
3 The model and econometric issues

Human capital literature suggests, and descriptive statistics (figures 1.1 and 1.2) appear to confirm, that higher levels of education and good health have a positive relationship with wages and, by implication, productivity. However, it may also be the case that high wages contribute to better health and higher levels of education as they provide the funding to access related goods and services. This section briefly describes the multivariate model that was used to estimate the effects of education and the target health conditions on wages. It also sets out some of the econometric issues associated with this type of research. More detail is provided in appendix A.

3.1 The basic model

The model used to estimate the effects of education and health on wages is based on Mincer's (1974) specification, in which the natural logarithm of hourly wages is expressed as a linear function of years of schooling and a quadratic function of potential experience. Potential experience was used because of a lack of reliable data on actual labour market experience.¹ The quadratic function of potential experience implies that over time returns to experience diminish and eventually could become negative.

The basic form of the model is:

$$\ln w_i = \beta_0 + S_i' \beta_1 + \beta_2 e_i + \beta_3 e_i^2 + H_i' \beta_4 + X_i' \beta_5 + \varepsilon_i$$

where:

- S_i' represents a vector of dummy variables indicating 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 denoting labour market and demographic characteristics; and

¹ Mincer measured experience as a person's age, minus the number of years spent in school, minus the number of years prior to school (generally assumed to be five).

-
- ε_i is an error term.

The variables are explained in more depth in chapter 4 and appendix B.

The model is estimated separately for women and men, to allow for gender differences.²

3.2 Sample selection bias and the Heckman approach

Data on wages are only available for people in employment, which raises the possibility of bias in the data used to estimate the wage model. The potential for bias arises because people with observed wages — the employed — may be systematically different from working-age people without observed wages — people who are unemployed or not seeking employment. If they are systematically different, a model that only uses data from employed people could be biased because it does not account for the potential wages of people not currently working. Regression analysis of wages and their determinants that is restricted to the working population is likely to return coefficient estimates that are inconsistent with their true population values (including those who are working and those who are not currently working) (Greene 2003).

Potential sample selection bias is addressed by applying an approach devised by Heckman (1979). This approach involves estimating two equations: a ‘selection’ equation that estimates the likelihood that a person with a given set of characteristics will be employed; and a ‘principal’ or ‘wage’ equation that includes an adjustment factor based on the selection equation to estimate a wage for everybody in the sample, employed or otherwise.

This approach is well-established and commonly used in labour market research. For example, Breusch and Gray (2004) used HILDA data and a Heckman model to estimate the relationship between wages and a number of individual characteristics, including education. Pelkowski and Berger (2004) estimated the effects of health problems on individuals’ labour market participation and wages. They used the Heckman approach to account for the fact that the sample of people who are earning a wage is non-random, and health status has a significant effect on people’s decision to participate in the labour market.

² An alternative approach was tested in which a single model was estimated for men and women, using dummy variables and interaction. Results showed that there were statistically significant differences between genders in the effects of a range of human capital variables, including education and health status.

The results of econometric estimation carried out for this paper show that there is sample selection bias present for the men in the sample, but not for women (section C.1).

3.3 Other econometric issues

As well as sample selection bias, there are a number of other econometric issues that may lead to bias in the results. Two of the more significant issues are endogeneity bias and unobserved heterogeneity. These issues are briefly discussed below, with further detail presented in appendix A.

Endogeneity bias

‘Endogeneity bias’ arises where the dependent variable (in this case, wages) has a causal effect on one or more of the explanatory variables. This could occur if higher levels of education and good health lead to higher wages and, at the same time, higher wages contribute to better health and higher levels of education. Failing to account for the feedback effects of wages on health and education can lead to biased estimates of the effects of health and education on wages.

Endogeneity between health and wages can arise because of the feedback between wages and health, or from unobserved factors that affect both health and wages. Cai’s (2007) study into the relationship between health and wages found that reverse causality (wages driving changes in health status) was not statistically significant. Cai does find, however, that there is evidence of endogeneity of health resulting from unobserved factors.

A key difference between Cai’s study and this study is the measures of health used. Cai used self-reported health (poor to excellent) as a general measure of health status. This study uses summary indexes constructed from a short-form health survey to measure health. This is a similar approach to the construction of a ‘health stock’ in Disney, Emerson and Wakefield (2006). As Disney, Emerson and Wakefield explain, the construction of such a health measure should ‘strip the health term in the labour force participation equation of possible subjectivity and endogeneity in individual response to general health-related questions’ (Disney, Emerson and Wakefield 2006, p. 626).

Given the findings by Cai (2007) for the HILDA data, and the construction of the health variable by Disney, Emerson and Wakefield (2006), the model used in this study does not adjust for the possibility of endogeneity between wages and health. This is a similar approach to that taken by Brazenor (2002). If endogeneity were

present in the data, it would potentially lead to results that overstate the positive effects of good health on wages.

Endogeneity bias with regard to education remains a potential problem. Card (1999) states:

[s]ince people with a higher return to education will tend to acquire more schooling, a cross-sectional regression of earnings on schooling yields an upward-biased estimate of the average marginal return to schooling ... (p. 1814)

This suggests that the modelling framework used for this project might overstate the positive effects of education on labour productivity. This should be taken into account when interpreting the results of the analysis.

Unobserved heterogeneity

In econometric terms, ‘unobserved heterogeneity’ describes a situation where some unobserved characteristic (such as a person’s innate ability or their work ethic) is related to both the dependent variable (in this case wages) and one or more independent variables (such as health or education). Unobserved heterogeneity can cause endogeneity bias.

Unobserved heterogeneity could arise in the context of the relationship between health and wages. If an unobserved variable (such as self discipline) leads to better health and higher wages, estimated coefficients for the effects of health on wages might be biased and not reflect the true underlying effects of health on wages.

Unobserved heterogeneity is also a potential problem when estimating the relationship between education and wages. ‘Ability bias’ is a specific form of unobserved heterogeneity that refers to the possibility that some people have innate abilities (such as cognitive ability) that would make it easier for them to complete education. Even in the absence of formal education, these characteristics would be sought after in the labour market and rewarded with higher wages. Therefore, some of the benefits that are associated with education might have more to do with the person’s innate characteristics than their level of education, and estimates of the effects of education on wages might be biased.

Laplagne et al. (2007) used HILDA data to estimate the effects of education and health status on labour force participation. They used a series of econometric tests to test for the presence of unobserved heterogeneity, and found statistically significant evidence of unobserved heterogeneity in the data. They concluded that ‘unobserved heterogeneity means that the coefficients from the standard

multinomial logit model are likely to be biased upward' (Laplagne et al. 2007, p. 45).

To the extent that labour productivity is explained by inherent ability (rather than by education), the ability of governments to increase labour productivity by increasing average education levels is lower than would be implied by estimates of the effects of education on wages (as a proxy for productivity).

Leigh (2007) estimated the returns to education in Australia using HILDA data. As part of his analysis Leigh reviewed Australian and overseas literature on ability bias — that is, the extent to which unobserved characteristics account for both the level of education and the measure of performance. Depending on the method used, Australian estimates of ability bias were between 9 per cent and 39 per cent. Overseas estimates ranged from 10 per cent to 60 per cent. For the purposes of his analysis, Leigh assumed that ability bias meant that estimates of the returns to education were biased upward by 10 per cent.

Based on the literature, including Leigh (2007) and the Laplagne et al. (2007) results, it is likely that endogeneity bias would cause the results estimated for this project to be biased upward. That is, the actual positive effects of education and improved health status on wages might be less than implied by this model. However, the use of wages as an indicator of labour productivity could lead to understatement of the effects of education and health status on productivity. It was not possible to determine which of these biases has a more significant effect on the results, and therefore not possible to determine whether the results in this paper under- or overstate the effects of education and health status on labour productivity.

Some researchers have used panel data models to correct for unobserved heterogeneity. This was not possible in this case because of the adjustment required to address sample selection bias in the data. Techniques to correct for sample selection bias in panel data are experimental and beyond the scope of this study.

3.4 Estimating the potential wages of persons not currently employed

The model required to address sample selection bias has the advantage that it can be used to inform other policy questions. One question of interest to policy makers is the potential effect of labour market reforms on macroeconomic indicators such as unemployment rates, gross domestic product (GDP) and labour force productivity. To determine the macroeconomic effects of policies, it is useful to understand the potential productivity that could be expected of people who are unemployed or not

in the labour force if they were to become employed. Estimating the potential wages of these groups — as an indicator of potential productivity — was a secondary objective of this paper.

The potential wages of people who are unemployed or not in the labour force are likely to systematically vary by age and gender (for reasons related to experience, for example). To account for this, the potential wages of men and women were estimated separately. And for each gender the model included binary variables to account for different age groups (15–24 years; 25–44 years; and 45–64 years), and for recipients of the Disability Support Pension³.

The potential wages of non-working men and women in the various age groups were estimated relative to the average wages of employed men and women in the same age groups. Technical details of the approach to estimating the relative wages of the demographic groups are provided in appendix A.

³ Recipients of the Disability Support Pension were a target group for the NRA reforms.