

## Selectivity & Treatment – Heckman 2-Step Correction

The data set *select.dta* contains information on a sample of married women taken from the 2003 General Household Survey.

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	4270	47.20047	11.89064	17	69
sex	4270	2	0	2	2
ndepchld	4270	.8386417	1.118618	0	8
hw	2296	884.8653	870.429	.0061538	21875
treated	4270	.5377049	.4986347	0	1
married	4270	1	0	1	1
lhw	2296	6.6076	.6470898	-5.090678	9.9931
educ	4270	12.10023	2.658739	6	35

the variable *treated* takes the value 1 if the women are observed in work and 0 otherwise. Hence the summary statistics show that around 53.8% of the sample of married women are in work. This may be a non-random sample of all married women if there are variables which affect participation in the labour force. If so, then OLS on the sample of working women will be biased and inconsistent.

Suppose we are interested in the determinants of wages for married women. The uncorrected OLS regression on the sample of *working* women is

```
. reg lhw educ age
```

Source	SS	df	MS	Number of obs = 2296		
Model	90.9340802	2	45.4670401	F( 2, 2293) = 119.83		
Residual	870.040215	2293	.379433151	Prob > F = 0.0000		
Total	960.974295	2295	.418725183	R-squared = 0.0946		
				Adj R-squared = 0.0938		
				Root MSE = .61598		
lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0791092	.005117	15.46	0.000	.0690748	.0891435
age	.0037069	.0013281	2.79	0.005	.0011024	.0063113
_cons	5.468114	.0960707	56.92	0.000	5.279719	5.656508

If selectivity exists then these coefficients may not be applicable to *all* married women (working and non-working). To determine whether selection is a problem, first estimate the probability of being in work, (the probability of being treated) as a function of the original control variables and an additional identifying variable – in this case the number of dependent children. This variable is assumed to affect the probability of participation in work (negatively), but is assumed not to influence wages on offer once in work, (in practice this assumption should be tested).

## A probit estimate of the probability of being treated gives

```
. probit treated educ age ndepchld
```

```
Iteration 0:   log likelihood = -2947.5859
Iteration 1:   log likelihood = -2698.3709
Iteration 2:   log likelihood = -2696.5695
Iteration 3:   log likelihood = -2696.5692
```

```
Probit regression                               Number of obs   =       4270
                                                LR chi2(3)      =       502.03
                                                Prob > chi2     =       0.0000
Log likelihood = -2696.5692                    Pseudo R2      =       0.0852
```

treated	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
educ	.0116392	.0077145	1.51	0.131	-.003481	.0267595
age	-.0442952	.0022067	-20.07	0.000	-.0486202	-.0399702
ndepchld	-.1714651	.0219398	-7.82	0.000	-.2144664	-.1284638
_cons	2.19459	.1614445	13.59	0.000	1.878165	2.511015

Remember these coefficients have no direct interpretation (being simply the values that maximize the likelihood function), but the marginal effects =  $\phi(Z\gamma) \gamma_j$  do and are given by

```
. dprobit treated educ age ndepchld
```

```
Iteration 0:   log likelihood = -2947.5859
Iteration 1:   log likelihood = -2698.3709
Iteration 2:   log likelihood = -2696.5695
Iteration 3:   log likelihood = -2696.5692
```

```
Probit regression, reporting marginal effects   Number of obs   =       4270
                                                LR chi2(3)      =       502.03
                                                Prob > chi2     =       0.0000
Log likelihood = -2696.5692                    Pseudo R2      =       0.0852
```

treated	dF/dx	Std. Err.	z	P> z	x-bar	[ 95% C.I. ]	
educ	.0046198	.003062	1.51	0.131	12.1002	-.001382	.010621
age	-.0175815	.0008756	-20.07	0.000	47.2005	-.019298	-.015865
ndepchld	-.0680575	.0087036	-7.82	0.000	.838642	-.085116	-.050999
obs. P	.5377049						
pred. P	.540176 (at x-bar)						

z and P>|z| correspond to the test of the underlying coefficient being 0

So 1 extra year of education raises the probability of being in work by **0.46 percentage points** (not per cent) and 1 extra child reduces the probability of being in work by around 6.8 percentage points. Note that the identifying variable in this case is highly significant, (as it should be if it is to make the technique worthwhile).

The next step is to construct the inverse mills ratio. To do this we need  $\phi(Z\gamma)$  and  $\Phi(Z\gamma)$ . The  $Z\gamma$  (the sum of each variable evaluated at its mean value multiplied by its probit

estimate =  $\bar{z}_1 \hat{\gamma}_1 + \bar{z}_2 \hat{\gamma}_2 + \dots + \bar{z}_k \hat{\gamma}_k$  can be obtained using the following command in stata.

```
. predict zg if e(sample), xb
```

To calculate the standard normal pdf

```
. g phi=normalden(zg)
```

To calculate the standard normal cdf

```
. g PHI=normal(zg)
```

and to calculate lambda

```
. g lambda=phi/PHI
```

The final step is to include this lambda term as an additional regressor in the original model

```
. reg lhw educ age lambda if sex==2 & married==1
```

Source	SS	df	MS			
Model	92.733141	3	30.911047	Number of obs =	2296	
Residual	868.241154	2292	.378813767	F( 3, 2292) =	81.60	
Total	960.974295	2295	.418725183	Prob > F =	0.0000	
				R-squared =	0.0965	
				Adj R-squared =	0.0953	
				Root MSE =	.61548	

  

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	.0772027	.0051871	14.88	0.000	.0670308	.0873746
age	.0103897	.0033414	3.11	0.002	.0038373	.0169421
lambda	-.3269318	.1500193	-2.18	0.029	-.6211196	-.032744
_cons	5.418822	.098621	54.95	0.000	5.225426	5.612218

Can see that the lambda term is significant and negatively signed – which suggests that the error terms in the selection and primary equations are negatively correlated (since The coefficient on lambda =  $\rho_{eu}\sigma_u$ ). So (unobserved) factors that make participation more likely tend to be associated with lower wages.

Notice also that the estimated coefficient on age is now significantly larger than before, indicating the selection was biasing down the wage returns to age.

In practice all this can be estimated using a single command in stata. Just ensure that you understand the process before using this command.

```
. heckman lhw educ age, select(ndepchld educ age) twostep
```

```
Heckman selection model -- two-step estimates      Number of obs      =      4270
(regression model with sample selection)          Censored obs       =      1974
                                                    Uncensored obs     =      2296

                                                    Wald chi2(4)       =      651.41
                                                    Prob > chi2        =      0.0000
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
-----					
lhw					
educ	.0772027	.0053843	14.34	0.000	.0666496 .0877557
age	.0103897	.0034559	3.01	0.003	.0036162 .0171632
_cons	5.418822	.1021588	53.04	0.000	5.218595 5.61905
-----					
select					
ndepchld	-.1714651	.0219398	-7.82	0.000	-.2144664 -.1284638
educ	.0116392	.0077145	1.51	0.131	-.003481 .0267595
age	-.0442952	.0022067	-20.07	0.000	-.0486202 -.0399702
_cons	2.19459	.1614445	13.59	0.000	1.878165 2.511015
-----					
mills					
lambda	-.3269318	.1556167	-2.10	0.036	-.631935 -.0219286
-----					
	rho	-0.49324			
	sigma	.66282721			
	lambda	-.32693179	.1556167		
-----					