Chapter 7

Observation histories

The observation histories (OHs) arose from a desire for a practical way of obtaining empirical hazard functions from the NESPD. The hazard function is an indication of the likelihood of an individual dropping out of the data set at a particular time, given that the individual is still in the dataset at that point. Empirical hazard functions, reflecting the observed rate of attrition, can be constructed from

$$\frac{\text{(number in period T) - (number in period T+1)}}{\text{(number in period T)}}$$

The obvious way to create these figures for each period is simply to read the whole dataset and sum the relevant numbers for each period. However, a far more efficient and informative aggregation of the data is possible by reversing the type of information collected: instead of saving information on observations for individuals, the characteristics of individuals with particular "observation histories" or patterns of observation in the dataset are recorded[1].

7.1Creating the observation histories

The key to the OH analysis is to note that a pattern of observation in the dataset constitutes a binary number. Individuals are identified in the NES by their national insurance numbers (NINos), represented internally as a six digit number. A missing observation has a missing NINo. Create an observation flag, representing a missing NINo by a zero and any valid number by a one. Then the record of observations translates into a sixteen-bit number (a "flag vector").

[1] The methods for storage and access of information described in this chapter are variants on a programming technique called "key transformation" or "scatter storage". A discussion can be found in most computer science texts dealing with system-level programming or data manipulation techniques; see, for example, Page and Wilson (1983) pp173-183.

For example, consider the record of an individual with the NINo "000100" over the sixteen years of the NES[2]. This is represented as Table 7.1.

This individual was observed first in 1978 and on seven subsequent occasions, the last being in 1988. The longest continual period of observation was four years, and the person had three periods of consecutive observation.

Year	NinoOther	Flag
	data	
1975	(missing)	0
1976	(missing)	0
1977	(missing)	0
1978	000100.....	1
1979	000100.....	1
1980	000100.....	1
1981	000100.....	1
1982	(missing)	0
1983	000100.....	1
1984	000100.....	1
1985	000100.....	1
1986	(missing)	0
1987	(missing)	0
1988	000100.....	1
1989	(missing)	0
1990	(missing)	0

Table 7.1 Creating the flag vector

This could have been collected to give totals for the dataset for each possible start year and end year, but there is no feasible way to store, for example, the information on run length so that is accessible.

However, consider the creation of a "hashing vector" of constants in powers of two:

$$h \equiv [1 \;\; 2 \;\; 4 \;\; 8 \;\; 16 \;\; 32 \;\; 64 \;\; 128 \;\; 256 \;\; 512 \;\; 1024 \;\; 2048 \;\; 4096 \;\; 8192 \;\; 16384 \;\; 32768]'$$

Take the inner product of this vector and f, the flag vector from table 7.1:

$$h' \,.\, f = (8+16+32+64+256+512+1024+8192) = 10104$$

This is a unique reference, only generated by a particular pattern in the flag vector. No other combination of missing/observed flags will give this number when multiplied by the hash

[2] This is an example pattern and does not reflect the actual characteristics of NESPD individual "000100".

vector. Now consider a 65535x1 column vector, initially set to zero. Adding one to the 10104-th element in this vector records that this individual has the above pattern. If this cell now contains, for example, ten, then nine other people have also been found with that same pattern of observations.

Running through the entire database, the end result will be a vector containing the numbers of people with each of the possible 65535 patterns of observation[3]. Because the observation pattern itself provides a unique index into the dataset, there is no need to separately record which patterns were used to generate which totals. Similarly, for any cell in the vector, the row number of that cell allows the identification of the observation pattern experienced by the people counted by that cell as the transformation between pattern and index is a one-to-one mapping.

For example, if the cell in row number 10104 contained "ten", then ten people have same original observation pattern. That pattern can be recovered by repeated modulo division of the row index by the elements of the hashing vector, starting with the largest number[4].

[3] The number of patterns is 2^{16}-1 because there is obviously no possibility of an individual being in the dataset but having no observations at all. For observation histories looking at only a subset of the period of the NES, this is a feasible alternative and account must be taken of this.

[4] Modulo division returns the remainder from an integer division; that is

$$x\%y \equiv \text{remainder}(x/y)$$

where "%" represents modulo division and "/" integer division.

Taking the number 10104, the result of this calculation is given in Table 7.2; it can be seen that the pattern of ones and zeros obtained, once inverted, replicates the original selection pattern in Table 7.1.

Index	Divisor	Result	Remainder ⇒ index
10104	32768	0	10104
10104	16384	0	10104
10104	8192	1	1912
19124096	0	1912	
1912	2048	0	1912
1912	1024	1	888
888	512	1	376
376	256	1	120
120	128	0	120
120	64	1	56
56	32	1	24
24	16	1	8
8	8	1	0
0	4	0	0
0	2	0	0
0	1	0	0

Table 7.2 Recovering the observation pattern

Just as the total number of people can be stored, so can other information: for example, age, wage, numbers in the private sector, and so on. However, these are now aggregated for all people with a particular pattern. These may be stored for each year or for the whole period, as long as they are stored on the correct row of the output matrix so that the pattern can be reconstituted. For example, one extraction run might store age in 1975 and then wages for each period.

This is an extremely efficient way of storing the pattern based information, but it has three drawbacks. Firstly, these observation matrices tend to be large objects, which double in size with every increase in the period under review[5]. Secondly, there is less scope to make individual inferences. Data is only disaggregated down to the level of the OH, and so only total or average figures are available for groups of people with particular observation patterns. Thirdly, as a result of this last point, the variables stored are less informative, particularly with regard to interactive and qualitative variables.

[5] The "table lookup" algorithm described here is the fastest key transformation technique, and also the simplest. More complex methods using less storage space are well-documented, but these rely on a sparse storage vector for their effectiveness and are therefore not often appropriate for the NES. The analytical routines are also greatly complicated.

However, these drawbacks are relatively minor, and the OHs do allow a very different analysis to the cross-product matrices. Although not completely disaggregated, the information available is far more detailed than is possible with any aggregate statistics. Moreover, the techniques described here are easily extended to cope with differing data requirements; for example, allowing for multiple "destinations" (part-time work, full-time work, unemployment). An incidental benefit, but one which is important for the NESPD, is that the data, being aggregated, are not subject to the confidentiality restrictions and so may be removed from the DE and analysed at the researcher's leisure.

Section 7.3 describes potential applications for the OHs. Chapter eight is based on analyses of OHs, and a large number of the statistics dotted throughout the thesis have been generated from them.

7.2Analysing the observation histories

Taking the compacted information and listing the variables associated with each pattern is of limited interest; more useful would be to specify and analyse subsets of the patterns. For example, it could be desirable to select only those who appear in the dataset in 1975 and have at least two consecutive observations.

As for the cross-product matrices, this analysis requires software designed for the purpose. Unlike XPReg, this software is, to some extent, specific to the particular type of OH created, although the same programs may be used for several OHs. However, the basic principle is essentially the same. Similar compression techniques to those outlined above may be used to store useful information about the patterns: number of observations, first observation, longest period of continuous observation, and so on. By reversing the compression operation, as detailed above, vectors of characteristics of the observation patterns are created which may be used in logical operations to select patterns with particular characteristics. The key

transformation technique described above thus provides an effective way of both storing data and analysing it[6].

7.3Using the observation histories

7.3.1Descriptive analyses

The most obvious use of the OHs is in descriptive statistics of the dataset, such as those presented in chapter eight. As information is identified for every possible pattern, it is relatively easy to form simple statistics such as frequencies of observations and absences, lengths of continuous observation, and so on. When more data then just numbers of individuals for each pattern is recorded, the scope for this analysis increases. Suppose the numbers of individuals changing their work location was noted for each observation pattern. Then it is a simple matter to calculate the frequency of moves for people with a particular OH. Aggregated dataset statistics are straightforward to calculate for years or combinations of years.

A useful feature is the ability to follow cohorts within the dataset. For example, a cohort could be constructed of all those observed the whole time from 1975 to 1990 and a second of all those observed in 1975 and 1990 but with at least one missing observation in the meantime. This could then provide a basis for the comparative analysis of wage growth and the effect of absence.

7.3.2Analysis of transitions

The above example used the OHs to map an individual's pattern of observation. However, the principle is clearly extensible to the analysis of other changes of state. Two examples illustrate

[6] The mechanics of extracting information in simple cases is described elsewhere.

this use.

The first extension is to allow for multiple destinations rather than a simple observed/not observed split. The NESPD has been aligned with the Department of Social Security's Juvos dataset, which records periods of unemployment. Observations or missing values may then be classified as part-time work, full-time work, known unemployment or just missing. This allows transition probabilities (and associated changes in average wages) to be calculated for all possible combinations of the four states. The potential for state dependence in the labour market has been pointed out by many authors and is central to a number of theories of segmented or two-tier labour markets.

A second extension is to split observations into categories; for example, covered or not covered by collective agreement. This would allow transitions between union and non-union status to be modelled much more accurately than is possible with summary statistics. Transitions between observed states have received much less attention than moves between full-time and part-time work or employment and unemployment, for example. However, some authors (for example, Card (1994) on union membership) have attempted to analyse and allow for changes in state rather than just using current status. The OH approach is ideally suited to this.

Both of the examples given move away from the binary observed/not observed decision of section 7.1. However, the only significant change is that the hashing vector needs a higher base number: base four in the first example, base three in the second. The OH principle remains the same however many possibilities are analysed.

7.3.3Estimation: the pseudo-panel dataset

As the OHs amount to a grouped dataset they can be used to specify and tests models in the

same way as any other dataset. The necessary corrections for efficient linear estimation on grouped data are straightforward, although for non-linear models the aggregation can cause problems if the number of individuals in the groups is small - which it often is.

However, from an econometric viewpoint, an extremely appealing feature of the OHs is that the same individuals appear (or are absent) in each year for a given pattern. This means that there is a constant panel effect for each pattern, and so panel estimation techniques may be employed as if the OHs constituted a "true" panel dataset.

Consider the usual relationship for individual i in period t allowing for individual heterogeneity:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \qquad \textbf{(7.1)}$$

Collecting information on all individuals, using some known characteristic grouping variable, means that the model for a pattern p in period t is

$$\bar{y}_{pt} = \bar{x}_{pt}\beta + \bar{\alpha}_{pt} + \bar{u}_{pt} \qquad \textbf{(7.2)}$$

where

$$\bar{y}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i\in p} y_{it} \quad \bar{x}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i\in p} x_{it} \quad \bar{\alpha}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i\in p} \alpha_{it} \quad \bar{u}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i\in p} u_{it} \qquad \textbf{(7.3)}$$

There has been some interest in recent years in using repeated cross-sections to construct "pseudo-panels" embodying the relationship in (7.2); Verbeek (1992) surveys these. The attraction of these models is that they enable some panel analysis techniques to be applied to non-panel datasets. However, unlike true panel datasets, the pattern-specific effect α_{pt} is not constant over time. A pattern or group will be formed of different individuals in different years and the number of individuals in a group may vary over time. If

$$plim\ \bar{x}_{pt}\bar{\alpha}_{pt} = 0 \qquad \textbf{(7.4)}$$

then OLS estimation of (7.2) is consistent and unbiased; but if (7.2) does not hold then OLS is inconsistent.

The option of true panel models, to avoid the correlation by treating the pattern effect as a fixed parameter, runs up against the identifiability problem as there are PT α-terms to be identified in PT equations. One solution is to consider (7.2) as an errors-in-variables problem, for which appropriate instrumental variable techniques have been developed (see Verbeek (1992); Bowden and Turkington (1984))[7]. However, this reduces the appeal of these pseudo-panels over cross-section specifications.

In the case of the OHs, this problem does not arise. Each pattern represents a particular OH; only individuals with the same OH over the whole period will be marked as belonging to that pattern; and the same individuals appear every year. Therefore, for any one OH,

$$\bar{\alpha}_{pt} = \bar{\alpha}_{p} \qquad \textbf{(7.5)}$$

and so (7.2) becomes

$$\bar{y}_{pt} = \bar{x}_{pt}\beta + \bar{\alpha}_{p} + \bar{u}_{pt} \qquad \textbf{(7.6)}$$

Thus the pattern effect can be validly treated as a fixed parameter which is identifiable; either the covariance estimator or using dummy variables will allow consistent estimation of β. Alternatively, it can be treated as a random effect, with the appropriate estimation method being used.

7.3.4Hazard and survivor functions

As mentioned, the desire to construct hazard and survival functions was a motivating factor for this work. Consider using the OHs to group individuals. The most obvious choice of "group" is to consider each yearly cohort, but these functions could also be calculated for only those sub-groups have no missing observations; or for those with only start and end

[7] An alternative is to assume that N_{pt} is large and the α term is relatively small and so ignorable.

observations; or for the whole dataset rebased to a common start period; and so on.

Let N_{gt} represent the number of individuals still deemed to be in group g in year t, a number which is easily retrieved from the OHs. Then empirical hazard and survivor functions for the group g are determined from

$$hazard\ rate \equiv \frac{N_{gt+1} - N_{gt}}{N_{gt}} \qquad survival\ rate \equiv \frac{N_{gt}}{N_g} 1 \tag{7.7}$$

Alternatively, one could create functions which reflect the probability of going missing in a particular year rather than leaving the dataset for good, or the likelihood of returning after a missing observation, and so on. Clearly, a wide variety of probabilistic measures are available from manipulation of the OHs.

7.4Summary

The above sections illustrate ways in which the OHs can be constructed and used. Together with the cross-product matrices, the OHs provide an efficient way of extracting large amount of information from the data in an easily digestible form. Clearly, all the information required could also be calculated by reading the dataset and storing the desired totals. However, by using the OHs, one pass through the dataset provides the potential to construct many more statistics. For example, it may be that, following an analysis of the effect of absence on the wages of the 1975 cohort, it becomes desirable to separate the effects of the timing and length of absence. This information is already available in the OHs without the need to return to the NESPD. The principles of the OH can easily be extended to encompass efficient storage of other information on discrete states.

One final advantage of using the OHs is that, being semi-aggregated data, they do not suffer from the confidentiality restrictions on the NESPD or the practical difficulties caused by its size. They are therefore appealing as a compact panel data set which can be removed from the

DE premises and analysed using standard econometric packages under the direct control of the researcher.

This chapter ends the description of the data collection methods used to analyse the NESPD at the University of Stirling. There are a number of other ways of analysing the NESPD, of which the most obvious is in the construction of simple transition matrices (which are to some extent already embodied in the cross-product matrices) and labour market cohorts. All of these are used to some extent in the following chapters, but the bulk of the analysis is performed using the cross-product and the OHs.