BHPS Work-life History Files, Version 2

Brendan Halpin

January 27, 2000

1 Introduction

This is the second release of a set of files, derived from BHPS work-life history data, designed to facilitate use of this data. The first release (in 1997) is described in Halpin (1998), alternatively Halpin (1997). This document focusses mainly on the differences between the two releases.

The first set of files has been replaced for three reasons:

1. It covers only 5 waves (8 are currently available)
2. It contains some errors
3. Its design masks certain measurement problems inherent in longitudinal data, and makes it harder for the analyst to deal with them.

In preparing the new release I have fundamentally redesigned the algorithms used to construct the derived files, implementing the same specification but doing so in a way that focuses on episodes rather than 'calendar vectors'. This is internally simpler (and makes it easier to correct errors present in the first version) and also, in particular for information derived from `wINDRESP` and `wJOBHIST`, provides far more information to the analyst about measurement problems.

2 Design differences

The main difference in the design is the focus on episodes rather than calendar vectors. A calendar vector is a set of variables describing the state in each consecutive month, whereas an episode is a period of time, with a defined start and end, during which the state is constant. The previous work-life history project was oriented to state calendars as the central way of representing the 'reconciled' history (that is, the single representation derived from multiple reports). In some ways this is easier than working with episodes (in particular, it is easier to decide which value to attribute to a given month, given multiple potentially conflicting reports, than to choose between spell-oriented reports). However, it is inefficient both practically (very slow and very large file sizes) and in programming terms: a lot of the time we *do* need to work in terms of spells, and it is tedious and error-prone to recover spell information from calendar vectors. For instance, if we want to determine 'how long has this unemployment spell lasted?' we need to iterate back through the vector until we find an observation with a different value. If we want to know 'how long has this job lasted?' we need extra information in order to distinguish between successive jobs with similar characteristics (that is, we need to carry a spell-number vector as well, in order to mark transitions where the substantive information does not change).

The focus on spells makes for easier programming, and it also makes it easier to carry through to the end-use files information necessary for dealing with seam effects, an inevitable measurement problem associated with repeated measurement of spell data. Briefly, seam effects are observed as disproportionate numbers of apparent changes occurring immediately after an observation point (*i.e.*, the interview). They arise either because of measurement error in the state information, such that a spell which overlaps two successive interviews is reported with different substantive information at each interview, making it appear to change at an unknown date between the two (usually defaulting to immediately after the first, depending on how the problem is resolved) or because of measurement error in the start date of a spell that in reality starts after the earlier interview but which is mistakenly reported or recorded at the later interview as start-

ing before the earlier.

Seam effects are a problem of the data, but different representations of the data make it harder or easier to cope with them. The core of the new release of the work-life history files, *newpan*, is designed to carry through to the user as much information as possible about seam effects arising on observations during the panel period (as distinct from the retrospective histories). *newpan* has a very specific structure for this purpose, with a record for every relevant observation (that is, every distinct report of either current status at the time of interview or of status changes between interviews). Groups of these records (which I will refer to throughout as 'splits') can be defined as episodes, but to make such a definition involves making decisions about seam effects (*i.e.*, which problematic transitions to ignore and which to accept as real changes). The design of *newpan* retains the split-level information (which is independent of decisions about seam effects) and superimposes on it a structure of episode-level information (which is dependent on the specific rules used to define which transitions to ignore). The particular rules used in the released version of *newpan* are permissive, that is, they tend to define successive splits as part of the same episode unless the state changes excessively (*e.g.*, employed to unemployed). However, the point of this structure is to allow the end-user to impose a different set of rules.

2.1 Differences affecting the rest of the file set

In brief, the previous release was constructed according to the following sequence: first, three calendar files were created, representing

- data observed in *wINDRESP* and *wJOBHIST* (pnlempc)
- data observed in *BLIFEMST* (lempc), and
- data observed in *CLIFEJOB* (ljobc).

These were then combined to extend the retrospective histories using panel data, as follows:

- pnlempc + lempc ⇒ xlempc
- pnlempc + ljobc ⇒ xljobc.

The third step involved combining xlempc and xljobc to create a combined extended retrospective history file, ljempc, which contained both occupational and employment information.

A final step was to take each of the foregoing files and to derive a spell-oriented file from it, pnlempe from pnlempc, and so on.

In this version the same overall framework is used, but the routes through it are radically different. That is, we generate *newpan* as a replacement for pnlempc, merge it with spell-structured files derived from *BLIFEMST* and *CLIFEJOB*[1] to create new versions of xlempe and xljobe respectively.

Once these files are created (*newpan*, xljobe and xlempe) it is trivially easy to create calendar files from them and thus the final file, ljempc, can be generated as before.

3 The structure of *newpan*

The keystone of this enterprise is *newpan*. This contains the richest data, and the least extent of recall error. The files derived from the long-term retrospective histories (*i.e.*, from *BLIFEMST* and *CLIFEJOB*), including those with information from *newpan* have a lesser level of detail (for instance, in xljobe job changes within employers – which are recorded in *newpan* – are ignored). Moreover, the 'split' structure of *newpan* is not retained in the other files, which means that the definition of an ignorable transition is frozen.

The structure of *newpan* is therefore more complex than that of the other files:

- it contains one record for each independent observation, and
- it has a superimposed episode structure, that is, a set of variables whose values depend on changeable rules about grouping consecutive observations into substantive employment status episodes.

This means that we have two types of variable: 'split' variables, whose values depend directly on observations in the main BHPS data and the rules used to reconcile these, and the superimposed 'episode' variables.

[1] At present, these spell-structured files are the old lempe and ljobe. It would be appropriate but not urgent to create new episode files derived directly from the original files without going through a calendar phase.

Split variables are in the main direct reproductions of the values of their corresponding variables in the main database. For instance, empstat is (largely) derived from wJBSTAT and wJHSTAT in *wINDRESP* and *wJOBHIST* respectively. One important exception are the split-start and split-end dates (spdate and spend). These indicate the start and end dates of the period of time on which this observation contributes independent information. This is distinct from the current spell, because in many cases the previous split will also report on the current spell. This will occur, for instance, when the respondent is in the same spell at two successive interviews: in this case the split corresponding to the second interview will begin contributing new information directly after (in practice, the month after) the first interview. Thus that split will have a start date the month after the first interview and an end date of the month of the second interview.

This is in accordance with one of the basic rules of the construction of the combined data set: to give priority to the earlier report. Even though the second report overlaps the first, in covering the period between the start date and the earlier interview, we use the first report exclusively for that period, and only begin to use the second after the first runs out. This makes no difference when the reports agree, but in the event of a disagreement it gives us a rule for deciding what to believe when.

Table 1 lists some of the variables in *newpan*, differentiating between split variables and episode variables, and within split variables, between 'substantive' variables, which are directly derived from variables in the main BHPS database, and 'derived' split variables, which are created in the process of generating the file.

Derived split variables include the split start and end dates, described above. The split start date is either the reported start date of the spell reported in that observation (if there is no previous observation) or the end date of the previous observation (precisely, the end date plus one month) if there is one.

The source variable is very important: this links every record in *newpan* to the specific file and record within the file in the main BHPS database. source is a 2-digit number with a decimal portion. The first digit indicates the wave, the second the type of record (0 for *wINDRESP*, 1 for *wJOBHIST*) and the decimal portion indicates, for *wJOBHIST* records only, the record number within an individual's observations in *wJOBHIST*, with the numbers counting backwards. Thus 30 indicates the split derived from *CINDRESP* (*i.e.*, wave 3), while 41.02 refers to the second record in *DJOBHIST*, that is, the spell before the spell immediately before the spell reported in *DINDRESP*, the state at the wave 4 interview.

Split level left censoring is indicated by lcens. A split is left censored when we do not know (from that data source) when the spell it is describing began. Typically this arises when the start date is missing. When this is a *wINDRESP* observation, the variable spdate is set to the interview date unless the respondent answers that this spell started on or before the reference date (1 September the year before the fieldwork began), in which case it is set to the month of the reference date. Right censored splits are *wINDRESP* observations where there is no next observation or where there is a time gap before the next observation (the next split's spdate is more that one greater than this split's spend).

indfirst and indlast are boolean (0/1) variables indicating an individual's first and last split, respectively.

3.1 Episode-level variables

I'll say more about the rules by which splits are grouped into episodes, below, but here I'll describe some of the episode-level variables. episno is the basic indicator of episodes: for a consecutive set of splits deemed to constitute a single episode, episno stays the same, and increments between such episodes (resetting to one for each new individual). epfirst and eplast are booleans indicating first and last splits in an episode. The variable djse is a member of another class, which might be described as substantive episode level variables: certain facts about spells are only collected at their end, in the inter-wave retrospective section (*i.e.*, in *wJOBHIST*). djse is based on the *wJOBHIST* variable, wJHSTAT, which differentiates between jobs with a different employer and those with the same employer as the next job. Where an episode spans several interviews, and is observed to end, it will contribute several splits, all but the last of which will be from *wINDRESP*s, and the last from

Table 1: Some variables in *newpan*

Split variables	
	(i) some substantive variables
EMPSTAT	Employment status
SIC	Industry (SIC)
SOC	Occupation (SOC)
SIZE	No. employed at workplace
PAYG	Monthly Gross Pay
PAYN	Monthly Net Pay
	(ii) derived split variables
SPDATE	Split start date
SPEND	Split end date
SOURCE	Info source this split
INDFIRST	Resp's first record
INDLAST	Resp's last record
LCENS	Split left censored
SPCENS	Split right censored
Episode variables	
EPISNO	Episode seq number
EPFIRST	1st split in episode
EPLAST	last split in episode
EPDATE	Episode start date
EPEND	Episode end date
EPCENS	Epis right censored
DJSE	Whether this is a job with the same employer as the next job

a *wJOBHIST*. Only the last split will have direct access to the djse information, but if we decide that one or more prior splits constitute part of the same episode, we can copy back the relevant information. Thus, the value of variables such as djse depends on how episodes are defined (at least, for all but the *wJOBHIST* split from which it originates).

4 Other files

newpan is the core of the work-life history data set, but there are other important files. These include lemp and ljob, derived directly from *BLIFEMST* and *CLIFEJOB* as before, and their extended equivalents, xlemp and xljob, where the long-term retrospective information is supplemented from the panel.

References

HALPIN, B. (1997). 'Unified BHPS work-life histories: combining multiple sources into a user-friendly format', *Technical Paper 13*, ESRC Research Centre on Micro-social Change, University of Essex.

HALPIN, B. (1998). 'Unified BHPS Work-Life Histories: Combining Multiple Sources into a User-Friendly Format', *Bulletin de Méthodologie Sociologique*, No. 60.

File name	Sources	Available epis	Available cal	note
newpan	*wINDRESP*, *wJOBHIST*	y	n	only spell-oriented
lemp	*BLIFEMST*	y	y	episode derived from calendar
ljob	*CLIFEJOB*	y	y	episode derived from calendar
xlemp	lemp, newpan	y	y	calendar derived from episode
xljob	ljob, newpan	y	y	calendar derived from episode
ljemp	xlemp, xljob	y	y	calendar format, uses calendar versions of xlemp and xljob

Table 2: Files in the WLHP data set