kept up with the economics profession's expanding emphasis on data and empirical analysis." Whether bright and motivated students at the leading institutions of higher education can be expected to get all or close to all 33 multiple-choice questions correct on either the micro or macro parts of the TUCE (because they figure out what the test designers want for an answer) or score poorly (because they know more than what the multiple-choice questions assume) is open to question and empirical testing. What is not debatable is that the TUCE 4 is based on a censored sample that excludes those at and exposed to thinking at the forefront of the science of economics.

[ii]. Because Becker and Powers (2002) do not have any data before the start of the course, they condition on those who are already in the course and only adjust their change-score model estimation for attrition between the pretest and posttest. More recently, Huynh, Jacho-Chavez, and Self (2010) account for selection into, out of and between collaborative learning sections of a large principles course in their change-score modeling.

[iii]. Although Δy_i is treated as a continuous variable this is not essential. For example, a bivariate choice (probit or logit) model can be specified to explicitly model the taking of a posttest decision as a "yes" or "no" for students who enrolled in the course. The selection issue is then modeled in a way similar to that employed by Greene (1992) on consumer loan default and credit card expenditures. As with the standard Heckman selection model, this two-equation system involving bivariate choice and selection can be estimated in a program like LIMDEP.

[iv]. The procedure is not "parametric" in that it is not fully based on a parametric model. It is not "nonparametric" in that it does employ a particular binary choice model to describe participation, or receiving the treatment. But the binary choice model functions as an aggregator of a vector of variables into a single score, not necessarily as a behavioral relationship. Perhaps "partially parametric" would be appropriate here, but we have not seen this term used elsewhere.

[v] The example mentioned at several points in this discussion will be presented in much greater detail in Part 2. The data will be analyzed with LIMDEP, Stata and SAS. We note at this point, there are some issues with duplication of the results with the three programs and with the studies done by the original authors. Some of these are numerical and specifically explainable. However, we do not anticipate that results in Step 5 can be replicated across platforms. The reason is that Step 5 requires generation of random numbers to draw the bootstrap samples. The pseudorandom number generators used by different programs vary substantially, and these differences show up in, for example, in bootstrap replications. If the samples involved are large enough, this sort of random variation (chatter) gets averaged out in the results. The sample in our real world application is not large enough to expect that this chatter will be completely averaged out. As such, as will be evident later, there will be some small variation across programs in the results that one obtains with our or any other small or moderately sized data set.