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.

MODULE FOUR, PART TWO: SAMPLE SELECTION

IN ECONOMIC EDUCATION RESEARCH USING LIMDEP (NLOGIT)

Part Two of Module Four provides a cookbook-type demonstration of the steps required to use LIMDEP (NLOGIT) in situations involving estimation problems associated with sample selection. Users of this model need to have completed Module One, Parts One and Two, but not necessarily Modules Two and Three. From Module One users are assumed to know how to get data into LIMDEP, recode and create variables within LIMDEP, and run and interpret regression results. Module Four, Parts Three and Four demonstrate in STATA and SAS what is done here in LIMDEP.

THE CASE, DATA, AND ROUTINE FOR EARLY HECKMAN ADJUSTMENT

The change score or difference in difference model is used extensively in education research. Yet, before Becker and Walstad (1990), little if any attention was given to the consequence of missing student records that result from: 1) "data cleaning" done by those collecting the data, 2) student unwillingness to provide data, or 3) students self-selecting into or out of the study. The implications of these types of sample selection are shown in the work of Becker and Powers (2001) where the relationship between class size and student learning was explored using the third edition of the Test of Understanding in College Economics (TUCE), which was produced by Saunders (1994) for the National Council on Economic Education (NCEE), since renamed the Council for Economic Education.

Module One, Part Two showed how to get the Becker and Powers data set "beck8WO.csv" into LIMDEP (NLOGIT). As a brief review this was done with the read command:

```
READ; NREC=2837; NVAR=64; FILE=k:\beck8WO.csv; Names=
A1,A2,X3, C,AL,AM,AN,CA,CB,CC,CH,CI,CJ,CK,CL,CM,CN,CO,CS,CT,
CU,CV,CW,DB,DD,DI,DJ,DK,DL,DM,DN,DQ,DR,DS,DY,DZ,EA,EB,EE,EF,
EI,EJ,EP,EQ,ER,ET,EY,EZ,FF,FN,FX,FY,FZ,GE,GH,GM,GN,GQ,GR,HB,
HC,HD,HE,HF $
```

where

A1: term, where 1= fall, 2 = spring
A2: school code, where 100/199 = doctorate,
200/299 = comprehensive,
300/399 = lib arts,
400/499 = 2 year
hb: initial class size (number taking preTUCE)

hc: final class size (number taking postTUCE)
dm: experience, as measured by number of years teaching
dj: teacher's highest degree, where Bachelors=1, Masters=2, PhD=3
cc: postTUCE score (0 to 30)
an: preTUCE score (0 to 30)
ge: Student evaluation measured interest
gh: Student evaluation measured textbook quality
gm: Student evaluation measured regular instructor's English ability
gq: Student evaluation measured overall teaching effectiveness
ci: Instructor sex (Male = 1, Female = 2)
ck: English is native language of instructor (Yes = 1, No = 0)
cs: PostTUCE score counts toward course grade (Yes = 1, No = 0)
ff: GPA*100
fn: Student had high school economics (Yes = 1, No = 0)
ey: Student's sex (Male = 1, Female = 2)
fx: Student working in a job (Yes = 1, No = 0)

In Module One, Part Two the procedure for changing the size of the work space in earlier versions of LIMDEP and NLOGIT was shown but that is no longer required for the 9th version of LIMDEP and the 4th version of NLOGIT. Starting with LIMDEP version 9 and NLOGIT version 4 the required work space is automatically determined by the "Read" command and increased as needed with subsequent "Create" commands.

Separate dummy variables need to be created for each type of school (A2), which is done with the following code:

```
recode; a2; 100/199 = 1; 200/299 = 2; 300/399 = 3; 400/499 =4$
create; doc=a2=1; comp=a2=2; lib=a2=3; twoyr=a2=4$
```

To create a dummy variable for whether the instructor had a PhD we use

```
Create; phd=dj=3$
```

To create a dummy variable for whether the student took the postTUCE we use

```
final=cc>0;
```

To create a dummy variable for whether a student did (noeval = 0) or did not (noeval = 1) complete a student evaluation of the instructor we use

```
Create evalsum=ge+gh+gm+gq; noeval=evalsum=-36$
```

"Noeval" reflects whether the student was around toward the end of the term, attending classes, and sufficiently motivated to complete an evaluation of the instructor. In the Saunder's data set evaluation questions with no answer where coded -9; thus, these four questions summing to -36 indicates that no questions were answered.

And the change score is created with

```
Create; change=cc-an$
```

Finally, there was a correction for the term in which student record 2216 was incorrectly recorded:

```
recode; hb; 90=89$
```

All of these recoding and create commands are entered into LIMDEP command file as follows:

```
recode; a2; 100/199 = 1; 200/299 = 2; 300/399 = 3; 400/499 =4$
create; doc=a2=1; comp=a2=2; lib=a2=3; twoyr=a2=4; phd=dj=3;final=cc>0;
evalsum=ge+gh+gm+gq; noeval=evalsum=-36$
Create; change=cc-an$
recode; hb; 90=89$ #2216 counted in term 2, but in term 1 with no posttest
```

To remove records with missing data the following is entered:

```
Reject; AN=-9$
```

```
Reject; HB=-9$
Reject; ci=-9$
Reject; ck=-9$
Reject; cs=0$
Reject; cs=-9$
Reject; a2=-9$
Reject; phd=-9$
```

The use of these data entry and management commands will appear in the LIMDEP (NLOGIT) output file for the equations to be estimated in the next section.

THE PROPENSITY TO TAKE THE POSTTEST AND THE CHANGE SCORE EQUATION

To address attrition-type sample selection problems in change score studies, Becker and Powers first add observations that were dropped during the early stage of assembling data for TUCE III. Becker and Powers do not have any data on students before they enrolled in the course and thus cannot address selection into the course, but to examine the effects of attrition (course withdrawal) they introduce three measures of class size (beginning, ending, and average) and argue that initial or beginning class size is the critical measure for assessing learning over the entire length of the course.[i] To show the effects of initial class size on attrition (as discussed in Module Four, Part One) they employ what is now the simplest and most restrictive of sample correction methods, which can be traced to James Heckman (1979), recipient of the 2000 Nobel Prize in Economics.

From Module Four, Part One, we have the data generating process for the difference between post and preTUCE scores for the i^{th} student (Δy_i):

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

where the data set of explanatory variables is matrix $\mathbf{X}$, 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. Sample selection associated with students' unwillingness to take the postteest (dropping the course) results in population error term and regressor correlation that biases and makes coefficient estimators in this change score model inconsistent.

The data generating process for the i^{th} student's propensity to take the posttest is:

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

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.

$\mathbf{T}^*$ is the vector of all students' propensities to take a posttest.

$\mathbf{H}$ is the matrix of explanatory variables that are believed to drive these propensities.

$\boldsymbol{\alpha}$ is the vector of slope coefficients corresponding to these observable variables.

$\boldsymbol{\omega}$ is the vector of unobservable random shocks that affect each student's propensity.

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 a Heckman adjustment for the resulting bias caused by excluding those students from the change-score regression requires estimation of equation (2) and the calculation of an inverse Mill's ratio for each student who has a pretest. This inverse Mill's ratio is then added to the change-score regression (1) as another explanatory variable. In essence, this inverse Mill's ratio adjusts the error term for the missing students.

For the Heckman adjustment for sample selection each disturbance 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:

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

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

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

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 regression for this censored sample of $n_{T=1}$ students who took the posttest is now:

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

which suggests the Heckman adjusted regression to be estimated:

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

where λ_i is the inverse Mill's ratio (or hazard) such that $\lambda_i = f(-T_i^*)/[1-F(-T_i^*)]$, and $f(.)$ and $F(.)$ are the normal density and distribution functions. 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.

The probit command for the selection equation to be estimated in LIMDEP (NLOGIT) is

```
probit;lhs=final;rhs=one,an,hb,doc,comp,lib,ci,ck,phd,noeval;hold results$
```

where the "hold results" extension tells LIMDEP to hold the results for the change equation to be estimated by least squares with the inverse Mill's ratio used as regressor.

The command for estimating the adjusted change equation using both the inverse Mills ratio as a regressor and maximum likelihood estimation of the ρ and σ_ε is written

```
selection;lhs=change;rhs=one,hb,doc,comp,lib,ci,ck,phd,noeval;mle$
```

where the extension "mle" tells LIMDEP (NLOGIT) to use maximum likelihood estimation.

As described in Module One, Part Two, entering all of these commands into the command file in LIMDEP (NLOGIT), highlighting the bunch and pressing the GO button yields the following output file:

```
Initializing NLOGIT Version 4.0.7
--> READ; NREC=2837; NVAR=64; FILE=k:\beck8WO.csv; Names=
    A1,A2,X3, C,AL,AM,AN,CA,CB,CC,CH,CI,CJ,CK,CL,CM,CN,CO,CS,CT,
    CU,CV,CW,DB,DD,DI,DJ,DK,DL,DM,DN,DQ,DR,DS,DY,DZ,EA,EB,EE,EF,
    EI,EJ,EP,EQ,ER,ET,EY,EZ,FF,FN,FX,FY,FZ,GE,GH,GM,GN,GQ,GR,HB,
    HC,HD,HE,HF $
--> recode; a2; 100/199 = 1; 200/299 = 2; 300/399 = 3; 400/499 =4$
--> recode; hb; 90=89$ #2216 counted in term 2, but in term 1 with no posttest
--> create; doc=a2=1; comp=a2=2; lib=a2=3; twoyr=a2=4; phd=dj=3; final=cc>0;
    evalsum=ge+gh+gm+gq; noeval=evalsum=-36$
--> Create; change=cc-an$
--> Reject; AN=-9$
--> Reject; HB=-9$
--> Reject; ci=-9$
--> Reject; ck=-9$
--> Reject; cs=0$
--> Reject; cs=-9$
--> Reject; a2=-9$
--> Reject; phd=-9$

--> probit;lhs=final;rhs=one,an,hb,doc,comp,lib,ci,ck,phd,noeval;hold results$
```

```
Normal exit:   6 iterations. Status=0. F=    822.7411

+--------------------------------------------+
| Binomial Probit Model                      |
| Dependent variable                FINAL    |
| Log likelihood function       -822.7411    |
| Restricted log likelihood     -1284.216    |
| Chi squared [   9 d.f.]       922.95007    |
| Significance level             .0000000    |
| McFadden Pseudo R-squared      .3593438    |
| Estimation based on N =   2587, K =  10    |
| AIC =      .6438  Bayes IC =      .6664    |
| AICf.s. =      .6438  HQIC =      .6520    |
| Model estimated: Dec 08, 2009, 12:12:49    |
| Results retained for SELECTION model.      |
| Hosmer-Lemeshow chi-squared =  26.06658    |
| P-value=  .00102 with deg.fr. =       8    |
+--------------------------------------------+

+--------+-------------+----------------+--------+--------+----------+
|Variable| Coefficient | Standard Error |b/St.Er.|P[|Z|>z]| Mean of X|
+--------+-------------+----------------+--------+--------+----------+
+--------+Index function for probability                              |
|Constant|     .99535***       .24326        4.092   .0000           |
|AN      |     .02204**        .00948        2.326   .0200     10.5968|
|HB      |    -.00488**        .00192       -2.538   .0112     55.5589|
|DOC     |     .97571***       .14636        6.666   .0000      .31774|
|COMP    |     .40649***       .13927        2.919   .0035      .41786|
|LIB     |     .52144***       .17665        2.952   .0032      .13568|
|CI      |     .19873**        .09169        2.168   .0302     1.23116|
|CK      |     .08779          .13429         .654   .5133      .91998|
|PHD     |    -.13351          .10303       -1.296   .1951      .68612|
|NOEVAL  |   -1.93052***       .07239      -26.668   .0000      .29068|
+--------+-------------------------------------------------------------+
| Note: ***, **, * = Significance at 1%, 5%, 10% level.               |
+---------------------------------------------------------------------+

+----------------------------------------+
| Fit Measures for Binomial Choice Model |
| Probit   model for variable FINAL      |
+----------------------------------------+
|                 Y=0       Y=1     Total|
| Proportions  .19714    .80286   1.00000|
| Sample Size     510      2077      2587|
+----------------------------------------+
| Log Likelihood Functions for BC Model  |
|              P=0.50    P=N1/N   P=Model|
| LogL =     -1793.17  -1284.22   -822.74|
+----------------------------------------+
| Fit Measures based on Log Likelihood   |
| McFadden = 1-(L/L0)          =   .35934|
| Estrella = 1-(L/L0)^(-2L0/n) =   .35729|
| R-squared (ML)               =   .30006|
| Akaike Information Crit.     =   .64379|
| Schwartz Information Crit.   =   .66643|
+----------------------------------------+
| Fit Measures Based on Model Predictions|
| Efron                        =   .39635|
| Ben Akiva and Lerman         =   .80562|
| Veall and Zimmerman          =   .52781|
| Cramer                       =   .38789|
+----------------------------------------+
+-------------------------------------------------------+
```

```
|Predictions for Binary Choice Model.  Predicted value is |
|1 when probability is greater than  .500000, 0 otherwise.|
|Note, column or row total percentages may not sum to      |
|100% because of rounding. Percentages are of full sample.|
+------+---------------------------------+----------------+
|Actual|         Predicted Value         |                |
|Value |       0               1         | Total Actual   |
+------+---------------+-----------------+----------------+
|  0   |    342 ( 13.2%)|    168 (  6.5%)|    510 ( 19.7%)|
|  1   |    197 (  7.6%)|   1880 ( 72.7%)|   2077 ( 80.3%)|
+------+----------------+----------------+----------------+
|Total |    539 ( 20.8%)|   2048 ( 79.2%)|   2587 (100.0%)|
+------+----------------+----------------+----------------+
+-----------------------------------------------------------+

|Crosstab for Binary Choice Model.  Predicted probability  |
|vs. actual outcome. Entry = Sum[Y(i,j)*Prob(i,m)] 0,1.    |
|Note, column or row total percentages may not sum to      |
|100% because of rounding. Percentages are of full sample.|
+------+---------------------------------+----------------+
|Actual|      Predicted Probability      |                |
|Value |    Prob(y=0)        Prob(y=1)   | Total Actual   |
+------+----------------+----------------+----------------+
| y=0  |    259 ( 10.0%)|    250 (  9.7%)|    510 ( 19.7%)|
| y=1  |    252 (  9.7%)|   1824 ( 70.5%)|   2077 ( 80.2%)|
+------+----------------+----------------+----------------+
|Total |    512 ( 19.8%)|   2074 ( 80.2%)|   2587 ( 99.9%)|
+------+----------------+----------------+----------------+

======================================================================
Analysis of Binary Choice Model Predictions Based on Threshold =  .5000
----------------------------------------------------------------------
Prediction Success
----------------------------------------------------------------------
Sensitivity = actual 1s correctly predicted                     87.819%
Specificity = actual 0s correctly predicted                     50.784%
Positive predictive value = predicted 1s that were actual 1s    87.946%
Negative predictive value = predicted 0s that were actual 0s    50.586%
Correct prediction = actual 1s and 0s correctly predicted       80.518%
----------------------------------------------------------------------
Prediction Failure
----------------------------------------------------------------------
False pos. for true neg. = actual 0s predicted as 1s            49.020%
False neg. for true pos. = actual 1s predicted as 0s            12.133%
False pos. for predicted pos. = predicted 1s actual 0s          12.054%
False neg. for predicted neg. = predicted 0s actual 1s          49.219%
False predictions = actual 1s and 0s incorrectly predicted      19.405%
======================================================================
```

--> selection;lhs=change;rhs=one,hb,doc,comp,lib,ci,ck,phd,noeval;mle$

```
+-----------------------------------------------------------+
| Sample Selection Model                                    |
| Probit selection equation based on FINAL                  |
| Selection rule is: Observations with FINAL    =  1        |
| Results of selection:                                     |
|                 Data points     Sum of weights            |
| Data set            2587            2587.0                |
| Selected sample     2077            2077.0                |
+-----------------------------------------------------------+
```

```
+----------------------------------------------------+

| Sample Selection Model                             |
| Two step    least squares regression               |
| LHS=CHANGE   Mean                 =   5.456909     |
|              Standard deviation   =   4.582964     |
|              Number of observs.   =       2077     |
| Model size   Parameters           =         10     |
|              Degrees of freedom   =       2067     |
| Residuals    Sum of squares       =   39226.14     |
|              Standard error of e  =   4.356298     |
| Fit          R-squared            =   .0960355     |
|              Adjusted R-squared   =   .0920996     |
| Model test   F[  9,  2067] (prob) =  24.40 (.0000) |
| Diagnostic   Log likelihood       =  -5998.683     |
|              Restricted(b=0)      =  -6108.548     |
|              Chi-sq [  9]  (prob) = 219.73 (.0000) |
| Info criter. LogAmemiya Prd. Crt. =   2.948048     |
|              Akaike Info. Criter. =   2.948048     |
|              Bayes Info. Criter.  =   2.975196     |
| Not using OLS or no constant. Rsqd & F may be < 0. |
| Model was estimated Dec 08, 2009 at 00:12:49PM     |
| Standard error corrected for selection..   4.36303 |
| Correlation of disturbance in regression           |
| and Selection Criterion (Rho).....    .11132       |
+----------------------------------------------------+
+--------+--------------+----------------+--------+--------+----------+
|Variable| Coefficient  | Standard Error |b/St.Er.|P[|Z|>z]| Mean of X|
+--------+--------------+----------------+--------+--------+----------+
|Constant|    6.74123***       .75107        8.976   .0000            |
|HB      |    -.01022*         .00563       -1.815   .0695     55.7429|
|DOC     |    2.07968***       .57645        3.608   .0003      .33558|
|COMP    |    -.32946          .44269        -.744   .4567      .40924|
|LIB     |    2.27448***       .53733        4.233   .0000      .14011|
|CI      |     .40823          .25929        1.574   .1154     1.22773|
|CK      |   -2.73074***       .37755       -7.233   .0000      .91815|
|PHD     |     .63345**        .29104        2.177   .0295      .69957|
|NOEVAL  |    -.88434         1.27223        -.695   .4870      .15744|
|LAMBDA  |     .48567         1.59683         .304   .7610      .21796|
+--------+-------------------------------------------------------------+
| Note: ***, **, * = Significance at 1%, 5%, 10% level.               |
+---------------------------------------------------------------------+

Normal exit:  25 iterations. Status=0. F=    6826.467


---------------------------------------------------------------------
ML Estimates of Selection Model
Dependent variable               CHANGE
Log likelihood function     -6826.46734
Estimation based on N =   2587, K =  21
Information Criteria: Normalization=1/N
             Normalized   Unnormalized
AIC             5.29375    13694.93469
Fin.Smpl.AIC    5.29389    13695.29492
Bayes IC        5.34131    13817.95802
Hannan Quinn    5.31099    13739.52039
Model estimated: Mar 31, 2010, 15:17:41
FIRST 10 estimates are probit equation.
--------+------------------------------------------------------
        |                 Standard          Prob.
  CHANGE| Coefficient       Error       z    z>|Z|
```

```
--------+------------------------------------------------------------
        |Selection (probit) equation for FINAL
Constant|     .99018***      .24020     4.12  .0000
      AN|     .02278**       .00940     2.42  .0153
      HB|    -.00489**       .00206    -2.37  .0178
     DOC|     .97154***      .15076     6.44  .0000
    COMP|     .40431***      .14433     2.80  .0051
     LIB|     .51505***      .19086     2.70  .0070
      CI|     .19927**       .09054     2.20  .0277
      CK|     .08590         .11902      .72  .4705
     PHD|    -.13208         .09787    -1.35  .1772
  NOEVAL|   -1.92902***      .07138   -27.03  .0000
        |Corrected regression, Regime 1
Constant|    6.81754***      .72389     9.42  .0000
      HB|    -.00978*        .00559    -1.75  .0803
     DOC|    1.99729***      .55348     3.61  .0003
    COMP|    -.36198         .43327     -.84  .4034
     LIB|    2.23154***      .50534     4.42  .0000
      CI|     .39401         .25339     1.55  .1199
      CK|   -2.74337***      .38031    -7.21  .0000
     PHD|     .64209**       .28964     2.22  .0266
  NOEVAL|    -.63201        1.26902     -.50  .6185
SIGMA(1)|    4.35713***      .07012    62.14  .0000
RHO(1,2)|     .03706         .35739      .10  .9174
--------+------------------------------------------------------------
```

The estimated probit model (as found on page 7) is

Estimated propensity to take the posttest = 0.995 + 0.022*(preTUCE score)*

− 0 .005*(initial class size)* + 0.976*(Doctoral Institution)*

+ 0.406 *(Comprehensive Institution)* + 0.521*(Liberal Arts Institution)*

+ 0.199 *(Male instructor)* + 0.0878*(English Instructor Native Language)*

− 0.134*(Instructor has PhD)* − 1.930*(No Evaluation of Instructor)*

The beginning or initial class size is negatively and highly significantly related to the propensity to take the posttest, with a one-tail p value of 0.0056.

The corresponding change-score equation employing the inverse Mills ratio is on page 9:

Predicted Change = 6.741 − 0.010(initial class size) + 2.080(Doctoral Institution)

− 0.329 (Comprehensive Institution) + 2.274 Liberal Arts Institution)

+ .408(Male instructor) − 2.731(English Instructor Native Language)

+ 0.633(Instructor has PhD) − 0.88434(No Evaluation of Instructor) + 0 .486 λ

The change score is negatively and significantly related to the class size, with a one-tail p value of 0.0347, but it takes an additional 100 students to lower the change score by a point.

Page 10 provides maximum likelihood estimation of both the probit equation and the change score equation with separate estimation of ρ and σ_ε. The top panel provides the probit coefficients for the propensity equation, where it is shown that initial class size is negatively and significantly related to the propensity to take the posttest with a one-tail p value of 0.009. The second panel gives the change score results, where initial class size is negatively and significantly related to the change score with a one-tail p value of 0.040. Again, it takes approximately 100 students to move the change score in the opposite direction by a point.

As a closing comment on the estimation of the Heckit model, it is worth pointing out that there is no unique way to estimate the standard errors via maximum likelihood computer routines. Historically, LIMDEP used the conventional second derivatives matrix to compute standard errors for the maximum likelihood estimation of the two-equation Heckit model. In the process of preparing this module, differences in standard errors produced by LIMDEP and STATA suggested that STATA was using the alternative outer products of the first derivatives. To achieve consistency, Bill Greene modified the LIMDEP routine in April 2010 so that it also now uses the outer products of the first derivatives.

AN APPLICATION OF PROPENSITY SCORE MATCHING

Unfortunately, we are not aware of a study in economic education for which propensity score matching has been used. Thus, we looked outside economic education and elected to redo the example reported in Becker and Ichino (2002). This application and data are derived from Dehejia and Wahba (1999), whose study, in turn was based on LaLonde (1986). The data set consists of observed samples of treatments and controls from the National Supported Work demonstration. Some of the institutional features of the data set are given by Becker and Ichino. The data were downloaded from the website http://www.nber.org/~rdehejia/nswdata.html. The data set used here is in the original text form, contained in the data file "matchingdata.txt." They have been assembled from the several parts in the NBER archive.

Becker and Ichino report that they were unable to replicate Dehejia and Wahba's results, though they did obtain similar results. (They indicate that they did not have the original authors' specifications of the number of blocks used in the partitioning of the range of propensity scores, significance levels, or exact procedures for testing the balancing property.) In turn, we could not precisely replicate Becker and Ichino's results – we can identify the reason, as discussed below. Likewise, however, we obtain similar results.

There are 2,675 observations in the data set, 2490 controls (with t = 0) and 185 treated observations (with t = 1). The variables in the raw data set are

t = treatment dummy variable
age = age in years
educ = education in years
black = dummy variable for black
hisp = dummy variable for Hispanic
marr = dummy variable for married
nodegree = dummy for no degree (not used)
*re*74 = real earnings in 1974
*re*75 = real earnings in 1975
*re*78 = real earnings in 1978 – the outcome variable

We will analyze these data following Becker and Ichino's line of analysis. We assume that you have completed Module One, Part Two, and thus are familiar with placing commands in the text editor and using the GO button to submit commands, and where results are found in the output window. In what follows, we will simply show the commands you need to enter into LIMDEP (NLOGIT) to produce the results that we will discuss.

To start, the data are imported by using the command (where the data file is on the C drive but your data could be placed wherever):

READ ; file=C:\matchingdata.txt;
names=t,age,educ,black,hisp,marr,nodegree,re74,re75,re78;nvar=10;nobs=2675$

Transformed variables added to the equation are

age2 = age squared
educ2 = educ squared
re742 = re74 squared
re752 = re75 squared
blacku74 = black times 1(re74 = 0)

In order to improve the readability of some of the reported results, we have divided the income variables by 10,000. (This is also an important adjustment that accommodates a numerical problem with the original data set. This is discussed below.) The outcome variable is re78.

The data are set up and described first. The transformations used to create the transformed variables are

CREATE ; age2 = age^2 ; educ2 = educ^2 $
CREATE ; re74 = re74/10000 ; re75 = re75/10000 ; re78 = re78/10000 $
CREATE ; re742 = re74^2 ; re752 = re75^2 $
CREATE ; blacku74 = black * (re74 = 0) $

The data are described with the following statistics:

DSTAT ; Rhs = * $

Descriptive Statistics
All results based on nonmissing observations.

Variable	Mean	Std.Dev.	Minimum	Maximum	Cases	Missing
All observations in current sample						
T	.691589E-01	.253772	.000000	1.00000	2675	0
AGE	34.2258	10.4998	17.0000	55.0000	2675	0
EDUC	11.9944	3.05356	.000000	17.0000	2675	0
BLACK	.291589	.454579	.000000	1.00000	2675	0
HISP	.343925E-01	.182269	.000000	1.00000	2675	0
MARR	.819439	.384726	.000000	1.00000	2675	0
NODEGREE	.333084	.471404	.000000	1.00000	2675	0
RE74	1.82300	1.37223	.000000	13.7149	2675	0
RE75	1.78509	1.38778	.000000	15.6653	2675	0
RE78	2.05024	1.56325	.000000	12.1174	2675	0
AGE2	1281.61	766.842	289.000	3025.00	2675	0
EDUC2	153.186	70.6223	.000000	289.000	2675	0

```
   RE742|  5.20563      8.46589      .000000      188.098         2675       0
   RE752|  5.11175      8.90808      .000000      245.402         2675       0
BLACKU74|  .549533E-01  .227932      .000000      1.00000         2675       0
```

We next fit the logit model for the propensity scores. An immediate problem arises with the data set as used by Becker and Ichino. The income data are in raw dollar terms – the mean of re74, for example is $18,230.00. The square of it, which is on the order of 300,000,000, as well as the square of re75 which is similar, is included in the logit equation with a dummy variable for Hispanic which is zero for 96.5% of the observations and the blacku74 dummy variable which is zero for 94.5% of the observations. Because of the extreme difference in magnitudes, estimation of the logit model in this form is next to impossible. But rescaling the data by dividing the income variables by 10,000 addresses the instability problem.[ii] These transformations are shown in the second CREATE command above. This has no impact on the results produced with the data, other than stabilizing the estimation of the logit equation. We are now quite able to replicate the Becker and Ichino results except for an occasional very low order digit.

The logit model from which the propensity scores are obtained is fit using

```
NAMELIST ; X = age,age2,educ,educ2,marr,black,hisp,
               re74,re75,re742,re752,blacku74,one $
LOGIT ; Lhs = t ; Rhs = x ; Hold $
```

(Note: Becker and Ichino's coefficients on re74 and re75 are multiplied by 10,000, and coefficients on re742 and re752 are multiplied by 100,000,000. Some additional logit results from LIMDEP are omitted. Becker and Ichino's results are included in the results for comparison.)

```
------------------------------------------------------------------
Binary Logit Model for Binary Choice
Dependent variable                    T     Becker/Ichino
Log likelihood function      -204.97536      (-204.97537)
Restricted log likelihood    -672.64954      (identical)
Chi squared [  12 d.f.]       935.34837
Significance level               .00000
McFadden Pseudo R-squared      .6952717
Estimation based on N =   2675, K =  13
Information Criteria: Normalization=1/N
              Normalized   Unnormalized
AIC               .16297      435.95071
Fin.Smpl.AIC      .16302      436.08750
Bayes IC          .19160      512.54287
Hannan Quinn      .17333      463.66183
Hosmer-Lemeshow chi-squared =  12.77381
P-value=  .11987 with deg.fr. =       8
--------+---------------------------------------------------
        |                 Standard           Prob.       Mean
      T | Coefficient       Error       z    z>|Z|       of X
--------+---------------------------------------------------    Becker/Ichino
        |Characteristics in numerator of Prob[Y = 1]             Coeff.    |t|
    AGE |     .33169***      .12033     2.76  .0058    34.2258   .3316904  (2.76)
```

```
    AGE2|    -.00637***      .00186    -3.43  .0006    1281.61  -.0063668  (3.43)
    EDUC|     .84927**       .34771     2.44  .0146    11.9944   .8492683  (2.44)
   EDUC2|    -.05062***      .01725    -2.93  .0033    153.186  -.0506202  (2.93)
    MARR|   -1.88554***      .29933    -6.30  .0000     .81944  -1.885542  (6.30)
   BLACK|    1.13597***      .35179     3.23  .0012     .29159   1.135973  (3.23)
    HISP|    1.96902***      .56686     3.47  .0005     .03439   1.969020  (3.47)
    RE74|   -1.05896***      .35252    -3.00  .0027    1.82300  -.1059000  (3.00)
    RE75|   -2.16854***      .41423    -5.24  .0000    1.78509  -.2169000  (5.24)
   RE742|     .23892***      .06429     3.72  .0002    5.20563   .2390000  (3.72)
   RE752|     .01359         .06654      .20  .8381    5.11175   .0136000  (0.21)
BLACKU74|    2.14413***      .42682     5.02  .0000     .05495   2.144129  (5.02)
Constant|   -7.47474***     2.44351    -3.06  .0022             -7.474742  (3.06)
--------+-------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
------------------------------------------------------------------
+-----------------------------------------------------------+
|Predictions for Binary Choice Model.  Predicted value is |
|1 when probability is greater than  .500000, 0 otherwise.|
|Note, column or row total percentages may not sum to     |
|100% because of rounding. Percentages are of full sample.|
+------+---------------------------------+----------------+
|Actual|         Predicted Value         |                |
|Value |       0               1        | Total Actual   |
+------+---------------+-----------------+----------------+
|  0   |   2463 ( 92.1%)|     27 (  1.0%)|   2490 ( 93.1%)|
|  1   |     51 (  1.9%)|    134 (  5.0%)|    185 (  6.9%)|
+------+---------------+-----------------+----------------+
|Total |   2514 ( 94.0%)|    161 (  6.0%)|   2675 (100.0%)|
+------+---------------+-----------------+----------------+
```

The first set of matching results uses the kernel estimator for the neighbors, lists the intermediate results, and uses only the observations in the common support.[iii]

MATCH ; Lhs = re78 ; Kernel ; List ; Common Support $

The estimated propensity score function is echoed first. This merely reports the earlier estimated binary choice model for the treatment assignment. The treatment assignment model is not reestimated. (The ;Hold in the LOGIT or PROBIT command stores the estimated model for this use.)

```
+---------------------------------------------------+
| ******* Propensity Score Matching Analysis ****** |
| Treatment variable = T       , Outcome = RE78     |
| Sample In Use                                     |
| Total number of observations     =   2675         |
| Number of valid (complete) obs.  =   2675         |
| Number used (in common support)  =   1342         |
| Sample Partitioning of Data In Use                |
|                     Treated   Controls     Total  |
| Observations            185       1157      1342  |
| Sample Proportion    13.79%     86.21%   100.00%  |
+---------------------------------------------------+

+-----------------------------------------------------------+
```

```
| Propensity Score Function = Logit  based on T               |
| Variable   Coefficient  Standard Error  t statistic         |
| AGE             .33169       .12032986       2.757          |
| AGE2           -.00637       .00185539      -3.432          |
| EDUC            .84927       .34770583       2.442          |
| EDUC2          -.05062       .01724929      -2.935          |
| MARR          -1.88554       .29933086      -6.299          |
| BLACK          1.13597       .35178542       3.229          |
| HISP           1.96902       .56685941       3.474          |
| RE74          -1.05896       .35251776      -3.004          |
| RE75          -2.16854       .41423244      -5.235          |
| RE742           .23892       .06429271       3.716          |
| RE752           .01359       .06653758        .204          |
| BLACKU74       2.14413       .42681518       5.024          |
| ONE           -7.47474      2.44351058      -3.059          |
| Note:Estimation sample may not be the sample analyzed here. |
| Observations analyzed are restricted to the common support =|
| only controls with propensity in the range of the treated.  |
+-------------------------------------------------------------+
```

The note in the reported logit results reports how the common support is defined, that is, as the range of variation of the scores for the treated observations.

The next set of results reports the iterations that partition the range of estimated probabilities. The report includes the results of the *F* tests within the partitions as well as the details of the full partition itself. The balancing hypothesis is rejected when the p value is less than 0.01 within the cell. Becker and Ichino do not report the results of this search for their data, but do report that they ultimately found seven blocks, as we did. They do not report the means by which the test of equality is carried out within the blocks or the critical value used.

```
Partitioning the range of propensity scores
================================================================================
Iteration  1. Partitioning range of propensity scores into  5 intervals.
================================================================================
    Range               Controls              Treatment
                  # Obs. Mean PS S.D. PS   # obs. Mean PS S.D. PS     F     Prob
---------------  ---------------------   ---------------------  -------------
  .00061  .19554    1081  .02111  .03337       17  .07358  .05835   13.68  .0020
  .19554  .39047      41  .28538  .05956       26  .30732  .05917    2.18  .1460
  .39047  .58540      15  .49681  .05098       20  .49273  .06228     .05  .8327
  .58540  .78033      13  .68950  .04660       19  .64573  .04769    6.68  .0157
  .78033  .97525       7  .96240  .00713      103  .93022  .05405   29.05  .0000
Iteration  1  Mean scores are not equal in at least one cell
================================================================================

Iteration  2. Partitioning range of propensity scores into  6 intervals.
================================================================================
    Range               Controls              Treatment
                  # Obs. Mean PS S.D. PS   # obs. Mean PS S.D. PS     F     Prob
---------------  ---------------------   ---------------------  -------------
  .00061  .09807    1026  .01522  .02121       11  .03636  .03246    4.64  .0566
  .09807  .19554      55  .13104  .02762        6  .14183  .02272    1.16  .3163
  .19554  .39047      41  .28538  .05956       26  .30732  .05917    2.18  .1460
  .39047  .58540      15  .49681  .05098       20  .49273  .06228     .05  .8327
  .58540  .78033      13  .68950  .04660       19  .64573  .04769    6.68  .0157
```

```
  .78033  .97525       7  .96240  .00713      103  .93022  .05405   29.05  .0000
Iteration  2  Mean scores are not equal in at least one cell
================================================================================
Iteration  3. Partitioning range of propensity scores into  7 intervals.
================================================================================
    Range               Controls              Treatment
                  # Obs. Mean PS S.D. PS   # obs. Mean PS S.D. PS     F     Prob
---------------  ---------------------   ---------------------  -------------
  .00061  .09807    1026  .01522  .02121       11  .03636  .03246    4.64  .0566
  .09807  .19554      55  .13104  .02762        6  .14183  .02272    1.16  .3163
  .19554  .39047      41  .28538  .05956       26  .30732  .05917    2.18  .1460
  .39047  .58540      15  .49681  .05098       20  .49273  .06228     .05  .8327
  .58540  .78033      13  .68950  .04660       19  .64573  .04769    6.68  .0157
  .78033  .87779       0  .00000  .00000       17  .81736  .02800     .00 1.0000
  .87779  .97525       7  .96240  .00713       86  .95253  .01813    8.77  .0103
Mean PSCORES are tested equal within the blocks listed below
```

After partitioning the range of the propensity scores, we report the empirical distribution of the propensity scores and the boundaries of the blocks estimated above. The values below show the percentiles that are also reported by Becker and Ichino. The reported search algorithm notwithstanding, the block boundaries shown by Becker and Ichino shown below are roughly the same.

```
+-------------------------------------------------------------+
| Empirical Distribution of Propensity Scores in Sample Used  | Becker/Ichino
|   Percent      Lower   Upper   Sample size =   1342         | Percentiles (lower)
|   0% -   5%   .000611  .000801 Average score  .137746       | .0006426
|   5% -  10%   .000802  .001088 Std.Dev score  .274560       | .0008025
|  10% -  15%   .001093  .001378 Variance       .075383       | .0010932
|  15% -  20%   .001380  .001809 Blocks used to test balance  |
|  20% -  25%   .001815  .002355      Lower    Upper   # obs  |
|  25% -  30%   .002355  .003022  1  .000611  .098075   1037  | .0023546
|  30% -  35%   .003046  .004094  2  .098075  .195539     61  |
|  35% -  40%   .004097  .005299  3  .195539  .390468     67  |
|  40% -  45%   .005315  .007631  4  .390468  .585397     35  |
|  45% -  50%   .007632  .010652  5  .585397  .780325     32  |
|  50% -  55%   .010682  .015103  6  .780325  .877790     17  | .0106667
|  55% -  60%   .015105  .022858  7  .877790  .975254     93  |
|  60% -  65%   .022888  .035187                              |
|  65% -  70%   .035316  .051474                              |
|  70% -  75%   .051488  .075104                              |
|  75% -  80%   .075712  .135218                              | .0757115
|  80% -  85%   .135644  .322967                              |
|  85% -  90%   .335230  .616205                              |
|  90% -  95%   .625082  .949302                              | .6250832
|  95% - 100%   .949302  .975254                              | .949382 to .970598
+-------------------------------------------------------------+
```

The blocks used for the balancing hypothesis are shown at the right in the table above. Becker and Ichino report that they used the following blocks and sample sizes:

	Lower	Upper	Observations
1	0.0006	0.05	931
2	0.05	0.10	106
3	0.10	0.20	3
4	0.20	0.40	69
5	0.40	0.60	35

6	0.60	0.80	33
7	0.80	1.00	105

At this point, our results begin to differ somewhat from those of Becker and Ichino because they are using a different (cruder) blocking arrangement for the ranges of the propensity scores. This should not affect the ultimate estimation of the ATE; it is an intermediate step in the analysis that is a check on the reliability of the procedure.

The next set of results reports the analysis of the balancing property for the independent variables. A test is reported for each variable in each block as listed in the table above. The lines marked (by the program) with "*" show cells in which one or the other group had no observations, so the *F* test could not be carried out. This was treated as a "success" in each analysis. Lines marked with an "o" note where the balancing property failed. There are only four of these, but those we do find are not borderline. Becker and Ichino report their finding that the balancing property is satisfied. Note that our finding does not prevent the further analysis. It merely suggests to the analyst that they might want to consider a richer specification of the propensity function model.

```
Examining exogenous variables for balancing hypothesis
* Indicates no observations, treatment and/or controls, for test.
o Indicates means of treated and controls differ significantly.
================================================================
Variable  Interval Mean Control   Mean Treated    F     Prob
--------  -------- ------------   ------------  ------  -----
AGE            1       31.459064      30.363636     .41  .5369
AGE            2       27.727273      26.500000     .10  .7587
AGE            3       28.170732      28.769231     .07  .7892
AGE            4       26.800000      25.050000     .44  .5096
AGE            5       24.846154      24.210526     .10  .7544
AGE            6          .000000      30.823529     .00 1.0000 *
AGE            7       23.285714      23.837209     .55  .4653
AGE2           1     1081.180312     953.454545    1.43  .2576
AGE2           2      822.200000     783.833333     .02  .8856
AGE2           3      873.341463     906.076923     .05  .8202
AGE2           4      774.400000     690.350000     .25  .6193
AGE2           5      644.230769     623.789474     .03  .8568
AGE2           6         .000000    1003.058824     .00 1.0000 *
AGE2           7      543.857143     596.023256    1.99  .1666
EDUC           1       11.208577      11.545455     .37  .5575
EDUC           2       10.636364      10.166667     .40  .5463
EDUC           3       10.414634      10.076923     .31  .5819
EDUC           4       10.200000      10.150000     .01 1.0000
EDUC           5       10.230769      11.000000    1.03  .3218
EDUC           6          .000000      11.058824     .00 1.0000 *
EDUC           7       10.571429      10.046512     .86  .3799
EDUC2          1      132.446394     136.636364     .11  .7420
EDUC2          2      117.618182     106.166667     .60  .4624
EDUC2          3      113.878049     107.769231     .31  .5829
EDUC2          4      108.066667     107.650000     .00 1.0000
EDUC2          5      109.923077     124.263158     .83  .3703
EDUC2          6         .000000     124.705882     .00 1.0000 *
EDUC2          7      113.714286     104.302326     .70  .4275
MARR           1         .832359        .818182     .01  .9056
MARR           2         .563636        .833333    2.63  .1433
MARR           3         .268293        .269231     .00 1.0000
MARR           4         .200000        .050000    1.73  .2032
```

```
MARR          5         .153846        .210526     .17  .6821
MARR          6         .000000        .529412     .00 1.0000 *
MARR          7         .000000        .000000     .00 1.0000
BLACK         1         .358674        .636364    3.63  .0833
BLACK         2         .600000        .500000     .22  .6553
BLACK         3         .780488        .769231     .01  .9150
BLACK         4         .866667        .500000    6.65  .0145
BLACK         5         .846154        .947368     .81  .3792
BLACK         6         .000000        .941176     .00 1.0000 *
BLACK         7        1.000000        .953488     .00 1.0000 *
HISP          1         .048733        .000000   52.46  .0000 o
HISP          2         .072727        .333333    1.77  .2311
HISP          3         .048780        .000000    2.10  .1547
HISP          4         .066667        .150000     .66  .4224
HISP          5         .153846        .052632     .81  .3792
HISP          6         .000000        .058824     .00 1.0000 *
HISP          7         .000000        .046512    4.19  .0436
RE74          1        1.230846       1.214261     .00 1.0000
RE74          2         .592119        .237027   10.63  .0041 o
RE74          3         .584965        .547003     .06  .8074
RE74          4         .253634        .298130     .16  .6875
RE74          5         .154631        .197888     .44  .5108
RE74          6         .000000        .002619     .00 1.0000 *
RE74          7         .000000        .000000     .00 1.0000
RE75          1        1.044680        .896447     .41  .5343
RE75          2         .413079        .379168     .09  .7653
RE75          3         .276234        .279825     .00 1.0000
RE75          4         .286058        .169340    2.39  .1319
RE75          5         .137276        .139118     .00 1.0000
RE75          6         .000000        .061722     .00 1.0000 *
RE75          7         .012788        .021539     .37  .5509
RE742         1        2.391922       2.335453     .00 1.0000
RE742         2         .672950        .092200    9.28  .0035 o
RE742         3         .638937        .734157     .09  .7625
RE742         4         .127254        .245461    1.14  .2936
RE742         5         .040070        .095745    1.31  .2647
RE742         6         .000000        .000117     .00 1.0000 *
RE742         7         .000000        .000000     .00 1.0000
RE752         1        1.779930       1.383457     .43  .5207
RE752         2         .313295        .201080    1.48  .2466
RE752         3         .151139        .135407     .14  .7133
RE752         4         .128831        .079975     .97  .3308
RE752         5         .088541        .037465     .51  .4894
RE752         6         .000000        .037719     .00 1.0000 *
RE752         7         .001145        .005973    2.57  .1124
BLACKU74      1         .014620        .000000   15.12  .0001 o
BLACKU74      2         .054545        .000000    3.17  .0804
BLACKU74      3         .121951        .192308     .58  .4515
BLACKU74      4         .200000        .100000     .66  .4242
BLACKU74      5         .230769        .315789     .29  .5952
BLACKU74      6         .000000        .941176     .00 1.0000 *
BLACKU74      7        1.000000        .953488     .00 1.0000 *
Variable BLACKU74 is unbalanced in block  1
Other variables may also be unbalanced
You might want to respecify the index function for the P-scores
```

This part of the analysis ends with a recommendation that the analyst reexamine the specification of the propensity score model. Because this is not a numerical problem, the analysis continues with estimation of the average treatment effect on the treated.

The first example below shows estimation using the kernel estimator to define the counterpart observation from the controls and using only the subsample in the common support. This stage consists of *nboot* + 1 iterations. In order to be able to replicate the results, we set the seed of the random number generator before computing the results.

CALC ; Ran(1234579) $
MATCH ; Lhs = re78 ; Kernel ; List ; Common Support $

The first result is the actual estimation, which is reported in the intermediate results. Then the *nboot* repetitions are reported. (These will be omitted if **; List** is not included in the command.) Recall, we divided the income values by 10,000. The value of .156255 reported below thus corresponds to $1,562.55. Becker and Ichino report a value (see their section 6.4) of $1537.94. Using the bootstrap replications, we have estimated the asymptotic standard error to be $1042.04. A 95% confidence interval for the treatment effect is computed using $1537.94 ± 1.96(1042.04) = (-$325.41,$3474.11).

```
+---------------------------------------------------------------------+
| Estimated Average Treatment Effect (T       )  Outcome is RE78      |
| Kernel            Using Epanechnikov kernel with bandwidth =  .0600 |
| Note, controls may be reused in defining matches.                   |
| Number of bootstrap replications used to obtain variance    =     25 |
+---------------------------------------------------------------------+
  Estimated average treatment effect =       .156255
  Begin bootstrap iterations ****************************************
  Boootstrap estimate   1            =       .099594
  Boootstrap estimate   2            =       .109812
  Boootstrap estimate   3            =       .152911
  Boootstrap estimate   4            =       .168743
  Boootstrap estimate   5            =      -.015677
  Boootstrap estimate   6            =       .052938
  Boootstrap estimate   7            =      -.003275
  Boootstrap estimate   8            =       .212767
  Boootstrap estimate   9            =      -.042274
  Boootstrap estimate  10            =       .053342
  Boootstrap estimate  11            =       .351122
  Boootstrap estimate  12            =       .117883
  Boootstrap estimate  13            =       .181123
  Boootstrap estimate  14            =       .111917
  Boootstrap estimate  15            =       .181256
  Boootstrap estimate  16            =      -.012129
  Boootstrap estimate  17            =       .240363
  Boootstrap estimate  18            =       .201321
  Boootstrap estimate  19            =       .169463
  Boootstrap estimate  20            =       .238131
  Boootstrap estimate  21            =       .358050
  Boootstrap estimate  22            =       .199020
  Boootstrap estimate  23            =       .083503
  Boootstrap estimate  24            =       .146215
  Boootstrap estimate  25            =       .266303
  End bootstrap iterations   ****************************************
+---------------------------------------------------------------------+
| Number of Treated observations =    185  Number of controls =   1157 |
| Estimated Average Treatment Effect   =        .156255               | (.153794)
| Estimated Asymptotic Standard Error  =        .104204               | (.101687)
| 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                |
+-------------------------------------------------------------------+
```

Note that the estimated asymptotic standard error is somewhat different. As we noted earlier, because of differences in random number generators, the bootstrap replications will differ across programs. It will generally not be possible to exactly replicate results generated with different computer programs. With a specific computer program, replication is obtained by setting the seed of the random number generator. (The specific seed chosen is immaterial, so long as the same seed is used each time.)

The next set of estimates is based on all of the program defaults. The single nearest neighbor is used for the counterpart observation; 25 bootstrap replications are used to compute the standard deviation, and the full range of propensity scores (rather than the common support) is used. Intermediate output is also suppressed. Once again, we set the seed for the random number generator before estimation.

CALC ; Ran(1234579) $
MATCH ; Rhs = re78 $

```
Partitioning the range of propensity scores
Iteration  1  Mean scores are not equal in at least one cell
Iteration  2  Mean scores are not equal in at least one cell
Mean PSCORES are tested equal within the blocks listed below.

+-------------------------------------------------------------+
| Empirical Distribution of Propensity Scores in Sample Used  |
|   Percent      Lower   Upper   Sample size =   2675         |
|   0% -   5%   .000000  .000000 Average score  .069159       |
|   5% -  10%   .000000  .000002 Std.Dev score  .206287       |
|  10% -  15%   .000002  .000006 Variance       .042555       |
|  15% -  20%   .000007  .000015 Blocks used to test balance  |
|  20% -  25%   .000016  .000032      Lower    Upper   # obs  |
|  25% -  30%   .000032  .000064  1  .000000  .097525   2370  |
|  30% -  35%   .000064  .000121  2  .097525  .195051     60  |
|  35% -  40%   .000121  .000204  3  .195051  .390102     68  |
|  40% -  45%   .000204  .000368  4  .390102  .585152     35  |
|  45% -  50%   .000368  .000618  5  .585152  .780203     32  |
|  50% -  55%   .000618  .001110  6  .780203  .877729     17  |
|  55% -  60%   .001123  .001851  7  .877729  .975254     93  |
|  60% -  65%   .001854  .003047                              |
|  65% -  70%   .003057  .005451                              |
|  70% -  75%   .005451  .010756                              |
|  75% -  80%   .010877  .023117                              |
|  80% -  85%   .023149  .051488                              |
|  85% -  90%   .051703  .135644                              |
|  90% -  95%   .136043  .625082                              |
|  95% - 100%   .625269  .975254                              |
+-------------------------------------------------------------+
Examining exogenous variables for balancing hypothesis
Variable BLACKU74 is unbalanced in block  1
Other variables may also be unbalanced
You might want to respecify the index function for the P-scores

+-------------------------------------------------------------------+
| Estimated Average Treatment Effect (T       )  Outcome is RE78    |
```

```
| Nearest Neighbor  Using average of  1 closest neighbors              |
| Note, controls may be reused in defining matches.                    |
| Number of bootstrap replications used to obtain variance    =     25 |
+----------------------------------------------------------------------+
  Estimated average treatment effect =       .169094
  Begin bootstrap iterations ****************************************
  End bootstrap iterations   ****************************************
+----------------------------------------------------------------------+
| Number of Treated observations =    185  Number of controls =     54 |
| Estimated Average Treatment Effect   =        .169094                |
| Estimated Asymptotic Standard Error  =        .102433                |
| t statistic (ATT/Est.S.E.)           =       1.650772                |
| Confidence Interval for ATT = (     -.031675  to        .369864) 95% |
| Average Bootstrap estimate of ATT    =        .171674                |
| ATT - Average bootstrap estimate     =       -.002579                |
+----------------------------------------------------------------------+
```

Using the full sample in this fashion produces an estimate of $1,690.94 for the treatment effect with an estimated standard error of $1,093.29. Note that from the results above, we find that only 54 of the 2490 control observations were used as nearest neighbors for the 185 treated observations. In comparison, using the 1,342 observations in their estimated common support, and the same 185 treateds, Becker and Ichino reported estimates of $1,667.64 and $2,113.59 for the effect and the standard error, respectively and use 57 of the 1,342 controls as nearest neighbors.

The next set of results uses the caliper form of matching and again restricts attention to the estimates in the common support.

```
CALC    ; Ran(1234579) $
MATCH ; Rhs = re78 ; Range = .0001 ; Common Support $
CALC    ; Ran(1234579) $
MATCH ; Rhs = re78 ; Range = .01    ;  Common Support $
```

The estimated treatment effects are now very different. We see that only 23 of the 185 treated observations had a neighbor within a range (radius in the terminology of Becker and Ichino) of 0.0001. The treatment effect is estimated to be only $321.95 with a standard error of $307.95. In contrast, using this procedure, and this radius, Becker and Ichino report a nonsense result of -$5,546.10 with a standard error of $2,388.72. They state that this illustrates the sensitivity of the estimator to the choice of radius, which is certainly the case. To examine this aspect, we recomputed the estimator using a range of 0.01 instead of 0.0001. This produces the expected effect, as seen in the second set of results below. The estimated treatment effect rises to $1433.54 which is comparable to the other results already obtained

```
+----------------------------------------------------------------------+
| Estimated Average Treatment Effect (T       )  Outcome is RE78       |
| Caliper           Using distance of  .00010 to locate matches        |
| Note, controls may be reused in defining matches.                    |
| Number of bootstrap replications used to obtain variance    =     25 |
+----------------------------------------------------------------------+
  Estimated average treatment effect =       .032195
```

```
  Begin bootstrap iterations ****************************************
  End bootstrap iterations   ****************************************
+------------------------------------------------------------------+
| Number of Treated observations =     23  Number of controls =     66 |
| Estimated Average Treatment Effect   =        .032195               |
| Estimated Asymptotic Standard Error  =        .030795               |
| t statistic (ATT/Est.S.E.)           =       1.045454               |
| Confidence Interval for ATT = (     -.028163  to        .092553) 95% |
| Average Bootstrap estimate of ATT    =        .018996               |
| ATT - Average bootstrap estimate     =        .013199               |
+------------------------------------------------------------------+


+------------------------------------------------------------------+
| Estimated Average Treatment Effect (T       )  Outcome is RE78       |
| Caliper           Using distance of  .01000 to locate matches        |
| Note, controls may be reused in defining matches.                    |
| Number of bootstrap replications used to obtain variance    =     25 |
+------------------------------------------------------------------+
  Estimated average treatment effect =       .143354
  Begin bootstrap iterations ****************************************
  End bootstrap iterations   ****************************************
+------------------------------------------------------------------+
| Number of Treated observations =    146  Number of controls =   1111 |
| Estimated Average Treatment Effect   =        .143354               |
| Estimated Asymptotic Standard Error  =        .078378               |
| t statistic (ATT/Est.S.E.)           =       1.829010               |
| Confidence Interval for ATT = (     -.010267  to        .296974) 95% |
| Average Bootstrap estimate of ATT    =        .127641               |
| ATT - Average bootstrap estimate     =        .015713               |
+------------------------------------------------------------------+
```

CONCLUDING COMMENTS

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. On the other hand, 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. Someone using Heckman's original selection adjustment method can easily have their results replicated in LIMDEP, STATA and SAS, although standard error estimates may differ somewhat because of the difference in routines used. Such is not the case with propensity score matching. Propensity score matching results are highly sensitive to the computer program employed while Heckman's original sample selection adjustment method can be relied on to give comparable coefficient estimates across programs.

REFERENCES

Becker, William and William Walstad. "Data Loss From Pretest to Posttest As a Sample Selection Problem," *Review of Economics and Statistics*, Vol. 72, February 1990: 184-188,

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

Becker, S. and A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, Vol. 2, November 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*, Vol. 94, 1999: 1052-1062.

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

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).

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

Saunders, Phillip. The TUCE III Data Set: Background information and file codes (documentation, summary tables, and five 3.5-inch double-sided, high density disks in ASCII format). New York: National Council on Economic Education, 1994.

ENDNOTES

[i] Huynh, Jacho-Chavez, and Self (2010) have a data set that enables them to account for selection into, out of and between collaborative learning sections of a large principles course in their change-score modeling.

[ii] An attempt to compute a linear regression of the original RE78 on the original unscaled other variables is successful, but produces a warning that the condition number of the X matrix is 6.5 times 10^9. When the data are scaled as done above, no warning about multicollinearity is given.

[iii] The Kernel density estimator is a *nonparametric* estimator. Unlike a parametric estimator (which is an equation), a non-parametric estimator has no fixed structure and is based on a histogram of all the data. Histograms are bar charts, which are not smooth, and whose shape depends on the width of the bin into which the data are divided. In essence, with a fixed bin width, the kernel estimator smoothes out the histogram by centering each of the bins at each data point rather than fixing the end points of the bin. The optimum bin width is a subject of debate and well beyond the technical level of this module.

MODULE FOUR, PART THREE: SAMPLE SELECTION

IN ECONOMIC EDUCATION RESEARCH USING STATA

Part Three of Module Four provides a cookbook-type demonstration of the steps required to use STATA in situations involving estimation problems associated with sample selection. Users of this model need to have completed Module One, Parts One and Three, but not necessarily Modules Two and Three. From Module One users are assumed to know how to get data into STATA, recode and create variables within STATA, and run and interpret regression results. Module Four, Parts Two and Four demonstrate in LIMDEP (NLOGIT) and SAS what is done here in STATA.

THE CASE, DATA, AND ROUTINE FOR EARLY HECKMAN ADJUSTMENT

The change score or difference in difference model is used extensively in education research. Yet, before Becker and Walstad (1990), little if any attention was given to the consequence of missing student records that result from: 1) "data cleaning" done by those collecting the data, 2) student unwillingness to provide data, or 3) students self-selecting into or out of the study. The implications of these types of sample selection are shown in the work of Becker and Powers (2001) where the relationship between class size and student learning was explored using the third edition of the Test of Understanding in College Economics (TUCE), which was produced by Saunders (1994) for the National Council on Economic Education (NCEE), since renamed the Council for Economic Education.

Module One, Part Three showed how to get the Becker and Powers data set "beck8WO.csv" into STATA. As a brief review this was done with the insheet command:

```
. insheet a1 a2 x3 c al am an ca cb cc ch ci cj ck cl cm cn co cs ct cu  ///
> cv cw db dd di dj dk dl dm dn dq dr ds dy dz ea eb ee ef               ///
> ei ej ep eq er et ey ez ff fn fx fy fz ge gh gm gn gq gr hb            ///
> hc hd he hf using "F:\BECK8WO2.csv", comma
(64 vars, 2849 obs)
```

where

A1: term, where 1= fall, 2 = spring
A2: school code, where 100/199 = doctorate,
200/299 = comprehensive,
300/399 = lib arts,
400/499 = 2 year
hb: initial class size (number taking preTUCE)
hc: final class size (number taking postTUCE)

dm: experience, as measured by number of years teaching
dj: teacher's highest degree, where Bachelors=1, Masters=2, PhD=3
cc: postTUCE score (0 to 30)
an: preTUCE score (0 to 30)
ge: Student evaluation measured interest
gh: Student evaluation measured textbook quality
gm: Student evaluation measured regular instructor's English ability
gq: Student evaluation measured overall teaching effectiveness
ci: Instructor sex (Male = 1, Female = 2)
ck: English is native language of instructor (Yes = 1, No = 0)
cs: PostTUCE score counts toward course grade (Yes = 1, No = 0)
ff: GPA*100
fn: Student had high school economics (Yes = 1, No = 0)
ey: Student's sex (Male = 1, Female = 2)
fx: Student working in a job (Yes = 1, No = 0)

Separate dummy variables need to be created for each type of school (A2), which is done with the following code:

```
recode a2 (100/199=1) (200/299=2) (300/399=3) (400/499=4)
generate doc=(a2==1) if a2!=.
generate comp=(a2==2) if a2!=.
generate lib=(a2==3) if a2!=.
generate twoyr=(a2==4) if a2!=.
```

To create a dummy variable for whether the instructor had a PhD we use

```
generate phd=(dj==3) if dj!=.
```

To create a dummy variable for whether the student took the postTUCE we use

```
generate final=(cc>0) if cc!=.
```

To create a dummy variable for whether a student did (noeval = 0) or did not (noeval = 1) complete a student evaluation of the instructor we use

```
generate noeval=(ge + gh + gm + gq == -36)
```

"Noeval" reflects whether the student was around toward the end of the term, attending classes, and sufficiently motivated to complete an evaluation of the instructor. In the Saunder's data set evaluation questions with no answer where coded -9; thus, these four questions summing to -36 indicates that no questions were answered.

And the change score is created with

```
generate change=cc-an
```

Finally, there was a correction for the term in which student record 2216 was incorrectly recorded:

```
recode hb (90=89)
```

All of these recoding and create commands are entered into the STATA command file as follows:

```
recode a2 (100/199=1) (200/299=2) (300/399=3) (400/499=4)
gen doc=(a2==1) if a2!=.
gen comp=(a2==2) if a2!=.
gen lib=(a2==3) if a2!=.
gen twoyr=(a2==4) if a2!=.
gen phd=(dj==3) if dj!=.
gen final=(cc>0) if cc!=.

gen noeval=(ge+gh+gm+gq==-36)

gen change=cc-an
recode hb (90=89)
```

To remove records with missing data the following is entered:

```
drop if an==-9
drop if hb==-9
drop if ci==-9
drop if ck==-9
drop if cs==0
drop if cs==-9
drop if a2==-9
drop if phd==-9
```

The use of these data entry and management commands will appear in the STATA output file for the equations to be estimated in the next section.

THE PROPENSITY TO TAKE THE POSTTEST AND THE CHANGE SCORE EQUATION

To address attrition-type sample selection problems in change score studies, Becker and Powers first add observations that were dropped during the early stage of assembling data for TUCE III. Becker and Powers do not have any data on students before they enrolled in the course and thus cannot address selection into the course, but to examine the effects of attrition (course withdrawal) they introduce three measures of class size (beginning, ending, and average) and argue that initial or beginning class size is the critical measure for assessing learning over the entire length of the course.[i] To show the effects of initial class size on attrition (as discussed in Module Four, Part One) they employ what is now the simplest and most restrictive of sample correction methods, which can be traced to James Heckman (1979), recipient of the 2000 Nobel Prize in Economics.

From Module Four, Part One, we have the data generating process for the difference between post and preTUCE scores for the i^{th} student (Δy_i):

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

where the data set of explanatory variables is matrix $\mathbf{X}$, 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. Sample selection associated with students' unwillingness to take the postteest (dropping the course) results in population error term and regressor correlation that biases and makes coefficient estimators in this change score model inconsistent.

The data generating process for the i^{th} student's propensity to take the posttest is:

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

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.

$\mathbf{T}^*$ is the vector of all students' propensities to take a posttest.

$\mathbf{H}$ is the matrix of explanatory variables that are believed to drive these propensities.

$\boldsymbol{\alpha}$ is the vector of slope coefficients corresponding to these observable variables.

$\boldsymbol{\omega}$ is the vector of unobservable random shocks that affect each student's propensity.

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 a Heckman adjustment for the resulting bias caused by excluding those students from the change-score regression requires estimation of equation (2) and the calculation of an inverse Mill's ratio for each student who has a pretest. This inverse Mill's ratio is then added to the change-score regression (1) as another explanatory variable. In essence, this inverse Mill's ratio adjusts the error term for the missing students.

For the Heckman adjustment for sample selection each disturbance 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:

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

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

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

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 regression for this censored sample of $n_{T=1}$ students who took the posttest is now:

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

which suggests the Heckman adjusted regression to be estimated:

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

where λ_i is the inverse Mill's ratio (or hazard) such that $\lambda_i = f(-T_i^*)/[1 - F(-T_i^*)]$, and $f(.)$ and $F(.)$ are the normal density and distribution functions. 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.

STATA's built-in "heckman" command estimates both the selection and outcome equation using either the full-information maximum likelihood or Heckman's original two-step estimator (which uses the Mills ratio as a regressor). The default "heckman" command implements the maximum likelihood estimation, including ρ and σ_ε, and is written:

```
heckman change hb doc comp lib ci ck phd noeval, ///
select (final = an hb doc comp lib ci ck phd noeval) vce(opg)
```

while the Mills ratio two-step process can be implemented by specifying the option "twostep" after the command. The option "vce(opg)" specifies the outer-product of the gradient method to estimate standard errors, as opposed to STATA's default Hessian method.

As described in Module One, Part Three, entering all of these commands into the command window in STATA and pressing enter (or alternatively, highlighting the commands in a do file and pressing ctrl-d) yields the following output file:

```
. insheet ///
> A1 A2 X3 C AL AM AN CA CB CC CH CI CJ CK CL CM CN CO CS CT ///
> CU CV CW DB DD DI DJ DK DL DM DN DQ DR DS DY DZ EA EB EE EF ///
> EI EJ EP EQ ER ET EY EZ FF FN FX FY FZ GE GH GM GN GQ GR HB ///
> HC HD HE HF ///
> using "C:\BECK8WO.csv", comma
(64 vars, 2837 obs)


. recode a2 (100/199=1) (200/299=2) (300/399=3) (400/499=4)
(a2: 2837 changes made)


. gen doc=(a2==1) if a2!=.
. gen comp=(a2==2) if a2!=.
. gen lib=(a2==3) if a2!=.
. gen twoyr=(a2==4) if a2!=.
. gen phd=(dj==3) if dj!=.
. gen final=(cc>0) if cc!=.
. gen noeval=(ge+gh+gm+gq==-36)
. gen change=cc-an
. recode hb (90=89)
(hb: 96 changes made)


. drop if an==-9 | hb==-9 | ci==-9 | ck==-9 | cs==0 | cs==-9 | a2==-9 |
phd==-9
(250 observations deleted)
```

```
. heckman change hb doc comp lib ci ck phd noeval, select (final = an hb doc
comp lib ci ck phd noeval) vce(opg)

Iteration 0:   log likelihood =  -6826.563
Iteration 1:   log likelihood = -6826.4685
Iteration 2:   log likelihood = -6826.4674
Iteration 3:   log likelihood = -6826.4674

Heckman selection model                         Number of obs      =      2587
(regression model with sample selection)        Censored obs       =       510
                                                Uncensored obs     =      2077

                                                Wald chi2(8)       =    211.39
Log likelihood = -6826.467                      Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |                 OPG
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
change       |
          hb |  -.0097802   .0055923    -1.75   0.080    -.0207408    .0011805
         doc |   1.997291   .5534814     3.61   0.000      .912487    3.082094
        comp |   -.361983   .4332653    -0.84   0.403    -1.211167    .4872015
         lib |    2.23154    .505341     4.42   0.000      1.24109     3.22199
          ci |   .3940114   .2533859     1.55   0.120    -.1026158    .8906386
          ck |  -2.743372   .3803107    -7.21   0.000    -3.488767   -1.997976
         phd |   .6420888   .2896418     2.22   0.027     .0744013    1.209776
      noeval |  -.6320101   1.269022    -0.50   0.618    -3.119248    1.855227
       _cons |   6.817536   .7238893     9.42   0.000     5.398739    8.236332
-------------+----------------------------------------------------------------
final        |
          an |   .0227793    .009396     2.42   0.015     .0043634    .0411953
          hb |  -.0048868   .0020624    -2.37   0.018     -.008929   -.0008447
         doc |   .9715436    .150756     6.44   0.000     .6760672     1.26702
        comp |   .4043055   .1443272     2.80   0.005     .1214295    .6871815
         lib |   .5150521   .1908644     2.70   0.007     .1409648    .8891394
          ci |   .1992685   .0905382     2.20   0.028     .0218169      .37672
          ck |   .0859013   .1190223     0.72   0.470    -.1473781    .3191808
         phd |  -.1320764   .0978678    -1.35   0.177    -.3238939     .059741
      noeval |  -1.929021   .0713764   -27.03   0.000    -2.068916   -1.789126
       _cons |   .9901789    .240203     4.12   0.000     .5193897    1.460968
-------------+----------------------------------------------------------------
     /athrho |   .0370755   .3578813     0.10   0.917    -.6643589      .73851
    /lnsigma |   1.471813   .0160937    91.45   0.000      1.44027    1.503356
-------------+----------------------------------------------------------------
         rho |   .0370585   .3573898                     -.581257    .6282441
       sigma |   4.357128   .0701223                     4.221836    4.496756
      lambda |   .1614688    1.55763                     -2.89143    3.214368
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0):   chi2(1) =     0.03   Prob > chi2 = 0.8612
------------------------------------------------------------------------------
```

The above output provides maximum likelihood estimation of both the probit equation and the change score equation with separate estimation of ρ and σ_ε. The bottom panel provides the probit coefficients for the propensity equation, where it is shown that initial class size is negatively and significantly related to the propensity to take the posttest with a one-tail p value of 0.009. The tob panel gives the change score results, where initial class size is negatively and

significantly related to the change score with a one-tail p value of 0.04. Again, it takes approximately 100 students to move the change score in the opposite direction by a point.

Alternatively, the following command estimates the Heckman model using the Mills ratio as a regressor:

```
. heckman change hb doc comp lib ci ck phd noeval, select (final = an hb doc comp lib
ci ck phd noeval) twostep


Heckman selection model -- two-step estimates   Number of obs      =      2587
(regression model with sample selection)        Censored obs       =       510
                                                Uncensored obs     =      2077

                                                Wald chi2(16)      =    931.46
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
change       |
          hb |  -.0102219   .0056305    -1.82   0.069    -.0212575    .0008137
         doc |   2.079684   .5764526     3.61   0.000     .9498578     3.20951
        comp |   -.329457   .4426883    -0.74   0.457     -1.19711    .5381962
         lib |   2.274478   .5373268     4.23   0.000     1.221337    3.327619
          ci |   .4082326   .2592943     1.57   0.115    -.0999749    .9164401
          ck |  -2.730737    .377552    -7.23   0.000    -3.470725   -1.990749
         phd |   .6334483   .2910392     2.18   0.030      .063022    1.203875
      noeval |  -.8843357   1.272225    -0.70   0.487    -3.377851     1.60918
       _cons |   6.741226   .7510686     8.98   0.000     5.269159    8.213293
-------------+----------------------------------------------------------------
final        |
          an |    .022039   .0094752     2.33   0.020      .003468      .04061
          hb |  -.0048826   .0019241    -2.54   0.011    -.0086537   -.0011114
         doc |   .9757148   .1463617     6.67   0.000     .6888511    1.262578
        comp |   .4064945   .1392651     2.92   0.004       .13354     .679449
         lib |   .5214436   .1766459     2.95   0.003      .175224    .8676632
          ci |   .1987315   .0916865     2.17   0.030     .0190293    .3784337
          ck |     .08779   .1342874     0.65   0.513    -.1754085    .3509885
         phd |   -.133505   .1030316    -1.30   0.195    -.3354433    .0684333
      noeval |  -1.930522   .0723911   -26.67   0.000    -2.072406   -1.788638
       _cons |   .9953498   .2432624     4.09   0.000     .5185642    1.472135
-------------+----------------------------------------------------------------
mills        |
      lambda |   .4856741   1.596833     0.30   0.761    -2.644061     3.61541
-------------+----------------------------------------------------------------
         rho |    0.11132
       sigma |  4.3630276
      lambda |  .48567415   1.596833
------------------------------------------------------------------------------
```

The estimated probit model (in the bottom portion of the above output) is

Estimated propensity to take the posttest = 0.995 + 0.022*(preTUCE score)*

− 0 .005*(initial class size)* + 0.976*(Doctoral Institution)*

+ 0.406 *(Comprehensive Institution)* + 0.521*(Liberal Arts Institution)*

+ 0.199 *(Male instructor)* + 0.0878*(English Instructor Native Language)*

− 0.134*(Instructor has PhD)* − 1.930*(No Evaluation of Instructor)*

The beginning or initial class size is negatively and highly significantly related to the propensity to take the posttest, with a one-tail p value of 0.011.

The corresponding change-score equation employing the inverse Mills ratio is in the upper portion of the above output:

Predicted Change = 6.741 − 0.010(initial class size) + 2.080(Doctoral Institution)

− 0.329 (Comprehensive Institution) + 2.274 Liberal Arts Institution)

+ .408(Male instructor) − 2.731(English Instructor Native Language)

+ 0.633(Instructor has PhD) − 0.88434(No Evaluation of Instructor) + 0 .486 λ

The change score is negatively and significantly related to the class size, with a one-tail p value of 0.0345, but it takes an additional 100 students to lower the change score by a point.

AN APPLICATION OF PROPENSITY SCORE MATCHING

Unfortunately, we are not aware of a study in economic education for which propensity score matching has been used. Thus, we looked outside economic education and elected to redo the example reported in Becker and Ichino (2002). This application and data are derived from Dehejia and Wahba (1999), whose study, in turn was based on LaLonde (1986). The data set consists of observed samples of treatments and controls from the National Supported Work demonstration. Some of the institutional features of the data set are given by Becker and Ichino. The data were downloaded from the website http://www.nber.org/~rdehejia/nswdata.html. The data set used here is in the original text form, contained in the data file "matchingdata.txt." They have been assembled from the several parts in the NBER archive.

Becker and Ichino report that they were unable to replicate Dehejia and Wahba's results, though they did obtain similar results. (They indicate that they did not have the original authors' specifications of the number of blocks used in the partitioning of the range of propensity scores, significance levels, or exact procedures for testing the balancing property.) In turn, we could not precisely replicate Becker and Ichino's results – we can identify the reason, as discussed below. Likewise, however, we obtain similar results.

There are 2,675 observations in the data set, 2,490 controls (with t = 0) and 185 treated observations (with t = 1). The variables in the raw data set are

t = treatment dummy variable
age = age in years
educ = education in years
black = dummy variable for black
hisp = dummy variable for Hispanic
marr = dummy variable for married
nodegree = dummy for no degree (not used)
*re*74 = real earnings in 1974
*re*75 = real earnings in 1975
*re*78 = real earnings in 1978 – the outcome variable

We will analyze these data following Becker and Ichino's line of analysis. We assume that you have completed Module One, Part Three, and thus are familiar with placing commands in the command window or in a do file. In what follows, we will simply show the commands you need to enter into STATA to produce the results that we will discuss.

First, note that STATA does not have a default command available for propensity score matching. Becker and Ichino, however, have created the user-written routine *pscore* that implements the propensity score matching analysis underlying Becker and Ichino (2002). As described in the endnotes of Module Two, Part Three, users can install the *pscore* routine by typing *findit pscore* into the command window, where a list of information and links to download this routine appears. Click on one of the download links and STATA automatically

downloads and installs the routine for use. Users can then access the documentation for this routine by typing *help pscore*. Installing the *pscore* routine also downloads and installs several other routines useful for analyzing treatment effects (i.e., the routines *attk*, *attnd* and *attr*, discussed later in this Module).

To begin the analysis, the data are imported by using the command (where the data file is on the C drive but your data could be placed wherever):

```
insheet ///
t age educ black hisp marr nodegree re74 re75 re78 ///
using "C:\matchingdata.txt"
```

Transformed variables added to the equation are

age2 = age squared
educ2 = educ squared
re742 = re74 squared
re752 = re75 squared
blacku74 = black times 1(re74 = 0)

In order to improve the readability of some of the reported results, we have divided the income variables by 10,000. (This is also an important adjustment that accommodates a numerical problem with the original data set. This is discussed below.) The outcome variable is re78.

The data are set up and described first. The transformations used to create the transformed variables are

```
gen age2=age^2
gen educ2=educ^2
replace re74=re74/10000
replace re75=re75/10000
replace re78=re78/10000
gen re742=re74^2
gen re752=re75^2
gen blacku74=black*(re74==0)
global X age age2 educ educ2 marr black hisp re74 re75 re742 re752 blacku74
```

The data are described with the following statistics:

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           t |      2675    .0691589    .2537716          0          1
         age |      2675    34.22579    10.49984         17         55
        educ |      2675    11.99439    3.053556          0         17
       black |      2675    .2915888    .4545789          0          1
        hisp |      2675    .0343925    .1822693          0          1
-------------+--------------------------------------------------------
        marr |      2675    .8194393    .3847257          0          1
    nodegree |      2675    .3330841    .4714045          0          1
        re74 |      2675       1.823    1.372225          0   13.71487
        re75 |      2675    1.785089    1.387778          0   15.66532
        re78 |      2675    2.050238    1.563252          0   12.11736
-------------+--------------------------------------------------------
        age2 |      2675     1281.61    766.8415        289       3025
       educ2 |      2675    153.1862    70.62231          0        289
       re742 |      2675    5.205628    8.465888          0   188.0976
       re752 |      2675    5.111751    8.908081          0   245.4024
    blacku74 |      2675    .0549533    .2279316          0          1
```

We next fit the logit model for the propensity scores. An immediate problem arises with the data set as used by Becker and Ichino. The income data are in raw dollar terms – the mean of re74, for example is $18,230.00. The square of it, which is on the order of 300,000,000, as well as the square of re75 which is similar, is included in the logit equation with a dummy variable for Hispanic which is zero for 96.5% of the observations and the blacku74 dummy variable which is zero for 94.5% of the observations. Because of the extreme difference in magnitudes, estimation of the logit model in this form is next to impossible. But rescaling the data by dividing the income variables by 10,000 addresses the instability problem. These transformations are shown in the *replace* commands above. This has no impact on the results produced with the data, other than stabilizing the estimation of the logit equation.

The following command estimates the logit model from which the propensity scores are obtained and tests the balancing hypothesis. The logit model from which the propensity scores are obtained is fit using:[ii]

```
. global X age age2 educ educ2 marr black hisp re74 re75 re742 re752 blacku74
. pscore t $X, logit pscore(_pscore) blockid(_block) comsup
```

where the *logit* option specifies that propensity scores should be estimated using the logit model, the *blockid* and *pscore* options define two new variables created by STATA representing each observation's propensity score and block id, and the *comsup* option restricts the analysis to observations in the common support.

(Note: Becker and Ichino's coefficients on re74 and re75 are multiplied by 10,000, and coefficients on re742 and re752 are multiplied by 100,000,000. Otherwise, the output presented here matches that of Becker and Ichino)

```
. pscore t $X, logit pscore(_pscore) blockid(_block) comsup


**************************************************
Algorithm to estimate the propensity score
**************************************************


The treatment is t

          t |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,490       93.08       93.08
          1 |        185        6.92      100.00
------------+-----------------------------------
      Total |      2,675      100.00


Estimation of the propensity score

Iteration 0:   log likelihood = -672.64954
Iteration 1:   log likelihood = -506.34385
Iteration 2:   log likelihood = -385.59357
Iteration 3:   log likelihood = -253.47057
Iteration 4:   log likelihood = -239.00944
Iteration 5:   log likelihood = -216.46206
Iteration 6:   log likelihood = -209.42835
Iteration 7:   log likelihood = -205.15188
Iteration 8:   log likelihood = -204.97706
Iteration 9:   log likelihood = -204.97537
Iteration 10:  log likelihood = -204.97536
Iteration 11:  log likelihood = -204.97536

Logistic regression                               Number of obs   =       2675
                                                  LR chi2(12)     =     935.35
                                                  Prob > chi2     =     0.0000
Log likelihood = -204.97536                       Pseudo R2       =     0.6953

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3316903   .1203299     2.76   0.006     .0958482    .5675325
        age2 |  -.0063668   .0018554    -3.43   0.001    -.0100033   -.0027303
        educ |    .849268   .3477058     2.44   0.015     .1677771    1.530759
       educ2 |  -.0506202   .0172493    -2.93   0.003    -.0844282   -.0168122
        marr |  -1.885542   .2993309    -6.30   0.000    -2.472219   -1.298864
       black |   1.135972   .3517854     3.23   0.001     .4464852    1.825459
        hisp |    1.96902   .5668594     3.47   0.001      .857996    3.080044
        re74 |  -1.058961   .3525178    -3.00   0.003    -1.749883   -.3680387
        re75 |  -2.168541   .4142324    -5.24   0.000    -2.980422    -1.35666
       re742 |   .2389164   .0642927     3.72   0.000      .112905    .3649278
       re752 |   .0135926   .0665375     0.20   0.838    -.1168185    .1440038
    blacku74 |    2.14413   .4268152     5.02   0.000     1.307588    2.980673
       _cons |  -7.474743   2.443511    -3.06   0.002    -12.26394    -2.68555
------------------------------------------------------------------------------
Note: 22 failures and 0 successes completely determined
```

```
Note: the common support option has been selected
The region of common support is [.00061066, .97525407]
Description of the estimated propensity score
in region of common support

                 Estimated propensity score
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0006426       .0006107
 5%     .0008025       .0006149
10%     .0010932       .0006159       Obs                1342
25%     .0023546        .000618       Sum of Wgt.        1342

50%     .0106667                      Mean           .1377463
                        Largest       Std. Dev.      .2746627
75%     .0757115        .974804
90%     .6250822       .9749805       Variance       .0754396
95%      .949302       .9752243       Skewness       2.185181
99%      .970598       .9752541       Kurtosis       6.360726
```

The next set of results summarizes the tests of the balancing hypothesis. By specifying the *detail* option in the above *pscore* command, the routine will also report the separate results of the *F* tests within the partitions as well as the details of the full partition itself. The balancing hypothesis is rejected when the p value is less than 0.01 within the cell. Becker and Ichino do not report the results of this search for their data, but do report that they ultimately found seven blocks. They do not report the means by which the test of equality is carried out within the blocks or the critical value used.

```
**************************************************
Step 1: Identification of the optimal number of blocks
Use option detail if you want more detailed output
**************************************************

The final number of blocks is 7

This number of blocks ensures that the mean propensity score
is not different for treated and controls in each blocks
```

```
**********************************************************
Step 2: Test of balancing property of the propensity score
Use option detail if you want more detailed output
**********************************************************

Variable black is not balanced in block 1

The balancing property is not satisfied

Try a different specification of the propensity score
```

```
  Inferior |
  of block |           t
of pscore  |         0          1 |     Total
-----------+----------------------+----------
         0 |       924          7 |       931
       .05 |       102          4 |       106
        .1 |        56          7 |        63
        .2 |        41         28 |        69
        .4 |        14         21 |        35
        .6 |        13         20 |        33
        .8 |         7         98 |       105
-----------+----------------------+----------
     Total |     1,157        185 |     1,342

Note: the common support option has been selected


****************************************
End of the algorithm to estimate the pscore
****************************************
```

The final portion of the *pscore* output presents the blocks used for the balancing hypothesis. Again, specifying the *detail* option will report the results of the balancing property test for each of the independent variables, which are excluded here for brevity. This part of the analysis also recommends that the analyst reexamine the specification of the propensity score model. Because this is not a numerical problem, the analysis continues with estimation of the average treatment effect on the treated.

The first example below shows estimation using the kernel estimator to define the counterpart observation from the controls and using only the subsample in the common support.[iii] This stage consists of *nboot* + 1 iterations. In order to be able to replicate the results, we set the seed of the random number generator before computing the results:

```
set seed 1234579
attk re78 t $X, pscore(_pscore) bootstrap comsup reps(25)
```

Recall, we divided the income values by 10,000. The value of .153795 reported below thus corresponds to $1,537.95. Becker and Ichino report a value (see their section 6.4) of $1,537.94. Using the bootstrap replications, we have estimated the asymptotic standard error to be $856.28. A 95% confidence interval for the treatment effect is computed using $1537.95 ± 1.96(856.27) = (-$229.32,$3,305.22).

```
. attk re78 t $X, pscore(_pscore) bootstrap comsup reps(25)

 The program is searching for matches of each treated unit.
 This operation may take a while.
```

```
ATT estimation with the Kernel Matching method

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      185        1157       0.154           .           .

------------------------------------------------------------
Note: Analytical standard errors cannot be computed. Use
the bootstrap option to get bootstrapped standard errors.



Bootstrapping of standard errors

command:      attk re78 t age age2 educ educ2 marr black hisp re74 re75 re742 re752
blacku74 , pscore(_pscore) comsup bwidth(.06)
statistic:    attk       = r(attk)
note: label truncated to 80 characters

Bootstrap statistics                                Number of obs    =      2675
                                                    Replications     =        25

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        attk |    25  .1537945 -.0050767  .0856277  -.0229324   .3305215   (N)
             |                                      -.0308111    .279381   (P)
             |                                      -.0308111   .2729317  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected



ATT estimation with the Kernel Matching method
Bootstrapped standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      185        1157       0.154       0.086       1.796

------------------------------------------------------------
```

Note that the estimated asymptotic standard error is somewhat different. As we noted earlier, because of differences in random number generators, the bootstrap replications will differ across programs. It will generally not be possible to exactly replicate results generated with different computer programs. With a specific computer program, replication is obtained by setting the seed of the random number generator. (The specific seed chosen is immaterial, so long as the same seed is used each time.)

The next set of estimates is based on all of the program defaults. The single nearest neighbor is used for the counterpart observation; 25 bootstrap replications are used to compute the standard deviation, and the full range of propensity scores (rather than the common support) is used. Intermediate output is

also suppressed. Once again, we set the seed for the random number generator before estimation. In this case, the *pscore* calculation is not used, and we have instead estimated the nearest neighbor matching and the logit propensity scores in the same command sequence by specifying the *logit* option rather than the *pscore* option. Skipping the *pscore* routine essentially amounts to ignoring any test of the balancing hypothesis. For the purposes of this Module, this is a relatively innocuous simplification, but in practice, the *pscore* routine should always be used prior to estimating the treatment effects.

```
. attnd re78 t $X, logit bootstrap reps(25)


 The program is searching the nearest neighbor of each treated unit.
 This operation may take a while.



ATT estimation with Nearest Neighbor Matching method
(random draw version)
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT    Std. Err.          t
------------------------------------------------------------

      185          57       0.167        0.211      0.789

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches



Bootstrapping of standard errors

command:      attnd re78 t age age2 educ educ2 marr black hisp re74 re75 re742 re752
blacku74 , pscore() logit
statistic:    attnd      = r(attnd)
note: label truncated to 80 characters

Bootstrap statistics                              Number of obs    =      2675
                                                  Replications     =        25

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
       attnd |    25  .1667644   .012762  .1160762  -.0728051   .4063339   (N)
             |                                      -.1111108   .3704965   (P)
             |                                      -.1111108   .2918935  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected
```

```
ATT estimation with Nearest Neighbor Matching method
(random draw version)
Bootstrapped standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

      185          57       0.167       0.116       1.437

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
nearest neighbour matches
```

Using the full sample in this fashion produces an estimate of $1,667.64 for the treatment effect with an estimated standard error of $1,160.76. In comparison, using the 1,342 observations in their estimated common support, and the same 185 treated observations, Becker and Ichino reported estimates of $1,667.64 and $2,113.59 for the effect and the standard error, respectively and use 57 of the 1,342 controls as nearest neighbors.

The next set of results uses the radius form of matching and again restricts attention to the estimates in the common support.

```
. attr re78 t $X, logit bootstrap comsup radius(0.0001) reps(25)


The program is searching for matches of treated units within radius.
 This operation may take a while.


ATT estimation with the Radius Matching method
Analytical standard errors

------------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.           t
------------------------------------------------------------

       23          66      -0.555       0.239      -2.322

------------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

```
Bootstrapping of standard errors

command:      attr re78 t age age2 educ educ2 marr black hisp re74 re75 re742 re752
blacku74 , pscore() logit comsu
> p radius(.0001)
statistic:    attr       = r(attr)
note: label truncated to 80 characters

Bootstrap statistics                                Number of obs    =      2675
                                                    Replications     =        25

------------------------------------------------------------------------------
Variable     |  Reps  Observed      Bias  Std. Err. [95% Conf. Interval]
-------------+----------------------------------------------------------------
        attr |    25  -.554614 -.0043318  .5369267  -1.662776   .5535483   (N)
             |                                       -1.64371    .967416   (P)
             |                                      -1.357991    .967416  (BC)
------------------------------------------------------------------------------
Note:  N   = normal
       P   = percentile
       BC  = bias-corrected


ATT estimation with the Radius Matching method
Bootstrapped standard errors

-------------------------------------------------------
n. treat.   n. contr.         ATT   Std. Err.          t
-------------------------------------------------------

       23          66      -0.555       0.537      -1.033

-------------------------------------------------------
Note: the numbers of treated and controls refer to actual
matches within radius
```

The estimated treatment effects are now very different. We see that only 23 of the 185 treated observations had a neighbor within a range (radius in the terminology of Becker and Ichino) of 0.0001. Consistent with Becker and Ichino's results, the treatment effect is estimated to be -$5,546.14 with a standard error of $5,369.27. Becker and Ichino state that that these nonsensical results illustrate both the differences in "caliper" vesus "radius" matching as well as the sensitivity of the estimator to the choice of radius. In order to implement a true caliper matching process, the user-written *psmatch2* routine should be used.

After installing the *psmatch2* routine, caliper matching with logit propensity scores and common support can be implemented with the following command:

```
. psmatch2 t $X, common logit caliper(0.0001) outcome(re78)

Logistic regression                                     Number of obs   =       2675
                                                        LR chi2(12)     =     935.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -204.97536                             Pseudo R2       =     0.6953

------------------------------------------------------------------------------
           t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3316903   .1203299     2.76   0.006     .0958482    .5675325
        age2 |  -.0063668   .0018554    -3.43   0.001    -.0100033   -.0027303
        educ |    .849268   .3477058     2.44   0.015     .1677771    1.530759
       educ2 |  -.0506202   .0172493    -2.93   0.003    -.0844282   -.0168122
        marr |  -1.885542   .2993309    -6.30   0.000    -2.472219   -1.298864
       black |   1.135972   .3517854     3.23   0.001     .4464852    1.825459
        hisp |    1.96902   .5668594     3.47   0.001      .857996    3.080044
        re74 |  -1.058961   .3525178    -3.00   0.003    -1.749883   -.3680387
        re75 |  -2.168541   .4142324    -5.24   0.000    -2.980422    -1.35666
       re742 |   .2389164   .0642927     3.72   0.000      .112905    .3649278
       re752 |   .0135926   .0665375     0.20   0.838    -.1168185    .1440038
    blacku74 |    2.14413   .4268152     5.02   0.000     1.307588    2.980673
       _cons |  -7.474743   2.443511    -3.06   0.002    -12.26394    -2.68555
------------------------------------------------------------------------------
Note: 22 failures and 0 successes completely determined.
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
-------------------------------------------------------------------------------
      Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+-------------------------------------------------------
          re78  Unmatched | .634914353    2.1553921  -1.52047775   .115461434   -13.17
                      ATT | .672171543   .443317968   .228853575   .438166333     0.52
-----------------------------+-------------------------------------------------------
Note: S.E. for ATT does not take into account that the propensity score is estimated.

 psmatch2: |   psmatch2: Common
 Treatment |        support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |         0      2,490 |     2,490
   Treated |       162         23 |       185
-----------+----------------------+----------
     Total |       162      2,513 |     2,675
```

The "difference" column in the "ATT" row of the above results presents the estimated treatment effect. Using a true caliper matching process, the estimates of $2,228.85 and $4,381.66 for the effect and the standard error, respectively, are much more comparable to the results previously obtained.

CONCLUDING COMMENTS

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. On the other hand, 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. Someone using Heckman's original selection adjustment method can easily have their results replicated in LIMDEP, STATA and SAS. Such is not the case with propensity score matching. Propensity score matching results are highly sensitive to the computer program employed while Heckman's original sample selection adjustment method can be relied on to give comparable results across programs.

REFERENCES

Becker, William and William Walstad. "Data Loss From Pretest to Posttest As a Sample Selection Problem," *Review of Economics and Statistics*, Vol. 72, February 1990: 184-188,

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

Becker, S. and A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, Vol. 2, November 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*, Vol. 94, 1999: 1052-1062.

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

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).

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

Saunders, Phillip. The TUCE III Data Set: Background information and file codes (documentation, summary tables, and five 3.5-inch double-sided, high density disks in ASCII format). New York: National Council on Economic Education, 1994.

ENDNOTES

[i] Huynh, Jacho-Chavez, and Self (2010) have a data set that enables them to account for selection into, out of and between collaborative learning sections of a large principles course in their change-score modeling.

[ii] Users can also estimate the logit model with STATA's default logit command. The predicted probabilities from the logit estimation are equivalent to the propensity scores automatically provided with the *pscore* command. Since STATA does not offer any default matching routine to use following the default logit command, we adopt the use of the pscore routine (the download of which includes several matching routines to calculate treatment effects). The *pscore* routine also tests the balancing hypothesis and provides other relevant information for propensity score matching which is not provided by the default logit command.

[iii] The Kernel density estimator is a *nonparametric* estimator. Unlike a parametric estimator (which is an equation), a non-parametric estimator has no fixed structure and is based on a histogram of all the data. Histograms are bar charts, which are not smooth, and whose shape depends on the width of the bin into which the data are divided. In essence, with a fixed bin width, the kernel estimator smoothes out the histogram by centering each of the bins at each data point rather than fixing the end points of the bin. The optimum bin width is a subject of debate and well beyond the technical level of this module.

MODULE FOUR, PART FOUR: SAMPLE SELECTION

IN ECONOMIC EDUCATION RESEARCH USING SAS

Part Four of Module Four provides a cookbook-type demonstration of the steps required to use SAS in situations involving estimation problems associated with sample selection. Unlike LIMDEP and STATA, SAS does not have a procedure or macro available from SAS Institute specifically designed to match observations using propensity scores. There are a few user-written codes but these are not well suited to replicate the particular type of sample-selection problems estimated in LIMDEP and STATA. As such, this segment will go as far as SAS permits in replicating what is done in Parts Two and Three in LIMDEP and STATA. Users of this model need to have completed Module One, Parts One and Four, but not necessarily Modules Two and Three. From Module One users are assumed to know how to get data into SAS, recode and create variables within SAS, and run and interpret regression results. Module Four, Parts Two and Three demonstrate in LIMDEP and STATA what is done here in SAS.

THE CASE, DATA, AND ROUTINE FOR EARLY HECKMAN ADJUSTMENT

The change score or difference in difference model is used extensively in education research. Yet, before Becker and Walstad (1990), little if any attention was given to the consequence of missing student records that result from: 1) "data cleaning" done by those collecting the data, 2) student unwillingness to provide data, or 3) students self-selecting into or out of the study. The implications of these types of sample selection are shown in the work of Becker and Powers (2001) where the relationship between class size and student learning was explored using the third edition of the Test of Understanding in College Economics (TUCE), which was produced by Saunders (1994) for the National Council on Economic Education (NCEE), since renamed the Council for Economic Education.

Module One, Part Four showed how to get the Becker and Powers data set "beck8WO.csv" into SAS. As a brief review this was done with the read command:

```
data BECPOW;
infile 'C:\Users\gregory.gilpin\Desktop\BeckerWork\BECK8WO.CSV'
delimiter = ',' MISSOVER DSD lrecl=32767 ;
informat A1 best32.; informat A2 best32.; informat X3 best32.;
informat C best32. ; informat AL best32.; informat AM best32.;
informat AN best32.; informat CA best32.; informat CB best32.;
informat CC best32.; informat CH best32.; informat CI best32.;
informat CJ best32.; informat CK best32.; informat CL best32.;
informat CM best32.; informat CN best32.; informat CO best32.;
informat CS best32.; informat CT best32.; informat CU best32.;
informat CV best32.; informat CW best32.; informat DB best32.;
```

```
informat DD best32.; informat DI best32.; informat DJ best32.;
informat DK best32.; informat DL best32.; informat DM best32.;
informat DN best32.; informat DQ best32.; informat DR best32.;
informat DS best32.; informat DY best32.; informat DZ best32.;
informat EA best32.; informat EB best32.; informat EE best32.;
informat EF best32.; informat EI best32.; informat EJ best32.;
informat EP best32.; informat EQ best32.; informat ER best32.;
informat ET best32.; informat EY best32.; informat EZ best32.;
informat FF best32.; informat FN best32.; informat FX best32.;
informat FY best32.; informat FZ best32.; informat GE best32.;
informat GH best32.; informat GM best32.; informat GN best32.;
informat GQ best32.; informat GR best32.; informat HB best32.;
informat HC best32.; informat HD best32.; informat HE best32.;
informat HF best32.;

format A1 best12.; format A2 best12.; format X3 best12.;
format C best12. ; format AL best12.; format AM best12.;
format AN best12.; format CA best12.; format CB best12.;
format CC best12.; format CH best12.; format CI best12.;
format CJ best12.; format CK best12.; format CL best12.;
format CM best12.; format CN best12.; format CO best12.;
format CS best12.; format CT best12.; format CU best12.;
format CV best12.; format CW best12.; format DB best12.;
format DD best12.; format DI best12.; format DJ best12.;
format DK best12.; format DL best12.; format DM best12.;
format DN best12.; format DQ best12.; format DR best12.;
format DS best12.; format DY best12.; format DZ best12.;
format EA best12.; format EB best12.; format EE best12.;
format EF best12.; format EI best12.; format EJ best12.;
format EP best12.; format EQ best12.; format ER best12.;
format ET best12.; format EY best12.; format EZ best12.;
format FF best12.; format FN best12.; format FX best12.;
format FY best12.; format FZ best12.; format GE best12.;
format GH best12.; format GM best12.; format GN best12.;
format GQ best12.; format GR best12.; format HB best12.;
format HC best12.; format HD best12.; format HE best12.;
format HF best12.;
input
A1 A2 X3 C AL AM AN CA CB CC CH CI CJ CK CL CM CN CO CS CT CU
CV CW DB DD DI DJ DK DL DM DN DQ DR DS DY DZ EA EB EE EF
EI EJ EP EQ ER ET EY EZ FF FN FX FY FZ GE GH GM GN GQ GR HB
HC HD HE HF; run;
```

where

A1: term, where 1= fall, 2 = spring
A2: school code, where 100/199 = doctorate,
200/299 = comprehensive,
300/399 = lib arts,
400/499 = 2 year
hb: initial class size (number taking preTUCE)
hc: final class size (number taking postTUCE)
dm: experience, as measured by number of years teaching
dj: teacher's highest degree, where Bachelors=1, Masters=2, PhD=3
cc: postTUCE score (0 to 30)

an: preTUCE score (0 to 30)
ge: Student evaluation measured interest
gh: Student evaluation measured textbook quality
gm: Student evaluation measured regular instructor's English ability
gq: Student evaluation measured overall teaching effectiveness
ci: Instructor sex (Male = 1, Female = 2)
ck: English is native language of instructor (Yes = 1, No = 0)
cs: PostTUCE score counts toward course grade (Yes = 1, No = 0)
ff: GPA*100
fn: Student had high school economics (Yes = 1, No = 0)
ey: Student's sex (Male = 1, Female = 2)
fx: Student working in a job (Yes = 1, No = 0)

Separate dummy variables need to be created for each type of school (A2), which is done with the following code:

```
if  99 < A2 < 200 then a2 = 1;
if 199 < A2 < 300 then a2 = 2;
if 299 < A2 < 400 then a2 = 3;
if 399 < A2 < 500 then a2 = 4;
doc = 0; comp = 0; lib = 0; twoyr = 0;
if a2 = 1 then doc = 1;
if a2 = 2 then comp = 1;
if a2 = 3 then lib = 3;
if a2 = 4 then twoyr = 4;
```

To create a dummy variable for whether the instructor had a PhD we use

```
phd = 0;
if    dj = 3 then phd = 1;
```

To create a dummy variable for whether the student took the postTUCE we use

```
final = 0;
if cc > 0 then final = 1;
```

To create a dummy variable for whether a student did (noeval = 0) or did not (noeval = 1) complete a student evaluation of the instructor we use

```
evalsum = ge+gh+gm+gq;
noeval= 0;
if evalsum = -36 then noeval = 1;
```

"Noeval" reflects whether the student was around toward the end of the term, attending classes, and sufficiently motivated to complete an evaluation of the instructor. In the Saunder's data set

evaluation questions with no answer where coded -9; thus, these four questions summing to -36 indicates that no questions were answered.

And the change score is created with

```
change = cc - an;
```

Finally, there was a correction for the term in which student record 2216 was incorrectly recorded:

```
if hb = 90 then hb = 89;
```

All of these recoding and create commands are entered into SAS editor file as follows:

```
data becpow;
      set beckpow;
      if  99 < A2 < 200 then a2 = 1;
      if 199 < A2 < 300 then a2 = 2;
      if 299 < A2 < 400 then a2 = 3;
      if 399 < A2 < 500 then a2 = 4;
      doc = 0; comp = 0; lib = 0; twoyr = 0;
      if a2 = 1 then doc = 1;
      if a2 = 2 then comp = 1;
      if a2 = 3 then lib = 1;
      if a2 = 4 then twoyr = 1;
      phd = 0;
      if    dj = 3 then phd = 1;
      final = 0;
      if cc > 0 then final = 1;
      evalsum = ge+gh+gm+gq;
      noeval= 0;
      if evalsum = -36 then noeval = 1;
      change = cc - an;
      if hb = 90 then hb = 89;
run;
```

To remove records with missing data the following is entered:

```
data becpow;
      set beckpow;
      if AN=-9 then delete;
      if HB=-9 then delete;
      if ci=-9 then delete;
      if ck=-9 then delete;
      if cs=0 then delete;
```

```
      if cs=-9 then delete;
      if a2=-9 then delete;
      if phd=-9 then delete;
run;
```

The use of these data entry and management commands will appear in the SAS output file for the equations to be estimated in the next section.

THE PROPENSITY TO TAKE THE POSTTEST AND THE CHANGE SCORE EQUATION

To address attrition-type sample selection problems in change score studies, Becker and Powers first add observations that were dropped during the early stage of assembling data for TUCE III. Becker and Powers do not have any data on students before they enrolled in the course and thus cannot address selection into the course, but to examine the effects of attrition (course withdrawal) they introduce three measures of class size (beginning, ending, and average) and argue that initial or beginning class size is the critical measure for assessing learning over the entire length of the course.[i] To show the effects of initial class size on attrition (as discussed in Module Four, Part One) they employ what is now the simplest and most restrictive of sample correction methods, which can be traced to James Heckman (1979), recipient of the 2000 Nobel Prize in Economics.

From Module Four, Part One, we have the data generating process for the difference between post and preTUCE scores for the i^{th} student (Δy_i):

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

where the data set of explanatory variables is matrix $\mathbf{X}$, 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. Sample selection associated with students' unwillingness to take the postteest (dropping the course) results in population error term and regressor correlation that biases and makes coefficient estimators in this change score model inconsistent.

The data generating process for the i^{th} student's propensity to take the posttest is:

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

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.

$\mathbf{T}^*$ is the vector of all students' propensities to take a posttest.

$\mathbf{H}$ is the matrix of explanatory variables that are believed to drive these propensities.

$\boldsymbol{\alpha}$ is the vector of slope coefficients corresponding to these observable variables.

$\boldsymbol{\omega}$ is the vector of unobservable random shocks that affect each student's propensity.

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 a Heckman adjustment for the resulting bias caused by excluding those students from the change-score regression requires estimation of equation (2) and the calculation of an inverse Mill's ratio for each student who has a pretest. This inverse Mill's ratio is then added to the change-score regression (1) as another explanatory variable. In essence, this inverse Mill's ratio adjusts the error term for the missing students.

For the Heckman adjustment for sample selection each disturbance 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:

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

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

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

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 regression for this censored sample of $n_{T=1}$ students who took the posttest is now:

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

which suggests the Heckman adjusted regression to be estimated:

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

where λ_i is the inverse Mill's ratio (or hazard) such that $\lambda_i = f(-T_i^*)/[1 - F(-T_i^*)]$, and $f(.)$ and $F(.)$ are the normal density and distribution functions. 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.

The probit command for the selection equation to be estimated in SAS is

```
proc qlim data =becpow;
model final= an hb doc comp lib ci ck phd noeval / discrete;
quit;
```

where the "/ discrete" extension tells SAS to estimated the model by probit.

The command for estimating the adjusted change equation using both the inverse Mills ratio as a regressor and maximum likelihood estimation of the ρ and σ_ε is written

```
proc qlim data=becpow;
model final = an hb doc comp lib ci ck phd noeval / discrete;
model change =   hb doc comp lib ci ck phd noeval / select(final=1);
quit;
```

where the extension " / select (final = 1)" tells SAS that the selection is on observations with the variable final equal to 1.

As described in Module One, Part Four, entering all of these commands into the editor window in SAS and pressing the RUN button yields the following output file:

Discrete Response Profile of final

Index	Value	Frequency	Percent
1	0	510	19.71
2	1	2077	80.29

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	final
Number of Observations	2587
Missing Values	12
Log Likelihood	-822.74107
Maximum Absolute Gradient	0.0000765
Number of Iterations	17
Optimization Method	Newton-Raphson
AIC	1665
Schwarz Criterion	1724

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	922.95	2 * (LogL - LogL0)
Upper Bound of R (U)	2568.4	- 2 * LogL0
Aldrich-Nelson	0.263	R / (R+N)
Cragg-Uhler 1	0.3001	1 - exp(-R/N)
Cragg-Uhler 2	0.4767	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.3573	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.3495	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.3593	R / U
Veall-Zimmermann	0.5278	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.4564	

N = # of observations, K = # of regressors

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	0.995350	0.243263	4.09	<.0001
AN	1	0.022039	0.009475	2.33	0.0200
HB	1	-0.004883	0.001924	-2.54	0.0112
doc	1	0.975715	0.146361	6.67	<.0001
comp	1	0.406495	0.139265	2.92	0.0035
lib	1	0.521444	0.176646	2.95	0.0032
CI	1	0.198732	0.091687	2.17	0.0302
CK	1	0.087790	0.134287	0.65	0.5133
phd	1	-0.133505	0.103032	-1.30	0.1951
noeval	1	-1.930522	0.072391	-26.67	<.0001

The QLIM Procedure

Summary Statistics of Continuous Responses

Variable	N	Mean	Standard Error	Type	Lower Bound
change	2077	5.456909	4.582964	Regular	

Discrete Response Profile of final

Index	Value	Frequency	Percent
1	0	510	19.71
2	1	2077	80.29

Model Fit Summary

Number of Endogenous Variables	2
Endogenous Variable	final change
Number of Observations	2587
Missing Values	12
Log Likelihood	-6826
Maximum Absolute Gradient	0.08290
Number of Iterations	95
Optimization Method	Newton-Raphson
AIC	13695
Schwarz Criterion	13818

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
change.Intercept	1	6.846570	0.685111	9.99	<.0001
change.HB	1	-0.009696	0.005338	-1.82	0.0693
change.doc	1	1.969444	0.471441	4.18	<.0001
change.comp	1	-0.379481	0.422623	-0.90	0.3692
change.lib	1	2.211300	0.508472	4.35	<.0001
change.CI	1	0.385456	0.252689	1.53	0.1272
change.CK	1	-2.749171	0.373792	-7.35	<.0001
change.phd	1	0.649805	0.288707	2.25	0.0244
change.noeval	1	-0.587762	0.768104	-0.77	0.4441
_Sigma.change	1	4.356734	0.067846	64.22	<.0001
final.Intercept	1	0.991434	0.245062	4.05	<.0001
final.AN	1	0.022555	0.010361	2.18	0.0295

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
final.HB	1	-0.004885	0.001924	-2.54	0.0111
final.doc	1	0.973056	0.148459	6.55	<.0001
final.comp	1	0.405223	0.139923	2.90	0.0038
final.lib	1	0.517302	0.180438	2.87	0.0041
final.CI	1	0.199186	0.091770	2.17	0.0300
final.CK	1	0.086526	0.134763	0.64	0.5208
final.phd	1	-0.132627	0.103368	-1.28	0.1995
final.noeval	1	-1.929498	0.072913	-26.46	<.0001
_Rho	1	0.025573	0.211925	0.12	0.9040

The estimated probit model (as found on the top of page 8) is

Estimated propensity to take the posttest = 0.995 + 0.022*(preTUCE score)*

− 0 .005*(initial class size)* + 0.976*(Doctoral Institution)*

+ 0.406 *(Comprehensive Institution)* + 0.521*(Liberal Arts Institution)*

+ 0.199 *(Male instructor)* + 0.0878*(English Instructor Native Language)*

− 0.134*(Instructor has PhD)* − 1.930*(No Evaluation of Instructor)*

The beginning or initial class size is negatively and highly significantly related to the propensity to take the posttest, with a one-tail p value of 0.0056.

The corresponding change-score equation employing the inverse Mills ratio is on page 8-9:

Predicted Change = 6.847 − 0.010(initial class size) + 1.970(Doctoral Institution)

− 0.380 (Comprehensive Institution) + 2.211 Liberal Arts Institution)

+ .386(Male instructor) − 2.749(English Instructor Native Language)

+ 0.650(Instructor has PhD) − 0.588(No Evaluation of Instructor) + 0 .486 λ

The change score is negatively and significantly related to the class size, with a one-tail p value of 0.0347, but it takes an additional 100 students to lower the change score by a point. The maximum likelihood results also contain separate estimates of ρ and σ_{ε}. Note that the coefficients are slightly different then those provided by LIMDEP. This is due to the maximization algorithm of used in proc qlim – that of Newton–Raphson maximization method. Currently SAS does not have any other standard routine to perform Heckman's two-step procedure. It should be noted that there are a few user written codes which can be implemented.

AN APPLICATION OF PROPENSITY SCORE MATCHING

Unfortunately, we are not aware of a study in economic education for which propensity score matching has been used. Thus, we looked outside economic education and elected to redo the example reported in Becker and Ichino (2002). This application and data are derived from Dehejia and Wahba (1999), whose study, in turn was based on LaLonde (1986). The data set consists of observed samples of treatments and controls from the National Supported Work demonstration. Some of the institutional features of the data set are given by Becker and Ichino. The data were downloaded from the website http://www.nber.org/~rdehejia/nswdata.html. The data set used here is in the original text form, contained in the data file "matchingdata.txt." They have been assembled from the several parts in the NBER archive.

Becker and Ichino report that they were unable to replicate Dehejia and Wahba's results, though they did obtain similar results. (They indicate that they did not have the original authors' specifications of the number of blocks used in the partitioning of the range of propensity scores, significance levels, or exact procedures for testing the balancing property.) In turn, we could not precisely replicate Becker and Ichino's results – we can identify the reason, as discussed below. Likewise, however, we obtain similar results.

There are 2,675 observations in the data set, 2490 controls (with t = 0) and 185 treated observations (with t = 1). The variables in the raw data set are

t = treatment dummy variable
age = age in years
educ = education in years
black = dummy variable for black
hisp = dummy variable for Hispanic
marr = dummy variable for married
nodegree = dummy for no degree (not used)
*re*74 = real earnings in 1974
*re*75 = real earnings in 1975
*re*78 = real earnings in 1978 – the outcome variable

We will analyze these data following Becker and Ichino's line of analysis. We assume that you have completed Module One, Part Two, and thus are familiar with placing commands in the editor and using the RUN button to submit commands, and where results are found in the output window. In what follows, we will simply show the commands you need to enter into SAS to produce the results that we will discuss.

To start, the data are imported by using the import wizard. The file is most easily imported by specifying the file as a 'delimited file *.*': When providing the location of the file, click 'options' and then click on the Delimiter 'space' and unclick the box for 'Get variable

names from first row'. In what follows, I call the imported dataset 'match'. As 'match' does not have proper variables names, this is easily corrected using a datastep:

```
data match (keep = t age educ black hisp marr nodegree re74 re75 re78);
      rename var3 = t var5 = age var7 = educ var9 = black var11 = hisp
             var13 = marr var15 = nodegree var17 = re74 var19 = re75
             var21 = re78;
      set match;
            run ;
```

Transformed variables added to the dataset are

age2 = age squared
educ2 = educ squared
re742 = re74 squared
re752 = re75 squared
blacku74 = black times 1(re74 = 0)

In order to improve the readability of some of the reported results, we have divided the income variables by 10,000. (This is also an important adjustment that accommodates a numerical problem with the original data set. This is discussed below.) The outcome variable is re78.

The data are set up and described first. The transformations used to create the transformed variables are

```
data match;
      set match;
age2 = age*age; educ2 = educ*educ;
re74 = re74/10000; re75 = re75/10000; re78 = re78/10000;
re742 = re74*re74; re752 = re75*re75;
blacku74 = black*(re74 = 0);
run;
```

The data are described with the following code and statistics:

```
proc means data = match;
      var t age educ black hisp marr nodegree re74 re75 re78 age2 educ2 re742
          re752 blacku74;
      quit;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
t	2675	0.0691589	0.2537716	0	1.0000000
age	2675	34.2257944	10.4998419	17.0000000	55.0000000
educ	2675	11.9943925	3.0535556	0	17.0000000
black	2675	0.2915888	0.4545789	0	1.0000000
hisp	2675	0.0343925	0.1822693	0	1.0000000
marr	2675	0.8194393	0.3847257	0	1.0000000
nodegree	2675	0.3330841	0.4714045	0	1.0000000
re74	2675	1.8230003	1.3722252	0	13.7148680
re75	2675	1.7850894	1.3877777	0	15.6653230
re78	2675	2.0502376	1.5632520	0	12.1173580
age2	2675	1281.61	766.8415075	289.0000000	3025.00
educ2	2675	153.1861682	70.6223147	0	289.0000000
re742	2675	5.2056281	8.4658880	0	188.0976043
re752	2675	5.1117511	8.9080813	0	245.4023447
blacku74	2675	0.0549533	0.2279316	0	1.0000000

We next fit the logit model for the propensity scores. An immediate problem arises with the data set as used by Becker and Ichino. The income data are in raw dollar terms – the mean of re74, for example is $18,230.00. The square of it, which is on the order of 300,000,000, as well as the square of re75 which is similar, is included in the logit equation with a dummy variable for Hispanic which is zero for 96.5% of the observations and the blacku74 dummy variable which is zero for 94.5% of the observations. Because of the extreme difference in magnitudes, estimation of the logit model in this form is next to impossible. But rescaling the data by dividing the income variables by 10,000 addresses the instability problem.[ii] These transformations are shown in the second set of commands above. This has no impact on the results produced with the data, other than stabilizing the estimation of the logit equation. We are now quite able to replicate the Becker and Ichino results except for an occasional very low order digit.

The logit model from which the propensity scores are obtained is fit using

```
proc qlim data = match;
      model t = age age2 educ educ2 marr black hisp re74 re75 re742 re752
            blacku74 / discrete (dist = logit);
quit;
```

(Note: Becker and Ichino's coefficients on re74 and re75 are multiplied by 10,000, and coefficients on re742 and re752 are multiplied by 100,000,000.)

The QLIM Procedure

Discrete Response Profile of t

Index	Value	Frequency	Percent
1	0	2490	93.08
2	1	185	6.92

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	t
Number of Observations	2675
Missing Values	1
Log Likelihood	-204.97536
Maximum Absolute Gradient	0.0001075
Number of Iterations	184
Optimization Method	Newton-Raphson
AIC	435.95071
Schwarz Criterion	512.54287

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	935.35	2 * (LogL - LogL0)
Upper Bound of R (U)	1345.3	- 2 * LogL0
Aldrich-Nelson	0.2591	R / (R+N)
Cragg-Uhler 1	0.2951	1 - exp(-R/N)
Cragg-Uhler 2	0.7466	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.4499	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.4326	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.6953	R / U
Veall-Zimmermann	0.7742	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.9531	

N = # of observations, K = # of regressors

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	-7.474730	2.433330	-3.07	0.0021
age	1	0.331690	0.119278	2.78	0.0054
age2	1	-0.006367	0.001835	-3.47	0.0005
educ	1	0.849267	0.347572	2.44	0.0145
educ2	1	-0.050620	0.017239	-2.94	0.0033
marr	1	-1.885541	0.299056	-6.30	<.0001
black	1	1.135973	0.351814	3.23	0.0012
hisp	1	1.969023	0.566775	3.47	0.0005
re74	1	-1.058962	0.352476	-3.00	0.0027
re75	1	-2.168542	0.414191	-5.24	<.0001
re742	1	0.238917	0.064275	3.72	0.0002
re752	1	0.013593	0.066518	0.20	0.8381
blacku74	1	2.144130	0.426518	5.03	<.0001

The above results provide the predicted probabilities to be used in matching algorithms. As discussed in the Introduction of this part, SAS does not have such a procedure or macro specifically designed to match observations to estimate treatment effects. We refer the reader to Parts Two and Three of this module to for further understanding on how to implement matching procedures in LIMDEP and STATA.

CONCLUDING COMMENTS

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. On the other hand, 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. Someone using Heckman's original selection adjustment method can easily have their results replicated in LIMDEP, STATA and SAS. Such is not the case with propensity score matching. Propensity score matching results are highly sensitive to the computer program employed while Heckman's original sample selection adjustment method can be relied on to give comparable results across programs.

REFERENCES

Becker, William and William Walstad. "Data Loss From Pretest to Posttest As a Sample Selection Problem," *Review of Economics and Statistics*, Vol. 72, February 1990: 184-188,

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

Becker, S. and A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores," *The Stata Journal*, Vol. 2, November 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*, Vol. 94, 1999: 1052-1062.

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

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).

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

Saunders, Phillip. The TUCE III Data Set: Background information and file codes (documentation, summary tables, and five 3.5-inch double-sided, high density disks in ASCII format). New York: National Council on Economic Education, 1994.

ENDNOTES

[i] Huynh, Jacho-Chavez, and Self (2010) have a data set that enables them to account for selection into, out of and between collaborative learning sections of a large principles course in their change-score modeling.

[ii] An attempt to compute a linear regression of the original RE78 on the original unscaled other variables is successful, but produces a warning that the condition number of the X matrix is 6.5 times 10^9. When the data are scaled as done above, no warning about multicollinearity is given.