MODELLING ANNUAL EARNINGS WITH UNEMPLOYMENT:

NON-RANDOM SELECTION IN FEMALE WORKERS

By

Oliver Robertson

Many thanks to Dr. Dean Hyslop and Dr. Yiğit Sağlam for advice and feedback.

Contact e-mail: oliver.robertson@otago.ac.nz

Abstract

Female earnings are underrepresented in the earnings and earnings dynamics literature. This underrepresentation is largely a result of the differences in participation rates between male and female workers. Female workers' higher rates of unemployment increase the risk of sample selection bias. If selection into employment is non-random, then estimating earnings equations based on only those in the workforce will result in biased estimates. This paper focuses on modelling the annual earnings of female workers using data from the Survey of Families, Income, and Employment (SoFIE), while accounting for non-random selection into the workforce. We apply a correction for non-random selection in dynamic panel data proposed by Semykina and Wooldridge (2013). Semykina and Wooldridge's model corrects for non-random selection, while also dealing with a number of other issues that frequently arise in dynamic models of annual earnings. This paper finds that there is considerable evidence of non-random selection into the workforce. The differences between the estimated models that correct for non-random selection and those that do not are relatively large. Likewise, differences in participation patterns and regression results between the models that use the entire sample and those that only use individuals that worked in the first wave of SoFIE, indicates that there are systematic differences between workers and non-workers. This paper underscores the existence and importance of sample selection bias in the context of annual earnings, and presents an application of a new correction methodology in the New Zealand context.

1 Introduction

This paper focuses on modelling the annual earnings dynamics of female workers using data from the Survey of Families, Income, and Employment (SoFIE), while accounting for non-random selection into the workforce. Previous analyses of annual earnings dynamics have largely focused on male earnings, while for female workers the focus has been on extensive margin adjustments (Abowd and Card, 1989; Eckstein and Lifshitz, 2011). This has been driven by the lower rate of female workforce participation, leading to a larger number of zero observations in earnings data and increased risk of non-random selection resulting in estimation bias (Moffitt and Gottschalk, 2011). In his seminal work, Heckman (1979) outlined the causes and potential impacts of sample selection bias, illustrating how to deal with sample selection bias in a cross-sectional setting by treating it as a case of omitted variable bias. His method has been extended and adapted by many authors, with corrections that function with panel data, address sample attrition, are non-parametric, and deal with a range of related issues (Kyriazidou, 1997; Das et al., 2003; Kniesner and Ziliak, 1996).

In this paper we apply a newly developed correction for non-random selection in dynamic panel data proposed by Semykina and Wooldridge (2013). This model corrects for non-random selection, while also dealing with a number of other issues that frequently arise in dynamic models of annual earnings. Our earnings equation includes a lagged dependent variable, which can cause estimation issues if the auto-regressive coefficient is close to one and imposes stricter data requirements. Semykina and Wooldridge's correction addresses both of theses issues using backwards substitution, and the potential effect of unobserved individual specific fixed effects is removed by modelling the conditional expectation of the unobserved effect. The potential effects of selection are removed in a similar way to Heckman's correction; the selection equation is modelled separately using a probit model.

This paper proceeds as follows. First, Section 2 introduces sample selection bias, with a focus on non-random selection in panel earnings data, and outlines the earnings dynamics structure we will use. In this Section we will also review Semykina and Wooldridge's correc-

tion for non-random selection. Section 3 describes the SoFIE data set, and the extract that is used to estimate the earnings model. In Section 4 we present the results of applying both a selection of standard models that do not take into account non-random selection, as well as the results achieved with the application of Semykina and Wooldridge's correction. This allows for an analysis of how controlling for the potential non-random selection of female workers into the workforce has altered the estimated earnings dynamics. Section 5 concludes.

2 Female Earnings and Selection Bias

There is great interest in accurately modelling and understanding earnings and earnings dynamics (Hause, 1972; Esping-Andersen, 2007). How much an individual or family earns affects many aspects of their lives, such as consumption, schooling, and retirement decisions (Meghir and Pistaferri, 2011; Mitchell and Fields, 1981). Earnings are a stochastic process, determined to some extent by factors that are known to researchers, but also by seemingly random shocks (Altonji et al., 2013). The evolution of annual earnings over the life cycle is an inherently inter-temporal process. The choices an individual makes, changes in employment status, and earnings shocks have effects that continue to influence earnings over time (Hause, 1977; Jacobson et al., 1993). This means that attempts to model the earnings process must take these dynamic effects into account (Altonji et al., 2013).

Equation (2.1a) illustrates a simple dynamic model of annual earnings. Here Y_{it} is the annual earnings of individual i at time t, ρ is the auto-regressive coefficient reflecting the persistence of earnings and earnings shocks, X_{it} is a matrix of individual specific characteristics, β is the corresponding slope coefficients, C_i is unobserved, individual specific effect, and ϵ_{it} are idiosyncratic shocks.

$$Y_{it} = \rho Y_{it-1} + X_{it}\beta + C_i + u_{it} \ . \tag{2.1a}$$

Many workers will have periods of time where they do not work, and thus have zero earnings. Logarithmic transformations of wage, income, and earnings are frequently used in the labour economics literature but, as the logarithm is not defined for zero values, the researcher must decide how to treat these observations where earnings are zero.

The issues caused by periods of workforce non-participation are especially problematic when looking at female workers (Martins, 2001). Workforce participation for female workers is more likely to be punctuated by periods of non-participation, and a higher portion of the female population never works in comparison to the male population (Blau and Kahn, 2005; Killingsworth and Heckman, 1986; Hyslop, 2001). Much of the work on earnings dynamics has focused on male workers (Altonji et al., 2013; Abowd and Card, 1989). This is in large part due to the differences in workforce participation between male and female workers, where by focusing on male workers there is a larger sample to work with, and the risk of sample selection bias is minimised (Moffitt and Gottschalk, 2011).

Female levels of workforce participation have risen dramatically since World War II. Killingsworth and Heckman (1986) show that the participation rate for married women in the United States increased from 21.6% in 1950, to 40.8% by 1980. Likewise, Blau and Kahn (2005) show that participation rates for all women in the United States approximately doubled between 1947 and 1999. In the same time span, participation rates for male workers have decreased (Pencavel, 1986). Similar results have been observed in New Zealand, with increasing rates of female workforce participation and decreasing male rates (Jaumotte, 2004; Bryant et al., 2004). The increase in female participation rates to some extent reduces the issue of zero earnings observations, as more females will work in any given period and sample selection bias will be less of an issue. But higher female participation rates are to some extent off-set by the decrease in male participation rates, making dealing with the issue more relevant even when working with male earnings data.

2.1 Sample Selection Bias

Ignoring female earnings dynamics leaves out a large portion of the population that potentially has very different earnings dynamics. There is also no reason to believe that the results of studies on male earnings extend to females. Studies that do examine female annual earnings must decide how to deal with the zero earnings observations. In this context sample selection bias is a potentially serious issue. When modelling earning dynamics, how workers select into the workforce must be taken into account (Dustmann and Rochina-Barrachina, 2007). Sample selection bias occurs when the data available is of a non-random subset of the population, and from this researchers attempt to infer underlying relationships that extend beyond the available sample to the entire population (Stolzenberg and Relles, 1997). When modelling annual earnings, earnings are only observed for the subset of the sample that work in a given year. If the variables that affect the probability of working also influence the earnings an individual receives, then this non-random selection will lead to biased coefficient estimates (Heckman, 1979; Kassouf, 1994).

Heckman (1979) introduced what has become the standard approach to dealing with sample selection issues. Heckman proposed treating the problem as one of omitted variable bias, where the specification error leads to biased estimators. Heckman's solution to this involves first modelling the decision to work with a probit model, and then using this to generate the Inverse Mills ratio (IMR) which is included as an additional regressor in the main earnings equation Heckman (1979). Intuitively, correlation between the unobserved error term in the earnings equation, and both the observed and unobserved factors determining selection results in a sample that is systematically different from the population as a whole.

Heckman's original correction has been adapted in a number of ways that extend the model and loosen the assumptions required for estimation[1]. The distributional assumptions made in estimating the model have been relaxed, with a number of semi-parametric and non-

[1]See Vella (1998) for a survey of models that correct for sample selection bias. Dustmann and Rochina-Barrachina (2007) focuses on three methods for correcting for sample selection bias in panel data, while accounting for unobserved fixed effects.

parametric models (Ahn and Powell, 1993; Martins, 2001; Das et al., 2003). Longitudinal data introduces further issues that can complicate estimation, and a number of models developed in the literature extend or adapt Heckman's correction to function with panel data (Wooldridge, 1995; Kyriazidou, 1997; Vella and Verbeek, 1999). Next we introduce a correction for sample selection bias recently proposed by Semykina and Wooldridge (2013). This method uses elements of Heckman's approach, while including aspects from many of the extensions that allow for the estimation of a dynamic model. It models selection through the use of a probit model, and reduces the data requirements of differencing while still taking into account unobserved fixed effects.

2.2 Semykina and Wooldridge Correction

Semykina and Wooldridge (2013) propose a method of correcting for sample selection bias in dynamic panel data models. They assume the data has the underlying structure in Equation (2.2a), but that the dependent variable Y_{it} is only observed for a subset of the observations. Equation (2.2b) is the selection equation that determines if a particular observation has an observed dependent variable in time period t. In the context of annual earnings the dependent variable in the main equation would be annual log(earnings), and the selection equation determines if individual i participates in the workforce in year t.

$$Y_{it}^{*} = \rho Y_{i,t-1} + X_{it}\beta + C_{i1} + u_{it1} \ , \tag{2.2a}$$

$$S_{it} = 1[Z_{it}\beta_{2t} + C_{i2} + u_{it2} > 0] \ , \tag{2.2b}$$

$$Y_{it} = S_{it}Y_{it}^{*} \ . \tag{2.2c}$$

In this model, earnings are a function of lagged earnings (or lagged log(earnings)), some observed regressors in X_{it}, an unobserved individual specific fixed effect C_{i1}, and an idiosyncratic shock u_{it1}. The observed regressors X_{it} are assumed to be strictly exogenous conditional on the unobserved fixed effect, but may be correlated with C_{i1}. Denote

$X_i \equiv (X_{it}, X_{i2}, ..., X_{iT})$ and $Z_i \equiv (Z_{i1}, Z_{i2}, ..., Z_{iT})$. Here, Z_i is a set of regressors containing X_i and at least one additional time varying regressor that affects selection but is correctly excluded from the main earnings equation, C_{i2} is an individual specific unobserved fixed effect that impacts selection into/out of the workforce, and u_{it2} is an idiosyncratic shock. The variable Y_{it}^* is a latent variable, and its observeability depends on the outcome of the selection equation, where S_{it} acts as an indicator variable $S_{it} \in \{0, 1\}$. Y_{it} is the observed earnings of individual i in period t, and equals Y_{it}^* if $S_{it} = 1$ and the individual selects into the workforce, and zero otherwise.

If there were no selection, so Y_{it} was observed in every period for every individual, the unobserved individual specific fixed effects C_{i1} could still cause issues if they are correlated with X_i. This results in endogeneity which can bias coefficient estimates (Nickell, 1981). A common method for dealing with this is first differencing, which removes the potentially problematic fixed effects, and then estimating the new differenced equation using additional lags of Y_{it} as instruments for $\Delta Y_{i,t-1}$ to identify ρ.

There are some issues with first differencing as a correction method that Semykina and Wooldridge's model deals with. First, if ρ is close to one then the correlation between periods decreases, leading to weak instruments and poor identification (Blundell and Bond, 1998). Second, the presence of a lagged dependent variable increases the data requirements of first differencing, so that three consecutive periods must be observed. If there is selection, so only a portion of the sample have observed earnings in each period, this requirement may reduce the available data considerably. Lastly, if the behavioural decision to work/not work is correlated with earnings, then not taking this into account can lead to biased coefficient estimates, and first differencing does not remove the selection effect (Heckman, 1979).

The model proposed in Semykina and Wooldridge (2013) avoids these issues. Backwards substitution is used to replace the lagged dependent variable, removing the requirement to observe consecutive time periods. This results in (2.3), which includes the initial level of the dependent variable Y_{i0}, and a summation of past and present X_i. Unfortunately this

does introduce the requirement that Y_{i0} is observed for all individuals, but Semykina and Wooldridge suggest a way to avoid this, which will be covered in detail later in this section.

$$Y_{it} = \rho Y_{i,t-1} + X_{it}\beta + c_{i1} + u_{it1} \ , \tag{2.3a}$$

$$= \rho^t Y_{i0} + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + c_{i1}\sum_{j=0}^{t-1}\rho^j + e_{it1} \ , \tag{2.3b}$$

$$\text{where} \quad e_{it1} = \sum_{j=0}^{t-1} \rho^j u_{i,t-j,1} \ . \tag{2.3c}$$

While (2.3b) has removed the lagged dependent variable, the potential endogeneity of X_i with C_{i1} is still problematic. Semykina and Wooldridge propose following Chamberlain (1984) in modelling the conditional mean of the unobserved fixed effects in both the main and selection equations as a function of the observed exogenous regressors, including those in Z_i that are not in X_i, and Y_{i0}.

$$C_{i1} = \eta_1 + \sum_{s=1}^{T} Z_{is}B_{s1} + \gamma_1 Y_0 + U_{i1} \ , \tag{2.4a}$$

$$C_{i2} = \eta_2 + \sum_{s=1}^{T} Z_{is}B_{s2} + \gamma_2 Y_0 + U_{i2} \ . \tag{2.4b}$$

Explicitly modelling the unobserved fixed effect should remove the endogeneity, and thus that source of potential bias. This relies on the assumption that the unobserved fixed effect, or at least the portion of it that is correlated with X_i, is a linear function of the exogenous regressors from every time period. In Equations (2.4a) and (2.4b), η is an intercept, B_{s1} and B_{s2} are vectors containing the slope coefficient for the effect of each regressor in each time period on C_1 and C_2 respectively, γ is the effect of the initial earnings Y_0 on the fixed effects, and U_i is an individual specific unobserved effect. Modelling the fixed effects in this way results in the earnings and selection Equations (2.5a) and (2.5b). It should be noted

that if C_1 and C_2 take this form, the errors of each equation will be serially correlated.

$$Y_{it} = \rho^t Y_{i0} + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}B_{s1} + \gamma_1 Y_0\Big) + \epsilon_{it1} \ , \tag{2.5a}$$

$$S_{it} = 1[Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0 + \epsilon_{it2} > 0] \ , \tag{2.5b}$$

$$\epsilon_{it1} = \sum_{s=0}^{t-1} \rho^s (u_{i,t-s,1} + U_{i1}), \ \epsilon_{it2} = U_{i2} + u_{it2}, \ \mathbf{E}(\epsilon_{it1}|Y_{i0}, Z_i, S_{it} = 1) = \upsilon_{2t}\epsilon_{it2} \ . \tag{2.5c}$$

While using the model in Equation (2.5a) removes the endogeneity issue, and (potentially) reduces the data requirements, it doesn't model the behavioural decision to select into or out of the workforce. If the panel were balanced with no selection, then (2.5a) can be used in the place of a first-differencing approach. If non-random selection is present, then ignoring this could lead to biased coefficient estimates.

Semykina and Wooldridge propose to model the selection effect using a two-step estimator; first modelling selection, then estimating the IMR and including this in the main equation as an additional regressor. They focus on the fully parametric case, using a probit to model selection, but also state that it is possible to estimate the selection effect semi-parametrically, including h_t as a regressor in (2.5a).

$$h_{it} = \upsilon_{2t}\lambda_{it} \ , \tag{2.6a}$$

$$\lambda_{it} = \text{IMR}_{it}(Z_{it}, Y_0) = \frac{\phi[-(Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0)]}{\Phi[Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0]} \ . \tag{2.6b}$$

A large advantage of removing the lagged dependent variable through backwards substitution is that the correction does not have to be conditioned on observing an individual in three consecutive periods, it can be treated by modelling only contemporaneous selection. This does require the additional assumption that selection is a static, rather than a dynamic, process. As such, a probit model of selection is estimated for each time period separately, the estimated IMR, $\hat{\lambda}_{it}$, is calculated using (2.6b) and the slope estimates from the probit

model. Estimating selection separately for each year also allows the variance of the error term to vary. Alternatively, a single pooled selection equation could be estimated. The IMR value generated for each observation is then added to the main equation as an additional regressor. As a separate selection equation is estimated for each time period, Semykina and Wooldridge allow the slope of the IMR in the main equation to vary over different time periods. If selection were instead modelled with a single equation, the slope coefficient on the IMR could be constrained to be constant over the different time periods.

Equation (2.7a) is the full model that corrects for sample selection, deals with the potential for endogeneity through modelling the unobserved fixed effect, and doesn't require three periods to be consecutively observed in order to use an observation:

$$Y_{it} = \rho^t Y_{i0} + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}B_{s1} + \gamma_1 Y_0\Big) + \phi_t\lambda_{it} + \zeta_{it1}\ , \quad (2.7a)$$

$$\text{where} \quad \mathbf{E}(\zeta_{it1}|Z_i, Y_{io}, S_{it} = 1) = 0 \ \forall\ t \in 1, .., T. \quad (2.7b)$$

This formulation of the model requires that Y_0 is observed for all individuals in the sample, a strict requirement that undermines the model's ability to make inferences that extend to the greater population. The premise that there is non-random selection into the workforce means that selecting a sample based on only individuals that work in the first period will result in a non-representative sample. An alternative to this proposed by Semykina and Wooldridge is using Chamberlain's modelling device to model Y_0, similar to the way in which C_{i1} is modelled, as a function of all of the observed exogenous variables (Chamberlain, 1984). In this case Y_0 is modelled as in Equation (2.8a), where k_s is a vector

of slope coefficients, and this is added to the main equation resulting in (2.8b).

$$Y_{i0} = \sum_{s=1}^{T} Z_{is}k_s + b_i \ , \tag{2.8a}$$

$$Y_{it} = \rho^t \sum_{s=1}^{T} Z_{is}k_s + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}\delta_{s1}\Big) + \phi_t\lambda_{it} + q_{it1} \ , \tag{2.8b}$$

$$\text{where} \quad \mathbf{E}(b_i|Z_i) = 0, \ q_{it1} = \zeta_{it1} + \rho^t b_i \ . \tag{2.8c}$$

This model can be estimated, both using Y_0 or modelling it with Chamberlain's device, using Non-linear Least Squares (NLS) or Generalised Method of Moments (GMM) estimation. Here we focus on estimating the model using GMM, as this was shown by Semykina and Wooldridge (2013) to be more efficient. Let the vector of parameters be $\theta \equiv (\rho, \beta, k_1, ..., k_T, \eta_1, \delta_{11}, ..., \delta_{T1}, \phi_1, ..., \phi_T)$[2]. Equation (2.9) defines $m_{it}(\theta)$, the conditional expectation of Y_{it},

$$m_{it}(\theta) = m_{it}(z_i, Y_{i0}, S = 1; \theta) \ , \tag{2.9a}$$

$$= \rho^t \sum_{s=1}^{T} Z_{is}k_s + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}\delta_{s1}\Big) + \phi_t\lambda_{it} \ . \tag{2.9b}$$

To specify the GMM estimator we define a vector of instruments, $\omega_{it} \equiv \omega_{it}(\pi_{it}) \equiv (1, Z_{i1}, ..., Z_{iT}, \hat{\lambda}_{it2})$. This is a $1 \times (LT + 2)$ vector, where T is the number of time periods, and L is the number of regressors appearing in Z. If Y_0 is observed, then this is also included, making it a $1 \times (LT + 3)$ vector. Taking the vector of instruments for each time period allows us to construct the block diagonal instrument matrix (2.10), which will be $T \times T(LT + 2)$.

[2]This is when modelling Y_0. If Y_0 were observed $k_1, ..., K_T$ would not be present, and γ_1 would be.

$$W_i = \begin{pmatrix} \omega_{i1} & 0 & 0 & \cdots & 0 \\ 0 & \omega_{i2} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \omega_{iT} \end{pmatrix} . \tag{2.10}$$

Then define a $T \times 1$ vector $\hat{g}_i \equiv (\hat{g}_{i1}, \hat{g}_{i2}, ..., \hat{g}_{iT})$, where $\hat{g}_{i1} = S_{i1}(Y_{i1} - m_{i1})$, and the moment conditions $\mathbf{E}[W_i' g_i] = 0$ are available to use in estimation. They can be used in the estimator in (2.11), where $\hat{\Omega}^{-1}$ is a consistent estimator of a positive semi-definite weighting matrix Ω^{-1}.

$$\min_{\theta} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) \hat{\Omega}^{-1} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) . \tag{2.11}$$

The first order conditions for this GMM model are given in (2.12), where $\nabla_\theta g_i(\theta)$ indicates the first derivative of g_i with respect to the vector of parameters θ,

$$\Big(\sum_{i=1}^{N} W_i' \nabla_\theta g_i(\theta) \Big) \hat{\Omega}^{-1} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) = 0 . \tag{2.12}$$

The GMM estimator will be consistent when any positive semi-definite is used for Ω. However, Semykina and Wooldridge do specify a form that is preferred, and this is outlined in an online supplement to their paper. If the Ω specified by Semykina and Wooldridge is used, then Equation (2.13) outlines the asymptotic variance of the GMM estimator,

$$\text{Avar}(\hat{\theta}) = \frac{(\hat{G}' \Omega^{-1} \hat{G})}{N} , \tag{2.13a}$$

$$\text{where} \quad \hat{G} = \frac{1}{N} \sum_{i=1}^{N} W_i' \nabla_\theta g_i(\theta) . \tag{2.13b}$$

It should be noted that time constant regressors in both X_i and Z_i do not have separately identified slope coefficients when Z_i is used to model C_{i1}. If we examine Equation (2.9),

when X_i is time constant the only thing differentiating its direct effect from the indirect effect through modelling C_{i1} is the slope coefficients, which will not be separately identified. This is illustrated in Equation (2.14a), where X_i is a vector of time constant variables,

$$(\sum_{j=o}^{t-1} \rho^j X_i)\beta = X_i\beta \sum_{j=0}^{t-1} \rho^j = \frac{1-\rho^t}{1-\rho} X_i\beta \ . \tag{2.14a}$$

This means that the effect of time invariant variables is not separable from that of unobserved heterogeneity, so when estimating the model time invariant variables are removed from X_i, and only included in Z_i.

Generally when modelling selection using a probit model, Z_i is the matrix of regressors used in estimating the selection model and contains X_i, the regressors from the main equation, and at least one additional time varying regressor. This is to aid in identification, and helps avoid multicollinearity between the IMR and the other regressors in X_i when it is added to the main equation. In the case of Semykina and Wooldridge's correction, we include the full set of Zs in the main equation, in modelling C_1 and potentially Y_0. It is not clear if this will introduce identification issues.

3 Survey of Families, Income, and Employment

The Survey of Families, Income, and Employment (SoFIE) is a longitudinal survey carried out in New Zealand by Statistics New Zealand (SNZ). It contains a range of data on individuals and households, including demographic information, employment status, and annual earnings. The longitudinal nature of SoFIE allows for modelling of the dynamics and persistence of annual earnings for New Zealanders over time (Statistics New Zealand, 2001). The survey included eight waves of data collection, and ran between 2002 and 2010.

This paper uses an extract from SoFIE selected to focus on the prime aged female population. This extract includes women aged 24-54 in wave one that participated in every wave of the survey. This does mean that the portion of the sample that did not reply in every

wave of the survey is not represented in the extract. The extract only includes those aged 24-54 in an attempt to focus on the working age population. Many of those under 24 are still in education, and those starting older than 54 in wave one have a much higher chance of retiring during the course of the survey.

The survey includes both males and females, but our focus was solely on the female portion of the sample. As outlined in Section 2, female earnings dynamics have historically been neglected in the literature, with a greater focus on the extensive margin for female workers. This is due in part to the lower rate of female workforce participation, where men are more likely to work in a given year, and are much more likely to work in every year of a sample (Killingsworth and Heckman, 1986). This difference in participation rates appears to hold in SoFIE as well, where 78.5% of the males in our extract work in each of the eight periods, and only 3.4% never work, while in comparison, 59.5% of the females in our SoFIE extract work in every period, and 7.1% never work.

Table 1 lists some summary statistics for the SoFIE extract. A portion of the sample reported having negative earnings in some years. These individuals have been removed from the extract. While this introduces an additional risk of sample selection bias, the number removed was only 108 so the risk is relatively low.

The education variables in Table 1 indicate the fraction of the sample that has that level of education as their highest qualification, versus a baseline of having no education. 'School' indicates finishing high school, 'vocational' indicates vocational training, and 'university' indicates they have a bachelors degree or higher. While SoFIE includes a distinction between a bachelors and a higher degree, due to the small number of individuals that have a higher degree the two categories have been combined.

Table 1: *SoFIE summary statistics*

Sample size	4464
Fraction working	0.80
Mean(earnings*)	27.693 (49.604)
Mean(earnings*|working)	34.441 (53.179)
Age	42.17
Age | working	42.28
Fraction with children	0.54
Fraction with children | working	0.51
Fraction with partner	0.76
Fraction with partner | working	0.76
No qualification	0.14
No qualification | working	0.12
School	0.28
School | working	0.28
Vocational	0.35
Vocational | working	0.35
University	0.23
University | working	0.25

** in thousands*
Individuals that have negative earnings have been removed
Standard deviations in parenthesis

4 Results

In this section we estimate models for the earnings of female workers from SoFIE. First we estimate three simple models that ignore the potential for non-random selection into the workforce, and following this we estimate selection corrected models using Semykina and Wooldridge's correction. Equation (2.1a), the primary earnings equation from Section 2, is still the equation of interest, and Equation (4.1a) is the first differenced earnings equation that removes any unobserved individual specific effects. Here Y_{it} is the log(earnings) of individual i in time period t, the model includes a lagged dependent variable in Y_{it-1}, and ρ is the auto-regressive coefficient that determines the persistence of earnings over time.

$$\Delta Y_{it} = \rho \Delta Y_{it-1} + \Delta X_{it}\beta + \Delta u_{it1} \quad . \tag{4.1a}$$

First, Ordinary Least Squares (OLS) will be used to estimate the model in (2.2b). This ignores the presence of unobserved fixed effects, potentially leading to biased coefficient estimates if elements of X_{it} are correlated with C_{i1}. Second, the data will be first differenced as in (4.1a). This will remove the individual specific fixed effect, and the model is then estimated using First Differenced Ordinary Least Squares (FDOLS). While removing the bias due to the unobserved fixed effects, this method also introduces endogeneity through the correlation of ΔY_{it-1} with Δu_{it}. Third, the First Differenced with Instrumental Variables (FDIV) model will be used to estimate the earnings equation. This will correct for the endogeneity introduced through differencing by using Y_{it-2} as an instruments for ΔY_{it-1}.

For each of the three models, an unbalanced panel where periods of unemployment have been removed is used. For the ethnicity dummy variables, European is the base ethnicity; for education no qualification is treated as the base case. When first differencing is applied in the FDOLS and FDIV models, the time invariant regressors for education and ethnicity are removed as their fixed effects have been differenced out, and as Age increases by one in every period its effect is captured in the intercept.

Table 2 contains the results of applying these models to the SoFIE data. The auto-regressive coefficient $\hat{\rho}$ is statistically different from zero for all three models, but the estimate changes dramatically depending on the model used. Using OLS, annual earnings appear relatively persistent with $\hat{\rho} = 0.659$. When first differencing is used to remove the fixed effects the estimate changes to $\hat{\rho} = -0.293$. This indicates that earnings depends negatively on the previous period in the FDOLS model. When the correlation between ΔY_{it-1} and Δu_{it} is corrected for using the FDIV model the results change again. In this case the estimate of $\hat{\rho} = 0.257$ indicates a lower level of earnings persistence than when OLS was used, but it is positive unlike the FDOLS model.

The results from the OLS model have the effect of age on log(earnings) following the inverted U shape that is familiar from the annual earnings and wage literature (Cardoso et al., 2011), although the slope coefficient on age^2 is very small and not statistically different

Table 2: *SoFIE regression results*

	OLS	FDOLS	FDIV
Intercept	2.665** (0.1425)	0.035 (0.028)	-0.085** (0.026)
ρ	0.659** (0.0048)	-0.293** (0.013)	0.257** (0.021)
Age	0.0003** (0.00006)		
Age^2	-0.0000 (.0000)	-0.0000 (0.0000)	0.0000 (0.0000)
Partner	-0.029* (0.0116)	0.025 (0.031)	0.056 (0.037)
School	0.066** (0.0173)		
Vocational	0.092** (0.0166)		
University	0.252** (0.0181)		
Asian	-0.041 (0.0238)		
Maori	0.019 (0.0163)		
Pacific Islander	0.031 (0.0263)		
Other	0.027 (0.0406)		
Wave 3	-0.029 (0.0191)		
Wave 4	-0.015 (0.0190)	0.012 (0.021)	0.028 (0.027)
Wave 5	0.023 (0.0190)	0.066 (0.037)	0.089* (0.041)
Wave 6	0.018 (0.0191)	0.11* (0.05)	0.123* (0.057)
Wave 7	-0.006 (0.0192)	0.123 (0.065)	0.133 (0.073)
Wave 8	0.011 (0.0193)	0.131 (0.079)	0.161 (0.089)

** significant at 5% level. ** Significant at 1% level*

from zero. The coefficient on age is positive and statistically significant, while the slope of age^2 is negative. On the other hand, the size of the effect is very small for both age and age squared. For the other two models; age is captured in the intercept, and the effect of age^2 is again very small and not statistically different from zero.

The partner dummy variable is negative and statistically significant at the 5% level in the OLS model, but positive and not significant in each of the other two models. This could imply that there is some level of correlation between the unobserved individual specific effects and having a partner, and that removing the fixed effects through first differencing produces a more reliable estimate of the effect of having a partner.

The OLS model is the only one that includes coefficients for the time constant ethnicity and education variables. All of the ethnicity dummy variables are small, and none are significantly different from zero. The effect of being Asian is negative, while all the other ethnicities have a positive impact on annual earnings as compared to being European. All of the education dummy variables are positive and significantly different from zero, with the effect increasing as the level of education increases from no qualification (the base level), to high school level, vocational training, and then a university level of education.

While the FDIV model should have removed any unobserved fixed effects from the residuals, the potential for sample selection bias has so far been ignored. If selection into the workforce is non-random, then these results may well be biased. In the next section the sample selection bias correction introduced in Section 2.2 is applied to the SoFIE data, and the results are compared to those from this section.

4.1 Selection Corrected Results

In this section the sample selection bias correction proposed by Semykina and Wooldridge (2013) and introduced in Section 2.2 is applied to the sample of prime aged females from SoFIE. By correcting for non-random selection into the workforce in each wave, the model estimation should be more robust. Some adaptations have been made to Semykina and

Wooldridge's correction to make it more tractable, and results will be presented for four different versions of the correction. In each case, the results of the selection and main earnings equation will be presented.

The earnings and selection equations of interest are (2.2b) and (2.2c), introduced in Section 2.2. In the estimation that follows, the model is estimated using two sets of data. The first uses the subset of individuals that work in the first period, using their wave one earnings as the initial value and estimating the model using the remaining seven waves of data. The second set of data uses all individuals in the sample, and models Y_{i0} as a function of the observed variables.

In this application of Semykina and Wooldridge's correction, the selection model is simplified in two ways. First, selection is estimated using a single selection equation, instead of one for each wave of SoFIE. This means that the slopes of each coefficient are constrained to be constant over the different waves of the sample, and differences between waves are controlled for by including time dummies. Second, we assume that the IMR will have the same effect in each year, so that there is only a single slope coefficient. Selection is modelled conditional on being in the sample, so when working with the subset where all individuals work in the first period, that selection is ignored.

The sample selection bias correction proposed by Semykina and Wooldridge requires an additional, time varying, regressor that is correctly excluded from the primary earnings equation to be included in the selection equation. In the literature looking at the wages of female workers, the number of children a woman has is frequently used in this role (Baldwin and Johnson, 1992; Dankmeyer, 1996). In that context the intuition is that if a women does work, the fact that she has one or more children should not influence her wages, but having children will influence her decision to enter/exit the workforce. This does not transfer perfectly to the context of annual earnings. Annual earnings is essentially the number of hours worked multiplied by an individual's hourly wage. Even if it accepted that the number of children a women has does not affect the wage she receives if she works, it is

quite possible that it does influence the number of hours she works in a given year, and thus her annual earnings (Kaufman and Uhlenberg, 2000). Nevertheless, lacking a more appropriate instrument in the data set, the child dummy variables will be included as an exogenous variable to control for selection. This follows Semykina and Wooldridge's empirical application of the correction (although they are looking at average annual hourly earnings). The child dummy variables will also be included in the main earnings equation when they are used in modelling Y_{i0} and/or C_{i1}.

As in Section 2.2, let $X_i \equiv (X_{i1}, X_{i2}, ..., X_{iT})$ and $Z_i \equiv (Z_{i1}, Z_{i2}, ..., Z_{iT})$. When modelling Y_{i0} and/or the unobserved, individual specific fixed effect C_{i1}, Semykina and Wooldridge model each of them as a function of Z_i, which generally contains X_i, and any additional variables that are used in modelling selection. The effect of time invariant regressors are not separable from individual heterogeneity, so when modelling C_{i1} any time invariant regressors are removed from X_i, but remain in Z_i. Equations (4.2a) and (4.3a) are the equations for modelling Y_0 and C_{i1} respectively. This method assumes that these unobserved variables are functions of all variables from all time periods, and that a particular variable can have a differing effect depending on the time period.

$$Y_{i0} = \sum_{j=1}^{j=T} Z_{ij} k_j \ , \tag{4.2a}$$

$$Y_{i0} = \overline{Z_i^Y} k \ . \tag{4.2b}$$

$$C_{i1} = \eta + \sum_{j=1}^{j=T} Z_{ij} \delta_{j1} + \gamma Y_0 \ , \tag{4.3a}$$

$$C_{i1} = \eta + \overline{Z_i^C} \delta_1 + \gamma Y_0 \ . \tag{4.3b}$$

An intuitive way to understand this method is that, for an individual over the data sample, their observed characteristics reveal something about their unobserved characteristics, as

represented by C_{i1}. Even if this doesn't fully model C_{i1}, hopefully it captures the portion of it that is correlated with X_i, removing the issue of endogeneity.

In applying the Semykina and Wooldridge correction, our application simplifies the method of modelling both Y_0 and C_{i1}. In both cases, instead of assuming that Y_0 and C_{i1} are functions of each variable in each time period, we follow Mundlak (1978) and assume that they are functions of the mean value of each variable. This implicitly assumes that the slope coefficients are constant over the waves of the survey. This assumption does not change the affect that time invariant variables such as ethnicity and education have on Y_0 and C_{i1}, but will potentially alter the influence of time varying variables such as partner, or the number of children a woman has. We also assume that mean(age) and mean(age^2) are uncorrelated with the unobserved fixed effects. This follows Semykina and Wooldridge's empirical application, and essentially assumes that there are no systematic differences in unobserved ability across different age cohorts. These adjustments are shown in (4.2b) and (4.3b), where $\overline{Z_i}$ is a vector containing the means of the various variables, and the superscript indicates which variable is being modelled, so $\overline{Z_i^C}$ does not include the age related variables.

We apply four different versions of Semykina and Wooldridge's correction, starting with a basic model that corrects for only sample selection bias, and progressing in complexity to a model that corrects for sample selection bias while also modelling Y_{i0} and C_{i1}. A summary of the different models and the assumptions made in estimating them is displayed in Table 3. This progression will illustrate the effects of applying the sample selection bias corrections that Semykina and Wooldridge recommend, the impact of conditioning on working in the first period, and how modelling C_{i1} compares to differencing away the individual specific fixed effects.

The first model uses the subset of the sample that work in the first wave, treating Y_{i1} as the initial level of earnings, and does not model the individual specific fixed effects. The second model uses the full sample, modelling Y_{i0} as a function of the observed variables, but still ignores the potential impact of the unobserved fixed effects. The third model again uses

Table 3: *Summary of models*

Model 1:	Conditions on $Y_1 > 0$	Ignores C_{i1}	$t = 2, ..., 8$
	$Y_{it} = \rho^{t-1} Y_{i1} + \sum_{j=0}^{t-2} (\rho^j X_{t-j})\beta + \phi\lambda_{it} + \epsilon_{it}$		
Model 2:	Models Y_0	Ignores C_{i1}	$t = 1, ..., 8$
	$Y_{it} = \rho^t \overline{Z_i^Y} k + \sum_{j=0}^{t-1} \left(\rho^j X_{it-j}\right)\beta + \phi\lambda_{it} + \epsilon_{it}$		
Model 3:	Conditions on $Y_1 > 0$	Models C_{i1}	$t = 2, ..., 8$
	$Y_{it} = \rho^{t-1} Y_{i1} + \sum_{j=0}^{t-2} (\rho^j X_{t-j})\beta + \frac{(1-\rho^{t-1})}{(1-\rho)} \left(\overline{Z_i^C}\delta_1 + \gamma Y_0\right) + \phi\lambda_{it} + \epsilon_{it}$		
Model 4:	Models Y_0	Models C_{i1}	$t = 1, ..., 8$
	$Y_{it} = \rho^t \overline{Z_i^Y} k + \sum_{j=0}^{t-1} \left(\rho^j X_{it-j}\right)\beta + \frac{(1-\rho^t)}{(1-\rho)} \left(\overline{Z_i^C}\delta_1\right) + \phi\lambda_{it} + \epsilon_{it}$		

the subset that work in the first period, but introduces modelling the fixed effect C_{i1} as a function of the observed variables. The fourth model uses the full data set, modelling both Y_{i0} and C_{i1} as functions of all the observed variables.

As noted above, different samples are used in estimating the models. Selecting the sub-sample of individuals that work in the first period results in a population with very different participation patterns to those observed in the entire sample. Tables 5 and 4 show, for the sub and full samples respectively, the fraction of the sample that works in each wave, and the fraction that works in any pair of waves. There are a few significant differences, most noticeably the sub-sample that works in the first wave has a much higher rate of participation in future waves. For example, the fraction of the sample working in wave two is much higher in the sub-sample, with a 95% participation rate. This then decreases steadily over the survey to 0.89 by wave eight. In comparison, in the full sample participation in each wave is relatively constant over the different waves, staying close to 0.80. This indicates that working in wave one is a very good predictor of working in future waves, but that its predictive power decreases over time. In both samples, the fraction of the population working in any two waves decreases the further apart the waves are. This implies that there is inter-temporal correlation in workforce participation, and this correlation is higher in the

sub-sample that works in the first period.

Table 4: *Full sample: Fraction of sample used*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	0.77							
Wave 2	0.73	0.80						
Wave 3	0.72	0.76	0.81					
Wave 4	0.71	0.74	0.77	0.82				
Wave 5	0.70	0.73	0.75	0.78	0.82			
Wave 6	0.70	0.73	0.74	0.77	0.78	0.82		
Wave 7	0.69	0.72	0.73	0.75	0.76	0.78	0.82	
Wave 8	0.69	0.71	0.73	0.74	0.75	0.76	0.78	0.81
	$N = 4146$							

Table 5: *Wave one workers sample: Fraction of sample used*

	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 2	0.95						
Wave 3	0.91	0.93					
Wave 4	0.89	0.89	0.92				
Wave 5	0.88	0.88	0.89	0.91			
Wave 6	0.87	0.87	0.87	0.88	0.91		
Wave 7	0.87	0.86	0.86	0.86	0.87	0.90	
Wave 8	0.86	0.85	0.85	0.85	0.86	0.87	0.89
	$N = 3414$						

4.1.1 Selection models

Table 6 contains the results of the selection models. In each case, the selection equation is treated in the same way as the corresponding earnings equations. For example, in models one and three where only the individuals that work in the first period are included, Y_{i1} is included as a regressor in the selection equation. Likewise, when modelling the unobserved fixed effects but not Y_{i0}, mean(age) and mean(age^2) are excluded from the selection equation. As the variables that are used to model C_{i1} are a subset of those used to model Y_{i0}, the

variables used in the selection equation are identical between models two and four. This means that, while in the respective earnings equations we differentiate between modelling Y_{i0} and C_{i1}, in the selection model the coefficients represent the combined effect.

In each of the three models of selection in Table 6, having a partner increases the probability of working. Likewise, every level of education increases the probability of working when compared to the baseline of having no education. The impact of education was largest in models two and four, although it is impossible to know if that is due to the differences in model specification, or the difference in samples used, as selection into the sub-sample is not controlled for. Having a university level of education had the largest effect in all of the models, with vocational training having the second largest impact in models two and four, and the smallest in models one and three.

Each of the ethnicity dummy variables has a negative effect on the probability of being in the workforce as compared to the base line of being European. The effect of each ethnicity is quite similar between models one and three, and the negative effect is much stronger in the selection model for models two and four. Interestingly, the effect of age is larger in models two and four, which both include mean(age) and mean(age^2) in modelling the other components. Both age and age squared have very small impacts in models one and three, with neither being significantly different from zero. In models two and four on the other hand, both age and age^2 have a larger impact, with age increasing the probability of working, while age^2 has a negative effect. In this case in order to understand the net effect of age, its influence on the modelled Y_{i0} must also be taken into account. In this case mean(age) and mean(age^2) are only included in the second and fourth models, and both have a negative impact on the probability of working, although this is only significant in the case of mean(age^2).

For all of the models, the dummy variables indicating that a woman in the sample has one, or two or more children aged 0-4 years old, are significant and negative. The effect is largest for model one, but the other three models also include the mean of each of these dummies, so it is hard to compare the net effects. The dummy variables for having children

Table 6: *Selection model results*

	Model 1	Models 2 and 4	Model 3
Intercept	-0.450	-0.429	0.032
	(0.359)	(0.268)	(0.367)
Partner	0.090**	0.080	0.127
	(0.030)	(0.041)	(0.069)
High school	0.092*	0.388**	0.094*
	(0.042)	(0.024)	(0.042)
Vocational	0.052	0.444**	0.054
	(0.040)	(0.023)	(0.040)
University	0.225**	0.753**	0.243**
	(0.045)	(0.027)	(0.045)
Asian	-0.058	-0.506**	-0.066
	(0.061)	(0.033)	(0.061)
Maori	-0.005	-0.115**	-0.019
	(0.040)	(0.025)	(0.041)
Other	-0.251**	-0.347**	-0.261**
	(0.095)	(0.060)	(0.095)
Pacific	-0.163**	-0.190**	-0.164**
	(0.059)	(0.037)	(0.059)
Age	0.012	0.125**	-0.008
	(0.017)	(0.036)	(0.017)
Age^2	0.000	-0.001*	0.000
	(0.000)	(0.000)	(0.000)
Wave 3	-0.200**	-0.050	-0.185**
	(0.053)	(0.055)	(0.053)
Wave 4	-0.286**	-0.094	-0.259**
	(0.052)	(0.083)	(0.052)
Wave 5	-0.364**	-0.198	-0.328**
	(0.051)	(0.113)	(0.052)
Wave 6	-0.397**	-0.262	-0.355**
	(0.051)	(0.143)	(0.051)
Wave 7	-0.451**	-0.367*	-0.404**
	(0.051)	(0.174)	(0.051)
Wave 8	-0.504**	-0.473*	-0.449**
	(0.051)	(0.205)	(0.051)
Child 0-4: 1	-0.755**	-0.379**	-0.486**
	(0.040)	(0.035)	(0.057)
Child 0-4: 2+	-1.195**	-0.744**	-0.850**
	(0.052)	(0.050)	(0.077)
Child 5-17: 1	-0.060	-0.068*	-0.045
	(0.034)	(0.033)	(0.050)
Child 5-17: 2+	-0.036	-0.116**	-0.086
	(0.035)	(0.041)	(0.066)
Y_0	0.219**		0.219**
	(0.009)		(0.009)
Mean(Part)		0.300**	-0.025
		(0.046)	(0.078)
Mean(Age)		-0.041	
		(0.038)	
Mean(Age^2)		-0.001**	
		(0.000)	
Mean(Child 0-4: 1)		-0.672**	-0.517**
		(0.057)	(0.090)
Mean(Child 0-4: 2+)		-0.683**	-0.618**
		(0.083)	(0.126)
Mean(Child 5-17: 1)		-0.187**	-0.032
		(0.046)	(0.070)
Mean(Child 5-17: 2+)		-0.292**	0.157*
		(0.048)	(0.074)
Sample size	$N = 3414$	$N = 4146$	$N = 3414$

** significant at 5% level. ** Significant at 1% level*

aged 5-17 are also negative, but are only significant for models two and four, and the effect is much smaller than having children aged 0-4.

The mean values included in the selection equations are supposed to model C_{i2}, for models three and four, and Y_{i0} for models two and four. Since partner and the child variables are dummy variables, the mean will be in the range of zero to one, so the listed coefficient is the maximum effect that variable can have. For models two and four, mean(partner) is positive and significant, while all the child variables are negative and significant. For model four, partner is negative, but small and statistically insignificant, having any number of children aged $0-4$ has a negative and significant impact on the probability of working, while having more than two children aged $5-17$ actually slightly increases the probability of working.

4.1.2 Model Results

Table 7 present the results of applying the four models to the SoFIE data. As the number of regressors used is large only the primary results are presented here (β, ρ, ϕ), with the time dummies and parameters used to estimate Y_{i0} and C_{i1} available in the Appendix[3].

The most interesting result in Table 7 is that $\hat{\rho}$ is large and statistically significant in each of the four models. In the previous models in Section 4, OLS had the highest level of earnings persistence with $\hat{\rho}$ approximately equal to 0.65, while here model three has the lowest level of earnings persistence with $\hat{\rho} = 0.699$. In each case, controlling for non-random selection seems to have increased the persistence of annual earnings. This is similar to the results in Semykina and Wooldridge (2013), where applying the sample selection bias correction using GMM resulted in the largest value for $\hat{\rho}$. Comparing models one and three, it seems that including the individual effects explicitly by modelling C_{i1} has reduced the persistence of earnings. This is similar to the earlier results, where differencing out the individual specific fixed effects also resulted in lower levels of ρ, potentially indicating that the unobserved fixed effects were incorrectly being attributed to the auto-regressive coefficient.

[3]See Tables 8 and 9

Table 7: *SoFIE sample selection bias corrected models*

	Model 1	Model 2	Model 3	Model 4
Intercept	1.105			
	(0.734)			
ρ	0.794**	0.953**	0.699**	1.006**
	(0.029)	(0.070)	(0.067)	(0.129)
ϕ	-0.854*	-0.887	-0.720	-0.919
	(0.411)	(0.756)	(0.690)	(0.791)
Partner	-0.034**	-0.034*	0.005	-0.059
	(0.010)	(0.014)	(0.128)	(0.105)
Age	0.044*	0.074**	5.686	1.854
	(0.0203)	(0.018)	(0.049)	(0.049)
Age^2	-0.0005*	-0.0009**	-0.0007	-0.0002
	(0.0002)	(0.0002)	(0.0006)	(0.0006)
High school	0.066	0.007		
	(0.044)	(0.051)		
Vocational	0.075	0.013		
	(0.042)	(0.051)		
University	0.163**	0.039		
	(0.046)	(0.070)		
Asian	0.037	-0.004		
	(0.061)	(0.063)		
Maori	0.030	0.026		
	(0.028)	(0.032)		
Other	0.043	0.020		
	(0.168)	(0.179)		
Pacific	0.052	0.035		
	(0.072)	(0.079)		

** significant at 5% level. ** Significant at 1% level*
Time dummies, and variables used to model Y_{i0} and C_{i1} are in Tables 9 and 8 in Appendix **??**.

Including the full sample versus using only those individual that work in the first wave leads to much higher levels of earnings persistence, with both model two and four having higher levels of $\hat{\rho}$. It is interesting that including those with lower levels of participation (as shown previously the full sample has a lower participation rate) increases the level of earnings persistence. If periods of unemployment had a negative impact on future earnings, we might have expected that the full sample would in fact have lower levels of persistence (Arulampalam et al., 2001; Gregory and Jukes, 2001). In fact, for both models two and four, $\hat{\rho}$ is not significantly different from one, so earnings may be non-stationary.

Contrary to model three, modelling the unobserved fixed effects with the full sample in model four does not lead to a lower value for $\hat{\rho}$, in fact it increases slightly as compared to model two. While this does undermine the argument that controlling for the unobserved fixed effects lowers the modelled level of persistence in earnings, this difference could be due to the samples used in each case.

The slope coefficient on the IMR, ϕ, is relatively similar across the four models. In each case it is negative, ranging from -0.720 for model three, to -0.919 for model four. It has a larger effect on annual earnings when the full sample is used, potentially due to the lower rate of participation leading to a larger selection effect. In model one ϕ is significant at the 5% level, but it is not statistically different from zero in models two, three, or four.

An interesting result is that almost none of the slope coefficients used in modelling either Y_{i0} or C_{i1} are statistically significant. The intercept for Y_{i0} in model two is the only time any of the parameters used to model Y_{i0} are significantly different from zero. Likewise, the university dummy variable is the only significant parameter used in estimating C_{i0} and that is only in model three, when Y_{i0} is also modelled it becomes insignificant. In fact, as the models are progressed through in order, fewer and fewer variables are significant. The AR(1) parameter on lagged log(Earnings), ρ, is the only variable that remains significant throughout all of the models, and is the only statistically significant parameter in model four. It seems that controlling for earnings in the previous period has subsumed the effects

of the other parameters.

5 Conclusion

This chapter focused on modelling the earnings dynamics of females from the SoFIE data set. Female earnings have been neglected in the earnings literature, largely due to female workers' more frequent periods of unemployment and therefore greater risk of sample selection bias. This chapter contributes to the literature by modelling female annual earnings using a new model proposed by Semykina and Wooldridge (2013) that corrects for non-random selection, and also adds to the relatively limited existing literature on the earnings dynamics of female workers in New Zealand.

This paper has a number of interesting results. While the IMR was only significant in model one, the sample bias correct models produced results that were quite different from the simple models. Similar to the results in Semykina and Wooldridge (2013), we found that the models that corrected for sample selection bias had much higher levels of earnings persistence (although Semykina and Wooldridge are examining annual hourly earnings). We extended the model applied in Semykina and Wooldridge (2013) by modelling the earnings of the full SoFIE sample of females, where as Semykina and Wooldridge focus on the subsample that worked in the first year. We found that using the full sample had a large impact on the results, with the level of earnings persistence rising considerably, and not being significantly different from one. This indicates that there are systematic differences between the full and subsamples, and that using only the subset of the sample that work in the first period may lead to results that do not generalise to the greater population. The participation patterns produced in this paper also clearly showed that there are very different participation patterns between the full and subsample. Individuals that work in the first wave have much higher rates of participation in future waves, and there was evidence of intertemporal correlation in participation patterns.

The two models that used the full sample had values for ρ close to one. This indicates that the annual earnings of female workers in SoFIE may be non-stationary. Also, if this is true, makes the use of the Semykina and Wooldridge more important. The FDIV model suffers from the problem of weak instruments when ρ is close to one, while the correction model applied here does not rely on differencing and thus avoid the issue (Blundell and Bond, 1998).

While our results are similar to those in Semykina and Wooldridge (2013), it is harder to compare them to the wider earnings and earnings dynamics literature. Our model specification includes lagged log(earnings) in the earnings equation, which is less common than specifications which focus on the components of the residuals. We can however, still compare the primary results. When working with the full sample we found that earnings may be non-stationary. In the literature that examines male earnings there are a number of important papers that make similar findings, but generally they are examining the residual rather than a lagged component of earnings (MaCurdy, 1982; Topel and Ward, 1992; Browning et al., 2010). When working with the subset that worked in the first period, we found that earnings were persistent, but there was no evidence of non-stationarity. Again, there are a number of paper examining male earnings that make similar findings, though again they are usually focus on the residual of the earnings equation (Baker, 1997; Guvenen, 2009). Again, it should be noted that these papers are examining the residuals, and thus the AR(1) coefficient is measuring the persistence of a shock, while we are estimating the persistence of earnings. Regardless, it is interesting to compare the results when it comes to the question of if earnings are stationary.

This paper has shown that the correction proposed by Semykina and Wooldridge has some interesting properties. Using their correction, we have found levels of earnings persistence much higher than was evident in more established models that ignore sample selection bias. This application also extends the literature by examining the earnings dynamics of female workers in New Zealand, which is an area that has previously been under-explored.

The high levels of earnings persistence and intertemporal correlation of participation patterns indicate that individuals with otherwise similar characteristics can have quite different earnings profiles that persist over time. While we have not found definitive proof of sample selection bias, there is considerable evidence that there is non-random selection into the workforce, which makes controlling for this selection vital.

References

Abowd, J. M. and Card, D. (1989). On the covariance structure of earnings and hours changes. *Econometrica*, 57(2):411–445.

Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics*, 58(1-2):3–29.

Altonji, J. G., Smith, A. A., and Vidangos, I. (2013). Modeling earnings dynamics. *Econometrica*, 81(4):1395–1454.

Arulampalam, W., Gregg, P., and Gregory, M. (2001). Unemployment scarring. *The Economic Journal*, 111(475):577–584.

Baker, M. (1997). Growth-rate heterogeneity and the covariance structure of life-cycle earnings. *Journal of Labor Economics*, pages 338–375.

Baldwin, M. and Johnson, W. G. (1992). Estimating the employment effects of wage discrimination. *The Review of Economics and Statistics*, pages 446–455.

Blau, F. D. and Kahn, L. M. (2005). Changes in the labor supply behavior of married women: 1980-2000. Technical report, National Bureau of Economic Research.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1):115–143.

Browning, M., Ejrnaes, M., and Alvarez, J. (2010). Modelling income processes with lots of heterogeneity. *The Review of Economic Studies*, 77(4):1353–1381.

Bryant, J., Jacobsen, V., Bell, M., and Garrett, D. (2004). Labour force participation and gdp in new zealand. *Labour*, 4(07).

Cardoso, A. R., Guimarães, P., and Varejão, J. (2011). Are older workers worthy of their pay? An empirical investigation of age-productivity and age-wage nexuses. *De Economist*, 159(2):95–111.

Chamberlain, G. (1984). Panel data. *Handbook of econometrics*, 2:1247–1318.

Dankmeyer, B. (1996). Long run opportunity-costs of children according to education of the mother in the Netherlands. *Journal of Population Economics*, 9(3):349–361.

Das, M., Newey, W. K., and Vella, F. (2003). Nonparametric estimation of sample selection models. *The Review of Economic Studies*, 70(1):33–58.

Dustmann, C. and Rochina-Barrachina, M. E. (2007). Selection correction in panel data models: An application to the estimation of females' wage equations. *The Econometrics Journal*, 10(2):263–293.

Eckstein, Z. and Lifshitz, O. (2011). Dynamic female labor supply. *Econometrica*, 79(6):1675–1726.

Esping-Andersen, G. (2007). Sociological explanations of changing income distributions. *American Behavioral Scientist*, 50(5):639–658.

Gregory, M. and Jukes, R. (2001). Unemployment and subsequent earnings: Estimating scarring among British men 1984–94. *The Economic Journal*, 111(475):607–625.

Guvenen, F. (2009). An empirical investigation of labor income processes. *Review of Economic dynamics*, 12(1):58–79.

Hause, J. C. (1972). Earnings profile: Ability and schooling. *Journal of Political Economy*, 80(3):S108–S138.

Hause, J. C. (1977). The covariance structure of earnings and the on-the-job training hypothesis. In *Annals of Economic and Social Measurement, Volume 6, number 4*, pages 335–365. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, pages 153–161.

Hyslop, D. R. (2001). Rising US earnings inequality and family labor supply: The covariance structure of intrafamily earnings. *American Economic Review*, pages 755–777.

Jacobson, L. S., LaLonde, R. J., and Sullivan, D. G. (1993). Earnings losses of displaced workers. *The American economic review*, pages 685–709.

Jaumotte, F. (2004). Labour force participation of women. *OECD Economic studies*, 2003(2):51–108.

Kassouf, A. L. (1994). The wage rate estimation using the Heckman procedure. *Brazilian Review of Econometrics*, 14(1):89–107.

Kaufman, G. and Uhlenberg, P. (2000). The influence of parenthood on the work effort of married men and women. *Social Forces*, 78(3):931–947.

Killingsworth, M. R. and Heckman, J. J. (1986). Female labor supply: A survey. *Handbook of Labor Economics*, 1(1):103–204.

Kniesner, T. J. and Ziliak, J. P. (1996). The importance of sample attrition in life cycle labor supply estimation. *Available at SSRN 1795*.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica: Journal of the Econometric Society*, pages 1335–1364.

MaCurdy, T. E. (1982). The use of time series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of econometrics*, 18(1):83–114.

Martins, M. F. O. (2001). Parametric and semiparametric estimation of sample selection models: an empirical application to the female labour force in Portugal. *Journal of Applied Econometrics*, 16(1):23–39.

Meghir, C. and Pistaferri, L. (2011). Earnings, consumption and life cycle choices. *Handbook of Labor Economics*, 4:773–854.

Mitchell, O. S. and Fields, G. S. (1981). The effects of pensions and earnings on retirement: A review essay. Working Paper 772, National Bureau of Economic Research.

Moffitt, R. A. and Gottschalk, P. (2011). Trends in the covariance structure of earnings in the US: 1969–1987. *The Journal of Economic Inequality*, 9(3):439–459.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the Econometric Society*, pages 1417–1426.

Pencavel, J. (1986). Labor supply of men: a survey. *Handbook of labor economics*, 1:3–102.

Semykina, A. and Wooldridge, J. M. (2013). Estimation of dynamic panel data models with sample selection. *Journal of Applied Econometrics*, 28(1):47–61.

Statistics New Zealand (2001). A longitudinal survey of income, employment and family dynamics. feasibility project final report. Technical report, Statistics New Zealand.

Stolzenberg, R. M. and Relles, D. A. (1997). Tools for intuition about sample selection bias and its correction. *American Sociological Review*, pages 494–507.

Topel, R. H. and Ward, M. P. (1992). Job mobility and the careers of young men. *The Quarterly Journal of Economics*, 107(2):439–479.

Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources*, pages 127–169.

Vella, F. and Verbeek, M. (1999). Two-step estimation of panel data models with censored endogenous variables and selection bias. *Journal of Econometrics*, 90(2):239–263.

Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics*, 68(1):115–132.

Appendix

Table 8: *Semykina and Wooldridge models: coefficients for modelling* Y_{i0}

	Model 1	Model 2	Model 3	Model 4
Y_0: High school		0.080 (0.309)		0.037 (0.297)
Y_0: Vocational		0.102 (0.309)		0.079 (0.299)
Y_0: University		0.415 (0.378)		0.361 (0.367)
Y_0: Asian		-0.022 (0.407)		0.016 (0.383)
Y_0: Maori		-0.121 (0.191)		-0.070 (0.184)
Y_0: Other		0.012 (1.027)		0.070 (0.956)
Y_0: Pacific		0.090 (0.441)		0.181 (0.415)
Y_0: Mean(partner)		0.202 (0.184)		0.154 (0.190)
Y_0: Mean(age)		-0.287 (0.327)		.035 (0.356)
Y_0: Mean(age^2)		0.0032 (0.0039)		−0.0004 (0.0041)
Y_0: Mean(child 0-4: 1)		-0.478 (0.531)		-0.378 (0.615)
Y_0: Mean(child 0-4: 2+)		-0.534 (0.802)		0.065 (1.108)
Y_0: Mean(child 5-17: 1)		-0.245 (0.278)		-0.467 (0.327)
Y_0: Mean(child 5-17: 2+)		-0.485 (0.384)		-0.688 (0.385)
Y_0: Intercept		15.252* (6.502)		9.096 (6.864)

Table 9: *Semykina and Wooldridge models: time dummies and variables used to model* C_{i1}

	Model 1	Model 2	Model 3	Model 4
Wave 2		-1.067		-0.455
		(0.912)		(1.414)
Wave 3	-0.051**	-1.075	-0.035	-0.460
	(0.013)	(0.908)	(0.018)	(1.411)
Wave 4	-0.036*	-1.056	-0.012	-0.441
	(0.015)	(0.906)	(0.022)	(1.411)
Wave 5	-0.011	-0.994	0.021	-0.378
	(0.012)	(0.901)	(0.020)	(1.408)
Wave 6	-0.017	-1.024	0.029	-0.414
	(0.013)	(0.907)	(0.025)	(1.418)
Wave 7	-0.033**	-1.031	0.024	-0.424
	(0.010)	(0.905)	(0.025)	(1.419)
Wave 8	-0.019*	-1.016	0.047	-0.413
	(0.009)	(0.904)	(0.028)	(1.421)
C_{i1}: Intercept			1.598	
			(1.259)	
C_{i1}: Mean(partner)			-0.030	0.029
			(0.134)	(0.103)
C_{i1}: Mean(child 0-4: 1)			-0.114	0.007
			(0.087)	(0.091)
C_{i1}: Mean(child 0-4: 2+)			-0.191	-0.079
			(0.164)	(0.145)
C_{i1}: Mean(child 5-17: 1)			0.002	0.055
			(0.042)	(0.050)
C_{i1}: Mean(child 5-17 2+)			-0.059	0.062
			(0.043)	(0.068)
C_{i1}: High school			0.049	0.010
			(0.054)	(0.052)
C_{i1}: Vocational			0.063	0.010
			(0.052)	(0.052)
C_{i1}: University			0.179**	0.023
			(0.058)	(0.078)
C_{i1}: Asian			0.016	-0.010
			(0.075)	(0.066)
C_{i1}: Maori			0.012	0.019
			(0.035)	(0.034)
C_{i1}: Other			0.012	0.008
			(0.202)	(0.184)
C_{i1}: Pacific			0.058	0.009
			(0.087)	(0.086)
C_{i1}: gamma			0.032	
			(0.034)	