A Specifying a wage model

This appendix contains a description of the econometric model used to estimate the effects of education and health status on wages, and sets out the approach used to estimate the potential wages of people not currently employed. Also described are some of the econometric issues associated with estimating wage functions and the techniques that were used to overcome some of those problems, as well as the potential implications of unresolved econometric issues.

A.1 Specifying a human capital earnings function

The effects of human capital characteristics on wages are commonly estimated using a human capital earnings function based on the model specified by Mincer (1974). In Mincer's model, the natural logarithm of wages is expressed as a linear function of years of schooling and a quadratic function of potential experience:

$$\ln w_i = \beta_0 + \beta_1 s_i + \beta_2 x_i + \beta_3 x_i^2 \qquad [A.1]$$

where: w_i is the wage rate of the *i*th individual; s_i represents years of schooling; and x_i is potential years of work experience. Experience is included as a proxy for the accumulation of human capital that occurs after formal education (such as on-the-job training). The quadratic term is included to allow for a possible decline in the returns to this form of human capital over the individual's life. (For example, technological can change render redundant the skills accumulated early in a person's working life.)

Conveniently, coefficients in the log-linear wage equation can be interpreted as approximations of percentage effects. That is, β_1 can be read as an approximation of the effect on wages of an additional year of schooling in percentage terms.[1]

[1] The larger the value of the coefficient, the less accurate it is as an approximation of the percentage effect — coefficients are converted into actual percentages by taking the inverse natural logarithm of the coefficient and subtracting 1: [exp(coefficient) – 1] x 100.

Augmented specification

To estimate the effects of health and education on wages, and to predict the wages of those not currently employed, a number of adjustments are made to Mincer's basic model:

- The explanatory variables are augmented with a vector of health variables (mental and physical health) (appendix B).
- To allow for different returns across different types of education, the continuous measure of schooling is replaced with a series of dummy variables indicating the highest level of educational attainment (appendix B).
- A measure of actual experience is used in place of Mincer's proxy of potential experience.[2]
- A vector of control variables, denoting labour market and demographic characteristics that can influence wages, is included in the specification. The control variables are outlined below.

The basic form of the model is:

$$\ln w_i = \beta_0 + S_i^{'}\beta_1 + \beta_2 e_i + \beta_3 e_i^2 + H_i^{'}\beta_4 + X_i^{'}\beta_5 + \varepsilon_i \qquad \text{[A.2]}$$

where $S_i^{'}$ represents a vector of dummy variables indicating the individual's highest level of education; e_i is a measure of experience; $H_i^{'}$ is a vector of mental and physical health variables; $X_i^{'}$ is a vector of control variables; and ε_i is the error term.

Gender

The very different labour market experiences of women and men require that separate models be estimated for women and men. Miller (1982) shows that women typically earn less than men, the growth of their earnings over their lifetime is much slower, and their age–earnings profile 'dips' in the middle, in contrast to the steadily concave male earnings profile. Gender-wage differences have persisted over time (Le and Miller 2001).

2 A lack of reliable data on labour market experience led Mincer to use *potential* labour market experience as a proxy for actual experience. That is, an individual's potential labour market experience was equal to their age, minus years spent in school, minus years prior to school (typically assumed as 5). Use of an actual measure of experience removes an upward bias associated with measures of potential experience.

Differences in the wages paid to men and women could result from factors including different reservation wages (for example, the opportunity cost of working might be lower for some women than for men), different human capital investment decisions or different treatment in the labour market (including gender discrimination), or a combination of these factors (Preston 2000).

Nevertheless, Australian research shows that there is a persistent ‘gap’ of around 12–15 per cent between male and female wages, even after taking into account differences that may affect productivity (Eastough and Miller 2004, Le and Miller 2001, Miller 2005, Preston 1997). This suggests that different labour market outcomes are not due entirely to human capital investment decisions.[3]

The fact that men and women have systematically different labour market outcomes, and that these differences cannot be explained in terms of human capital theory, implies that there is a different structure of wage determinants for each gender. Wage equations were estimated separately for men and women to account for these differences.

Control variables

To estimate the returns to education and health, it is appropriate to further augment Mincer’s equation with a number of control variables that relate to characteristics that can affect an individual’s earning capacity. Broadly, these can be considered as labour market and demographic variables.

Labour market variables

An individual’s previous and current labour market status can affect their wages. A number of variables are included in the model to account for these effects:

- Full-time study is likely to impact on an individual’s earning capacity, because students’ job opportunities are likely to be constrained and they may be paid wages that are not commensurate with their level of human capital. To account for this, the model includes a binary variable to indicate whether the person is studying.
- Unemployment history is included as a control variable, describing the proportion of time that an individual has spent unemployed since completing

[3] Preston (1997) shows that around 72 per cent of the gender-wage gap is ‘unexplained’, and this is relatively stable between 1981 and 1991. Le and Miller (2001) estimate that the ‘systematic unequal treatment’ of women accounts for around 84 per cent of the difference in male and female wages in the late 1990s.

school. This is included to account for possible scarring effects that might limit an individual's ability to find employment (Knights, Harris and Loundes 2002), or limit their earnings capacity when they are employed (Arulampalam 2001, Gregg and Tominey 2005).

- Whether or not a worker is full- or part-time might also affect their wages. Booth and Wood (2006) present evidence of a premium afforded to part-time workers in Australia. In contrast, Hirsch (2005) finds that part-time workers in the United States are penalised. To account for this, an indicator of part-time work is included in the wage equation.
- Some researchers include in their models variables identifying the industry that people are employed in. All else being equal, wages can differ across industries, due to a range of factors, including: the desirability of the work involved; the level of competition; and the capital intensity of the industry. However, excluding industry variables makes it more likely that education coefficients are representative of the average returns to education across the entire labour market (Chapman, Rodrigues and Ryan 2007). Given the economy-wide focus of this project, no industry variables were included. In addition, excluding industry variables makes it possible to estimate the potential wages of people who are unemployed or not in the labour force (one of the objectives of this study).

Demographic variables

In estimating the relationship between human capital and wages, there are a number of demographic factors to be considered:

- Changes in the age-education profile over time are likely to alter the wage-education relationship, and need to be taken into account when considering returns to education. To this end the human capital model is augmented with age dummy variables to account for the possibility of age, cohort and period effects that might cloud actual returns to education.
- An indicator of indigenous status is included to control for different employment opportunities that may result from cultural differences, discrimination, or specific government policies that may apply to Indigenous Australians (such as the Community Development Employment Projects program).
- Language difficulties can affect an individual's ability to participate in the workforce. A non-English speaking background indicator is included to account for this effect.
- Marital status has been found to be related to wages. For example, Cai (2007) found that married men earn approximately 9 per cent more than unmarried men.

- Geography (including state of residence and whether the respondent lives in a regional area).

Sample selection bias

Data on wages are only available for people in employment, which raises the prospect of bias in the sample of persons in the database used to estimate the wage model. The potential for bias arises because people with observed wages — those employed — are likely to be systematically different from those without observed wages — those unemployed or not in the labour force. Restricting the sample to people who earn a wage is likely to introduce biases into the estimation. Regression analysis of wages and their determinants that is restricted to this non-random sample is likely to return estimates inconsistent with their true population values (Greene 2003).

The problem of sample selection bias can be taken into account by explicitly incorporating into the model an adjustment for the risk that certain people will not be included in the sample (that is, they will not be employed). Heckman (1979) devised an approach where two equations are estimated: a 'selection' (employment) equation and a 'principal' (wage) equation.[4]

The Heckman approach begins with a latent variable E_i^* that is a function of each individual's characteristics z_i:

$$E_i^* = \gamma' z_i + u_i \qquad \text{[A.3]}$$

The latent variable can not be directly observed, but if its value exceeds zero the person will be employed:

$E_i = 1$ if $E_i^* > 0$

$E_i = 0$ if $E_i^* \leq 0$

Turning now to the wage equation, the natural logarithm of hourly wages can be expressed as a function of a vector of human capital, labour market and demographic characteristics $\mathbf{x}_i$:

$$\ln w_i = \beta' \mathbf{x}_i + \varepsilon_i \qquad \text{[A.4]}$$

---

[4] This section is based on Laplagne, Glover and Fry (2005) (unpublished).

The error terms in equations A.3 and A.4 have the following properties:

$u_i \sim N(0, \sigma_u)$

$\varepsilon_i \sim N(0, \sigma_\varepsilon)$

$\mathrm{corr}(u_i, \varepsilon_i) = \rho$

In order to correct for the fact that w_i is observed only when $E_i=1$, the expected wage of each individual must be adjusted by the expected value of the error from the selection equation. Thus, the conditional expected wage is given by:

$$E(w_i \mid E_i = 1) = \beta'\mathbf{x}_i + \rho\sigma_\varepsilon\lambda_i(\alpha_u) \qquad \text{[A.5]}$$

where $\lambda_i(\alpha_u) = \dfrac{\phi(\gamma'\mathbf{z}_i / \sigma_u)}{\Phi(\gamma'\mathbf{z}_i / \sigma_u)}$ is the inverse Mills ratio and ϕ and Φ are the normal density function and the cumulative normal distribution function, respectively.

Rewriting $\rho\sigma_\varepsilon$ as ψ allows equation (A.5) to be rewritten as:

$$E(w_i \mid E_i = 1) = \beta'\mathbf{x}_i + \psi\lambda_i(\alpha_u) \qquad \text{[A.6]}$$

λ is the term in the principal (or wage) equation that corrects for self-selection. The coefficient for λ is the covariance between the error term in the selection and principal equations (equations A.3 and A.4 respectively). A positive and significant coefficient for the correction term implies that employees have unobserved characteristics, such as innate ability, that result in their observed wages being higher than wage predictions based on their observed characteristics. It should be noted that the consistency and unbiasedness of the estimators in this model depend on the validity of the assumption that the disturbance term ε_i is normally distributed.

A.2 Predicting wages for those not employed

The wage model developed for this paper is well suited to the task of estimating the wages of people who are not currently employed, relative to those who are employed. The relative wages of people not currently employed is of particular relevance when using economy-wide models to estimate the effects of proposed human capital and labour market reforms.

The productivity (and hence wages) of people who are not currently employed is likely to systematically vary according to age and sex. For that reason, relative wages are estimated separately for men and women of different ages. Specifically, the model is estimated separately for men and women aged 15–24, 25–44 and 45–64. It is also estimated for male and female recipients of the Disability Support Pension, and for the labour force as a whole.

There are three steps involved in estimating the potential wages of people who are not employed relative to the wages of those who are:

1. Estimate the average wage of each demographic group, conditional on them being employed.
2. Estimate the 'offer wage' of each demographic group. The offer wage is a hypothetical wage that would be offered to a person who is not currently working if they were to start work. The offer wage is estimated based on the person's observed human capital and labour market characteristics.
3. Calculate the ratio of the wage (conditional on employment) and the offer wage to estimate the potential wage of people who are unemployed or not in the labour force relative to the employed population in each demographic group.

Estimating wages conditional on employment status is complicated in this application by the log transformation of the dependent variable (hourly wages). Yen and Rosinski (2008) show that, where the dependent variable is in log form, using the functional form specified in equation A.5 to estimate expected log wages and then taking the exponent of the expected log wage can lead to systematic underestimation of the conditional wage. To account for this possibility, Yen and Rosinski derive an alternative approach to estimating the wage (in dollars, not log form):

$$E(w_i \mid E_i = 1) = \exp\left(\beta'\mathbf{x}_i + \sigma^2/2\right)\frac{\Phi\left(z'\alpha + \rho\sigma\right)}{\Phi\left(z'\alpha\right)} \qquad [A.7]$$

Following the same reasoning, the expected wage conditional on the person not being employed (the offer wage) is given by:

$$E(w_i \mid E_i = 0) = \exp\left(\beta'\mathbf{x}_i + \sigma^2/2\right)\frac{\Phi\left(-z'\alpha - \rho\sigma\right)}{\Phi\left(-z'\alpha\right)} \qquad [A.8]$$

The ratio of the two wages is reported as an estimate of the relative wage of people who are not currently working.