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 lnWf is only observed if Y>0 and where E(Uf) = E(Vf) = 0

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

\[ E(ln W_f | X_f, Y>0) = X_f \beta_f + E(U_f | Y>0) \text{  equation 1} \]

Provided that the error terms Uf and Vf are normally distributed, we have:

\[ U_f = \left[ \frac{\sigma_{0.1}}{\sigma_0^2} \right] V_f + v_i \]

Where vi is uncorrelated with Vf

\( \sigma_{0.1} \) is the covariance between Uf and Vf meaning that \( \sigma_{0.1} = \rho \sigma_0 \sigma_1 \)

\( \sigma_0^2 \) is the variance of Vf

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.

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

- **ologwage**  Log Gross weekly pay ($\ln W_f$)

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:
The Special Issue on Humanities and Social Science © Centre for Promoting Ideas, USA www.ijbssnet.com

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.

198
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:

\[ 0.2 \quad 0.4 \quad 0.6 \]

\[ 2 \quad 4 \quad 6 \quad 8 \]

Density

Log Gross Weekly Pay

As seen from the above graph, the probability of participation increases with log gross weekly pay.
. 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
Pseudo R2 = 0.0178

Log likelihood = -1398.6788

| emp          | Coef.    | Std. Err. | z      | P>|z|   | [95% Conf. Interval] |
|--------------|----------|-----------|--------|-------|---------------------|
| marr         | 0.248869 | 0.0542576 | 4.59   | 0.000 | 0.1425262 to 0.3552118 |
| onchild      | -0.1131442 | 0.0276155 | -4.10 | 0.000 | -0.1672696 to 0.050188 |
| hed1         | 0.0198785 | 0.1353889 | 0.15   | 0.883 | -0.2454789 to 0.2852359 |
| hed2         | 0.1128858 | 0.0775139 | 1.46   | 0.145 | -0.0390387 to 0.2648103 |
| hed6         | 0.0437787 | 0.0813121 | 0.54   | 0.590 | -0.1155899 to 0.2031472 |
| hed7         | 0.0664842 | 0.0732073 | 0.91   | 0.364 | -0.0769993 to 0.2099678 |
| hed8         | -0.254337 | 0.151507 | -1.68  | 0.093 | -0.5512852 to 0.0426112 |
| excellent-h  | 0.4174451 | 0.2696709 | 1.55   | 0.122 | -0.1114528 to 0.9463431 |
| goodhealth   | 0.4101651 | 0.2672895 | 1.53   | 0.125 | -0.1137127 to 0.9404353 |
| fairhealth   | 0.3711425 | 0.2724089 | 1.36   | 0.173 | -0.1627692 to 0.9050542 |
| poorhealth   | 0.0350291 | 0.2836652 | 0.12   | 0.902 | -0.5209446 to 0.5910028 |
| _cons        | 0.7796884 | 0.2673598 | 2.92   | 0.004 | 0.2556730 to 1.303704  |

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

\[
\phi \left( \frac{Z_\theta}{\sigma_0} \right) = \frac{Z_\theta}{\sigma_0} \Phi \left( \frac{Z_\theta}{\sigma_0} \right)
\]

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:
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 per week. 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 heteroscedasticity 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 Likelihood 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


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\bhsps2005.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);
```