

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:\BECK8W02.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 were 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.¹ 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):

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

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 posttest (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:

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

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.

ω 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 ϵ , equation (1), is assumed to be distributed bivariate normal with the corresponding disturbance term in the ω vector of the selection equation (2). Thus, for the i^{th} student we have:

$$(\epsilon_i, \omega_i) \sim \text{bivariate normal}(0, 0, \sigma_\epsilon, I, \rho) \quad (3)$$

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

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

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:

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

which suggests the Heckman adjusted regression to be estimated:

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

where λ_i is the inverse Mill's ratio (or hazard) such that $\lambda_i = f(-T_i^*)/[1 - F(-T_i^*)]$, and $f(\cdot)$ and $F(\cdot)$ 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
                                           Prob > chi2     =     0.0000
```

```
Log likelihood = -6826.467
```

	Coef.	OPG 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 top 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

$$\begin{aligned} \text{Estimated propensity to take the posttest} &= 0.995 + 0.022(\text{preTUCE score}) \\ &- 0.005(\text{initial class size}) + 0.976(\text{Doctoral Institution}) \\ &+ 0.406(\text{Comprehensive Institution}) + 0.521(\text{Liberal Arts Institution}) \\ &+ 0.199(\text{Male instructor}) + 0.0878(\text{English Instructor Native Language}) \\ &- 0.134(\text{Instructor has PhD}) - 1.930(\text{No Evaluation of Instructor}) \end{aligned}$$

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:

$$\begin{aligned} \text{Predicted Change} &= 6.741 - 0.010(\text{initial class size}) + 2.080(\text{Doctoral Institution}) \\ &- 0.329(\text{Comprehensive Institution}) + 2.274(\text{Liberal Arts Institution}) \\ &+ .408(\text{Male instructor}) - 2.731(\text{English Instructor Native Language}) \\ &+ 0.633(\text{Instructor has PhD}) - 0.88434(\text{No Evaluation of Instructor}) + 0.486\lambda \end{aligned}$$

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)
re74 = real earnings in 1974
re75 = real earnings in 1975
re78 = 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
Pseudo R2 = 0.6953
```

Log likelihood = -204.97536

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 of pscore	t		Total
	0	1	
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 $1537.95 \pm 1.96(856.27) = (-\$229.32, \$3,305.22)$.

```
. 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 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" versus "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: Treatment assignment	psmatch2: Common support		Total
	Off suppo	On suppor	
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.
- Dehejia, 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.