MODELLING ANNUAL EARNINGS WITH UNEMPLOYMENT:

NON-RANDOM SELECTION IN FEMALE WORKERS

By

Oliver Robertson

Supervisors

Dr. Dean Hyslop and Dr. Yiğit Sağlam

A thesis

submitted to the Victoria University of Wellington

in fulfilment of the requirements for the degree of

Master of Arts

in Economics

Victoria University of Wellington

2019

The results in this masters thesis are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Statistics New Zealand.

The opinions, findings, recommendations, and conclusions expressed in this thesis are those of the author(s), not Statistics NZ.

Access to the anonymised data used in this study was provided by Statistics NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this thesis have been confidentialised to protect these groups from identification and to keep their data safe. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

Acknowledgements

I would like to my family, Ikey, Chris, and Sarah, for their support while I have been writing this thesis. My supervisors Dean Hyslop and Yigit Saglam have both been vital in every step of this process, thank you both for all of your advice and support in what became a very long project. And many thanks to my loving partner Anna, who provided both moral support and significant advice.

I am also grateful for a scholarship funded by the Royal Society of New Zealand Marsden Fund Grant MEP1301.

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 tend to have more frequent changes in employment status, and more periods of unemployment than their male counterparts. These periods of unemployment result in observations with zero earnings, and common transformations such as the logarithm are not defined for zero values. This means that any analysis of the logarithm of earnings is forced to exclude periods where an individual does not work, and cannot take into account the effect of moving into or out of employment. The higher rate of unemployment in female workers also increases the risk of sample selection bias. If selection into employment is non-random, then estimating earnings equations based on only workers will result in biased estimates. This thesis takes a novel approach by focusing on the annual earnings of females, and in doing so introduces two methods for addressing the issues associated with zero earnings observations. First, the Inverse Hyperbolic Sine (IHS) function is introduced as an alternative to the logarithm. The IHS is defined for zero values, allowing for the creation of descriptive statistics that take into account periods of unemployment and changes in employment status. While the IHS has many properties that are useful when working with annual earnings, this thesis also highlights a number of estimation issues that can arise when using the function that have not previously been mentioned in the literature. Second, a new correction for sample selection bias that has been proposed by Semykina and Wooldridge (2013) is used to model the annual earnings of female workers. Both the sample selection bias correction and the IHS are applied to data on prime aged females from the Survey of Families, Income, and Employment (SoFIE) data set.

Contents

List of Figures

List of Tables

Acronyms

DGP Data Generating Process

ECM Error Components Model

FDIV First Differenced with Instrumental Variables

FDOLS First Differenced Ordinary Least Squares

GLS Generalised Least Squares

GMM Generalised Method of Moments

HILDA Household, Income, and Labour Dynamics in Australia

HS Hyperbolic Sine

IHS Inverse Hyperbolic Sine

IMR Inverse Mills ratio

MC Monte-Carlo

MDE Minimum Distance Estimation

MLE Maximum Likelihood Estimation

NLS Non-linear Least Squares

OLS Ordinary Least Squares

PSID Panel Survey of Income Dynamics

SoFIE Survey of Families, Income, and Employment

SNZ Statistics New Zealand

Chapter 1

Introduction

The literature on earnings dynamics has largely avoided modelling the annual earnings of female workers. Females have more frequent changes in employment status than males, and more years where they do not work. These periods of unemployment result in zero earnings, and the logarithm, a useful transformation when working with earnings, is undefined for zero values. The inability to include zero observations when using the logarithm creates two major issues when using annual earnings data. First, the forced exclusion of zero observations when working with log(earnings) can result in misleading descriptive statistics that ignore periods of unemployment and the earnings variation caused by changes in employment status. Second, if selection into or out of employment is not random, then modelling annual earnings using only the subset of individuals that work will result in biased results that do not generalise to the greater population.

The literature examining annual earnings and earnings dynamics has historically focused on the annual earnings of male workers. The trend in the literature has been to focus on changes in wage or hours worked for male workers, and changes in employment status for female workers. This is largely driven by the differences in workforce participation rates between males and females. The higher participation rates of male workers imply that non-random selection is less of an issue, and the infrequency of changes in employment status

for men makes it harder to study the participation decision. For females, the often more frequent changes in employment status have made the participation decision a primary focus, and the larger number of periods of unemployment lowers the quantity of data available and increases the risk of sample selection bias when modelling earnings dynamics.

This thesis contributes to the existing literature on earnings dynamics by exploring different ways of analysing and modelling the annual earnings of female workers, taking into account periods of non-participation, and correcting for non-random selection into the workforce. Two methods for dealing with the issues caused by the lower participation rates of female workers are proposed. First, the Inverse Hyperbolic Sine (IHS) function is considered as a potential alternative to the logarithmic transformation. The IHS has many of the advantages of the logarithm, such as reducing the impact of outliers, but is also defined for zero (as well as negative) values. This allows for the inclusion of observations where an individual does not work, enabling the creation of descriptive statistics that take into account changes in employment status and periods of unemployment. Second, a new correction for sample selection bias proposed by Semykina and Wooldridge (2013) is used to correct for the potentially non-random selection of females workers into the workforce.

The IHS function was originally introduced to the finance and economics literature by Burbidge et al. (1988) for use as an alternative to the logarithm in empirical research. The function shares many of the characteristics of the logarithm, but functions with zero values. The IHS is linear at the origin, and becomes log-like in the tails, with the speed at which it changes from linear to log-like determined by a scaling parameter θ. The IHS has been used in a range of different areas, including modelling savings and wealth, where the presence of zero or negative values restricts the use of the logarithm (Pence, 2006). The IHS has not been widely used in studies of annual earnings dynamics, with its main application being in studies of the self-employed (Bucks and Moore, 2006; Rybczynski, 2009). This thesis extends the IHS to the analysis of the annual earnings of female workers, by applying the transformation to both simulated and empirical data.

While the IHS allows for the inclusion of the zero earnings observations, it does not take into account the potential for systematic differences between the working and non-working populations. If workers non-randomly select into employment, and this selection is correlated with factors that also influence the earnings a worker receives, then using only the fraction of the sample that works will lead to biased and inconsistent regression estimates. In this thesis, a new correction for sample selection bias proposed by Semykina and Wooldridge (2013) is used to correct for the potentially non-random selection of females workers into and out of the worforce. The selection correction follows the methodology from Heckman (1979) with some extensions, and Semykina and Wooldridge's model enables its use in a dynamic panel setting. Semykina and Wooldridge's model also deals with a number of other common issues that arise when estimating dynamic earnings equations, such as the presence of unobserved, individual specific fixed effects, and avoids the weak instrument issue associated with differencing.

This thesis makes use of a combination of Monte-Carlo (MC) simulation and empirical data. The simulated data is used to evaluate the properties of the IHS as a potential transformation, highlighting its advantages and disadvantages, and explores how the expected results may change if it used. Both the IHS and Semykina and Wooldridge's sample selection bias correction are also applied to empirical data; the Survey of Families, Income, and Employment (SoFIE) longitudinal survey carried out by Statistics New Zealand. SoFIE contains data on a number of variables such as annual earnings, employment status, ethnicity, and education level, and an extract from SoFIE is used that contains data on prime aged females.

The structure of this thesis is as follows: Chapter 2 formally introduces the IHS function as a potential replacement for the logarithm. The function's properties are detailed, as well as how it has been used in the literature and how the function is estimated. MC simulation is used to generate simple data that is used to test the estimation procedure, and in doing so demonstrates that estimation of the IHS can be unreliable. Depending on the underlying

structure of the data used, the estimated parameters can be very different from those used to generate the data. The potential reasons for this unreliability, and what effect incorrect parameter estimation has on a models predictions are explored using the MC simulations, as well as how IHS estimation performs when there are censored observations such as is observed in empirical earnings data.

Chapter 3 focuses on applying the IHS to panel data of annual earnings. MC simulation is used to generate data designed to emulate the moments and participation patterns observed in empirical data. The IHS and logarithm are both applied to the simulated data to produce a range of summary statistics and auto-covariance matrices, and the variance of earnings is decomposed into its intensive and extensive components. Following this, the SoFIE data set is introduced, and the same methods previously used on the simulated data are then applied to the female sample from SoFIE. This application of the IHS demonstrates some of the strengths and weaknesses of the transformation. By including the zero observations, the IHS allows us to decompose the variance of annual earnings into its intensive and extensive components while also reducing the impact of extreme observations. On the other hand, the results depend on the value used for the scaling parameter in the IHS, and as we show in Chapter 2, estimation of this parameter is unreliable, especially when applied to censored data.

Chapter 4 then formally introduces sample selection bias, and explicitly models the annual earnings of female workers. This chapter explains the causes of sample selection bias, as well as how it has been corrected for in the literature, with a focus on the annual earnings of female workers. A number of simple models are used to model the earnings of prime aged females from SoFIE. These models do not take into account sample selection bias, but illustrate some of the standard methods that are used to model earnings. The Semykina and Wooldridge correction is then applied to the same data, and a number of strengths and weaknesses of the model emerge. The results are broadly consistent with the findings in Semykina and Wooldridge (2013), in which the annual hourly earnings of female workers

from the Panel Survey of Income Dynamics (PSID) were modelled and found that using Semykina and Wooldridge's method estimated higher levels of earnings persistence than the other models. The most striking result of the full Semykina and Wooldridge correction is the high level of earnings persistence estimated for female workers in SoFIE, indicating that the annual earnings of female workers from the SoFIE sample may be non-stationary. This differs from the results in Semykina and Wooldridge (2013), where there was no evidence of non-stationarity.

Chapter 2

The Inverse Hyperbolic Sine Function

2.1 Introduction

The logarithm is a frequently used transformation in the earnings and income literature, as it reduces the impact of extreme observations and results in a transformed distribution of earnings that is approximately normal (Lillard and Panis, 1998; Davies and Shorrocks, 2000). However, annual earnings data is characterised by periods of workforce non-participation. The logarithm is not defined for zero and negative values, forcing the researcher to decide how to deal with observations where the individual doesn't work in a data set. Approaches such as the exclusion of zero or negative observations can be problematic, as ignoring periods of zero earnings leads to a misleading understanding of earnings, and potentially results in sample selection bias and inconsistent coefficient estimators.

The Inverse Hyperbolic Sine (IHS) function is a potential alternative to the logarithmic transformation; it shares many of the logarithms properties, but is also defined for zero and negative input values. This allows for the inclusion of zero observations when using earnings data (Pence, 2006). More broadly, the IHS can be used in any case where the properties of the logarithmic transformation are desired, but the data contains negative or zero values (Bali and Theodossiou, 2008; Zhang et al., 2000).

This chapter formally introduces the IHS. The properties of the function are outlined, along with how it has been used in the existing literature. This will highlight why its ability to use zero and negative values is desirable. Estimation of the function is demonstrated using relatively simple, simulated data. This is used to highlight some potential estimation issues that can arise when the IHS function is applied, that have not been mentioned to date in the literature.

This chapter is organised as follows: Section 2.2 will introduce the IHS, defining the functional form used as well as its derivative and asymptotic properties. Section 2.3 provides a survey of the literature that makes use of the IHS. This will include research from a number of different areas, such as cheese consumption and wealth, but finds that there is very little that focuses on earnings or earnings dynamics.

Section 2.4 will cover the estimation and interpretation of a model that uses the IHS. The focus is on the estimation of θ, a scaling parameter in the IHS function which determines how quickly the function changes from linear to log-like. This section also reviews how other researchers have estimated or selected a value for θ. Section 2.5 uses Monte-Carlo (MC) simulation to evaluate the effectiveness of Maximum Likelihood Estimation (MLE) in estimating a model with the IHS. These results will highlight under which circumstances the IHS produces reliable results, and when care must be taken in applying the transformation.

Section 2.6 contains discussion of the results. It covers the potential advantages and disadvantages to applying the IHS in empirical research, as well as the estimation issues we have encountered. This section also discusses some potential solutions to the issues raised by a censored dependent variable.

2.2 Background and Motivation

The natural logarithm is a popular transformation used in the earnings literature, having many attractive properties in this area of research. However, the logarithm is undefined

for negative and zero values, and this can be problematic in the context of earnings. For instance, many individuals will have periods where they do not work and thus have zero earnings. The primary focus of this thesis is in examining the earnings dynamics of female workers, and as females have more frequent periods of workforce non-participation this issue is especially pertinent (Killingsworth and Heckman, 1986; Hyslop, 2001).

The nature of the process through which workers select into the workforce has a large impact on empirical research. If selection is random, inference based only on the individuals that work is valid, where as if selection in or out of the workforce is non-random then ignoring non-workers will lead to sample selection bias. This issue is explored in more depth in Chapters 3 and 4.

This section will introduce the IHS as a potential substitute for the logarithm. First exploring why the logarithm is a popular transformation in the earnings literature. Second, introducing the IHS and its functional form. The IHS function's properties are outlined, and we define its asymptotics.

2.2.1 The Logarithmic Transformation

The logarithmic transformation is frequently used by econometricians as it has a number of properties that make it attractive in economic research. It can be applied to both the dependent and independent variables in a regression, but here we will focus on its characteristics in a log-linear model. The transformation has a larger dampening effect as the variable to be transformed increases, so it reduces the effect of large outliers. The log transformation also allows linear estimation of otherwise non-linear relationships. For example, when individuals care about changes in their relative wealth, rather than changes in its absolute level, a linear regression of log(wealth) on the independent variables will be more appropriate than one using the untransformed level of wealth (Greene, 2003; Pence, 2006; Davidson and MacKinnon, 2004).

Earnings data is often approximately log-normally distributed, and thus transforming it

with the logarithm results in data with an approximately normal distribution (Lydall, 2013; Battistin et al., 2009; Moene and Wallerstein, 2003; Davies and Shorrocks, 2000). Use of the log also leads to models that are easy to interpret and understand, with elasticities that are simple to calculate (Benito and Hernando, 2008).

The Box-Cox transformation, which is a generalisation of the logarithm and has it as a special case, is an alternative transformation that is frequently used (Sakia, 1992). The standard functional form of the Box-Cox is, like the logarithm, only defined for strictly positive values. Burbidge et al. (1988) use a modified version of the Box-Cox that allows the function to work with negative values. This modified Box-Cox is still not defined for zero values however. This means that its use is also problematic in contexts such as earnings data, where zero observations are expected. Burbidge et al. apply both the IHS as well as the modified Box-Cox, finding that the IHS performs better in modeling net worth in their context.

2.2.2 The Inverse Hyperbolic Sine Transformation

The IHS offers a potential solution to the issue of zero value observations by preserving many of the logarithms attractive features, while also allowing the inclusion of non-positive values. The IHS function was first introduced by Johnson (1949) as a method for transforming a variable so that it becomes normally distributed. Its use in empirical economic and finance research was proposed by Burbidge et al. (1988) as a substitute for the Box-Cox transformation.

The functional form used for the IHS by Burbidge et al. can be seen in Equation (2.2.1a),

$$\text{IHS}(Y,\theta) = a\sin h^{-1}(Y\theta)/\theta = \frac{\ln(\theta Y + \sqrt{1+\theta^2 Y^2})}{\theta}\ , \tag{2.2.1a}$$

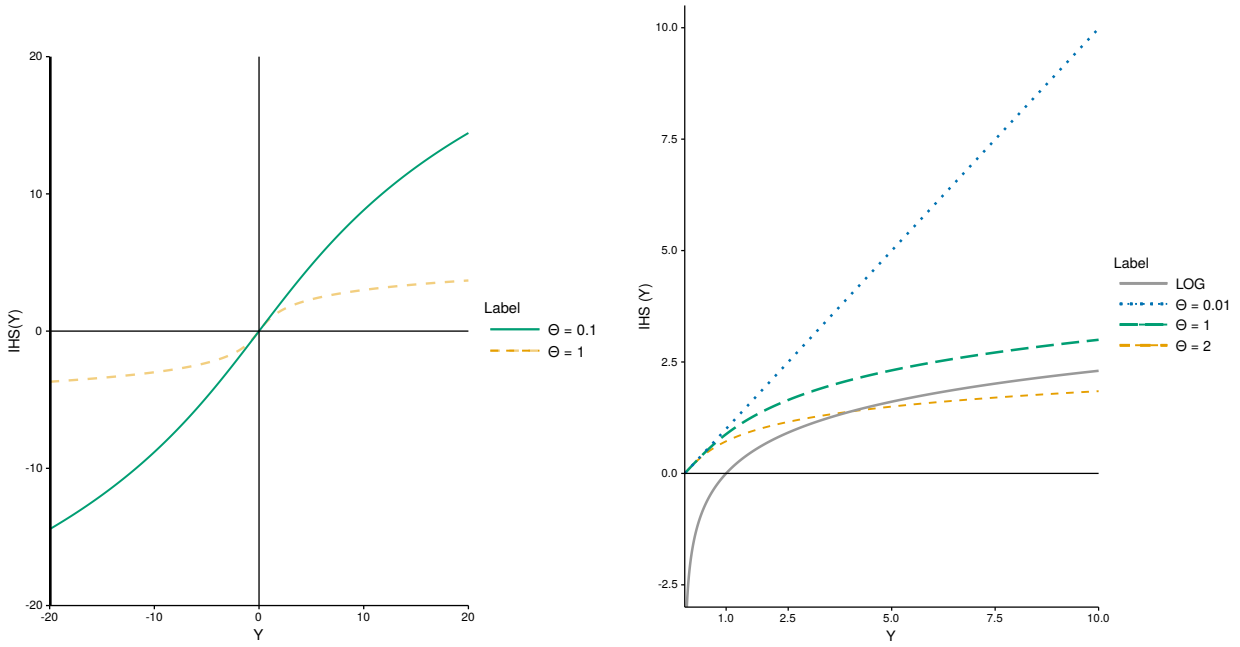
$$\frac{\partial \text{IHS}(Y,\theta)}{\partial Y} = \frac{1}{\sqrt{1+\theta^2 Y^2}}\ , \tag{2.2.1b}$$

$$\frac{\partial^2 \text{IHS}(Y,\theta)}{\partial Y^2} = \frac{-Y\theta^2}{(1+\theta^2 Y^2)^{1.5}}\ , \tag{2.2.1c}$$

where Y is the variable being transformed, and θ acts as a scaling parameter. θ can take on any value, but as the function's output is unaffected by its sign we will focus on $\theta \geq 0$. Likewise, while the IHS is defined for all real values, we will focus on $Y \geq 0$ for clarity, as the function is rotationally symmetric around the origin. Figure 2.1a illustrates this symmetry, showing the IHS with two values of θ over the domain $Y \in [-20, 20]$. Equation (2.2.1b) is the slope of the IHS, and (2.2.1c) is the second derivative with respect to Y. Unlike the logarithm, the IHS is defined for zero values, for which it will always return zero regardless of the value θ takes, ($\text{IHS}(0,\theta) = 0, \forall\theta \in \mathbb{R}$). It is a monotonically increasing function, that is defined over all real numbers and concave for $Y \geq 0$.

First we will examine the properties of the IHS with $\theta = 1$, then we will look at how the function changes with different values of θ. When $\theta = 1$, the IHS function changes from Equation (2.2.1a) to (2.2.2a), the slope becomes (2.2.2b), and the second derivative (2.2.2c). At the origin, when $Y = 0$, the output of the function will be $\ln(1) = 0$, its slope will be one, and the second derivative will be zero, so the function is linear through the origin. As Y becomes large, when $\sqrt{1+Y^2} \approx Y$, the slope will be approximately equal to $1/Y$ and the IHS function will be approximately $\ln(2Y) = \ln(2) + \ln(Y)$. Therefore, the function returns zero for an input of zero, and its output approaches the logarithm, but shifted upwards by $\ln(2)$, as Y becomes large (Pence, 2006). This vertical shift can be seen in Figure 2.1b when comparing the IHS with $\theta = 1$ to the logarithm. Equation (2.2.2c) shows that $\forall Y > 0$ the second derivative is negative and the function is concave. At the origin the slope of the IHS function is one, and the second derivative is equal to zero so the function is linear. As

Figure 2.1: *Plot of the inverse hyperbolic sine with different values of θ, and versus the logarithm*



(a) (b)

Y becomes large the function becomes an approximation of the logarithm, and the slope approaches that of the logarithm.

$$\text{IHS}(Y) = \ln(Y + \sqrt{1 + Y^2}) \quad , \tag{2.2.2a}$$

$$\frac{\partial \text{IHS}(Y)}{\partial Y} = 1/\sqrt{1 + Y^2} \quad , \tag{2.2.2b}$$

$$\frac{\partial^2 \text{IHS}(Y)}{\partial Y^2} = \frac{-Y}{(1 + Y^2)^{1.5}} \quad . \tag{2.2.2c}$$

Second, we consider the general IHS where θ is not constrained to be equal to one. The ability to have different values of θ increases the flexibility of a model. It allows individuals/observations to have different preferences depending on the level of the transformed variable. For example, if the researcher were interested in wealth, the use of the IHS would allow individuals with low levels of wealth to value changes in the absolute value of their wealth, while individuals with a higher level of wealth could value changes in their relative

wealth (Pence, 2006). Empirically, the scaling factor can be treated as a parameter to be estimated along with the rest of the model using MLE (Burbidge et al., 1988; Bucks and Moore, 2006), and the estimation procedure is detailed in Section 2.4.

Equation (2.2.1b) defines the slope of the IHS function with respect to Y. When either Y or θ are equal to zero, the slope is equal to one (although Equation (2.2.1a) is undefined for $\theta = 0$). When $Y = 0$ the second derivative, (2.2.1c), is equal to zero and therefore the function is linear. Likewise, for $\theta = 0$ the second derivative is equal to zero, so the IHS function is always linear at the origin and nests an approximation of the untransformed variable as θ approaches zero. Burbidge et al. (1988) show this by applying l'Hopital's rule. Testing if θ is significantly different from zero allows us to determine if the IHS transformation is required, or if the untransformed level of Y is appropriate. Examining Equation (2.2.1c), it is clear that $\forall\theta \in \mathbb{R}$, when $Y > 0$ the second derivative is negative, so the function is concave.

Away from the origin, the values of Y and θ jointly determine the slope of the IHS. When θ^2Y^2 is close to zero, the slope in (2.2.1b) will be close to one and the second derivative will be close to zero, so the function will be approximately linear. As such, the function will remain linear for a greater portion of the domain as θ approaches zero (Pence, 2006).

For large values of Y (when $\sqrt{\theta^2Y^2} \approx \sqrt{1+\theta^2Y^2}$), (2.2.1a) approaches $\ln(2\theta Y)/\theta = (\ln(Y)+\ln(2\theta))/\theta$. For all values of θ, when $0 < Y < 1$ the IHS of Y will be greater than the log of Y, as can be seen in Figure 2.1b. When θ is smaller than or equal to one, the IHS of Y will remain greater than $log(Y)\ \forall\ Y \in \mathbb{R}^+$. The vertical shift upwards (as seen in Figure 2.1b) increases as θ decreases, and approaches the difference between Y and the log of Y as θ approaches zero. Conversely, when θ is greater than one the IHS is initially larger than the equivalent log, but the function will cross the log exactly once and then remain below it, as seen in Figure 2.1b. For any value of Y, as θ approaches zero the output of the function approaches Y, while as θ increases the output of the function will approach zero.

For all values of θ (except zero), as Y increases the function becomes non-linear. When

$\theta^2 Y^2$ is large compared to one, so that $\sqrt{\theta^2 Y^2} \approx \sqrt{1 + \theta^2 Y^2}$, the slope will approach $1/(\theta Y)$. This is the slope of the log transformation of Y, scaled by $1/\theta$. As θ increases and becomes greater than one, the slope for large Y values becomes smaller than the slope of the log for the same Y. Likewise, as θ decreases and approaches zero, the slope for large Y values becomes steeper than that for the log transformation. In this way θ acts as a scaling factor on the slope of the function.

Thus, in summary, the IHS is always linear at the origin and approximately a logarithmic transformation in the tails, with the scaling parameter θ determining how quickly the function changes from linear to approximately logarithmic. Figure 2.1b illustrates the relationship between the IHS and logarithmic transformations by plotting them together over the domain $Y \in [0, 10]$ with $\theta = \{2, 1, 0.01\}$. With a scaling factor of 1 the IHS becomes an approximation of the logarithm very quickly, while when it is set to 0.01 the function remains approximately linear over a larger portion of the domain. When $\theta = 2$ the function is larger than the logarithm at the origin, and then crosses it once and remains below it.

2.3 Applications of the Inverse Hyperbolic Sine in the Literature

The IHS function has been used in a broad range of research areas. This includes modeling the risk of financial portfolios (Bali and Theodossiou, 2008), measuring consumption and the effect of welfare reforms (Brzozowski, 2007; Bellemare et al., 2013), measuring the effectiveness of mental health care expenditure (Zhang et al., 2000), modeling crop yields (Moss and Shonkwiler, 1993; Ramirez et al., 1994), and research focused on individual or family wealth (Pence, 2006; Carroll et al., 2003). In the majority of these cases the log transformation's properties are desired, but the presence of zero or negative values in the variable to be transformed renders it unworkable (Zhang et al., 2000; Brzozowski, 2007; Bellemare et al., 2013).

The IHS function was introduced into the empirical finance and economics literature by Burbidge et al. (1988) as a method of transforming the dependent variable in a regression. Burbidge et al. compared the Box-Cox and IHS, finding that both reduced the impact of extreme values on regression estimation, but that the IHS was defined for all real values. They applied both the IHS and Box-Cox to data on the net worth of households, finding that the IHS performed better in this case, even though there was only a small number of zero observations that had to be modified to function with the Box-Cox.

The IHS is frequently used in place of the logarithm, a special case of the Box-Cox transformation, in the literature examining individual and family wealth. The log transformation has attractive properties when examining wealth, as it helps to mitigate highly skewed data, and allows estimation of a linear model when preferences are not linear in the level of wealth (Davidson and MacKinnon, 2004). However, the logarithm's inability to handle zero and negative values is problematic, as households can have non-positive levels of wealth (Friedline et al., 2015; Gale and Pence, 2006).

The IHS can reduce the impact of outliers in the data in a similar manner to the logarithm (Robb and Burbidge, 1989; Kennickell and Sunden, 1997), with the level of dampening determined by θ. θ also determines how quickly the function changes from linear to an approximation of the logarithm. This allows the unit of observation to have a linear relationship with the dependent variable when close to the origin, and a relative relationship as the dependent variable increases (Pence, 2006; Brown et al., 2015).

The IHS transformation nests the original levels of the transformed variable for portions of the domain, determined by the value of θ (Pence, 2006). By estimating θ through MLE we can determine if transforming the dependent variable is necessary. If we do not reject the hypothesis that $\theta = 0$, then this indicates that it might be correct to use the original level of the untransformed variable as the regressor (Pence, 2006; Zhang et al., 2000; Burbidge et al., 1988).

The use of the IHS therefore allows the incorporation of non-positive values, preserves

some of the logarithm function's attractive properties, and has additional beneficial properties, such as allowing elasticities to vary with the level of the transformed variable (Gale and Pence, 2006; Bucks and Moore, 2006; Carroll et al., 2003). The IHS is most frequently used to transform the dependent variable (Browning and Crossley, 2009; MacKinnon and Magee, 1990; Robb and Burbidge, 1989), but is also used with independent variables (Bellemare et al., 2013), and to transform the residuals of a regression (Moss and Shonkwiler, 1993).

Pence (2006) was interested in the effect of US retirement 401(k) accounts on household saving, and applies the IHS function to household wealth. Pence takes advantage of the IHS function's ability to operate with zero or negative levels of wealth, and its ability to account for differences in initial wealth through the scale parameter θ. θ allows the model to include a varying relationship between household wealth and the regressors, where wealth grows in absolute levels when close to zero and the function is linear, and changes relatively as wealth increases and it approximates the logarithm (Pence, 2006).

Bloemen (2016) examines the IHS in the context of wealth in older individuals, examining the effect of wealth on job exit. Counter to some other wealth papers that make use of the IHS, Bloemen does not mention zero or negative values of wealth as a motivating factor, instead focusing on correcting for the skewness in his data.

Carroll et al. (2003) look at household wealth, and how it changes based on unemployment risk. Carroll et al. apply the IHS to the ratio of household wealth to a measure of permanent income, then use this ratio as their dependent variable. They state that this deals with the inherent skewness of the data in a similar fashion to the log, but doesn't require drop or altering households with non-positive wealth, and does not impose constant elasticities. Kapteyn and Panis (2003) are interested in the differences institutions make in wealth accumulation/saving. They apply the IHS to the wealth while examining individuals aged fifty or older in the United States, Italy, and the Netherlands. They use the IHS as wealth data is skewed, and it is approximately logarithmic while also functioning with zero or negative values. Kapteyn and Panis also point out that the IHS is not invariant to a

change in scale or units.

Friedline et al. (2015) examine the relationship between household wealth and youth's math achievement. They cite many of the same reasons as previous authors for using the IHS, specifically that it accounts for skewness in the data, allows both absolute and relative relationships between variables, and functions with non-positive values. Separating it from other wealth research that utilizes the IHS, Friedline et al. are using transformed wealth as an independent variable while looking at youth math achievement. Bellemare et al. (2013) also use the IHS for more than just transforming the dependent variable. In their research looking at the welfare impacts of commodity price volatility in Ethiopia, they use the IHS to transform wealth as an independent variable, and also apply it to commodity prices and each family's marketable surplus of goods, independent variables in the regression.

The IHS is also used outside the income and wealth literature to reduce the effect of highly skewed data. Brown et al. (2015) compensate for the non-normality of donations data while also allowing for the presence of zero values by using the IHS. They also include distinct sub-populations with different group-specific preferences by estimating multiple values of θ. Bali and Theodossiou (2008) employ the IHS in modeling risk and portfolio returns, and conclude that it does a comparable job to the extreme value distributions in modeling skewed and long-tailed distributions. Zhang et al. (2000) use the IHS in analyzing the change in health care costs and lost wages for individuals that either receive or do not receive mental health care. They take advantage of the symmetry of the IHS around zero, so gains and losses are valued equally, and its ability to reduce the impact of outliers. Zhang et al. also warn that while the IHS is defined for zero values, care should be taken if zero is a special case. In those situations they state a two-part model may be more appropriate.

Benito and Hernando (2008) estimated an employment equation for temporary contracts using data on firms in Spain. The data includes many firm observations with no temporary contracts, making the use of the logarithm of temporary unemployment problematic. They estimate their model using the IHS of temporary unemployment as well as its logarithm,

but state that they prefer the log as it is more familiar, easier to derive elasticities, and can be derived theoretically. Muhumuza (2012) applies the IHS to hours worked for children in rural Uganda, and says that θ controls for kurtosis in the model, but does not specify the value used or how it was estimated.

While a broad range of topics have papers that utilise the IHS, it is relatively unexplored in the context of income and earnings. When the IHS has been used with earnings data, it has been in research looking at the self employed, a demographic which has observations with negative earnings (Bucks and Moore, 2006). Various papers have used the IHS in the context of wealth, but the literature appears to lack an exploration of its use with earnings cross-sectional or panel data. In this context the logarithm has many attractive properties, but can introduce numerous issues due to periods of workforce non-participation.

2.4 Estimation of the Inverse Hyperbolic Sine

Estimating a model in which the dependent variable is transformed with the IHS requires a value for the scaling parameter θ. MLE is the main method used in estimating θ, although a number of researchers select a value manually. There are different formulations of the log-likelihood function in the literature, based on different assumptions about the underlying data, but this section focuses on the method suggested by Burbidge et al. (1988). The estimation procedure assumes that there is a non-linear relationship between Y_i and X_i, but that it can be linearized by applying the IHS to Y_i. First the MLE method proposed by Burbidge et al. is explained. Second the model and coefficient interpretation are compared to the standard log model, and lastly we will look at how other researchers have selected or estimated θ.

2.4.1 Estimating θ

Equation (2.4.1) outlines the model used in estimating the IHS, where Y_i is the dependent variable, X_i is a matrix of observed regressors, and ϵ_i is an unobserved error term. In order to estimate θ, and from there the other parameters of interest, we assume that the IHS transformed Y has a linear relationship with the observed regressors. This assumption, combined with the additional assumption that the unobserved error term has a normal distribution, allows estimation of the model,

$$\text{IHS}(Y_i, \theta) = X\beta + \epsilon_i \ , \tag{2.4.1a}$$

$$\epsilon_i \sim N(0, \sigma^2 I) \ . \tag{2.4.1b}$$

Based on the model in Equation (2.4.1) Burbidge et al. suggest the scaling parameter θ can be estimated along with the other parameters of interest using MLE on the log likelihood function, seen in (2.4.2). In this case the slope coefficients β, θ, and the variance-covariance matrix are estimated simultaneously. Y_i is the dependent variable being transformed with the IHS, σ^2 is the variance-covariance matrix, β is a vector of slope coefficients and the intercept, and X is a matrix of observed regressors.

$$\begin{aligned}\mathcal{L}(\theta, \beta, \sigma^2) =& C - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\Big(\text{IHS}(Y, \theta) - X\beta\Big)'\Big(\text{IHS}(Y, \theta) - X\beta\Big) - \\ & \frac{1}{2}\sum \ln(1 + Y_i^2\theta^2).\end{aligned} \tag{2.4.2}$$

Alternatively, a two step procedure can be used where θ is estimated separately by maximising the concentrated log likelihood function, (2.4.3a), and the resulting $\hat{\theta}$ can be used in estimating the full model through either MLE or Ordinary Least Squares (OLS). M is defined as $I - X(X'X)^{-1}X'$, where I is the identity matrix. In both (2.4.2) and (2.4.3a) C is a constant, and we can restrict our search to $\theta > 0$ due to the symmetry of the function

$(\text{IHS}(Y,\theta) = \text{IHS}(Y,-\theta))$.

$$\mathcal{L}^c(\theta) = C - \frac{n}{2}\ln(\text{IHS}(Y,\theta)'M\ \text{IHS}(Y,\theta)) - \frac{1}{2}\sum \ln(1+Y_i^2\theta^2). \tag{2.4.3a}$$

$$M = I - X(X'X)^{-1}X' \tag{2.4.3b}$$

2.4.2 Model Interpretation

The coefficients in a IHS-linear model are not as easy to interpret as the coefficients in either a linear-linear or log-linear model. In the standard log-linear regression model the slope coefficients can be interpreted as approximately the percentage change in Y from a unit increase in X, while in a linear-linear model the coefficients represent the unit change in Y for a unit change in X. In a regression model where the dependent variable is transformed with the IHS, coefficient interpretation varies from linear to relative, and requires knowing which portion of the domain a particular observation is in.

As explained in Section 2.2.2 the IHS is linear at the origin and becomes approximately logarithmic away from the origin. Therefore, coefficient interpretation requires the knowledge of where the function is linear and where it is approximately logarithmic. For observations that lie in the linear portion the slope coefficients can be interpreted in the same way as in the linear-linear model. For observations that lie in the approximately logarithmic portion of the function, $\theta\beta$ has the same interpretation as β does in a log-linear model, being approximately equal to the percentage change in Y for a unit change in X. We can see why this relationship holds in Equation 2.4.4a. This relies on θY being large relative to 1, so that $\sqrt{\theta^2 Y^2} \approx$

$\sqrt{1+\theta^2Y^2}$, which occurs in the approximately logarithmic portion of the function.

$$\text{IHS}(Y) = \alpha + X\beta + \epsilon \ , \tag{2.4.4a}$$

$$\frac{\ln(\theta Y + \sqrt{1+\theta^2Y^2})}{\theta} = \alpha + X\beta + \epsilon \ , \tag{2.4.4b}$$

$$\approx \frac{\ln(2\theta Y)}{\theta} = \alpha + X\beta + \epsilon \ , \tag{2.4.4c}$$

$$\ln(2\theta) + \ln(Y) = \alpha\theta + X\beta\theta + \epsilon\theta \ , \tag{2.4.4d}$$

$$\ln(Y) = -\ln(2\theta) + \alpha\theta + X\beta\theta + \epsilon\theta \ . \tag{2.4.4e}$$

As the coefficients themselves do not have a simple interpretation, unless you know if a particular point lies in the approximately logarithmic or linear regions, a model produced using the IHS can be analyzed by examining the marginal effects and elasticities. This relies on an estimate for θ, an issue we will explore in depth in Section 2.5.

The coefficients can be transformed to give the marginal effect of a change in one of the regressors on the untransformed Y. To do this we use the inverse of the IHS, the Hyperbolic Sine (HS) function. The HS, seen in Figure 2.2, is similar to the IHS in that it is linear at the origin and the variable θ acts as a scaling parameter. In the tails it becomes similar to the exponential function, with the speed at which it changes from linear to exponential being determined by θ. The functional form of the HS can be seen in Equation 2.4.5,

$$\text{HS}(Y,\theta) = \frac{1}{2\theta}\big(\exp^{\theta Y} - \exp^{-\theta Y}\big) \ . \tag{2.4.5}$$

The marginal effect of a change in a regressor on the untransformed Y is not constant, as the relationship between X and Y is non-linear. To retrieve the marginal effect of a change in a particular regressor, say X_k, we apply the transformation in equation 2.4.6a. $\frac{\partial \text{IHS}(Y)}{X_k}$ is the marginal effect of X_k on IHS(Y), which is constant due to the linear relationship between the two variables. This marginal effect is equal to the β_k estimated with the linear model

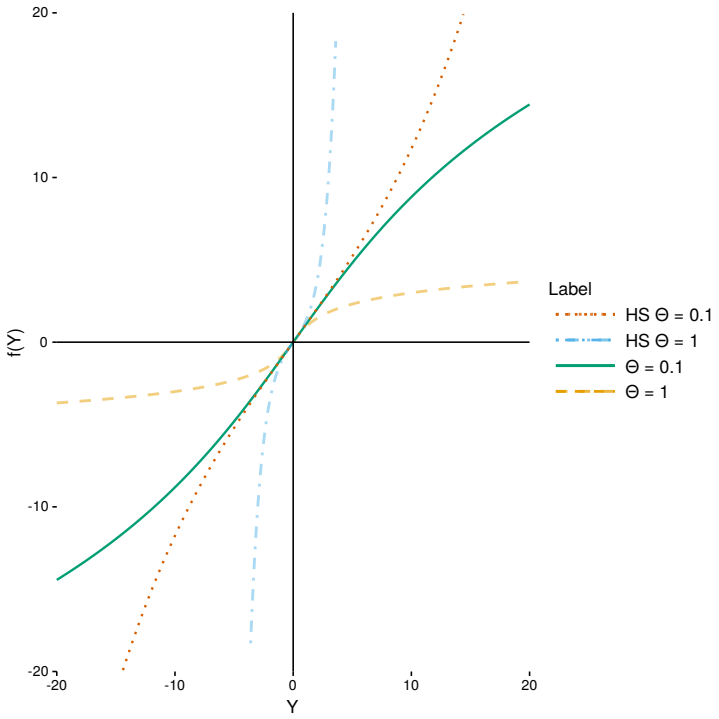


Figure 2.2: *Plot of the inverse hyperbolic sine and hyperbolic sine functions with* $\theta = \{1, 0.1\}$.

$\text{IHS}(Y) = X\beta + \epsilon$, and gives us the second component of the right side of Equation (2.4.6a). Y can also be written as $\text{HS}(\text{IHS}(Y))$, and so the first component of Equation (2.4.6a) is the derivative of the HS with respect to $\text{IHS}(Y, \theta)$. Intuitively, this transformation is taking the marginal effect of X_k on the $\text{IHS}(Y, \theta)$ and backing out the marginal effect on Y at a particular level of Y. This is required due to the non-linear effect X has on Y.

$$\frac{\partial Y}{\partial X_k} = \frac{\partial Y}{\partial \text{IHS}(Y, \theta)} \frac{\partial \text{IHS}(Y, \theta)}{X_k} \ , \tag{2.4.6a}$$

$$\frac{\partial Y}{\partial X_k} = \frac{\partial Y}{\partial \text{IHS}(Y, \theta)} \beta_k \ , \tag{2.4.6b}$$

$$\frac{\partial Y}{\partial X_k} = \frac{1}{2}\big(\exp^{\theta \text{IHS}(Y,\theta)} + \exp^{-\theta \text{IHS}(Y,\theta)} \big) \beta_k \ . \tag{2.4.6c}$$

Once we have calculated the marginal effects over all the observations in the data, it is straight forward to find the elasticity of Y with respect to any regressor. The elasticities are $e_k = \frac{\partial Y}{\partial X_k} \frac{X_k}{Y}$, where e_k is the elasticity of Y with respect to changes in X_k.

As discussed in Section 2.3, the IHS has been used in a range of scientific and social science fields, but its use in modeling earnings has been limited. The next section we will examine how the IHS/θ has been estimated or selected by other researchers.

2.4.3 Estimation in the Literature

There is no consistent approach to estimating models using the IHS in the literature. An interesting aspect of this is the estimation/specification of θ. There is a split, with some papers estimating it along with the other parameters (frequently using MLE), some assuming it is equal to one, and others assuming it is equal to some other number. An important point is that the estimation issues that we detail in the next section have not been explored in the current literature.

Bellemare et al. (2013) assume $\theta = 1$, basing this on previous literature, and claim that this specification is robust to different parameter values. Browning and Crossley (2009) also assume $\theta = 1$, and they state that their preliminary investigation suggests that their results are insensitive to this choice. Brzozowski (2007) follows Browning and Crossley in assuming that θ is equal to one. Brown et al. (2015) estimate θ using MLE, and additionally allow it to vary over different groups in the data.

Bucks and Moore (2006) looked at earnings data for self-employed workers, applying the IHS to both annual and hourly earnings. Bucks and Moore used $\theta = 0.0001$ for the annual, and $\theta = 0.2$ for hourly earnings, basing the first estimate on both the results of previous papers (Burbidge et al., 1988; Pence, 2006; Kennickell and Sunden, 1997), and their own MLE, and the second estimate based solely on their own estimation. Benito and Hernando (2008) assume that $\theta = 1$, and provide no justification for this assumption. Bloemen (2016) estimates θ using MLE to be 0.733. Bloemen also mentions that θ is often assumed to be equal to one in the literature, and that this is potentially not justified. Friedline et al. (2015) state values of θ that have been used previously in the literature, and show the distribution of wealth when transformed with those values, but choose to use $\theta = 1$ without any estimation,

as they claim it better suits their data.

Some authors such as Grabka et al. (2015) and Haushofer and Shapiro (2013) do not specify the value they use for θ, although in the case of Haushofer and Shapiro it is implicitly equal to one. Interestingly, both of these papers cite Pence (2006) and Burbidge et al. (1988), both of which explicitly estimate θ and find it much smaller than one. Zhang et al. (2000) use MLE to estimate their model, and find $\theta = 0.009$. Pence (2006) estimates θ using MLE, but assumes a LaPlace distribution for the error term, leading to a different log-likelihood function. Pence estimated $\theta = 0.0003$, and makes some comparison between the results based on this value, and the results they get with different θ values.

All the values of θ mentioned so far, whether estimated or chosen, have been between 0 and 1, but this is not a requirement. Yen and Jones (1997), in their work on household consumption of cheese, estimate θ to be just over one. Carroll et al. (2003) apply the IHS to wealth and estimate $\theta = 3.76$ using MLE, which is much larger than other values estimated or selected in the literature we have seen.

2.5 Estimation with Simulated Data

Section 2.4 introduced the MLE method proposed by Burbidge et al. (1988) for estimating a model with the IHS. In order to evaluate this estimator, a simple model where the dependent variable is transformed with the IHS is generated using MC simulation. The generated data is used to estimate the model's underlying parameters and elasticities, and to compare how different Data Generating Processes (DGPs) and values of θ affect estimation. Equations (2.5.1a) and (2.5.1b) outline the structure of the DGP,

$$IHS(Y_n) = \alpha + X_n'\beta + \epsilon_n \ , \tag{2.5.1a}$$

$$Y_n = HS(\alpha + X_n'\beta + \epsilon_n, \theta) \ , \tag{2.5.1b}$$

$$\epsilon_n \sim N(0, \sigma^2) \ , \tag{2.5.1c}$$

where IHS(Y_n) is simulated and then transformed using the HS to produce the "observed" Y_n. In each simulation cross-sectional data is generated with N observations. The variables and parameters take the values or distributions outlined in Table 2.1. When more than one value is listed, multiple simulations were conducted with each value/combination of values.

Equation (2.5.1a) models the underlying linear relationship between IHS(Y_n) and X_n. The intercept and slope coefficients, α and β respectively, are fixed over observations. An observation specific variable, X_n, is generated that is "observed" and used in estimation, as well as an idiosyncratic unobserved error term, ϵ_n. If both IHS(Y_n) and X_n were observed, α and β could be estimated using a linear model. However, IHS(Y_n) is not directly observed. As part of the DGP, IHS(Y_n) is transformed using the HS function to generate the observed variable, Y_n. If this does not have a linear relationship with X_n then the parameters cannot be estimated using a linear model. With correct estimation of θ the observed Y_n can be transformed using IHS(Y, θ), and used to estimate the parameters of the model.

Table 2.1: *Simulation parameters*

Variable	Value or distribution
Intercept	$\alpha = \{1, 5\}$
Slope	$\beta = 1$
Scaling parameter	$\theta = \{1, 0.5, 0.1, 0.05, 0.01\}$
Observed variable	$X \sim U[1, 10]$
Error	$\epsilon \sim N(0, 1)$
Sample size	$N = \{250, 1000, 5000\}$
Simulation reps	$R = 500$

Across all simulations the error term ϵ_n has a standard normal distribution, and the observed regressor X_n is distributed uniformly between one and ten. Two alternative values are used for the intercept, $\alpha = 1$ and $\alpha = 5$. The slope coefficient on X, β, is equal to one in all simulations. Three sample sizes were used, setting N equal to $250, 1000$, and 5000. This allows the small and large sample properties of the estimator to be explored. In generating the observed Y the HS is applied with four different values, $\theta = \{1, 0.1, 0.05, 0.01\}$.

Table 2.2 outlines some summary statistics of the simulated dependent variable. This is the dependent variable that is 'observed', and used to estimate θ. Untransformed indicates the data has not been transformed with the HS, and outlines the distribution pre-transformation. In each other case the dependent variable has been transformed with the HS and the value of θ shown. The mean value over the 500 simulations is shown for the minimum value, the maximum value, the level of skewness, and the mean of the simulated dependent variable. The value of θ used has a large impact on the observed dependent variable. The larger values of θ lead to mean and maximum values that are much larger than the untransformed distribution, and significantly more skewed. The impact of changing the intercept, α, is also much greater with larger values of θ. As θ becomes smaller, the observed dependent variable approaches the underlying, untransformed distribution. For example, with $\theta = 0.01$ the observed distribution is very close to that of the underlying DGP.

Table 2.2: *Simulation Summary Statistics*

	Mean(Y)	Min(Y)	Max(Y)	Skewness
Untransformed: $\alpha = 1$	6.50 (0.09)	-0.17 (0.43)	13.16 (0.42)	0 (0.05)
Untransformed: $\alpha = 5$	10.50 (0.09)	3.86 (0.45)	17.17 (0.45)	0 (0.05)
HS$(\theta = 1), \alpha = 1$	$5,480$ (544)	-0.15 (0.50)	$283,231$ $(153,197)$	8.20 (3.14)
HS$(\theta = 1), \alpha = 5$	$298,773$ $(33,320)$	25.62 (10.37)	$15,650,169$ $(8,882,850)$	8.33 (3.41)
HS$(\theta = 0.5), \alpha = 1$	60.89 (2.74)	-0.19 (0.48)	741.30 (187.87)	2.82 (0.55)
HS$(\theta = 0.5), \alpha = 5$	448.98 (20.31)	6.76 (1.47)	$5,501.89$ $(1,413.97)$	2.84 (0.54)
HS$(\theta = 0.1), \alpha = 1$	7.24 (0.11)	-0.18 (0.44)	17.32 (0.85)	0.27 (0.05)
HS$(\theta = 0.1), \alpha = 5$	13.03 (0.15)	3.95 (0.44)	26.81 (1.18)	0.36 (0.05)
HS$(\theta = 0.05), \alpha = 1$	6.67 (0.09)	-0.16 (0.42)	14.13 (0.50)	0.08 (0.05)
HS$(\theta = 0.05), \alpha = 5$	11.10 (0.10)	3.85 (0.41)	19.29 (0.57)	0.11 (0.05)
HS$(\theta = 0.01), \alpha = 1$	6.52 (0.09)	-0.18 (0.43)	13.18 (0.44)	0 (0.05)
HS$(\theta = 0.01), \alpha = 5$	10.52 (0.08)	3.81 (0.44)	17.23 (0.44)	0 (0.05)

Notes: Each value is the mean over the 500 simulations for each DGP

Estimation follows the two-step procedure outlined in Section 2.4. When estimating θ in the first step, we have used the R function *optimise* to minimise the negative of the log-likelihood. As $\hat{\theta}$ can be constrained to be non-negative we have exponentialised it in the concentrated log-likelihood, replacing θ with $\exp(\mu)$ where $\mu = \ln(\theta)$. The function *optimse* minimises or maximises over a single dimension, and requires a specified domain to search over. For each simulation we have specified the domain to be $\mu \in [\ln(\theta) - 5, \ln(\theta) + 5]$, as this provides a broad range of μ values around the true value, while constraining it to be positive. In Section 2.5.1 we show that our results are robust to estimation based on different optimisation methods, including unconstrained optimisation, both within R and using Matlab.

As stated previously, our main interest is in applying the IHS to female earnings data. Ideally, the simulations would all have sample moments based on those observed in empirical earnings data. In that case, X_n could be thought of as some measure of individual n's human capital and Y_n as their annual earnings. Creating Y_n that is distributed in a manner similar to empirical earnings is problematic in this case. This is an issue because we fix the underlying DGP used to generate $\text{IHS}(Y, \theta)$, and then transform it with the HS and various values of θ to produce the observed Y. We want to understand how, holding the DGP constant, different values of θ affect our estimation. The value of θ used has a large impact on the data created in this way, so given the DGP of a particular simulation, Y_n with $\theta = 1$ will look very different to Y_n with $\theta = 0.1$. While this means our generated data will not always be an accurate representation of empirical earnings, it does allow us to explore the general features and estimation issues of models that use the IHS. In Chapter 3 we focus on generating panel data that closely resembles empirical earnings data, and investigate how the use of the IHS versus the logarithm affects the results.

While we can't consistently create data that mimics those observed empirically due to the effect of the HS transformation on the observed variable, there are still advantages in creating data that has similar features to empirical earnings. This is especially true when

those features are at odds with the assumptions required in estimating a model that uses the IHS. A common feature of earnings data, especially with regard to female workers, is periods of workforce non-participation where earnings are zero. This is one of the motivating factors that makes the IHS an attractive replacement to the logarithm, but also seems likely to violate one of the core assumptions made in the MLE by Burbidge et al.. As explained in Section 2.4, model estimation relies on the assumption that the errors are normally distributed after the dependent variable is transformed with the IHS. If there are many observations where the observed earnings are zero, then this mass of data seems likely to lead to errors that are not normally distributed. In addition to the simulations outlined above, simulations are performed with the dependent variable randomly censored and set equal to zero. In empirical research, such as when working with earnings data, it is likely that this censoring will be non-random, an issue we cover in more detail in Chapters 3 and 4. Introducing random censoring allows us to test a best case scenario, and to evaluate how well the MLE performs when the normality assumption is breached.

In each simulation the model will be estimated using the two-step estimator proposed by Burbidge et al. (1988). First estimating $\hat{\theta}$ using the concentrated log-likelihood, and second using OLS to estimate $\text{IHS}(Y_n, \hat{\theta}) = \alpha + X_n'\beta + \epsilon_n$. As outlined in Section 2.4.2, model interpretation is not straightforward when applying the IHS. Interpretation of the slope coefficient depends on both θ and the level of the variable being transformed. In order to evaluate how well the estimated model is performing, the estimated parameters will be used to find the elasticity of Y with respect to a change in X at each observation, as well as points above and below the minimum and maximum points of the observed data. These results will be compared to those achieved using the true θ, as well as the other values used in the DGP's of the other simulations. This allows the evaluation of how well the estimated model matches the true model when applied to the data used in estimation, as well as how they compare when applied to data outside the range used in the estimation process. Comparing the results of the model estimated with $\hat{\theta}$ to the true model, and also models estimated with

the other values of θ, will illustrate how much of an issue misestimation of θ is likely to be. We also check the region around each estimated $\hat{\theta}$ with the concentrated log-likelihood to see if it is a local maximum. If it is not, that indicates that $\hat{\theta}$ is not the value that maximises the concentrated log-likelihood, and that the maximisation process has stopped for some other reason.

2.5.1 Simulation Results

This section reports the results of the simulations performed to test the performance of the IHS estimator. The results from the simulations are briefly summarised in Table 2.3. The simulations generated a large quantity of output, so the following sections focus on a subset of these simulations. The results of the simulations with $\theta = 1$ and $\theta = 0.01$ are explored in detail, while the other simulations are summarised. Tables of results for those simulations not covered in detail in this section are included in Appendix A.

Table 2.3: *Simulation Results Summary*

		$\alpha = 1$			$\alpha = 5$		
	N	250	1000	5000	250	1000	5000
$P(\text{Converge} \mid \theta = 1)$		0.552	0.854	1	0.164	0.240	0.268
$P(\text{Converge} \mid \theta = 0.5)$		0.944	1	1	0.386	0.478	0.534
$P(\text{Converge} \mid \theta = 0.1)$		1	1	1	1	1	1
$P(\text{Converge} \mid \theta = 0.05)$		0.860	0.994	1	0.966	0.998	1
$P(\text{Converge} \mid \theta = 0.01)$		0.500	0.506	0.584	0.492	0.530	0.644
$P(\text{Converge} \mid Censored, \theta = 0.1)$		0.006	0	0	0	0	0
$\bar{\hat{\theta}} \mid \theta = 1, \text{converged}$		1.077	1.225	1.039	0.028	0.060	0.106
$\bar{\hat{\theta}} \mid \theta = 0.5, \text{converged}$		0.647	0.516	0.505	0.182	0.271	0.419
$\bar{\hat{\theta}} \mid \theta = 0.1, \text{converged}$		0.102	0.100	0.100	0.103	0.100	0.100
$\bar{\hat{\theta}} \mid \theta = 0.05, \text{converged}$		0.054	0.050	0.050	0.051	0.050	0.050
$\bar{\hat{\theta}} \mid \theta = 0.01, \text{converged}$		0.038	0.027	0.018	0.033	0.022	0.015
$\bar{\hat{\theta}} \mid Censored, \theta = 0.1, \text{converged}$		0.020	-	-	-	-	-

NOTES: Probability of convergence based on fraction of simulations which reached a local maximum.

The values reported in P(Converge) are the fraction of the simulations with the specified DGP that result in a local maximum in the concentrated log-likelihood. When a local

maximum is not reached, the optimiser has halted due to reaching one of the boundary conditions. The results in Table 2.3 indicate that larger samples are more likely to successfully converge, but that even with a relatively large sample convergence is not guaranteed with some combinations of θ and α. If $\hat{\theta}$ does result in a local maximum in the concentrated log-likelihood, the estimated value is reported in $\bar{\hat{\theta}} \mid$ converged. Again, as the sample size increases the mean estimates become closer to the true value, but even with N=5000 there are relatively big differences in some cases. From Table 2.3, it appears that random censoring of the dependent variable results in an estimator that almost never converges successfully. The following sections explore the results from Table 2.3 in more detail, focusing on particular values of θ and why estimation is not always reliable.

2.5.2 $\theta = 1$

Table 2.4 summarises the results of the simulations with $\theta = 1$. *MLE max* indicates the fraction of simulations in which $\hat{\theta}$ results in a local maximum in the concentrated log-likelihood, and $\bar{\hat{\theta}}_{\text{MLE}}$ reports the mean estimate for θ conditional on the concentrated log-likelihood being maximised. The rows with $\%\Delta$ show the mean percentage difference between the elasticity of Y with respect to X, estimated with $\hat{\theta}$ versus the true θ value at selected points in the data.

,

When $\alpha = 1$ the results seem to indicate that a large sample is required for accurate estimation of θ. With $N = 250$, $\hat{\theta}_{\text{MLE}}$ resulted in a local maximum of the concentrated log-likelihood in only 55.2% of the simulations. When it was at a local maximum the mean estimate is $\bar{\hat{\theta}} = 1.077$, with a standard error of 0.042. In the remaining simulations where the concentrated log-likelihood is not maximised, the estimated $\hat{\theta}$ values approach the upper boundary of 148.4 before the estimation halts. This can be seen in Figure 2.3, where the distribution of $\hat{\theta}$ is shown both for all values, and the subset that are local maximums.

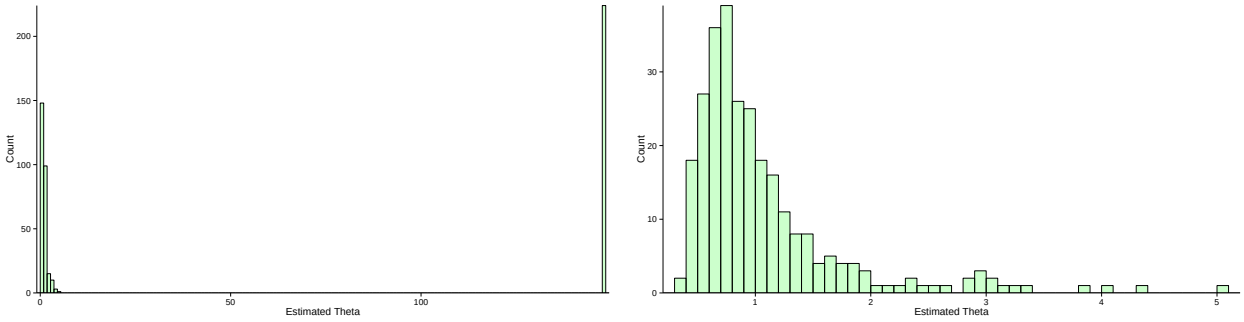
As the sample size increases, the standard error of $\bar{\hat{\theta}}$ decreases, and the proportion of

Table 2.4: *Summary of simulations with* $\theta = 1$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max[1]	0.552	0.854	1	0.164	0.240	0.268
$\bar{\hat{\theta}}_{\text{MLE}}$[2]	1.077 (0.042)	1.225 (0.033)	1.039 (0.009)	0.028 (0.0014)	0.060 (0.0047)	0.106 (0.0073)
%Δ: median[3]	0 (0.013)	0 (0.009)	0 (0.004)	−0.1 (0.004)	0 (0.004)	0 (0.000)
%Δ: 10th percentile[3]	-0.1 (0.036)	-0.1 (0.018)	0 (0.004)	0 (0.004)	0 (0.000)	0 (0.000)
%Δ: 90th percentile[3]	0 (0.013)	0 (0.009)	0 (0.004)	−0.1 (0.004)	0 (0.004)	0 (0.000)
%Δ: below min[3]	−2.7 (0.787)	−2.3 (0.456)	0 (0.197)	2.0 (0.277)	0.9 (0.098)	0.3 (0.031)
%Δ: above max[3]	0 (0.013)	0 (0.009)	0 (0.004)	−0.1 (0.004)	0 (0.004)	0 (0.000)

[1] *Fraction of simulations resulting in maximised concentrated log-likelihood*
[2] *Mean* $\hat{\theta}$ *based on simulations where concentrated log-likehood is maximised.*
[3] *% difference between elasticities calcualted using true model versus the estimated model.*

simulations that result in a local maximum of the concentrated log-likelihood function increases. With $N = 1000$ the estimated $\hat{\theta}$ value results in a local maximum in 85.4% of simulations, while with $N = 5000$ this occurs in all simulations. The standard error of the mean estimate, conditional on it being a local maximum, decreases slightly to 0.033 with $N = 1000$, and then to 0.009 with $N = 5000$.



(a) (b)

Figure 2.3: *Histogram of* $\hat{\theta}$*. Data based on simulations with* $\theta = 1, \quad \alpha = 1$ *and* $N = 250$*. Left panel is all estimates, right panel is limited to estimates where concentrated log-likelihood is maximised.*

The second panel of Table 2.4 shows the results of the simulations with $\theta = 1$ and $\alpha = 5$. In this case the MLE appears to be performing worse than when $\alpha = 1$. For each sample

size used, the model is successfully converging in fewer simulations. With $N = 250$, $\hat{\theta}$ results in a local maximum in the concentrated log-likelihood in only 16.4% of the simulations, and even with $N = 5000$ convergence only occurs 26.8% of the time. When the model does successfully converge to a local maximum, our mean estimate for $\hat{\theta}$ is lower than in the $\alpha = 1$ cases, and doesn't seem to approach the true value even when $N = 5000$. Examining the results in more depth reveals that when the estimate is not a local maximum it is again stopping at or near the upper boundary set when running *optimise*. Contrary to the first panel of Table 2.4, when $\hat{\theta}$ is a local maximum it is not distributed around the true value of one. The estimated values are instead very low, none larger than 0.072 when $N = 250$, 0.41 for $N = 1000$, and 0.51 for $N = 5000$, and the standard errors increase with N.

With all three sample sizes and $\alpha = 1$, the mean estimated value of θ conditional on the successful convergence of the MLE seems relatively close to the true value, although the standard error is quite large with $N = 250$ and $N = 1000$. On the other hand, only in the $N = 5000$ case does the MLE successfully converge in every simulation. For both $\alpha = 1$ and $\alpha = 5$, when the MLE does not converge, it instead estimates a value for $\hat{\theta}$ very close to the upper boundary specified in the optimiser. When these large values of θ are used to transform Y and the model is estimated with OLS, it results in $\hat{\beta}$ estimates much lower than the true value.

As detailed in Section 2.4.2, model interpretation is not straightforward when using the IHS, and different $\hat{\beta}$ estimates cannot be directly compared without taking into account the value of $\hat{\theta}$. Transforming $\hat{\beta}$ by multiplying it by $\hat{\theta}$ produces a value that, when in the approximately logarithmic portion of the IHS, has the same interpretation as the slope coefficient in a log-linear model. Figure 2.4 shows the result of applying this transformation to the simulations generated with $N = 250, \alpha = 1$. This includes the simulations in which the estimated $\hat{\theta}$ did not result in a local maximum in the concentrated log-likelihood function, and the value estimated for θ was close to the upper boundary of 148.4. This distribution of values is very tightly grouped around the true value of one, indicating that even with a very

large value for $\hat{\theta}$, the estimated model will make very similar predictions about the marginal effect when in the approximately logarithmic region of the IHS. A very similar distribution is observed when applying the same transformation to the data generated with $N = 1000$, and when $\alpha = 5$.

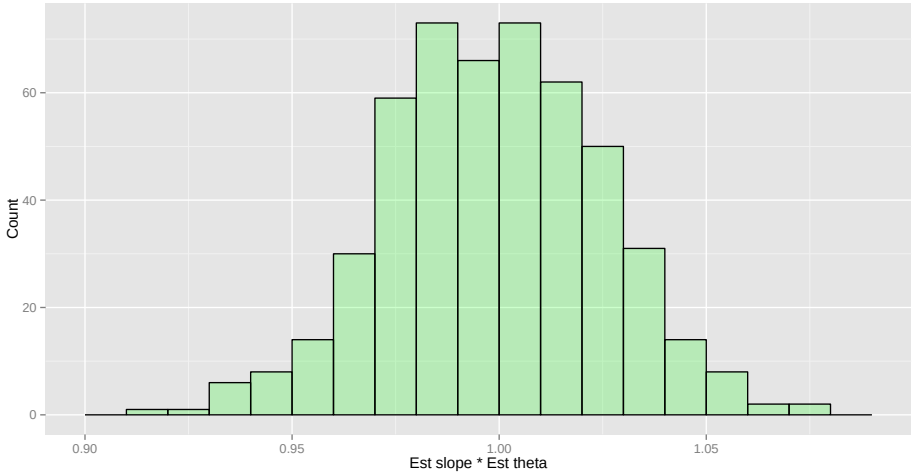


Figure 2.4: *Histogram of* $\hat{\beta}\hat{\theta}$ *for* $N = 250, \quad \theta = 1,$ *and* $\alpha = 1$ *simulations.*

These results seem to indicate that incorrect estimation of θ does not impact the model's predictions, at least with this DGP and in the portion of the function which is approximately logarithmic. This raises three questions: 1) are there any areas where this mis-estimation matters, 2) would choosing a value for θ rather than estimating it produce a model that makes the same predictions, 3) will the results be the same with a different DGP.

Table 2.4 also lists the percentage difference between the elasticities of Y with respect to X calculated with the estimated value $\hat{\theta}$, and those calculated using the true value. Comparing the elasticities provides a way to compare the predictions made by models based on different values of θ. With $\alpha = 1$, for all three sample sizes, the percentage difference is very small except when looking below the range of data used in estimation, and it is still small here for $N = 5000$. This reinforces the results illustrated in Figure 2.4, that the marginal effect in the approximately logarithmic portion of the function predicted by the models are approximately normally distributed around the true value, and it is only the lower value

that is potentially not in the log-like portion of the functions domain that produces different results.

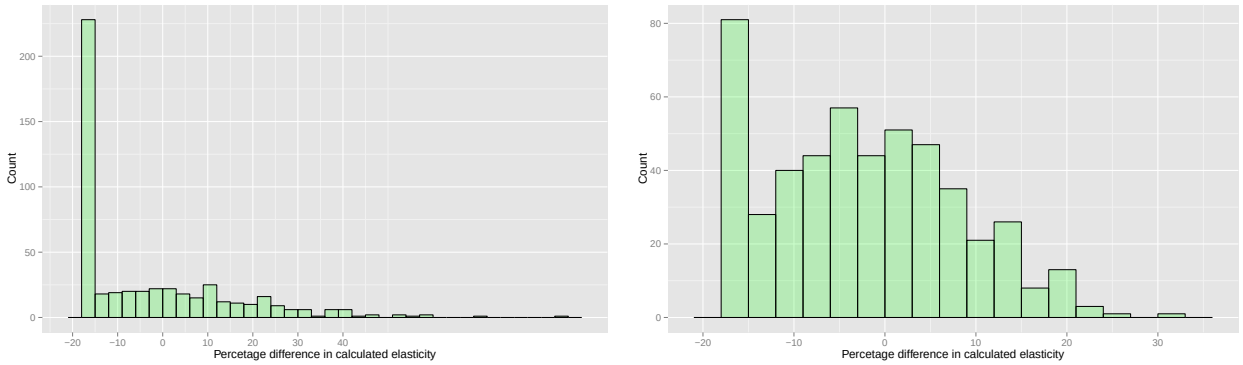
%Δ: below min in Table 2.4 indicates a mean difference of -2.7% in the elasticity calculated at a synthetic observation located one fifth of the distance from zero to the smallest observation for $N = 250$, and -2.3% for $N = 1000$ when $\alpha = 1$. This is an observation that did not exist in the data used to estimate the model, but is based on the same underlying DGP except shifted closer to the origin. These percentage differences in estimated elasticity can be see in Figure 2.5. In each case there is a large cluster of observations, more than half of the 500 repetitions for $N = 250$, where $\hat{\theta}$ estimates an elasticity between 13% and 16% lower than that found using the true value of θ. All the $\hat{\theta}$ estimates that stopped near the upper boundary are in this group for both N=250 and N=1000, with estimated elasticity approximately 16% lower than those estimated with the true θ value. This explains why there is no difference between the elasticities calculated with $\hat{\theta}$ versus the true value when $N = 5000$. The concentrated log-likelihood is maximised in every simulation in that case, so the large values near the upper boundary are never used to estimate the model.

There is also a long right tail in the $N = 250$ simulations, with the elasticities calculated using $\hat{\theta}$ up to 90% higher than those estimated with the true value. These higher elasticities seem to be related to $\hat{\theta}$ estimates that are lower than the true value. There are 41 simulations where the elasticities calculated with $\hat{\theta}$ at the point below the observed data are at least 25% larger than the elasticities as caluculated with the true value, and all of these simulations have $\hat{\theta} < 0.6$. The percentage differences are distributed more tightly around zero in the $N = 1000$ case, but there is still the large mass of observations approximately 16% lower, corresponding to the simulations with $\hat{\theta}$ near the upper boundary [1].

Similar to the $\alpha = 1$ DGP, *% Diff below min* with $N = 250$ has the largest difference between estimated elasticities when $\alpha = 5$. The histogram of this can be seen in Figure 2.6a, and it looks very similar to the equivalent histogram for the $\alpha = 1$ simulations, except in this

[1] *optim* does not halt right on the specified boundary, but very close to it.

Figure 2.5: *Percentage difference in estimated elasticities*

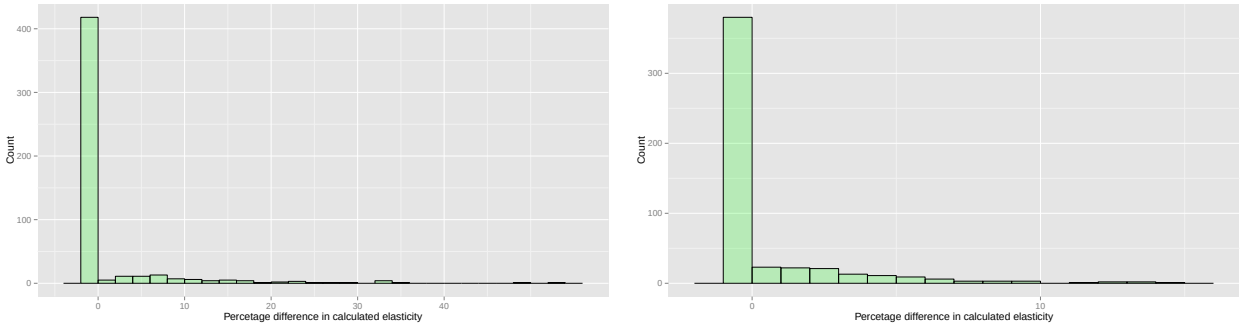


(a) (b)

% difference between elasticities estimated with $\hat{\theta}$ versus the true value. Data generated with $\theta = 1$, $\alpha = 1$. For a) $N = 250$, for b) $N = 1000$

case the largest grouping of observations is around zero rather than negative sixteen. Here the difference in estimated elasticities between $\hat{\theta}$ and the true value seems to be caused by the long right tail and low $\hat{\theta}$ estimates, rather than the values that halted near the boundary condition. These findings indicate that when $\alpha = 5$ the elasticities predicted by the model are relatively robust to the MLE θ mis-estimation, but that estimating a value lower than the true θ might cause issues.

Figure 2.6: *Differences between true versus estimated elasticities*



(a) (b)

Histogram of percentage difference in elasticity calculated with true θ versus $\hat{\theta}$ for each simulation at a point below the range of the data. (a) based on simulations with $N = 250$, $\theta = 1$, and $\alpha = 5$, (b) identical except $N = 1000$.

Table A.11 records the level of correlation between the elasticities at each observation cal-

Table 2.5: *Elasticity correlations:* $\theta = 1$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	0.997	0.998	0.999
$\theta = 0.5$	0.988	0.981	0.967
$\theta = 0.1$	0.663	0.604	0.613
$\theta = 0.05$	0.311	0.323	0.465
$\theta = 0.01$	-0.121	0.031	0.327
		$\alpha = 5$	
$\hat{\theta}$	0.999	1	1
$\theta = 0.5$	1	1	1
$\theta = 0.1$	1	1	1
$\theta = 0.05$	1	1	1
$\theta = 0.01$	0.994	0.994	0.994

culated with the true value of $\theta = 1$, and the values used in generating the other simulations as well as the $\hat{\theta}$ estimated using MLE. In each case, the value reported is the mean over the 500 simulations. It should be noted that while a low correlation does indicate that two sets of elasticities are not similar, a correlation close to one does not neccesserily indicate that they are the same, as it does not take into account differing magnitudes. The correlation between the elasticities estimated with $\hat{\theta}$ and those calculated with the true value of $\theta = 1$ is very high in all cases. Even in the $\alpha = 1$ and $N = 250$ simulations, where $\hat{\theta} \approx 148.4$ in almost half of the simulations, the mean correlation is 0.997. Looking at the results in more detail, there doesn't seem to be a relationship between $\hat{\theta}$ and the level of correlation.

Examining the other potential values for θ, the level of correlation between the elasticities estimated with the true value versus the other potential values appears to depend on the value of α used. When $\alpha = 5$, the level of correlation between the estimated elasticities is high for all values of θ. The lowest level of correlation when $\alpha = 5$ occurs when $\theta = 0.01$, and even there it is very high (0.994). When $\alpha = 1$ the other values of θ do not perform as well. Generally, as the value of θ falls the level of correlation between the estimated elasticities

also decreases. While the level of correlation is quite high when $\theta = 0.5$, with a minimum of 0.967, it is much smaller when $\theta = 0.1$ and continues to fall as the value of θ used decreases. This matches the results in Figure 2.5, where $\hat{\theta}$ estimates lower than the true value did lead to quite different elasticities.

As stated above, looking at only the correlation between estimated elasticites can be misleading. To measure how similar the various models are, Table 2.6 lists the average absolute percentage difference in elasticities calculated with either $\hat{\theta}$ or the various values used in the DGPs. This is done by first estimating each model, either with a selected value of θ or using $\hat{\theta}_{\text{MLE}}$, estimating the elasticities at each observation, and finding the absolute percentage difference between the true model and the other values of θ. This is then averaged over the observations of a single simulation, and Table 2.6 presents the mean of these values over the 500 simulations.

Table 2.6: *Average absolute percentage difference: $\theta = 1$ versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	1.0	0.6	0.2
$\theta = 0.5$	2.2	2.2	2.2
$\theta = 0.1$	27.9	27.5	27.6
$\theta = 0.05$	62.8	62.2	62.3
$\theta = 0.01$	310.3	308.2	308.7
		$\alpha = 5$	
$\hat{\theta}$	0.2	0.1	0.0
$\theta = 0.5$	0.0	0.0	0.0
$\theta = 0.1$	0.1	0.1	0.1
$\theta = 0.05$	0.2	0.2	0.2
$\theta = 0.01$	3.2	3.2	3.2

The results reported in Table 2.6 are broadly similar to those in the tables of correlations, with very small values when comparing $\theta = 1$ to $\hat{\theta}$ and $\theta = 0.5$, and the average percentage

difference increasing as the comparison value becomes smaller. When $\alpha = 1$ the differences become very large when using the smaller values of θ. For example, when $\theta = 0.05$, the average difference is slightly larger than 62%. When $\alpha = 5$ the general pattern is the same, with the average percentage difference increasing as θ falls, but the difference is much smaller. For example, when $\theta = 0.01$ the average difference is only 3.2%.

It is interesting to note that, while the mean correlations and average percentage differences between the elasticities calculated with the true value of $\theta = 1$ and $\hat{\theta}$ would indicate that misestimation of θ does not impact the calculated elasticities, the correlations between the elasticities calculated with the other values of θ indicate that it can be an issue. When $\alpha = 5$ the estimated elasticities were very similar for all values of θ, while when $\alpha = 1$ there were large differences in the estimated elasticities when comparing the true value to the smaller θ values. Even with the extremely high estimates of $\hat{\theta}$, the correlation between elasticities calculated with the true value versus $\hat{\theta}$ are very high and the average percentage difference low. Reducing the value of θ used on the other hand, seems to result in quite different estimated elasticities. In this case it seems that estimating a value for θ higher than the true value has little impact on the calculated elasticities, except for observations below the data used in estimation, while using a lower value of θ to estimate the model would impact the elasticities.

2.5.3 $\theta = 0.01$

Table 2.7 contains the summary of the results from the simulations with $\theta = 0.01$. With this DGP the MLE does not converge reliably with any of the sample sizes with either value of α. In each case, as the sample size increases the proportion of simulations which do successfully converge rises. *MLE max* is similar with $N = 250$ for both values of α, but seems to increase more as the sample size rises with $\alpha = 5$. This results in 64.4% of simulations successfully converging with $N = 5000, \alpha = 5$, where as with $\alpha = 1$ successful convergence only occurs in 58.4% of simulations.

Table 2.7: *Summary of simulations with* $\theta = 0.01$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max	0.500	0.506	0.584	0.492	0.530	0.644
$\bar{\hat{\theta}}_{\text{MLE}}$	0.038 (0.001)	0.027 (0.001)	0.018 (0.001)	0.033 (0.001)	0.022 (0.001)	0.015 (0.0004)
%Δ: median	−0.2 (0.018)	−0.1 (0.009)	0 (0.004)	-0.1 (0.013)	0 (0.004)	0 (0.4)
%Δ: 10th percentile	−1.4 (0.107)	−0.6 (0.054)	−0.2 (0.022)	−1.5 (0.116)	-0.6 (0.063)	−0.2 (0.031)
%Δ: 90th percentile	1.9 (0.152)	0.9 (0.076)	0.3 (0.036)	1.8 (0.143)	0.7 (0.076)	0.3 (0.040)
%Δ: below min	−1.6 (0.125)	−0.7 (0.063)	−0.3 (0.027)	−1.9 (0.152)	-0.8 (0.080)	−0.3 (0.040)
%Δ: above max	4.3 (0.326)	2.1 (0.179)	0.8 (0.080)	3.8 (0.295)	1.6 (0.161)	0.7 (0.085)

The mean estimate for $\hat{\theta}$, conditional on the log-likelihood being at a local maximum, has quite different results here compared to the simulations with $\theta = 1$. In the earlier simulations with $\theta = 1$ the mean estimated value was approximately correct with $\alpha = 1$, and smaller than the true value when $\alpha = 5$. With $\theta = 0.01$ the results are similar for both values of α, with the estimated value larger than the true value. As the sample size increases the mean estimate decreases, and the standard error is very small in all cases.

The most substantial difference between these results versus those for the $\theta = 1$ simulations is what occurs when the model doesn't successfully converge. With $\theta = 1$, non-convergence resulted in estimates very close to the upper boundary specified in *optimise*. With this DGP, $\hat{\theta}$ instead approaches the lower boundary (in this case 6.77×10^{-5}). As we will show in Section 2.6, when this occurs it appears that the concentrated log-likelihood is maximised as $\hat{\theta} \to 0$.

Examining Table 2.7, the only observations at which the percentage difference in elasticities is greater than 2% is the synthetic observation that lies above the data used in estimation. All points except %Δ *above max* show small differences, although larger than those observed in the $\theta = 1$ case, the largest being a 1.9% difference at the 90th percentile for the $N = 250, \alpha = 1$ simulations. The results are very similar over the two potential values

of α, and all of the differences decrease as the sample size increases. $\%\Delta$ *above max* shows a mean difference of 4.3% in the elasticities calculated in the simulations with $\alpha = 1$ and $N = 250$, and a 3.8% difference with that sample size when $\alpha = 5$. This decreases to 2.1% and 1.6% respectively when the sample size increases to $N = 1000$, and decreases again for $N = 5000$.

This result is quite different to that observed in the $\theta = 1$ simulations, where the largest difference in estimated elasticities occured below the range of the data. A closer examination of the results shows that in the $N = 250, \alpha = 1$ simulations, all repititions where the concentrated log-likelihood is not maximised by $\hat{\theta}$, (And $\hat{\theta}$ is therefore very close to the lower bound) have elasticity differences that are very small and close to zero. It is the simulations that are local maximums, which have a mean $\hat{\theta}$ estimate above the true value of 0.01, that result in the higher percentage difference in estimated elasticities. This result is repeated for both values of α, and all sample sizes. High values of $\hat{\theta}$ seem to produce elasticities that are larger than those calculated with the true value of θ. This results in $\%\Delta$ *above max* becoming smaller as the mean estimate for $\hat{\theta}$ decreases towards to true value as the sample size increases.

Table 2.8 contains the level of correlation between the elasticities calculated with the true value of $\theta = 0.01$, versus $\hat{\theta}$ and a range of other potential values. When $\alpha = 5$, the elasticities produced by all of the values of θ used are highly correlated with those estimated using the true value, with the level of correlation increasing as the value of θ decreases. This leads to $\hat{\theta}$ producing the elasticities that are most highly correlated with those produced by true value of $\theta = 0.01$. When $\alpha = 1$, for all three sample sizes $\theta = 0.1$, $\theta = 0.05$, and $\hat{\theta}$ produce elasticities that are very closely correlated with those produced by the true value, and generally the level of correlation increases with the sample size. The remaining values, $\theta = 1$ and $\theta = 0.5$, both produce elasticities that have much lower levels of correlation with those produced using the true model.

These results are supported by Table 2.9, which reports the mean average absolute per-

Table 2.8: *Elasticity correlations:* $\theta = 0.01$ *versus other values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	0.996	1	1
$\theta = 1$	0.321	0.379	0.480
$\theta = 0.5$	0.433	0.533	0.667
$\theta = 0.1$	0.920	0.958	0.984
$\theta = 0.05$	0.992	0.996	0.999
		$\alpha = 5$	
$\hat{\theta}$	0.998	0.999	1
$\theta = 1$	0.918	0.918	0.918
$\theta = 0.5$	0.921	0.920	0.920
$\theta = 0.1$	0.963	0.963	0.963
$\theta = 0.05$	0.991	0.991	0.991

Table 2.9: *Average absolute percentage difference:* $\theta = 0.01$

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	1.2	0.6	0.3
$\theta = 1$	45.6	45.7	45.7
$\theta = 0.5$	37.8	37.9	37.9
$\theta = 0.1$	10.0	10.0	10.0
$\theta = 0.05$	3.2	3.2	3.2
		$\alpha = 5$	
$\hat{\theta}$	1.3	0.7	0.4
$\theta = 1$	24.2	24.2	24.2
$\theta = 0.5$	23.3	23.3	23.3
$\theta = 0.1$	11.4	11.4	11.4
$\theta = 0.05$	4.5	4.5	4.5

centage difference in estimated elasticities. Both tables 2.8 and 2.9 indicate that the high levels of non-convergence, which lead to very low estimates for $\hat{\theta}$, still result in estimated elasticities that are very close to those estimated using the true value of θ. Conversely, if a larger value of θ is selected, Table 2.9 shows that the elasticities estimated are relatively different from those produced by the true model, even when the level of correlation reported in Table 2.8 was high. These results are the reverse of those observed in the $\alpha = 1, \theta = 1$ case, where over estimation of $\hat{\theta}$ did not lead to different elasticities, but selecting a value lower than the true value did.

2.5.4 Other θ Values

As reported in Section 2.5, further simulations were carried out with a range of θ values used in the DGP. In this section we will summarise these simulations, with the full tables of results available in Appendix A. The simulations with $\theta = 0.1$ perform best, with the model successfully converging 100% of the time and a $\bar{\hat{\theta}}$ very close to the true value. Likewise, the differences in estimated elasticity at each observation are very small.

With $\theta = 0.5$ and $\alpha = 1$ the estimator converges in almost 100% of the simulations, while with $\alpha = 5$ convergence is less reliable. Likewise $\bar{\hat{\theta}}$ appears to be more accurately estimated with $\alpha = 1$, with the mean estimate higher and with a large standard error for $N = 250$, but very close to the true value and with a much smaller standard error for the larger sample sizes.

For $\alpha = 5$ the estimated values are all below the true θ. As the sample size increases $\bar{\hat{\theta}}$ moves closer to the true value, but the standard errors also increases substantially. Examining the results in more depth, it appears there is a large clump of observations with $\hat{\theta}$ below the true value, and then a long tail that extends to the right. As the sample size increases the right tail extends, increasing both the mean estimate as well as the standard error. The difference in estimated elasticities for the true θ versus $\hat{\theta}$ is very small at all points examined except at the synthetic point below the data used in estimation. Even here, the

mean difference is only 4.7% for the simulations with $N = 250$, and this decreases to 2.5% and 0.7% when $N = 1000$ or $N = 5000$ respectively.

The simulations with $\theta = 0.05$ seem to perform quite well. With $\alpha = 1, N = 250$ the model successfully converges in 86% of the simulations. With the other sample sizes and including $\alpha = 5$, successfull convergence occurs in at least 96% of the simulations. $\bar{\hat{\theta}}$ is close to the true value of 0.05 in all of the simulations, with $\alpha = 1, N = 250$ again performing the worst, but still with a mean estimate of 0.054. The difference in estimated elasticity between the true θ and $\hat{\theta}$ is small at all points examined.

We also estimated models where the observed Y values were scaled before estimation. In these simulations we found the results similar to the non-scaled simulations with the same θ value, at least with regards to if the model successfully converges or not. The estimated θ values were different from the 'true' value used in the DGP, but this is due to the non-linear nature of the IHS transformation. The scaling also makes it hard to estimate the true models elasticities, as the value of θ that should be used to calculate them is now different to the value used in the DGP.

2.5.5 Censored Data

In order to simulate data that better matches the empirical earnings data, we now introduce simulations where a portion of the observations are censored. As detailed previously, it is common for individuals to have periods where they do not work. The main concern for economists is if this censoring is the result of random or non-random selection. If the decision to work or not work is correlated with earnings then OLS will be biased.

We create data with similar features to those observed empirically by randomly censoring 10% of the observed Y values. While in this case using OLS on the subset that is observed would result in an accurate model, our interest is in establishing how the IHS estimation procedure performs in the simplest scenario with censored data.

Table 2.10 presents the results for the simulations with $\theta = 0.1$ and 10% censoring. We

Table 2.10: *Summary of censored simulations:* $\theta = 0.1$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max	0.006	0	0	0	0	0
$\bar{\hat{\theta}}_{\text{MLE}}$	0.020 (0.011)	- -	- -	- -	- -	- -
%Δ: median	5.4 (0.08)	5.5 (0.04)	5.4 (0.02)	0.069 (0.16)	0.067 (0.08)	0.066 (0.04)
%Δ: 10th percentile	-	-	-	-	-	-
%Δ: 90th percentile	−19.5 (0.05)	−19.6 (0.03)	-19.6 (0.01)	−23.8 (0.10)	−24.1 (0.04)	−24.2 (0.02)
%Δ: below min	24.1 (0.05)	24.1 (0.02)	24.1 (0.01)	44.7 (0.14)	44.6 (0.07)	44.5 (0.03)
%Δ: above max	−39.6 (0.04)	−39.7 (0.01)	−39.8 (0.00)	−44.7 (0.06)	−44.9 (0.03)	−45.0 (0.01)

selected $\theta = 0.1$ as in the previous simulations this value resulted in the most stable model[2]. Unlike the previous simulations we have examined, here the model almost never successfully converges. For $\alpha = 1, N = 250$, $\hat{\theta}$ results in a local maximum in the concentrated log-likelihood in only 0.6% of simulations, which is three out of the 500 performed with this DGP. The model never converges with any of the other sample sizes, with either value of α. These results are very different than the ones achieved with $\theta = 0.1$ and no censoring. In that case the model successfully converged in all simulations with both values of α for all sample sizes.

The three times the model does converge, the estimated θ value is lower than the true value of 0.1. With this DGP non-convergence results in $\hat{\theta}$ values very close to the lower bound of 6.7×10^{-5}. The differences in estimated elasticites are very consistent over the different sample sizes for each given α, remaining almost unchanged but with the standard error decreasing as N increases. The mean % differences are quite large, with the smallest differences occuring at the median observation and being approximately 5.4% when $\alpha = 1$, and 6.7% for $\alpha = 5$. The differences at all other points reported in Table 2.10 are larger, and indicate that the estimated θ does a poor job of producing a model with elasticities similar

[2]Results for $\theta = 1$ with censoring are available in Appendix A

to those generated with the true model.

It is difficult to know if this failure to accurately estimate the models elasticities is dirrectly due to the failure to estimate θ, or that including the zero values in the second stage regression leads to bias in the estimation of α and β which itself leads to different elasticities. Either way it seems that in the presence of censored data, where zero observations are a special case rather than a point in a continuum of continuous outcomes, a model that uses the IHS can report estimated elasticities quite different to the true model.

2.5.6 Alternative Optimisation Methods

We have tested a number of additional estimation methodologies that can be used to estimate θ, in addition to the single dimensional optimisation using the R function *optimise* detailed in Section 2.5. Additional methods attempted include; maximising the concentrated log likelihood, both with and without the analytical gradient, using the package *maxLik* (Henningsen and Toomet, 2011) in R, and also using multi-dimensional optimisation on the full log-likelihood, again with *maxLik*. *MaxLik* is an R function designed for maximum likelihood estimation, with its main advantages over *optimise* being that it can perform multi-dimensional optimisation, doesn't require a specified region to search over, and allows specification of the analytical gradient.

Some of the simulations performed in Section 2.5.1 were also replicated in Matlab, using the same data that were generated in R . The estimation in Matlab used the function *fmincon*, which is an optimiser that minimises the objective function subject to user specified constraints. In this case we estimated $\hat{\theta}$ using the concentrated log-likelihood and with the same upper and lower boundaries used in the matching simulation performed in R . This is designed as a robustness check, ensuring that the results and estimation issues encountered in R are not artifacts caused by the language/optimisers, but an econometric issue.

There are no major differences in the achieved results based on which optimisation method is used, none of the methods outlined avoid the estimation issues encountered in

Section 2.5.1. The single dimensional optimisation with *optimise* and *maxLik* on the concentrated log likelihood without the gradient is very fast. Adding the analytical gradient to the optimisation with *maxLik* slows it down considerably, but also seems to get more precise estimates (when θ is accurately estimated).

Estimation with *maxLik* on the full log-likelihood at first seems to estimate the correct θ even when the other methods do not, but this only occurs when the true values of the parameters are used as the starting conditions in the optimiser. Changing these by even a little bit gives a very different θ estimate. This seemingly correct estimation seems to be caused by a very flat log likelihood function, so the optimiser stops immediately at a point which is not a local maximum.

2.6 Discussion

According to the literature, θ should be consistently estimated, and with it we should be able to estimate the other parameters of interest. The results from Section 2.5 of this chapter indicate that the reliability of this estimator depends on the the value of θ used in the transformation, and the particular data available. The DGP, sample size, and value used for θ all seem to affect the probability that θ can be successfully estimated, and in the case that estimation is problematic, whether $\hat{\theta}$ approaches the lower or upper boundary. The issue seems to be one of poor identification, potentially arising when the generated data is largely contained in either the region of the function that is linear, or the region where it is log-like.

As θ becomes more extreme, either moving towards zero or increasing in size, the probability of successful estimation decreases. In the simulations conducted in Section 2.5 the DGPs with $\theta = 1$ and $\theta = 0.01$ were the least reliable, while the values between these seemed to perform better. Holding θ fixed and shifting the underlying distribution of IHS(Y) by altering α also seems to affect the probability of successful estimation. Changing the value of

θ alters the portions of the domain that are approximately linear and approximately log-like. Increasing or decreasing α shifts the data, potentially moving it between these two regions.

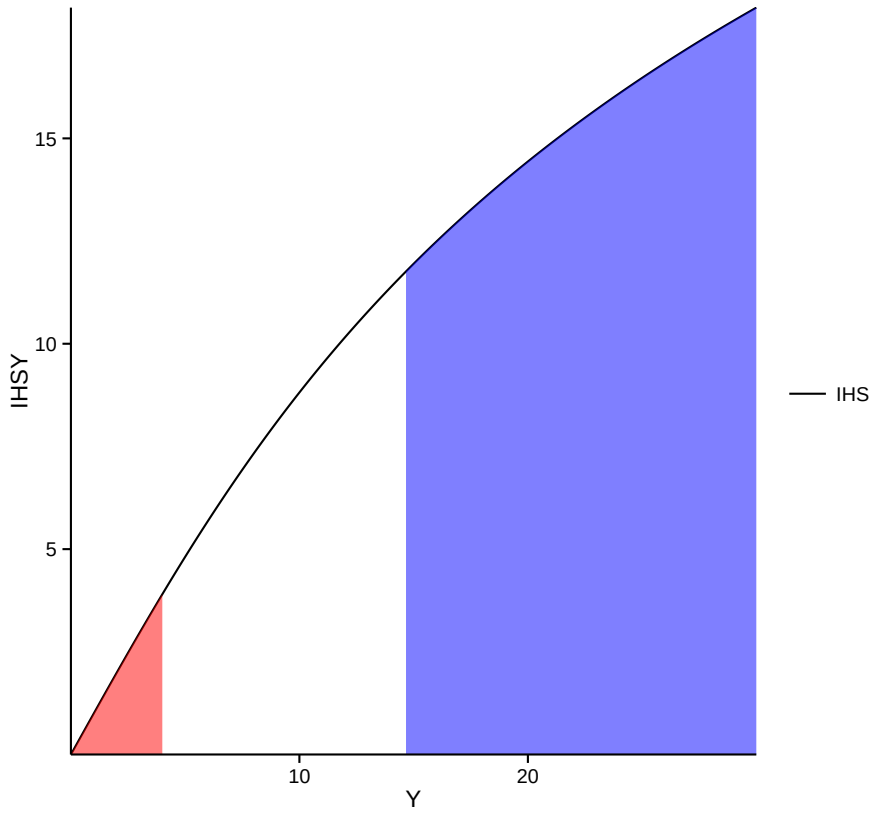


Figure 2.7: *Plot the the IHS with $\theta = 0.1$. The shading under the plotline shows where the function is approximately linear (red region) and loglike(blue region).*

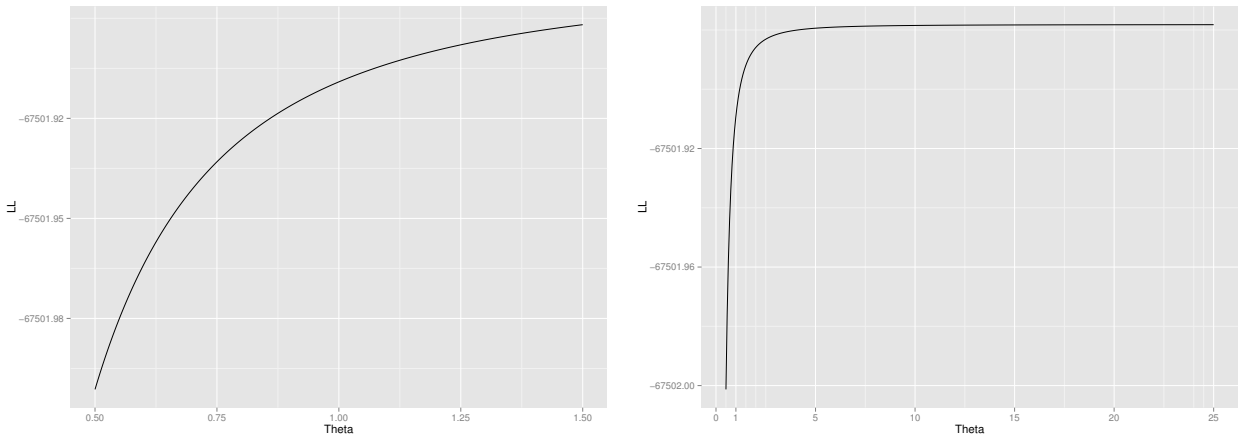
Figure 2.7 plots the IHS with $\theta = 0.1$, and highlights portions of the domain that are approximately linear and logarithmic respectively. As θ increases, the loglike portion will expand and the linear portion will contract towards zero. Alternatively, as θ decreases the linear portion will expand outwards, along with the beginning of the loglike region. In the simulations detailed above, the value that α takes shifts where the data lies, and can influence how much of the data lies in the linear or log like regions.

In the simulations we have conducted, we have observed two distinct outcomes when $\hat{\theta}$ does not successfully converge. It either tends towards the upper or lower boundary, with these boundaries specified in the optimiser. These two outcomes do not occur within the

same DGP. It seems that, given a data set, $\hat{\theta}$ will tend to only the upper or lower boundary. We will examine these two results seperately in an attempt to understand what is causing the estimation issues.

We observe $\hat{\theta}$ failing to converge and approaching the upper bound with two DGPs, when $\theta = 1$ or $\theta = 0.5$. Compared to the other values used, these θ values result in the function becoming approximately logarithmic the earliest. We also find that non-convergence occurs more frequently in these DGPs when $\alpha = 5$, shifting the distribution towards the approximately logarithmic portion of the function. Likewise, with $\alpha = 1$ this issue occurs more frequently with $\theta = 1$, where the relative portion of the domain which is approximately logarithmic is larger and closer to zero.

In this situation, the concentrated log-likelihood becomes very flat and is upwards sloping, as in Figure 2.8. In panel (a) it is clear that the function is not maximised at the true value of one. Panel (b) shows how flat the function becomes. This can cause issues with some estimation techniques, with the extreme flatness causing the optimiser to halt at the starting parameters. Using optimise it results in an estimate very close to the upper boundary as increasing the value of $\hat{\theta}$ continues to increase the objective function.



(a) (b)

Figure 2.8: *Concentrated log-likelihood of model with $\alpha = 1,\ \theta = 1$. Left panel is focused on region around the true value, right panel shows flatness of the concentrated log-likelihood as $\hat{\theta}$ increases.*

If we assume that all observations lie in the portion of the domain that is log-like, then we can substitute in the asymptotic equivalent and by rearranging we get an equation that is not fully identified, as in Equation (2.6.1). Here we find that, asymptotically, a model estimated with the IHS generates a slope coefficient that is approximately equal to that generated in a log-linear model when β is multiplied by θ. This reinforces the earlier claim that $\theta\beta$ is equivalent to β in the log-linear model. It also illustrates that there are multiple combinations of θ and β that produce the same net effect, as indicated by the results of calculating elasticities when the data lies in the log like region. This was shown in Section 2.5.2, where different values of θ lead to very different βs, that still produced elasticities very similar to the true model.

$$\text{IHS}(Y_n) = \alpha + X_n\beta + \epsilon_n \quad , \tag{2.6.1a}$$

$$\frac{\log(2\theta Y)}{\theta} = \alpha + X_n\beta + \epsilon_n \quad , \tag{2.6.1b}$$

$$\log(Y) + \log(2\theta) = \theta\alpha + \theta X_n\beta + \theta\epsilon_n \quad , \tag{2.6.1c}$$

$$\log(Y) = \big(\theta\alpha - \log(2\theta)\big) + \theta X_n\beta + \theta\epsilon_n \quad . \tag{2.6.1d}$$

Manual experimentation with individual data sets from the sets of simulations which do not successfully converge reinforces this result. By manually selecting values of θ and then estimating the second stage regression, very different values of β are estimated. However, the estimated elasticites are very close to those produced by both the true θ value, and $\hat{\theta}$. The transformation $\theta\beta$ also produces values that appear to be distributed around the true value, as illustrated previously in Figure 2.4.

This identification issue is highlighted by the following example. First we generate a linear sequence from five to twenty, y, and then transform it with the HS and a value of θ such that all the values lie in the exponential portion of the domain (This results in a transformed sample where the values lie in the approximately logarithmic portion of the IHS

for the same value of θ). Second, we apply the IHS with the true of θ used in step one, as well as larger values. Using larger values ensures that HS(y) will always be in the approximately logarithmic portion for each transformation. If we plot these together, as in Figure 2.9, we find that each value of θ used transforms HS(y) into a strainght line, with a different slope in each case.

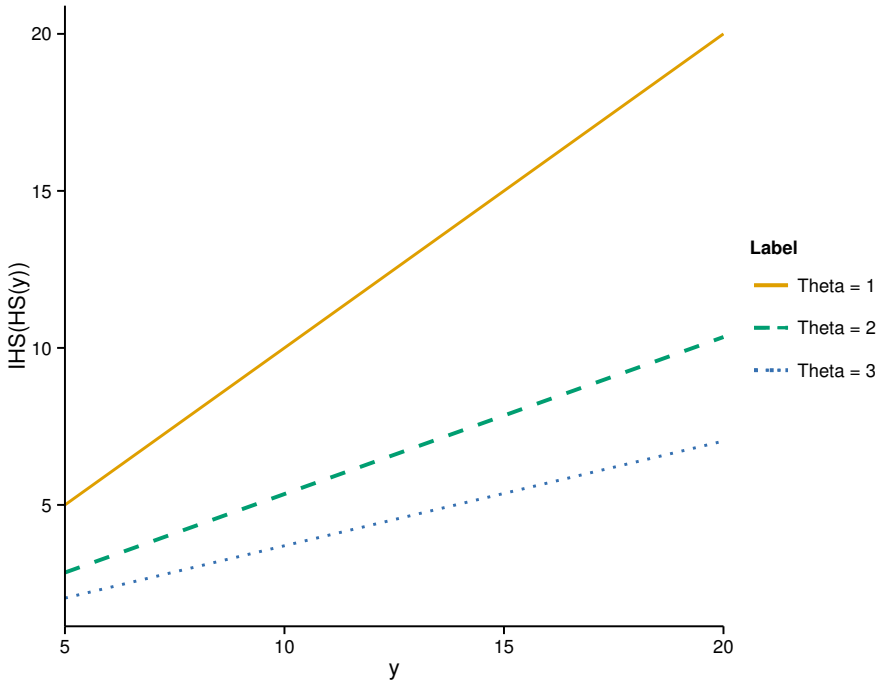


Figure 2.9: *Plot of IHS transformation applied with various values of θ to data in the approximately logarithmic portion of the functions domain.*

An important question is, does the misestimation of θ lead to a model that makes inaccurate or biased predictions? In Section 2.5.1 we found that in the simulations where $\hat{\theta}$ approached the upper boundary, the estimated elasticities were still close to those estimated with the true value. If the over estimation is caused by the data lying in the approximately logarithmic portion of the functions domain, then it seems that there is a range of θ values that will produce models that make very similar predictions. If this is the case then those models will give accurate estimates of the elasticities, at least for observations that are similar to those used in the estimation procedure. On the other hand, the value of θ used

determines where the function moves from approximately logarithmic to linear. If the value used in estimating the model is different from the true value, then forecasting or modeling of observations that lie below the data used in estimation may be inaccurate. If the true value for θ implies a linear relationship at a particular point, but the estimated value implies an approximately logartihmic one, then elasticities calculated for observations in that region will be incorrect. This point is reinforced by the differences we observed in estimated elasticities below the range of data for $\theta = 1$.

For smaller values of θ, non-convergence implies that $\hat{\theta}$ will be very close to the lower bound. As detailed in Section 2.2.2, a θ of zero implies that there is a linear relationship without transforming the dependent variable, and testing if $\hat{\theta} = 0$ is akin to testing if the IHS transformation is required (Burbidge et al., 1988; Pence, 2006).

This issue seems to arise when the data used in estimation all lies in the region which is approximately linear. Examining $\hat{\beta}$ in the simulations where convergence does not occur and the parameter approaches the lower bound, we find that they are almost identical to the true values used in the DGP before the HS transformation is applied. This implies that given the data and value of θ, applying the HS has almost no impact on Y.

$$\mathcal{L}^c(\theta) = C - \frac{n}{2}\ln(\text{IHS}(Y,\theta)'M\ \text{IHS}(Y,\theta)) - \frac{1}{2}\sum \ln(1 + Y_i^2\theta^2). \tag{2.6.2a}$$

$$M = I - X(X'X)^{-1}X' \tag{2.6.2b}$$

Equation (2.6.2a) is the concentrated log-likelihood first introduced in Section 2.4. Let H be the middle component of the concentrated log-likelihood, so

$$H(\theta, Y) = -\frac{n}{2}\ln(\text{IHS}(Y,\theta)'M\ \text{IHS}(Y,\theta)) \quad . \tag{2.6.3}$$

H will remain approximately constant for all values of θ, such that the data lies in the linear portion of the function. For example let $j \in Y$, and $j \geq y \forall y \in Y$. If $\text{IHS}(a, j) \approx \text{IHS}(b, j)$,

then $H(a, Y) \approx H(b, Y)$. In this case the concentrated log likelihood will estimate the θ that maximises the third component of the equation, $-\frac{1}{2}\sum \log(1 + Y_i^2\theta^2)$, leading to $\hat{\theta} = 0$.

2.7 Conclusion

This chapter introduced the IHS as an alternative transformation to the logarithm. The IHS has many properties that make it attractive, and is defined for negative and zero values, unlike the logarithm. The IHS has been used in a range of applications, but its use has not been extensively explored in the income and earnings dynamics literature. Chapter 3 will look at applying the IHS to both data simulated to closely match empirical earnings data, as well as empirical earnings data from New Zealand.

The simulation exercises conducted in this chapter have shown that estimation of models using the IHS is not always reliable. To date, the literature has not explored these estimation issues. The value estimated for the scaling parameter $\hat{\theta}$ appears to be especially volatile, depending on the DGP used it is sometimes estimated at the upper or lower boundaries. However, this misestimation seems to result in models that make predictions very similar to the true model. Within the range of the data used in estimation, differences in elasticities between models based on $\hat{\theta}$ and the true θ value are generally small.

One of the main strengths of the IHS is its ability to function with zero and negative values. The simulations in this chapter showed that, if zero is a special case and not part of a continuous distribution, then MLE estimation of the model does not seem to be accurate or effective. The estimated $\hat{\theta}$ was very small, and the elasticities were very different from the true model. A potential solution to this issue is manually selecting a value of θ to use, which is a common approach in the literature. However, we have shown that even when misestimation of θ resulted in a model that makes predictions that are very similar to the true model, manually selecting a value can lead to a model that is very different.

While the models estimated seem relatively robust to θ misestimation within the range

of data used, the estimated elasticities can be quite different outside of this range. For any two values of θ there will be portions of the domain of the IHS where they are each log-like, and portions where they are both approximately linear. Generally, θ misestimation seems to occur when all of the data lies in either the approximately logarithmic or linear portions of the function with the true θ. If a different θ value is used to estimate the model, the elasticities will be approximately equal where the two functions overlapping regions are synchronised. Point estimates of elasticities may differ from the true model when applied to data outside of the range used in estimation if the θ value used is different from the true value.

This chapter shows some of the strengths and weaknesses of the IHS as a replacement for the logarithm. Not only does the IHS display many of the desirable properties of the logarithm, but it is also defined for zero or negative values. These properties are particularly useful in the context of female earnings, which the literature has largely ignored due to the high frequency of zero values. However, care must be taken in situations where zero observations are a special case, and when using the model to make predictions outside of the data used in estimation. Chapter 3 will focus on applying the IHS to annual earnings data, while taking into account the issues encountered in this chapter.

Chapter 3

Earnings Dynamics With The Inverse Hyperbolic Sine Function

3.1 Introduction

This chapter focuses on the dynamics of annual earnings, and the results of applying the Inverse Hyperbolic Sine (IHS) function to panel earnings data. Panel data provides economists with a powerful tool that can be used to estimate and quantify causal relationships and intertemporal correlations. By examining repeated measurements taken from a sample population over time, panel data allows the researcher to observe trends, estimate dynamic relationships, and account for individual heterogeneity in a way which is difficult or impossible when using only cross-sectional data (Hausman and Taylor, 1981; Bond, 2002).

In the context of annual earnings, examining the distribution of earnings at a moment in time doesn't reveal the relative importance of permanent versus transitory variation, how persistent earnings are over time, or how the distribution of earnings changes over time (Baker, 1997; Meghir and Pistaferri, 2011). Many of the factors that contribute to and affect annual earnings are inherently intertemporal, and understanding their impact necessitates the use of longitudinal data (Hause, 1977).

Many workers will have periods of time where they do not work, and thus have zero earnings. Women are especially likely to have years where they do not work, and have lower overall participation rates than men. In this context, sample selection bias can be a serious issue. This occurs when the data available are not a random sample of the population, but are the result of selection based on the decisions of the sampled population, or on the decisions of the researcher (Kyriazidou, 1997). Logarithmic transformations of wage, income, and earnings are frequently used in the labour economics literature, and as explained in Chapter 2 the logarithm is not defined for zero values. Thus the researcher must decide how to treat these observations where earnings are zero.

One common method is to exclude observations where earnings are zero. This is done by either removing only the observations where an individual does not work, resulting in an unbalanced panel, or removing any individual that has any periods where they have earnings of zero, leading to a balanced panel (MaCurdy, 1982; Meghir and Pistaferri, 2004; Heathcote et al., 2014). With both balanced and unbalanced panels, if selection into/out of the workforce is non-random, then excluding these observations results in the observed sample being a non-representative portion of the entire population (Vella, 1998). This could lead to biased coefficient estimates, and as such inference based on the sample will not generalise to the wider population (Dustmann and Rochina-Barrachina, 2007).

Removing periods of workforce non-participation also ignores a large component of life-time earnings. The variance of earnings is made up of two dimensions; extensive and intensive variation. Variation of earnings at the intensive margin is caused by changes in hourly wage or changes in the number of hours worked (Heckman, 1993). The intensive margin for male workers has been the focus of the majority of literature looking at earnings dynamics (Abowd and Card, 1989; Lillard and Willis, 1978). Extensive margin variation, that is variation in earnings associated with moving into or out of the workforce, results in large contemporaneous shocks to earnings and can impact long-term earnings dynamics (Carroll et al., 2003; Gregory and Jukes, 2001). Excluding observations with zero earnings, or individuals that

have periods of workforce non-participation, ignores the dynamic effects of extensive margin variation, and might lead to incorrect inference about a range of topics relating to earnings, such as consumption decisions, risk management, and the persistence of earnings over time.

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).

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. A focus on only periods where females work, or on only females that work in every period, could lead to biased coefficient estimation, and a flawed understanding of female earnings dynamics.

Chapter 2 introduced the IHS function and explored its properties. It was applied to simple simulated cross-sectional data that was not modelled on the empirical moments of earnings data. In this chapter the IHS is tested as an alternative to the logarithm in studies of earnings and earnings dynamics. There are essentially two empirical issues caused by individuals with zero earnings. First, these observations cannot be included in any kind of estimation or statistical analysis that uses the logarithm. This means that standard measures such as the mean or variance of log(earnings) do not take into account periods of zero earnings. Second the exclusion of these observations can lead to sample selection bias if selection into or out of the workforce is non-random. While the IHS does not take

into account the behavioural decision to work, and thus might not correct for non-random selection, it does allow for the inclusion of periods where earnings are zero. This allows for the inclusion of all observations, making greater use of the data available, and allows us to examine the effect that extensive margin changes and prolonged unemployment can have on individual earnings.

This chapter proceeds as follows: Section 3.2 reviews the earnings dynamics literature. A number of methods exist to analyse the dynamics of earnings, and this section introduces some of them, including variance decomposition using an Error Components Model (ECM), and decomposing the variance of earnings into extensive and intensive margins. Section 3.3 employs Monte-Carlo simulation in generating data that is designed to be similar to empirically observed earnings. The IHS and logarithm are each applied to simulated data produced by two models from the literature, and their impact on the descriptive statistical properties of earnings such as the auto-covariances and the decomposition of its variance into intensive and extensive portions are compared. Finally, Section 3.4 presents an empirical application. Statistics New Zealand's Survey of Families, Income, and Employment (SoFIE) data set is introduced, and the same methods used in Section 3.3 are applied using both the IHS and logarithm of annual earnings for female workers.

3.2 Earnings Dynamics

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 by individuals, but also by seemingly random shocks (Altonji et al., 2013). An examination of the dynamics of earnings can reveal how persistent earnings shocks are, and the impact they have on lifetime earnings and

individual decision making(Lillard and Weiss, 1979; Meghir and Pistaferri, 2011; Jappelli and Pistaferri, 2006). Over the life cycle of an individual, the distribution and variance of earnings can also change. How they change, and in what magnitude, informs the decisions that individuals make.

The evolution of annual earnings 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). Likewise, understanding the distribution of earnings and what it means requires taking into account how it evolves over time. For example, a cross-sectional analysis that finds a high level of earnings inequality could be driven by an underlying system with perfect income mobility, or permanent stratification (Moffitt and Gottschalk, 2011; Lillard and Willis, 1978). Cross-sectional analysis of earnings restricts the inter-temporal effects, allowing only a snap-shot of the earnings distribution at a moment in time (Hsiao, 2014). Expanding this with the use of panel data allows the researcher to gain an understanding of how the distribution changes, how persistent earnings are for an individual over time, and how the variance of earnings changes over the life cycle (Baker, 1997; Meghir and Pistaferri, 2011).

Earnings are frequently modelled as a function of observable variables, and a sum of some number of random unobserved components (Meghir and Pistaferri, 2011). A simple example of this formulation is the earnings function in (3.2.1), where Y_{it} is some measure of annual earnings, X_{it} is a vector of observed variables that affect earnings, and ϵ_{it} is an unobserved error term made up of some unspecified components. Thus the measured variables are modelled explicitly by the earnings function, while unmeasured components appear in the residual term (Lillard and Willis, 1978). Frequently, Y_{it} is the logarithm of annual earnings, and the variance of this can interpreted as a measure of earnings inequality (Lillard and Willis, 1978). This means that understanding the components of the error term tells us

something about the dynamic structure and level of inequality (Baker and Solon, 1999) .

$$Y_{it} = X_{it}\beta + \epsilon_{it} \ . \tag{3.2.1}$$

ECM are used in the earnings dynamics literature to decompose ϵ_{it} into its separate components. Typically some earnings function is specified, as in (3.2.1), and estimated through Ordinary Least Squares (OLS), Generalised Least Squares (GLS), or some other method. Then ϵ_{it}, the residuals from this regression, are analysed and/or fit to some theoretical model, and the inter-temporal correlation patterns of earnings are investigated. MaCurdy (1982) provides a good explanation of this, as well as detailing a number of ECM that are commonly estimated. Frequently the auto-covariance matrix of the residuals (or differenced residuals) is examined, and used to fit the desired model using Minimum Distance Estimation (MDE) (Chamberlain, 1984; Heathcote et al., 2014). In their seminal paper, Abowd and Card (1989) examine the covariance structure of experience adjusted earnings and hours changes. They present auto-covariance matrices, analyse what they imply about the structure of earnings, and fit the results to a number of ECMs.

Equations (3.2.2a) and (3.2.2b) are some basic formulations for the error term. Equation (3.2.2a) contains an error term made up of an unobserved, individual specific fixed effect α_i, while (3.2.2b) contains a unit root (δ_{it}). In each case, e_{it} is an idiosyncratic error, frequently described as measurement error (Duncan and Hill, 1985).

$$\epsilon_{it} = \alpha_i + e_{it} \ , \tag{3.2.2a}$$

$$\epsilon_{it} = \delta_{it} + e_{it} \ , \quad \text{where } \delta_{it} = \delta_{it-1} + u_{it} \ . \tag{3.2.2b}$$

Frequently the error term is assumed to contain some combination of permanent effects, such as a unit root component or individual specific fixed effects, and transitory shocks in the form of MA(k) or AR(l) components (Meghir and Pistaferri, 2011). For example, Lillard and Weiss (1979) used an ECM containing an individual specific fixed effect, an AR(1)

process, and individual specific growth factors, to estimate an earnings model for American scientists. Examining the auto-covariance structure of the residuals can provide evidence as to the structure of the error term.

The distinction between permanent and transitory shocks is important. First, in Friedman's permanent income hypothesis, individuals react differently to permanent versus transitory shocks to earnings, with differing effects on individual decision making and the level of consumption (Baker and Solon, 1999; Meghir, 2004). Second, many studies have found that both consumption and earnings have increasing variances over the course of the life cycle, potentially indicating a unit root in the residual (Meghir and Pistaferri, 2011; Deaton and Paxson, 1994). Third, the relative magnitude of the permanent and transitory shocks has a large impact on the distribution and mobility of earnings, inequality, and how these evolve over time (Baker and Solon, 1999; Hyslop, 2001). Thus this distinction allows for the separation of earnings variability that occurs across individuals, as opposed to variability for an individual over time (MaCurdy, 1982).

Studies of earnings dynamics have generally been focused on the the intensive margin, and the earnings of male workers (Abowd and Card, 1989; Heathcote et al., 2014; Meghir and Pistaferri, 2004), although there are some exceptions that have examined the earnings dynamics of female workers (Dolton and Makepeace, 1986; Hyslop, 2001). The focus on male earnings is largely due to the differences in participation rates between males and females, where historically males have been more likely to work in any given year (Heckman, 1993; Haider, 2001). This has meant that male earnings were a larger component of family earnings, and using male earnings data reduced the potential for sample selection bias. Moffitt and Gottschalk (2011) for example, focus on white males so that periods of zero earnings are less of an issue.

While the lower level of female workforce participation has been one of the driving factors behind the focus on male earning dynamics, female levels of workforce participation have risen dramatically since World War II. Killingsworth and Heckman (1986) show data that

indicates 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). 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.

Modelling earning dynamics while ignoring extensive margin variation also removes a large component of the variance of life time earnings (Couch and Lillard, 1998). A focus on only the intensive margin variation excludes groups that have lower rates of participation, such as female workers, and presents models of earnings dynamics that apply to only a subset of the population.

Descriptive statistics of annual earnings based on only workers also gives a potentially misleading understanding. For example, looking at the average of log(earnings) for a sample ignores the unemployed, overstating average earnings. Likewise, estimating the variance of log(earnings) gives us a measure of how earnings vary for those who work, but ignores movements into and out of the workforce, which potentially constitute a large portion of life time earnings variation, especially for female workers.

Variation in earnings can be decomposed into its component parts, allowing the impact of extensive margin variation to be captured. Equation (3.2.3) outlines this decomposition, as proposed by Hyslop and Card (2016). Here v_i is the variance of earnings for individual i, Y_{it} is some measure of earnings for individual i in period t, p_i is the fraction of years in the sample in which i works, v_{ic} is the variance of earnings in the years in which i works, and

$(\overline{Y}_i^c)$ is the mean of earnings for those years.

$$v_i = \frac{1}{T}\sum_{i=1}^{T}(Y_{it} - \bar{Y}_i)^2 = p_i v_{ic} + p_i(1-p_i)(\bar{Y}_i^c)^2 \ . \tag{3.2.3}$$

This decomposition allows for the comparison of the extensive and intensive margin contributions to the variance of earnings.

3.3 Monte-Carlo Simulation

In this section Monte-Carlo simulation is used to investigate the effects of applying the IHS transformation to data simulated to match empirical annual earnings. Panel data is simulated using a model from the literature, with a focus on matching the variance-covariance structure observed in annual female earnings. An ECM is used, where the simulation generates residuals from a regression of earnings on some set of observable features (Amemiya, 1971; Meghir and Pistaferri, 2011).

The model is of a stationary process, with constant cross-sectional variance of earnings. To match the participation patterns observed empirically for female workers, workforce non-participation is introduced where, based on individual specific characteristics or shocks, an individual will not work in a given period and thus have earnings of zero. We explore different levels and mechanisms of non-participation that introduce both permanent and transitory unemployment, and apply these to the simulated data.

The simulated earnings data are analysed using the untransformed level of earnings, balanced and unbalanced panels of log(earnings), and IHS(earnings). The sample statistics for each combination of model, censoring type, and earnings/transformation of earnings are presented. The variance of the simulated earnings is also decomposed into extensive and intensive portions, to illustrate the effect of different participation patterns on earnings variation, and how transforming earnings with the IHS with different values of θ will affect the observed dynamics. We also examine the auto-covariance matrix of earnings, and how

this changes with the different censoring methods and transformations of earnings.

These results are used to compare the IHS to both the untransformed level, and the logarithm of earnings. The analysis of how the IHS affects the sample statistics, autocovariances, and extensive and intensive variations of earnings with the simulated data will inform the analysis of the empirical data.

3.3.1 Model Description

This model assumes that the unobserved component of annual earnings can be decomposed into an individual specific fixed effect, a transitory shock with some level of persistence, and measurement error. It is is loosely based on other models used previously in the literature (Hyslop, 2001; Lillard and Willis, 1978). Equation (3.3.1) outlines the model and Table 3.1 the parameters used in generating the data,

$$\log(W_{nt}) = w_{nt} = \alpha_n + \delta_{nt} + \epsilon_{nt} \ , \tag{3.3.1a}$$

$$\text{where} \quad \delta_{nt} = \rho\delta_{nt-1} + e_{nt} \ . \tag{3.3.1b}$$

This model is used to generate panel data of annual earnings, where W_{nt} is the annual earnings of individual n for year t, and has a log-normal distribution, while w_{nt} is log(earnings). Each individual's log(earnings) in each time period is made up of three separate components; a time invariant individual specific element α_n, a persistent shock modelled as an AR(1) process δ_{nt}, and an idiosyncratic, serially uncorrelated purely transitory shock ϵ_{nt} that can be thought of as capturing random measurement error. We assume that these components are orthogonal and additively log-normal, such that log(earnings) is normally distributed.

Our focus is analysing the effect that the IHS transformation will have on panel data of annual earnings, with an emphasis on the earnings of female workers. As such, the autocovariance structure of the model is fit to that observed empirically for females. Hyslop (2001)

Table 3.1: *Simulation parameters*

Variable	Value or distribution
Auto-regressive coefficient	$\rho = 0.843$
Individual specific fixed effect	$\alpha_n \sim N(10, 1)$
Measurement error	$\epsilon_{nt} \sim N(0, 0.38)$
Transitory shock	$e_{nt} \sim N(0, 0.04)$
Sample size	$N = 1,000,000$
Time periods	$T = 10$

estimates a similar (although more complex and non-stationary) model for wages, and our Data Generating Process (DGP) is based on his results. Hyslop estimated ρ, the correlation parameter associated with the AR(1) process, as equal to 0.843, with an auto-covariance structure of earnings where 66% of the total variance is associated with the individual specific component, 9% from the persistent shocks, and 24% resulting from measurement error in the first period. To mimick this covariance structure in the simulation, σ_α^2 is normalised to one, and the other variances are calibrated based on this to be $\sigma_\epsilon^2 = 0.38$, and $\sigma_e^2 = 0.04$. The AR(1) component ρ is also fitted based on Hyslop's results and set equal to 0.843.

Two mechanisms are used to censor the data. Here, censoring means that the individual does not participate in the workforce in a period, and has zero earnings. Firstly, non-participation can be based on the individual fixed effects. With this method, individuals in the panel with a sufficiently low α_n will not work in any period. In the context of an ECM, this says that individuals with a fixed effect that results in earnings significantly below their expected earnings given observeables will never work. Secondly, the AR(1) process can be used to determine periods where an individual will not work. This is intended to model individuals choosing not to work in periods where their earnings are significantly lower than their expected earnings, but that they will return to work when their earnings increase. The censoring based on the individual specific effect leads to persistent selection and permanent unemployment, while the AR(1) component introduces temporary unemployment and intertemporal correlation into participation patterns. A combination of these methods

is also used, where temporary unemployment is responsible for half of the workforce non-participation, and permanent unemployment the other half.

A single simulation is performed with $N = 1,000,000$ and $T = 10$. Each of the two censoring methods are applied to the same simulated data set in order to create multiple data sets with different levels of individual participation. Both models of censoring have been applied separately, as well as in conjunction with each responsible for 50% of non-participation. In all cases, a lower bound is specified for each variable based on the percentage of observations to be censored. The lower bounds are specified so that first approximately 10%, and then 25% of all observations will be censored, roughly fitting the participation patterns for male and female workers respectively. This results in seven data sets in total. One with full participation, and then for each level of censoring there is workforce non-participation based on the individual specific component, the persistent shock, or a combination of both.

In order to compare the IHS and logarithmic transformations, each is applied to the simulated data. As illustrated in Chapter 2, when dealing with censored data where there is a large mass of observations at zero, θ estimation is unreliable. As such, we do not estimate θ dirrectly, but apply the IHS with $\theta = \{1, 0.1, 0.001\}$. This allows us to analyse how the IHS performs with a range of potential values, when Maximum Likelihood Estimation (MLE) may not be feasible.

When applying the logarithm to earnings data where zero observations are present, the researcher has a choice as to how to proceed. They can exclude any individual that has zero earnings in one or more periods from the data set entirely, resulting in a balanced panel where every individual works in every period. Alternatively, they can remove any periods where an individual does not work, while retaining the periods where they do work, resulting in an unbalanced panel. The unbalanced panel retains a larger fraction of the observations, but can make the use of some models and estimators more complicated. The balanced panel, on the other hand, results in a data set that is easier to work with, but removes more observations, potentially increasing the effects of selection bias. Along with the IHS

we generate both balanced and unbalanced panels of log(earnings) using the simulated data, with the balanced and unbalanced panels being identical when there is no censoring or when the censoring is solely based on the individual specific fixed effect α_n.

For each simulation we present sample statistics comparing the results of the untransformed level, the balanced and unbalanced logarithm, and IHS, as well as a decomposition of the variance of annual earnings showing the portion of variance from extensive versus intensive margins for different years of workforce participation, and the auto-covariance matrices for earnings. This analysis allows us to compare the logarithmic and IHS functions, and will inform our work with the empirical data in Section 3.4.

3.3.2 Summary Statistics

Table 3.3 contains the results and summary statistics for the stationary model simulation. The table is split into three sections, the first contains the uncensored data, where earnings are observed for every individual in every time period. The second and third sections contain the data with 10% and 25% censoring respectively. When the data are censored, the left column indicates which variables is used to select the individuals that do not work, either the AR(1) persistent shock, the individual specific permanent component α_n, or a combination of the two.

The table presents are a range of summary statistics that can be used to understand the distribution of earnings. Results are included for untransformed earnings (in thousands), log(earnings), and IHS(earnings). For IHS(earnings) three values of θ are used to generate the results, $\theta = \{1, 0.1, 0.001\}$, with IHS(earnings,$\theta = 0.001$) measured in thousands. For each of these transformations of earnings, $\overline{Y}$ presents the mean value over all individuals, $\text{Var}(Y_{it})$ is the mean cross-sectional variance, $\text{Var}(\overline{Y_i})$ is the mean intertemporal variance an individual experiences over the sample, $\text{Cor}(Y_2, Y_1)$ and $\text{Cor}(Y_{10}, Y_1)$ are the first and ninth order auto-correlations respectively, and Skew measures the level of skewness in the data.

When the data is not censored, each model uses the same observations. When censoring

Table 3.2: *Fraction of sample used in the balanced and unbalanced panels of log(earnings)*

Censoring method	Balanced panel	Unbalanced panel
	10% censoring	
AR(1)	0.66	1
α_n	0.90	0.90
Combination	0.76	0.95
	25% censoring	
AR(1)	0.38	0.97
α_n	0.75	0.75
Combination	0.53	0.87

is introduced, the untransformed level of earnings as well as IHS(earnings) still make use of all of the observations, but log(earnings) is forced to exclude the zero earnings observations, either through a balanced or unbalanced panel. For censoring based on only α_n, this results in identical balanced and unbalanced panels of log(earnings). When censoring is based on either the AR(1) process, or a combination of both the AR(1) process and α_n, the balanced panel has a lower sample size than the unbalanced panel. Table 3.2 shows the fraction of the sample used for both the balanced and unbalanced panels with the different levels and types of censoring. The unbalanced panel uses different individuals in each time period, and pairwise complete observations to calculate the auto-correlations. Table 3.4 show the fraction of the sample used in each time period or period pair for the unbalanced panel models with 25% censoring, illustrating the participation patterns of the sample. The 10% censored table is included in Appendix B, the participation patterns are similar to the 25% case, but with more individuals working in each time period.

The way in which $\overline{Y}$ changes in Table 3.3 when the IHS is applied illustrates the main difference between it and the logarithm. The uncensored data produces the result expected based on Chapter 2, the mean of IHS(earnings, $\theta = 1$) approaches the mean of log(2) + log(earnings). The introduction of censoring changes the results significantly, with $\overline{log(Y)}$

Table 3.3: *Simulation sample statistics*

		$\overline{Y}$	Var(Y_{it})	Var($\overline{Y_i}$)	Cor(Y_2, Y_1)	Cor(Y_{10}, Y_1)	Skew
	Levels*	47.0	7920	3580	0.57	0.50	12.43
	Log	10.00	1.52	0.44	0.74	0.68	0.00
	IHS($\theta = 1$)	10.69	1.52	0.44	0.74	0.68	0.00
	IHS($\theta = 0.1$)	83.90	151.77	43.73	0.74	0.68	0.00
	IHS($\theta = 0.001$)*	3.79	1.48	0.427	0.74	0.68	0.07
	10% censoring						
	Levels*	44.7	7920	3710	0.57	0.47	12.43
	Balanced log	10.14	1.47	0.43	0.73	0.69	0.01
AR(1)	Unbal log	10.07	1.48	0.43	0.73	0.69	0.01
	IHS($\theta = 1$)	9.69	11.77	7.20	0.61	0.16	-2.12
	IHS($\theta = 0.1$)	76.15	778.43	457.11	0.62	0.20	-1.85
	IHS($\theta = 0.001$)*	3.48	2.65	1.24	0.68	0.37	-0.59
	Levels*	46.5	7960	3570	0.58	0.51	12.34
	Log	10.19	1.23	0.44	0.68	0.60	0.21
	IHS($\theta = 1$)	9.80	11.78	0.39	0.97	0.96	-2.20
α_n	IHS($\theta = 0.1$)	77.26	774.35	39.36	0.95	0.94	-1.97
	IHS($\theta = 0.001$)*	3.58	2.53	.390	0.86	0.83	-0.74
	Levels*	45.8	7950	3640	0.57	0.49	12.38
	Balanced log	10.20	1.29	0.43	0.69	0.64	0.18
Combo	Unbal log	10.15	1.30	0.43	0.69	0.64	0.18
	IHS($\theta = 1$)	9.78	11.55	4.03	0.77	0.54	-2.20
	IHS($\theta = 0.1$)	77.04	761.03	262.10	0.77	0.55	-1.95
	IHS($\theta = 0.001$)*	3.55	2.53	0.840	0.75	0.58	-0.68
	25% censoring						
	Levels*	40.1	7730	3880	0.56	0.41	12.66
	Balanced log	10.26	1.46	0.42	0.73	0.69	0.00
AR(1)	Unbal log	10.16	1.46	0.42	0.73	0.69	0.01
	IHS($\theta = 1$)	8.14	23.17	13.79	0.64	0.15	-0.99
	IHS($\theta = 0.1$)	64.09	1480	866.28	0.64	0.16	-0.91
	IHS($\theta = 0.001$)*	2.96	3.99	2.11	0.67	0.25	-0.31
	Levels*	44.8	8080	3560	0.58	0.52	12.11
	Log	10.42	1.05	0.44	0.62	0.54	0.26
	IHS($\theta = 1$)	8.34	23.98	0.33	0.99	0.98	-1.04
α_n	IHS($\theta = 0.1$)	66.09	1540	32.81	0.98	0.98	-0.98
	IHS($\theta = 0.001$)*	3.16	4.11	0.326	0.93	0.91	-0.51
	Levels*	43.3	7950	3740	0.57	0.46	12.35
	Balanced log	10.40	1.14	0.43	0.65	0.60	0.25
Combo	Unbal log	10.32	1.15	0.42	0.65	0.60	0.25
	IHS($\theta = 1$)	8.43	22.67	7.77	0.79	0.52	-1.11
	IHS($\theta = 0.1$)	66.68	1450	495.23	0.78	0.53	-1.04
	IHS($\theta = 0.001$)*	3.14	3.91	1.33	0.77	0.54	-0.50

** measured in thousands*

increasing for all methods and levels of censoring as individuals that do not work are cut from the sample, either entirely in the case of the balanced panel, or on a case-wise basis for the unbalanced. The increase in $\overline{log(Y)}$ is due to the non-random selection; individuals with earnings at the lower end of the distribution (or at least one component of the residual of their earnings is at the low end of its distribution) are not working, increasing the average of

log(earnings). For the IHS on the other hand, the zero earnings observations are still included when calculating the mean. This leads to mean(IHS(earnings) falling, with the magnitude of the decrease increasing with high levels of censoring. In this way, the IHS incorporates the period of unemployment when estimated the average earnings for an individual in the sample, while log(earnings) ignores the periods where individuals do not work.

In the uncensored data, both measures of earnings variance are identical between log(earnings) and IHS(earnings, $\theta = 1$). As θ decreases both estimates of variance increase, reflecting that the IHS nests the untransformed earnings as $\theta \to 0$. Interestingly, even though the different values of θ have different cross-sectional and intertemporal variances, the ratio of $\text{Var}(\overline{Y_i})$ to $\text{Var}(Y_{it})$ is very similar between each of the IHS transformations and the logarithm, being approximately 0.28. This changes with the introduction of censoring. All of the estimates for both of the variances of log(earnings) have either decreased or stayed approximately the same. The impact is relatively small when censoring is based on the AR(1) process, and there is a large decrease in cross-sectional variance when censoring is based on α_n. For the IHS of earnings the cross-sectional variance increases dramatically when censoring is introduced, and the increase is larger with a higher fraction of the sample censored. For example, the cross-sectional variance for IHS(earnings, $\theta = 1$) when the data was uncensored was 1.52, and this increased to approximately 23 for each of the censoring methods when 25% censoring was introduced. When censoring is based on α_n there is a small decrease in the intertemporal variance for IHS(earnings), where as when it is based on the AR(1) component it increases by a large amount.

For both the first and ninth order auto-correlations in the uncensored data, Table 3.3 shows that the logarithm and IHS of earnings have identical results for all three values of θ. When censoring is introduced, the auto-correlations of log(earnings) are relatively stable. When censoring is based on the AR(1) component, the logarithm produces almost identical auto-correlations with both levels of censoring. When censoring is based on α_n the auto-correlations of log(earnings) decrease, for example with the 25% level of censoring they fall

from 0.74 to 0.62 and 0.68 to 0.54 for the first and ninth order auto-correlations respectively.

There are larger changes to the auto-correlations of IHS(earnings). When censoring is based on the AR(1) process, the first order auto-correlations decrease by about 0.1 when the data is censored at the 25% level. The ninth order auto-correlations however are much lower than both the uncensored IHS of earnings, or the logarithm of censored earnings. When the data is censored by 25%, $\text{Cor}(Y_{10}, Y_1)$ falls to 0.15, 0.16, and 0.25 for θ=1, 0.1, and 0.001 respectively. When censoring is instead based on the α_n the opposite occurs, and the auto-correlations of IHS(earnings) become much higher. For example, when 25% censoring is applied the first order auto-correlations are all over 0.9, with the lowest being 0.93 when $\theta = 0.001$. The ninth order auto-correlations are slightly lower, but the decrease is smaller than what occurs in both the uncensored and censored based on α_n data.

As noted previously, the simulated earnings data is log-normally distributed so as to mimic the observed empirical distribution (Mayer, 1960). This results in uncensored data that has a long right tail, represented in Table 3.3 by the positive level of skewness present in the untransformed earnings. Both the logarithm and the IHS remove the skewness present in the uncensored data. The logarithm continues to perform well when censoring is introduced, still removing the skewness almost entirely when censoring is based on the AR(1) component, and reducing it to a low level when censoring is based on α_n. With the IHS, the introduction of censoring leads to a distribution of IHS(earnings) with negative skewness. Censoring based on α_n leads to a more negatively skewed distribution than when it is based on the AR(1) component, but the difference is relatively small. As θ decreases the level of skewness in the distribution grows larger, although it is still negative with all values used here. Based on Chapter 2, it seems that if θ continued to decrease, the level of skewness in the data would approach that seen in the untransformed earnings.

Table 3.4: *Unbalanced log(earnings) fraction of sample used:* 25% *censoring*

T	1	2	3	4	5	6	7	8	9	10
				Censored based on AR(1) component						
1	0.77									
2	0.70	0.77								
3	0.67	0.70	0.77							
4	0.65	0.67	0.70	0.77						
5	0.64	0.65	0.67	0.70	0.77					
6	0.63	0.64	0.65	0.67	0.70	0.77				
7	0.62	0.63	0.64	0.65	0.67	0.70	0.77			
8	0.61	0.62	0.63	0.64	0.65	0.67	0.70	0.77		
9	0.61	0.61	0.62	0.63	0.64	0.65	0.67	0.70	0.77	
10	0.60	0.61	0.6	0.62	0.63	0.64	0.65	0.67	0.70	0.77
				Censored on α_i *and transitory shock*						
1	0.77									
2	0.73	0.77								
3	0.71	0.73	0.77							
4	0.70	0.71	0.72	0.77						
5	0.69	0.70	0.71	0.72	0.77					
6	0.69	0.70	0.70	0.71	0.73	0.77				
7	0.69	0.69	0.70	0.70	0.71	0.73	0.77			
8	0.68	0.69	0.69	0.70	0.70	0.71	0.73	0.77		
9	0.68	0.68	0.69	0.69	0.70	0.70	0.71	0.73	0.77	
10	0.68	0.68	0.68	0.69	0.69	0.70	0.70	0.71	0.73	0.77

3.3.3 Variance Decomposition

As outlined in Section 3.2, the variance of annual earnings can be decomposed into intensive and extensive margin components. Equation (3.3.2) can be used to estimate the fraction of earnings variation that is attributable to movements into and out of the workforce, and the fraction that is due to changes in earnings while remaining in the workforce. In this section, the variance decomposition is applied to the simulated data. The variance of untransformed earnings and IHS(earnings) with each level of θ used in Section 3.3.2 are decomposed, and log(earnings) is not. While $p_i v_{ic} + p_i(1-p_i)(\overline{Y}_i^c)^2$ would function with log(earnings), it is derived from $\sum_{i=1}^{T}(Y_{it} - \bar{Y}_i)^2$ which will not work with log(earnings) in

the presence of zero observations.

$$v_i = \frac{1}{T}\sum_{i=1}^{T}(Y_{it} - \bar{Y}_i)^2 = p_i v_{ic} + p_i(1-p_i)(\overline{Y}_i^c)^2 \ . \tag{3.3.2}$$

The variance decomposition is applied to the simulated data when they are censored based on the AR(1) component, or the combination of the AR(1) and α_n components. It is not applied to the uncensored data, as in that model all individuals work in every period, so all variation in earnings comes from the intensive margin. Likewise, when the data is censored based on only the individual specific fixed effect α_n, the population is bi-modal with some individuals working in every period, and some never working. This again results in earnings variation caused solely by the intensive margin, and therefore we do not apply the decomposition to this data.

In each case where the decomposition is applied, we break the sample down by the number of years worked. For each number of years, we report n, which is the fraction of the population that worked that many years, *Ext Margin* which is the fraction of earnings variation coming from the extensive margin,[1] average earnings in the years worked, [2] and the mean standard deviation of earnings over the entire ten year sample. The average earnings reported are only for the years that the individual works, while the mean standard deviation includes the effect of moving to/from zero earnings.

Table 3.5 reports the results of decomposing the variance of untransformed annual earnings and IHS(earnings) using the simulated data where workforce participation is based on the AR(1) component. The top panel contains the results for untransformed earnings, with both 10% and 25% censoring. The second panel contains the results for IHS(earnings, $\theta = 1$), and the bottom two rows contain summaries of the mean results for IHS(earnings, $\theta = 0.1$) and IHS(earnings, $\theta = 0.001$). As they use the same raw data, each model with censoring based on only on the AR(1) component has an identical fraction of individuals working in

[1] We do not include the fraction from the intensive margin, but it is implicitly equal to 1 − (Ext Margin)
[2] Mean annual earnings conditional on working

each year. When the 10% level of censoring is applied, most (66%) of the individuals in the simulated data work in every period. The fraction of individuals working for a given number of years decreases as the number of years worked falls. Only 11% of the sample works for nine periods, and approximately 3% of the individuals in the sample work for either one, two, or three periods. When the level of censoring is increased to 25% the fraction of individuals working in every period decreases, changing from 66% to 38%, while the proportion of individuals working for the remaining numbers of years increases. While previously no one worked in zero periods, now 3% of the sample never work.

Table 3.5: *Variance decomposition: Selection on AR(1)*

	Years worked	n^1	Extensive margin	Mean earnings[2]	Avg SD[3]	n^1	Extensive margin	Mean earnings[2]	Avg SD[3]
		10% Censoring				25% Censoring			
Earnings	10	0.66	0.00	53.25	66.26	0.38	0.00	59.44	73.28
	9	0.11	0.17	41.96	53.67	0.13	0.17	50.40	63.64
	8	0.07	0.28	40.27	52.6	0.10	0.29	48.26	61.56
	7	0.05	0.38	38.46	49.30	0.08	0.38	46.62	59.92
	6	0.04	0.46	37.09	47.43	0.07	0.47	45.57	56.48
	5	0.03	0.52	35.60	42.43	0.06	0.52	43.60	52.73
	4	0.02	0.55	35.68	41.95	0.05	0.57	42.57	47.75
	3	0.01	0.64	33.49	33.17	0.04	0.63	41.66	42.31
	2	0.01	0.69	31.62	25.50	0.03	0.66	39.50	33.17
	1	0.01	1.00	29.87	17.26	0.03	1.00	37.80	23.43
	0	0.00	-	-	-	0.03	-	-	-
	Average	-	0.12	49.70	61.07	-	0.26	53.45	63.09
IHS($\theta = 1$)	10	0.66	0.00	10.84	0.66	0.38	0.00	10.96	0.65
	9	0.11	0.96	10.61	3.43	0.13	0.96	10.80	3.49
	8	0.07	0.98	10.57	4.51	0.10	0.98	10.76	4.59
	7	0.05	0.99	10.54	5.14	0.08	0.99	10.73	5.23
	6	0.04	0.99	10.50	5.47	0.07	0.99	10.71	5.58
	5	0.03	0.99	10.48	5.56	0.06	0.99	10.67	5.67
	4	0.02	0.99	10.46	5.44	0.05	0.99	10.65	5.54
	3	0.01	0.99	10.41	5.07	0.04	0.99	10.63	5.17
	2	0.01	1.00	10.37	4.40	0.03	1.00	10.59	4.49
	1	0.01	1.00	10.33	3.29	0.03	1.00	10.55	3.36
	0	0.00	-	-	-	0.03	-	-	-
	Average	-	0.33	10.76	2.69	-	0.60	10.85	3.76
IHS($\theta = 0.1$)			0.33	84.62	21.42	-	0.59	85.47	29.83
IHS($\theta = 0.001$)			0.29	3863.00	1113.55	-	0.55	3946.80	1473.09

[1] *Fraction of sample working this many years*
[2] *Average earnings conditional on working*
[3] *Average standard deviation of earnings over the ten time periods.*

Mean annual earnings generally decreases as the number of years worked falls, with a small increase when moving from five to four years. For both levels of censoring the largest decrease occurs when moving from working in all periods, to only working in nine of the

ten. In that case, with 10% censoring mean earnings conditional on working decreases from $53,250$ to $41,960$.

As the level of censoring increases, so do mean earnings for every level of workforce participation. This is due to the censoring being based on the AR(1) component. As the level of censoring increases, observations with the lowest levels of the transitory shock move from working to non-participation in a given year, and all else being equal these observations will have lower earnings on average. This means individuals with below average earnings given their number of years worked will move down/work in fewer periods.

By definition, the extensive margin contributes nothing to earnings variation for individuals that work every year. Likewise, individuals that work for only a single year have variance of earnings that is made up of only extensive margin variation. Between these extremes, the fraction of earnings variance that is caused by changes to the extensive margin is relatively similar for both levels of censoring. The largest increase in the portion of variance attributable to the extensive margin occurs when moving from working in every year to working in only nine, where it increases from making up 0% to 17% of earnings variation. After this it increases monotonically as the number of years worked falls, with the size of the increase generally decreasing with years worked.

The average standard deviation of untransformed earnings monotonically decreases as the number of years worked falls for both levels for censoring. In each case, the average standard deviation is larger than mean earnings for the majority of years worked. For 10% censoring only three, two, and one year(s) worked have standard deviations of earnings lower than mean earnings. When 25% of the sample is censored, the average standard deviation is lower than mean earnings for those working only one or two years in the sample. The large standard deviation of earnings reflects the long right tailed nature of the log-normal earnings DGP, with outliers having a large impact on the results.

Table 3.5 also presents the results of decomposing IHS(earnings, $\theta = 1$) using the simulation censored based on the AR(1) component. As noted previously, the fraction of the sample

working for any given number of years is identical, as the same simulated data is used. The portion of the variance of IHS(earnings, $\theta = 1$) attributable to extensive margin variation is very different to that estimated using untransformed earnings, and is identical for both levels of censoring. [3] Again, for individuals working in every period all earnings variation is due to changes in the intensive margin, and the largest change comes when moving from working in every period to working in only nine periods. There the fraction of earnings variance attributable to the extensive margin increases from 0% to 96%. This large change is due to the size of the average standard deviation in relation to mean earnings. For example, with 10% censoring the mean annual IHS($earnings, \theta = 1$) for an individual working in all ten periods is 10.84 while the average standard deviation is only 0.66. When censoring is introduced, the extensive margin contribution from moving from IHS(earnings,$\theta = 1$) = 10.84 to zero, dominates the intensive margin contribution to the variance of earnings. In comparison, the decomposition of untransformed earnings showed a much smaller contribution by the extensive margin due to the much larger standard deviations.

Mean IHS(earnings, $\theta = 1$) shows a similar trend to mean earnings in the untransformed decomposition. For both levels of censoring, it is highest for individual's working in every year, and monotonically decreases as the number of years worked falls. As the level of censoring increases, so do mean IHS(earnings, $\theta = 1$).

As noted previously, the average standard deviation of IHS(earnings, $\theta = 1$) over the sample is very different to that observed with untransformed earnings. The difference in standard deviations is driven by the IHS's dampening effect on extreme values, as illustrated in Chapter 2, and this counteracts the long right tail present in the untransformed data. While in the untransformed decomposition the average standard deviation was highest for individuals that worked every year and decreased as years worked fell, for IHS(earnings) it is at its lowest point for individuals working in every period. The shift from working ten to nine years produces a large increase in the standard deviation of IHS(earnings, $\theta = 1$),

[3] For this level of precision.

from 0.66 to 3.43 with 10% censoring. With both levels of censoring, the average standard deviation of IHS(earnings, $\theta = 1$) is very similar and follows the same general trend. It first increases as the number of years worked falls, reaching its largest point for those working in five years. Below five years worked the standard deviation falls again, approximately reversing its original increases, although for those working in only a single year the average standard deviation is still much larger than for those working every year. This decrease is caused by the lack of variation in periods where an individual does not work overriding the effect of the extensive margin changes.

The bottom two rows of Table 3.5 show the summary results of decomposing IHS(earnings, $\theta = 0.1$) and IHS(earnings, $\theta = 0.001$) respectively. These results are broadly similar to those for IHS(earnings, $\theta = 1$). In each of these cases, mean IHS(earnings) for years in which an individual works decreases as the years of workforce participation fall. As θ becomes smaller, mean earnings increases, and the decreasing the number of years worked leads to a smaller increase in the fraction of earnings variance attributed to the extensive margin. Likewise, the standard deviation of earnings over the simulation increases as θ approaches zero, but still follows the same pattern as for $\theta = 1$. There is a large increase when moving from working in all periods to working in only nine, and it continues to increase and plateaus at five years worked before decreasing.

Table 3.6 decomposes the variance of untransformed earnings and IHS(earnings) using the sample censored based on both the AR(1) transitory shock and the permanent, individual specific effect α_n. The introduction of censoring based on the individual specific fixed effect means that there will be permanent unemployment for a portion of the population. The censoring is split evenly between the AR(1) and individual specific components.

With the combination of censoring methods, the proportion of the population that works for each number of years is higher than observed in Table 3.6. In the case of 10% censoring, the proportion working in every period rises from 66% to 76%. The proportion working in zero periods is also much higher, rising from 0% to 5%, and 3% to 13%, for 10% and 25%

Table 3.6: *Variance decomposition: Selection on AR(1) and α_i*

	Years worked	n^1	Extensive margin	Mean earnings^2	Avg SD^3	n^1	Extensive margin	Mean earnings^2	Avg SD^3
		10% Censoring				25% Censoring			
Earnings	10	0.76	0.00	53.20	65.12	0.53	0.00	61.17	72.18
	9	0.08	0.16	39.96	50.89	0.11	0.17	49.21	59.33
	8	0.04	0.28	37.79	49.50	0.07	0.28	46.70	57.18
	7	0.03	0.38	35.88	44.72	0.05	0.38	45.34	55.41
	6	0.02	0.44	35.11	44.16	0.04	0.46	43.73	52.63
	5	0.01	0.53	33.82	38.99	0.03	0.51	41.97	48.58
	4	0.01	0.57	31.96	34.64	0.02	0.56	41.15	45.72
	3	0.00	0.62	30.65	29.58	0.01	0.64	39.83	37.95
	2	0.00	0.71	29.88	24.54	0.01	0.67	36.83	28.44
	1	0.00	1.00	28.90	17.12	0.01	1.00	36.02	20.78
	0	0.05	-	-	-	0.13	-	-	-
	Average	-	0.06	50.79	61.97	-	0.14	56.62	65.57
IHS($\theta = 1$)	10	0.76	0.00	10.89	0.66	0.53	0.00	11.10	0.66
	9	0.08	0.96	10.62	3.43	0.11	0.96	10.89	3.51
	8	0.04	0.98	10.57	4.51	0.07	0.98	10.84	4.62
	7	0.03	0.99	10.53	5.14	0.05	0.99	10.82	5.27
	6	0.02	0.99	10.50	5.47	0.04	0.99	10.78	5.61
	5	0.01	0.99	10.47	5.56	0.03	0.99	10.75	5.70
	4	0.01	0.99	10.42	5.42	0.02	0.99	10.72	5.57
	3	0.00	0.99	10.38	5.04	0.01	0.99	10.70	5.19
	2	0.00	1.00	10.37	4.40	0.01	1.00	10.65	4.51
	1	0.00	1.00	10.32	3.28	0.01	1.00	10.61	3.37
	0	0.05	-	-	-	0.13	-	-	-
	Average	-	0.20	10.84	2.06	-	0.39	11.02	2.99
IHS($\theta = 0.1$)			0.19	85.39	16.62	-	0.38	87.13	23.86
IHS($\theta = 0.001$)			0.17	3937.35	940.74	-	0.35	4109.41	1236.93

[1] *Fraction of sample*
[2] *Average earnings conditional on working*
[3] *Average standard deviation of earnings over the ten time periods.*

censoring respectively. This approximately matches the expected level, based on the fixed component α_n making up half of the censoring. Apart from ten and zero years worked, all other combinations of workforce participation have lower fractions of the sample than in the previous model, due to a larger fraction working in either every or none of the periods.

For untransformed earnings, the fraction of earnings variation caused by extensive margin changes is approximately the same in this sample when compared to earnings variation in the sample censored solely on the AR(1) component. Both levels and methods of censoring produce very similar extensive margin effects with untransformed earnings.

Mean earnings in the sample created using the combination of the two censoring methods is lower for both 10% and 25% censoring for all numbers of years worked, except for those working every year in the 25% censored model. Again, there is a large drop in mean earnings

when moving from working in every period to working in nine, and after this mean earnings monotonically decrease as years worked falls. Mean earnings are higher for every number of years worked in the 25% censored sample. The average standard deviation of earnings is lower when compared to the sample censored on the AR(1) component only, and follows a similar pattern in decreasing monotonically as years worked falls.

Table 3.6 also presents the results of decomposing IHS(earnings, $\theta = 1$) in the model censored based on both the AR(1) and individual specific components. For all potential years worked, the fraction of IHS(earnings, $\theta = 1$) variance attributable to extensive margin variation is identical to the earlier model censored solely on the AR(1) component, for both levels of censoring. Due to the changes in the fraction of the population working each number of years, this results in a decrease in the average effect of extensive margin variation, from 0.33 and 0.60 in the original model, to 0.20 and 0.39 for the samples with 10% and 25% censoring respectively.

For the 10% censored sample of IHS(earnings, $\theta = 1$), mean annual earnings conditional on working are approximately equal to the model censored based on only the individual specific fixed effect. The sample with 25% censoring has consistently higher mean earnings that the equivalent model censored based on the AR(1) component. Based on this, and a greater fraction of the population working in every year, the average mean earnings over all number of years worked is higher with both levels of censoring, when censoring is based on both the AR(1) and individual specific components.

The average standard deviation of IHS(earnings, $\theta = 1$) in the sample censored based on the combination of censoring types is very close to those in Table 3.5, where the data was censored based on only the AR(1) component. While the average standard deviations are very similar for the different levels of workforce participation, the differences in the distribution of workers over years worked leads to very different results when looking at the standard deviation over all years. In this sample more individuals work in every period, and more work in none, and these are the two levels of workforce participation with the lowest

standard deviation of IHS(earnings). For both levels of censoring, over all years the average standard deviation of IHS(earnings, $\theta = 1$) is lower when the data is censored based on both the AR(1) and individual specific components.

The final two rows of Table 3.6 list a summary of the results for the decompositions of IHS(earnings, $\theta = 0.1$) and IHS(earnings, $\theta = 0.001$) respectively, each censored based on the combination of individual specific and transitory parameters. The results here follow those for IHS(earnings, $\theta = 1$), in that they closely match the results from the earlier model where censoring was applied based solely on the AR(1) component. The extensive margin contributions to the variance of IHS(earnings) are almost identical for any particular number of years worked, while over the entire sample the extensive margin portion is lower due to the larger share of individuals working in every period. Likewise, mean earnings per year for individuals working any particular number of years are very similar between the different censoring methods, being slightly larger when the data is censored based on the combination of parameters.

There are a few interesting results here. First, the application of the IHS does seem to significantly reduce the impact of extreme values. The standard deviations are much smaller relative to mean earnings in all cases with the IHS, and the extensive margin effects are much stronger. This shows that application of the IHS changes the estimated effects of extensive and intensive margin changes on lifetime earnings. However, the fraction of earnings variation attributed to the extensive margin does depend of the value of θ used and, as shown in Chapter 2, estimation is unreliable in the presence of a clump at zero.

3.3.4 Auto Covariance Matrices

This section examines the variances, auto-covariances, and correlations produced using data generated by the stationary model, with the different levels and methods of censoring. In each case, auto-covariance matrices are produced of untransformed earnings, balanced and unbalanced panels of log-earnings, and IHS(earnings). Due to the large number of tables

produced, not all of them are included in this section, with some excluded tables available in Appendix B.

Table 3.7 is the auto-covariance matrix of log(earnings) with the uncensored simulated data. The auto-covariance matrices are not presented for the IHS with the uncensored data, as they are almost identical to the logarithm, the only difference being higher variances and co-variances for $\theta = 0.1$ and $\theta = 0.001$, as is evident in Table 3.3. The main diagonal elements of the table are the variances of log(earnings) in each time period. For example, the variance of log(earnings) in time period one is 1.516, it then increases to 1.520 in time period two as we move down the diagonal. The variances are relatively stable over time, due to the stationary nature of the model used.

Table 3.7: *Auto-covariance matrix of log(earnings).*

T	1	2	3	4	5	6	7	8	9	10
1	1.516 (0.0021)	0.74	0.72	0.71	0.71	0.70	0.69	0.69	0.68	0.68
2	1.12 (0.0019)	1.520 (0.0021)	0.74	0.72	0.71	0.71	0.70	0.69	0.69	0.68
3	1.10 (0.0019)	1.12 (0.0019)	1.517 (0.0021)	0.74	0.72	0.71	0.71	0.70	0.69	0.69
4	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.517 (0.0021)	0.74	0.72	0.71	0.71	0.70	0.69
5	1.07 (0.0019)	1.09 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.518 (0.0021)	0.74	0.72	0.71	0.71	0.70
6	1.06 (0.0018)	1.07 (0.0019)	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.518 (0.0021)	0.74	0.72	0.71	0.71
7	1.05 (0.0018)	1.06 (0.0019)	1.07 (0.0019)	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.519 (0.0021)	0.74	0.72	0.71
8	1.04 (0.0018)	1.05 (0.0018)	1.06 (0.0018)	1.07 (0.0019)	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.517 (0.0021)	0.74	0.72
9	1.04 (0.0018)	1.04 (0.0018)	1.05 (0.0018)	1.06 (0.0019)	1.07 (0.0019)	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.519 (0.0022)	0.74
10	1.03 (0.0018)	1.04 (0.0018)	1.04 (0.0018)	1.05 (0.0018)	1.06 (0.0018)	1.07 (0.0019)	1.08 (0.0019)	1.10 (0.0019)	1.12 (0.0019)	1.518 (0.0021)

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, and standard errors in parentheses.

The upper triangle of the matrix contains the auto-correlation between earnings in each pair of years. For example, examining the first row shows the auto-correlation between earnings in period one and each other time period, e.g. the auto-correlation between earnings in $T = 1$ and $T = 2$ is 0.74. Moving further along the row, each element contains the auto-correlation of progressively higher orders. Moving down diagonally from any element in this

row are all of the correlations of a certain order. For example, the first order auto-correlation for $T = 1$ is 0.74, and moving down diagonally from that element of the table are all of the other first order auto-correlations.

The lower triangle contains the auto-covariances between earnings in the different time periods. For example, the first column contains the auto-covariance between earnings in $T = 1$ and each other time period. Similar to the auto-correlation, moving diagonally down from any element of the first row (after the first element, which is the variance) are all of the auto-correlations of that order. The standard errors for each estimate are in parentheses, and in the case of 3.7 every element is significantly different from zero.

The results in Table 3.7 are as expected given the DGP and lack of censoring. The variance of log(earnings) stays approximately constant, due to the underlying model's stationarity. Given the model used (see Equation (3.3.1)), when comparing two time periods, say t and t+k, as k increases and the two periods become more distant from each other we expect $\text{Cor}\left(\log(\text{earns}_t), \log(\text{earns}_{t+k})\right)$ to approach the fraction of variance attributable to α_n, while $\text{Cov}\left(\log(\text{earns}_t), \log(\text{earns}_{t+k})\right)$ tends to $\text{Var}(\alpha_n)$. Both of these occur here, the co-variance approaches one as the periods examined are further apart, and the correlation approaches 0.66, the fraction of variance attributable to the individual specific component. Equation (3.3.3a) shows why the auto-correlation approaches the fraction of variation attributable to the individual specific component in this case,

$$\begin{aligned}
\text{Cor}(y_t, y_{t+k}) &= \frac{\text{Cov}(y_t, y_{t+k})}{\sigma_t \sigma_{t+k}} \ , && \text{(3.3.3a)} \\
&= \frac{\text{Cov}(y_t, y_{t+k})}{\sigma_y^2} \ , && \text{(3.3.3b)} \\
&= \frac{\text{Cov}(\alpha_i + \delta_t + \epsilon_t, \alpha_i + \delta_{t+k} + \epsilon_{t+k})}{\sigma_y^2} \ , && \text{(3.3.3c)} \\
&= \frac{\text{Var}(\alpha)}{\sigma_y^2} \ . && \text{(3.3.3d)}
\end{aligned}$$

So here, the positive and persistent auto-correlations indicate that there is an individual

specific fixed effect present in the simulated earnings data.

Previously, the entire sample was used to produce each auto-covariance matrix, but the introduction of censored data changes this. While the IHS is defined for zero values, and thus makes use of the entire sample, as detailed in Chapter 2 the logarithm is not. Auto-covariance matrices are created for log(earnings) based on both balanced and unbalanced panels. In this case, for both censoring on the AR(1) component, and a combination of that with the individual specific fixed effects, the results were almost identical between the balanced and unbalanced panels, so results are presented for only the balanced panels. As the substantive interest of this piece is female workers, the analysis of the auto-covariance matrices will focus on the samples with 25% censoring, as this best matches the participation rates for females.

Table 3.8: *Auto-covariance matrix of log(earnings):* 25% *censoring**

T	1	2	3	4	5	6	7	8	9	10
1	1.458 (0.0033)	0.73	0.72	0.71	0.70	0.70	0.70	0.69	0.69	0.69
2	1.06 (0.0029)	1.460 (0.0033)	0.73	0.72	0.71	0.71	0.70	0.70	0.70	0.69
3	1.05 (0.0029)	1.06 (0.0029)	1.456 (0.0033)	0.73	0.72	0.71	0.71	0.70	0.70	0.70
4	1.03 (0.0029)	1.05 (0.0029)	1.06 (0.0029)	1.454 (0.0033)	0.73	0.72	0.71	0.70	0.70	0.70
5	1.03 (0.0029)	1.04 (0.0029)	1.05 (0.0029)	1.06 (0.0029)	1.459 (0.0033)	0.73	0.72	0.71	0.70	0.70
6	1.02 (0.0029)	1.03 (0.0029)	1.04 (0.0029)	1.04 (0.0029)	1.06 (0.0029)	1.456 (0.0033)	0.73	0.72	0.71	0.71
7	1.02 (0.0029)	1.02 (0.0029)	1.03 (0.0029)	1.03 (0.0029)	1.04 (0.0029)	1.06 (0.0029)	1.457 (0.0033)	0.73	0.72	0.71
8	1.01 (0.0029)	1.02 (0.0029)	1.02 (0.0029)	1.03 (0.0029)	1.03 (0.0029)	1.04 (0.0029)	1.06 (0.0029)	1.455 (0.0033)	0.73	0.72
9	1.01 (0.0029)	1.02 (0.0029)	1.02 (0.0029)	1.02 (0.0029)	1.03 (0.0029)	1.04 (0.0029)	1.05 (0.0029)	1.06 (0.0029)	1.457 (0.0033)	0.73
10	1.01 (0.0029)	1.01 (0.0029)	1.02 (0.0029)	1.02 (0.0029)	1.02 (0.0029)	1.03 (0.0029)	1.03 (0.0029)	1.05 (0.0029)	1.06 (0.0029)	1.462 (0.0033)

** Censored based on the AR(1) component.*
Diagonal shows the variance of wages in each period, lower triangle is co-variances, upper triangle is correlations, standard errors in parenthesis.

Table 3.8 presents the auto-correlation matrix for the balanced log with 25% censoring based on the AR(1) component. As stated previously, the results are almost identical for

the unbalanced panel. The variance of log(earnings) in each period is slightly lower than in the uncensored sample. reflecting that removing individuals with lower levels of earnings has narrowed the overall distribution of log(earnings). Most interesting in this case, the auto-covariances and correlations are very similar to those observed in the uncensored data. The first order auto-correlations are very slightly smaller, and the ninth slightly higher, but the difference is minimal and general patterns the same. It seems that when the data is censored based on the AR(1) process, both the balanced and unbalanced log(earnings) panels generate auto-covariance matrices very similar to those observed in the uncensored data, and still reflect the presence of the individual specific effect.

Table 3.9: *Auto-covariance matrix of IHS(earnings, $\theta = 1$): 25% censoring**

T	1	2	3	4	5	6	7	8	9	10
1	23.185 (0.026)	0.64	0.50	0.41	0.34	0.29	0.24	0.20	0.18	0.15
2	14.78 (0.027)	23.167 (0.026)	0.64	0.51	0.41	0.34	0.29	0.24	0.21	0.18
3	11.71 (0.027)	14.82 (0.027)	23.205 (0.026)	0.64	0.51	0.41	0.34	0.29	0.24	0.21
4	9.55 (0.026)	11.73 (0.027)	14.79 (0.027)	23.198 (0.026)	0.64	0.50	0.41	0.34	0.28	0.24
5	7.91 (0.026)	9.54 (0.026)	11.72 (0.027)	14.81 (0.027)	23.199 (0.026)	0.64	0.50	0.41	0.34	0.29
6	6.64 (0.026)	7.93 (0.026)	9.55 (0.026)	11.68 (0.027)	14.77 (0.027)	23.179 (0.026)	0.64	0.50	0.41	0.34
7	5.59 (0.025)	6.63 (0.026)	7.9 (0.026)	9.53 (0.026)	11.66 (0.027)	14.76 (0.027)	23.164 (0.026)	0.64	0.50	0.41
8	4.74 (0.025)	5.61 (0.025)	6.62 (0.026)	7.89 (0.026)	9.52 (0.026)	11.67 (0.027)	14.74 (0.027)	23.13 (0.026)	0.64	0.51
9	4.06 (0.025)	4.8 (0.025)	5.58 (0.025)	6.58 (0.026)	7.88 (0.026)	9.52 (0.026)	11.68 (0.027)	14.78 (0.027)	23.155 (0.026)	0.64
10	3.46 (0.025)	4.08 (0.025)	4.75 (0.025)	5.57 (0.025)	6.6 (0.026)	7.89 (0.026)	9.51 (0.026)	11.69 (0.027)	14.75 (0.027)	23.132 (0.026)

** Censored based on AR(1) component*
Diagonal shows the variance of wages in each period, lower triangle is co-variances, upper triangle is correlations. Standard errors in parentheses

Table 3.9 presents the auto-covariance matrix of IHS(earnings, $\theta = 1$) when the sample has 25% censoring based on the AR(1) component. There are two main differences between this table and the results with log(earnings) in Table 3.8. First, the variances and auto-covariances are much higher. The increase in variance is driven by including the extensive

margin changes, so movements into and out of the workforce are contributing to the measured variance, as shown in Table 3.5. The auto-covariances are much larger than for log(earnings), but steadily decrease as the order examined increases. Second, The auto-correlation is lower at all levels, and seems to steadily decrease rather than approaching the fraction of variance caused by the individual specific fixed effects. This is driven by the AR(1) process, as two periods become further apart the correlation in the participation decision for an individual decreases, and earnings are less likely to be correlated. This is an interesting result, it implies that if selection in the empirical data is driven by a similar mechanism, then the IHS may under estimate the level of the individual specific effect (if there is one in the empirical data).

The results are similar for both other values of θ, with the variances and auto-covariances increasing as θ decreases. The biggest difference occurs with $\theta = 0.001$, where both the first and ninth order auto-correlations are larger than in the other IHS cases, but still smaller than for log(earnings).

Table 3.10: *Auto-covariance matrix of log(earnings):* 25% *censoring**

T	1	2	3	4	5	6	7	8	9	10
1	1.145 (0.002)	0.65	0.64	0.63	0.62	0.61	0.61	0.60	0.60	0.60
2	0.75 (0.002)	1.143 (0.002)	0.65	0.64	0.63	0.62	0.61	0.61	0.60	0.60
3	0.73 (0.002)	0.74 (0.002)	1.140 (0.002)	0.65	0.64	0.63	0.62	0.61	0.61	0.61
4	0.72 (0.002)	0.73 (0.002)	0.74 (0.002)	1.141 (0.002)	0.65	0.64	0.63	0.62	0.61	0.61
5	0.71 (0.002)	0.72 (0.002)	0.73 (0.002)	0.74 (0.002)	1.143 (0.002)	0.65	0.64	0.63	0.62	0.61
6	0.70 (0.002)	0.71 (0.002)	0.72 (0.002)	0.73 (0.002)	0.74 (0.002)	1.139 (0.002)	0.65	0.64	0.63	0.62
7	0.69 (0.002)	0.70 (0.002)	0.71 (0.002)	0.71 (0.002)	0.73 (0.002)	0.74 (0.002)	1.141 (0.002)	0.65	0.64	0.63
8	0.69 (0.002)	0.70 (0.002)	0.70 (0.002)	0.71 (0.002)	0.72 (0.002)	0.73 (0.002)	0.74 (0.002)	1.141 (0.002)	0.65	0.64
9	0.69 (0.002)	0.69 (0.002)	0.70 (0.002)	0.70 (0.002)	0.71 (0.002)	0.72 (0.002)	0.73 (0.002)	0.74 (0.002)	1.142 (0.002)	0.65
10	0.68 (0.002)	0.69 (0.002)	0.69 (0.002)	0.70 (0.002)	0.70 (0.002)	0.71 (0.002)	0.72 (0.002)	0.73 (0.002)	0.75 (0.002)	1.146 (0.002)

** Censored using combination of AR(1) and α_n*
Diagonal contains variances, upper triangle the auto-correlations, lower triangle the auto-covariances, and standard errors in parentheses.

Table 3.10 contains the auto-covariance matrix of log(earnings) with 25% censoring, when censoring is based on both the AR(1) component and the individual specific effect α_n. With the combination of censoring, the logarithm produces results that are further away from those produced with the uncensored sample than when censoring was based solely on the AR(1) component. The variance is lower in every period, from approximately 1.5 in every period with the uncensored sample, to approximately 1.1 here. The auto-correlations are also lower than those in both the uncensored sample and the sample censored based on the AR(1) process. Both the first order, and ninth order auto-correlations are smaller, but again it does appear to reach a steady state point, rather than continuing to decrease as was the case in IHS(earnings). The auto-covariances also start and end lower than previously.

These results imply that, if empirical workforce participation follows a similar process, log(earnings) would still reflect the presence of the individual specific effect in the auto-correlations, but that this estimate may be biased downwards. Likewise, the auto-covariances would imply a lower level of $\text{Var}(\alpha_n)$ than is used in the DGP. It is also interesting that the balanced and unbalanced panels of log(earnings) produce very similar results, although this will not necessarily be true in the empirical data.

Table 3.11 presents the auto-covariance matrix of IHS(earnings, $\theta = 1$) with the 25% censored sample, where censoring is based on the combination of components. There are a few significant differences between these results and those seen with the IHS when censoring was based on only the AR(1) component. The variances are slightly smaller here, although the difference is relatively small. The auto-covariances are larger at all orders, and while they also decrease as the periods examined are further apart, the decreases are smaller, and at the ninth order the auto-covariances are much larger in this sample. The auto-correlations are much larger in the sample censored based on the combination of components. The first order auto-correlations are actually larger than in the uncensored data, and while they decrease steadily as the order of auto-correlation examined increases they are still remain relatively large (0.52 at the ninth order). This result is interesting, as compared to the results in Table

Table 3.11: *Auto-covariance matrix of IHS(earnings, $\theta = 1$): 25% censoring***

T	1	2	3	4	5	6	7	8	9	10
1	22.675 (0.027)	0.79	0.71	0.66	0.62	0.59	0.57	0.55	0.54	0.52
2	17.8 (0.028)	22.669 (0.027)	0.78	0.71	0.66	0.62	0.59	0.57	0.55	0.54
3	16.08 (0.028)	17.79 (0.028)	22.677 (0.027)	0.78	0.71	0.66	0.62	0.59	0.57	0.55
4	14.92 (0.028)	16.06 (0.028)	17.78 (0.028)	22.684 (0.027)	0.78	0.71	0.66	0.62	0.59	0.57
5	14.06 (0.028)	14.92 (0.028)	16.1 (0.028)	17.79 (0.028)	22.685 (0.027)	0.78	0.71	0.66	0.62	0.59
6	13.43 (0.028)	14.08 (0.028)	14.94 (0.028)	16.07 (0.028)	17.79 (0.028)	22.675 (0.027)	0.79	0.71	0.66	0.62
7	12.93 (0.028)	13.44 (0.028)	14.09 (0.028)	14.94 (0.028)	16.1 (0.028)	17.82 (0.028)	22.665 (0.027)	0.78	0.71	0.66
8	12.51 (0.027)	12.93 (0.028)	13.45 (0.028)	14.09 (0.028)	14.95 (0.028)	16.11 (0.028)	17.78 (0.028)	22.651 (0.027)	0.78	0.71
9	12.16 (0.027)	12.51 (0.027)	12.93 (0.028)	13.45 (0.028)	14.09 (0.028)	14.93 (0.028)	16.07 (0.028)	17.78 (0.028)	22.668 (0.027)	0.79
10	11.88 (0.027)	12.17 (0.027)	12.52 (0.027)	12.92 (0.028)	13.44 (0.028)	14.09 (0.028)	14.93 (0.028)	16.08 (0.028)	17.8 (0.028)	22.676 (0.027)

** Censored based on combination of AR(1) and α_n*
Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, and standard errors in parentheses.

3.9 the effect of the individual effect is stronger. The censoring based on α_n seems to produce very high levels of auto-correlation, as if a sufficiently low α_n causes an individual to not work in one period, they will not work in any, and will have perfectly correlated earnings (see the summary statistics in Table 3.3 where the auto-correlations produced in the α_n censored models are very large).

3.4 Empirical Data: The 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 includes both males and females, but this thesis will focus solely on the female portion of the sample. As outlined in Section 3.2, female earnings dynamics have historically been neglected in the literature, with a greater focus on the extensive margin for female workers. This is largely due 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,[4] while in comparison, 59.5% of the females in our SoFIE extract work in every period, and 7.1% never work.

As covered in Section 3.2, low participation rates complicate the modelling of earnings dynamics. Following on from Section 3.3, in this section we trial the use of the IHS as an alternative to the logarithm when working with empirical earnings data. The IHS allows us to include the observations where individuals do not work, where as the logarithm forces the exclusion of such observations. This allows for an examination of earnings data that includes the effects of extensive margin variation.

In this section we will apply similar methods as those used in Section 3.3 with the simulated models. Summary statistics will be presented, the variance of annual earnings will be decomposed into its extensive and intensive portions, and the auto-covariance matrices of annual earnings will be generated. Both the logarithm and the IHS will be used in generating these results. In the case of the logarithm, we will use both a balanced panel where only individuals that work in every period are included, and an unbalanced panel that excludes periods where an individual does not work. As in Section 3.3, the variance decomposition will not be applied to the logarithm of earnings. Following Section 3.3, the IHS will be applied with $\theta = \{1, 0.1, 0.001\}$.

[4]Male participation rates and summary statistics are presented in Appendix C

3.4.1 Survey of Families, Income, and Employment Summary

SoFIE was designed to provide longitudinal data similar to that gathered in overseas studies such as the Household, Income, and Labour Dynamics in Australia (HILDA), and the Panel Survey of Income Dynamics (PSID) (Carter et al., 2009; Statistics New Zealand, 2001). Data were collected annually over eight waves, with the survey starting in October of 2002 following a feasibility study commissioned in 1997 and presented in 2001 (Statistics New Zealand, 2001). The aim of the survey was to allow researchers to explore and analyse a range of economic and social factors affecting individuals and households. The longitudinal structure allows for the modeling of dynamic relationships which were previously difficult or impossible to capture using the existing cross-sectional data (Carter et al., 2014).

The target population for the study was the usually resident population of New Zealand, residing in a permanent private dwelling on one of the main islands in the North Island or the South Island, including Waiheke island, as of the first wave of the panel (Statistics New Zealand, 2008). The target population was adjusted over time, as individuals moved out of scope by moving overseas, into institutions, or died (Statistics New Zealand, 2008). In the first wave, a total of $15,100$ randomly selected households were contacted, and approximately 77% of these responded (Statistics New Zealand, 2008). This resulted in data being collected for $22,000$ adults and $7,500$ children aged under 15, in over $11,500$ households (Statistics New Zealand, 2008). In the following waves, the same individuals/families were contacted and re-interviewed, so their progression could be tracked over time (Statistics New Zealand, 2011).

Like other longitudinal studies, not all members of the survey responded in each wave (John Fitzgerald, 1998). Sample attrition meant that by the 4th wave of the survey 76% of the original sample were still taking part, and when SoFIE finished in 2010 the overall retention rate was 65% (Statistics New Zealand, 2011; Statistics New Zealand, 2008). Each wave of SoFIE included longitudinal weights created by SNZ. The goal of the longitudinal weights was to make the sample representative of the target population, compensating

for the potentially non-random attrition of individuals from the survey up to that point (Statistics New Zealand, 2011).

The data collected in SoFIE includes annual earnings and income, education, employment status, family status, and other demographic information (Carter et al., 2014). Every second wave (waves two, four, six, and eight) additional data was collected on assets and liabilities, and the SoFIE-Health sub-study collected more in depth information on the respondents health in waves three, five, and seven (Carter et al., 2014).

As SoFIE contains individual's personal information, SNZ has rules on how it can be used and presented. In all the results that follow, including those presented in Chapter 4, these rules have been applied. When presented, the sample size used is randomly rounded to base three. This means that each sample size is rounded to the nearest multiple of three with two thirds probability, and rounded to the next nearest one third of the time. Sample sizes that are already a multiple of three are left unchanged. Any proportions or means that are presented are based on these rounded counts. Regression results only need to be altered or censored if a small quantity of data is used, or if the results allow identification of individuals or firms. This does not occur in any of the regressions in this thesis, so all regression results are unchanged. The degrees of freedom presented for regressions have been adjusted where required, to match the rounded sample sizes.

3.4.2 Data Extract

This thesis 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. Similarly to sample selection bias, non-random attrition can lead to biased results, but the data includes longitudinal weights produced by SNZ that are intended to correct for this. 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.

Table 3.12 lists some summary statistics for the SoFIE extract. Where longitudinal weights are used in this analysis, it is the weight from the final wave of SoFIE that are used, as the weights are adjusted each wave to take into account sample attrition. As the extract used was selected based on the individuals responding in each year, using the final longitudinal weights was most appropriate. A portion of the sample reported having negative earnings in some years, and results are presented for both the full extract and a sub-sample that removes the individuals that have negative earnings. In each case, both weighted and unweighted summary statistics are presented.

Table 3.12: *SoFIE summary statistics*

	Full sample		Sub-sample$^+$	
	unweighted	weighted	unweighted	unweighted
Sample size	4572	4572	4464	4464
Weighted sample size	-	778,962.5	-	759,862.5
Fraction working	0.81	0.80	0.81	0.80
Mean(earnings*)	26.997 (48.293)	27.311 (49.346)	27.356 (48.557)	27.693 (49.604)
Mean(earnings*|working)	33.469 (51.718)	33.997 (52.953)	33.870 (51.947)	34.441 (53.179)
Age	42.87	42.21	42.85	42.17
Age | working	43.03	42.31	43.02	42.28
Fraction with children	0.55	0.54	0.55	0.54
Fraction with children | working	0.53	0.51	0.52	0.51
Fraction with partner	0.74	0.76	0.74	0.76
Fraction with partner | working	0.75	0.77	0.75	0.76
No qualification	0.16	0.15	0.16	0.14
No qualification | working	0.14	0.13	0.14	0.12
School	0.29	0.28	0.29	0.28
School | working	0.29	0.28	0.29	0.28
Vocational	0.35	0.35	0.34	0.35
Vocational | working	0.35	0.35	0.35	0.35
University	0.20	0.22	0.21	0.23
University | working	0.22	0.24	0.22	0.25

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

Table 3.12 shows that removing the individuals with negative earnings decreases the sample size by 108. Comparing the sample statistics, there are only very small differences between the sub and full samples. Removing these individuals leads to mean earnings and mean earnings conditional on working both rising in the sub-sample, while other changes are minor. Likewise, the weighted summary statistics are very similar to the unweighted results. Using the longitudinal weights increases both mean earnings and mean earnings conditional on working, while slightly reducing the fraction of the sample that is working. Weighting also slightly reduces both age and age conditional on working, indicating that the sample is slightly older than the population it is targeted at. The fraction of both the full and sub-sample with children falls by one percentage point after weighting, while the fraction that work and have children falls by two percentage points in the full sample, and one in the sub-sample. In both the full and sub samples, weighting increases the fraction with a partner.

The education variables in Table 3.12 indicate the fraction of the sample that has that level of education as their highest, 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. The weighting indicates that individuals in our extract are slightly more likely to have a high school level of education than the target population, have vocational training in approximately the same proportion, and are slightly less likely to have a university education, but in all cases the difference between weighted and unweighted is small.

While the longitudinal weights are designed to make the sample population more representative of the target population, the summary statistics in Table 3.12 show that applying the weights does not have a large impact on the results. As the differences between the weighted and unweighted samples are small, it seems that the attrition that occurred over

the course of the survey has had at most a small impact. Further, SoFIE does not oversample particular sub-populations in the way that some other longitudinal surveys such as the PSID do (Becketti et al., 1988; John Fitzgerald, 1998). As the unweighted extract seems relatively representative of the target population, the further analysis of both this chapter and Chapter 4 will focus on the unweighted data (Solon et al., 2015).

3.4.3 Descriptive statistics

Table 3.13 contains some descriptive statistics on the earnings, the IHS of earnings, and the logarithm of earnings for individuals in the sample. In this table, mean(earnings) includes zero observations for the untransformed earnings, as well as the IHS of earnings. The balanced log panel includes only those individuals that work in all of the eight waves, and the unbalanced log panel includes any observations where an individual works. This means that all $4,572$ individuals in the full sample, and $4,464$ from the sub-sample are used in estimating the sample statistics for earnings and IHS(earnings). The balanced log panel selects only those individuals that work in every period, and is not defined for negative earnings, with $N = 2,658$ in both cases. The unbalanced panel of log(earnings) uses any observation where an individual works, resulting in $N = 4,245$ for the full sample, and $N = 4,146$ for the sub-sample. As periods of non-participation are removed, different fractions of the sample are used in estimating the variances, auto-covariances, and auto-correlations in the unbalanced panel. Table 3.14 lists the fraction of the sub-sample used to generate the different statistics for each wave or wave pair.

For the other variables, $\text{Var}(Y_{it})$ is the average of the cross-sectional variances over the eight waves of the survey, $\text{Var}(\overline{Y}_i)$ is the average intertemporal variance of earnings experienced by individuals in the sample, $\text{Cor}(Y_t, Y_{t-1})$ is the average first order auto-correlation of earnings, $\text{Cor}(Y_1, Y_8)$ is the seventh order auto-correlation of earnings, and Skew is the level of skewness in the data. The descriptive statistics of earnings show similar patterns between the full and sub-sample, with the main difference being that removing individuals that have

Table 3.13: *Earnings summary statistics*

	N	Mean	Var(Y_{it})	Var($\overline{Y}_i$)	Cor(Y_t, Y_{t-1})	Cor(Y_1, Y_8)	Skew
	Full sample						
Earnings*	4572	26.997	2328.07	1466.894	0.760	0.243	75.432
Balanced log	2658	10.25	0.79	0.368	0.691	0.364	-1.897
Unbalanced log	4245	9.999	1.33	0.693	0.676	0.352	-2.021
IHS(earnings, $\theta = 1$)	4572	8.528	20.433	8.26	0.764	0.405	-1.476
IHS(earnings, $\theta = 0.1$)	4572	66.931	1295.85	507.728	0.774	0.415	-1.384
IHS(earnings, $\theta = 0.001$)*	4572	3.044	3.299	1.148	0.806	0.456	-0.884
	Sub-sample$^+$						
Earnings*	4464	27.36	2353.11	1482.475	0.769	0.247	75.995
Balanced log	2658	10.25	0.79	0.368	0.691	0.364	-1.897
Unbalanced log	4146	10.01	1.31	0.678	0.679	0.359	-2.043
IHS(earnings, $\theta = 1$)	4464	8.64	18.81	6.992	0.798	0.435	-1.381
IHS(earnings, $\theta = 0.1$)	4464	67.83	1197.96	431.891	0.807	0.445	-1.294
IHS(earnings, $\theta = 0.001$)*	4464	3.09	3.12	1.01	0.831	0.485	-0.819

** in thousands*
+ sub-sample with negative earnings removed

negative earnings reduces both measures of variance, and increases the auto-correlations for the IHS of earnings, while the balanced log is identical as in each case it uses the same portion of the sample. The unbalanced log has a small increase in mean(earnings), and slight decreases in most measurements of variance, while the variance of untransformed earnings actually increases in the non-zero sub-sample. Both Table 3.13 and 3.12 show that the full and sub-samples have very similar demographics and earnings patterns, so the rest of the analysis focusing on SoFIE will use the sub-sample with non-negative earnings.

Earnings in the sub-sample are positively skewed, potentially indicating the presence of outliers or a long right tail. The average cross-sectional variance of earnings is also much higher than the average inter-temporal variance, indicating that there may be some level of long term earnings inequality. The average first order auto-correlation is also quite high at 0.769, indicating that earnings in one wave of the survey are highly correlated with earnings in the previous or next wave. The max order auto-correlation is 0.247, so even for the waves seven years apart there is a reasonable level of auto-correlation in untransformed earnings, potentially indicating the presence of permanent components that influence earnings.

The balanced and unbalanced panels of log(earnings) have relatively similar results. The

balanced panel has a slightly higher level of mean(earnings), potentially indicating that similarily to the simulated results in Section 3.3, individuals that work in every period have on average higher earnings. The unbalanced panel has both a higher cross-sectional and inter-temporal variance of earnings than the balanced panel, but the ratio of the two is relatively similar. The balanced panel has a slightly higher level of auto-correlation in earnings, but the differences are small. Likewise, both the balanced and unbalanced panels remove the positive skewness and actually have negative skewness, with the balanced panel being slightly closer to zero.

While the IHS with $\theta = 1$ produces results most similar to the logarithm, the inclusion of the zero earnings observations means that mean(IHS(earnings, $\theta = 1$)) is lower than both the balanced and unbalanced log panels, and slightly larger than the results for the 25% censored simulations in Section 3.3.2. Mean earnings increase as the level of θ used decreases, as would be expected from both Chapter 2 and Section 3.3. Likewise, both measures of variance increase as θ decreases, with the ratio of cross-sectional to inter-temporal variance being relatively stable, but increasing slightly as θ decreases. Compared to the logarithm, the IHS estimates a slightly higher relative level of cross-sectional variation and auto-correlation in earnings. The level of auto-correlation in IHS(earnings) increases as θ falls, but again the change is relatively minor. The IHS generates a distribution that has removed the positive skewness, replacing it with a slightly negatively skewed distribution, but in all cases the level of skewness is closer to zero than with the logarithm, with $\theta = 0.001$ producing the distribution closest to zero.

As stated previously, the unbalanced panel of log(earnings) uses any observation where an individual works. This results in a different number of observations being used to estimate the variance in each wave and the auto-covariance between two waves, with Table 3.14 reporting the fraction of the sample used. The fractions presented in Table 3.14 are based on the sample of individuals that work in at least one time period, so while it can be treated as showing workforce participation in each wave and the intertemporal correlation of participation, it

should be noted that in the full sample the fractions would be slightly smaller, taking into account the individuals that never work.

Table 3.14: *Unbalanced panel: 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
$N = 4146$								

Fraction of observations used in each time period/time period pair

Examining the diagonal of Table 3.14, it seems that participation increases slightly over the first four waves of the survey and then remains relatively stable. This differs from the participation patterns observed in the simulated data, where participation was stable. The SoFIE sample does show evidence of intertemporal correlation in the participation patterns, similar to those observed in the simulated data in Section 3.3.1. Examining the first column in Table 3.14 shows the fraction of the sample that work in both wave one and each of the other waves, so 77% of the sample works in the first wave, and 73% work in both the first and second wave. Examining any two waves, the fraction of the sample that works in both decreases as the the waves are further apart, indicating that there is intertemporal correlation in the participation patterns. Comparing the empirical participation patterns to those from the simulated data, the fraction working in each wave matches up relatively well, with participation slightly higher in the empirical data. The simulated data with censoring based on only the individual specific fixed effect α_n has a much faster drop off in participation between periods, while the version censored on the combination of factors has very similar participation patterns, but a much larger fraction of the sample never working.

3.4.4 Results

Table 3.15 contains the auto-covariances of earnings[5] for the sub-sample that have non-negative earnings. Here, the diagonal elements are the variance of annual earnings in each wave of the survey. The variance of earnings in the first wave is 796, and it increases over the first three waves to 993 in wave 2, and then 1251 in wave three. Most notably, the variance of annual earnings spikes up dramatically in the final year of the survey to $11,246$, almost ten times higher than in any of the previous years. This spike in the variance appears to be driven by outlier(s) that have annual earnings in the eighth wave that are much higher than earnings in the previous years.

Table 3.15: *Auto-covariance matrix of earnings**

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	796 (121.7)	0.80	0.73	0.76	0.70	0.71	0.68	0.25
Wave 2	713.37 (115.2)	993 (155.6)	0.77	0.77	0.74	0.72	0.69	0.25
Wave 3	729.31 (135.7)	856.66 (153.7)	1251 (293.6)	0.80	0.76	0.76	0.70	0.25
Wave 4	679.16 (128.3)	770.66 (150.5)	902.71 (175.4)	1009 (182.8)	0.89	0.86	0.84	0.29
Wave 5	655.2 (122.6)	778.8 (154.3)	891.94 (181.3)	937.21 (187.8)	1102 (201.0)	0.89	0.82	0.28
Wave 6	696.81 (165.4)	792.46 (176.4)	934.67 (211.1)	948.17 (209.8)	1025.44 (212.0)	1215 (261.0)	0.91	0.34
Wave 7	671.92 (145.4)	762.66 (168.3)	862.86 (196.9)	924.72 (203.4)	951.72 (209.1)	1105.68 (242.5)	1213 (238.5)	0.32
Wave 8	737.94 (173.9)	834.62 (192.9)	944.48 (222.2)	971.91 (223.7)	1002.21 (227.8)	1252.68 (329.5)	1195.76 (260.4)	11246 (10018.7)
	$N = 4464$	$T = 8$						

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, and standard errors are in parenthesis.
** in thousands*

The upper triangle of Table 3.15 contains the auto-correlation of annual earnings over the different waves of the survey. This shows that the first order auto-correlation between waves one and two is 0.80, so earnings in the first wave are highly correlated with earnings in the second. As we move to the right in the first row, the level of auto-correlation generally

[5]Earnings in thousands

decreases, falling to 0.73 when comparing wave one to wave three, rising slightly to 0.76 with wave 4, and then decreasing to 0.70 with wave five. The largest change is comparing the auto-correlation between earnings in wave one and seven, to that between earnings in wave one and wave eight, where the correlation falls from 0.68 to 0.25.

The results here are relatively different to those observed in the Monte-Carlo (MC) simulations. The auto-correlations are much larger and persistent than observed in any of the simulated models with untransformed earnings. The auto-covariances are also relatively persistent, not decreasing significantly as the order examined increases. The eighth wave appears to be an outlier. As mentioned previously, the variance of annual earnings in the eighth period is much higher than in all of the other survey waves. Likewise, the level of correlation between wave eight and the other waves is much lower than when comparing any other pair of waves. The highest level of auto-correlation involving wave eight is the 2nd order auto-correlation with wave six, 0.34, whereas the lowest auto-correlation not involving wave eight is 0.68. The standard error on the estimate of the variance of earnings in wave eight is also very large. All of the variances and covariances in Table 3.15 are significantly different from zero, except the variance of earnings in the 2009-2010 wave.

Table 3.16 is the auto-covariance matrix of log(earnings) based on the unbalanced panel. There are a few interesting results here. It seems that the logarithmic transformation has reduced the impact of the outliers that was evident in wave eight of the untransformed earnings. In fact, the variances follow a different pattern here as compared to the untransformed earnings, with variance highest in the first two waves, and then generally decreasing over the course of the panel. The auto-correlations are quite stable at each respective order, and shows some evidence of a individual specific effect in log(earnings). The first order auto-correlation is approximately equal to 0.59, and seems to approach approximately 0.3.

Table 3.17 is the auto-covariance matrix of log(earnings) using the balanced panel. The auto-correlations are similar to those for the unbalanced panel, but larger at every order examined. This still indicates the presence of a permanent component in earnings. The vari-

Table 3.16: *Auto-covariance matrix of log(earnings): unbalanced panel*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	1.406 (0.069)	0.60	0.46	0.42	0.39	0.35	0.35	0.32
Wave 2	0.86 (0.057)	1.453 (0.080)	0.57	0.47	0.45	0.38	0.37	0.35
Wave 3	0.65 (0.032)	0.81 (0.042)	1.398 (0.073)	0.57	0.48	0.42	0.39	0.35
Wave 4	0.59 (0.032)	0.67 (0.034)	0.8 (0.037)	1.438 (0.071)	0.59	0.49	0.44	0.40
Wave 5	0.54 (0.027)	0.62 (0.032)	0.65 (0.031)	0.82 (0.039)	1.328 (0.070)	0.60	0.50	0.44
Wave 6	0.44 (0.023)	0.49 (0.025)	0.53 (0.024)	0.63 (0.029)	0.74 (0.034)	1.135 (0.054)	0.59	0.48
Wave 7	0.45 (0.025)	0.48 (0.024)	0.51 (0.027)	0.57 (0.028)	0.63 (0.03)	0.69 (0.032)	1.198 (0.061)	0.58
Wave 8	0.41 (0.026)	0.45 (0.024)	0.45 (0.023)	0.52 (0.027)	0.54 (0.029)	0.55 (0.027)	0.68 (0.031)	1.148 (0.063)

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.
See Table 3.14 for sample size

ances of log(earnings) in each wave is smaller in the balanced panel versus the unbalanced panel, and follows a different pattern to both the unbalanced log(earnings) and untransformed earnings, as seen in Table 3.15. There the variance was lowest in wave one, and then much higher in wave eight, where as here wave one has the highest variance, and wave eight is approximately the same as the others. Likewise, wave eight doesn't seem to have an abnormal level of correlation with the other waves, so it seems that with both the balanced and unbalanced panels the logarithm has successfully removed the effects of the outliers.

Both the balanced and unbalanced panels of log(earnings) show evidence of an individual specific component to earnings. They each large large and positive first order auto-correlations, which are slightly larger for the balanced panel. As the order of auto-correlation examined increases, the level decreases, but in both cases it appears it may be approaching a stable point. The variance of earnings is larger for the unbalanced panel, reflecting the broader distribution of earnings evident in the individuals that work in less than the full eight waves. There is no evidence that the variance of log(earnings) increases over the course of the survey. In fact, it seems to decrease, with the earlier waves having much larger

Table 3.17: *Auto-covariance matrix of log(earnings): balanced panel*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	1.028 (0.062)	0.68	0.54	0.54	0.46	0.43	0.39	0.36
Wave 2	0.60 (0.028)	0.771 (0.035)	0.70	0.61	0.55	0.52	0.46	0.44
Wave 3	0.49 (0.025)	0.54 (0.024)	0.795 (0.052)	0.70	0.60	0.55	0.46	0.44
Wave 4	0.49 (0.028)	0.48 (0.024)	0.55 (0.026)	0.788 (0.052)	0.71	0.62	0.53	0.49
Wave 5	0.41 (0.023)	0.42 (0.02)	0.47 (0.024)	0.55 (0.028)	0.768 (0.056)	0.70	0.58	0.54
Wave 6	0.35 (0.019)	0.37 (0.019)	0.4 (0.02)	0.44 (0.021)	0.5 (0.023)	0.657 (0.036)	0.69	0.56
Wave 7	0.34 (0.021)	0.35 (0.019)	0.35 (0.019)	0.41 (0.022)	0.44 (0.022)	0.49 (0.023)	0.757 (0.050)	0.67
Wave 8	0.33 (0.021)	0.34 (0.019)	0.34 (0.019)	0.38 (0.022)	0.42 (0.028)	0.4 (0.020)	0.51 (0.029)	0.783 (0.055)
	$N = 2658$	$T = 8$						

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.

variances.

Table 3.18 is the auto-covariance matrix for IHS(earnings, $\theta = 1$). The results are relatively similar with IHS(earnings, $\theta = 0.1$) and IHS(earnings, $\theta = 0.001$), with lower values of θ leading to larger variances and slightly larger auto-correlations. The auto-covariance matrices for $\theta = 0.1$ and $\theta = 0.001$ are available in Appendix B. In each case, the estimated variance is much larger than either of the log models, but the IHS has also managed to remove the effect of the large outlier(s) that produced the extreme results in Table 3.15. This increase in variance reflects the extensive margin changes that are included when using the IHS, and which the logarithm ignored. This matches what was observed when censoring was introduced to the simulated models. Interestingly, there is evidence of a permanent component here as well. In fact, for each of the IHS models, the auto-correlations are larger than for either of the log models. This is different from what we observed in any of the simulated models, implying that what ever selection mechanism is at work, it is different from any of the models we applied. For example, in the simulated model when censoring was based on the AR(1) component, this resulted in the IHS models producing much lower

Table 3.18: *Auto-covariance matrix of IHS(earnings, $\theta = 1$)*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	20.875	0.76	0.65	0.56	0.52	0.48	0.45	0.43
	(0.378)							
Wave 2	15.42	19.463	0.79	0.65	0.6	0.54	0.5	0.48
	(0.373)	(0.397)						
Wave 3	12.77	15.05	18.634	0.78	0.68	0.6	0.55	0.52
	(0.367)	(0.381)	(0.404)					
Wave 4	10.94	12.22	14.21	17.982	0.81	0.7	0.63	0.58
	(0.36)	(0.37)	(0.382)	(0.411)				
Wave 5	10.13	11.3	12.58	14.68	18.361	0.82	0.7	0.63
	(0.361)	(0.370)	(0.377)	(0.391)	(0.416)			
Wave 6	9.23	10.09	10.93	12.62	14.83	17.978	0.81	0.7
	(0.356)	(0.362)	(0.367)	(0.381)	(0.398)	(0.424)		
Wave 7	8.81	9.46	10.17	11.38	12.85	14.79	18.342	0.82
	(0.356)	(0.360)	(0.363)	(0.372)	(0.385)	(0.399)	(0.424)	
Wave 8	8.62	9.09	9.65	10.69	11.68	12.88	15.17	18.808
	(0.358)	(0.361)	(0.362)	(0.369)	(0.379)	(0.387)	(0.403)	(0.424)
$N = 4464$ $T = 8$								

Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, and standard errors in parentheses.

auto-correlations.

These results imply that there is an individual fixed effect present in annual earnings. Even in the presence of censored observations, the logarithm and IHS both have large auto-correlations, even at the 7th order. Given the auto-covariance matrices observed, there is no evidence of an MA(1) process, the first order auto-correlation is large and positive, and does not move significantly as the order examined increases. While not a robust analysis, the variances of each of the models show no signs of increasing.

Table 3.19 reports the results of decomposing the variance of untransformed annual earnings and IHS(earnings). For both untransformed and the IHS of earnings, mean earnings conditional on working generally decreases as the number of periods worked falls. There is an increase for individuals working in only one period for untransformed earnings, but this is probably due to the small sample working in only a single period being influenced by some number of outliers, and the effect disappears when the IHS is used. This is similar to the results of the MC simulations in Section 3.3.3, where the expected earnings an individual receives if they work decreases with the number of periods worked. This could be interpreted

Table 3.19: *Variance decomposition of earnings*

	Fraction of sample	mean(earn)	Mean(earn|working)	Waves worked	Intensive variance	Extensive variance	Fraction Intensive
			Earnings*				
8 waves	0.595	38.375	38.375	8	2328.67	0	1
7 waves	0.107	21.803	24.918	7	220.56	118.22	0.62
6 waves	0.067	15.523	20.698	6	176.57	120.31	0.54
5 waves	0.047	12.292	19.668	5	130.96	152.64	0.47
4 waves	0.036	7.811	15.621	4	63.87	97.61	0.44
3 waves	0.034	5.228	13.940	3	31.41	84.15	0.34
2 waves	0.026	2.796	11.182	2	16.36	55.17	0.28
1 wave	0.018	2.222	17.778	1	-	578.74	-
0 waves	0.071	0	-	0	-	-	-
Average	-	27.356	31.712	6.46	1572.39	49.57	0.83
			IHS(earnings, $\theta = 1$)				
8 waves	0.595	10.95	10.95	8	0.37	0	1
7 waves	0.107	8.99	10.28	7	0.92	11.63	0.07
6 waves	0.067	7.54	10.06	6	0.96	19.11	0.05
5 waves	0.047	6.16	9.86	5	0.91	23.09	0.04
4 waves	0.036	4.81	9.62	4	0.94	23.43	0.04
3 waves	0.034	3.55	9.46	3	0.46	21.37	0.03
2 waves	0.026	2.26	9.04	2	0.23	15.92	0.03
1 wave	0.018	1.09	8.75	1	-	8.73	-
0 waves	0.071	0	-	0	-	-	-
Average	-	8.64	10.55	6.46	0.53	6.15	0.67
IHS(earnings, $\theta = 0.1$)	4464	67.83	82.49	6.46	52.01	363.97	0.68
IHS(earnings, $\theta = 0.001$)*	4464	3.086	3.686	6.46	0.387	0.631	0.74

** in thousands*
$N = 4464$

a number of ways. It could be that individuals that receive higher wage offers are more likely to work in a given year, and therefore more likely to work over more periods in the survey data. It could also be the result of individuals having different characteristics that influence both their probability of working in a given wave, and if they do work also influencing the earnings they receive.

Almost 60% of sample work in all eight waves, and this then drops off to approximately 11% working in seven of the eight. The fraction working for each number of waves continues to decrease from there as the number of waves worked falls, with approximately 1.8% of the sample working in only one wave. There is a larger clump of observations, about 7.1% of the sample, that never work over the course of the SoFIE survey. These results are also quite similar to those from the MC simulations, where working in all periods consistently has the largest fraction of the sample.

The level of intensive margin variance in earnings is much higher for those working in

every period than any other number of waves worked. In fact, at 2,328.67, it is more that ten times larger the next biggest value. This is probably caused by outliers, as observed in the auto-covariance matrix of annual earnings in Table 3.15. The application of the IHS certainly seems to have removed this effect here, as the variance attributed to intensive margin variation in IHS(earnings, $\theta = 1$) for those that work in every period is more reasonable relative to the other values. In untransformed earnings the level of intensive margin variance steadily decreases with the number of waves an individual works in. For IHS(earnings, $\theta = 1$) the fraction of variation attributable to intensive margin variation actually increases as the number of waves worked decrease from eight, to seven, to six, and then stays relatively stable for those working in five or four waves, before decreasing for those working for three or two waves. This could indicate that, after reducing the effect of outliers, those working consistently have lower variation in their earnings over time, and that individuals working in only some periods have less stable earnings.

Both untransformed and IHS(earnings, $\theta = 1$) show a similar pattern in the level of extensive margin variance. In each case it follows an inverted U pattern, first increasing as the number of periods worked falls, and then decreasing. Again, outliers seems to be influencing the results for the untransformed earnings. The level of variance attributable to the extensive margin is 578.74 for those working in only one of the waves, and this is much larger than all of the other values. As seen in the intensive margin case, the application of the IHS seems to have removed this effect, the extensive margin variation is actually lowest for those working in only a single wave with IHS(earnings, $\theta = 1$).

The relative importance of the intensive margin compared to the extensive margin when looking at earnings variation changes dramatically when using the IHS versus untransformed earnings. Untransformed earnings attributes a much larger fraction of total earnings variation to the intensive margin among individuals that do not work in every period. For example, for those working in seven of the eight waves, the variance decomposition using the untransformed earnings attributes 62% of earnings variation to the intensive margin,

where as when working with IHS(earnings, $\theta = 1$) only 7% is attributed to intensive margin variation. As the majority of the sample work in every year, the difference in the fraction of variance attributed to the extensive and intensive margins over the entire sample is not as different as might be expected when examining the breakdown by number of waves worked. When decomposing the variance of untransformed earnings 83% of the variation is attributed to the intensive margin over the entire sample, where as when working with IHS(earnings, $\theta = 1$) 67% is attributed to the intensive margin. The overall results are similar with the other values of θ, with the level of earnings and variance rising as θ approaches zero, and relatively similar results with the fraction of variance attributed to the intensive margin, with that fraction being a bit larger with $\theta = 0.001$.

3.5 Conclusion

This chapter has introduced some methods for analysing annual earnings, including a method of decomposing the variance of earnings into its intensive and extensive portions. The MC simulations in this chapter seem to have produced simulated annual earnings data that matches that observed empirically relatively well. Applying the logarithm and the IHS to both the simulated and empirical data sets has helped demonstrated some of the strengths and weaknesses of the IHS.

Unlike the logarithm, the IHS is defined for zero values. This allows the function to be applied to earnings data where there are observations where individuals do not participate in the workforce. The inclusion of the zero earnings observations allows for the creation of descriptive statistics, such as the mean and variances, that include the periods where individuals do not work, potentially giving a more accurate understanding of earnings over the sample and incorporating extensive margin changes.

The participation patterns observed in the empirical data are similar to those from the simulated data. In both cases, there was evidence of intertemporal correlation in participa-

tion patterns, so working in one period increased the probability of also working in the next. The empirical auto-covariance matrices provided some evidence that there is a persistent individual specific effect in annual earnings, matching the results observed in the simulated auto-covariance matrices relatively well. The differences between the auto-covariance matrices for the balanced and unbalanced panels of log(earnings), combined with the differences observed in the variance decompositions indicate that the individuals that work are not the same as those that don't. Restricting the sample to those that work in every period produced quite different results, and examining the earnings of individuals that work different numbers of years, we see that there is a strong correlation between the number of years worked, and an individuals annual earnings. These results imply that there may be non-random selection into the workforce, and that any model of earnings should take that into account or risk producing biased and inconsistent results.

A strength of the IHS demonstrated in this chapter was its ability to reduce the impact of outliers. When using the untransformed level of annual earnings, both the variance decomposition and the auto-covariance matrix showed some evidence that extreme values were influencing the results, leading to much larger variances and lower auto-covariances in some cases. The application of the IHS appeared to successfully deal with this, with both variances and auto-covariances that were in line with the other results. In this case, each of the three values of θ used seemed to deal with the outliers.

The application of the IHS lead to quite different results when decomposing the variance of earnings into its intensive and extensive portions. With untransformed earnings, the intensive margin always made up a significant fraction of total variance, having the smallest impact for those working in only two periods where it was responsible for 28% of earnings variation. When the variance of IHS(earnings) is decomposed, the fraction of earnings variation attributed to the extensive margin is much lower. It is possible that removing the impact of extreme outliers is producing a more accurate model of the relative impact of intensive versus extensive margin variation, but problematically the value of θ chosen impacts

the results. It seems that as the value of θ falls, the fraction of variance attributed to the intensive margin increases. As shown in Chapter 2, θ estimation is unreliable when there is a concentration of data at zero, so it is unclear how a value for θ should be selected.

Chapter 4

Sample Selection Bias

4.1 Introduction

This chapter 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. Chapter 3 applied the logarithmic and Inverse Hyperbolic Sine (IHS) transformations to simulated and empirical annual earnings data, and found that while the use of the IHS helped to capture the extensive margin effects on annual earnings that the logarithm ignores, it does not take into account the underlying characteristics and behaviours that determine selection.

As detailed in Chapter 3, 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 increasing the 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 chapter we apply a newly developed 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. Their 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. Their 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 chapter proceeds as follows. First, Section 4.2 formally introduces sample selection bias, with a focus on non-random selection in panel earnings data, the issues it can cause, and how it has been corrected for in the literature. Section 4.3 introduces the new method proposed by Semykina and Wooldridge (2013) for correcting for sample selection bias in dynamic panel models. Section 4.4 models the earnings of New Zealand females from the SoFIE data set, using a range of standard models. Section 4.4.2 applies Semykina and Wooldridge's correction model to the SoFIE data, allowing for a comparison between the naive models that ignore selection, and the new model that takes it into account. Section 4.4.3 discusses the results, and analyses how controlling for non-random selection has altered the output of the models. Finally, Section 4.4.4 suggests some extensions to the model that could produce interesting future research.

4.2 Sample Selection Bias

Sample selection bias is a potentially serious issue when modelling the annual earnings of female workers. When examining the intensive margin of labour supply, which is the decision of how many hours to work, or modelling intertemporal 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).

Equation 4.2.1 is the outcome equation of interest, a cross-sectional model of earnings that includes a selection process. Heckman (1979) used a very similar, but more general, model to illustrate how non-random selection can lead to biased estimates. This is a two equation model, where (4.2.1a) determines the annual earnings individual i would receive if they worked, and (4.2.1c) is the selection equation that determines whether they work in a given year,

$$Y_i^* = X_i\beta + \epsilon_i \ , \tag{4.2.1a}$$

$$Y_i = d_i Y_i^* = d_i(X_i\beta + \epsilon_i) \ , \tag{4.2.1b}$$

$$d_i = 1(Z_i\gamma + u_i > 0) \ . \tag{4.2.1c}$$

Here Y_i^* is the earnings individual i would receive if they work, d_i is an indicator variable which equals one if individual i works and zero otherwise, and Y_i is observed earnings, which is equal to zero if they do not work and equal to Y_i^* if they do. The vector X_i contains the observed variables that influence the individual's potential annual earnings, Z_i

is a vector of observable variables that influence the probability of an individual working and usually includes X_i and some additional variables, and ϵ_i and u_i are unobserved shocks for the earnings and selection equations respectively, with the presence of u_i making selection probabilistic. β and γ are vectors of coefficients for the earnings and selection equations.

The regression equation for the entire population with no selection, so every individual works and $d_i = 1,\ \forall i\ \in\ I$, is $\mathbb{E}(Y_i^*) = X_i\beta + \epsilon_i$, which by construction has $\mathbb{E}(\epsilon_i) = 0$. On the other hand, with a selection process in which positive earnings are only observed for the subset of the population that works, and observations with zero earnings are excluded from the model, the regression equation will be $\mathbb{E}(Y_i|X_i, d_i = 1) = X_i\beta + \mathbb{E}(\epsilon|Z_i + u_i > 0)$. If employment selection is random with regards to the observed regressors and the error term ϵ_i, so there is no connection between the probability of working and an individual's earnings, $\mathbb{E}(\epsilon_i|Z_i + u_i > 0) = 0$ and our estimator will be unbiased, with the only consequence of the limited sample being a reduction in efficiency (Heckman, 1979). With a non-random sample there is no guarantee that this condition will hold, and there is significant evidence that selecting into employment is not random (Killingsworth and Heckman, 1986; Das et al., 2003; Eckstein and Lifshitz, 2011). If there is non-zero correlation between u_i and both ϵ_i and X_i, so that earnings are correlated with the probability of working, then the expected value of ϵ_i conditional on individual i working will not be zero and the conditional mean of Y_i will be misspecified, leading to biased coefficient estimates (Winship and Mare, 1992; Vella, 1998).

Intuitively, correlation between the unobserved error term in the earnings equation, ϵ_i, and both the observed and unobserved factors determining selection will result in a sample that is systematically different from the population as a whole. As the correlation occurs through ϵ_i, which is unobserved by the researcher, these systematic differences in earnings are not explained through the observed characteristics of the sample. The observed characteristics that affect the earnings equation (X_i) are correlated with the unobserved errors ϵ_i, through u_i. The correlation of ϵ_i and u_i essentially acts as an ommited variable that is

correlated with X_i, leading to biased Ordinary Least Squares (OLS) estimates (Vella, 1998). When there is a sample selection effect, if elements of Z_i that are not in X_i are included in the primary earnings equation, they may appear significant even when they do not truly belong in the equation (Heckman, 1979).

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 relies on the joint distribution of the errors of the primary and selection equations being normal. The solution is a two-step procedure, where first a selection equation is estimated using a probit model, and then the coefficients and regressors from this are used to generate the Inverse Mills ratio (IMR) as in Equation (4.2.2),

$$\lambda_i = \frac{\phi(-Z_i\gamma)}{\Phi(Z_i\gamma)} \ . \tag{4.2.2}$$

Where Z_i is the vector of regressors influencing selection, γ is the vector of slope coefficients from the probit model, and λ_i is the IMR for individual i. The IMR is then added to the primary equation as an additional regressor, where it measures the selection bias effects caused by the lack of earnings observations for those that do not work. The inclusion of the IMR should correct for the omitted variable bias caused by the non-random selection, by controlling for the correlation between ϵ_i and u_i (Heckman, 1979; Dolton and Makepeace, 1986).

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-parametric models used (Ahn and Powell, 1993; Martins, 2001; Das et al., 2003). Longitudinal data introduces further issues that can complicate estimation, and a number of

[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.

models developed in the literature extend or adapt Heckman's correction to function with panel data (Wooldridge, 1995; Kyriazidou, 1997; Vella and Verbeek, 1999). Section 4.3 introduces 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. 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.

4.3 Wooldridge and Semykina Sample Selection Bias 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 (4.3.1a), but that the dependent variable Y_{it} is only observed for a subset of the observations. Equation (4.3.1b) 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.

$$Y_{it}^{*} = \rho Y_{i,t-1} + X_{it}\beta + C_{i1} + u_{it1} \ , \tag{4.3.1a}$$

$$S_{it} = 1[Z_{it}\beta_{2t} + C_{i2} + u_{it2} > 0] \ , \tag{4.3.1b}$$

$$Y_{it} = S_{it}Y_{it}^{*} \ . \tag{4.3.1c}$$

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}. Selection into the workforce is determined by Equation (4.3.1b). 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, 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 (4.3.2), 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.

$$Y_{it} = \rho Y_{i,t-1} + X_{it}\beta + c_{i1} + u_{it1} \ , \tag{4.3.2a}$$

$$= \rho^t Y_{i0} + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + c_{i1} \sum_{j=0}^{t-1} \rho^j + e_{it1} \ , \tag{4.3.2b}$$

$$\text{where} \quad e_{it1} = \sum_{j=0}^{t-1} \rho^j u_{i,t-j,1} \ . \tag{4.3.2c}$$

While (4.3.2b) 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}.

$$C_{i1} = \eta_1 + \sum_{s=1}^{T} Z_{is} B_{s1} + \gamma_1 Y_0 + U_{i1} \ , \tag{4.3.3a}$$

$$C_{i2} = \eta_2 + \sum_{s=1}^{T} Z_{is} B_{s2} + \gamma_2 Y_0 + U_{i2} \ . \tag{4.3.3b}$$

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 (4.3.3a) and (4.3.3b), 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 (4.3.4a) and (4.3.4b). It should be noted

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

$$Y_{it} = \rho^t Y_{i0} + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}B_{s1} + \gamma_1 Y_0\Big) + \epsilon_{it1} \ , \tag{4.3.4a}$$

$$S_{it} = 1[Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0 + \epsilon_{it2} > 0] \ , \tag{4.3.4b}$$

$$\epsilon_{it1} = \sum_{s=0}^{t-1} \rho^s (u_{i,t-s,1} + U_{i1}), \ \epsilon_{it2} = U_{i2} + u_{it2}, \ \mathbf{E}(\epsilon_{it1}|Y_{i0}, Z_i, S_{it} = 1) = \upsilon_{2t}\epsilon_{it2} \ . \tag{4.3.4c}$$

While using the model in Equation (4.3.4a) 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 (4.3.4a) 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.

As shown in Section 4.2, $\mathbf{E}(\epsilon_{it1}|Y_{i0}, Z_i, S_{it} = 1) = \mathbf{E}(\upsilon_{2t}\epsilon_{it2}|Y_{i0}, Z_i, S_{it} = 1) = h_{it}(Z_i, Y_0)$, where h_{it} is an unknown function. 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 (4.3.4a).

$$h_{it} = \upsilon_{2t}\lambda_{it} \ , \tag{4.3.5a}$$

$$\lambda_{it} = \text{IMR}_{it}(Z_{it}, Y_0) = \frac{\phi[-(Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0)]}{\Phi[Z_{it}\beta_{2t} + \eta_2 + \sum_{s=1}^{T} Z_{is}B_2 + \gamma_2 Y_0]} \ . \tag{4.3.5b}$$

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 (4.3.5b) and the slope estimates from the probit estimation. 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 (4.3.6a) 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 to use an observation:

$$Y_{it} = \rho^t Y_{i0} + \Big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\Big)\beta + \Big(\frac{1-\rho^t}{1-\rho}\Big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}B_{s1} + \gamma_1 Y_0\Big) + \phi_t \lambda_{it} + \zeta_{it1} \ , \quad (4.3.6a)$$

$$\text{where} \quad \mathbf{E}(\zeta_{it1}|Z_i, Y_{io}, S_{it} = 1) = 0 \ \forall \ t \in 1, .., T. \quad (4.3.6b)$$

This formulation of the model requires that Y_0 is observed for all individuals in the sample, a strict requirement that undermines the models 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 (4.3.7a), where k_s is a vector of slope coefficients,

and this is added to the main equation resulting in (4.3.7b).

$$Y_{i0} = \sum_{s=1}^{T} Z_{is}k_s + b_i \ , \tag{4.3.7a}$$

$$Y_{it} = \rho^t \sum_{s=1}^{T} Z_{is}k_s + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}\delta_{s1}\Big) + \phi_t\lambda_{it} + q_{it1} \ , \tag{4.3.7b}$$

$$\text{where} \quad \mathbf{E}(b_i|Z_i) = 0, \ q_{it1} = \zeta_{it1} + \rho^t b_i \ . \tag{4.3.7c}$$

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 (4.3.8) defines $m_{it}(\theta)$, the conditional expectation of Y_{it},

$$m_{it}(\theta) = m_{it}(z_i, Y_{i0}, S = 1; \theta) \ , \tag{4.3.8a}$$

$$= \rho^t \sum_{s=1}^{T} Z_{is}k_s + \big(\sum_{j=0}^{t-1} \rho^j X_{i,t-j}\big)\beta + \big(\frac{1-\rho^t}{1-\rho}\big)\Big(\eta_1 + \sum_{j=s}^{T} Z_{is}\delta_{s1}\Big) + \phi_t\lambda_{it} \ . \tag{4.3.8b}$$

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 (4.3.9), 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.

$$W_i = \begin{pmatrix} \omega_{i1} & 0 & 0 & \cdots & 0 \\ 0 & \omega_{i2} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & \vdots \\ \vdots & \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & \omega_{iT} \end{pmatrix} . \tag{4.3.9}$$

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 (4.3.10), where $\hat{\Omega}^{-1}$ is a consistent estimator of a positive semi-definite weighting matrix Ω^{-1}.

$$\min_{\theta} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) \hat{\Omega}^{-1} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) . \tag{4.3.10}$$

The first order conditions for this GMM model are given in (4.3.11), where $\nabla_\theta g_i(\theta)$ indicates the first derivative of g_i with respect to the vector of parameters θ,

$$\Big(\sum_{i=1}^{N} W_i' \nabla_\theta g_i(\theta) \Big) \hat{\Omega}^{-1} \Big(\sum_{i=1}^{N} W_i' g_i(\theta) \Big) = 0 . \tag{4.3.11}$$

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 (4.3.12) outlines the asymptotic variance of the GMM estimator,

$$\text{Avar}(\hat{\theta}) = \frac{(\hat{G}' \Omega^{-1} \hat{G})}{N} , \tag{4.3.12a}$$

$$\text{where} \quad \hat{G} = \frac{1}{N} \sum_{i=1}^{N} W_i' \nabla_\theta g_i(\theta) . \tag{4.3.12b}$$

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 (4.3.8),

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 (4.3.13a), where X_i is a vector of time constant variables,

$$(\sum_{j=o}^{t-1} \rho^j X_i)\beta = X_i\beta \sum_{j=0}^{t-1} \rho^j = \frac{1-\rho^t}{1-\rho} X_i\beta \ . \tag{4.3.13a}$$

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.

4.4 SoFIE Analysis

This section will focus on modelling the annual earnings of females workers using the SoFIE data set. Chapter 3 introduced SoFIE, a longitudinal survey data set from New Zealand, and examined descriptive statistics of the demographics and annual earnings of prime aged females in the sample. The variance of their annual earnings was decomposed into its intensive and extensive components, and auto-covariance matrices were used to examine the intertemporal correlation of annual earnings.

In this section annual earnings are explicitly modelled, first using some relatively simple models, and then using Semykina and Wooldridge's new correction for sample selection bias. The simple models will illustrate potential issues that arise when using panel data, and allow

for a comparison between the models that ignore non-random selection, and Semykina and Wooldridge's model that takes it into account. The residuals of each regression will be used to generate auto-covariance matrices similar to those from Section 3.3, allowing us to analyse the structure of the residuals.

The SoFIE extract used is the same as the one used in Section 3.4.2. A few adjustments have been made to the data to ease estimation of the models. While education was largely time invariant, for a small subset of the population it did change over the duration of the sample. To avoid issues this causes, each individual's education in all periods has been set equal to their "highest" level (from lowest to highest: no qualifications, high school level, vocational training, university level). Also, due to either measurement error or the timing of the survey changing for an individual wave to wave, there is a small number of individuals in the sample that do not have their ages increasing by one in each wave. To avoid issues, especially when using the first differenced data, the age in wave one has been assumed correct, and thereafter it increases by one in each wave. Both of theses changes affect only a small number of individuals.

4.4.1 Simple models

In this section three models are used to estimate the annual earnings equation for females from SoFIE. Equation (4.3.1b), the primary earnings equation from Section 4.3, is still the equation of interest, and Equation 4.4.1a is the first differenced earnings equation. Here Y_{it} is the log(earnings) of individual i in time period t, X_{it} is a vector of observed variables that influence annual earnings, and β is the corresponding slope coefficients. 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. There are unobserved, individual specific fixed effects present in C_{i1}, and u_{it1} is the error term.

$$\Delta Y_{it} = \rho \Delta Y_{it-1} + \Delta X_{it}\beta + \Delta u_{it1} \quad . \tag{4.4.1a}$$

First OLS will be used to estimate the model in (4.3.1b). 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 (4.4.1a). 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 4.1 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, although there was no evidence of this in the auto-covariance matrices of annual earnings shown in Section 3.4.4. 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

Table 4.1: *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*

et al., 2011). 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.

Table 4.2 is the auto-covariance matrix of the residuals from the FDIV regression, the auto-covariance matrices for the OLS and FDOLS model residuals are available in Appendix C. The variance of the residuals from the FDIV regression peak in the second wave, and steadily decreases after that. The first order auto-correlations are approximately -0.4, and then very close to zero at higher orders. Likewise, the first order auto-covariances are negative and statistically significant. For all higher orders the auto-correlations are approximately zero.

As demonstrated in Section 3.3, an individual specific effect manifests as persistently positive auto-correlations. The auto-covariances generated with the residuals from the OLS model showed some evidence of a small individual effect (see Table C.1). Examining Table 4.2 there is no evidence of a persistent individual specific effect in the residuals from the

Table 4.2: *Auto-covariance matrix of log(earnings) residuals: unbalanced panel FDIV*

	wave 3	wave 4	wave 5	wave 6	wave 7	wave 8
wave 3	0.818 (0.065)	-0.39	0.00	0.00	-0.03	0.00
wave 4	-0.33 (0.049)	0.851 (0.074)	-0.43	-0.01	0.02	-0.02
wave 5	0.00 (0.022)	-0.37 (0.045)	0.842 (0.066)	-0.41	0.01	0.00
wave 6	0.00 (0.019)	-0.01 (0.019)	-0.33 (0.043)	0.788 (0.069)	-0.40	-0.040
wave 7	-0.02 (0.020)	0.02 (0.020)	0.01 (0.022)	-0.30 (0.037)	0.718 (0.060)	-0.39
wave 8	0.00 (0.014)	-0.01 (0.014)	0.00 (0.020)	-0.03 (0.023)	-0.28 (0.040)	0.715 (0.068)
	$N = 3876$					

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.

FDIV model, so it seems that the differencing has successfully removed any unobserved individual specific component. If ϵ_{it}, the residual from the undifferenced earnings equation, is essentially white noise then we would expect the differenced residuals to have an auto-covariance matrix with first order auto-correlations equal to approximately -0.5, and higher orders that are zero. The first order auto-correlations in Table 4.2 are approximately -0.4, implying that the original errors are not pure white noise, but are relatively close to it.

While the FDIV models seem to 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 4.3 is applied to the SoFIE data, and the results are compared to those from this section.

4.4.2 Wooldridge and Semykina correction

In this section the sample selection bias correction proposed by Semykina and Wooldridge (2013) and introduced in Section 4.3 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 equation, main earnings equation, and auto-covariance matrix of the residuals will be presented.

The earnings and selection equations of interest are (4.3.1b) and (4.3.1c), introduced in Section 4.3. 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 4.3, 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 (4.4.2a) and (4.4.3a) 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.

$$Y_{i0} = \sum_{j=1}^{j=T} Z_{ij} k_j \ , \tag{4.4.2a}$$

$$Y_{i0} = \overline{Z_i^Y} k \ . \tag{4.4.2b}$$

$$C_{i1} = \eta + \sum_{j=1}^{j=T} Z_{ij} \delta_{j1} + \gamma Y_0 \ , \tag{4.4.3a}$$

$$C_{i1} = \eta + \overline{Z_i^C} \delta_1 + \gamma Y_0 \ . \tag{4.4.3b}$$

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, this section 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 (4.4.2b) and (4.4.3b), 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.

The models

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}. 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}

Table 4.3: *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}	$t = 1, ..., 8$
	$Y_{it} = \rho^t \overline{Z_i^Y} k + \sum_{j=0}^{t-1} \left(\rho^j X_{it-j}\right)\beta + \frac{(1-\rho^t)}{(1-\rho)} \left(\overline{Z_i^C}\delta_1\right) + \phi\lambda_{it} + \epsilon_{it}$		

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 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. A summary of the different models and the assumptions made in estimating them is displayed in Table 4.3.

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 4.5 and 4.4 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, similar to the simulated models in Section 3.3, there is inter-temporal correlation in workforce participation, and this correlation is higher in the sub-sample that works in the first period.

Table 4.4: *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 4.5: *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
	$N = 3414$						

Selection models

Table 4.6 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 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 4.6, 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).

Table 4.6: *Selection model results*

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

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

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 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.

Results

Table 4.7 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 Appendix C[3].

The most interesting result in Table 4.7 is that $\hat{\rho}$ is large and statistically significant in each of the four models. In the previous models in Section 4.4, 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,

[3]See Tables C.3 and C.4

Table 4.7: *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 C.4 and C.3 in Appendix C.

it seems that including the individual effects explicitly by modelling C_{i1} has reduced the persistence of earnings. This is similar to the results from Section 4.4, 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.

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 as the estimated coefficient on ρ gets larger, it subsumes the effects of the other parameters.

Auto-covariance matrices

The residuals from each of the four models have been used to generate auto-covariance matrices. Similar to those in Chapter 3 and Section 4.4.1 of this chapter, the upper triangle contains the auto-correlations, the lower triangle the auto-covariances, and the diagonal the variance of the residuals in each wave. The standard errors of the estimates are in parentheses. The sample used seems to have a large impact on the auto-covariance matrix. The patterns observed are relatively similar between the models that use the same sample, so only the auto-covariance matrices for models one and four are presented in this Section. The auto-covariance matrices for models two and three are available in Appendix C.

Table 4.8: *Auto-covariance matrix of model 1 residuals*

	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 2	0.665 (0.046)	0.46	0.31	0.29	0.26	0.22	0.22
Wave 3	0.33 (0.029)	0.799 (0.05)	0.48	0.39	0.35	0.30	0.25
Wave 4	0.23 (0.025)	0.40 (0.027)	0.865 (0.055)	0.50	0.39	0.33	0.31
Wave 5	0.23 (0.023)	0.33 (0.025)	0.44 (0.028)	0.881 (0.062)	0.53	0.42	0.36
Wave 6	0.18 (0.019)	0.27 (0.022)	0.32 (0.020)	0.43 (0.027)	0.753 (0.047)	0.53	0.42
Wave 7	0.17 (0.020)	0.24 (0.021)	0.28 (0.020)	0.36 (0.023)	0.42 (0.023)	0.842 (0.054)	0.56
Wave 8	0.16 (0.017)	0.21 (0.018)	0.26 (0.020)	0.31 (0.024)	0.33 (0.018)	0.46 (0.028)	0.822 (0.054)
	$N = 3414$	$T = 7$					

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, standard errors in parenthesis.

Table 4.8 presents the auto-covariance matrix of the residuals from model one. The first order auto-correlations increase over the duration of the panel, starting at 0.46 when comparing Wave two to wave three, and increasing steadily to 0.56 when comparing wave seven and eight. This increase could be due to economic conditions changing over the course of the panel, leading to the residuals of the model becoming more highly correlated for later waves. It could also be that in later waves a different sample is being used to estimate the correlations, and that this sample has different characteristics.

There is some evidence in Table 4.8 of a persistent, individual specific effect. The auto-correlations steadily decrease as the order examined increases, but appear to be approaching a steady state level (for example, 0.22 for $\text{Cor}(\epsilon_2, \epsilon_8)$, rather than zero. This is interesting, as the level of auto-correlation is much higher than in any of the simple models from Section 4.4.1, even when compared to OLS where the fixed effects have not been differenced out. The auto-covariances follow a similar pattern to the correlations, decreasing steadily as the order examined increases but again do not appear to be approaching zero. All of the variances and co-variances are statistically significant.

As stated above, the auto-covariance matrix for model three[4] is very similar to that of model one. The main differences are that; explicitly modelling the individual specific fixed effects in model three seems to have slightly lowered the variance of the residuals in each wave, and the auto-correlations are slightly lower. This could reflect that explicitly modelling the individual specific fixed effects removes some portion of it from the residuals, lowering the level of correlation. That said, the change is quite small, and the auto-correlations remain relatively large (and still much larger than when OLS was used). It seems that even when the fixed effects are explicitly modelled, that does not remove the permanent component from the residuals.

Table 4.9 is the auto-covariance matrix of the residuals from model four. Model two produces a very similar auto-covariance matrix[5], with slightly higher variances and slightly

[4]See Table C.6
[5]See Table C.5

Table 4.9: *Auto-covariance matrix of model 4 residuals*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	1.276 (0.067)	0.59	0.44	0.40	0.36	0.33	0.33	0.30
Wave 2	0.77 (0.055)	1.310 (0.077)	0.55	0.44	0.42	0.37	0.35	0.33
Wave 3	0.55 (0.030)	0.70 (0.037)	1.240 (0.065)	0.54	0.45	0.40	0.37	0.33
Wave 4	0.51 (0.030)	0.58 (0.031)	0.69 (0.034)	1.295 (0.067)	0.56	0.46	0.41	0.37
Wave 5	0.45 (0.025)	0.53 (0.028)	0.55 (0.029)	0.70 (0.035)	1.206 (0.065)	0.58	0.48	0.41
Wave 6	0.38 (0.022)	0.43 (0.023)	0.45 (0.023)	0.53 (0.027)	0.64 (0.033)	1.020 (0.051)	0.58	0.45
Wave 7	0.39 (0.024)	0.42 (0.022)	0.43 (0.024)	0.49 (0.025)	0.55 (0.028)	0.61 (0.030)	1.090 (0.060)	0.56
Wave 8	0.35 (0.024)	0.39 (0.023)	0.38 (0.021)	0.43 (0.024)	0.46 (0.026)	0.46 (0.024)	0.60 (0.030)	1.036 (0.060)
	$N = 4146$	$T = 8$						

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, standard errors in parenthesis.

lower auto-correlations. The variance of the residuals in each period is approximately twice as large as models one and three. The auto-covariances are also larger than the previous models, and are all significantly different from zero. The auto-correlations follow a similar pattern to models one and three, with the level of auto-correlation decreasing as the order examined increases. There are two primary differences to the earlier models however. First, the initial level of auto-correlation is higher, 0.59 for both models two and four. Second, the level of auto-correlation for a given order does not increase as later waves are examined. For example, the first order auto-correlations show no sign of systematically increasing over the course of the survey.

In Section 4.4, the OLS models showed some evidence of both an MA(1) and an unobserved, individual specific fixed effect in the residuals. The FDIV model appeared to successfully remove the fixed effect, with auto-correlations higher than order two approximately equal to zero. The four models estimated using the Semykina and Wooldridge sample selection bias correction do not show evidence of an MA(1) in the residuals, but

there is evidence of an individual specific fixed effect. Even when C_{i1} is explicitly modelled, the auto-covariance matrices show strong evidence of a permanent component, and actually have significantly higher levels of auto-correlation than the OLS models, which did not attempt to take the fixed effects into account.

It is possible that non-random selection into the workforce led to earlier models where the residuals had lower levels of auto-correlation. But it is interesting that even when the fixed effects are explicitly modelled through C_{i1}, there is almost no change in the auto-correlations. It is possible that the portion of C_{i1} that is correlated with the observed variables is being controlled for, resulting in unbiased coefficient estimates, and that there is a remaining fraction that is individual specific, but uncorrelated with the observed regressors.

$$C_{i1} = \overline{Z_i^C}\delta_1 + \alpha_i \ . \tag{4.4.4a}$$

This is illustrated in 4.4.4, where α_i is the portion of the individual specific fixed effect that is uncorrelated with the observed variables. If this is correct, the estimation would not be biased by endogeneity, but individual fixed effects would still be evident in the auto-correlations of the residuals.

4.4.3 Discussion

This chapter 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 sub-sample 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 chapter 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.

An interesting result from this model was the evidence of permanent components in the auto-covariance matrices of the residuals from the sample selection correct models. While there was some evidence of an unobserved fixed effect in the OLS model, the effect was much weaker. As we do not have to difference in the Semykina and Wooldridge correction, we have not removed all of the fixed effects, but we are explicitly modelling the conditional expectation of C_{i0}. This means that modelling the individual component in this way has not removed the permanent component entirely. As long as this method controls for the fraction of the individual specific component that is correlated with the regressors, this should still remove the endogeneity. If this is correct, it provides a way to estimate the model without endogeneity, similar to the FDIV model, but also allows us to examine the individual specific fixed effect, where differencing would have removed it entirely.

4.4.4 Extensions

There are a number of changes and potential extensions that could be made to Section 4.4.2s application of Semykina and Wooldridge's sample selection bias correction. In this chapter, a simplified version of the sample selection bias correction was applied. The full model, where all observed regressors are used to model both Y_{i0} and C_{i1}, had a large number of parameters and had numerous estimation issues. A good first extension would be applying the full model, and comparing the results to see if using Chamberlain's method changes the results versus using the Mundlak method as we did in this Chapter.

There has been a range of work investigating if unemployment "scars" workers, reducing their probability of working in future periods and lowering their expected wage/earnings (Ruhm, 1991; Arulampalam et al., 2001; Gangl, 2006). An interesting extension of the Semykina and Wooldridge correction would be adding in employment dynamics, potentially to

both the selection and earnings equations. This could show how unemployment spells affect future employment prospects, and lead to lower earnings. The auto-regressive coefficient in the models of this section represents the persistence of earnings, and it would be an interesting extension to see how unemployment or demographic characteristics influence earnings persistence.

This chapter has presented the results of applying the Semykina and Wooldridge correction to the SoFIE data, including auto-covariance matrices of the residuals of the models. While this has allowed for some simple analysis, a more in-depth investigation that estimates the residual structure would be of great interest. The residuals of the GMM estimation showed evidence of an individual specific effect remaining, even after introducing modelling of C_{i1}. Investigating this by estimating an Error Components Model (ECM) could reveal the structure and magnitude of the remaining unobserved permanent component.

4.5 Conclusions

This chapter focused on modelling the annual earnings 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.

A number of simple models were applied to the data sample from SoFIE in order to estimate the earnings equation for the female workers. These methods (OLS, FDOLS, and FDIV) did not control for sample selection bias, but their results did indicate the potential presence of unobserved, individual specific fixed effects. Removing the individual specific effects through first differencing revealed a relatively low, but positive and significant level of earnings persistence.

The sample selection bias correct regression results are broadly similar to those in Se-

mykina and Wooldridge (2013). Semykina and Wooldridge focused on a subset of females from the Panel Survey of Income Dynamics (PSID) that worked in the first period, while in this chapter we estimated the earnings equation using both the subset of females from SoFIE that worked in the first period, and the full sample. When using the subsample that work in the first period, our results are very similar to those in Semykina and Wooldridge (2013). On the other hand, when we instead use the full sample of females from SoFIE we find a much higher level of earnings persistence, with $\hat{\rho}$ not significantly different from one. This indicates that annual earnings may not be stationary for females in the SoFIE sample, and also provides evidence of systematic differences between the different samples.

The coefficient on the IMR was only significant in the most simple of the selection correct models. In each case it was negative, with its impact increasing slightly when the full sample was used. While the lack of statistical significance may indicate that sample selection bias is not a large issue for female workers in SoFIE, the differences between the models that take non-random selection into account, and those that do not, are relatively large. Likewise, the 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, indicates that there are systematic differences between workers and non-workers that should be taken into account.

The auto-correlations of the residuals from each of the sample selection bias correct models indicate the presence of an individual specific fixed effect. The level of auto-correlation was much higher than in any of the simple models even when C_{i0}, the individual specific fixed effect, is explicitly modelled. The level of auto-correlation was also higher in the models that used the entire sample, again implying significant differences between the samples. It is an interesting result that even when the individual specific fixed effects were explicitly modelled, the auto-covariance matrices still showed strong evidence of a permanent component in the residuals. If this method does remove the endogeneity that can be caused by unobserved, individual specific fixed effects, then it is an interesting alternative to differencing that allows

for some analysis of how strong the individual specific component is.

This chapter 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. The auto-covariance matrices produced from the residuals of the models strongly indicate the presence of unobserved, individual specific fixed effects, even after C_{i0} is explicitly modelled. In combination with the high levels of earnings persistence, this indicates 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.

Chapter 5

Conclusion

This thesis has examined the challenges of modelling the annual earnings of female workers. The previous earnings literature has tended to focus on the annual earnings of male workers, and the workforce participation decision for females. This has largely been driven by the data available. Female workers have lower participation rates, so more periods where their earnings are zero. This means that there is a greater risk of sample selection bias when working with female earnings, and that understanding the selection process is more important.

Female workers' lower participation rates lead to two key issues. First, when using the logarithm of earnings zero earnings observations can not be included. The exclusion of zero earnings observations leads to descriptive statistics that are misleading, and ignores the large impact that periods of unemployment have on an individuals earnings. Second, if selection into the workforce is non-random then modelling annual earnings based on only the portion of the sample that works will result in biased coefficient estimates. This thesis applied two methods that alleviate these issues. The Inverse Hyperbolic Sine (IHS) function was introduced as a potential replacement to the logarithm. This function has many properties that are similar to the logarithm, but is also defined for zero values, allowing for descriptive statistics that take into account periods of unemployment. However, the IHS does not model

the behavioural workforce participation decision. To correct for sample selection bias, a new method proposed by Semykina and Wooldridge (2013) was introduced, and applied to female workers from the Survey of Families, Income, and Employment (SoFIE) data set.

Chapter 2 introduced the IHS. This function has a number of useful properties, such as reducing the impact of extreme values, and becomes log-like in the right tail. Use of the IHS requires a value for θ, a scaling parameter that influences how quickly the function changes from linear to log-like. Typically in the literature that uses the IHS, θ is either estimated using Maximum Likelihood Estimation (MLE), or selected by the researcher. The results in Chapter 2 clearly show that, contrary to the existing literature, both of these methods for selecting θ can be problematic. The estimation of θ is be unreliable when the bulk of the data lies solely in either the log-like or linear regions of the function, with estimated $\hat{\theta}$ values quite different to those used in generating the data. Our results also showed that this misestimation did not lead to incorrect elasticities when working with data in the same region as was used in estimation, but that extending the results to model observations that were above or below the data used could lead to incorrect inference. On the other hand, if θ is selected by the researcher this can result in a misspecified model that produces elasticities that are very different from the true values. Additionally, when using data that contains censored observations $\hat{\theta}$ estimation was shown to be very unreliable, and the results produced using $\hat{\theta}$ led to elasticities that were different from the true model.

Chapter 3 focused on annual earnings and earnings dynamics. The IHS and logarithm were applied to both simulated and empirical data on prime aged females from SoFIE, and were used to generate a range of descriptive statistics. The core findings from this application were that the IHS seems to perform as well as the logarithm in reducing the impact of outliers present in the untransformed earnings data, but also includes periods of unemployment in its generation of sample statistics, and allows for the decomposition of the variance of annual earnings into its extensive and intensive margin components. As compared to the decomposition of the variance of untransformed earnings, the IHS finds that extensive

margin changes have a much larger impact on the overall variance of earnings experienced by the females in the SoFIE sample. While the results indicate that the IHS could be very useful when working with earnings data by allowing for the inclusion of zero earnings observation, the choice of θ can lead to quite different results in the variance decomposition, with the fraction of variance attributed to the intensive and extensive margins changing with different values of θ. As was shown in Chapter 2, θ estimation can be unreliable, especially when some of the data is censored. At this stage it is not clear how the correct value for θ can be estimated or chosen by the researcher, and this undermines the IHS function's ability to be a useful transformation when working with annual earnings data.

Chapter 4 focused on explicitly modelling the earnings function of female workers. Given that female workers have much lower rates of workforce participation than male workers, sample selection bias is a major risk. This chapter introduces a new sample selection bias correction proposed by Semykina and Wooldridge (2013), that both builds on Heckman's original model and incorporates a number of extensions from the literature. Chapter 4 first applies a number of basic models to the SoFIE data. The results indicate that there is a reasonable level of earnings persistence. Four versions of Semykina and Wooldridge's correction are then applied to the data. Interestingly, even after modelling the fixed effects explicitly, the auto-covariance matrices of the residuals still indicate that the presence of an individual specific component. This result is quite different from that achieved with first differencing, but as long as modelling the fixed effects has removed the endogeneity, should not lead to biased coefficient estimation. All four versions of Semykina and Wooldridge's model that are applied show strong levels of earnings persistence, in each case the auto-regressive coefficient is larger than any of the simple models. In the most developed of these models, the auto-regressive coefficient is larger than one, so there is some evidence that earnings are non-stationary. While the Inverse Mills ratio (IMR), the component added to the regression to control for non-random selection, is not statistically significant in three of the four models, the results are quite different from the simple models with higher levels

of earnings persistence. Also, the earlier auto-covariance matrices from Chapter 3, and the participation patterns of the two samples used in Chapter 4 indicate that there are systematic differences between workers and non-workers. This means that controlling for non-random selection into the workforce is very important when modelling the annual earnings of female workers.

This thesis provides novel results in three areas. First, we have highlighted estimation issues that can arise when working with the IHS. The poor identification issues that complicate the estimation required to use the IHS have not been detailed in the existing literature, and this is an important issue going forward for any researchers that want to use the IHS. Second, we have shown how the IHS can be used to analyse the annual earnings of female workers. The use of the IHS allows for the creation of descriptive statistics that include periods of unemployment and extensive margin changes. It also allows for the decomposition of the variance of earnings, while controlling the impact that extreme values have. The caveat being that while it has useful properties, correct selection/estimation of θ is problematic. Third, we have applied the new sample selection bias correction proposed in Semykina and Wooldridge (2013) to the SoFIE data. The results found much higher levels of earnings persistence than was present in the simple models that ignored selection. When using the full sample of females, the AR(1) parameter ρ was approximately equal to one, indicating that annual earnings may be non-stationary. While the IMR, the parameter added to the earnings equation to control for sample selection bias, was only significant in one of the four models, the results were still significantly different from the simple models. Likewise, the different participation patterns observed between the two samples used indicated that there were systematic differences between workers and non-workers. These results indicate that when modelling the annual earnings of female workers, non-random selection could well be an issue, and should be corrected for.

Appendix A

Chapter 2

Table A.1: *Summary of simulations with* $\theta = 0.5$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max[1]	0.944	1	1	0.386	0.478	0.534
$\bar{\hat{\theta}}_{\text{MLE}}$[2]	0.647 (0.701)	0.516 (0.087)	0.505 (0.034)	0.182 (0.148)	0.271 (0.218)	0.419 (0.384)
%Δ: median[3]	0.002 (0.010)	0 (0.005)	0 (0.002)	-0.002 (0.004)	-0.001 (0.002)	0 (0.001)
%Δ: 10th percentile[3]	-0.008 (0.056)	0 (0.026)	-0.001 (0.012)	0.009 (0.021)	0.005 (0.010)	0.001 (0.004)
%Δ: 90th percentile[3]	0.002 (0.012)	0 (0.006)	0 (0.003)	-0.002 (0.005)	-0.001 (0.003)	0 (0.001)
%Δ: below min[3]	-0.047 (0.212)	-0.003 (0.099)	-0.004 (0.043)	0.047 (0.101)	0.025 (0.054)	0.007 (0.024)
%Δ: above max[3]	0.002 (0.012)	0 (0.006)	0 (0.003)	-0.002 (0.005)	-0.001 (0.003)	0 (0.001)

[1] *Fraction of simulations resulting in maximised concentrated log-likelihood*
[2] *Mean* $\hat{\theta}$ *based on simulations where concentrated log-likehood is maximised.*
[3] *% difference between elasticities calcualted using true model versus the estimated model.*

Table A.2: *Elasticity correlations:* $\theta = 0.5$ *versus alternative values*

	N=250	N=1000	N=5000
	$\alpha = 1$		
$\hat{\theta}$	0.998	0.999	1
$\theta = 1$	0.998	0.977	0.966
$\theta = 0.1$	0.589	0.636	0.715
$\theta = 0.05$	0.100	0.331	0.553
$\theta = 0.01$	−0.239	0.103	0.420
	$\alpha = 5$		
$\hat{\theta}$	1	1	1
$\theta = 1$	1	1	1
$\theta = 0.1$	1	0.999	1
$\theta = 0.05$	0.995	0.995	0.995
$\theta = 0.01$	0.367	0.353	0.382

Table A.3: *Average absolute percentage difference:* $\theta = 0.5$

	N=250	N=1000	N=5000
	$\alpha = 1$		
$\hat{\theta}$	2.0	0.9	0.4
$\theta = 1$	4.0	4.0	4.0
$\theta = 0.1$	37.4	37.3	37.5
$\theta = 0.05$	78.1	78.0	78.3
$\theta = 0.01$	222.4	222.2	223.1
	$\alpha = 5$		
$\hat{\theta}$	0.6	0.3	0.2
$\theta = 1$	0.1	0.1	0.1
$\theta = 0.1$	2.2	2.2	2.2
$\theta = 0.05$	7.4	7.4	7.4
$\theta = 0.01$	61.5	61.7	61.5

Table A.4: *Summary of simulations with* $\theta = 0.1$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max[1]	1	1	1	1	1	1
$\bar{\hat{\theta}}_{\text{MLE}}$[2]	0.102 (0.019)	0.100 (0.009)	0.100 (0.004)	0.103 (0.026)	0.100 (0.011)	0.100 (0.005)
%Δ: median[3]	0 (0.004)	0 (0.001)	0 (0)	0 (0.005)	0 (0.002)	0 (0.001)
%Δ: 10th percentile[3]	−0.005 (0.048)	0 (0.022)	0 (0.010)	0.002 (0.056)	0.001 (0.027)	0.001 (0.012)
%Δ: 90th percentile[3]	0.005 (0.049)	0 (0.023)	0 (0.011)	−0.001 (0.044)	−0.001 (0.021)	−0.001 (0.010)
%Δ: below min[3]	-0.007 (0.060)	0 (0.027)	0.001 (0.013)	0 (0.083)	0.001 (0.040)	0.001 (0.018)
%Δ: above max[3]	0.005 (0.082)	−0.001 (0.037)	−0.001 (0.018)	−0.004 (0.066)	−0.002 (0.032)	−0.001 (0.014)

[1] *Fraction of simulations resulting in maximised concentrated log-likelihood*
[2] *Mean* $\hat{\theta}$ *based on simulations where concentrated log-likehood is maximised.*
[3] *% difference between elasticities calcualted using true model versus the estimated model.*

Table A.5: *Elasticity correlations:* $\theta = 0.1$ *versus alternative values*

	N=250	N=1000	N=5000
	$\alpha = 1$		
$\hat{\theta}$	0.993	0.999	1
$\theta = 1$	0.598	0.548	0.572
$\theta = 0.5$	0.694	0.690	0.747
$\theta = 0.05$	0.942	0.970	0.988
$\theta = 0.01$	0.876	0.939	0.976
	$\alpha = 5$		
$\hat{\theta}$	0.998	1	1
$\theta = 1$	0.989	0.989	0.989
$\theta = 0.5$	0.989	0.989	0.989
$\theta = 0.05$	0.975	0.975	0.975
$\theta = 0.01$	0.860	0.860	0.859

Table A.6: *Average absolute percentage difference:* $\theta = 0.1$

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	2.5	1.1	0.5
$\theta = 1$	37.2	37.5	37.4
$\theta = 0.5$	29.7	29.8	29.8
$\theta = 0.05$	8.5	8.5	8.5
$\theta = 0.01$	12.9	12.9	12.9
		$\alpha = 5$	
$\hat{\theta}$	2.5	1.2	0.5
$\theta = 1$	13.6	13.6	13.6
$\theta = 0.5$	12.7	12.7	12.7
$\theta = 0.05$	9.6	9.6	9.6
$\theta = 0.01$	17.9	18.0	18.0

Table A.7: *Summary of simulations with* $\theta = 0.05$

	$\alpha = 1$			$\alpha = 5$		
	$N = 250$	$N = 1000$	$N = 5000$	$N = 250$	$N = 1000$	$N = 5000$
MLE max	0.860	0.994	1	0.966	0.998	1
$\bar{\hat{\theta}}_{\text{MLE}}$[2]	0.054 (0.019)	0.050 (0.011)	0.050 (0.005)	0.051 (0.019)	0.050 (0.009)	0.050 (0.004)
%Δ: median[3]	0 (0.005)	0 (0.002)	0 (0.001)	0 (0.004)	0 (0.001)	0 (0)
%Δ: 10th percentile[3]	−0.002 (0.038)	−0.001 (0.019)	0 (0.009)	−0.001 (0.046)	0.001 (0.023)	0 (0.010)
%Δ: 90th percentile[3]	0.003 (0.049)	0.002 (0.025)	0 (0.011)	0.001 (0.050)	-0.001 (0.025)	0 (0.011)
%Δ: below min[3]	−0.003 (0.045)	−0.002 (0.022)	0 (0.010)	−0.003 (0.062)	0.001 (0.031)	0 (0.013)
%Δ: above max[3]	0.002 (0.103)	0.002 (0.053)	−0.001 (0.024)	−0.002 (0.094)	−0.003 (0.048)	0 (0.020)

[1] *Fraction of simulations resulting in maximised concentrated log-likelihood*
[2] *Mean* $\hat{\theta}$ *based on simulations where concentrated log-likehood is maximised.*
[3] *% difference between elasticities calcualted using true model versus the estimated model.*

Table A.8: *Elasticity correlations:* $\theta = 0.05$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	0.993	0.999	1
$\theta = 1$	0.410	0.416	0.502
$\theta = 0.5$	0.521	0.564	0.687
$\theta = 0.1$	0.959	0.976	0.991
$\theta = 0.01$	0.991	0.995	0.998
		$\alpha = 5$	
$\hat{\theta}$	0.997	0.999	1
$\theta = 1$	0.954	0.954	0.954
$\theta = 0.5$	0.956	0.956	0.956
$\theta = 0.1$	0.987	0.987	0.987
$\theta = 0.01$	0.986	0.985	0.985

Table A.9: *Average absolute percentage difference:* $\theta = 0.05$

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	2.3	1.1	0.5
$\theta = 1$	43.4	43.5	43.5
$\theta = 0.5$	35.6	35.7	35.7
$\theta = 0.1$	7.2	7.2	7.2
$\theta = 0.01$	3.5	3.5	3.5
		$\alpha = 5$	
$\hat{\theta}$	2.5	1.2	0.5
$\theta = 1$	20.7	20.7	20.7
$\theta = 0.5$	19.8	19.8	19.8
$\theta = 0.1$	7.6	7.6	7.6
$\theta = 0.01$	5.3	5.3	5.3

Table A.10: *Summary of* 10% *censored simulations with* $\theta = 1$

	$\alpha = 1$			$\alpha = 5$		
	N=250	N=1000	N=5000	N=250	N=1000	N=5000
MLE max[1]	0.266	0.074	0.002	0	0	0
$\bar{\hat{\theta}}_{\text{MLE}}$[2]	0.615 (0.249)	0.612 (0.058)	0.622			
%Δ: median[3]	−0.5 (3.3)	0.5 (1.9)	0.9 (0.9)	−0.5 (4.1)	0 (2.0)	0 (0.9)
%Δ: 10th percentile[3]						
%Δ: 90th percentile[3]	−0.5 (3.3)	0.5 (1.9)	0.9 (0.9)	−0.5 (4.1)	0 (2.0)	0 (0.9)
%Δ: below min[3]	−4.1 (24.3)	−13.2 (10.5)	−15.8 (1.8)	−0.5 (4.1)	0 (2.0)	0 (0.9)
%Δ: above max[3]	−0.5 (3.3)	0.5 (1.9)	0.9 (0.9)	−0.5 (4.1)	0 (2.0)	0 (0.9)

[1] *Fraction of simulations resulting in maximised concentrated log-likelihood*
[2] *Mean* $\hat{\theta}$ *based on simulations where concentrated log-likehood is maximised.*
[3] *% difference between elasticities calcualted using true model versus the estimated model.*

Table A.11: *Elasticity correlations with censored data:* $\theta = 1$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	0.971	0.935	0.845
$\theta = 0.5$	0.988	0.981	0.967
$\theta = 0.1$	0.659	0.611	0.613
$\theta = 0.05$	0.302	0.330	0.464
$\theta = 0.01$	-0.143	0.036	0.324
		$\alpha = 5$	
$\hat{\theta}$	1	1	1
$\theta = 0.5$	1	1	1
$\theta = 0.1$	1	1	1
$\theta = 0.05$	1	1	1
$\theta = 0.01$	0.994	0.993	0.994

Table A.12: *Average absolute percentage difference:* 10% *censored sample* $\theta = 1$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	5.2	4.5	4.1
$\theta = 0.5$	2.3	2.2	2.2
$\theta = 0.1$	28.1	27.6	27.5
$\theta = 0.05$	63.4	62.3	62.2
$\theta = 0.01$	313.2	308.9	308.1
		$\alpha = 5$	
$\hat{\theta}$	3.4	1.6	0.7
$\theta = 0.5$	0.5	0.2	0.1
$\theta = 0.1$	1.5	0.8	0.4
$\theta = 0.05$	2.1	1.1	0.6
$\theta = 0.01$	5.6	4.1	3.5

Table A.13: *Elasticity correlations:* 10% *censored sample* $\theta = 0.1$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	0.874	0.930	0.972
$\theta = 1$	0.589	0.546	0.566
$\theta = 0.5$	0.685	0.685	0.739
$\theta = 0.05$	0.943	0.967	0.987
$\theta = 0.01$	0.877	0.932	0.973
		$\alpha = 5$	
$\hat{\theta}$	0.847	0.847	0.847
$\theta = 1$	0.989	0.989	0.989
$\theta = 0.5$	0.989	0.989	0.989
$\theta = 0.05$	0.975	0.976	0.975
$\theta = 0.01$	0.859	0.860	0.859

Table A.14: *Average absolute percentage difference:* 10% *censored sample* $\theta = 0.1$ *versus alternative values*

	N=250	N=1000	N=5000
		$\alpha = 1$	
$\hat{\theta}$	13.6	13.6	13.5
$\theta = 1$	37.8	37.6	37.5
$\theta = 0.5$	29.9	29.8	29.8
$\theta = 0.05$	8.5	8.5	8.5
$\theta = 0.01$	12.9	12.9	12.9
		$\alpha = 5$	
$\hat{\theta}$	18.5	18.5	18.5
$\theta = 1$	17.3	14.3	13.7
$\theta = 0.5$	14.3	13.0	12.8
$\theta = 0.05$	9.7	9.7	9.6
$\theta = 0.01$	18.1	18.0	18.0

Appendix B

Chapter 3

Table B.1: *Unbalanced log(earnings) participation rates:* 10% *censoring*

T	1	2	3	4	5	6	7	8	9	10
Censored based on transitory shock										
1	0.90									
2	0.86	0.90								
3	0.85	0.86	0.90							
4	0.84	0.85	0.86	0.90						
5	0.84	0.84	0.85	0.86	0.90					
6	0.83	0.84	0.84	0.85	0.86	0.90				
7	0.83	0.83	0.84	0.84	0.85	0.86	0.90			
8	0.82	0.83	0.83	0.84	0.84	0.85	0.86	0.90		
9	0.82	0.82	0.83	0.83	0.84	0.84	0.85	0.86	0.90	
10	0.82	0.82	0.82	0.83	0.83	0.84	0.84	0.85	0.86	0.90
Censored on α_i and transitory shock										
1	0.90									
2	0.88	0.90								
3	0.87	0.88	0.90							
4	0.87	0.87	0.88	0.90						
5	0.87	0.87	0.87	0.88	0.90					
6	0.86	0.87	0.87	0.87	0.88	0.90				
7	0.86	0.86	0.87	0.87	0.87	0.88	0.90			
8	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.90		
9	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.90	
10	0.86	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.88	0.90

Fraction of observations used in each time period/time period pair

Table B.2: *SoFIE summary statistics: Males*

	Full sample		Sub-sample^{+}	
	Unweighted	Weighted	Unweighted	Weighted
Sample size	3693	3693	3537	3537
Weighted sample size	-	714162.5	-	684775
Fraction working	0.91	0.91	0.91	0.91
Mean(earnings*)	53.006 (113.979)	52.758 (107.825)	53.834 (75.248)	53.731 (80.310)
Mean(earnings*|working)	58.230 (118.185)	57.855 (111.601)	58971.68 (76.808)	58783.79 (82.214)
Age	43.33	42.43	43.24	42.32
Age | working	43.12	42.24	43.03	42.14
Fraction with children	0.51	0.50	0.51	0.5
Fraction with children | working	0.52	0.51	0.52	0.51
Fraction with partner	0.80	0.79	0.80	0.79
Fraction with partner | working	0.82	0.81	0.82	0.81
School	0.22	0.23	0.22	0.23
School | working	0.22	0.23	0.22	0.23
Vocational	0.42	0.41	0.42	0.41
Vocational | working	0.42	0.41	0.42	0.41
University	0.19	0.21	0.20	0.21
University | working	0.20	0.21	0.20	0.21

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

Table B.3: *Variance decomposition of earnings: males from SoFIE*

	N	mean(earn)	mean(earn|working)	Periods worked	Intensve variance	Extensive variance	Fraction intensive
			Earnings*				
8 periods	0.785	60.149	60.149	8	2912.17	0	1
7 periods	0.083	46.932	53.636	7	1854.18	930.66	0.55
6 periods	0.037	44.039	58.718	6	5826.28	4311.54	0.43
5 periods	0.020	27.933	44.694	5	2709.38	1539.95	0.38
4 periods	0.014	21.668	43.336	4	373.94	1157.54	0.31
3 periods	0.012	11.642	31.045	3	1680.57	943.15	0.38
2 periods	0.008	4.317	17.269	2	50.86	129.25	0.32
1 periods	0.007	1.882	15.055	1	-	63.87	-
0 periods	0.034	0	-	0	-	-	-
Average		53.833	57.921	7.3	2855.93	306.36	0.9
			IHS(earnings, $\theta = 1$)				
8 periods	0.785	11.46	11.46	8	0.26	0	1
7 periods	0.083	9.7	11.09	7	0.64	13.51	0.04
6 periods	0.037	8.15	10.87	6	0.7	22.34	0.03
5 periods	0.020	6.67	10.68	5	0.42	26.99	0.02
4 periods	0.014	5.38	10.77	4	0.46	29.25	0.02
3 periods	0.012	3.71	9.89	3	0.41	23.33	0.02
2 periods	0.008	2.33	9.34	2	0.34	16.8	0.03
1 periods	0.007	1.16	9.28	1	-	9.72	-
0 periods	0.034	0	-	0	-	-	-
Average		10.39	11.33	7.3	0.32	3.48	0.83
IHS(earnings, $\theta = 0.1$)	N=3537	82.87	90.23	7.3	31.75	215.74	0.83
IHS(earnings, $\theta = 0.001$)*	N=3537	4.091	4.428	7.3	0.265	0.473	0.85

** in thousands*

Table B.4: *Auto-covariance matrix of IHS(earnings, $\theta = 0.1$): females from SoFIE*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	1319.99 (23.026)	0.78	0.66	0.58	0.53	0.49	0.46	0.45
Wave 2	991.8 (22.839)	1236.39 (24.242)	0.8	0.67	0.61	0.55	0.51	0.49
Wave 3	826.75 (22.545)	968.09 (23.367)	1185.25 (24.685)	0.79	0.69	0.61	0.56	0.53
Wave 4	713.8 (22.207)	794.64 (22.813)	918.06 (23.441)	1149.43 (25.12)	0.82	0.71	0.64	0.59
Wave 5	659.47 (22.358)	733.75 (22.892)	815.55 (23.238)	946.55 (24.066)	1173.10 (25.561)	0.82	0.71	0.64
Wave 6	602.97 (22.101)	659.16 (22.457)	712.08 (22.742)	818.16 (23.525)	954.71 (24.513)	1146.81 (26.062)	0.82	0.71
Wave 7	574.79 (22.145)	618.02 (22.361)	664.89 (22.525)	741.7 (23.022)	833.07 (23.816)	953.66 (24.651)	1171.88 (26.09)	0.82
Wave 8	560.67 (22.295)	593.01 (22.475)	630.03 (22.493)	697.45 (22.881)	758.95 (23.475)	835.83 (23.965)	977.18 (24.931)	1200.85 (26.206)
	$N = 4464$	$T = 8$						

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.

Table B.5: *Auto-covariance matrix of IHS(earnings, $\theta = 0.001$)*: females from SoFIE*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	3.315 (0.046)	0.81	0.70	0.63	0.58	0.54	0.51	0.48
Wave 2	2.63 (0.048)	3.183 (0.049)	0.82	0.71	0.65	0.60	0.56	0.53
Wave 3	2.25 (0.049)	2.58 (0.050)	3.073 (0.050)	0.82	0.73	0.66	0.61	0.57
Wave 4	2.00 (0.050)	2.19 (0.050)	2.5 (0.050)	3.04 (0.052)	0.84	0.74	0.68	0.64
Wave 5	1.85 (0.051)	2.03 (0.051)	2.24 (0.051)	2.56 (0.052)	3.083 (0.053)	0.84	0.74	0.68
Wave 6	1.70 (0.051)	1.86 (0.051)	2.00 (0.051)	2.25 (0.052)	2.57 (0.053)	3.017 (0.055)	0.84	0.75
Wave 7	1.62 (0.051)	1.75 (0.051)	1.88 (0.051)	2.09 (0.052)	2.30 (0.053)	2.58 (0.054)	3.096 (0.055)	0.84
Wave 8	1.57 (0.052)	1.67 (0.052)	1.78 (0.052)	1.97 (0.052)	2.12 (0.053)	2.31 (0.054)	2.63 (0.055)	3.148 (0.056)
$N = 4464$ $T = 8$								

Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.
** in thousands*

Appendix C

Chapter 4

Table C.1: *Auto-covariance matrix of log(earnings) residuals: Unbalanced panel using OLS*

	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 2	0.630 (0.044)	-0.09	0.01	0.08	0.06	0.07	0.09
Wave 3	-0.06 (0.022)	0.642 (0.045)	-0.09	0.03	0.05	0.04	0.04
Wave 4	0.01 (0.015)	-0.06 (0.026)	0.685 (0.051)	-0.11	0.01	0.07	0.08
Wave 5	0.05 (0.015)	0.02 (0.014)	-0.07 (0.025)	0.620 (0.045)	-0.15	0.05	0.06
Wave 6	0.04 (0.011)	0.03 (0.013)	0.01 (0.013)	-0.09 (0.026)	0.547 (0.039)	-0.10	-0.01
Wave 7	0.04 (0.012)	0.02 (0.013)	0.05 (0.013)	0.03 (0.015)	-0.06 (0.020)	0.568 (0.044)	-0.08
Wave 8	0.05 (0.011)	0.02 (0.010)	0.05 (0.013)	0.03 (0.016)	0.00 (0.013)	-0.04 (0.024)	0.539 (0.049)
Sample size:	$N = 4050$		$T = 1 - 7$		Total observations= $24,075$		

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.

Table C.2: *Auto-covariance matrix of log(earnings) residuals: unbalanced panel FDOLS*

	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 3	0.632 (0.048)	-0.09	-0.15	-0.03	-0.05	-0.03
Wave 4	-0.06 (0.031)	0.648 (0.051)	-0.10	-0.19	-0.04	-0.01
Wave 5	-0.10 (0.019)	-0.06 (0.027)	0.618 (0.044)	-0.08	-0.15	-0.04
Wave 6	-0.02 (0.014)	-0.11 (0.018)	-0.05 (0.028)	0.572 (0.043)	-0.05	-0.20
Wave 7	-0.03 (0.015)	-0.02 (0.014)	-0.09 (0.019)	-0.03 (0.024)	0.552 (0.043)	-0.07
Wave 8	-0.02 (0.011)	-0.01 (0.011)	-0.02 (0.016)	-0.11 (0.018)	-0.04 (0.026)	0.548 (0.053)
	$N = 3876$	$T = 6$				

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations.

Table C.3: *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 C.4: *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)	

Table C.5: *Auto-covariance matrix of model 2 residuals*

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 1	1.311 (0.067)	0.59	0.44	0.39	0.35	0.31	0.31	0.28
Wave 2	0.78 (0.054)	1.318 (0.077)	0.55	0.44	0.41	0.36	0.34	0.31
Wave 3	0.56 (0.030)	0.70 (0.037)	1.239 (0.066)	0.54	0.45	0.39	0.36	0.32
Wave 4	0.51 (0.030)	0.57 (0.031)	0.69 (0.034)	1.291 (0.066)	0.56	0.46	0.41	0.37
Wave 5	0.44 (0.025)	0.52 (0.028)	0.55 (0.029)	0.70 (0.035)	1.206 (0.064)	0.58	0.48	0.41
Wave 6	0.36 (0.022)	0.42 (0.023)	0.44 (0.023)	0.53 (0.027)	0.65 (0.033)	1.023 (0.051)	0.58	0.45
Wave 7	0.37 (0.024)	0.41 (0.022)	0.42 (0.024)	0.49 (0.025)	0.55 (0.028)	0.61 (0.030)	1.099 (0.060)	0.57
Wave 8	0.32 (0.025)	0.37 (0.023)	0.37 (0.021)	0.43 (0.025)	0.46 (0.026)	0.47 (0.024)	0.61 (0.030)	1.058 (0.060)
	$N = 4146$	$T = 8$						

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, standard errors in parenthesis.

Table C.6: *Auto-covariance matrix of model 3 residuals*

	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8
Wave 2	0.641 (0.044)	0.43	0.28	0.28	0.24	0.21	0.21
Wave 3	0.30 (0.024)	0.760 (0.048)	0.45	0.37	0.33	0.29	0.25
Wave 4	0.21 (0.022)	0.36 (0.023)	0.838 (0.054)	0.49	0.38	0.32	0.30
Wave 5	0.20 (0.021)	0.30 (0.023)	0.42 (0.027)	0.864 (0.062)	0.52	0.41	0.36
Wave 6	0.16 (0.016)	0.25 (0.020)	0.30 (0.019)	0.41 (0.027)	0.736 (0.047)	0.52	0.41
Wave 7	0.16 (0.018)	0.23 (0.019)	0.26 (0.018)	0.35 (0.022)	0.41 (0.022)	0.828 (0.053)	0.55
Wave 8	0.15 (0.016)	0.19 (0.017)	0.25 (0.019)	0.30 (0.023)	0.32 (0.018)	0.45 (0.028)	0.809 (0.054)
	$N = 3414$	$T = 7$					

NOTES: Diagonal shows the variance of earnings in each period, lower triangle the auto-covariances, upper triangle the correlations, standard errors in parenthesis.

Bibliography

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.

Amemiya, T. (1971). The estimation of the variances in a variance-components model. *International Economic Review*, 12(1):1–13.

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.

Baker, M. and Solon, G. (1999). Earnings dynamics and inequality among canadian men, 1976-1992: Evidence from longitudinal income tax records. Technical report, National bureau of economic research.

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

Bali, T. G. and Theodossiou, P. (2008). Risk measurement performance of alternative distribution functions. *Journal of Risk and Insurance*, 75(2):411–437.

Battistin, E., Blundell, R., and Lewbel, A. (2009). Why is consumption more log normal than income? Gibrat's law revisited. *Journal of Political Economy*, 117(6):1140–1154.

Becketti, S., Gould, W., Lillard, L., and Welch, F. (1988). The panel study of income dynamics after fourteen years: An evaluation. *Journal of Labor Economics*, 6(4):472–492.

Bellemare, M. F., Barrett, C. B., and Just, D. R. (2013). The welfare impacts of commodity price volatility: evidence from rural Ethiopia. *American Journal of Agricultural Economics*, 95(4):877–899.

Benito, A. and Hernando, I. (2008). Labour demand, flexible contracts and financial factors: Firm-level evidence from Spain. *Oxford Bulletin of Economics and Statistics*, 70(3):283–301.

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.

Bloemen, H. G. (2016). Private wealth and job exit at older age: a random effects model. *Empirical Economics*, 51(2):763–807.

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

Bond, S. R. (2002). Dynamic panel data models: a guide to micro data methods and practice. *Portuguese Economic Journal*, 1(2):141–162.

Brown, S., Greene, W. H., Harris, M. N., and Taylor, K. (2015). An inverse hyperbolic sine heteroskedastic latent class panel tobit model: An application to modelling charitable donations. *Economic Modelling*, 50:228 – 236.

Browning, M. and Crossley, T. F. (2009). Shocks, stocks, and socks: Smoothing consumption over a temporary income loss. *Journal of the European Economic Association*, 7(6):1169–1192.

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

Brzozowski, M. (2007). Welfare reforms and consumption among single mother households: evidence from Canadian provinces. *Canadian Public Policy*, 33(2):227–250.

Bucks, B. and Moore, K. (2006). Measuring the role of self-employment in earnings inequality. *Federal Reserve Board*.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127.

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.

Carroll, C. D., Dynan, K. E., and Krane, S. D. (2003). Unemployment risk and precautionary wealth: Evidence from households' balance sheets. *Review of Economics and Statistics*, 85(3):586–604.

Carter, K., Mok, P., and Le, T. (2014). Income mobility in New Zealand: A descriptive analysis. *New Zealand Treasury*.

Carter, K. N., Cronin, M., Blakely, T., Hayward, M., and Richardson, K. (2009). Cohort profile: Survey of Families, Income and Employment (SoFIE) and Health Extension (SoFIE-health). *International Journal of Epidemiology*.

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

Couch, K. A. and Lillard, D. R. (1998). Sample selection rules and the intergenerational correlation of earnings. *Labour Economics*, 5(3):313–329.

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.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric theory and methods*, volume 5. Oxford University Press New York.

Davies, J. B. and Shorrocks, A. F. (2000). The distribution of wealth. *Handbook of income distribution*, 1:605–675.

Deaton, A. and Paxson, C. (1994). Intertemporal choice and inequality. *Journal of Political Economy*, 102(3):437–467.

Dolton, P. J. and Makepeace, G. H. (1986). Sample selection and male-female earnings differentials in the graduate labour market. *Oxford Economic Papers*, 38(2):317–341.

Duncan, G. J. and Hill, D. H. (1985). An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, pages 508–532.

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.

Friedline, T., Masa, R. D., and Chowa, G. A. (2015). Transforming wealth: Using the inverse hyperbolic sine (IHS) and splines to predict youth's math achievement. *Social Science Research*, 49:264–287.

Friedman, M. (1957). The permanent income hypothesis. In *A theory of the consumption function*, pages 20–37. Princeton University Press.

Gale, W. G. and Pence, K. M. (2006). Are successive generations getting wealthier, and if so, why? evidence from the 1990s. *Brookings Papers on Economic Activity*, 2006(1):155–234.

Gangl, M. (2006). Scar effects of unemployment: An assessment of institutional complementarities. *American Sociological Review*, 71(6):986–1013.

Grabka, M. M., Marcus, J., and Sierminska, E. (2015). Wealth distribution within couples. *Review of Economics of the Household*, 13(3):459–486.

Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.

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.

Haider, S. J. (2001). Earnings instability and earnings inequality of males in the United States: 1967–1991. *Journal of labor Economics*, 19(4):799–836.

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.

Haushofer, J. and Shapiro, J. (2013). Household response to income changes: Evidence from an unconditional cash transfer program in Kenya. *Massachusetts Institute of Technology*.

Hausman, J. A. and Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric Society*, pages 1377–1398.

Heathcote, J., Storesletten, K., and Violante, G. L. (2014). Consumption and labor supply with partial insurance: An analytical framework. *The American Economic Review*, 104(7):2075–2126.

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

Heckman, J. J. (1993). What has been learned about labor supply in the past twenty years? *The American Economic Review*, 83(2):116–121.

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458.

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge University Press.

Hyslop, D. and Card, D. (2016). The extensive and intensive margins of earnings dynamics. Working paper.

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.

Jappelli, T. and Pistaferri, L. (2006). Intertemporal choice and consumption mobility. *Journal of the European Economic Association*, 4(1):75–115.

John Fitzgerald, Peter Gottschalk, R. M. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *The Journal of Human Resources*, 33(2):251–299.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, pages 149–176.

Kapteyn, A. and Panis, C. (2003). The size and composition of wealth holdings in the United States, Italy, and the Netherlands. Technical report, National Bureau of Economic Research.

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.

Kennickell, A. B. and Sunden, A. E. (1997). Pensions, social security, and the distribution of wealth. *FEDS Discussion Paper*.

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.

Lillard, L. A. and Panis, C. W. (1998). Panel attrition from the panel study of income dynamics: Household income, marital status, and mortality. *Journal of Human Resources*, pages 437–457.

Lillard, L. A. and Weiss, Y. (1979). Components of variation in panel earnings data: American scientists 1960-70. *Econometrica: Journal of the Econometric Society*, pages 437–454.

Lillard, L. A. and Willis, R. J. (1978). Dynamic aspects of earning mobility. *Econometrica: Journal of the Econometric Society*, pages 985–1012.

Lydall, H. F. (2013). Theories of the distribution of earnings. *The Personal Distribution of Incomes (Routledge Revivals)*, page 15.

MacKinnon, J. G. and Magee, L. (1990). Transforming the dependent variable in regression models. *International Economic Review*, pages 315–339.

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.

Mayer, T. (1960). The distribution of ability and earnings. *The Review of Economics and Statistics*, pages 189–195.

Meghir, C. (2004). A retrospective on Friedman's theory of permanent income. *The Economic Journal*, 114(496):F293–F306.

Meghir, C. and Pistaferri, L. (2004). Income variance dynamics and heterogeneity. *Econometrica*, 72(1):1–32.

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.

Moene, K. O. and Wallerstein, M. (2003). Earnings inequality and welfare spending: A disaggregated analysis. *World Politics*, 55(04):485–516.

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.

Moss, C. B. and Shonkwiler, J. (1993). Estimating yield distributions with a stochastic trend and nonnormal errors. *American Journal of Agricultural Economics*, 75(4):1056–1062.

Muhumuza, T. (2012). Market access and child labour: Survey evidence from rural Uganda.

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.

Pence, K. M. (2006). The role of wealth transformations: An application to estimating the effect of tax incentives on saving. *Contributions in Economic Analysis & Policy*, 5(1).

Ramirez, O. A., Moss, C. B., and Boggess, W. G. (1994). Estimation and use of the inverse hyperbolic sine transformation to model non-normal correlated random variables. *Journal of Applied Statistics*, 21(4):289–304.

Robb, A. L. and Burbidge, J. B. (1989). Consumption, income, and retirement. *Canadian Journal of Economics*, pages 522–542.

Ruhm, C. J. (1991). Are workers permanently scarred by job displacements? *The American economic review*, 81(1):319–324.

Rybczynski, K. (2009). Are liquidity constraints holding women back? an analysis of gender in self-employment earnings. *The Journal of Economic Asymmetries*, 6(1):141–165.

Sakia, R. (1992). The box-cox transformation technique: a review. *The statistician*, pages 169–178.

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

Solon, G., Haider, S. J., and Wooldridge, J. M. (2015). What are we weighting for? *Journal of Human resources*, 50(2):301–316.

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

Statistics New Zealand (2008). Survey of family, income and employment: Wave four – september 2006. Accessed 2016.

Statistics New Zealand (2011). New weighting methodology in longitudinal surveys: As applied in the survey of family, income and employment. 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.

Winship, C. and Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, pages 327–350.

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

Yen, S. T. and Jones, A. M. (1997). Household consumption of cheese: an inverse hyperbolic sine double-hurdle model with dependent errors. *American Journal of Agricultural Economics*, 79(1):246–251.

Zhang, M., Fortney, J. C., Tilford, J. M., and Rost, K. M. (2000). An application of the inverse hyperbolic sine transformation—a note. *Health Services and Outcomes Research Methodology*, 1(2):165–171.