Chapter 3

The New Earnings Survey and the NES Panel Dataset

3.1The history of the NES

The New Earnings Survey is a product of the UK Department of Employment (DE). It was originally a one-off survey of the workplace conducted in 1968, which became a yearly survey from 1970. In the early years the information was collected by the Department of Health and Social Security from the National Insurance (NI) deduction card records. In 1975 the survey was expanded, the sample selection method was changed and the responsibility for data collection was transferred to the Inland Revenue and their tax records. As there was only a 25% overlap between the participants in the 1974 and 1975 surveys the NES has been treated as a new survey from 1975.

The survey is composed of all those in employment (apart from the self-employed and some other exemptions) whose NI numbers end with a particular two-digit code. It has been held on computer since its inception. Each participant is identified by a search of the Inland Revenue's PAYE records. The identification code has been the same since 1975, and therefore individuals who remain in work will make a number of appearances in the survey. This fact was largely ignored during much of the life of the NES, its only useful contribution being a loose check on the integrity of data being entered; but in the early 1980s it was realised that the DE had the makings of a remarkable panel dataset in its computer banks. A decision was made to restructure many years of isolated survey data into a single integrated dataset which could be used for longitudinal studies. This was duly completed by the end of the decade.

The published NES remained unaffected by these changes, but researchers wishing to have more detailed information on hours and earnings could contact the DE with particular queries. This was very much on an ad hoc basis.

With the completion of the first panel dataset (covering the years 1975-1990), it was decided that public access to the data was to be encouraged in order to make the best use of this new resource. This access to this data is limited for a number of reasons to be outlined in the next chapter, but it has generally taken one of two routes. One method is to use SPSS: a copy of the dataset is held in SPSS format and the DE will run SPSS programs on the data, returning the results of regressions or cross-tabulations to the researcher. The alternative route was taken by the research team at Stirling University, in association with a number of other researchers. A second copy of the dataset is in the format of the GAUSS matrix language[1]. Programs written in GAUSS extract aggregate data which can be removed from the DE and analysed at leisure. The core of this dissertation focuses on two aspects of this, the cross-product matrix and the observation history; see chapters five and seven.

3.2Survey contents

The survey data is split into "fixed" and "variable" sections. The variable section holds information which is recorded every year: general information about hours, earnings and the employer. These questions are the same very year, and an individual will have one record for each year of the survey in which he appears[2]. These repeated questions concern:

- components of pay (weekly; basic, gross, overtime, shift, bonuses; affected by absence)
- hours of work (actual and normal)
- age and sex (which is surprisingly variable)
- occupation (in KOS, SIC and manual/non-manual breakdowns)
- time in the job (twelve months or more)
- location and business of employer
- coverage by Wages Board or collective agreement

[1] The raw NES data is held in ASCII format. The fact that copies are held in the SPSS and GAUSS formats reflects the uses made of the data, not any limitation of software or hardware.

[2] The physical split of the data into fixed and variable files does not correspond exactly to the breakdown used here.

- full-time or part-time
- sector (private, government, public corporation)

The repeated information thus contains a wealth of information on the workplace, allowing for some detailed analysis of work trends. The information available in the dataset is much richer than the published figures suggest; for example, location of employer is coded down to town/district level, although only county/regional figures are given in the published Survey.

In addition to this, occasional questions are asked. These are the "fixed" fields, as they are only requested once or are seldom repeated. There are a variety of reasons for asking these extra questions. Some are requested by researchers, some by the DE, and some by other bodies; for example, several extra questions were asked in 1979 on behalf of the Statistical Office of the European Community.

Information collected has included data on tenure, collective agreement, holiday entitlement, training, company size, and so on. This information is of limited use in panel studies (time-invariant variables are indistinguishable from individual heterogeneity in differencing or means-deviations models), but they have been used successfully in a number of cross-section studies; for example, Coleman (1994) used the occasional questions on tenure for an analysis of the relationship between tenure, sector and wages.

An obvious omission from this list of variables is any personal information about the employee other than age and sex. The NES contains no data on ethnicity, education, family background, et cetera. This is one of the major flaws in the NES, and particularly relevant to cross-sectional studies. To some extent, a panel analysis allowing for individual heterogeneity will reduce the impact of these influences, as they are generally time-invariant: individuals tend to complete their general education before commencing a career (see Elliott (1991); Dolton and Kidd(1994); Vella (1994), for example).

This is not a very satisfactory solution, as the panel analysis is being used to control for an "unmeasurable" effect which is clearly measurable in a more general sense (even if some of the measures used, such as ordered dummies for levels of education, are of debatable worth). An effective valuation of education could lead to much better cross-sectional analysis and more informative panel studies if the education variable changes over time. This is even more true of family background. Some authors have argued that, for example, earnings and participation of women are affected by the number and age of any children (Dolton and Makepeace (1987); Elias (1988); Elias and Main (1982); Joshi and Newell (1985)). Both of these may vary over time, and a panel analysis will have little more success than cross-sections in controlling for this unmeasured effect.

Some of the information on employers is also relatively limited: details of company size, establishment size, number of employees et cetera is only available for particular years. This information cannot be subsumed into individual heterogeneity unless the individual always works for companies of a similar type, although it could be argued that, once occupation, industry, region, sector and so on are taken into account, the remaining inter-company differences should be small.

3.3Survey coverage and missing data

By using a two-digit NI code for identification, the DE hoped to achieve a one percent sample of the labour force in employment. In fact, the overall participation rate is around 70% of this level (using Labour Force/General Household Survey estimates of the workforce). The survey forms are sent to the individual's last recorded employer, who is obliged to complete the form under the 1947 Statistics of Trade Act. Return rates are typically 95% or better, but not all of these returns contain usable information.

The missing data is due to a number of causes. Firstly, rather less than the full number of forms is sent out. Some employers are exempt, principally the armed forces or the self-employed. In theory, those earning insufficient amounts to pay NI or tax should have no records and so be left out of the survey; but because tax records are held over, even if no tax is paid in a particular year, rather more are included than might be expected[3]. There is also the possibility of the employer having changed address or name. Adams and Owen (1989, Table 1) report that, over the period 1975-1986, only around 90% of the desired number of forms are sent out.

When the forms are returned, about 80% are usable. Of the missing individuals, a number have moved "out of scope", into occupational pensions or unpaid work. However, by far the largest number of unusable responses is due to employers replying that the employee in question no longer worked for them (Adams and Owen (1989)). This could be due to a refusal of employers to co-operate, but it is more likely to be due to unemployment or a change of jobs. These last two arise because the form is sent to the last recorded employer, and there is a lag in updating records when a change of status occurs.

This missing data is a source of some concern. About 240,000 men and 190,000 women have appeared in the NES over the period 1975-1990[4]. Almost all of these have some missing observations, and around one-fifth have only one observation. Some 99,000 men and 56,000 women joined in 1975. Of these, only 9,200 men and 3,300 women have a complete set of observations up to 1990.

If the data are missing randomly (that is, observability does not vary systematically with the variables of interest in a study), the net effect of this will be to reduce the precision of estimates but not to invalidate them. Recent studies do not indicate that random attrition is the case. Discussion on the quality of the data in the NES is very sparse, with barely six papers in

[3] Studies by Bob Hart and Elizabeth Roberts on the micro-data do indicate that a substantial number of people earning below the NI limit are included in the NES.

[4] All the analysis in this thesis is based on the 1975-1990 dataset, as later versions are not yet available.

twenty years[5]. Bell and Ritchie (1993b, 1994) claim that non-observation is correlated with almost every variable in the dataset to some degree, although these findings are largely descriptive and so the magnitude of this correlation is difficult to assess.

The question of missing data is an enormous issue, but is largely ignored in the literature. This thesis follows the trend and avoids the issue too[6]. This is due to the difficulty of constructing realistic consistent dynamic non-linear models, as discussed in chapter two. However, chapters nine and ten on applied earnings analysis do attempt to allow for the missing data problem. As Heckman-type corrections for panels are restrictive in their assumptions and inappropriate for the NES, an ad hoc approach using proxy variables is taken.

3.4Validation and measurement errors

An issue related to that of missing data is measurement error. Much of the NES data is in the form of categorical variables, and it might be assumed that this data is reasonably accurate. However, the measurement of hours and earnings needs to be re-examined.

The DE does not carry out separate validation checks on the NESPD. Instead, the only works so far to carry out a comprehensive comparison of the NES and an independent data source (the FES and other household surveys) are the papers by Atkinson, Micklewright and Stern (1981, 1982)[7]. The most important finding is that hours for non-manual workers are

[5] Atkinson, Micklewright and Stern (1981, 1982); Micklewright and Trinder (1981); Adams and Owen (1989); Bell and Ritchie (1993b, 1994). Adams and Owen is an in-house report by the DE. Bell and Ritchie (1993b, Appendix) survey all four.

[6] An investigation into the problem of attrition in panels is being carried out by the author.

[7] The review of Atkinson *et al* gives some guidelines as to the accuracy of the NESPD but has some drawbacks. Firstly, the authors did not have access to the NES micro-data and so had to rely on published aggregates for NES numbers. Secondly, the various datasets did not record the same information and so only a limited analysis of the categorical variableswas possible. Most importantly, their analysis concentrated on the years 1971-1977 and seemed to indicate a change in the NES after the 1975 change in administration, and so extrapolating these results to the post-1975 NESPD may not be justified.

consistently lower in the NES than those reported by the household surveys. The reason is probably due to the fact that the household surveys ask employees what their "normal hours" are; in the NES the employer is asked. The NES response is significantly more concentrated around a standard working week of 38-40 hours, while the household surveys report higher normal hours on average. The suggestion is not that employers under-report their employees' hours, but that they may not have any clear idea about the normal hours for non-manual workers.

Atkinson *et al*'s comparison of earnings in the NES and the FES does not produce any clear results, and it may be expected that employers' reporting of earnings to the tax office is reasonably accurate. However, the potential for error in the hours worked will lead to error in the hourly wage rates reported by the NES, as these are calculated as reported earnings over reported hours. This increased measurement error in hourly earnings has already been noted for one major US dataset, the Panel Study of Income Dynamics (Bound, Brown, Duncan and Rodgers (1994), using a follow-up Validation Survey).

It is a well-known result (for example, Johnston (1984)) that measurement error in the dependent variable will reduce the efficiency of a model; and that measurement error in explanatory variables lead to biased and inconsistent OLS estimates of the true coefficients. The effect of measurement error may be insignificant (and indeed is often assumed to be). However, in panel models these errors become more important because of the nature of the estimators used. All the linear estimation methods in chapter two (including the differencing models) involve converting the data to deviations from some mean value. If the observed variables have relatively little variation, then these transformations may increase the measurement errors relative to transformed observables. In other words, the signal-to-noise ratio decreases, and bias and/or inefficiency may increase (see Biorn(1992); Bound *et al* (1994); Hsiao(1986)).

The estimates to be presented in chapters nine and ten do not include hours or earnings as explanatory variables, and so may be assumed to be relatively error-free. The dependent variable of log hourly earnings is liable to be subject to some measurement error, but earnings do vary over time, and so the efficiency loss due to measurement error and the covariance transformation should be small. However, it should be noted that while the covariance estimator is in general more efficient than the cross-sectional estimator, in the presence of significant measurement error this may not be the case.