"Work-in Progress"
Draft only, not for quotation.

Modeling zero-inflated longitudinal data: Changes in hours worked by women with dependent children.

Michele Haynes & Justine Gibbings*

The University of Queensland Social Research Centre
The University of Queensland

*Research and Analysis Branch
Australian Government
Department of Families, Community Services and Indigenous Affairs

Paper prepared for

ACSPRI Social Science Methodology Conference,
University of Sydney
December 2006

Abstract

Analysis of the first three waves of HILDA survey data has shown that the number of hours worked by women is strongly related to time-invariant characteristics of the individual such as level of education, as well as characteristics that change over time including the number of children (Gibbings & Heyworth, 2005). However, the distribution of the number of hours worked is characterised by a large number of zero observations (excess zeros) that arise for women who are not employed. It seems reasonable to propose that the covariates that influence the process of working or not may differ from those which influence the number of positive work hours (the intensity of work) and this should be accounted for in the statistical model.

In this study we consider data from four waves of the HILDA survey and apply a two-part or mixed regression model with correlated random effects to jointly examine the associations between selected covariates and the processes related to women's work hours. The statistical model is implemented using a SAS macro as described in Tooze et al. (2002).

In addition to demonstrating the application of this more sophisticated statistical model for analysing longitudinal semi-continuous data with excess zeros, we present substantive results that include the findings that the presence of children of any age is associated with a reduction in hours worked, the intensity of work is lowest when a child under five is present in the household, post-secondary education has a significant positive effect on both the likelihood of being employed and the number of hours worked and that higher education and good health have a protective effect on employment. We also found that for women who had children of any age, and no additional birth, the number of hours worked increased at a steady rate over the three years of the survey.

1 Introduction

Employment patterns of women have long been of interest to governments and academics. The large volume of literature on this topic shows that a range of social, psychological and economic factors may be important influences upon women's participation in paid employment. These factors include attitudes towards work and family, the role of paid work in the socio-economic status of individuals, and the marginal tax rates experienced by individuals when they enter or re-enter the workforce.

1.1 Women's employment participation

Women's participation can be conceptualised has having a number of dimensions. Employment status is classified by the Australian Bureau of Statistics (ABS) as: employed, unemployed and not in the labour force[1]. For those employed, another aspect is the intensity of work. Intensity is a rate of working and a variety of time periods can be used. For those with jobs that are characterised by regular hours and which have been held for a year, the rate per week or per year is equivalent. The more variety in an individual's work over a year, the more likely it is that the number of hours worked over a year will not be equivalent to the current hours worked in a week. A third characteristic is the nature of the tasks undertaken in the job – these can be classified in a systematic way using the ABS Australian Standard Classification of Occupations (ASCOII)[2].

In HILDA[3], the characteristics of employment are recorded in a number of ways. There is a very detailed calendar which provides details of employment status throughout each month for the previous financial year. It also provides information about either full or part-time study undertaken. There is no information about the intensity of work. However, hours (usually) worked per week is available for the current job at the time of interview. If the individual has more than one job, the total of hours worked in all jobs is recorded. Jobs are also classified according to the ASCO II standards.

The presence of dependent children and also their ages and number is a key variable in understanding women's employment patterns. In previous work (Gibbings & Heyworth, 2005) we showed that a life cycle approach which models employment participation over time provides greater insight into this issue than a model which describes employment as a simple linear process. Analysis of the first three waves of HILDA survey data has shown that time-invariant characteristics such as the level of education of women as well as characteristics that change over time such as the number and ages of children, have a strong relationship with the number of hours worked by women. To capture the effect of change in an individual's work hours over time with changes in the composition of a family with regards to number and age of children, we constructed a grouping variable which was based on these family characteristics measured over the first three waves of the HILDA survey. The construction of these groups is detailed below in Section 2.

In previous analyses the number of hours currently worked in a week has been analysed as a continuous variable with a normal probability distribution. Typically, however, the

[1] Labour Statistics: Concepts, Sources and Methods (cat. no. 6102.0.55.001)
[2] Catalogue no. 1220.0.30.001. Recently updated to the Australian New Zealand Standard Classification of Occupations (ANZCO) 2006, Catalogue no 1220.0
[3] Household, Income and Labour Dynamics in Australia, release 4.1

distribution of the number of hours worked is characterised by a large number of observations at zero which occur for women who are not employed at the time of interview. It seems reasonable to propose that the covariates that influence the process of working or not may differ from those which influence the number of positive work hours. If a covariate does influence both processes then it will likely differ in its effect size. To model both processes simultaneously we use a two-part regression model with a two-component mixture of probability distributions suitable for combining a binary outcome and a semi-continuous positive outcome. Because the data is also repeated for each individual in the survey over four waves, observations are considered to be clustered by individual and thus two separate random effects are included, one for each part of the modeling process.

1.2 Regression models for analysis of data with excess zeros

Two-part models are a combination of a discrete point-mass variable (where all of the mass occurs at zero) and a continuous random variable. A logistic regression model may be used to describe the probability of the zero part of the model and a conditional multiple linear regression model may be used to describe the mean of the non-zero continuous part of the variable (Duan et al. 1983, Lachenbruch, 2002). Duan et al. (1983) use a two-part model to analyse the demand for medical care where the first component is a probit model to describe the event of having zero medical expenses and the second component is a linear model on the log scale to describe the trend in positive expenses. They show that the likelihood for the two-part model can be separated into two functions so that maximising the overall likelihood is equivalent to combining the maximum likelihood estimates for the two parts.

In the econometrics literature, Tobit and sample selection models are often used to analyse cross-sectional data with excess zeros (Heckman 1976, Maddala 1983, Amemiya 1984). These models were developed to analyse censored or limited dependent variables where a latent variable is assumed to be censored by a random mechanism and where this variable and the probability of censoring is assumed to be jointly modelled as a function of the same covariates. Following Duan et al. (1983), Olsen and Schafer (2001, p.731) suggest that "two-part models are easier to interpret than selection models when zeros represent actual data, because the meaning of the underlying normal variable becomes dubious when zero is a valid response rather than a proxy for a negative or missing value". Because a zero is a valid response to hours worked, a two-part regression or mixed distribution model may provide a more valid approach to analysing the relationship of selected individual and household covariates with hours worked data.

The two-part regression models have also been extended to include random effects in both the logistic and linear stages of the model to capture unexplained heterogeneity among individuals in longitudinal and clustered data (Olsen & Schafer 2001; Tooze et al. 2002). Olsen and Schafer (2001) consider various strategies for estimating these models including Bayesian simulation using MCMC, Monte Carlo EM algorithm and a Laplace approximation method. Tooze et al. (2002) use adaptive Gaussian quadrature methods to maximise the likelihood of their models and implement this procedure using a SAS macro called MIXCORR that calls PROC GENMOD and PROC NLMIXED. The macro is available from the authors and we use this here to analyse the data from the HILDA survey.

In Section 2 we describe the sample of data extracted from the first four waves of the HILDA survey and outline the derivation of variables considered in the analysis. Section 3 presents the formulation of the two-part regression model with correlated random effects and outlines the interpretation of model parameters including regression coefficients and random effects. We also discuss a method for estimating the combined effect of a covariate on the number of hours worked by computing a ratio of mean hours worked using information from both the occurrence and intensity parts of the model. In Section 4 we present the results from fitting the model to the number of hours worked by women from the sub-sample of longitudinal HILDA data. Section 5 provides a summary of the results and a discussion of the methods used.

2 The Data

The data used in this study was based upon a dataset created for previous work done when there were three waves of data available[4]. The basis for selection for that dataset was that records were chosen that met the criteria of a person being female and who was a member of a couple with dependent children under 15, or a sole parent with dependent children under 15 for at least one wave of data. Using each woman's unique ID we then selected her records for the other waves.

Not every woman had a record at each wave and only those with a record for each of the three waves were selected providing us with a balanced dataset. For the fourth wave we extracted the next record for each of these women providing us with a balanced dataset of 2038 women and 8152 records over the four waves. A balanced file removes the need to consider missing data due to attrition in the analysis of longitudinal data. Although this may introduce a degree of bias into the analysis we begin with this data to illustrate the methodology used.

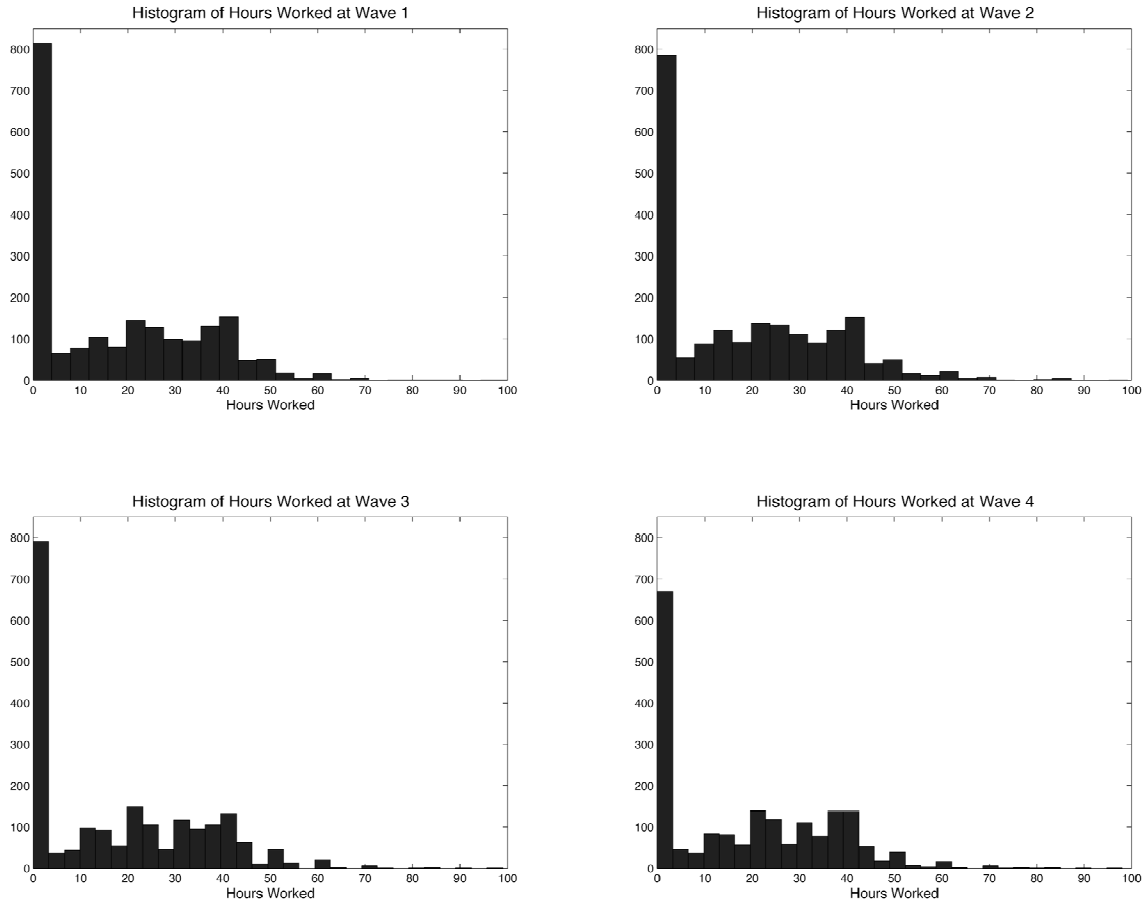
The key variables included in the regression model to assess their influence on the working hours of women were age, post-secondary school education, whether the woman was living in a couple or as a sole parent, the presence of a child under five years of age or over five but under 15 years of age in the household, presence of a serious health condition, the SEIFA[5] Index of Advantage/Disadvantage decile and dummy variables to represent the family structure group. Occupation associated with current or previous employment is also an important consideration in explaining the likelihood as well as intensity of work. A variable derived from the ABS ASCO codes will be included in future analyses.

Figure 1 shows the distribution of our dependent variable *Hours currently worked per week* for each of the four waves of data. The large number of zeros for individuals not in employment can be clearly seen but there is a gradual decline over the four waves from 39 to 34 per cent of the proportion of women in this category.

[4] Further details are provided in Gibbings et al (2005)

[5] Socio-Economic Indexes for Areas 2001, ABS Catalogue no. 2039.0 In the confidentialised version of HILDA, the SEIFA Index of Advantage/Disadvantage is reported as deciles

Fig 1 Histograms of number of hours worked per week at each of four waves



The family structure variable that we developed for inclusion in the model is divided into five groups as follows:

Initiating Group

No children under 15 present at wave 1, but in wave 2 or 3 at least one present. If present at wave 2 also present at wave 3.

Building Group

At least one child present at wave 1, with at least one additional child present at wave 2 and/or 3.

Stable Group

Same number of children under 15 present at each wave.

Teenage Group

Two or more children present at wave 1, but (at least) one less child under 15 at wave 2 and/or 3. If one less child under 15 at wave 2 then no more than this number or less at wave 3.

Grown-up Group

At least one child under 15 present at wave 1, in either wave 2 or 3 number reduces to zero. In most cases when the children turn 15 they stay in the family as dependent students.

The value for the variable is based on characteristics from the first three waves of data, but this value is held constant over all four waves.

3 Mixed distribution model with correlated random effects

In this section we provide an outline of the two-part mixed distribution model for analysing continuous positive repeated measures data with excess zeros as described in Tooze et al. (2002) and Olsen and Schafer (2001). For the first part we model the probability of being employed (or occurrence) using a logistic regression model with a random intercept. For women who work, we model the positive number of hours worked using a lognormal regression model also with a random intercept. A log transformation of hours worked for the positive part of the model is used as this reduces the level of skewness in the distribution of data that often occurs in positive variables of this type.

The inclusion of random intercepts in both parts of the model is one way to account for unobserved heterogeneity among individuals. A random intercept in the occurrence part of the model allows some individuals to have a consistently high or low propensity to work over all four waves of the survey, while a random intercept in the intensity part allows individuals to have a tendency to high or low mean hours worked. Specification of the correlation among these two random effects will capture the tendency of individuals with a high (low) propensity for employment to also work for longer (shorter) hours.

3.1 The model

Let the random variable Y_{ij} denote the number of hours worked in a week and take the observed value y_{ij} for individual *i* (*i = 1,…,2038*) at time *j* (*j = 1,2,3,4*). Let R_{ij} be a random variable denoting the occurrence of being employed where

$$R_{ij} = \begin{cases} 0, & \text{if } \mathrm{Y_{ij}} = 0 \\ 1, & \text{if } \mathrm{Y_{ij}} > 0 \end{cases}$$

with conditional probabilities

$$\Pr(R_{ij} = r_{ij} \mid \boldsymbol{\theta}_1) = \begin{cases} 1 - p_{ij}(\boldsymbol{\theta}_1), & \text{if } \mathrm{r_{ij}} = 0 \\ p_{ij}(\boldsymbol{\theta}_1), & \text{if } \mathrm{r_{ij}} = 1 \end{cases}$$

where $\boldsymbol{\theta}_1 = [\boldsymbol{\beta}_1', \mu_{1i}]$ is a vector of fixed-effects $\boldsymbol{\beta}_1$ and random intercept μ_{1i} in the occurrence part of the model. To model the probability of employment (occurrence or part one) we assume a logistic regression model such that

$$\text{logit}(p_{ij}(\boldsymbol{\theta}_1)) = \mathbf{X}_{\mathbf{1ij}}'\boldsymbol{\beta}_1 + u_{1i} \qquad (1)$$

where $\mathbf{X}_{\mathbf{1ij}}'$ is a vector of covariates.

Let S_{ij} denote the number of hours worked for those who are employed and who work a positive number of hours (intensity of work). The mean hours worked for women who are employed is $E(S_{ij} \mid \boldsymbol{\theta}_2) = \mu_{sij}(\boldsymbol{\theta}_2)$ where $\boldsymbol{\theta}_2 = [\boldsymbol{\beta}_2', \mu_{2i}]$ is a vector of fixed-effects $\boldsymbol{\beta}_2$ and random intercept μ_{2i} in the intensity part of the model. For the positive number of hours worked (intensity or part two) we assume a continuous lognormal model:

$$\log(S_{ij} \mid \boldsymbol{\theta}_2) \sim N(\mathbf{X}_{2ij}'\boldsymbol{\beta}_2 + \mu_{2i}, \sigma_e^2). \quad (2)$$

The random effects for the logistic and lognormal parts of the model are allowed to covary by assuming that they are generated by a bivariate normal distribution with variances σ_1^2 and σ_2^2, respectively, and with covariance $\rho\sigma_1\sigma_2$. This is the correlated logistic-lognormal regression model and in this study we estimate the model using the method of adaptive Gaussian quadrature implemented by the MIXCORR macro written for SAS by Tooze et al. (2002).

3.2 Interpreting the parameters

The effects of the covariates associated with the occurrence (whether employed or not) and intensity (number of hours worked if employed) parts of the model have the same interpretation as if the two parts of the model were fit to the data separately. The coefficients in the occurrence part are interpreted as they would be in a logistic regression analysis and the coefficients in the intensity part are interpreted as they would be in a continuous lognormal regression model. If a covariate is present in both the occurrence and intensity parts of the model then it is also plausible to compute the overall or combined effect of the covariate on the total number of hours worked. One way to do this is to consider the impact on the mean hours worked by a one unit change in the covariate of interest by computing the ratio of means for a one unit change in the value of the covariate, while holding the values of the remaining covariates in the model fixed (or constant). The ratio of means can then be computed for various combinations of values for the other covariates in the model such as combinations that yield a high or low likelihood of employment.

Following notation of Tooze et al. (2002) the ratio of means for a one unit change in a covariate Z that appears in both parts of the model is computed as follows:

$$\left[\frac{E(Y_{ij} \mid Z_{ij} = z+1)}{E(Y_{ij} \mid Z_{ij} = z)}\right] = \exp(\alpha_2)\exp(\alpha_1)\left[\frac{\Pr(R_{ij} = 0 \mid Z_{ij} = z+1)}{\Pr(R_{ij} = 0 \mid Z_{ij} = z)}\right] \quad (3)$$

where α_1 is the effect of the covariate in the occurrence (logistic) part of the model and α_2 is the effect of the covariate in the intensity (lognormal) part of the model. The probability of not working is computed as

$$\Pr(R_{ij} = 0 \mid Z_{ij} = z) = 1 - \Pr(R_{ij} = 1 \mid Z_{ij} = z) = \frac{1}{1 + \exp(\mathbf{X}_1'\boldsymbol{\beta}_1 + \alpha_1 z)} \quad (4)$$

where $\mathbf{X}_1$ is the vector of the covariates (excluding the covariate Z) in the occurrence part of the model and $\boldsymbol{\beta}_1$ is the vector of occurrence effects also excluding the effect of covariate Z.

4 Results

A logistic-lognormal regression model with correlated random effects was used to assess the association of selected covariates with working hours of women. In the first or binary part of the model the impact of age, post-secondary school education, whether the woman was living in a couple or as a sole parent, the presence of a child under five or over five and under 15 years of age in the household, presence of a serious health condition and SEIFA decile on the probability of working or not was assessed. In the second or intensity part of the model, age was replaced with wave (taking values 0, 1, 2, 3) and we introduced dummy variables to represent the family structure groups. To assess whether the intensity of work changed over time for each of the family types we included interaction terms for family structure by wave.

Both models with and without correlated random effects were fit to the data using the MIXCORR macro. We also implemented the model with uncorrelated random effects using Stata by separately fitting a logistic model with random effects to the presence of work hours and a linear regression model with random effects to positive work hours. The parameter estimates for the linear part of the model generated using SAS and xtreg in Stata were equivalent. However, as in previous work (Haynes & Western 2005) we found that to achieve similar estimates for the parameter estimates in the logistic part of the model we were required to increase the number of quadrature points beyond 24 when using the command xtlogit in Stata. Using the MIXCORR macro, the AIC from the model with correlated random effects (45458) was smaller than the AIC (45751) from the model with uncorrelated random effects indicating that that the model with correlated random effects is the better fitting model of the pair. Parameter estimates from both models are given in Table 1.

With reference to the results from the correlated model shown in the third column of Table 1, we firstly discuss the association of the covariates with the probability of a woman working or not. The estimated coefficients for age and age-squared indicate that the probability of working increases until the age of 46 and then begins to decline with age after that point. After controlling for age the probability of working is significantly greater for those women with post-secondary education. Sole parents are significantly less likely to work than women in a couple household and the probability of working for a woman with a child under five is significantly lower than for a woman who has no children. After accounting for these covariates, the probability of working is greater for women who do not have a serious health condition.

For women who do work a positive number of hours per week, the results in the lower part of Table 1 show that work intensity is highest for married women with post-secondary education, no serious health condition and who are in the initiating family group at wave one, at which time they have no children. Women in the building group work significantly less hours at wave one than women in the initiating group. This trend does not change over the four waves of the survey as women in the building group have an additional child during wave two or three. At wave one, women in the initiating group work significantly more hours than women in the stable and teenage groups as these women already have children at wave one. As time progresses the hours worked by these women in the stable and teenage group tends to increase, although they are still not working as many hours per week as women in the initiating group by wave four. Women in the grownup group also work significantly less hours than women in the initiating

group at wave one, but the difference in work hours is smaller than the difference for the other groups. The hours worked by women in the grownup group also tends to increase over time but at a slightly slower rate than for the stable and teenage groups.

From Table 1, the large significant random intercept variance for the occurrence part of the model shows that after accounting for the covariate differences among women, some women have a greater probability of being employed than others. From the lower part of Table 1, the significant random intercept variance for the intensity of work shows that for the women who do work and after accounting for the covariate differences among them, some women tend to work more hours per week than others. Furthermore, the correlation coefficient of 0.58 between the random intercepts of the occurrence and intensity parts of the model indicates that women with a greater propensity to work tended to also work longer than others, even after controlling for individual and family group characteristics. This result is not unexpected and in addition to the comparison of coefficients from both the uncorrelated and correlated models, illustrates the importance of specifying the correct covariance structure in the random effects model considered here.

If the same covariate is included in both the occurrence and intensity parts of the regression model it is possible to quantify the overall effect of the variable on the number of hours worked. Using equation (3) from Section 3.2 we can compute the ratio of the overall mean hours worked for a one unit change in a common covariate, conditional on fixed values of the other covariates in the model. Table 2 shows the ratio of mean hours worked for women with and without: post-secondary education; a child under five in the household; a serious health condition. In computing the ratio of mean hours worked from the regression coefficients, age was fixed to be 38 years (the average at wave three), the SEIFA decile was fixed at five (the median) and women were considered to live in a couple household. Two ratios were computed for each of the three variables. The first ratio assumes a scenario where the values of the remaining covariates are more likely to lead to a higher probability of work and a higher intensity of work (post-secondary education = 1, child under 5 = 0 and no health condition = 1). The second ratio in the pair assumes a scenario where the values of the remaining covariates are more likely to lead to a lower probability of work and a lower intensity of work (post-secondary education = 0, child under 5 = 1 and no health condition = 0).

From Table 2, the attribute of post-secondary education was associated with a greater mean number of hours worked. The mean hours worked was 1.4 times higher for women with post-secondary education if they had no children under five and no serious health condition. For women with at least one child under five and a serious health condition, the mean hours worked was 3.1 times higher if they had a post-secondary education. Having at least one child under five was associated with a lower mean number of hours worked. The mean hours worked for women with a child under five, a post-secondary education and no serious health condition was 0.6 times lower than the mean hours worked by women with the same attributes but without such a child. For women without post-secondary education and with a serious health condition the mean hours worked if they had a child under five was 0.2 times the mean hours worked for women with the same attributes but without such a child. These results indicate that the mean hours worked by women without post-secondary education and with a serious health condition is reduced more dramatically when there is a young child in the household, than for women who have post-secondary education and no health condition.

The presence of a serious health condition was associated with a lower number of hours worked. For women with post-secondary education and no children the mean hours worked in the absence of a health condition was no different from the mean hours worked in the presence of a health condition (ratio of means is 1.04). However, for women without post-secondary education and with children under five in the household, the mean hours worked was 1.7 times higher in the absence of a serious health condition. It appears that the presence of a serious health condition has more influence on lowering the mean number of hours worked for women with no post-secondary education and the presence of children under five in the household.

Table 1: Parameter estimates and model comparisons for logistic-lognormal mixed distribution model with uncorrelated and correlated random effects, respectively, fit to hours worked by women over four waves of the HILDA survey

Parameter	Uncorrelated RE Estimate (SE)	Correlated RE Estimate (SE)
Occurrence (Logistic)		
Intercept	-4.257** (1.492)	-4.286** (1.414)
Age	0.277*** (0.081)	0.277*** (0.078)
Age squared	-0.003** (0.001)	-0.003** (0.001)
Post-secondary education [yes=1]	2.027*** (0.200)	1.976*** (0.196)
Single [yes=1]	-0.366* (0.170)	-0.416* (0.167)
Children		
No children (baseline)	-	-
Child under 5	-2.707*** (0.220)	-2.769*** (0.221)
Children under 5 & over 5	-2.423*** (0.232)	-2.374*** (0.231)
All children over 5	-0.732*** (0.203)	-0.780*** (0.203)
No health condition [yes=1]	0.770*** (0.156)	0.761*** (0.153)
SEIFA decile	0.016 (0.029)	0.024 (0.029)
Variance random intercept (σ_1^2)	10.148*** (0.732)	10.173*** (0.721)
Intensity (Log-Normal)		
Intercept	3.427*** (0.069)	3.322*** (0.068)
Post-secondary education [yes=1]	0.180*** (0.031)	0.267*** (0.033)
Single [yes=1]	-0.039 (0.030)	-0.040 (0.029)
Children		
No children (baseline)	-	-
Child under 5	-0.368*** (0.047)	-0.409*** (0.046)
Children under 5 & over 5	-0.293*** (0.045)	-0.322*** (0.045)
All children over 5	-0.110** (0.038)	-0.122** (0.038)
No health condition [yes=1]	0.089*** (0.025)	0.092*** (0.025)
SEIFA decile	-0.004 (0.005)	-0.003 (0.005)
Wave	-0.078** (0.024)	-0.069** (0.024)
Family Type		
Initiating (baseline)	-	-
Building	-0.342*** (0.093)	-0.364*** (0.089)
Stable	-0.344*** (0.073)	-0.415*** (0.071)
Teenage	-0.376*** (0.083)	-0.411*** (0.081)
Grownup	-0.228** (0.089)	-0.281*** (0.085)
Building x Wave	0.047 (0.031)	0.036 (0.031)
Stable x Wave	0.109*** (0.026)	0.100*** (0.026)
Teenage x Wave	0.128*** (0.028)	0.119*** (0.028)
Grownup x Wave	0.096** (0.036)	0.085* (0.036)
Residual variance (σ_e^2)	0.142*** (0.004)	0.141*** (0.004)
Variance random intercept (σ_2^2)	0.300*** (0.013)	0.349*** (0.017)
Covariance ($\rho\sigma_1\sigma_2$)		1.090*** (0.085)
		(ρ = 0.58)
Model Statistic	**Value**	**Value**
AIC	45751	45458
-2log likelihood(ll)	45691	45396
	Difference in -2ll = 295	(p<0.0001)

*** p<0.001, **p<0.01, *p<0.05

Table 2: Effects on hours worked by women in HILDA data for covariates post-secondary education, presence of a child under 5 and presence of a health condition on probability of working or not (g), on intensity of work (f) and on mean amount (h)

Variable	(a) Post-sec education	(b) Child under 5	(c) No health condition	(d) Ratio of probabilities*	(e) $\exp(\alpha_1)$	(f) $\exp(\alpha_2)$	(g) $\exp(\alpha_1)$ x ratio (d)	(h) Ratio of means
Post-secondary edu	Y/N	0	1	0.147	7.214	1.306	1.060	1.385
(N=0/Y=1)	Y/N	1	0	0.332	7.214	1.306	2.395	3.128
Child under 5	1	Y/N	1	15.000	0.063	0.664	0.945	0.627
(No child = 0/Y=1)	0	Y/N	0	5.835	0.063	0.664	0.368	0.244
No health condition	1	0	N/Y	0.444	2.140	1.096	0.950	1.041
(N=0/Y=1)	0	1	N/Y	0.731	2.140	1.096	1.564	1.715

* Age fixed to mean at wave 3, 38 years
SEIFA score fixed to average of 5
Ratios considered for women in a couple relationship only

5 Discussion

Women's employment status can be classified as employed, unemployed and not in the labour force however there are very few unemployed women in our sample and these women have been combined with those not in the labour force. For women who are employed, participation in the workforce can also be measured by the intensity of work which will differ according to whether employment is on a part-time or full-time basis. In this study interest lies with assessing the level of impact that the number and age of children in the household have on a woman's workforce participation and how the intensity of work changes with family structure, while controlling for individual characteristics such as age and level of education. To undertake this assessment for Australian women we have used data from the first four waves of the HILDA survey collected from 2001-2004 and our measure of workforce participation is the variable 'number of hours usually worked in a week'.

As expected, the distribution of the number of hours worked for all women in the sub-sample is characterised by an excess of observations at zero which occur for women who are not employed at the time the data is collected. This variable therefore contains information on whether an individual works or not and given that she is employed, on the intensity of work. It is proposed that the covariates that influence the process of working or not may differ from those which influence the number of positive work hours and if a covariate does influence both processes then it will likely differ in its effect size. As the data reflects a combination of two processes it is inappropriate to use a linear regression model to assess the relationship between hours worked and other key variables. To model both processes simultaneously we have used a two-part regression model with mixed probability distributions suitable for combining a binary outcome and a semi-continuous positive outcome. Because the data is also repeated for each individual in the survey over four waves, observations are considered to be clustered by individual and thus two separate random effects are included, one for each part of the modeling process. A reasonable assumption is that some women have a greater propensity to be employed than others and that these women will also tend to work more hours per week, even after accounting for the covariate differences among them. To capture this process the model specification allows for correlated random effects.

The estimation of this complex correlated mixed distribution model is computationally intensive as the two parts of the model contain a common correlation parameter. Details of the model fitting techniques are described in Olsen and Schafer (2001) and Tooze et al. (2002). We chose to use the MIXCORR macro provided by Dr Janet Tooze that is relatively straight forward to implement in SAS computing software. The use of this methodology allows for a more sophisticated understanding of women's employment patterns as we can jointly consider the effects of individual and family characteristics on both the likelihood of working and the intensity of work. As initially demonstrated in Tooze et al. (2002) we have also computed the ratio of mean hours worked for a one unit change in a covariate where the calculation of the unconditional mean is based on both the probability of being unemployed and the intensity of work conditional on being employed.

Results from analysing the HILDA survey data using this methodology show that the probability of a woman working increases until about age 46 at which stage it begins to decline. This may be a cohort effect reflecting changing attitudes to women's

employment which will decrease in impact as younger women reach these ages. Of all the covariates included in part one of the model the presence of a child under five in the household has the largest influence on not working followed by the absence of post-secondary education. Post-secondary education has a significant positive effect on both the likelihood of being employed and the number of hours worked. It is also clear that the presence of a serious health condition has a negative effect on both aspects of working. Ratios of mean hours worked show that women who were more disadvantaged by the presence of both a medical condition and lower education, work much fewer hours when a child under five is in the household than a woman without those disadvantages. These results provide an indication of the protective effect of good education and health on employment.

Although the presence of children of any age is associated with a reduction in hours worked, the intensity of work is lowest when a child under five is present in the household. Inclusion of the family structure group over time in the intensity part of the model revealed that for women who were employed, the birth of an additional child resulted in lower work hours than the birth of a first child. For women who had a fixed number of children of any age the number of hours worked increased at a steady rate over the four years of the survey.

In this paper we have undertaken a longitudinal analysis of the occurrence and intensity of women's employment participation in Australia using the first four waves of the HILDA survey data. In addition to providing substantive results on the association of individual and family characteristics on employment our aim was to demonstrate an improved non-linear statistical method for analysing longitudinal data with an excess of zeros that is more appropriate than the use of linear regression techniques. The method which fits a logistic-lognormal mixed distribution with correlated random effects is easily implemented using the MIXCORR macro for SAS.

Acknowledgements

We thank Dr Janet Tooze from the Department of Public Health Sciences, Wake Forest University School of Medicine, for providing the MIXCORR macro to fit the two-part logistic-lognormal regression model with correlated random effects to our data using the SAS computing software. We also thank the Department of Families, Community Services and Indigenous Affairs, The University of Queensland Social Research Centre and the Melbourne Institute of Applied Economic and Social Research for supporting the research for the development of this paper and our attendance at the ACSPRI Social Science Methodology Conference in Sydney, December 2006.

References

Amemiya, T. (1984). Tobit Models: A Survey. *Journal of Econometrics*, 24, pp. 3-61.

Duan, N., Manning Jr, 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(2), pp. 115-126.

Gibbings J. and Heyworth, C. (2005). Family life events and mothers' employment transitions. Prepared for the HILDA Survey Research Conference, University of Melbourne, 29-30 September,2005. http://www.melbourneinstitute.com/hilda/Biblio/conf2005papers.html

Haynes, M., Western, M. and Spallek, M. (2005). Methods for categorical longitudinal survey data: Understanding employment transitions of Australian women. Prepared for the HILDA Survey Research Conference, University of Melbourne, 29-30 September,2005. http://www.melbourneinstitute.com/hilda/Biblio/conf2005papers.html

Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator of such models. *Annals of Economic and Social Measurement*. 5, pp. 475-92.

Lachenbruch, P. A. (2002). Analysis of data with excess zeros. *Statistical Methods in Medical Research*. 11, pp. 297-302.

Maddala, G.S. (1983). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press, Cambridge.

Olsen, M. C. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*. 96(454), pp. 730-745.

Tooze, J.A., Grunwald, G. K. and Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical Methods in Medical Research*. 11, pp. 341-55.