

Simulating Cohort Earnings for Australia

J. van de Ven*

March 3, 2005

Abstract

A dynamic microsimulation model of cohort earnings based on the Australian population aged between 20 and 55 years is described. A highly parsimonious modular structure is adopted to facilitate sensitivity analysis and enable additional characteristics to be added, should they be desired. Despite the restrictive specifications used, the model closely reflects the data used for calibration.

JEL Classification: N37, D31

Key Words: Microsimulation, Income, Lifetime, Inequality, Modelling.

1 Introduction

This paper describes a dynamic microsimulation model of cohort earnings developed to consider redistribution during the working-lifetime in Australia. Microsimulation models were first used for economic analysis by Orcutt (1957), and are now commonly employed to undertake policy analyses in many countries around the world. The feature that distinguishes microsimulation models from their macro based counterparts is that each micro-unit (also referred to as agent) from a given population is individually represented.¹ This feature is particularly useful for undertaking distributional analyses; see Creedy & van de Ven (1999), Creedy & van de Ven (2001) and van de Ven (2005) for distributional analyses based upon the simulation model described here.

Microsimulation models are classified as either dynamic or static, depending upon how (and whether) the population is aged. Unlike static microsimulation models, dynamic microsimulation is designed specifically to consider the effects of counterfactual conditions on a population of agents through time. The ability to consider counterfactual experiments means that dynamic models are capable of providing insights that survey data cannot. The limitations of survey data are compounded in Australia, where the few panel data sets are small, both in terms of duration and breadth, compared to those of many other countries. This makes the current model particularly useful for distributional analyses of earnings in Australia where income measured over a period in excess of a single year is desired.

*National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, UK. jvandeven@niesr.ac.uk. I should like to thank John Creedy and John Muellbauer for their extensive support and advice throughout the construction of the model. My thanks are also extended to the Henderson Foundation for its financial support, and to the HRD for making available the data sources used. Any omissions or errors are my own.

¹For macro-based models that study the impact of policy changes, see Dervis et al. (1982), Taylor (1990), and de Janvry et al. (1991). These are examples of Computable General Equilibrium models. Most micro-based models are constructed using a partial equilibrium framework. For examples of micro-based models that use a general equilibrium framework, see Meagher (1993), and Cogneau & Robilliard (2000).

Most microsimulation models that are currently in use are static. Prominent examples of these include, STINMOD (Australia; refer to NATSEM, Australia), POLIMOD (UK; see Redmond et al. (1998)), TRIM2 (US; see Giannarelli (1992)), SPSP (Canada; refer to Statistics Canada), GMOD (Germany), SWITCH (Ireland), LOTTE (Norway), FASIT (Sweden), and CSO (Hungary).² However, advances in computing power, and the availability of increasingly detailed survey data have led to an increase in both the number, and sophistication, of dynamic microsimulation models. Some of the dynamic models in use include ASPEN (US; see Basu et al. (1998)), CORSIM (US; see Caldwell (1997)), HARDING (Australia; see Harding (1993)), and SESIM (Sweden), while many more are currently being developed.

The model described in this paper is comprised of two components that generate labour force status and wage rates for a cohort of individuals aged 20 in 1970. Individual characteristics are generated at annual intervals for every cohort age between 20 and 55, thereby capturing the working-lifetime.³ Unlike the architecture of most microsimulation models, the current model has been developed to facilitate transparent sensitivity analysis. This objective has led to the adoption of a highly parsimonious structure.

The labour force component generates change using transition probability functions, which replace the transition matrices that are typical of microsimulation models.⁴ Transition probability functions enable sensitivity analysis to be undertaken by varying a few well defined parameters, as opposed to the relatively opaque element-by-element adjustment required for transition matrices. Wages from labour are also generated by a compact procedure compared to other microsimulation models using two meaningful functions that are shown to relate closely to standard earnings equations.

Most microsimulation models generate a large number of characteristics for each individual to make a broad range of analyses possible. Given the relatively few characteristics generated by the current model, the ability to include additional characteristics as required is a fundamental feature of the modular structure adopted. Following the addition of cohort demographics, for example, the model is capable of analysing the redistributive effect of income taxes and a range of transfer schemes that comprise approximately 70 per cent of Australian social security expenditure, excluding pensions for the retired.⁵

Dynamic microsimulation models can be distinguished by the extent to which they incorporate agent specific behavioural responses. Given the aging populations observed in many countries, attention has

²For useful surveys, refer to Sutherland (1995), and Merz (1991).

³Following the age of 55, retirement has a dominant effect upon annual measures of income inequality. See, for example, Figure 2 of Creedy & van de Ven (1999), which uses an earlier version of the present model.

⁴See, for example, Harding (1993).

⁵The model described here forms part of a larger microsimulation model developed to examine the redistributive effects of taxes in Australia. See van de Ven (2004) or van de Ven (1998) for a detailed description of the demographic components of the larger simulation model.

been focused in recent years on the responsiveness of labour supply, savings, and fertility to alternative tax and transfer systems.⁶ However, unlike the models used to examine these issues, the simulation procedure described in this paper makes no adjustment for behavioural responses. This property may be thought to call into question the extent to which analyses based upon the current model are of practical use. Specifically, given that many tax and transfer schemes are designed to affect agent behaviour, the predicted impact of such schemes derived from simulations that omit behavioural responses must be fundamentally inaccurate.

In response to this criticism, it may be noted that not all fiscal reforms cause the behaviour of agents to change. Furthermore, given that no model, however complex, can possibly capture the full extent of real-world diversity, any prediction derived from simulation methods must be treated with a degree of caution. With regard to microsimulation models that project labour supply responses, for example, it is possible to find reports of wage elasticities that range from small negative values to measures just over one. Although it is reasonable to suspect that a small positive wage elasticity is likely to reflect most applied cases, the accompanying uncertainty regarding the ‘true’ value cannot be ignored. In this sense, the first order effects generated by microsimulation models that omit behavioural responses provide a means of making unambiguous statements of the kind, “if behaviour remained unchanged...” Accompanying sensitivity analysis associated with any expected changes in behaviour can, of course, be subsequently undertaken. The extent to which the first order effects of policy change are of practical interest is indicated by the continued use of, and focus on, static microsimulation models, which include no behavioural responses.

Section 2 provides an overview of the simulation procedure. A detailed description of how labour and income characteristics are simulated is provided in sections 3 and 4. Concluding comments are made in section 5.

2 The Simulation Procedure

Heterogeneity between individuals is restricted to the following four characteristics:

1. employment status (identified as full-time employed / part-time employed / unemployed)
2. employment status of spouse (for spouses 17 years of age and over)
3. labour income
4. labour income of spouse

⁶see Macunovich (1998), and Hotz et al. (1997) for surveys of the fertility literature, and Auerbach (1997) on savings.

In any given year each of the characteristics are determined for each individual using the linear procedure depicted by the flow chart of Figure 1.

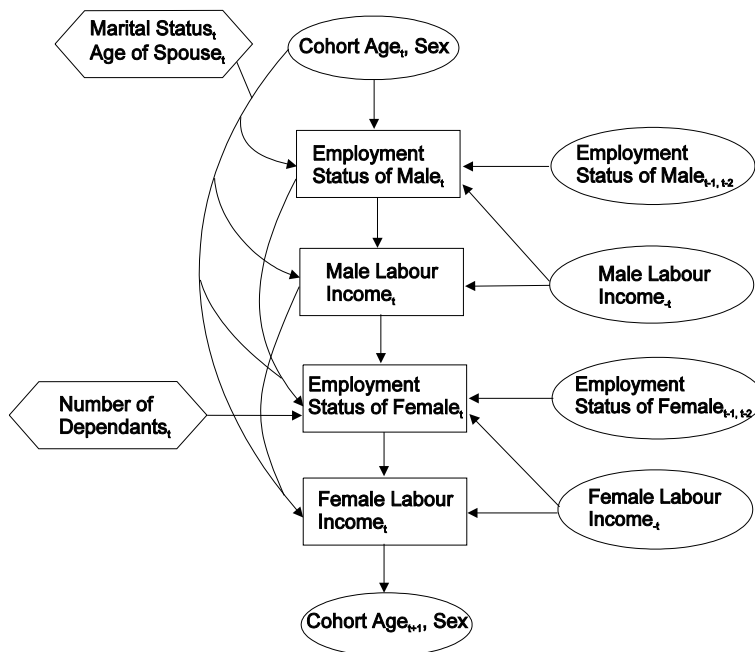


Figure 1: Stylised Earnings Simulation Procedure

In Figure 1, t subscripts refer to the reference period, and $-t$ subscripts refer to the entire simulated history up to period t . Characteristics included in elliptical frames are endogenous inputs used to generate period t characteristics, hexagon frames denote characteristics that must be exogenously specified, and arrows indicate links. Hence, male labour income in period t is generated with reference to the male's employment status in period t , the measures of income generated for the male for all periods previous to t , and the male's age and sex (specified as either a cohort member, or the spouse of a cohort member).

The following two sections describe the simulation procedures used to generate employment status and earnings respectively. They also detail the calculations undertaken to calibrate the model so that it reflects the Australian population.

3 Employment Status

The model uses a two stage Monte Carlo procedure to generate the employment status of any individual, where the employed population is determined first followed by a discrimination between the full-time and part-time employed. The probabilities required for the Monte Carlo procedure were derived from the Confidentialised Unit Record File (CURF) of the Survey of Employment and Unemployment Patterns (SEUP). The SEUP provides panel data on a range of demographic and labour characteristics between

September 1994 and September 1997 for 2311 individuals selected at random from the Australian population aged 15-59 years. After removing individuals with missing observations, 840 males and 993 females aged 20 years or older were used to estimate the four probit equations employed by the simulation model.⁷

The probit equations estimated for males are characterised by equations (1) and (2), which are used respectively to identify the employed and to discriminate between the full-time and part-time employed.

$$I_{i,t}^{emp} = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3 M_{i,t} + \beta_4 emp_{i,t-1} + \beta_5 emp_{i,t-2} + \beta_6 hy_{i,t-1} + \varepsilon_{i,t}^{emp} \quad (1)$$

$$I_{i,t}^{ft} = \beta_{10} + \beta_{11} t_i + \beta_{12} t_i^2 + \beta_{13} ft_{i,t-1} + \beta_{14} pt_{i,t-1} + \beta_{15} ft_{i,t-2} + \beta_{16} pt_{i,t-2} + \varepsilon_{i,t}^{ft} \quad (2)$$

where:

- $I_{i,t}^{emp}$ is the probit index associated with being employed for individual i at age t .
- $I_{i,t}^{ft}$ is the probit index of individual i at age t associated with being full-time employed, given that they are employed.
- $M_{i,t}$ is a marital status dummy which equals 1 if individual i is married at age t and zero otherwise.
- $emp_{i,t}$ is an employment status dummy which equals 1 if individual i is employed at age t and zero otherwise.
- $hy_{i,t-1}$ is a high income dummy which equals 1 if individual i has an income greater than 60 % of the total population (or \$27,040 per annum for 1997) and zero otherwise.
- $ft_{i,t}$ is a dummy which equals 1 if individual i is full-time employed at age t and zero otherwise.
- $pt_{i,t}$ is a dummy which equals 1 if individual i is part-time employed at age t and zero otherwise.

The regression coefficients and statistics are provided in Table 1.⁸

The estimated regressions for male employment / unemployment and full-time / part-time discrimination provide correct predictions respectively for 92.4 and 93.7 per cent of observations and all coefficients have the expected signs. It is evident from the standard errors provided in Table 1 that a relatively high degree of uncertainty is associated with the values of many of the estimated coefficients. This is, in part, attributable to the dominating effect of previous employment experience on the relationships estimated. To correct for this effect the constant and age coefficients, β_0 , β_1 , β_2 , β_{10} , β_{11} , and β_{12} , were adjusted to ensure that the employment characteristics simulated by the model reflect as

⁷Data for 328 individuals of the original 2311 surveyed were incomplete. A further 150 individuals were under 20 years of age at the time when the survey was first taken. When determining the labour status of spouses aged between 17 and 19 years, it is assumed that the regression estimates of the probit equations obtained remain applicable.

⁸Standard errors provided in brackets.

Table 1: Regression statistics for probit models of male employment

	β_0	β_1	β_2	β_3
Regression Coefficients	-1.6474	0.0542	-0.853E-03	0.2043
	(0.9861)	(0.0508)	(0.609E - 03)	(0.1478)
Adopted Coefficients	-1.4700	0.0542	-0.953E-03	0.2043
	β_4	β_5	β_6	
Regression Coefficients	1.4463	0.8930	0.3551	
	(0.1778)	(0.1879)	(0.1603)	
Adopted Coefficients	1.4463	0.8930	0.3551	
	β_{10}	β_{11}	β_{12}	β_{13}
Regression Coefficients	-0.8659	0.0826	-0.120E-02	0.9082
	(1.165)	(0.0584)	(0.7E - 03)	(0.2813)
Adopted Coefficients	-0.7029	0.0897	-0.135E-02	0.9114
	β_{14}	β_{15}	β_{16}	
Regression Coefficients	-0.5862	0.5244	-0.8397	
	(0.3219)	(0.2679)	(0.3375)	
Adopted Coefficients	-0.6112	0.4943	-0.8841	

closely as possible the data upon which they are based. The ‘adopted coefficient’ series listed in Table 1 are the result of the adjustments undertaken, where none of the adjusted coefficients vary significantly from their original regression estimates at the 95 per cent confidence interval. Figures 2 and 3 depict the relationship between the raw and simulated data.

Similarly, the two probit equations used to simulate female employment status are characterised by equations (3) and (4), and the associated regression results are presented in Table 2.

$$I_{i,t}^{emp} = \gamma_0 + \gamma_1 t_i + \gamma_2 t_i^2 + \gamma_3 child_{i,t} + \gamma_4 emp_{i,t-1} + \gamma_5 emp_{i,t-2} + \gamma_6 hy_{i,t-1} + \varepsilon_{i,t}^{emp} \quad (3)$$

$$I_{i,t}^{ft} = \gamma_{10} + \gamma_{11} t_i + \gamma_{12} spt_{i,t} + \gamma_{13} sft_{i,t} + \gamma_{14} child_{i,t} + \gamma_{15} ft_{i,t-1} + \gamma_{16} ft_{i,t-2} + \varepsilon_{i,t}^{ft} \quad (4)$$

where:

- $child_{i,t}$ is a dummy which equals 1 if individual i has at least one child five years or younger at age t and zero otherwise.
- $spt_{i,t}$ is a dummy which equals 1 if individual i at age t has a spouse who is employed part-time and zero otherwise.
- $sft_{i,t}$ is a dummy which equals 1 if individual i at age t has a spouse who is employed full-time and zero otherwise.

The estimated regressions for female employment / unemployment and full-time / part-time discrimination provide correct predictions respectively for 86.7 % and 85.6 % of observations and all coefficients have the expected signs. As for the male probit regressions, the relatively high degree of uncertainty

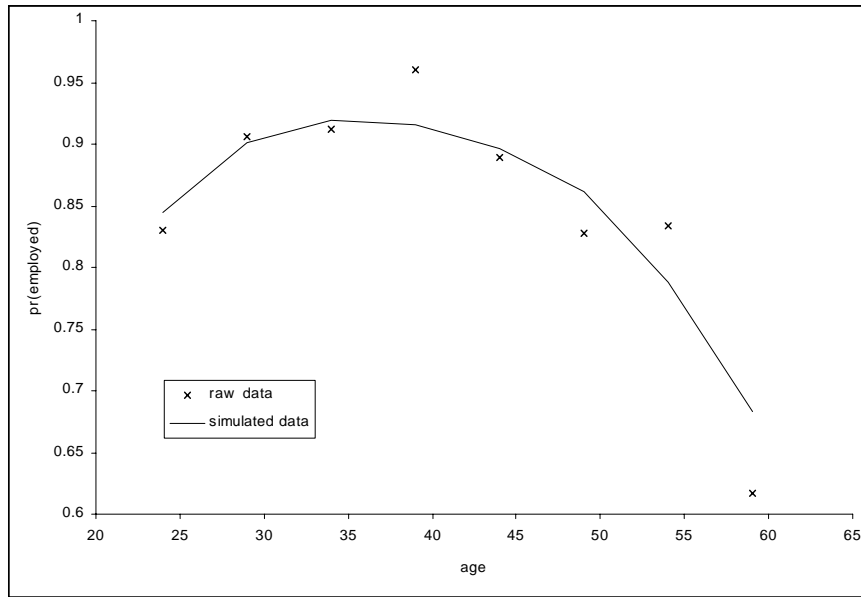


Figure 2: Probability of male employment versus age

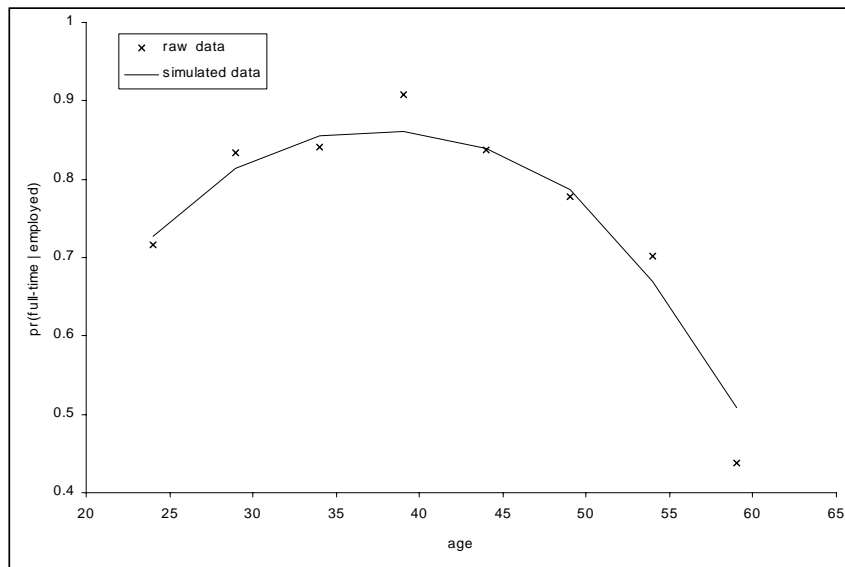


Figure 3: Probability for employed males of full-time employment versus age

Table 2: Regression statistics for probit models of female employment

	γ_0	γ_1	γ_2	γ_3
Regression Coefficients	-1.8727	0.07072	-0.1145E-02	-0.3478
	(0.8250)	(0.0415)	(0.50E-03)	(0.1356)
Adopted Coefficients	-1.9025	0.06343	-0.9052E-03	-0.0728
	γ_4	γ_5	γ_6	
Regression Coefficients	1.5964	0.64934	0.40742	
	(0.1323)	(0.1356)	(0.1479)	
Adopted Coefficients	1.5964	0.64934	0.40742	
	γ_{10}	γ_{11}	γ_{12}	γ_{13}
Regression Coefficients	0.14060	-0.01925	-0.95436	-0.52724
	(0.3147)	(0.74E-02)	(0.3934)	(0.1376)
Adopted Coefficients	-0.10601	-0.00725	-0.95436	-0.62724
	γ_{14}	γ_{15}	γ_{16}	
Regression Coefficients	-0.42774	1.7473	0.59431	
	(0.1769)	(0.1789)	(0.1792)	
Adopted Coefficients	-0.12774	1.7473	0.59431	

associated with the values of many of the estimated coefficients prompted an adjustment of the constant and age coefficients, which produced the ‘adopted coefficient’ series listed in Table 2. Figures 4 and 5 depict the relationship between the raw and simulated data.

4 Simulation of Labour Income

The primary impediment to producing a model that captures the essential dynamic characteristics of labour income for individuals in Australia is the scarcity of the required panel data. The most recent and comprehensive longitudinal data set that provides information suitable for income model estimation in Australia is the SEUP, which was described in section 3. This data set is, however, limited in two important respects:

1. SEUP provides unit record data for only three consecutive years. This implies that the fixed effects income models that are routinely examined for countries where more comprehensive data are available can not be estimated with a sufficient degree of precision.⁹
2. The data provided by SEUP do not enable wage and salary income to be isolated, where the closest approximate that can be obtained is total annual income less government benefits received. This figure includes income from investments, own businesses, superannuation, and so on, which relate to saving rather than to labour.

One model of individual income dynamics that can be estimated by the available data, such that both of these problems are mitigated, is found in Creedy (1985).

⁹See Appendix B for discussion regarding the estimation problems encountered for two of the fixed effects models considered using the SEUP data.

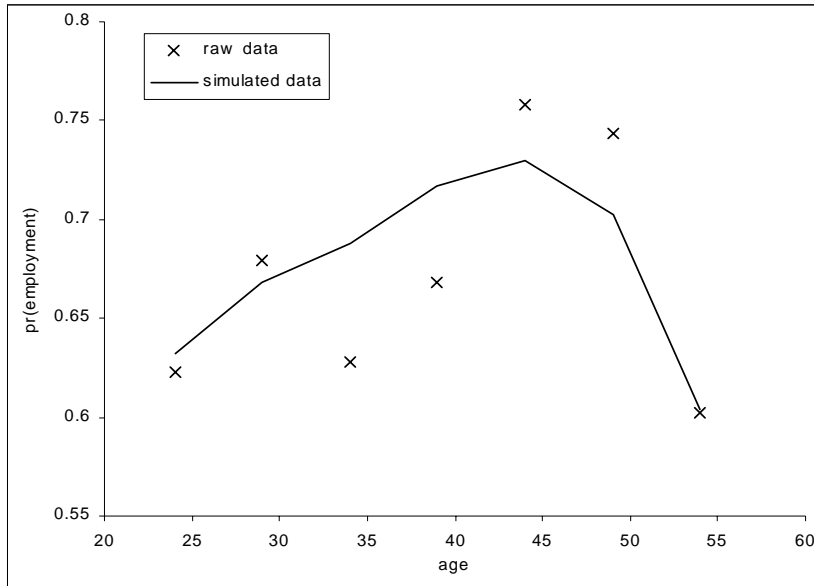


Figure 4: Probability of female employment versus age

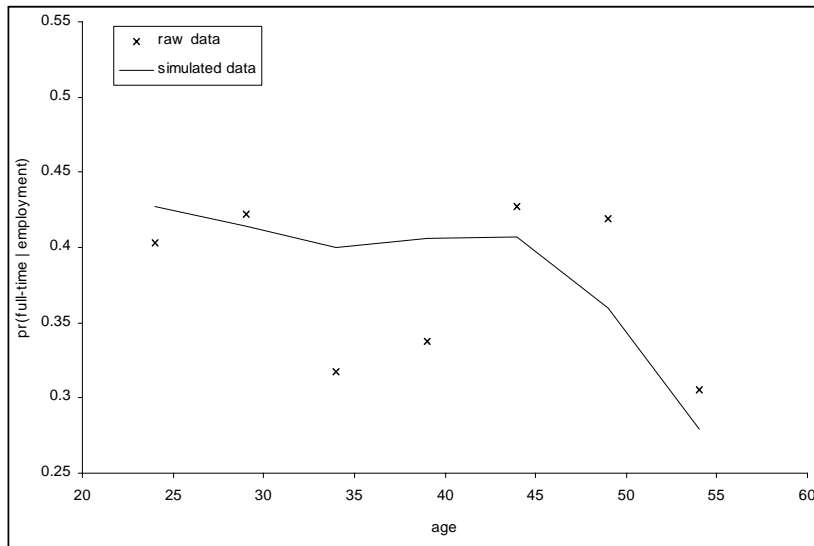


Figure 5: Probability for employed females of full-time employment versus age

4.1 Theory

The model described by Creedy (1985) divides the income simulation procedure into two separate parts; one that generates the underlying mean trend, and another that generates individual variation from the mean. Let the ‘underlying income’, y_{it} , define the income that individual i would earn at age t if they were employed. Defining m_t as the geometric mean of all y_{it} , then, for a total population of n individuals,

$$\begin{aligned} m_t &= \sqrt[n]{y_{1t} \cdot y_{2t} \cdots y_{nt}} \\ \log(m_t) &= \frac{1}{n} \sum_{i=1}^n \log(y_{it}) \end{aligned} \quad (5)$$

The central assumption of the model is that the proportional variation of any individual i ’s underlying income from one year to the next deviates from the proportional variation of the respective geometric mean by a random variable with a mean of zero. That is,

$$\frac{\dot{y}_{it}}{y_{it}} = \frac{\dot{m}_t}{m_t} + u_{it} \quad (6)$$

Defining $z_{it} = \log\left(\frac{y_{it}}{m_t}\right)$ and substituting into equation (6) obtains,

$$\dot{z}_{it} = u_{it} \quad (7)$$

Discretising equation (7) arrives at the following first order auto-regressive equation:

$$z_{it} = z_{i(t-1)} + u_{it} \quad (8)$$

Following Kalecki (1945), regression of incomes toward the mean implies that if $y_{it} > m_t$, then on average, $\frac{\dot{y}_{it}}{y_{it}} < \frac{\dot{m}_t}{m_t}$ and vice versa.¹⁰ When $y_{it} \geq m_t$ and $\beta < 1$, then $(1 - \beta) \log\left(\frac{y_{it}}{m_t}\right) \geq 0$. Regression toward the mean is allowed for in the model by subtracting $(1 - \beta) z_{i(t-1)}$ from the right-hand side of equation (8) to obtain,

$$z_{it} = \beta z_{i(t-1)} + u_{it} \quad (9)$$

The value of β consequently determines the variation of individual incomes relative to the geometric mean. When $\beta < 1$, regression toward the mean arises as described above. Regression away from the mean is characterised by $\beta > 1$, and when $\beta = 1$, the Gibrat process obtains.¹¹

In the simplest form of the model, u_{it} is a random variable independent of $z_{i(t-1)}$. Added complexity can, however, be incorporated into the model by redefining u_{it} . Adopting an auto-regressive form for u_{it} , for example, includes an allowance for the persistence of random effects that affect the growth of an

¹⁰See Bliss (1999) for a recent discussion of Galtonian regression toward the mean.

¹¹Equation (8) was first applied to income data by Gibrat (1931), and so in the above context is referred to as a Gibrat process, though more generally it is known as a Markov process.

individual's income. Assuming $u_{it} = \gamma u_{i(t-1)} + \varepsilon_{it}$, where ε_{it} is an independently distributed random variable with zero mean, and substituting into equation (9) obtains the following reduced form,

$$z_{it} = (\beta + \gamma) z_{i(t-1)} - \gamma\beta z_{i(t-2)} + \varepsilon_{it} \quad (10)$$

Equation (10) characterises the dynamic variation of individual incomes from the respective mean, based on the variation observed in the two preceding periods.

The second part of the model, which characterises the trend of mean income, uses a simple fourth order polynomial of age as described by equation (11):

$$m_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3 + \alpha_4 t^4 \quad (11)$$

With regard to simulation, mean income for any given age is derived from equation (11). Individual incomes are then generated relative to the prescribed mean via reference to the simulated variation. Specifically, base period variation (for some given minimum age) is exogenously imposed, where the relative values of individual incomes are allocated by generating a random variable. Equation (10) is then used to generate the relative values of individual incomes in subsequent years, based on the respective relative values of previous years and a random element, ε_{it} . Hence, the simulation model described by Creedy (1985) is composed of three elements; the mean generating component characterised by equation (11), the exogenously imposed variation of incomes for the minimum age, and the dynamic income variation component characterised by equation (10).

4.2 Model Estimation

The fact that the model described above separates simulation of mean income from simulation of income variation is advantageous in light of the aforementioned limitations of the SEUP data set. Specifically, only equation (10), the dynamic variation component of the model, needs to be estimated using SEUP. The 1996 Income Distribution Survey (IDS), which provides records for a representative cross-section of the Australian population, can be used to estimate both equation (11) (the component of the model that generates geometric mean incomes by age), and the standard deviation of incomes for the base age group. If it is assumed that the measure of income variation by age, sex, and employment status derived from SEUP is close to the associated variation of wage and salary income for the Australian population, then it is reasonable to adopt the coefficients estimated using SEUP as starting values for calibrating equation (10).

Data from the 1996 IDS were consequently used to obtain estimates for equation (11) and the standard deviation of log incomes for individuals between the ages of 15 and 19, σ_0 , which is adopted as the base age group.¹² Similarly, regression estimates for equation (10), including estimates of the

¹²In addition to forming the 'foundation' upon which individual variation from the population geometric mean is based, the values of σ_0 obtained are used to simulate the incomes of spouses between the ages of 17 and 19.

standard deviation of the associated error term, σ_ε , were obtained using the SEUP data.¹³ Four sub-populations were considered based on sex and labour force status where Tables 3 and 4 present estimates obtained respectively from IDS and SEUP data.

Table 3: Regression Results for Mean log Wage and Salary Income

	males		females	
	full time	part time	full time	part time
$\hat{\alpha}_0$	-1.1370 (1.1822)	-12.9337 (8.1421)	-2.5893 (1.5347)	-9.6795 (3.3397)
$\hat{\alpha}_1$	1.1054 (0.1494)	2.3533 (1.0291)	1.2973 (0.1940)	2.0433 (0.4221)
$\hat{\alpha}_2$	-0.0395 (0.0067)	-0.0908 (0.0459)	-0.0485 (0.0087)	-0.0808 (0.0188)
$\hat{\alpha}_3$	6.27E-04 (1.25E-04)	1.52E-03 (8.64E-04)	7.89E-04 (1.63E-04)	0.0014 (3.54E-04)
$\hat{\alpha}_4$	-3.70E-06 (8.46E-07)	-9.40E-06 (5.82E-06)	-4.72E-06 (1.10E-06)	-8.74E-06 (2.39E-06)
R Square	0.9968	0.9070	0.9925	0.9770
$\hat{\sigma}_0$	0.9524	1.0107	1.0249	1.0378

Table 4: Regression Results for Dynamic Income Variation Model

	males		females	
	full time	part time	full time	part time
$(\hat{\beta} + \hat{\gamma})$	0.71666 (0.0791)	0.73879 (0.3482)	0.59833 (0.1163)	0.60830 (0.1166)
$-\hat{\beta}\hat{\gamma}$	0.27101 (0.0775)	0.22343 (0.3127)	0.39610 (0.1156)	0.36663 (0.1154)
$\hat{\beta}$	0.99032 (0.0030)	0.96930 (0.0303)	0.99602 (0.0026)	0.98175 (0.0080)
$\hat{\gamma}$	-0.27366 (0.0784)	-0.23051 (0.3274)	-0.39769 (0.1161)	-0.37345 (0.1172)
R Squared	0.9923	0.9571	0.9961	0.9808
$\hat{\sigma}_\varepsilon$	0.0870	0.2226	0.0626	0.1392

All of the R squared values displayed in Tables 3 and 4 indicate that the respective equations adequately capture observed variation, and the hypothesis of heteroscedastic errors for equation (10) is rejected at the 95% confidence level using the White test.¹⁴ Furthermore, with the exception of $-\hat{\beta}\hat{\gamma}$ and hence, $\hat{\gamma}$, of the part-time employed male equation, all of the estimated coefficients displayed in Table 4 are highly significant, and all of the values of $\hat{\beta}$ and $\hat{\gamma}$ derived are consistent with expectations

¹³Estimates for β and γ were obtained following Creedy (1985), pp. 40-41. A typographical error exists in the expression to obtain the variance of β in Creedy (1985), p. 41. Using the notation of Creedy (1985), the following equation can be obtained for the variance following Goldberger (1964), p. 124, $var(\beta) = var(a)(\delta\beta/\delta a)^2 + var(b)(\delta\beta/\delta b)^2 + 2cov(a, b)(\delta\beta/\delta a)(\delta\beta/\delta b)$.

¹⁴A Chow parameter stability test with respect to age was performed for equation (10), and significant variation could not be rejected at the 95% confidence level. The effects of age, however, were found to be quite small and hence are neglected with regard to the model adopted.

and previous findings. The estimated standard errors in Table 3, however, indicate that some of the coefficient estimates are not significantly different from zero at the 95 % confidence interval. Nevertheless, a fourth order polynomial was required for equation (11) to capture the relatively flat relationship of mean log income with age between the ages of 30 and 50, as depicted by the associated graphs presented in Appendix A.

Following the initial estimation of model coefficients, four populations, each comprised of 5000 individuals, were generated to calibrate the income simulation model, where it's ability to reflect wage and salary data from the 1996 IDS was used as the basis for comparison.¹⁵ Table 5 provides the model coefficients that were subject to variation as part of the calibration procedure juxtaposed with their uncalibrated counterparts. Associated figures depicting the simulated distributions versus the raw distributions derived from the 1996 IDS are provided in Appendix A.

Table 5: Calibrated Model Coefficients				
	males		females	
	full time	part time	full time	part time
Calibrated Coefficients				
β	0.99032	0.98175	0.99602	0.98175
$\tilde{\gamma}$	-0.0737	-0.37350	-0.39769	-0.37345
$\tilde{\sigma}_0$	0.37245	0.4000	0.4249	0.5100
$\tilde{\sigma}_\varepsilon$	0.06901	0.1750	0.0586	0.1300
Uncalibrated Coefficients				
$\hat{\beta}$	0.99032	0.96930	0.99602	0.98175
$\hat{\gamma}$	-0.27366	-0.23051	-0.39769	-0.37345
$\hat{\sigma}_0$	0.9524	1.0107	1.0249	1.0378
$\hat{\sigma}_\varepsilon$	0.0870	0.2226	0.0626	0.1392

The most evident variation of the parameters listed in Table 5 is the reduction imposed on σ_0 . This result is to be expected given that the original estimate, $\hat{\sigma}_0$, was based on a small population due to the fact that a relatively small proportion of the population are defined as working in the lowest age group of the 1996 IDS.

4.3 The Model and Traditional Approaches to Income Estimation

Given the preceding discussion, it is useful to compare the model of Creedy (1985), with traditional approaches to income estimation. Wage models in the literature usually take the form:

$$\ln(y_{it\tau}) = \alpha_i + X_{it\tau}\delta + \xi_{it\tau} \quad (12)$$

¹⁵Due to small sample problems associated with the respective IDS sample sets, the parameters $\tilde{\sigma}_0$ and $\tilde{\sigma}_\varepsilon$ for part-time employed males and females were adjusted to ensure that the part-time income distributions take reasonable values compared with the full-time distributions obtained.

where α_i is an individual specific effect to allow for motivation and ability, for example.
 $X_{it\tau}$ is comprised of characteristics that are assumed to affect the income of individual i , aged t at time τ .
 $\xi_{it\tau}$ is a random variable that is independent across time, age, and individuals, and has a mean of zero. This component allows, for example, for the effect of luck

Given that the simulation model focuses on a single cohort aged 20 in 1970, the age, cohort (or vintage) and time effects can be aggregated. With regard to equation (12), this implies that the t and τ subscripts may be replaced by a single subscript, which for convenience may be defined as t (that is, the time and cohort effects associated with the τ subscript are subsumed by the age effect).

It is evident that extensive panel data are required to obtain accurate estimates of the coefficients associated with this model for any substantial heterogeneous population. Relatively few data are required to estimate the model proposed by Creedy (1985) because it omits the effects of individual characteristics, such that the i subscript can be dropped from the X_{it} variable in equation (12). Consider the model that incorporates individual dynamic variation characterised by equation (10), the reduced form for which is specified in terms of the random measures ε_{it} as indicated by equation (13):

$$z_{it} = F(t, \varepsilon_{it-}) \quad (13)$$

where ε_{it-} denotes the current and all past values of ε for individual i at age t . The following model for individual income is obtained from equation (13) by substituting in the identity $z_{it} = \ln(y_{it}) - \ln(m_t)$:

$$\ln(y_{it}) = \ln(m_t) + F(t, \varepsilon_{it-}) \quad (14)$$

Define χ_i as the random variable associated with base period variation, and redefine ε_{it} as the random variable used to generate variation in all subsequent periods, t . Substitution into equation (14) obtains:

$$\ln(y_{it}) = G(t, \chi_i) + \ln(m_t) + H(t, \varepsilon_{it-}) \quad (15)$$

The estimates for $\tilde{\beta}$ displayed in Table 5 are all close to one, while those for $\tilde{\gamma}$ are close to -0.3 . Consequently, from equation (10),

$$z_{it} \simeq 0.7z_{it-1} + 0.3z_{it-2} + \varepsilon_{it} \quad (16)$$

for all ages following the base year. The values of z_{it} for the first few years, including the base year can consequently be derived from equation (16) as follows:¹⁶

$$\begin{aligned} z_{i0} &\simeq \chi_i \\ z_{i1} &\simeq \chi_i + \varepsilon_{i1} \\ z_{i2} &\simeq \chi_i + \varepsilon_{i2} + 0.7\varepsilon_{i1} \\ z_{i3} &\simeq \chi_i + \varepsilon_{i3} + 0.7\varepsilon_{i2} + 0.79\varepsilon_{i1} \\ z_{i4} &\simeq \chi_i + \varepsilon_{i4} + 0.7\varepsilon_{i3} + 0.79\varepsilon_{i2} + 0.763\varepsilon_{i1} \\ z_{it} &\simeq \chi_i + H(t, \varepsilon_{it-}) \end{aligned}$$

¹⁶It is assumed that $z_{i0} = z_{i-1}$ when calculating z_{i1} .

Hence, given the coefficient estimates obtained for β and γ , $G(t, \chi_i) \simeq \chi_i$ in equation (15), so that χ_i may be interpreted as approximately equivalent to the individual specific fixed effect α_i in equation (12). The values of z_{it} presented above indicate that the individual random variables, ε_{it} , also affect income with a high degree of persistence. This observation is best considered in conjunction with the effect of $\ln(m_t)$ when comparing equation (15) to (12).

Assume that all of the variables except age included in X_{it} are measured with reference to the respective population means. In this case the coefficients on the age variables would determine the impact of age (for given time and cohort) on the mean income. This may be compared with the component $\ln(m_t)$ in model (15). Other variables included in X_{it} (such as work experience, education, marital status, health status *et cetera*) usually exhibit some persistence with age and consequently reflect the systematic variation of individual incomes from the population mean with respect to age. In contrast, ξ_{it} allows for the impact of pure chance events on individual income variation. These elements combined may consequently be compared with the persistent structure of $H(t, \varepsilon_{it-})$.

The three basic elements of the Creedy (1985) model described in section 4.1, can consequently be related to the fixed effects model usually estimated in the published literature. Specifically, for the estimates obtained, base period variation approximates the individual specific fixed effect. The relationship between mean income and age relate to the associated coefficients on age usually included in X_{it} of equation (12). The dynamic variation of individual income from the respective cohort mean, which is partially explained by underlying exogenous variables, and partially by the random component ξ_{it} in equation (12), is modelled implicitly in (15) by equation (10). Hence the reduced data requirements associated with the Creedy (1985) model are obtained at the cost of decreased explanatory power, a cost which is mitigated by the fact that the model is produced specifically for simulation and not for its explanatory power *per se*.

Adopting the above interpretation of the Creedy (1985) model has important implications regarding the relationship imposed between the incomes of husbands and wives. Specifically, there is general agreement in the published literature that a strong positive correlation exists between many of the underlying personal attributes typically included in X_{it} for spouses.¹⁷ Given the connection between these personal attributes and the values of ε_{it} included in the Creedy (1985) model, the values of ε_{it} generated for spouses are related via equations (17) and (18):

$$\varepsilon_{i(m)t} = \sigma_{(m)\varepsilon} \nu_{i(m)t} \tag{17}$$

$$\begin{aligned} \varepsilon_{i(f)t} &= \sigma_{(f)\varepsilon} \nu_{i(f)t} \\ &= \sigma_{(f)\varepsilon} \left[\frac{\lambda \nu_{i(m)t} + (1 - \lambda) \nu_{it}}{(2\lambda^2 + 1 - 2\lambda)^{0.5}} \right] \end{aligned} \tag{18}$$

¹⁷See for example, Winch (1958), Vandenberg (1972), and Alström (1961).

where $0 \leq \lambda \leq 1$ is an exogenously specified parameter

- $\varepsilon_{i(m)t}$ is the random variable used to generate the income of the male in family i at age t (where they exist).
- $\varepsilon_{i(f)t}$ is the random variable used to generate the income of the female in family i at age t (where they exist).
- $\sigma_{(m)\varepsilon}$ is the estimate of standard deviation relevant for males obtained from Table 5
- $\sigma_{(f)\varepsilon}$ is the estimate of standard deviation relevant for females obtained from Table 5
- $\nu_{i(m)t}$ is a standard normal variable generated to obtain the income of the male in family i at age t (where they exist).
- $\nu_{i(f)t}$ is a standard normal variable generated to obtain the income of the female in family i at age t (where they exist).
- ν_{it} is a standard normal variable generated independently from $\nu_{i(m)t}$.

From equation (18), it is evident that $\nu_{i(f)t}$ will be normally distributed with a mean of zero and variance of one, and that the correlation between the standard normal deviates used to generate male and female incomes is equal to the following:

$$\rho = \frac{\lambda}{(2\lambda^2 + 1 - 2\lambda)^{0.5}} \quad (19)$$

Consequently, as λ is increased from zero to one, so too does the correlation between the male and female standard normal deviates. This may be interpreted, following the previous discussion, as an increase in the correlation between the personal attributes of spouses that affect their respective labour incomes.

Adopting the above framework provides some exogenous control over the correlation of spouse wages. A value of $\lambda = 1$, for example, implies that if an individual should marry, their income relative to the distribution of those of the same age and sex should be close to that of their spouse. The actual correlation of incomes can not, however, be directly inferred from λ due to the effect of the relationship between the mean log incomes of males and females with age. To quantify the effect of varying λ on the correlation of spouse incomes, the incomes for an initial cohort of 10,000 individuals were generated with $\lambda = 0$, and then regenerated for $\lambda = 0.9$. The correlation between the incomes of spouses was found to increase from 0.2301 to 0.6658, and 0.4019 to 0.8962, for the full-time and part-time employed respectively.

5 Conclusion

This paper has outlined a microsimulation model of cohort earnings that is based on the Australian population. The model has been created to enable analyses to be undertaken for the working-lifetime and is of particular value given the scarcity of Australian panel data. Care has been taken to structure

the model to facilitate transparent sensitivity analysis, subject to the limitations imposed by the data used for calibration.

The annual variation of labour is modelled as a series of binomial events, where change is generated using transition probability functions. This method departs from the use of transition matrices that is common for microsimulation models and facilitates transparent sensitivity analysis. The probability functions adopted are statistical in nature and not based on theoretical maximising behaviour. The functions are, however, shown to describe the observed data well, so that little detail is lost due to the restrictions imposed.

The labour income component has two desirable features. First, consistent with the objective of transparency, it is highly parsimonious, using two meaningful functions to generate the income history of individuals. Second, the parameters of the two functions used by the model can be estimated by the limited data that are available. Although the labour income component of the model takes a non-standard form, a close relationship is shown to exist between it and standard earnings equations.

Despite the highly parsimonious nature of the microsimulation model, generated cohort earnings are shown to closely reflect the Australian data used for calibration.

References

- Alström, C. H. (1961), 'A study of inheritance of human intelligence', *Acta Psychiatrica et Neurologica Scandinavia* **36**, 175–202.
- Auerbach, A. J. (1997), *Fiscal Policy: Lessons from Economic Research*, MIT Press, London.
- Basu, N., Pryor, R. & Quint, T. (1998), 'ASPEN: A microsimulation model of the economy', *Computational Economics* **12**, 223–241.
- Bliss, C. (1999), 'Galton's fallacy and economic convergence', *Oxford Economic Papers* **51**, 4–14.
- Caldwell, S. (1997), *Corsim 3.0 User and Technical Documentation*, Ithaca, New York.
- Cameron, G. & Muellbauer, J. (2000), Earnings, unemployment, and housing: Evidence from a panel of British regions. CEPR Discussion Paper 2404.
- Cogneau, D. & Robilliard, A. S. (2000), Growth, distribution and poverty in Madagascar: Learning from a microsimulation model in a general equilibrium framework. Trade and Macroeconomics Division, International Food Policy Research Institute: 61.
- Creedy, J. (1985), *Dynamics of Income Distribution*, Basil Blackwell, Oxford.

- Creedy, J. & van de Ven, J. (1999), ‘The effects of selected Australian taxes and transfers on annual and lifetime inequality’, *Australian Journal of Labour Economics* **3**, 1–22.
- Creedy, J. & van de Ven, J. (2001), ‘Decomposing redistributive effects of taxes and transfers in Australia: Annual and lifetime measures’, *Australian Economic Papers* **40**, 185–198.
- de Janvry, A., Sadoulet, E. & Fargeix, A. (1991), ‘Politically feasible and equitable adjustment: Some alternatives for Ecuador’, *World Development* **19**, 1577–1594.
- Dervis, K., Melo, J. D. & Robinson, S. (1982), *General Equilibrium Models for Development Policy*, Cambridge University Press, Cambridge.
- Giannarelli, L. (1992), *An Analyst’s Guide to TRIM2*, Urban Institute Press, Washington D.C.
- Gibrat, R. (1931), *Les Inegalites Economiques*, Sirey, Paris.
- Goldberger, A. S. (1964), *Econometric Theory*, Wiley, New York.
- Harding, A. (1993), *Lifetime Income Distribution and Redistribution: Applications of a Microsimulation Model*, North-Holland, London.
- Hotz, V. J., Klerman, J. A. & Willis, R. J. (1997), The economics of fertility in developed countries, in M. R. Rosenzweig & O. Stark, eds, ‘Handbook of Population and Family Economics’, Elsevier Science, Oxford.
- Kalecki, M. (1945), ‘On the Gibrat distribution’, *Econometrica* **13**, 161–170.
- Macunovich, D. J. (1998), ‘Fertility and the Easterlin hypothesis: An assessment of the literature’, *Journal of Population Economics* **11**, 52–111.
- Meagher, G. A. (1993), Forecasting changes in the income distribution: An applied general equilibrium approach, in A. Harding, ed., ‘Microsimulation and Public Policy’, Elsevier, Amsterdam.
- Merz, J. (1991), ‘Microsimulation – A survey of principles, developments, and applications’, *International Journal of Forecasting* **7**, 77–104.
- Mincer, J. (1974), *Schooling, Experience, and Earnings*, Columbia University Press, New York.
- Orcutt, G. (1957), ‘A new type of socio-economic system’, *Review of Economics and Statistics* **58**, 773–797.
- Redmond, G., Sutherland, H. & Wilson, M. (1998), *The Arithmetic of Tax and Social Security Reform: A Users’ Guide to Microsimulation Methods and Analysis*, Cambridge University Press, Cambridge.

Sutherland, H. (1995), Static microsimulation models in Europe: A survey. University of Cambridge Department of Applied Economics Working Paper: 9523.

Taylor, L. (1990), *Socially Relevant Policy Analysis. Structural Computable General Equilibrium Models for the Developing World*, MIT Press, Cambridge, Mass.

van de Ven, J. (1998), A dynamic cohort microsimulation model. University of Melbourne Department of Economics Research Paper: 637.

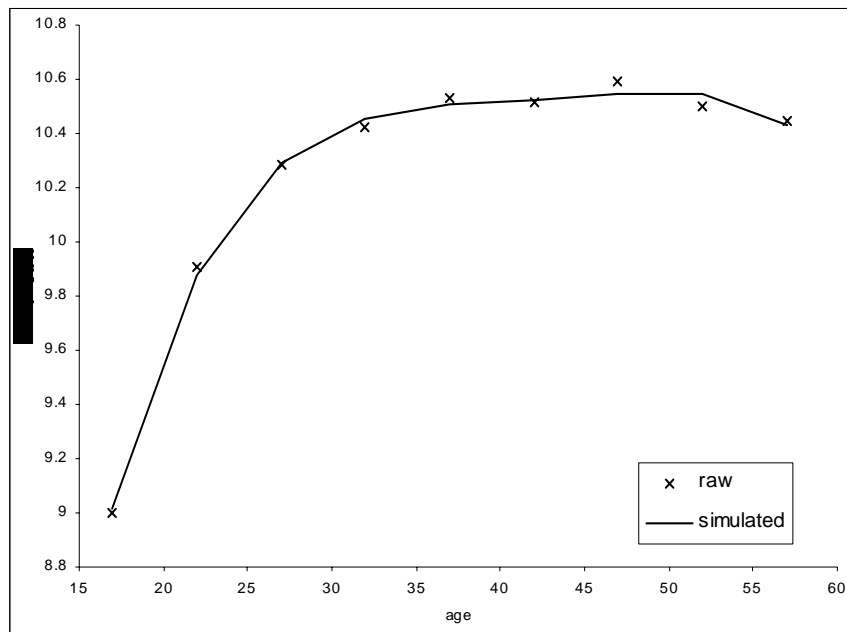
van de Ven, J. (2004), Redistribution During the Working Lifetime. Doctoral Thesis, University of Oxford, Part II.

van de Ven, J. (2005), Taxation and Redistribution in Australia and the UK – Evidence from Microsimulation Analyses. NIESR Working Paper.

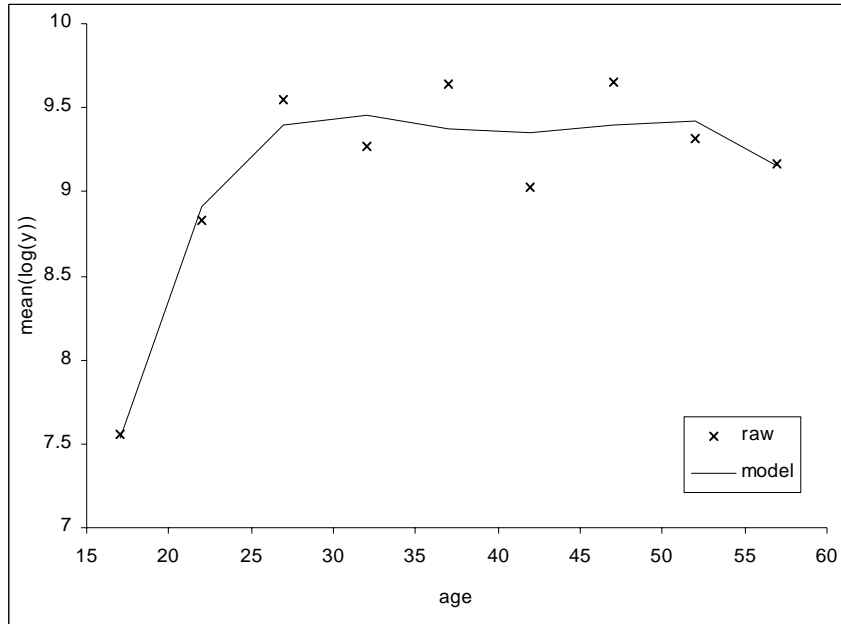
Vandenberg, S. G. (1972), ‘Assortative mating, or who marries whom?’, *Behavior Genetics* **2**, 127–157.

Winch, R. F. (1958), *Mate Selection*, Harper, New York.

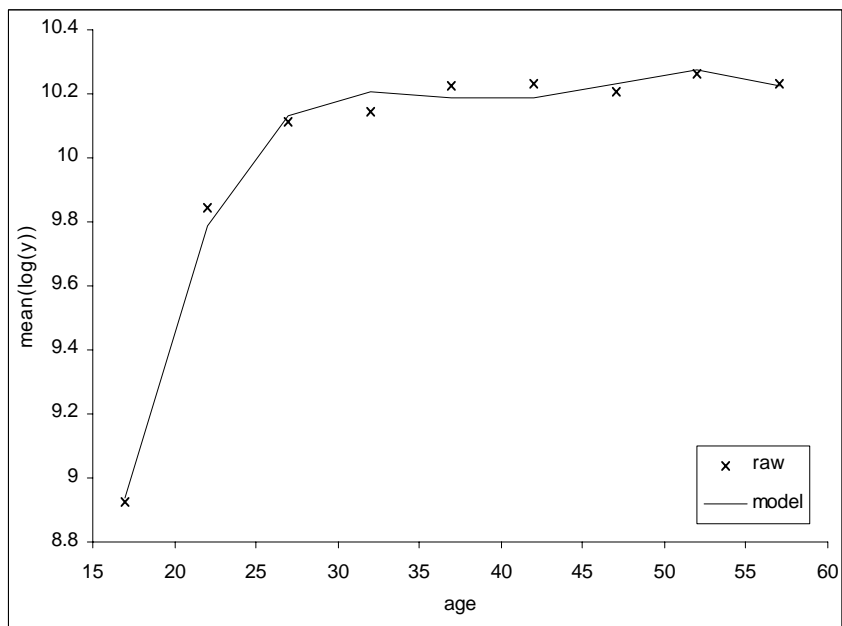
A Supplementary Figures - labour income



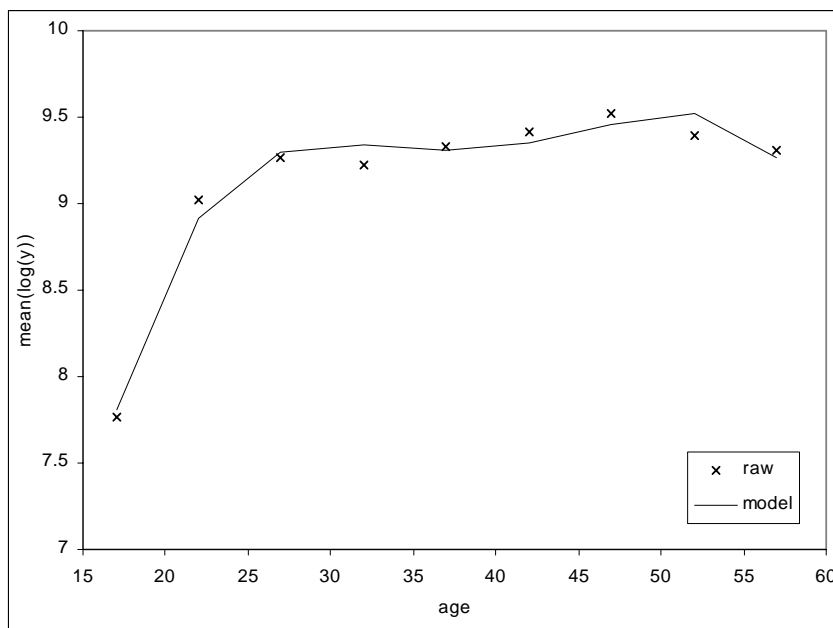
Full-time Employed Male Income by Age - Raw versus Simulated Data



Part-time Employed Male Income by Age - Raw versus Simulated Data



Full-time Employed Female Income by Age - Raw versus Simulated Data



Part-time Employed Female Income by Age - Raw versus Simulated Data

B Fixed Effects Estimation

Table 6 below provides the regression output from a fixed effects estimation of the following error correction model, which is based on Cameron & Muellbauer (2000):

$$\begin{aligned} \Delta \ln(y_{it}) = & \alpha(\mu + \theta_i + \beta_1 age_{it-1} + \beta_2 age_{it-1}^2 + \beta_3 mstat_{it-1} + \\ & + \beta_4 kids_{it-1} + \beta_5 phs_{it-1} - y_{it-1}) + \\ & + \gamma_1 \Delta mstat_{it} + \gamma_2 \Delta kids_{it} + \gamma_3 \Delta phs_{it} + \varepsilon_{it} \end{aligned} \quad (20)$$

where y_{it} denotes the income of individual i in year t , age_{it} denotes their age in years, $mstat_{it}$ is a marital status dummy equal to one if they are identified as married and zero otherwise, $kids_{it}$ is equal to their number of dependant children (under the age of 17), phs_{it} is a post high school education dummy, and θ_i is an individual specific effect. The p-values of the estimated coefficients are provided in brackets in Table 6.

The results displayed in Table 6 indicate that either the model is a poor representation of reality, or, more likely, that the data are insufficiently rich to obtain accurate estimates. Specifically, many of the coefficients have unexpected signs, the values obtained for α are excessively large, and the p-values indicate poor precision, although this last issue was expected given the small number of time-series observations used. It is also worth noting that an experience variable could not be included in the analysis because the panel provides only three time periods of data. This meant that the populations used to undertake each of the three regressions needed to be full or part time employed for all three

Table 6: Regression Coefficients of Error Correction Fixed Effects Model

employed	males	females	
	full time	full time	part time
α	1.2860 (0.0001)	1.3898 (0.0001)	1.3581 (0.0001)
μ	11.8121	9.4764	5.3971
θ_i	$N(0, 2.01)$	$N(0, 0.67)$	$N(0, 8.36)$
β_1	-0.1103 (0.5854)	0.00679 (0.9612)	-.42535 (0.2302)
β_2	0.0043 (0.3706)	0.0011 (0.6880)	0.0110 (0.0803)
β_3	0.1459 (0.5565)	-0.7594 (0.0057)	0.1940 (0.8420)
β_4	0.0014 (0.4180)	0.0025 (0.0873)	-0.0024 (0.5858)
β_5	-0.4069 (0.5407)	0.0025 (0.6338)	-1.5569 (0.4336)
γ_1	-0.0330 (0.8952)	-0.6277 (0.0212)	0.1682 (0.8889)
γ_2	0.0002 (0.8706)	0.0022 (0.1035)	-0.0022 (0.5944)
γ_3	0.0620 (0.9080)	0.1799 (0.6950)	1.0447 (0.3861)

sample periods, and consequently, including both an age and an experience variable would lead to multicollinearity.¹⁸ In addition, the three estimated models displayed in Table 6 are calculated using individuals who are either full time or part time employed for all three periods of the panel survey. This means that the estimates are calculated using populations that do not reflect the transiently employed, who form part of the simulated population. In light of these problems, the model characterised by equation (20) cannot be used.

Equation (21) is a highly restricted form of the fixed effect income models typically applied in the published literature.¹⁹ The associated regression results are provided in Table 7.

$$\ln y_{it} = \theta_i + \beta_1 \ln y_{it-1} + \beta_2 kids_{it} + \beta_3 mstat_{it} + \beta_4 phs_{it} + \beta_5 age_{it} + \beta_6 age_{it}^2 + \varepsilon_{it} \quad (21)$$

The results in Table 7 indicate that better estimates were obtained for the model characterised by equation (21) than that of equation (20). Notably, most of the estimated coefficients have the expected signs, although the associated p-values continue to indicate that the estimates obtained are subject to a high degree of uncertainty.²⁰ This is, however, not surprising given that the estimates are still calculated using only two time series observations. The low precision of the estimated coefficients

¹⁸ Although the effect of age is intuitively related to the fixed effect, the form that it takes in equation (20) implies that the coefficient on the first order age variable aggregates all of effects that change at a constant rate with time.

¹⁹ Mincer (1974)

²⁰ Although the age and education coefficients of the part time female regression remain problematic.

Table 7: Regression Coefficients of log Linear Fixed Effects Income Model

employed	males	females	
	full time	full time	part time
θ_i	$N(12.54, 1.26)$	$N(12.93, 0.64)$	$N(15.45, 2.32)$
β_1	-0.2801 (0.0001)	-0.4018 (0.0001)	-0.3631 (0.0001)
β_2	-0.0003 (0.8337)	0.0022 (0.1050)	-0.0081 (0.0330)
β_3	-0.0175 (0.9444)	-0.4598 (0.0806)	-0.0376 (0.9770)
β_4	0.0566 (0.9160)	0.2102 (0.6484)	-0.2330 (0.7388)
β_5	0.1371 (0.4005)	0.0928 (0.5627)	-0.0578 (0.8906)
β_6	-0.0028 (0.1744)	-0.0013 (0.5186)	0.0001 (0.9807)

displayed in Table 7, combined with the restricted form that is imposed by the limited nature of the SEUP data, suggest the use of another model for income simulation.