Chapter 1

Introduction

This thesis grew out of a number of research projects carried out by researchers at the Universities of Manchester and Stirling into the New Earnings Survey Panel Dataset (NESPD). The NESPD is the largest dataset of its type in Europe but its size presents some difficulties for research. It is too large for practical access by conventional statistical programs (except for generating descriptive statistics), and the data as supplied to the Department of Employment (DE) is not in a convenient form for analysis. The data is also subject to confidentiality restrictions and so cannot be analysed outwith the DE except in some aggregated form.

In 1991 the University of Stirling won a research contract to develop some basic software for the DE which would allow for the creation of data files suitable for the statistical package SPSS and the matrix programming language GAUSS. This software was devised and written by Elizabeth Roberts, a researcher at the University of Stirling. The SPSS files have been used by researchers at several institutions to create cross-tabulations, totals, and other descriptive statistics. In addition, some researchers have run cross-sectional analyses, and some attempt has been made to run non-linear models on a subset of the data.

The approach at Stirling was to use the GAUSS matrix programming language to develop a software package that could perform micro-level regressions on the dataset without the need to reread the data whenever a new set of variables is desired. The key is that the OLS regression is a linear combination of cross-product matrices; therefore, by appropriate construction of such matrices with all the variables of interest included, it is possible to run a large number of regressions with a single pass through the dataset to collect the data. In addition to the computational efficiency of this approach, it has the added benefit of avoiding the DE confidentiality restrictions.

Over time this program has been expanded to incorporate covariance estimation, differencing models, and linear instrumental variable specifications. Although these models in themselves are not new, the method of calculating them from cross-product matrices is unusual. Moreover, the covariance estimator in particular is unique in being the only software package to present time-varying coefficients for panel and cross-section models as the standard option. While models with time-varying coefficients can be specified in other packages by appropriate use of dummy variables, the default is to force a common set of slope coefficients on the model. The extra inconvenience involved in generating a time-varying-coefficients model, along with the requirement for more degrees of freedom, may explain the almost complete absence of this type of estimator in the literature on panel data. The only significant exception to this is Chamberlain's (1984) minimum-distance estimator, which has received some theoretical attention but little application.

The approach at Stirling University has been to make the time-varying coefficients model (for both panels and cross-sections) the basic specification. The methodological justification for this is provided by Hendry's general-to-specific approach; the empirical justification is that the hypothesis of constant slope coefficients is comprehensively rejected in almost all cases. Fortunately, the NES has sufficient observations to meet the demands on degrees of freedom with no difficulty.

The purpose of this thesis, then, is to present the estimation methods implemented in the software and to discuss some of the results to emerge from the application of these models to the NESPD. This is essentially a practical thesis, and the theoretical content is relatively small. However, the techniques to be discussed and the results obtained have a number of implications for other research in panel data studies and for the analysis of large datasets generally.

Also introduced in this thesis is a data structure called the "observation history", which presents semi-aggregated data in an extremely compact form. These allow for the quick calculation of an enormous number of descriptive statistics and the simple analysis of transitions between states (for example, from missing to observed or from union to non-union). Perhaps more importantly, they enable the creation of pseudo-panel datasets which may be removed from the DE. Unlike pseudo-panels created from aggregated data, these have the characteristics of true panels and so allow for straightforward analysis of a wide variety of models, including non-linear ones. However, these are a relatively new development and in this thesis are used mainly to generate descriptive statistics.

The structure of this work is as follows. Chapter two presents some of the basic ideas about estimation with panel datasets, paying particular attention to linear models and the choice of fixed- or random-effects specifications. Chapter three describes the NES and the difficulties associated with accessing the NESPD. Chapter four outlines the main solution taken at Stirling to the access problem. Chapter five describes in detail the construction of cross-product matrices for different specifications and estimation methods; chapter six describes the operation of the analytical software, the extension to instrumental variables, and the calculation of summary statistics. Chapter seven describes the construction and potential uses of the observation histories.

In the applied section, chapter eight uses the observation histories and some cohort data to provide insights into the data, the labour market, and the implications for various model specifications. Chapter nine estimates a Mincerian fixed-effects wage equation on the NESPD males and compares it with a cross-section. It also investigates the scope for more restricted specifications than the time-varying coefficients model. Chapter ten estimates a similar equation on female data, and uses this to discuss the gender gap in earnings. Both of these chapters are, I believe, the first estimates on the NESPD to allow for individual heterogeneity, and the evaluation of male earnings in particular has some pertinent comments to make about cross-sectional analyses. Finally, chapter eleven concludes the thesis.