ECONOMIC RECORD

THE ECONOMIC SOCIETY OF AUSTRALIA

THE ECONOMIC RECORD, 2010

# The Accuracy of Predicted Wages of the Non-Employed and Implications for Policy Simulations from Structural Labour Supply Models*

ROBERT BREUNIG
*Economics Program, Research School of Social Sciences, Australian National University, Canberra, Australia*

JOSEPH MERCANTE
*Tax Analysis Division, Australian Treasury*

*We examine the accuracy of predicted wages for the non-employed. We argue that unemployment, marginal attachment, and not in the labour force are three distinct states. Using panel data from Australia, we test the accuracy of predicted wages for these three groups of non-employed using sample selection models. Focusing on those individuals who subsequently enter employment, we find that predictions which incorporate the estimated sample selection correction perform poorly, particularly for the marginally attached and the not in the labour force. These results have important implications for policy simulations from structural labour supply models.*

## I Introduction

Labour supply models are often used to predict responses of individuals to changes in government tax and transfer systems. Of particular interest in many developed countries is the effect of such changes on individuals who are not currently working. Many government programmes around the world, such as earned income tax–credits and increased tax-free income thresholds for low earners, are specifically designed to attract new workers into the work force and into employment.

An important aspect of the predictions from these labour supply models is the predicted wages which are generated for non-employed individuals. These predicted wages directly determine the additional (predicted) utility that non-employed individuals will get from working and therefore the predicted changes in employment which will ensue from a policy change. For example, labour supply models will overstate (or understate) the employment benefits

* Robert Breunig thanks the Australian Treasury for hospitality and support. We benefitted from comments by Deborah Cobb-Clark, Jennifer Foster, Paul Frijters, Xiaodong Gong, Guyonne Kalb, Anthony King, Paul Miller and Marty Robinson. We are grateful to them. This paper uses confidentialised unit record data from the Household, Income and Labour Dynamics in Australia (HILDA) survey. The HILDA Project was initiated and is funded by the Commonwealth Department of Families, Community Services and Indigenous Affairs (FaCSIA) and is managed by the Melbourne Institute of Applied Economic and Social Research (MIAESR). The findings and views reported in this paper, however, are those of the authors and there is a non-countable set of people and organisations to whom the findings and views should not be attributed including FaCSIA, MIAESR and the Australian Treasury.

JEL Classifications: C52, J22, J30, J64

*Correspondence*: Robert Breunig, Economics Program, Research School of Social Sciences, Building 9, Australian National University, Canberra, ACT 0200, Australia. Email: Robert.Breunig@anu.edu.au




doi: 10.1111/j.1475-4932.2009.00619.x

from tax cuts if wages of the non-working are systematically over-predicted (or under-predicted).

Australian examples of the use of predicted wages in structural labour supply modelling include Duncan and Harris (2002), Kalb (2002), Kalb and Lee (2008) and Breunig *et al.* (2008). Predicted wages, corrected for sample selection, are used in a wide variety of other applications beyond structural labour supply modelling. For example, Rammohan and Whelan (2005) generate predicted wages for modelling the choice of child-care usage for working and non-working women.

The focus of this paper will be on two specific aspects of wage modelling for the non-working. First, we examine whether the unemployed, the marginally attached and those not in the labour force should be treated identically or separately in modelling the probability of employment. We propose a new test for determining whether the non-employed should be categorised as one, two or three groups. We find evidence that the unemployed, the marginally attached and the not in the labour force should be treated as three distinct groups for modelling purposes.

In the second part of the paper, we examine the wage predictions resulting from regressions which correct for selectivity bias using binomial and multinomial models of employment status. Specifically, we evaluate both conditional and unconditional wage predictions from these models. Using a panel of data from Australia, we compare predicted wages for the non-working to the wages they actually receive when they subsequently enter the labour market.

Overall, we find that wage predictions from wage equations which control for selection and which use information from the selection correction perform poorly. Selection correction terms are often poorly estimated and in small samples can be highly variable. For some groups that we consider, this results in very poor predictive performance. Including the estimated selection parameter in the wage equation leads to under-prediction of wages for the not in the labour force and marginally attached groups. For the unemployed, the results are more mixed, but it is clear that using the conditional (on selection) predictor sometimes produces very poor predictions. The main conclusion from the paper is that researchers should exercise caution in the use of conditional predictors for wages of the non-working, particularly in small samples.

The paper is organised as follows. In Section II, we discuss wage models which control for selection into employment. In Section III we discuss our data. In Section IV, we discuss our strategy for testing whether the non-employed should be pooled or considered separately. In Section V, we examine the wage predictions from our models and compare them with realised wages for those who transit from not working to employment. We test which models generate the most accurate wage predictions. In Section VI we discuss our results and conclude.

## *II Wage Models with Selection*

The standard approach in the literature is that proposed by Heckman (1979) whereby wages $w_i^*$ for all workers and non-workers depend upon a vector of observable human capital characteristics, $\mathbf{x}_i$ and some unobservable variables captured by $u_i$

$$\ln(w_i^*) = \mathbf{x}_i'\boldsymbol{\beta} + u_i. \tag{1}$$

The actual wage, $w_i$, is only observed if a latent variable $s_i^* > 0$ where

$$s_i^* = \mathbf{z}_i'\boldsymbol{\gamma} + v_i \tag{2}$$

$\boldsymbol{\beta}$ and $\gamma$ are vectors of parameters and Equation (2) provides a model for the probability of employment. This latter equation captures the benefits of employment and therefore $\mathbf{z}_i$ must contain all of the variables in $\mathbf{x}_i$. If we think of this model as arising in the context of the Heckman (1974) reservation wage model, it should also contain variables which affect the reservation wage, which is (at least partially) determined by the costs of employment. Importantly, $u_i$ and $v_i$ are assumed to be jointly, normally distributed.

The two-step empirical approach is to estimate $\gamma$ in Equation (2) and use those to estimate

$$\ln(w_i) = \mathbf{x}_i'\boldsymbol{\beta} + \rho\lambda(\mathbf{z}_i'\hat{\gamma}) + u_i \tag{3}$$

on the sample with observed wages. The inclusion of the inverse Mills ratio, $\lambda$, corrects for the fact that $E[v_i|s_i^* > 0] \neq 0$. In a reservation wage model, $\rho$ captures two effects. The first effect is that unobservable characteristics which result in a higher wage will also result in a higher probability of employment. $\rho$ will also capture the difference between the variance of wage offers and the covariance between wage offers and reservation wages. The first effect will be positive. The second effect will be negative if the covariance between reservation wages and wage offers, which one would expect to be positive, is greater than the variance of wage offers (see Ermisch & Wright, 1994). Empirically, it is not rare for the latter effect to dominate and produce negative estimates of $\rho$.[1]

To predict wages from Equation (3), one has several options. The unconditional predictor

$$E[\ln(w_i)] = \ln(\hat{w}_i) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} \tag{4}$$

gives the best estimate of the wage for the case where we do not know whether or not the individual is working. If we know that the individual is working, we can condition on this information and use our model estimates to generate a conditional predicted wage for working individuals

$$E[\ln(w_i)|s=1] = \ln(\hat{w}_i^{\mathrm{e}}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\rho}\frac{\phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}{\Phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}. \tag{5}$$

For those who are not employed, the conditional prediction of wages will be

$$E[\ln(w_i)|s=0] = \ln(\hat{w}_i^{\mathrm{ne}}) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{\rho}\frac{-\phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}{1-\Phi(\mathbf{z}_i'\hat{\boldsymbol{\gamma}})}. \tag{6}$$

Note that in using Equations (5) and (6) we are conditioning on unobservable human capital characteristics *and* on the relationship between the distributions of wage offers and reservation wages.[2]

If the model is correctly specified, the conditional predictor contains more information than the unconditional and Vella (1988) suggests its use in generating predicted wage gaps for black–white or Man–Woman differences which condition on the work decision variables and the estimate of the parameter $\rho$. Use of the unconditional predictor provides only an estimate of the wage gap experienced by those who work. Schaffner (1998) points out that using the unconditional predictor is only valid under very restrictive conditions. In particular, if there are unobserved traits that matter for one group and not for the other, then wage gap estimates will be biased. Our focus will be on prediction for individuals rather than groups and the key assumption in using the conditional predictor is that the distribution of unobservables, in particular, the variances and covariances captured by $\rho$, are reasonably constant over time.

Puhani (2000) reviews some of the main critiques of the Heckman selection approach. It has been criticised on the grounds of not providing an improvement in predictive power (for worker's wages) relative to ordinary least squares regression on the selected sample. It also suffers from potential collinearity problems when the variables in $\mathbf{z}_i'$ do not differ much from those in $\mathbf{x}_i'$. Lastly, the Heckman approach makes strong distributional assumptions which, when violated, may lead to poor model performance as has been validated in a number of Monte Carlo studies. These specification problems and the sensitivity of results to the strong model assumptions are generally found to be worse in small samples.[3]

In practice, one can estimate this model by pooling the unemployed, the marginally attached and the not in the labour force to form the category of non-workers or one can exclude one or more of these categories.[4] Flinn and Heckman (1983) find that, for young men, the unemployed and the not in the labour force are distinct groups and that the unemployment state facilitates job search in line with standard search theory models. Similar results are found by Tano (1991) for young people compared with older people, and Gonal (1992) for young women compared with young men. We will test whether the unemployed, the marginally attached and the not in the labour force are distinct groups and we will also check whether the distinction makes any difference in accurately predicting wages. These tests are described below.

If non-employment can best be described as a set of distinct categories, there may be predictive gains in modelling them as such. In that case, several methods have been suggested.[5] We begin with a multinomial model with $J$

[1] Dolton and Makepeace (1987) also discuss the difficulty of interpreting sample selection effects and point out that it is erroneous to argue that participants have lower earnings potential than non-participants when $\rho$ is negative.

[2] Hoffmann and Kassouf (2005) derive the marginal effects in a log earnings equation using the conditional predictor.

[3] Stolzenberg and Relles (1997) provide some intuition about specific mechanisms which can cause poor performance when using the Heckman selection approach.

[4] Most Australian studies treat the unemployed and the not in the labour force (including the marginally attached) as a combined group of non-workers. An exception is Ross (1986).

[5] In this paper, we do not consider nested models, where a sequence of choices are made.

states. Each state $j = 1,\ldots,J$ has an associated utility which is described as

$$s^*_{ij} = \mathbf{z}'_i\gamma_j + v_{ij}. \tag{7}$$

Without loss of generality, letting $j = 1$ be the employed state, wages are observed whenever

$$s^*_{i1} > \max_{j\neq 1}\{s^*_{ij}\}. \tag{8}$$

When the $v_{ij}$ are independently and identically Gumbel distributed this produces the multinomial logit model (see McFadden, 1973).

The approach of Lee (1983) is to specify a bivariate distribution between $u_i$ in Equation (1) and $\varepsilon_1$, defined as

$$\varepsilon_1 = \max_{j\neq 1}(s^*_j - s^*_1) \tag{9}$$

with no restriction on the parametric form of the bivariate distribution beyond standard regularity conditions. Lee (1983) further assumes that the joint distribution of $u$ and the inverse cumulative normal transformation of the cumulative distribution function of $\varepsilon_1$ do not depend upon the parameters of the distribution function of $\varepsilon_1$. In most applications, $u_i$ is assumed to be normally distributed which implies a linearity restriction on the conditional distribution of $u$ as discussed in Bourguignon *et al.* (2007, p. 177).

Schmertmann (1994) shows that these assumptions imply very strong restrictions on the correlation between $u$ and the $v_j$ from Equation (7). The correlations between the difference in unobservable determinants of the choice of alternative 1 against any other alternative and the unobservable determinants of wages must all have the same sign. If the unobservable determinants of utilities are identically distributed, as in the multinomial logit model, then these correlations must in fact be identical. Nonetheless, despite the restrictiveness of these assumptions, many empirical studies follow this route.

Combining the approach of Lee (1983), with the multinomial logit model, and the normality assumption on the unobservables in Equation (1) we estimate a wage equation, correcting for selection as

$$\ln(w_i) = \mathbf{x}'_i\boldsymbol{\beta} - \sigma\rho\frac{\phi(\Phi^{-1}(F_1(\hat{\gamma}_1,\ldots,\hat{\gamma}_j)))}{F_1(\hat{\gamma}_1,\ldots,\hat{\gamma}_j)} + u_i \tag{10}$$

where the $\gamma_j$ are the estimated coefficients of the multinomial logit model and $F_1$ is the cumulative distribution function of the first alternative (employment). $\phi$ and $\Phi$ are the probability density function and cumulative distribution function, respectively, of the standard normal. $\Phi^{-1}$ is the inverse cumulative distribution function of the standard normal. $w_i$ is observed only if workers are in the employed state. $\sigma$ is the standard deviation of the unobservables from Equation (1) and $\rho$ is the correlation between those unobservables and the translation of $v_{i1}$ from Equation (9). We cannot estimate $\rho$ and $\sigma$ separately, but the product of the two is estimated.

Once the parameters of Equation (10) are estimated, one can use the estimate of $\boldsymbol{\beta}$ to predict wages using the unconditional predictor of Equation (4). Alternately, one can create a conditional predictor for an individual's wage in state $j \neq 1$. The conditional predictor makes use of the extra information in $\widehat{\sigma\rho}$ and the estimates of $F_1$.

Another approach using the multinomial logit, proposed by Dubin and McFadden (1984), imposes a linearity assumption on the relationship between the error terms in the wage equation and the selection model. This gives rise to a wage equation, corrected for selection, as

$$\ln(w_i) = \mathbf{x}'_i\boldsymbol{\beta} + \sigma\frac{\sqrt{6}}{\pi}\sum_{j=2}^{M} r_j\left(\frac{P_j\ln(P_j)}{1-P_j} - r_1\ln(P_1)\right) + u_i. \tag{11}$$

$r_j$ is the correlation between $u_i$ in Equation (1) and $v_{ij}$ in Equation (7) for the $j$th alternative. This approach is less restrictive than the approach of Lee (which requires equal covariances between the unobservables in the wage equation and the unobservables which determine the utility for all $J$ states) and therefore more robust. However, it involves the complexity of estimating additional parameters which may be poorly estimated in typical samples.

Dubin and McFadden (1984), to address this problem and simplify estimation, also introduce a variant of their model which requires that the correlations sum to zero across all states. This provides a restricted model

$$\ln(w_i) = \mathbf{x}'_i\boldsymbol{\beta} + \sigma\frac{\sqrt{6}}{\pi}\sum_{j=2}^{M} r_j\left(\frac{P_j\ln(P_j)}{1-P_j} + \ln(P_1)\right) + u_i. \tag{12}$$

The linearity assumption proposed by Dubin and McFadden (1984) restricts the class of allowable distributions for $u$ and imposes a specific form of linearity between $u$ and Gumbel distributions (see Bourguignon *et al.*, 2007, p. 179). This restriction does not allow for $u$ to be normally distributed. Relative to Lee (1983), this provides a different set of assumptions which are not necessarily weaker or stronger.

Bourguignon *et al.* (2007) propose an alternative restriction which allows normality of $u$. This restriction requires that the expected value of $u$ conditional on $v_1$ to $v_j$ be a linear function of the correlations between $u$ and each $v$. This has the drawback of not providing a closed form solution for the conditional expectations of the $v_1$ to $v_j$, but the numerical computation is not particularly difficult.[6]

The wage, conditional on choosing to work, is

$$\ln(w_i) = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\left[r_1^* m(P_1) + \sum_{j=2}^{M} r_j^* m(P_j)\frac{P_j}{1-P_j}\right] + u_i \qquad (13)$$

where the $m(P)$ are defined as

$$m(P_j) = \int \Phi^{-1}(z - \ln(P_j))\, g(z)\, \mathrm{d}z \qquad (14)$$

and the $g$ are the probability density function of the $v$ which are assumed to be identically distributed. $r_j^*$ is the correlation between $u$ and $\Phi^{-1}(v_j)$.

For the predicted wages of individuals who are not working, we can again use an unconditional or a conditional predictor.

We will use the four methods discussed above to predict wages for those who are not working and compare them with the actual observed wages that those same individuals earn once they enter the labour force.[7] We do this using panel data, which we describe in the next section.

## III Data

The data are derived from the Household, Income and Labour Dynamics in Australia Survey (HILDA).[8] The HILDA Survey is a nationally representative annual panel survey of Australian households and we use the first five waves from 2001 to 2005. There are around 7500 households and around 13,000 responding individuals in each wave. After removing multi-family households, same-sex couple households and couple households where partner information is unavailable, there are, in wave five, 3954 married women and men, 695 lone parents, 1108 single women and 989 single men.

We further restrict our sample to persons between 25 and 59 years of age, to exclude those facing decisions about full-time study or retirement. We drop the self-employed, workers in family businesses, full-time students and the retired. Also dropped are those receiving disability support pension, Department of Veteran's Affairs disability pension or sickness allowance. Finally, persons who report working positive hours but state a zero wage are removed.[9] For couples, we drop the observation if either member satisfies one of these conditions. The analysis sample contains 1492 married women and married men. In the final sample of 484 lone parents, the majority (88 per cent) are women. Also there are 315 single women and 380 single men. The numbers are fairly similar for the earlier waves.

We discuss the definition of our key variables. Hours of labour supplied is defined as usual weekly hours of work in all jobs. The wage rate is defined as the person's gross weekly salary and wage income for all jobs divided by hours. For those not working, a wage of zero is assigned. Non-labour income is defined as the difference between gross income and salary and wage income over the financial year. Welfare income is income from pensions and benefits,

[6] In implementing this method in Section V below, we use the STATA code of Bourguignon *et al.* (2007) available at the link provided in their paper.

[7] Bourguignon *et al.* (2007) also discuss the semiparametric estimator of Dahl (2002). We do not consider this estimator here because we are interested in comparing the parametric multinomial models with the standard parametric Heckman model. An interesting project, beyond the scope of this paper, would be to compare wage predictions using the semi-parametric variants of the Heckman model with those based upon the model of Dahl.

[8] For more details, see Watson and Wooden (2002).

[9] Less than 1 per cent of the working sample reported zero wage.

TABLE 1
*Number of Individuals in Analysis Sample Entering Employment by Wave and by Employment State in Previous Wave (Full Sample)*

| Employed in wave | Status in previous wave | | | Total |
|---|---|---|---|---|
| | Unemployed | Marginally attached | NILF | |
| 2 | 58 (43.3%) | 58 (23.4%) | 47 (15.2%) | 163 (23.6%) |
| 3 | 53 (46.1%) | 46 (23.0%) | 40 (13.5%) | 139 (22.8%) |
| 4 | 42 (46.7%) | 42 (24.1%) | 47 (16.5%) | 131 (23.9%) |
| 5 | 46 (52.9%) | 52 (32.9%) | 50 (18.1%) | 148 (28.4%) |
| Total | 199 (46.7%) | 198 (25.4%) | 184 (15.8%) | 581 (24.5%) |

family tax benefit, maternity allowance and childcare benefit.[10]

We categorise people into four employment states: employed (E), unemployed (U), marginally attached (M) and not in the labour force (NILF). A person is considered to be marginally attached to the labour force if they want to work and are actively looking for work but not available to start work in the reference week; or want to work and are not actively looking for work but are available to start work within 4 weeks. In Australian official statistics, as in most countries, the marginally attached are included in the NILF group. There is a growing literature across a range of countries (e.g. Gray *et al.* (2005) for Australia, Brandolini *et al.* (2006) for Europe, and Jones and Riddell (1999, 2006) for Canada) showing that the three groups of non-employed behave quite differently in their propensity to transit to employment, with the marginally attached being less likely than the unemployed, but more likely than the NILF, to transit to employment. Table A1 provides details on the wave-by-wave sample sizes by labour force status.

Of particular interest in this study are the individuals who enter employment from one of the three non-employed categories. In our analysis sample, there are 581 cases (561 unique individuals) in the first five waves of HILDA where the individual is employed at time period $t + 1$ and not employed at time $t$.

The percentages in Table 1 indicate the fraction of individuals from the particular employment state who transitioned to employment. For example, of the unemployed in wave one, 43.3% were employed in wave two. Table 2 provides the transitions by gender and single/partnered status. Throughout, we treat those in *de facto* relationships as married. For non-partnered individuals, we separately consider lone parents. In our analysis, we pool single men and women due to the small sample sizes in those groups.

Tables A2 and A3 provide population estimates from Australian Bureau of Statistics (2007) for monthly transitions to employment. On average across the 6 years, approximately 22 per cent of individuals who are unemployed transit to employment in a given month and approximately 6.7 per cent of those not in the labour force transit to employment. One would expect annual transitions to be higher, which is what we find. Comparison is rendered difficult as, in the official statistics, the marginally attached and NILF are combined and we are not able to separate out the two categories. Another problem is that the official statistics have not been subjected to the various sample exclusions that we have applied to the HILDA data.

We are particularly interested in the wages of individuals who become employed after exiting the unemployed and/or not in the

[10] Where data for unearned income and salary and wage income were missing, we used the imputed values provided by HILDA. The imputation method for the first two waves is described in Watson (2004) and subsequent improvements based upon the method of Little and Su (1989) are discussed in Starick and Watson (2007). Following Frick and Grabka (2007), we included in our models dummy variables for each potentially imputed variable (wage, partner's wage and unearned income) which were set to one when we used an imputed value rather than the actual value. These imputed dummy variables made no difference to the results and we present the results without them.

TABLE 2
*Number of Individuals in Analysis Sample Entering Employment from Non-Employed State by Household Type*

| | Status in previous wave | | | |
|---|---|---|---|---|
| Subgroup | Unemployed | Marginally attached | NILF | Total |
| Married men | 67 (60.9%) | 21 (32.8%) | 10 (19.2%) | 98 (43.4%) |
| Married women | 49 (48.0%) | 112 (26.7%) | 145 (17.0%) | 306 (22.3%) |
| Single men | 35 (39.8%) | 12 (27.9%) | 3 (17.7%) | 50 (33.8%) |
| Single women | 17 (41.5%) | 7 (31.8%) | 5 (22.7%) | 29 (34.1%) |
| Lone parents | 31 (36.5%) | 46 (19.9%) | 21 (9.3%) | 98 (18.1%) |
| Total | 199 (46.7%) | 198 (25.4%) | 184 (15.8%) | 581 (24.5%) |

TABLE 3
*Mean (Median) Hourly Wages of Individuals in Analysis Sample Who Transit to Employment*

| From wave | To wave | Unemployed | Marginally attached | NILF | Employed |
|---|---|---|---|---|---|
| 1 | 2 | 19.6 (17.2) | 17.2 (15.2) | 18.6 (17) | 21.8 (19.6) |
| 2 | 3 | 16.4 (15) | 15.3 (14.6) | 17.8 (15.6) | 22.7 (20.2) |
| 3 | 4 | 20.3 (16.3) | 18.9 (16.7) | 21.8 (20) | 23.6 (21) |
| 4 | 5 | 20.1 (17.4) | 20.3 (16.7) | 20.4 (18.3) | 25.2 (22.2) |
| Total | | 19.0 (16.1) | 17.9 (15.6) | 19.7 (17.4) | 23.3 (20.7) |

labour force categories. Predicted wages are of interest for those who are not working and their transition to employment will provide an opportunity to compare their wages with those of the continually employed and their actual wages when employed to predictions prior to employment. Average hourly wages for our sample are given in Table 3 by wave and Table 4 by gender/partnered/lone parent split. The wages in Tables 3 and 4 are not corrected for inflation.

We can test, using *t*-tests, whether mean wages in Table 3 are statistically different depending upon previous labour force status, without consideration of any individual characteristics. For those working, wages for the individuals who were employed in the immediately preceding wave (the last column of Table 3) are statistically larger (at the 10% level in all cases, at much lower levels for most cases) than wages for those who transition to employment from any of the other labour force states. For the most part, wage differences between those who were previously not in employment are not statistically different from one another. The exception is that for waves 2 and 3 and the pooled data, we find that the employed who were previously NILF have statistically larger wages than the employed who were previously marginally attached.[11]

Turning to the outcomes classified by sex and marital status of Table 4, current mean wages for the previously employed are statistically greater than wages for the previously unemployed at the 6 per cent level or lower for all groups. For married men, married women and lone parents, wages for the previously employed are statistically greater than wages for the previously marginally attached. The sample sizes for single men and women are very small and it is difficult to make any statistical statement about these two groups of marginally attached. Wages for the previously employed are statistically greater than wages for the previously NILF for all groups except married women. Wages for the three groups of previously non-employed are not statistically different from one another for any

[11] If we conduct a non-parametric test of the equality of the medians, we find similar results. The median wages of the previously employed are significantly greater than those of the previously not employed for all three sub-groups.

Table 4
*Mean (Median) Hourly Wages of Individuals in Analysis Sample Who Transit to Employment*

| Subgroup | Unemployed | Marginally attached | NILF | Employed |
|---|---|---|---|---|
| Married men | 21.1 (16.3) | 18.4 (15.6) | 18.7 (17) | 26.0 (23) |
| Married women | 18.3 (15.6) | 18.3 (16.6) | 20.3 (18) | 21.0 (19) |
| Single men | 19.6 (17.5) | 20.1 (19.8) | 15.6 (14.4) | 22.7 (20) |
| Single women | 17.3 (15.8) | 21.5 (16.3) | 16.9 (17.8) | 21.5 (20) |
| Lone parents | 15.9 (14.3) | 15.7 (14.8) | 17.8 (16.4) | 20.7 (18.8) |
| Total | 19.0 (16.1) | 17.9 (15.6) | 19.7 (17.4) | 23.3 (20.7) |

of the sub-groups. The tests for equality of medians reveals the same patterns.

In summary, we draw several conclusions from the data on transitions into employment and wages for those who become employed. First, and in keeping with Flinn and Heckman (1983) and the subsequent literature that they inspired, we find that unemployed individuals have a higher probability of entering employment relative to those not in the labour force. We see this in both the monthly and the annual transitions. Second, it appears that our estimation sub-sample in HILDA has above-average propensity to become employed compared with population estimates of the Australian Bureau of Statistics. This is perhaps not surprising given the additional sample exclusions that we have made (full-time students and disabled) and the fact that we have annual, not monthly, transitions.

Married and single women have higher rates of movement from not in the labour force to employment relative to men. For married women this accords with our prior expectations. For married women and lone parents, average wages in employment when the previous state was not in the labour force are higher than wages when the previous state was either unemployment or marginal attachment, if we pool these two last categories. This result is significant at the 10% level in a one-sided test. This would be consistent with a model where married women and single parents who are caring for children at home have higher average labour productivity than unemployed women.

## *IV Should We Treat All Non-Workers Identically?*

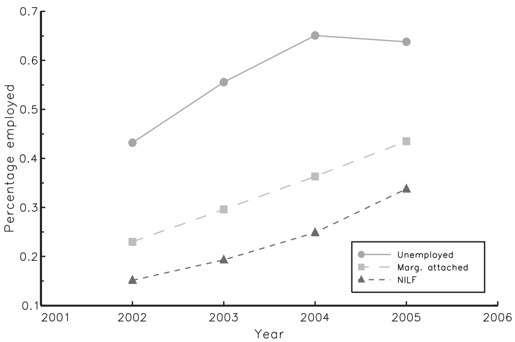
We want to examine whether the three groups of non-workers, the unemployed, the marginally attached and the not in the labour force, should be modelled separately or together. There is a growing literature which demonstrates that these three groups have very different propensities to become employed (see Jones & Riddell, 1999, & 2006 for Canada, Brandolini *et al.*, 2006 for Europe, and Gray *et al.*, 2005 for Australia). We wish to address a different but related question: should the non-employed be considered as one, two or three separate groups when estimating a wage equation which corrects for sample selection?

We offer a new method to address this question, which is to specifically look at models of employment probability for these three groups in combination with the employed. To our knowledge, the classification tests that we propose below are new.[12] The advantage of these tests is that they directly address the question of which modelling approach of those discussed in Section II is appropriate – a binomial classification of the employed and non-employed and the Heckman model or a richer multinomial classification in conjunction with the Lee or McFadden methods.

Gray *et al.* (2005) have applied the tests of transition probabilities proposed by Jones and Riddell (1999) to Australia using a different dataset which covers the period 1994 to 1997 and find that the marginally attached are distinct from both the unemployed and the not in the labour force. We applied these tests and the non-parametric tests of Brandolini *et al.* (2006) to our data and we also reject the hypothesis that the probabilities of transitioning into

[12] Since writing this paper we have become aware of the paper of Ahn and Low (2007) who propose a similar approach to distinguishing between the unemployed and the not in the labour force. They do not separately consider the marginally attached.

FIGURE 1
*Probability of Employment at Subsequent Waves Conditional on Initial Employment Status*




employment are identical for any of the groups of non-employed. We therefore confirm that the conclusions of Gray *et al.* (2005) are also found for the 2001–2006 period using the HILDA data. Figure 1 shows the transitions to employment by wave for all individuals in our analysis sample who are non-employed at wave 1.

Turning to our proposed classification tests, we examine three different possibilities: that the unemployed (U) and the marginally attached (M) can be pooled; that the unemployed and the not in the labour force (NILF) can be pooled, and that the marginally attached and the not in the labour force can be pooled. If we find that two of these groups can be pooled, we can subsequently test whether that pooled group can be pooled with the third remaining category.

For each of the three pairings which we test, we propose five different classification tests. We outline these below using the test for pooling the unemployed and the marginally attached as an example. Our testing approaches are based upon estimation of binomial and multinomial choice models. We estimate three probit models:

P1 Estimate probability of being employed using E, U and M.
P2 Estimate probability of being employed using E and M.
P3 Estimate probability of being employed using E and U.

If the model which determines non-employment is the same for the unemployed and the marginally attached, then P1, P2 and P3 should all (asymptotically) give similar answers. However, P2 and P3 should be inefficient relative to P1, as they only use a portion of the data. The basic principle underlying the Hausman (1978) test (comparison of two sets of coefficients, one of which is consistently estimated under the null and the other which is efficiently estimated under the null) therefore applies.[13]

Hence, our first two tests are:

T1 Hausman test comparing coefficients from P1 to those of P2.
T2 Hausman test comparing coefficients from P1 to those of P3.

We can also compare estimates from a multinomial choice model with those from a binary choice model. For this comparison, we estimate two logistic models:

L1 Binary logit for probability of being employed using E, U and M.
L2 Multinomial logit allowing U and M to be two distinct states.

Again, the Hausman principle applies and we have two Hausman-type tests that can be produced from these estimates:

T3 Hausman test comparing coefficients for unemployed from L1 and L2.
T4 Hausman test comparing coefficients for marginally attached from L1 and L2.

We can also use the multinomial logit estimates to conduct a Wald test to see if the coefficients for the unemployed and marginally attached states are equal.[14]

T5 $F$-test of equality of the coefficients for U and M from L2.

The probit and logit models are estimated using age, age squared, a dummy for poor English-speaking ability (self-assessed), a dummy variable for being in New South Wales, a dummy for living in a capital city, dummies for educational attainment, experience, experience squared, partner's wage, total unearned household income, number of resident children less than age 5, resident children aged 5–14, resident children aged >14, and non-resident children, a dummy variable if the individual

[13] Note that this is akin to the approach taken in Hausman and McFadden (1984)

[14] Another alternative would be the LR test of Cramer and Ridder (1991). In practice, this gives results very similar to test T5 and we do not report those results here.

owns their own home, a dummy if the individual is a public tenant, and dummies for imputed household income and imputed partner's wage.

Table A5 contains a list of all the variables used in these regressions and their means and standard deviations from the fifth wave of the data.[15] We exclude any variables which do not vary for the sub-sample of interest (e.g. we exclude *male* from the sub-sample of married men). We estimated all models with indicator variables if any of the wage or unearned income data were imputed (see footnote 10).

For each of our four sub-samples (married women, married men, lone parents and singles[16]) we conduct tests T1–T5 on each wave of data. We also conduct the tests on the data pooled across all five waves. For the pooled models, we conduct the Hausman tests in two different ways. We use the standard variance matrix of parameters uncorrected for the clustering which is created by the presence of multiple observations on the same individual in the pooled sample. We also conduct the Hausman tests using a variance matrix which is corrected for clustering using a standard outer-product correction. Neither are strictly correct, as the former does not account for the clustering and the latter is not strictly theoretically consistent with the Hausman test. Conclusions from the tests are consistent across both methods, however.[17]

The test results for married women are summarised in Table 5.[18] We focus on this group because as they are a majority of the non-employed and a majority of those who transit to employment. Second, they are a frequent focus of government policy. Given current high employment in Australia, recent reforms to the tax and transfer system have been designed, at least in part, to induce married women who are not in employment to enter the labour force and to enter employment (see Centrelink, 2008).

For married women, we find consistent evidence across all waves that the unemployed, the not in the labour force and the marginally attached are three distinct categories. Over 80 per cent of the wave-by-wave tests show significant differences and we find significant differences for all of the tests where we pool the data across waves.

In the last panel of Table 5, we combine the marginally attached and the NILF as is done in the official statistics and test whether this combined group can be pooled with the unemployed. We can conclude from those tests that this combined group is also statistically significantly different from the unemployed. As the wage predictions which we discuss in Section V are often estimated from models using ABS data which combine these two groups, we provide this test.

The conclusion we draw from these results is that it is a mistake to pool the unemployed, the marginally attached and the not in the labour force and to treat them identically in modelling the probability of employment. This conclusion points the way to two possible modelling strategies for wage equations which correct for selection into employment. The first is to model the employed with each of the non-employed groups separately. This would suggest separate estimation of three Heckman selection models for the three different groups. The problem with this strategy is that it is not clear which set of estimates one should use for understanding and predicting wages for the employed. A second modelling strategy which follows from these tests is to control for sample selection using the multinomial choice models discussed in Section II.[19]

In this paper, our main focus is on those who are not in employment. We examine in the next section whether the results we have presented have any implications for predicted wages for *non-workers*. For all three groups of

[15] Detailed descriptive statistics for other waves are available from the authors.

[16] As noted previously, we pool single men and women due to small sample sizes.

[17] We only report the results using the variance–covariance matrix which is not corrected for clustering.

[18] Results for lone parents, married men and singles are available from the authors. Because of the small sample sizes, we generally find no differences for the individual waves. However, we find that the three states are distinct in the pooled tests for all three groups.

[19] Instead of generating predicted wages from a selection model which are then plugged back into the labour supply model, another alternative is to jointly model labour supply and the wage equation, with four possible labour market states, and simultaneously estimate wages and labour supply. For a three-state example, see Breunig *et al.* (2008)

TABLE 5
*Married Women: Can We Pool the Unemployed, the Marginally Attached and Not in the Labour Force?*

| Test[a] | Wave 1 | 2 | 3 | 4 | 5 | Pooled |
|---|---|---|---|---|---|---|
| Can we pool the not in the labour force and the unemployed? | | | | | | |
| T1 | 0.00** | 0.05* | 0.21 | 0.01** | 0.00** | 0.00** |
| T2 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| T3 | 0.04** | 0.05** | 0.24 | 0.00* | 0.00** | 0.00** |
| T4 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| T5 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| Can we pool the marginally attached and the unemployed? | | | | | | |
| T1 | 0.01** | 0.08** | 0.66 | 0.00** | 0.00** | 0.00** |
| T2 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| T3 | 0.03** | 0.10* | 0.61 | 0.00* | 0.00** | 0.00** |
| T4 | 0.00** | 0.06* | 0.19 | 0.01** | 0.00** | 0.00** |
| T5 | 0.00** | 0.05* | 0.37 | 0.01** | 0.00** | 0.00** |
| Can we pool the marginally attached and the not in the labour force? | | | | | | |
| T1 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| T2 | 0.00** | 0.02** | 0.00** | 0.12 | 0.16 | 0.00** |
| T3 | 0.00** | 0.01** | 0.00** | 0.00** | 0.03** | 0.00** |
| T4 | 0.00** | 0.20 | 0.01** | 0.17 | 0.02** | 0.00** |
| T5 | 0.00** | 0.04** | 0.00** | 0.03** | 0.02** | 0.00** |
| Can we pool the unemployed and a combined group of the not in the labour force and the marginally attached? | | | | | | |
| T1 | 0.01** | 0.07* | 0.36 | 0.02** | 0.00** | 0.00** |
| T2 | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** | 0.00** |
| T3 | 0.12 | 0.15 | 0.41 | 0.15 | 0.09* | 0.00** |
| T4 | 0.00** | 0.00** | 0.01** | 0.00** | 0.00** | 0.00** |
| T5 | 0.00** | 0.00** | 0.03** | 0.00** | 0.00** | 0.00** |

*Notes*: Null hypothesis is that the two labour force states can be pooled. *P*-values for test of equality of labour force states are presented. ** and * indicate significance at the 5 and 10 per cent levels, respectively. [a]The five hypothesis tests, T1–T5, are described in detail in the text.

non-workers, we will examine the predicted wages from the different estimation strategies. We will then use the individuals who transit from non-work to employment to test which of these different estimation strategies provides the most accurate wage predictions for those non-workers who subsequently take up employment.

## *V The Accuracy of Predicted Wages Using Various Modelling Approaches*

In the previous section, we concluded that the unemployed, the marginally attached and the not in the labour force appeared to be distinct groups when modelling the probability of employment. In this section, we consider whether these results have any relevance to the accuracy of predicted wages for these three groups.

Our basic approach will be as follows. We will estimate a model for wages in a particular cross-sectional wave, say wave $t$. We will then use the estimated model to predict a wage, $\hat{w}_{it}$ for a non-employed individual. We then use an adjustment factor ($a_t$) to account for wage inflation between waves $t$ and $t + 1$ to generate a predicted wage for individual $i$ at time $t + 1$ as

$$\hat{w}_{i,t+1} = \hat{w}_{it}(1 + a_t). \qquad (15)$$

In the results presented below, we used the average increase in wages in our sample data between wave $t$ and $t + 1$ for the adjustment factor. We also experimented with using the inflation rate of average weekly earnings from the Australian Bureau of Statistics, but this did not affect our conclusions.

We examine 11 separate models for predicting the wages for married women.[20] For each model, we include all of the variables from

[20] Full results for lone parents, married men and singles are available from the authors.

Table A5. We exclude from the wage equation the variables relating to unearned income, partner's wage, resident and non-resident children, and home ownership status.

- M1 Linear regression using only the employed.
- M2 Heckman selection model using whole sample and conditional predictor of Equation (6).
- M3 Heckman selection model using whole sample and unconditional predictor of Equation (4).
- M4 Heckman selection model using only non-working population of interest (unemployed, marginally attached or not in the labour force) and conditional predictor of Equation (6).
- M5 Heckman selection model using only non-working population of interest and unconditional predictor of Equation (4).[21]
- M6 Lee selection model of Equation (10) and the conditional predictor of wages.
- M7 Lee selection model of Equation (10) and the unconditional predictor of wages.
- M8 The original multinomial model of Dubin and McFadden, Equation (12) and the conditional predictor of wages.
- M9 The original multinomial model of Dubin and McFadden, Equation (12) and the unconditional predictor of wages.
- M10 The restricted multinomial model of Bourguignon, Fournier and Gurgand, Equation (13), and the conditional predictor of wages.
- M11 The restricted multinomial model of Bourguignon, Fournier and Gurgand, Equation (13), and the unconditional predictor of wages.

For each of these, we test whether the average predicted wage ($\hat{w}_{i,t+1}$ above) is equal to the average realised wage for the three groups which transition into employment out of unemployment, marginal attachment or not in the labour force.

The results are summarised in Tables 6–8. The rows of the table present the average predicted wages for the group in question. The $p$-value of the test of equality between the predicted log wage and the actual, observed log wage for those that transition into employment are given just below the average predicted wages.[22] We also pool our predictions across all waves in column 6. Column 7 presents the pooled results, dropping wave 1. For married women, we find oddly large wages for those in wave 2 who were unemployed in wave 1 (see Table 6). There appears to be some variability in responses to wage and income questions which settles down in subsequent waves as respondents become more adept at accurately completing the questionnaire. We dropped the wave 1 to wave 2 changes to see if our results were sensitive to any potential problem. In our discussion, we will focus primarily on the pooled results rather than the wave-by-wave results. For the latter, sample sizes are sometimes fairly small and this introduces variability into the results.

### *(i) Discussion of Results*

We draw several conclusions from the results. The first conclusion is that the unconditional wage prediction from all of the models across all of the sub-groups is never statistically different from the wage prediction that one would make based upon a linear regression model estimated only on the sub-population of working individuals.

The second unambiguous conclusion from the results is that the conditional predictor which uses the estimated sample selection parameter in the prediction is highly variable. This is particularly true for the multinomial models where some of the conditional wage predictions are ludicrous. It is also true for the Heckman correction model. Looking at the pooled results in the row labelled M4 in Table 6, for example, we see that average predicted wages are nearly twice the average

[21] Note that M4 and M5 would only be sensible if the other non-working groups could be theoretically excluded from the model. M2 and M3 assume that the distribution of unobservables is the same for all the non-working groups. If the distribution of unobservables is different for the non-working groups, but the model applies to all of them, then the multinomial models are appropriate.

[22] For ease of reading, we present the wages in levels. We have used a consistent predictor of the wage level based upon the estimates of the log wage model without imposing any parametric assumptions. As the model is estimated in log wage, we present the $P$-values of the test which compares predicted with actual log wage. We do this so that our tests are not influenced by the noise generated in estimating the scaling factor which we use to inflate exp[ln(wage)] to wage level.

Table 6

*Average (Mean) Predicted Wages: Married Women Who Transit From Unemployed (U) to Employed (Predictions from Different Models)*

| From wave: / To wave: | 1 / 2 | 2 / 3 | 3 / 4 | 4 / 5 | Pooled / All waves | Pooled / W2–W5 only |
|---|---|---|---|---|---|---|
| Observed in data | 24.97 | 16.71 | 15.68 | 16.89 | 18.35 | 16.39 |
| M1 Linear regression | 20.88 (0.60) | 18.50 (0.48) | 20.00* (0.09) | 20.99 (0.16) | 20.43* (0.08) | 20.28** (0.02) |
| Heckman selection model using whole sample | | | | | | |
| M2 Conditional predictor | 14.42** (0.01) | 15.68 (0.18) | 18.44 (0.36) | 19.57 (0.35) | 16.67* (0.08) | 17.75 (0.75) |
| M3 Unconditional predictor | 20.14 (0.25) | 18.16 (0.90) | 19.82 (0.15) | 20.77 (0.21) | 19.91 (0.42) | 19.92* (0.07) |
| Heckman selection model using employed and unemployed | | | | | | |
| M4 Conditional predictor | 35.47** (0.01) | 27.53*** (0.00) | 15.68 (0.62) | 23.50** (0.04) | 39.48*** (0.00) | 38.87*** (0.00) |
| M5 Unconditional predictor | 21.08 (0.70) | 18.67 (0.37) | 19.89 (0.11) | 21.06 (0.15) | 20.72** (0.03) | 20.61*** (0.00) |
| Lee selection model | | | | | | |
| M6 Conditional predictor | 14.78** (0.02) | 15.82 (0.21) | 18.65 (0.30) | 19.42 (0.38) | 16.78* (0.10) | 17.78 (0.74) |
| M7 Unconditional predictor | 20.19 (0.26) | 18.18 (0.87) | 19.85 (0.14) | 20.75 (0.21) | 19.93 (0.40) | 19.93* (0.07) |
| Original Dubin–McFadden model | | | | | | |
| M8 Conditional predictor | 365.57*** (0.00) | 126.45*** (0.00) | 25.88*** (0.00) | 29.00*** (0.00) | 92.67*** (0.00) | 28.72*** (0.00) |
| M9 Unconditional predictor | 20.48 (0.34) | 18.91 (0.52) | 19.97 (0.11) | 21.01 (0.19) | 20.34 (0.20) | 20.45** (0.02) |
| Restricted Dubin–McFadden model | | | | | | |
| M10 Conditional predictor | 314.22*** (0.00) | 494.63*** (0.00) | 17.41 (0.70) | 12.36* (0.05) | 204.86*** (0.00) | 92.80*** (0.00) |
| M11 Unconditional predictor | 20.55 (0.40) | 19.40 (0.16) | 19.86 (0.16) | 20.68 (0.29) | 20.56** (0.05) | 20.76*** (0.00) |

*Notes*: Entries are average (mean) observed and predicted wages ($\hat{w}_{i,t+1}$). We estimate a model in ln(wage) but use a consistent predictor of the wage level from the ln(wage) model. Numbers in parentheses are *P*-values for tests of equality between average predicted log wage, $\ln(\hat{w}_{i,t+1})$ and observed log wage at time $t + 1$. *** Indicates significant difference between observed and predicted ln(wage) at 1 per cent level. ** and * indicate significance at the 5 and 10 per cent levels respectively.

Table 7

*Average (Mean) Predicted Wages: Married Women Who Transit from Not in the Labour Force (N) to Employed (Predictions from Different Models)*

| From wave:<br>To wave: | 1<br>2 | 2<br>3 | 3<br>4 | 4<br>5 | Pooled<br>All waves | Pooled<br>W2–W5 only |
|---|---|---|---|---|---|---|
| Observed in data | 18.79 | 17.89 | 22.80 | 20.77 | 20.14 | 20.63 |
| M1 Linear regression | 18.46 (0.64) | 18.70 (0.35) | 19.64 (0.18) | 21.67 (0.24) | 19.80 (0.74) | 20.31 (0.56) |
| Heckman selection model using whole sample | | | | | | |
| M2 Conditional predictor | 14.13*** (0.00) | 16.19 (0.15) | 18.43** (0.02) | 20.43 (0.64) | 16.79*** (0.00) | 18.16** (0.02) |
| M3 Unconditional predictor | 17.77** (0.04) | 18.37 (0.87) | 19.44* (0.09) | 21.52 (0.33) | 19.36 (0.20) | 20.01 (0.77) |
| Heckman selection model using employed and not in the labour force | | | | | | |
| M4 Conditional predictor | 13.57*** (0.00) | 15.40* (0.05) | 18.42** (0.02) | 19.35 (0.90) | 15.88*** (0.00) | 17.16*** (0.00) |
| M5 Unconditional predictor | 17.96 (0.12) | 18.38 (0.80) | 19.50 (0.11) | 21.43 (0.37) | 19.38 (0.27) | 19.98 (0.75) |
| Lee selection model | | | | | | |
| M6 Conditional predictor | 14.27*** (0.00) | 16.36 (0.21) | 18.57** (0.02) | 20.29 (0.70) | 16.87*** (0.00) | 18.18** (0.03) |
| M7 Unconditional predictor | 17.81** (0.05) | 18.39 (0.83) | 19.47* (0.09) | 21.51 (0.34) | 19.37 (0.22) | 20.01 (0.77) |
| Original Dubin–McFadden model | | | | | | |
| M8 Conditional predictor | 15.96*** (0.00) | 8.22*** (0.00) | 16.77*** (0.00) | 10.70*** (0.00) | 10.27*** (0.00) | 8.74*** (0.00) |
| M9 Unconditional predictor | 17.85* (0.07) | 18.73 (0.54) | 19.54 (0.12) | 21.49 (0.35) | 19.62 (0.49) | 20.31 (0.84) |
| Restricted Dubin–McFadden model | | | | | | |
| M10 Conditional predictor | 16.37*** (0.00) | 9.35*** (0.00) | 17.54*** (0.00) | 10.57*** (0.00) | 11.96*** (0.00) | 10.40*** (0.00) |
| M11 Unconditional predictor | 17.92 (0.14) | 19.03 (0.10) | 19.46* (0.07) | 21.31 (0.56) | 19.80 (0.58) | 20.56 (0.15) |

*Note*: See notes to Table 6.

TABLE 8

*Average (Mean) Predicted Wages: Married Women who Transit from Marginally Attached (M) to Employed (Predictions from Different Models)*

| From wave:<br>To wave: | 1<br>2 | 2<br>3 | 3<br>4 | 4<br>5 | Pooled<br>All waves | Pooled<br>W2–W5 only |
|---|---|---|---|---|---|---|
| Observed in data | 18.71 | 16.48 | 19.58 | 18.48 | 18.40 | 18.25 |
| M1 Linear regression | 18.51 (0.82) | 18.82* (0.09) | 18.57 (0.55) | 21.59* (0.09) | 19.52 (0.16) | 20.13* (0.06) |
| Heckman selection model using whole sample | | | | | | |
| M2 Conditional predictor | 14.10*** (0.00) | 16.21 (0.49) | 17.30* (0.08) | 20.30 (0.79) | 16.45*** (0.00) | 17.91 (0.23) |
| M3 Unconditional predictor | 17.68* (0.06) | 18.46 (0.31) | 18.38 (0.34) | 21.49 (0.14) | 19.08 (0.79) | 19.85 (0.32) |
| Heckman selection model using employed and marginally attached | | | | | | |
| M4 Conditional predictor | 10.90*** (0.00) | 14.85 (0.13) | 15.84*** (0.01) | 25.54*** (0.00) | 14.19*** (0.00) | 17.56 (0.21) |
| M5 Unconditional predictor | 17.64* (0.08) | 18.58 (0.19) | 18.34 (0.35) | 21.70* (0.06) | 19.15 (0.83) | 19.99 (0.13) |
| Lee selection model | | | | | | |
| M6 Conditional predictor | 14.28*** (0.00) | 16.35 (0.59) | 17.45 (0.11) | 20.16 (0.44) | 16.53*** (0.00) | 17.92 (0.24) |
| M7 Unconditional predictor | 17.74* (0.07) | 18.49 (0.28) | 18.40 (0.36) | 21.48 (0.14) | 19.10 (0.83) | 19.86 (0.31) |
| Original Dubin–McFadden model | | | | | | |
| M8 Conditional predictor | 4.51*** (0.00) | 45.46*** (0.00) | 20.98 (0.36) | 113.75*** (0.00) | 30.19*** (0.00) | 121.80*** (0.00) |
| M9 Unconditional predictor | 17.83 (0.11) | 18.88 (0.16) | 18.52 (0.45) | 21.73 (0.10) | 19.39 (0.67) | 20.24 (0.11) |
| Restricted Dubin–McFadden model | | | | | | |
| M10 Conditional predictor | 7.62*** (0.00) | 53.25*** (0.00) | 17.50* (0.09) | 91.96*** (0.00) | 30.87*** (0.00) | 99.00*** (0.00) |
| M11 Unconditional predictor | 17.92 (0.20) | 19.23** (0.03) | 18.40 (0.29) | 21.62 (0.22) | 19.57* (0.09) | 20.46*** (0.00) |

*Note*: See notes to Table 6.

actual wage. This problem arises in part because the sample selection term is often estimated with very low precision, at least in part due to small sample sizes. The estimates of the sample selection term are also unstable – switching between negative and positive for different waves of data using the same population.[23]

The third conclusion is that there is no obvious gain from using a multinomial model relative to a simple Heckman correction model. The conditional predictors from those models, as discussed above, are highly unstable. The unconditional predictors do not vary much from the unconditional predictor from the Heckman model nor from the linear predictor from an OLS regression on the selected sample.

Our fourth conclusion is that a simple linear predictor from a regression on the selected sample or the unconditional predictor from the sample selection model very often outperforms the conditional predictor which uses information from the sample selection correction. This is certainly the case for married women who move from not in the labour force to employment (M1, M3 and M5 in Table 7) and who move from marginal attachment to employment (M3 and M5 in Table 8).

For married women who move from unemployment to employment, the results are more mixed. The unconditional predictor works better (M3 of Table 6) across all waves, but if we consider only the last four waves, then the conditional predictor works better. Given the extreme observation for average wages for those who move from unemployment in wave 1 to employment in wave 2, we might prefer the conditional predictor for this group. However, if we estimate the model only on the employed and unemployed (dropping the not in the labour force and the marginally attached), then the conditional predictor performs very poorly. This is probably due to the small sample size, but it is somewhat disturbing that the conditional predictor performs so differently in rows M2 and M4 of Table 6.

### *(ii) Comparing the Distributions of Predicted and Actual Wages*

In Section V(i), we considered the differences in the mean of actual and predicted wages. These comparisons may be sensitive to influential observations. To check this, we tested whether the medians of the predicted and actual wages were different for all of the models of Tables 6–8.[24] If we consider the tests on the pooled data from the last two columns of Tables 6–8, the test of median equality and that of mean equality agree (in terms of whether the differences are significant) for every single case except for the unconditional predictor from the original Dubin–McFadden model for unemployed, married women. We find that the mean wage from this model is not significantly different from the mean actual wage ($P$-value of 0.2, see row M9, column 6 of Table 6), whereas we find that the medians are significantly different ($P$-value of 0.03). Given the over-whelming similarity between the results for medians and those for means, our main conclusions from Section V(i) also appear to apply to the medians.[25]

We also conducted Kolmogorov–Smirnov tests on the equality of the distributions of predicted and actual wages for the pooled models. For unemployed married women, the Kolmogorov–Smirnov test always agrees (in terms of statistical significance/insignificance) with the tests of means shown in Table 6. For the not in the labour force and the marginally attached, we always reject that the distributions are identical, even when we find that the means of the distribution and medians of the distribution are not statistically different.

To shed light on this latter result, Figures 2–4 provide non-parametric density estimates of the actual ln(wage) and the predicted ln(wage) from the Heckman selection model estimated using the whole sample and pooling waves 2–5. Figure 2 corresponds to models M2 and M3 of Table 6, Figure 3 corresponds to those models from Table 7, and Figure 4 corresponds to those models from Table 8. Typically, predictions from wage models provide much more concentrated distributions than that of actual wages. We can clearly see this in all three figures. It is this large peak and failure to capture

[23] For the Heckman selection models, Table A4 provides a summary of the sign and significance of the estimated sample selection correction parameter.

[24] We are grateful to an anonymous referee who suggested the additional tests and comparisons of this section.

[25] The vast majority of tests on the wave-by-wave comparisons were also very similar for means and medians. Versions of Tables 6–8 with median wages and the results of the tests are available from the authors.

FIGURE 2
*Actual and Predicted ln(wage) Distributions: Unemployed Married Women Who Become Employed*

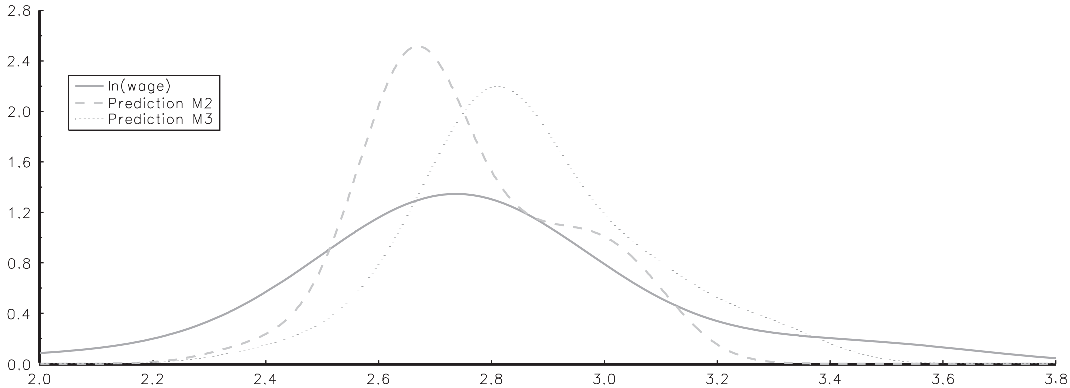
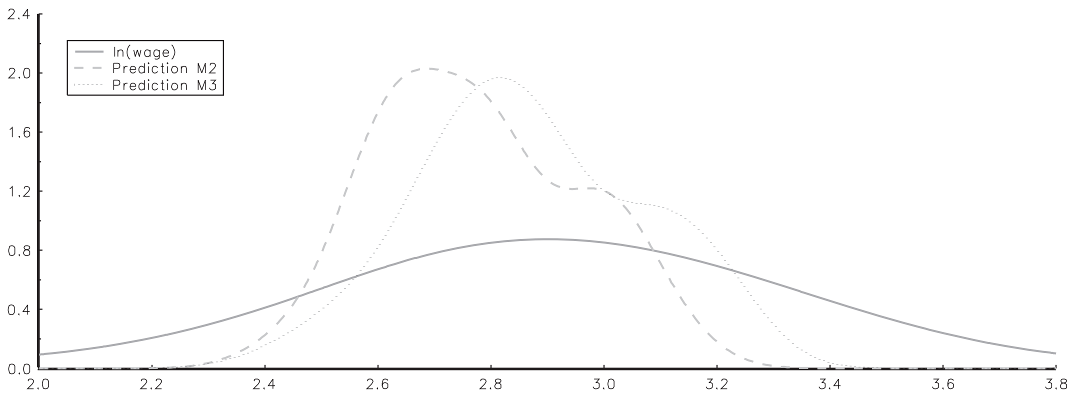



FIGURE 3
*Actual and Predicted ln(wage) Distributions: Not in the Labour Force Married Women Who Become Employed*



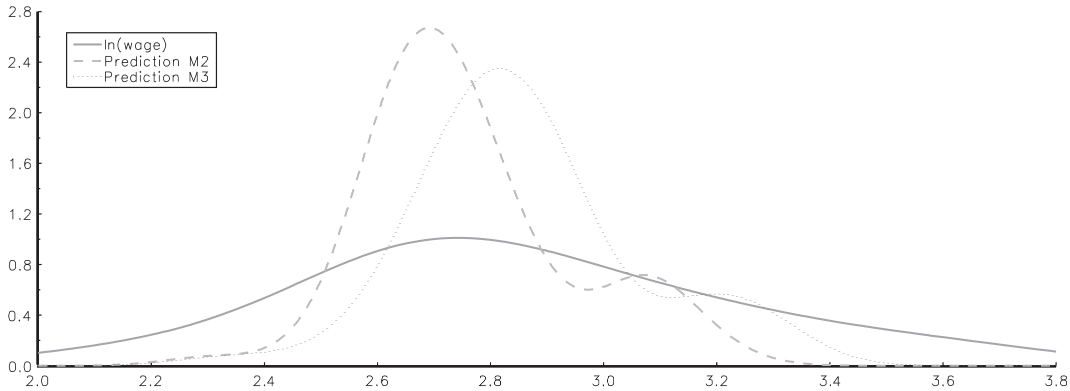

the tails of the actual wage distribution that lies behind the rejection of distribution equality we find in the Kolmogorov–Smirnov test. The fact that we do not reject this equality for the unemployed is, we belief, primarily driven by the smaller sample sizes for that sub-group. In both Figures 3 and 4, we can see that the predicted wages which use the conditional predictor lie well to the left of the actual distribution of wages and that the peak of the predicted and actual wage distributions are quite far apart. The peak of the distribution of wages using the unconditional predictor lies quite close to the peak of the actual wage distribution; hence, our failure to reject the null that the average predictions are equal to the average wages.

Recall that for the unemployed, we preferred the conditional predictor for this model on the basis of the test of means and medians. However, looking at Figure 2, the peak of the unconditional predictor actually appears closer to the peak of the actual wage distribution. The conditional predictor performs better on the tests of means and medians because of the relatively thicker left-hand tail in the wage distribution.

### *(iii) Selection Amongst the Newly Employed*

One might worry that those non-employed individuals at period $t$ who become employed at period $t + 1$ are not a random sample from the group of non-employed but are themselves

FIGURE 4
*Actual and Predicted ln(wage) Distributions: Marginally Attached Married Women Who Become Employed*




a selected sample with unobservable characteristics better than the average non-employed person. In that case, our tests may be interpreted as a test for the best predictor of wages conditional on actually taking up employment in subsequent periods. For some types of policy simulations, this may be the relevant predicted wage.

It is very difficult to get a good estimate of the unobservable characteristics for those who never take up employment. For those who move from non-employment to employment, we can estimate the unobservable effects on wages through the residual from the wage regression at time $t + 1$.[26] If we take the residuals from a wage regression estimated on the entire pooled sample of individuals who are employed and then run a regression on a set of dummy variables which indicate the previous employment status (one wave prior), we find significantly negative effects of having been either unemployed or marginally attached in the previous period.[27] The unobservables for the previously not in the labour force are less than those of the previously employed, on average, but the difference is not significant.

We find this result reassuring in regard to the amount of selection that might be present in our sample which moves from non-employment to employment. We expect, *a priori*, that the unemployed and the marginally attached might have poorer unobservable labour market characteristics than the employed and this is in fact what we find.

## *VI Discussion and Conclusions*

In a model of the probability of employment, we find that the unemployed, the marginally attached and the not in the labour force appear to be three distinct groups. This result is consistent across several different types of models and different specification tests. The implication is that these three groups should not be pooled together into one 'non-employed' group in a joint model of wages and employment. Our conclusion is based on specification tests of cross-sectional models which classify individuals into one or another category. Looking at transitions to employment, Gray *et al.* (2005) are led to similar conclusions for Australia using data covering the period 1994–1997. Applying similar tests to the transitions in our data, we come to the same conclusion for the 2001–2005 period.

Building upon these results, we examine the wage predictions from a variety of models beginning with a simple linear regression model which has no controls for selection into employment to more complicated models which allow for multiple non-employment states. We find that the linear predictor from a regression on the selected sample of workers almost always outperforms more complicated prediction strategies. Interestingly, this is the same conclusion that is reached by Duan *et al.* (1983) for the

[26] This will be independent of the estimate of the correlation between utility of employment and wages estimated at time $t$.

[27] We control for the clustering induced by the pooling of individuals over time.

problem of predicting the dependent variable for the *selected* sample. Our paper is the first to examine the question of predictive power for the *non-selected* sample.

The linear predictor does not always provide unbiased estimates of future wages, but it is less prone to very large errors than conditional prediction based upon a sample selection model. This is primarily driven by the instability and imprecision of the estimated coefficient on the sample selection term in the second stage of the two-step modelling procedure. More complicated multinomial models appear to suffer from these problems to a greater degree than the binary Heckman selection model.

For married women in the not in the labour force and marginally attached categories, the selection model does not seem to provide information about the average effect of unobservables on wages through the correlation with the selection equation. One possible explanation is that the decision to move from one of these non-employed states to employment is accompanied by a change in the relationship between the reservation wage and the distribution of wage offers as discussed in Section II above.

For unemployed, married women, however, there does seem to be information in the selection model regarding unobservables, although this result is somewhat sensitive to sample period and size. The result is consistent with a model that views unemployment as the state in which individuals better understand their reservation wage and truly are ready to take up employment if the right offer comes along.

One important caveat to our results is that they are based upon fairly small sample sizes. When pooling across all five waves, we have around 6000 married women but only approximately 200 each in the categories of unemployed, marginally attached and not in the labour force. These sample sizes are not out of line with uses of selection models in the Australian and overseas literature. Papers such as Duncan and Harris (2002), Rammohan and Whelan (2005) and Breunig *et al.* (2008) use smaller sample sizes in their selection models. Grogger (1998) uses a sample from the USA with approximately 1100 individuals of whom 5.2 per cent are non-employed. Beblo *et al.* (2003) use sample sizes ranging from 963 to 3235 in estimating selectivity-corrected gender wage gaps for a variety of European countries. Kalb (2002) pools across four waves of the Survey of Income and Housing Costs to get sample sizes that are approximately 50 per cent larger than those used here. The data we use, HILDA, is the only dataset in Australia which allows us to observe subsequent employment and wage outcomes for the previously non-working. It would be desirable to implement these tests on a much larger dataset, but this is currently not possible in Australia. Despite the small samples, the results are of interest as the HILDA data are widely used for research in labour economics in Australia.

All selection models start by making functional form assumptions about the relationship between the unobservables in the equations which determine wages and participation. We can think of sample selection correction terms as then being estimated upon the residuals from these equations. Two problems arise. The first is that, as in many models where we are estimating functions of inferred quantities, large sample sizes are required for the (desirable) asymptotic properties of the estimates to hold. Second, our procedure is dependent upon non-testable distributional assumptions, the failure of which may lead to biased or highly variable estimates. *A priori*, therefore, one might be suspicious about the use of the conditional predictor. Our results, although based on small samples, add an additional cautionary note to the use of wage prediction augmented with information from sample selection terms.

Some labour supply models use predicted wages in policy simulations to consider the likely employment outcomes from changes to the tax and transfer system. Our paper provides two lessons for individuals who are estimating such models using survey data. The first is that a one-size-fits-all approach to predicting wages for those who are not working may be inappropriate as the unemployed, the marginally attached and the not in the labour force appear to behave quite differently. The second conclusion is that prediction using a simple linear regression on the selected sample may be desirable on the grounds that it is less subject to assumptions about distributional assumptions and small sample sizes.

## REFERENCES

Ahn, S.C and Low, S. (2007), 'A Parametric Test for the Distinction Between Unemployed and Out of the Labour Force Statuses', *Seoul Journal of Economics*, **20**, 59–87.

Australian Bureau of Statistics (2007). Labour Force, Australia. Detailed Electronic Delivery, March 2007–Gross Flows (ST GM1), Australian Bureau of Statistics. Datacube 6291.0.55.001.

Beblo, M., Beninger, D., Heinze, A. and Laisney, F. (2003), 'Measuring Selectivity-Corrected Gender Wage Gaps in the EU', Centre for European Economic Research (ZEW) Discussion Paper No. 03-74, ZEW, Mannheim, Germany.

Bourguignon, F., Fournier, M. and Gurgand, M. (2007), 'Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons', *Journal of Economic Surveys*, **21**, 174–205.

Brandolini, A., Cipollone, P. and Viviano, E. (2006), 'Does the ILO Definition Capture all Unemployment?' *Journal of the European Economic Association*, **4**, 153–79.

Breunig, R., Cobb-Clark, D. and Gong, X. (2008), 'Improving the Modeling of Couples' Labour Supply', *Economic Record*, **84**, 466–85.

Centrelink, (2008). *Welfare to work*. Centrelink Publication, Commonwealth of Australia 2008. [Cited 17 July 2009.] Available from http://www.centrelink.gov.au/internet/internet.nsf/services/welfare_work.htm.

Cramer, J. and Ridder, G. (1991), 'Pooling States in the Multinomial Logit', *Journal of Econometrics*, **47**, 267–72.

Dahl, G.B. (2002), 'Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets', *Econometrica*, **70**, 2367–420.

Dolton, P. J. and Makepeace, G. H. (1987), 'Interpreting Sample Selection Effects', *Economics Letters*, **24**, 373–9.

Duan, N., Manning, W.G., Morris, C.N. and Newhouse, J.P. (1983), 'A Comparison of Alternative Models for the Demand for Medical Care', *Journal of Business and Economic Statistics*, **1**, 115–26.

Dubin, J.A. and McFadden, D.L. (1984), 'An Econometric Analysis of Residential Electric Appliance Holdings and Consumption', *Econometrica*, **52**, 345–62.

Duncan, A. and Harris, M. (2002), 'Simulating the Behavioural Effects of Welfare Reforms Among Sole Parents in Australia', *Economc Record*, **78**, 264–76.

Ermisch, J.F. and Wright, R.E. (1994), 'Interpretation of Negative Sample Selection Effects in Wage Offer Equations', *Applied Economics Letters*, **1**, 187–9.

Flinn, C.J. and Heckman, J.J. (1983), 'Are Unemployment and Out of the Labour Force Behaviorally Distinct Labor Force States?' *Journal of Labor Economics*, **1**, 28–42.

Frick, J.R. and Grabka, M.M. (2007), 'Item Non-Response and Imputation of Labor Income in Panel Surveys from a Cross-National Comparison', IZA Discussion Paper Number 3043, Institute for the Study of Labor, Bonn, Germany.

Gonal, F. (1992), 'New Evidence on Whether Unemployment and Out of the Labor Force are Distinct States', *Journal of Human Resources*, **27**, 329–61.

Gray, M., Heath, A. and Hunter, B. (2005), 'The Labour Force Dynamics of the Marginally Attached', *Australian Economic Papers*, **44**, 1–14.

Grogger, J. (1998), 'Market Wages and Youth Crime', *Journal of Labor Economics*, **16**, 756–91.

Hausman, J. (1978), 'Specification Tests in Econometrics', *Econometrica*, **46**, 1251–71.

Hausman, J.A. and McFadden, D. (1984), 'Specification Tests for the Multinomial Logit Model', *Econometrica*, **52**, 1219–31.

Heckman, J.J. (1974), 'Shadow Prices, Market Wages and Labor Supply', *Econometrica*, **42**, 679–94.

Heckman, J.J. (1979), 'Sample Selection Bias as a Specification Error', *Econometrica*, **47**, 153–61.

Hoffmann, R. and Kassouf, A.L. (2005), 'Deriving Conditional and Unconditional Marginal Effects in Log Earnings Equations Estimated by Heckman's Procedure', *Applied Economics*, **37**, 1303–11.

Jones, S.R.G. and Riddell, W.C. (1999), 'The Measurement of Unemployment: an Empirical Approach', *Econometrica*, **67**, 147–61.

Jones, S.R.G. and Riddell, W.C. (2006), 'Unemployment and Nonemployment: Heterogeneties in Labor Market States', *The Review of Economics and Statistics*, **88**, 314–23.

Kalb, G.R. (2002), 'Estimation of Labour Supply Models for Four Separate Groups in the Australian Population', Melbourne Institute Working Paper No. 24/02. [Cited 5 January 2010.] Available from: http://melbourneinstitute.com/wp/wp2002n24.pdf.

Kalb, G.R. and Lee, W.-S. (2008), 'Childcare Use and Parents Labour Supply in Australia', *Australian Economic Papers*, **47**, 272–95.

Lee, L.-F. (1983), 'Generalized Econometric Models with Selectivity', *Econometrica*, **51**, 507–12.

Little, R.J.A. and Su, H.L. (1989), 'Item Nonresponse in Panel Surveys', in Kasprzyk D., Duncan G.J., Kalton G. and Singh M.P. (eds), *Panel Surveys*. Wiley, New York; 400–25.

McFadden, D.L. (1973), 'Conditional Logit Analysis of Qualitative Choice Behavior', in Zarembka P. (ed.), *Frontiers in Econometrics*. Academic Press, New York; 105–42.

Puhani, P.A. (2000), 'The Heckman Correction for Sample Selection and its Critique', *Journal of Economic Surveys*, **14**, 53–68.

Rammohan, A. and Whelan, S. (2005), 'Child Care and Female Employment Decisions', *Australian Journal of Labour Economics*, **8**, 203–25.

Ross, R.T. (1986), 'Analysis of the 1980 Sydney Survey of Work Patterns of Married Women: Further Results', *Economic Record*, **62**, 325–37.

Schaffner, J.A. (1998), 'Generating Conditional Expectations from Models with Selectivity Bias: Comment', *Economics Letters*, **58**, 255–61.

Schmertmann, C.P. (1994), 'Selectivity Bias Correction Methods in Polychotomous Sample Selection Models', *Journal of Econometrics*, **60**, 101–32.

Starick, R. and Watson, N. (2007), 'Evaluation of Alternative Income Imputation Methods for the HILDA Survey', Technical Report. HILDA Project

Discussion Paper Series No. 1/07, Melbourne Institute of Applied Economic and Social Research, Melbourne.

Stolzenberg, R.M. and Relles, D.A. (1997), 'Tools for Intution about Sample Selection Bias and its Corrections', *American Sociological Review*, **62**, 494–507.

Tano, D.K. (1991), 'Are Unemployment and Out of the Labour Force Behaviorally Distinct Labor Force States? New Evidence from the Gross Change Data', *Economics Letters*, **36**, 113–7.

Vella, F. (1988), 'Generating Conditional Expectations from Models with Selectivity Bias', *Economics Letters*, **28**, 97–103.

Watson, N. (2004), 'Income and Wealth Imputation for Waves 1 and 2', Technical Report. HILDA Project Technical Paper Series No. 3/04, Melbourne Institute of Applied Economic and Social Research, Melbourne.

Watson, N. and Wooden, M. (2002), 'The Household, Income and Labour Dynamics in Australia (HILDA) Survey: Wave 1 Survey Methodology', Technical Report. HILDA Project Technical Paper Series No. 1/02, Melbourne Institute of Applied Economic and Social Research, Melbourne.

## *Appendix*

TABLE A1
*Sample Sizes by Wave, Employment Status, Gender and Marital/Parental Status*

| | | Wave | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Married men | E | 1802 | 1519 | 1484 | 1385 | 1418 |
| | U | 77 | 47 | 43 | 22 | 30 |
| | M | 24 | 31 | 24 | 24 | 20 |
| | N | 35 | 48 | 23 | 24 | 19 |
| Married women | E | 1365 | 1140 | 1114 | 1075 | 1131 |
| | U | 44 | 40 | 40 | 31 | 39 |
| | M | 205 | 148 | 133 | 92 | 86 |
| | N | 324 | 317 | 287 | 257 | 231 |
| Single men | E | 350 | 327 | 331 | 350 | 344 |
| | U | 40 | 32 | 24 | 25 | 17 |
| | M | 15 | 20 | 19 | 17 | 12 |
| | N | 10 | 19 | 5 | 8 | 6 |
| Single women | E | 288 | 284 | 289 | 275 | 284 |
| | U | 17 | 16 | 13 | 10 | 10 |
| | M | 9 | 6 | 9 | 10 | 10 |
| | N | 11 | 23 | 7 | 8 | 11 |
| Lone parents | E | 305 | 300 | 315 | 303 | 322 |
| | U | 37 | 26 | 24 | 27 | 33 |
| | M | 99 | 80 | 67 | 67 | 63 |
| | N | 70 | 76 | 74 | 75 | 65 |

*Notes*: E, employed; M, marginally attached; N, not in the labour force; U, unemployed.

TABLE A2
*Number (per cent) of Individuals Entering Employment by Year and Employment State in Previous Month*

| Year | From unemployed | From NILF | Total |
|---|---|---|---|
| 2001 | 59.7 (19.7) | 111.1 (7.0) | 170.8 (9.0) |
| 2002 | 61.4 (20.6) | 98.1 (6.2) | 159.6 (8.4) |
| 2003 | 60.2 (21.1) | 105.5 (6.6) | 165.7 (8.8) |
| 2004 | 56.8 (22.4) | 106.5 (6.6) | 163.3 (8.8) |
| 2005 | 56.6 (24.0) | 106.7 (6.9) | 163.3 (9.2) |
| 2006 | 57 (24.3) | 112.8 (7.4) | 169.8 (9.7) |

*Notes*: Average of monthly transitions over calendar year, individuals' ages 25–59.
*Source*: Australian Bureau of Statistics (2007) Labour Force Survey.

TABLE A3
*Number (per cent) of Individuals Entering Employment by Employment State in Previous Month*

| Subgroup | From unemployed | From NILF | Total |
|---|---|---|---|
| Men | 32.2 (22.3) | 34 (8.2) | 66.2 (11.8) |
| Women | 26.4 (21.4) | 72.8 (6.3) | 99.1 (7.7) |
| Total | 58.6 (21.8) | 106.8 (6.8) | 165.4 (9.0) |

*Notes*: Average of monthly transitions from 2001–2006 by sex, individuals' ages 25–59.
*Source*: Australian Bureau of Statistics (2007) Labour Force Survey.

Table A4
*Sign and Significance of Heckman Correction Term in Models of Tables 6–8 (Married Women)*

| | Wave | | | | | Pooled 1–5 | Pooled 2–5 |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | |
| Whole sample (E, U, M, N) (Tables 6–8) | +*** | +** | + | + | + | +*** | +*** |
| Reduced sample (E, U) (Table 6) | − | − | + | − | − | −*** | −** |
| Reduced sample (E, M) (Table 7) | +*** | + | + | − | + | +*** | + |
| Reduced sample (E, N) (Table 8) | +*** | +*** | + | + | +** | +*** | +*** |

*Notes*: *** indicates significance at the 1 per cent level. ** and * indicate significance at the 5 and 10 per cent levels respectively.

Table A5
*Key Variables: Wave 5 Averages and Standard Deviations*

| Subgroup | Married men | Married women | Single men | Single women | Lone parents |
|---|---|---|---|---|---|
| Observations | 1487 | 1487 | 379 | 315 | 483 |
| Proportion working | 0.954 (0.21) | 0.761 (0.427) | 0.908 (0.29) | 0.902 (0.298) | 0.667 (0.472) |
| Age/100 | 0.416 (0.087) | 0.395 (0.084) | 0.398 (0.093) | 0.43 (0.107) | 0.42 (0.082) |
| $(\text{Age}/100)^2$ | 0.181 (0.073) | 0.163 (0.068) | 0.167 (0.076) | 0.196 (0.09) | 0.183 (0.068) |
| Poor English | 0.008 (0.089) | 0.012 (0.109) | 0 (0) | 0.003 (0.056) | 0.017 (0.128) |
| NSW | 0.305 (0.46) | 0.305 (0.46) | 0.28 (0.449) | 0.26 (0.44) | 0.302 (0.46) |
| Capital city | 0.633 (0.482) | 0.633 (0.482) | 0.652 (0.477) | 0.692 (0.462) | 0.594 (0.492) |
| University degree | 0.298 (0.457) | 0.319 (0.466) | 0.23 (0.421) | 0.375 (0.485) | 0.203 (0.403) |
| Trade, diploma, or certificate | 0.426 (0.495) | 0.247 (0.431) | 0.414 (0.493) | 0.308 (0.462) | 0.35 (0.477) |
| Less than Year 12 schooling | 0.179 (0.383) | 0.28 (0.449) | 0.24 (0.428) | 0.203 (0.403) | 0.321 (0.467) |
| Experience | 22.961 (9.581) | 16.763 (8.849) | 19.913 (10.44) | 20.451 (10.87) | 17.538 (10.388) |
| $\text{Experience}^2/100$ | 6.189 (4.542) | 3.593 (3.395) | 5.052 (4.638) | 5.36 (4.631) | 4.153 (3.877) |
| Partner's wage/100 | 5.438 (4.786) | 11.842 (7.24) | n/a | n/a | n/a |
| Unearned income/1000 | 4.174 (24.02) | 4.174 (24.02) | 2.887 (10.268) | 2.391 (19.809) | 4.301 (14.321) |
| Resident children age (years) | | | | | |
| 0–4 | 0.258 (0.437) | 0.26 (0.439) | n/a | n/a | 0.164 (0.37) |
| 5–14 | 0.41 (0.492) | 0.455 (0.498) | n/a | n/a | 0.619 (0.486) |
| 15–24 | 0.217 (0.412) | 0.241 (0.428) | n/a | n/a | 0.476 (0.5) |
| Has non-resident children | 0.191 (0.393) | 0.147 (0.354) | 0.343 (0.475) | 0.149 (0.357) | 0.255 (0.436) |
| Public tenant | 0.009 (0.093) | 0.009 (0.093) | 0.021 (0.144) | 0.044 (0.206) | 0.104 (0.305) |
| Outright home owner | 0.247 (0.432) | 0.247 (0.432) | 0.158 (0.366) | 0.235 (0.425) | 0.172 (0.378) |
| = 1 if variable is imputed | | | | | |
| Unearned income | 0.072 (0.259) | 0.069 (0.254) | 0.095 (0.294) | 0.111 (0.315) | 0.118 (0.323) |
| Wage | 0.02 (0.141) | 0.017 (0.129) | 0.018 (0.135) | 0.041 (0.199) | 0.017 (0.128) |
| Partner's wage | 0.017 (0.129) | 0.02 (0.141) | n/a | n/a | n/a |
| Male | 1 | 0 | 1 | 0 | 0.124 (0.33) |