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.