Modeling wages of females in the UK

Saadia Irfan
NUST Business School
National University of Sciences and Technology
Islamabad, Pakistan
E-mail: saadiakhan87@hotmail.com

Abstract

This study analyses the wage equation for women in Britain. The aim of this study is to analyse the determinants of the wages of British women so as to make a statement about them. Data is collected from the BHPS 2005. In order to overcome the sample selection problem, Heckman correction procedure is applied. The findings of the study are generally consistent with previous research on determinants of wages of women.

Introduction

The most obvious analysis of wages of women would be to use the regression model like the following.

$$ln\, W_f = X_f'\beta_f + U_f$$

Where X_f' is a vector of regressors and the error term U_f has zero mean and constant variance. However, estimating the above equation using OLS will give biased results as the OLS does not allow for the sample 'selection problem'. This problem may occur during the collection of the sample and afterwards when for example, the selected females can, and frequently do, refuse to participate. This makes the sample biased if the females who do not participate are systematically different from those who do. This is known as "sample selection bias."Moreover, the sample can also be biased if the females agree to participate but then are "lost" over time due to transience, death, or any other reasons. This is known as "attrition bias."I will focus on sample selection bias only.

Selection bias threatens both the internal as well as external validity of the study. Under selection bias, the independent variables are correlated with the error term and thus the analyses based on such a sample does not give accurate estimates of the relationship between variables(e.g. Regression coefficients). For example, consider the relationship between 'wages of women' and 'years of experience at work'. Now if data for years of work experience of women is missing systematically for women with more years of experience, then the effect of years of work experience on wages of women will be underestimated as quantified using, for example, a regression coefficient. In this way, the internal validity of the study is threatened.

Turning towards the external validity, it is also threatened because the biased sample might not be generalizable to the intended population (Cuddeback et al, 2004). Consider another example of the results of a study that evaluates a high school dropout prevention program based on an analysis of a random sample of students who completed the program. Now the sample might under represent the high-risk students and over represent the low or medium risk students because the students most at risk dropped out of school prior to completing (or even starting) the program. And thus any conclusion that the prevention program is successful for all students irrespective of their level of risk, drawn from the sample might not be generalizable to the students most in danger of dropping out of school. The article proceeds as follows. Section 2 is devoted to the explanation of the technique proposed by Heckman to solve the above mentioned selection problem. Section 3 describes the data used in the study and Section 4 gives an explanation of the implementation. Section 5 discusses the results and presents some suggestions. Finally Section 6 gives the conclusion.

Heckman's solution

The most common technique used to tackle the above problem has been developed by Heckman, 1976, 1978, 1979. Heckman (1979) argues that the given the above problem, it is possible to estimate the variable which when omitted from a regression analysis give rise to the specification error. The estimated value of the omitted variable can be used as a regressor such that it is possible to estimate the functions of interest by simple methods**.** He proposes a two-step estimator where 'outcome' is the woman's wage and 'treatment' is her decision to work in the labour market. The sample selection model works as follows:

The outcome variable W_f is only observed if some criterion, defined with respect to variable Y, is met. Now the participation (treatment) decision of the women in this sample can be modelled using a variable Y to represent their participation.

This variable Y is positive in case where the woman decides to work and negative in case where the woman decides not to participate in work. The participation equation can be written as follows:

$$Y = Z_f^{'}\theta_f + V_f$$

Where lnW_f *is* only observed if *Y>0* and where $E(U_f)$ = $E(V_f)$ = *0*

Now the expected value of $Ln\, W_f$ of only the women who choose to work, can be written as:

$E(ln\, W_f \mid X_f, Y>0) = X_f^{'}\beta_f + E(U_f \backslash Y>0)$ *equation 1*

Provided that the error terms U_f and V_f *are* normally distributed, we have:

$$U_f = \left[\frac{\sigma_{0,1}}{\sigma_0^2}\right]V_f + v_i$$

Where v_i is uncorrelated with V_f

$\sigma_{0,1}$ is the covariance between U_f and V_f meaning that $\sigma_{0,1} = \rho\sigma_0\sigma_1$

σ_0^2 is the variance of V_f

Selectivity bias occurs whenever $\sigma_{0,1} \neq 0$ $i.e\ \rho \neq 0$

Data

The data is collected from BHPS 2005. Since we are only concerned with the wages of females, the observations for males are dropped via STATA. Moreover, a few more variables have been generated, the details of which are given in the Appendix.

Implementation

Suppose I am interested in finding about the determinants of the wages of females in order to make a statement about the determinants of wages of females. The wage equation formulated in this study is as follows:

$$ln\, W_f = X_f^{'}\beta_f + U_f$$

Where U_f *is the error term and* $X_f^{'}$ *is a set of the following variables thought to influence the wages of females in the UK.*

VARIABLE	***DESCRIPTION***
• ojbhrs	Number of hours normally worked per week
• oage	Age at the date of interview
• white	Dummy variable (0/1) equal to 1 if white
• unionmember	Dummy variable (0/1) equal to 1 if member of trade union
• unionatworkplace	Dummy variable (0/1) equal to 1 if union or staff association at workplace
• fsize4	Dummy variable (0/1) equal to 1 if working in a firm with 1-2 employees
• fsize5	Dummy variable (0/1) equal to 1 if working in firm with 3-9 employees
• fsize6	Dummy variable (0/1) equal to 1 if working in firm with 10-24 employees
• fsize7	Dummy variable (0/1) equal to 1 if working in firm with 25-49 employees
• fsize8	Dummy variable (0/1) equal to 1 if working in firm with 50-99 employees
• fsize9	Dummy variable (0/1) equal to 1 if working in firm with 100-199 employees
• fsize10	Dummy variable (0/1) equal to 1 if working in a firm with 200-499 employees
• fsize11	Dummy variable (0/1) equal to 1 if working in a firm with 500-999 employees
• fsize12	Dummy variable (0/1) equal to 1 if working in a firm with more than 1000 employees
• jobtenure	Number of years in current employment
• reg2	Dummy variable (0/1) equal to 1 if residing in inner London
• reg3	Dummy variable (0/1) equal to 1 if residing in outer London
• reg4	Dummy variable (0/1) i equal to 1 f residing in South East
• reg5	Dummy variable (0/1) equal to 1 if residing in South West
• reg6	Dummy variable (0/1) equal to 1 if residing in East Anglia
• reg7	Dummy variable (0/1) equal to 1 if residing in East Midland
• reg8	Dummy variable (0/1) equal to 1 if residing in West Midland conurbation
• reg9	Dummy variable (0/1) equal to 1 if residing in West Midland
• reg10	Dummy variable (0/1) equal to 1 if residing in Manchester
• reg11	Dummy variable (0/1) equal to 1 if residing in Merseyside
• reg12	Dummy variable (0/1) equal to 1 if residing in North West
• reg13	Dummy variable (0/1) equal to 1 if residing in South Yorkshire
• reg14	Dummy variable (0/1) equal to 1 if residing in West Yorkshire

• reg15	Dummy variable (0/1) equal to 1 if residing in York or Humberside
• reg16	Dummy variable (0/1) equal to 1 if residing in Tyne and Wear
• reg17	Dummy variable (0/1) equal to 1 if residing in North
• reg18	Dummy variable (0/1) equal to 1 if residing in Whales
• reg19	Dummy variable (0/1) equal to 1 if residing in Scotland
• reg20	Dummy variable (0/1) equal to 1 if residing in Northern Island
• seg3	Dummy variable (0/1) equal to 1 if employer of a large firm
• seg4	Dummy variable (0/1) equal to 1 if manager of a large firm
• seg5	Dummy variable (0/1) equal to 1 if employer of a small firm
• seg6	Dummy variable (0/1) equal to 1 if manager of a large firm
• seg7	Dummy variable (0/1) equal to 1 if professional self-employed
• seg8	Dummy variable (0/1) equal to 1 if professional employees
• seg9	Dummy variable (0/1) equal to 1 if professional non-manual worker
• seg10	Dummy variable (0/1) equal to 1 if professional non man, foreman
• seg11	Dummy variable (0/1) equal to 1 if junior non manual
• seg12	Dummy variable (0/1) equal to 1 if personal service worker
• seg13	Dummy variable (0/1) equal to 1 if foreman manual
• seg14	Dummy variable (0/1) equal to 1 if skilled manual worker
• seg15	Dummy variable (0/1) equal to 1 if semi-skilled manual worker
• seg16	Dummy variable (0/1) equal to 1 if un-skilled manual worker
• seg17	Dummy variable (0/1) equal to 1 if own account worker
• seg18	Dummy variable (0/1) equal to 1 if farmer-employer, manager
• seg19	Dummy variable (0/1) equal to 1 if farmer-own account
• seg20	Dummy variable (0/1) equal to 1 if agricultural worker
• seg21	Dummy variable (0/1) equal to 1 if members of armed forces
• marr	Dummy variable (0/1) equal to 1 if married

The dependent variable is:

- ologwage Log Gross weekly pay(LnW_f)

In the classical theory, the wage of a female worker can be easily expressed as a function of variables such as office job hours, age, work experience, marital status. In addition to these, I have used variables such as 'unionmember' and 'unionatworkplace' as a host of studies shows (for example,Blanchflower and Bryson; 2002) that wages are strongly affected if there exists a trade union at workplace or if the worker belongs to a trade union. I hypothesize that there is a positive relationship between log wage and the fact that there exists a trade union at workplace or if the worker belongs to a trade union.

Moreover, I have included the variable 'white' in the regression as despite the non-discrimination laws that operate in Britain, a number of studies have documented that white people are receiving higher wages than the non-whites. Also, I have included the variable 'firm size' as generally one would expect a larger firm to pay more wages (including benefits) as compared to a smaller firm. Moreover, the variable 'region' is included because given today's conditions, one would expect a person living in London to be earning more than a person in the same profession in, for example, Yorkshire.

I have obtained the regression estimates using OLS, ignoring the sample selection in order to make a comparison later with Heckman's solution. The estimates are as follows:

```
. drop if male==1;
(5258 observations deleted)

. reg  ologwage oage ojbhrs white unionmember unionatworkplace fsize* jobtenure reg*
> seg* marr if emp==1;
```

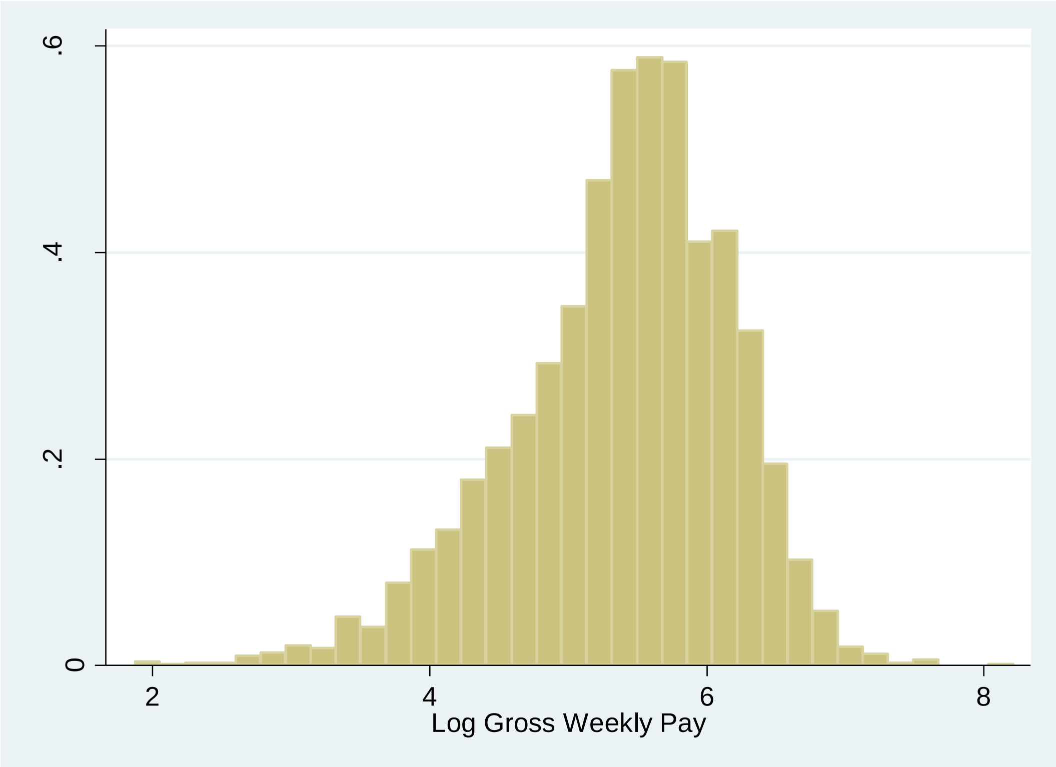
Source	SS	df	MS
Model	1276.60901	47	27.1618938
Residual	617.872396	3597	.171774366
Total	1894.48141	3644	.519890617

Number of obs	=	3645
F(47, 3597)	=	158.13
Prob > F	=	0.0000
R-squared	=	0.6739
Adj R-squared	=	0.6696
Root MSE	=	.41446

ologwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
oage	.0016188	.0007019	2.31	0.021	.0002427	.002995
ojbhrs	.0362456	.0007195	50.38	0.000	.034835	.0376563
white	.0739348	.0267724	2.76	0.006	.0214443	.1264254
unionmember	.1813019	.0200589	9.04	0.000	.141974	.2206297
unionatwor~e	.0687792	.0194576	3.53	0.000	.0306302	.1069283
fsize4	-.0913881	.0673813	-1.36	0.175	-.2234974	.0407212
fsize5	.0054168	.057658	0.09	0.925	-.1076288	.1184623
fsize6	.0728063	.0574647	1.27	0.205	-.0398603	.185473
fsize7	.0812659	.0578641	1.40	0.160	-.0321839	.1947157
fsize8	.0851151	.0589449	1.44	0.149	-.0304537	.2006839
fsize9	.1502036	.059689	2.52	0.012	.033176	.2672312
fsize10	.0963539	.0592131	1.63	0.104	-.0197406	.2124484
fsize11	.1346749	.0633199	2.13	0.033	.0105285	.2588214
fsize12	.1376076	.0591957	2.32	0.020	.0215472	.2536681
jobtenure	.0039338	.0013423	2.93	0.003	.0013021	.0065655
reg2	.0357843	.1673642	0.21	0.831	-.2923539	.3639226
reg3	.036184	.1623381	0.22	0.824	-.2821	.354468
reg4	-.1390006	.1589172	-0.87	0.382	-.4505775	.1725764
reg5	-.3291325	.1602063	-2.05	0.040	-.6432369	-.0150282
reg6	-.3110558	.1633422	-1.90	0.057	-.6313085	.0091969
reg7	-.2152297	.160594	-1.34	0.180	-.530094	.0996346
reg8	-.3057379	.1684516	-1.81	0.070	-.636008	.0245322
reg9	-.1991131	.1623146	-1.23	0.220	-.5173511	.1191248
reg10	-.144955	.1634379	-0.89	0.375	-.4653953	.1754853
reg11	-.3243071	.1689542	-1.92	0.055	-.6555627	.0069484
reg12	-.241339	.1630873	-1.48	0.139	-.5610918	.0784137
reg13	-.2159472	.1651713	-1.31	0.191	-.539786	.1078916
reg14	-.3072095	.1650523	-1.86	0.063	-.630815	.0163959
reg15	-.2712833	.165688	-1.64	0.102	-.596135	.0535685
reg16	-.2852035	.1691402	-1.69	0.092	-.6168238	.0464167
reg17	-.2956611	.1639479	-1.80	0.071	-.6171012	.025779
reg18	-.2776407	.1584849	-1.75	0.080	-.5883701	.0330886
reg19	-.2288928	.1583326	-1.45	0.148	-.5393235	.0815379
reg20	-.2375623	.1585369	-1.50	0.134	-.5483935	.073269
seg3	(dropped)					
seg4	.0947217	.0484403	1.96	0.051	-.0002514	.1896949
seg5	(dropped)					
seg6	-.0646911	.0515599	-1.25	0.210	-.1657806	.0363985
seg7	(dropped)					
seg8	.1742916	.0546837	3.19	0.001	.0670774	.2815057
seg9	-.1251586	.0449511	-2.78	0.005	-.2132907	-.0370264
seg10	-.3674108	.0530204	-6.93	0.000	-.4713639	-.2634577
seg11	-.4695167	.0440665	-10.65	0.000	-.5559144	-.3831189
seg12	-.721037	.0472634	-15.26	0.000	-.8137029	-.6283712
seg13	-.4783496	.0716551	-6.68	0.000	-.6188383	-.3378608
seg14	-.4753782	.0797118	-5.96	0.000	-.6316631	-.3190934
seg15	-.5431121	.0497305	-10.92	0.000	-.6406148	-.4456093
seg16	-.8136605	.0590577	-13.78	0.000	-.9294505	-.6978705
seg17	(dropped)					
seg18	(dropped)					
seg19	(dropped)					
seg20	-.1568136	.1454778	-1.08	0.281	-.4420408	.1284137
seg21	(dropped)					
marr	.0230436	.0154682	1.49	0.136	-.0072838	.053371
_cons	4.614255	.174587	26.43	0.000	4.271956	4.956555

Now the use of household micro data is complicated here as there are some female heads of household who receive no wage at all. This means that wages are only observed for those who work and are unobserved for those who do not work. Thus the sample of women who work in the labour market is not a random sample of women. The following graph shows the

Wage distribution of the sample. Clearly, this distribution would have been different if we could observe those unobserved wages too. Thus, it is appropriate here to use a sample correction method.



In order to correct for this sample bias problem, I have applied the Heckman's two-step estimation procedure.

In the first stage, I have gained probit estimates of the treatment equation. The treatment (participation) equation can be expressed as;
$Y = Z_f^{'}\theta_f + V_f$ *where* V_f *is the error term and* $Z_f^{'}$ *is a set of the following variables thought to influence the probability of participation of females in employment in the UK.*

• Emp	Dummy variable (0/1) equal to 1 if employed
• marr	Dummy variable (0/1) equal to 1 if married
• onchild	Number of children in household
• hed1	Dummy variable (0/1) equal to 1 if highest qualification is higher degree
• hed2	Dummy variable (0/1) equal to 1 if highest qualification is first degree
• hed6	Dummy variable (0/1) equal to 1 if highest qualification is alevels
• hed7	Dummy variable (0/1) equal to 1 if highest qualification is olevels
• hed8	Dummy variable (0/1) equal to 1 if highest qualification is commercial
• othlabstat	Dummy variable (0/1) equal to 1 if retired/maternity leave/ family care/ student/ govt. training/other
• excellenthealth	Dummy variable (0/1) equal to 1 if excellent health
• goodhealth	Dummy variable (0/1) equal to 1 if good health
• fairhealth	Dummy variable (0/1) equal to 1 if fair health
• poorhealth	Dummy variable (0/1) equal to 1 if poor health

As seen from above, the 'marital status' variable is present in both the participation equation as well as the wage equation, since I hypothesize that the fact that a woman is married has an inverse relationship with the both. Moreover, it makes sense to add 'onchild' variable in the participation equation, as it is likely that if there are dependent children in the household, then the woman household head will prefer not to work. Moreover, the type of degree that the female is holding will determine whether she is likely to do work or not that is why I have included the 'highest degree' variables. In addition to this the 'othlabstat' variable shall indicate whether the woman is retired or on maternity leave etc. Last but not least, the health four variables are included as I believe health is a very important factor that determines the likelihood of whether an individual can work or not. The omitted dummy variable for health is 'verypoorhealth'.

The probit estimates of the participation equation are as follows:

```
. probit  emp marr  onchild hed1 hed2 hed6 hed7 hed8 othlabstat excellenthealth  good
> health    fairhealth   poorhealth;

note: othlabstat != 0 predicts failure perfectly
      othlabstat dropped and 2220 obs not used

Iteration 0:   log likelihood = -1424.0522
Iteration 1:   log likelihood = -1398.7956
Iteration 2:   log likelihood = -1398.6788
Iteration 3:   log likelihood = -1398.6788

Probit regression                                 Number of obs   =       4147
                                                  LR chi2(11)     =      50.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -1398.6788                       Pseudo R2       =     0.0178
```

emp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
marr	.248869	.0542576	4.59	0.000	.1425262	.3552118
onchild	-.1131442	.0276155	-4.10	0.000	-.1672696	-.0590188
hed1	.0198785	.1353889	0.15	0.883	-.2454789	.2852359
hed2	.1128858	.0775139	1.46	0.145	-.0390387	.2648103
hed6	.0437787	.081312	0.54	0.590	-.1155899	.2031472
hed7	.0664842	.0732073	0.91	0.364	-.0769993	.2099678
hed8	-.254337	.151507	-1.68	0.093	-.5512852	.0426112
excellenth~h	.4174451	.2698509	1.55	0.122	-.1114528	.9463431
goodhealth	.4101651	.2672895	1.53	0.125	-.1137127	.934043
fairhealth	.3711425	.2724089	1.36	0.173	-.1627692	.9050542
poorhealth	.0350291	.2836652	0.12	0.902	-.5209446	.5910028
_cons	.7796884	.2673598	2.92	0.004	.255673	1.303704

These will help me to generate ‘Inverse Mills ratio’ which is given by the following equation:

$$= \frac{\phi\left[{Z_f\theta}/{\sigma_0}\right]}{\Phi\left[{Z_f\theta}/{\sigma_0}\right]}$$

Where $\phi(.)$ is the standard normal density and $\Phi(.)$ its cumulative distribution function.

Heckman (1979) shows that the Inverse Mills ratio is a proxy variable for the probability of participation and when it is added to the wage equation as an additional regressor, it measures the sample selection effect due to the lack of observations on the earnings of non-participants. Thus its inclusion as an additional regressor, results in the consistent estimation of the remaining coefficients of the wage equation. The estimates including the Inverse Mills ratio(its coefficient gives an estimate of ${\sigma_{0,1}}/{\sigma_0}$) are as follows:

```
. reg  ologwage oage ojbhrs white unionmember unionatworkplace fsize* jobtenure reg*
> seg* marr mills if emp==1;
```

Source	SS	df	MS
Model	1277.85568	48	26.6219933
Residual	616.62573	3596	.171475453
Total	1894.48141	3644	.519890617

Number of obs =	3645
F(48, 3596) =	155.25
Prob > F =	0.0000
R-squared =	0.6745
Adj R-squared =	0.6702
Root MSE =	.4141

ologwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
oage	.0015262	.0007021	2.17	0.030	.0001496	.0029028
ojbhrs	.035931	.0007283	49.34	0.000	.0345031	.0373589
white	.076437	.0267651	2.86	0.004	.0239606	.1289134
unionmember	.181194	.0200414	9.04	0.000	.1419003	.2204877
unionatwor~e	.0682892	.0194415	3.51	0.000	.0301717	.1064067
fsize4	-.0912509	.0673226	-1.36	0.175	-.2232452	.0407434
fsize5	.0042013	.0576095	0.07	0.942	-.1087493	.117152
fsize6	.0711005	.0574182	1.24	0.216	-.0414749	.183676
fsize7	.0781041	.0578257	1.35	0.177	-.0352702	.1914785
fsize8	.0817803	.0589066	1.39	0.165	-.0337133	.197274
fsize9	.1457528	.0596598	2.44	0.015	.0287823	.2627234
fsize10	.0954938	.0591624	1.61	0.107	-.0205013	.211489
fsize11	.1346188	.0632648	2.13	0.033	.0105804	.2586573
fsize12	.1375607	.0591441	2.33	0.020	.0216013	.2535201
jobtenure	.003825	.0013417	2.85	0.004	.0011944	.0064556
reg2	.0028944	.1676629	0.02	0.986	-.3258294	.3316182
reg3	.0106009	.1624741	0.07	0.948	-.3079498	.3291515
reg4	-.1637199	.1590434	-1.03	0.303	-.4755441	.1481043
reg5	-.3565468	.1603895	-2.22	0.026	-.6710102	-.0420834
reg6	-.338057	.163507	-2.07	0.039	-.6586328	-.0174813
reg7	-.2365936	.1606497	-1.47	0.141	-.5515672	.07838
reg8	-.3231423	.1684287	-1.92	0.055	-.6533676	.0070829
reg9	-.2231012	.1624172	-1.37	0.170	-.5415402	.0953379
reg10	-.172367	.1636118	-1.05	0.292	-.4931482	.1484143
reg11	-.3544201	.1691761	-2.09	0.036	-.6861109	-.0227293
reg12	-.2641055	.1631639	-1.62	0.106	-.5840086	.0557976
reg13	-.2392983	.1652546	-1.45	0.148	-.5633004	.0847039
reg14	-.3307854	.1651403	-2.00	0.045	-.6545634	-.0070075
reg15	-.2980694	.1658415	-1.80	0.072	-.6232223	.0270835
reg16	-.3087091	.1692177	-1.82	0.068	-.6404813	.0230631
reg17	-.3206962	.1640681	-1.95	0.051	-.642372	.0009797
reg18	-.2992479	.1585496	-1.89	0.059	-.6101041	.0116083
reg19	-.2519739	.1584262	-1.59	0.112	-.5625882	.0586403
reg20	-.2609328	.1586359	-1.64	0.100	-.5719581	.0500926
seg3	(dropped)					
seg4	.0938179	.0483993	1.94	0.053	-.0010749	.1887107
seg5	(dropped)					
seg6	-.0646788	.051515	-1.26	0.209	-.1656804	.0363228
seg7	(dropped)					
seg8	.1701462	.0546577	3.11	0.002	.062983	.2773094
seg9	-.1262121	.0449137	-2.81	0.005	-.2142709	-.0381533
seg10	-.3656202	.0529784	-6.90	0.000	-.469491	-.2617494
seg11	-.4686635	.0440293	-10.64	0.000	-.5549883	-.3823387
seg12	-.7187282	.0472301	-15.22	0.000	-.8113286	-.6261278
seg13	-.4797823	.0715947	-6.70	0.000	-.6201526	-.3394119
seg14	-.4730219	.0796472	-5.94	0.000	-.6291802	-.3168637
seg15	-.5409456	.0496937	-10.89	0.000	-.6383762	-.4435149
seg16	-.8117742	.0590105	-13.76	0.000	-.9274716	-.6960768
seg17	(dropped)					
seg18	(dropped)					
seg19	(dropped)					
seg20	-.1590363	.1453535	-1.09	0.274	-.4440199	.1259473
seg21	(dropped)					
marr	-.000574	.0177643	-0.03	0.974	-.0354032	.0342552
mills	-.4182769	.1551279	-2.70	0.007	-.7224244	-.1141295
_cons	4.749522	.1815056	26.17	0.000	4.393658	5.105386

From the above, it can be seen that the coefficient of the Inverse Mills Ratio is -0.4182 and significant. Thus $\sigma_{0,1} \neq 0$ and so selection problem is apparent in this model and as a result it would have been incorrect to estimate the wage equation for females using OLS. The negative coefficient of the Inverse Mills ratio signifies that OLS would produce downwardly biased estimates.

Results

Some notable results of the above regression are as follows:

As we would have expected and had hypothesised, age, office hours, being white, the fact that there is a trade union at workplace, and if the worker is a trade union member, job tenure, all have a positive and significant impact upon the Log weekly wage of a female. For example, if the number of office hours of female rises by 1, her wage rises by 3.59%. Likewise, a white female has 7.64 % higher wages than a non white female. Thus the fact that the female is white has a positive and significant impact upon her wages. Moreover, as hypothesised, the fact that the female is married has a negative relationship (although insignificant) with her Log weekly wage. The OLS on the other hand, had produced a positive relationship between the two.

Concluding remarks

For the above model, if we assume the following three,

$$Z_f' = X_f'$$

$$\theta_f = B_f$$

$$V_f = U_f$$

Then we have a standard Tobit model. However, clearly this might be incorrect as covariates affect the participation decision differently from the way they would affect the Log amount of wages that a female gets perweek.Hence,

$$\theta_f \neq B_f$$

Literature suggests that corrections using the Heckman's two step method can sometimes worsen rather than improve estimates, even under ordinary circumstances. For example, Winship & Mare (1992) show that the model is sensitive to hetroscedasity and non-normality. The probit estimation above assumes that the error term (V_f) is homoscedastic and when this assumption is violated, then the Heckman's procedure yields inconsistent estimates. The assumed bivariate normality of V_f and U_f is needed for two reasons. Firstly, normality of V_f is needed for consistent estimation in the probit model. Secondly, normality implies a non-linear relationship for the effect of Z_f' on $ln\ W_f$ through the coefficient on the Inverse Mills ratio. Thus, if V_f is not normal, then the coefficient on the Inverse Mills ratio mis-specifies the relationship between and $ln\ W_f$ and Z_f' and thus the model may yield biased results. An alternative to the above would be to use the 'Heckman' command in the Stata. This uses the Maximum Liklihood approach and corrects for the standard errors. However, to conclude, given that no technique or a set of techniques can offer a universal escape from the sometimes severe problems of selection bias, Heckman's two-step technique offers a useful sample selection correction model.

References

Blanchflower, D. And Bryson, A. (2002).Changes over Time in Union Relative Wage Effects in the UK and the US Revisited. Available from SSRN

Cuddeback, G. Cuddeback. Wilson, E. Orme, G. Combs-Orme, T. (2004). Detecting and Statistically Correcting Sample Selection Bias.

Heckman, J.J. 1979. Sample Selection bias as a specification bias. *Econometrical*, 47:53-161.

Winship, C. and Mare, R. (1992). Models for sample selection bias. Annual review of sociology. Volume 19, pp 327-350.

Appendix

I have generated 5 variables for health, a new variable for trade union member and whether there are any trade union or association at workplace. Copy of the do file is as follows:

```
#delimit;
use "U:\ManXP\Desktop\bhps2005.dta", clear;
gen excellenthealth=1 if  ohlstat==1;
replace excellenthealth=0 if  ohlstat!=1;
gen goodhealth=1 if ohlstat==2;
replace goodhealth=0 if  ohlstat!=2;
gen fairhealth=1 if ohlstat==3;
replace fairhealth=0 if ohlstat!=3;
gen poorhealth=1 if ohlstat==4;
replace poorhealth=0 if ohlstat!=4;
gen verypoorhealth=1 if ohlstat==5;
replace verypoorhealth=0 if ohlstat!=5;
gen unionmember=1 if  otuin1==1;
replace unionmember=0 if  otuin1!=1;
gen unionatworkplace=1 if  otujbpl==1;
replace unionatworkplace=0 if  otujbpl!=1;
drop if male==1;
reg  ologwage oage ojbhrs white unionmember unionatworkplace fsize* jobtenure reg* seg*  marr if emp==1;
probit  emp marr  onchild hed1 hed2 hed6 hed7 hed8 othlabstat excellenthealth  goodhealth    fairhealth   poorhealth;
predict y, xb;
gen n1=normalden(y);
gen n2=normprob(y);
gen mills=n1/n2;
reg  ologwage oage ojbhrs white unionmember unionatworkplace fsize* jobtenure reg* seg*  marr mills if emp==1;
heckman ologwage oage ojbhrs white unionmember unionatworkplace fsize* jobtenure reg* seg*  twostep select (emp=
marr  onchild hed1 hed2 hed6 hed7 hed8 othlabstat excellenthealth  goodhealth    fairhealth   poorhealth);
```