```
STOP



*******************************************************.
***  Longitudinal Data Analysis for Social Science Researchers
**
**
** ESRC Researcher Development Initiative training programme:
**
**
**   Training materials lab 2:
**     PANEL SURVEY DATA ANALYSIS AND DATA MANAGEMENT  .
**
**
**      www.longitudinal.stir.ac.uk
**      Paul Lambert / Vernon Gayle, 26 August 2007
*******************************************************.


**** STATA VERSION ************************************



*********************************************************.
*********************************************************.
** The file below covers examples of data managment and data analysis
**       primarily using data from the British Household Panel Survey 1991-2004:
**
***    Exercise 1:  Reviewing the bhps data resources
***    Exercise 2:  Deriving 'long format' panel data from bhps component files
***    Exercise 3:  Deriving 'wide format' panel data from bhps component files
***    Exercise 4:  Translating between wide and long format panel data
***    Exercise 5:  Understanding the bhps: sample origins, and using weights
***    Exercise 6:  Dealing with fixed-in-time bhps data
***    Exercise 7:  Examples using the bhps youth files
***    Exercise 8:  Further examples in matching panel data files
***    Exercise 9:  Long format data: descriptive analyses
***    Exercise 10: Panel data analysis: looking at transitions / sequences
***    Exercise 11: Long format data: panel data models
***    Exercise 12: Panel data models: looking at estimation issues
**
*********************************************************.



*******************************************************.
** GENERAL INSTRUCTIONS ON THESE FILES
**
** Work through this file in the interactive do-file editor, replicating
**  the STATA do-file commands. Further help on working with STATA is
**  available from the LDA web site.
**
**
*** This lab file assumes you have a number of files downloaded to your
**    machine. You will need the following:
*
*
*
** 1) EITHER supplied by the workshop session instructors (taught workshops)
*      OR as a product of carrying out the lab2 stata commands:
*
* bh1to15_long.dta
* bh1to15_wide.dta
*
* (derived BHPS data files obtained from merging source BHPS files in lab 2 below).
*
*
** 2) Downloadable from the UK Data Archive:
*
* -All BHPS Waves 1-15 component files in Stata format (UK Data Archive Study number
*   5151 (June 2007 release) (extracted from the zip file 5151STATA8.ZIP)
*      + warning - these are a large volume of files, ~152 different files, ~ 600MB)
```



```
*
* - The six Stata format 'episode' files from the BHPS Derived life history files
*     (UKDA study number 3954, 5th Edition) (covering waves 1-14 only) (you want to access
*      the 3 files data files on the top directory of the '3954.zip' archive,
*      called newpan.dta, xlempe.dta, and xljobe.dta, plus the 3 files in the
*      'episode' subfolder of the '3954.zip' archive, called l*.dta)
*
*- All BHPS derived net income files (UKDA study number 3909) (cover waves 1-12 only)
*
*
*
*******************************************************.

** .


*******************************************************.
** NOTIFICATION OF FILE LOCATIONS / DIRECTORIES AND STATA SETUP
**
**
**
** i) File location declarations:
*** For the commands below to work, you should begin by running the following
**   macros, which tell Stata where to look for the relevant data files (mentioned
**   above) on your machine : .


global path1 "d:\lda\work\"
* (the location of your working directory -  where you will save
*    newly created data files and output) .


global path2 "d:\data\lda\"
* (the location of a folder where you have saved the
*   WEBCT sourced data files mentioned above ) .


global path3 "d:\data\bhps\w1to15\"
* (the location of a folder where you have saved the BHPS
*   panel data files mentioned above) .

global path3b "d:\data\bhps\derived\"
* (the location of a folder where the BHPS derived
*   data files bhps1to15_long.dta and bhps1to15_wide.dta are or will be stored) .
* (That is, lab 2 creates these files, but also can use them if they have been created elsewhere).

global path3c "d:\data\bhps\inc\"
* (the location of a folder where you have saved the BHPS derived
*   income data files [study number 3909] mentioned above) .

global path3d "d:\data\bhps\lifehist\"
* (the location of a folder where you have saved the BHPS derived
*   life history data files [study number 3954] mentioned above) .



global path9 "d:\temp\"
* (a location of a temporary folder where you can save intermediate files) .


**
************************************************


************************************************
**
** Stata session management:
*** The commands below are used to set some general preferences within Stata,
**    it is usually good advice to run these but it is not essential
clear
set more off
set memory 128M
capture log close
capture log using $path1\log_lab2.txt, replace text
**
**
```

```
*******************************************************.


*********************************************************************
*
*   Reminder: other support materials in working with Stata in the context
*    of longitudinal survey datasets are available from the LDA website,
*    http://www.longitudinal.stir.ac.uk/Stata_support.html
*
************************************************************
************************************************************


** ..finally, here is the lab exercise...


************************************************************









************************************************************.
***  EXERCISE 1) REVIEWING THE BHPS DATA RESOURCES  .
************************************************************.



*  BEFORE YOU DO ANYTHING ELSE...
*   - Check the online BHPS documentation at http://iserwww.essex.ac.uk/ulsc/bhps/doc/
*     (see especially volume B)



** BHPS : Getting started in STATA :

****** i) Core files: Opening an indresp file :.

* Most BHPS individual data is on the 'indresp' files for adult interviews in each wave.
use $path3\lindresp.dta, clear
summarize
** This is a huge dataset.
summarize pid lhid lpno.
** these are key index variables.
numlabel _all, add
tab ldoim
tab ldoid
* Most of the time these variables are ignored - but for some analyses the wide window
*  of observation could be important.

tab lsex
tab lregion
* Note the unusual distribution: the BHPS includes regional 'booster' samples.
* If you want to do a nationally representative analysis within a particular
*  individual year, use an appropriate cross-sectional weight :.
tab lregion [aweight=lxrwtuk1]

** Rather than use the whole file, it is better to make subset extractions: .
use pid lhid lsex lregion lxrwtuk1 using $path3\lindresp.dta, clear
summarize
codebook
* (Choose the variable names you want from http://iserwww.essex.ac.uk/ulsc/bhps/doc/ ).
* (Footnote: PanelWhiz.eu offers useful functionality here, in searching for variables)


***** ii)  Some different individual files from 2004 :.
use pid nage nivfio using $path3\nindresp.dta, clear
summarize pid nage
```



```
numlabel _all, add
tab nivfio
* 15791 adults who were given some sort of interview.
use pid nage nivfio using $path3\nindall.dta, clear
summarize pid nage
numlabel _all, add
tab nivfio
* 22127 people in BHPS households (most non-adult interviews are children) .
use pid nypdoby4 using $path3\nyouth.dta, clear
summarize pid nypdoby4
* 1397 youths aged 11-15 in BHPS households were interviewed in the 2004 youth sample.



******* iii) Core files: Household files in 2004 .

use $path3\nhhresp.dta, clear
summarize nhid
* 8897 different households contacted in wave 14.
numlabel _all, add
tab nhstype
tab nhsprbi
* These household files contain a variety of household level data.


**** iv) Core files: Event oriented files :.
* (we will return to these event oriented files in lab 3)
use $path3\njobhist.dta, clear

summarize pid
gen one=1
xttab one, i(pid)

sort pid
gen first=0
replace first=1 if (pid~=pid[_n-1])
tab first
** Each record here is an employment spell over the last 12 months .
** There are 4418 records of spells, covering 3289 different individuals.
** In fact, these are the additional spells _only_ for people who have changed jobs
*    since their last scheduled BHPS interview - ie, people in the same situation as
*    when interviewed approx 12 months ago don't have a record.
* ('xttab' and the 'gen' command above both get to the same end result)



****  v) Core files:  cross-wave files :.
* (these have selected useful data on the same individuals from multiple time points)
use $path3\xwaveid.dta, clear
sort pid
summarize pid
list pid sex aivfio bivfio fivfio livfio nivfio oivfio in 200/220
* these show response status
numlabel _all, add
tab sex
tab aivfio
tab oivfio
tab oivfio aivfio
** There's one record per person who has ever been interviewed in the BHPS,
**  but lots of movement in and out.
** The cross-wave files are usually used to get summary information on people over time.


*** vi) Subsidiary files : Combined life history files (Study number 3954).
* (we will return to these event oriented files in lab 3)
use $path3d\ljempe.dta, clear
sort pid date
list pid date spellno stemp in 350/374
summarize
numlabel _all, add
tab spellno
tab stemp

* This file has a single record for every recorded distinct employment history event in each
*  BHPS respondent's life history (assuming the data was collected for them).
*   The length of the record is derived from information on the starting and ending time.
```

```
* There are 84011 events in all, covering 21815 individuals (spellno=1).
* Spellno is the number of event for each individual .
* In fact, this file extends only as far as the wave 10 data collection, though some other
*  derived life history files do extend further in time.

sort pid spellno
collapse (max) spellno , by (pid)
tab spellno

* Most people have 10 spells or less, but a few have more, with the highest being
* a respondent with 41 distincitive employement status events recorded.


******* vii) Subsidiary files: Derived income files (Study number 3909).

use $path3c\l_nethh.dta, clear
summarize
** this is a household level file with information on the household income circumstances
**  additional to that released in the main survey (after derivations accounting, for instance,
*    for local tax rates).
graph bar (mean) lloctax (mean) lhcost, over(lmemorig)



************************************************************.

************************************************************.
************************************************************.



************************************************************.
***  EXERCISE 2) DERIVING 'LONG FORMAT' PANEL DATA FROM BHPS COMPONENT FILES   .
************************************************************.




** THIS EXERCISE GENERATES THE COMBINED FILES BHPS1TO15_WIDE.DTA AND
**     BHPS1TO15_LONG.DTA WHICH ARE USED IN OTHER PARTS OF THESE LABS



************************************************************.
***   This Excercise concerns extracting related data from multiple
**    BHPS Waves (1-15 individual level + 1-15 household level data)


* (The first bit of syntax just produces 15 smaller wave-specific extracts,
*   which are then used, in the next section, to derive merged longitudinal files).



************************************************************
***** Segment 2.1) Make 15 successive temporary extract files:
********************

** BHPS wave 1 : .
use ahid apsu astrata using $path3\ahhsamp.dta , clear
sort ahid
save $path9\mtch1.dta, replace
use pid ahid apno asex aage aregion aqfedhi avote ///
   axrwght axewght ajbcssm ajbrgsc afimn ahlghq1    ///
   using $path3\aindresp.dta , clear
sort ahid
merge ahid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=1
gen year=1991
sort pid
```



```
summarize
save $path9\aindresp_ext1.dta , replace

** BHPS wave 2 : .
use bhid bpsu bstrata using $path3\bhhsamp.dta , clear
sort bhid
save $path9\mtch1.dta, replace
use pid bhid bpno bsex bage bregion bqfedhi bvote4 ///
   bxrwght bxewght blrwght blewght bjbcssm bjbrgsc bfimn bhlghq1   ///
   using $path3\bindresp.dta , clear
rename bvote4 bvote
sort bhid
merge bhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=2
gen year=1992
sort pid
summarize
save $path9\bindresp_ext1.dta , replace

** BHPS wave 3 : .
use chid cpsu cstrata using $path3\chhsamp.dta , clear
sort chid
save $path9\mtch1.dta, replace
use pid chid cpno csex cage cregion cqfedhi cvote ///
   cxrwght cxewght clrwght clewght cjbcssm cjbrgsc cfimn chlghq1   ///
   using $path3\cindresp.dta , clear
sort chid
merge chid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=3
gen year=1993
sort pid
summarize
save $path9\cindresp_ext1.dta , replace

** BHPS wave 4 : .
use dhid dpsu dstrata using $path3\dhhsamp.dta , clear
sort dhid
save $path9\mtch1.dta, replace
use pid dhid dpno dsex dage dregion dqfedhi dvote ///
   dxrwght dxewght dlrwght dlewght djbcssm djbrgsc dfimn dhlghq1   ///
   using $path3\dindresp.dta , clear
sort dhid
merge dhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=4
gen year=1994
sort pid
summarize
save $path9\dindresp_ext1.dta , replace

** BHPS wave 5 : .
use ehid epsu estrata using $path3\ehhsamp.dta , clear
sort ehid
save $path9\mtch1.dta, replace
use pid ehid epno esex eage eregion eqfedhi evote ///
   exrwght exewght elrwght elewght ejbcssm ejbrgsc efimn ehlghq1   ///
   using $path3\eindresp.dta , clear
sort ehid
merge ehid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=5
gen year=1995
sort pid
summarize
save $path9\eindresp_ext1.dta , replace

** BHPS wave 6 : .
```

```
use fhid fpsu fstrata using $path3\fhhsamp.dta , clear
sort fhid
save $path9\mtch1.dta, replace
use pid fhid fpno fsex fage fregion fqfedhi fvote ///
   fxrwght fxewght flrwght flewght fjbcssm fjbrgsc ffimn fhlghq1   ///
   using $path3\findresp.dta , clear
sort fhid
merge fhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=6
gen year=1996
sort pid
summarize
save $path9\findresp_ext1.dta , replace

** BHPS wave 7 : .
use ghid gpsu gstrata using $path3\ghhsamp.dta , clear
sort ghid
save $path9\mtch1.dta, replace
use pid ghid gpno gsex gage gregion gqfedhi gvote ///
   gxrwght gxewght glrwght glewght gjbcssm gjbrgsc gfimn ghlghq1   ///
   using $path3\gindresp.dta , clear
sort ghid
merge ghid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=7
gen year=1997
sort pid
summarize
save $path9\gindresp_ext1.dta , replace

** BHPS wave 8 : .
use hhid hpsu hstrata using $path3\hhhsamp.dta , clear
sort hhid
save $path9\mtch1.dta, replace
use pid hhid hpno hsex hage hregion hqfedhi hvote ///
   hxrwght hxewght hlrwght hlewght hjbcssm hjbrgsc hfimn hhlghq1   ///
   using $path3\hindresp.dta , clear
sort hhid
merge hhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=8
gen year=1998
sort pid
summarize
save $path9\hindresp_ext1.dta , replace

** BHPS wave 9 : .
use ihid ipsu istrata using $path3\ihhsamp.dta , clear
sort ihid
save $path9\mtch1.dta, replace
use pid ihid ipno isex iage iregion iqfedhi ivote ///
   ixrwght ixewght ilrwght ilewght ijbcssm ijbrgsc ifimn ihlghq1   ///
   using $path3\iindresp.dta , clear
sort ihid
merge ihid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=9
gen year=1999
sort pid
summarize
save $path9\iindresp_ext1.dta , replace
*
** BHPS wave 10 : .
use jhid jpsu jstrata using $path3\jhhsamp.dta , clear
sort jhid
save $path9\mtch1.dta, replace
use pid jhid jpno jsex jage jregion jqfedhi jvote ///
   jxrwght jxewght jlrwght jlewght jjbcssm jjbrgsc jfimn jhlghq1   ///
```



```
   using $path3\jindresp.dta , clear
sort jhid
merge jhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=10
gen year=2000
sort pid
summarize
save $path9\jindresp_ext1.dta , replace

** BHPS wave 11 : .
use khid kpsu kstrata using $path3\khhsamp.dta , clear
sort khid
save $path9\mtch1.dta, replace
use pid khid kpno ksex kage kregion kqfedhi kvote ///
   kxrwght kxewght klrwght klewght kjbcssm kjbrgsc kfimn khlghq1   ///
   using $path3\kindresp.dta , clear
sort khid
merge khid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=11
gen year=2001
sort pid
summarize
save $path9\kindresp_ext1.dta , replace

** BHPS wave 12 : .
use lhid lpsu lstrata using $path3\lhhsamp.dta , clear
sort lhid
save $path9\mtch1.dta, replace
use pid lhid lpno lsex lage lregion lqfedhi lvote ///
   lxrwght lxewght llrwght llewght ljbcssm ljbrgsc lfimn lhlghq1   ///
   using $path3\lindresp.dta , clear
sort lhid
merge lhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=12
gen year=2002
sort pid
summarize
save $path9\lindresp_ext1.dta , replace

** BHPS wave 13 : .
use mhid mpsu mstrata using $path3\mhhsamp.dta , clear
sort mhid
save $path9\mtch1.dta, replace
use pid mhid mpno msex mage mregion mqfedhi mvote ///
   mxrwght mxewght mlrwght mlewght mjbcssm mjbrgsc mfimn mhlghq1   ///
   using $path3\mindresp.dta , clear
sort mhid
merge mhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=13
gen year=2003
sort pid
summarize
save $path9\mindresp_ext1.dta , replace

** BHPS wave 14 : .
use nhid npsu nstrata using $path3\nhhsamp.dta , clear
sort nhid
save $path9\mtch1.dta, replace
use pid nhid npno nsex nage nregion nqfedhi nvote ///
   nxrwght nxewght nlrwght nlewght njbcssm njbrgsc nfimn nhlghq1   ///
   using $path3\nindresp.dta , clear
sort nhid
merge nhid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
```

```
drop _merge
gen wave=14
gen year=2004
sort pid
summarize
save $path9\nindresp_ext1.dta , replace
*

** BHPS wave 15 : .
use ohid opsu ostrata using $path3\ohhsamp.dta , clear
sort ohid
save $path9\mtch1.dta, replace
use pid ohid opno osex oage oregion oqfedhi ovote ///
   oxrwght oxewght olrwght olewght ojbcssm ojbrgsc ofimn ohlghq1   ///
   using $path3\oindresp.dta , clear
sort ohid
merge ohid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
drop _merge
gen wave=15
gen year=2005
sort pid
summarize
save $path9\oindresp_ext1.dta , replace
*


** We have produced 15 temporary extract files:
dir $path9\*ext1.dta



************************************************************.
*** Segment 2.2) Merging Panel data in long format .
************************.



*****************.
**  Via the extract files created above:

* (note the use of the stata command 'renpfix' to harmonise variable names)



use $path9\aindresp_ext1.dta , clear
renpfix a z
summarize
save $path9\mtch1.dta, replace
use $path9\bindresp_ext1.dta , clear
renpfix b z
summarize
save $path9\mtch2.dta, replace
use $path9\cindresp_ext1.dta , clear
renpfix c z
summarize
save $path9\mtch3.dta, replace
use $path9\dindresp_ext1.dta , clear
renpfix d z
summarize
save $path9\mtch4.dta, replace
use $path9\eindresp_ext1.dta , clear
renpfix e z
summarize
save $path9\mtch5.dta, replace
use $path9\findresp_ext1.dta , clear
renpfix f z
summarize
save $path9\mtch6.dta, replace
use $path9\gindresp_ext1.dta , clear
renpfix g z
summarize
save $path9\mtch7.dta, replace
use $path9\hindresp_ext1.dta , clear
renpfix h z
```



```
summarize
save $path9\mtch8.dta, replace
use $path9\iindresp_ext1.dta , clear
renpfix i z
summarize
save $path9\mtch9.dta, replace
use $path9\jindresp_ext1.dta , clear
renpfix j z
summarize
save $path9\mtch10.dta, replace
use $path9\kindresp_ext1.dta , clear
renpfix k z
summarize
save $path9\mtch11.dta, replace
use $path9\lindresp_ext1.dta , clear
renpfix l z
summarize
save $path9\mtch12.dta, replace
use $path9\mindresp_ext1.dta , clear
renpfix m z
summarize
save $path9\mtch13.dta, replace
use $path9\nindresp_ext1.dta , clear
renpfix n z
summarize
save $path9\mtch14.dta, replace
use $path9\oindresp_ext1.dta , clear
renpfix o z
summarize
save $path9\mtch15.dta, replace

* (we have created 15 temporary 'mtch' files, with the same variable names)
dir $path9\mtch*.dta


****

use $path9\mtch15.dta, clear
append using $path9\mtch1.dta
append using $path9\mtch2.dta
append using $path9\mtch3.dta
append using $path9\mtch4.dta
append using $path9\mtch5.dta
append using $path9\mtch6.dta
append using $path9\mtch7.dta
append using $path9\mtch8.dta
append using $path9\mtch9.dta
append using $path9\mtch10.dta
append using $path9\mtch11.dta
append using $path9\mtch12.dta
append using $path9\mtch13.dta
append using $path9\mtch14.dta

*(comment - we could append in any order, however, begining with the newest file
*  means that the value labels of wave 15 will usually overwrite earlier labels)

tab wave
tab year

sort pid wave
list pid wave year zsex zage zvote in 1/25
* Observe: the BHPS is an UNBALANCED PANEL  - not everyone
*   is present in every wave
** There are 194322 individual interview records across waves 1-15.
**  Typically there are ~10k interviews waves 1-8, and 16k waves 9-15.

** Use stata's longitudinal commands to describe this data

xtdes, i(pid) t(wave)

** This is very informative :
** 30588 different people are ever given an adult interview over waves 1-15.
** Of these, 4661 are balanced panel (present all 15 waves)
** The next largest groups are :
**    - waves 9-10-11-12-13-14-15 (Scottish and Welsh boosts);
**    - waves 11-12-13-14 (N. Irish boosts);
**    - waves 7-11 (ECHP)
```

```
**  (These three represent the BHPS boost samples).
**    - wave 1 (ie interviewed in first survey and never re-interviewed)

** After that, there's any number of alternative sample inclusions, eg...
xtdes, i(pid) t(wave) patterns(60)
xtdes, i(pid) t(wave) patterns(1000)


sort pid wave
sav $path3b\bh1to15_long.dta , replace

*[THIS DERIVED FILE IS USED AT SEVERAL POINTS ELSEWHERE IN THESE LAB EXERCISES]

*****************************************




*****************************************

** Extension: Merging panel data direct from the source data :
*   (the production of the intermediate temporary 15 'ext1' files was used in order
*      to document the merging process more clearly, but it was not strictly necessary)

** Here's a similar operation with one less intermediate step:

use pid ahid apno asex aage aregion aqfedhi avote ///
        using $path3\aindresp.dta , clear
renpfix a z
gen year=1991
summarize
save $path9\mtch1.dta, replace
use pid bhid bpno bsex bage bregion bqfedhi bvote4 ///
     using $path3\bindresp.dta , clear
rename bvote4 bvote
renpfix b z
gen year=1992
summarize
save $path9\mtch2.dta, replace
use pid chid cpno csex cage cregion cqfedhi cvote ///
       using $path3\cindresp.dta , clear
renpfix c z
gen year=1993
summarize
save $path9\mtch3.dta, replace
use pid dhid dpno dsex dage dregion dqfedhi dvote ///
    using $path3\dindresp.dta , clear
renpfix d z
gen year=1994
summarize
save $path9\mtch4.dta, replace
use pid ehid epno esex eage eregion eqfedhi evote ///
    using $path3\eindresp.dta , clear
renpfix e z
gen year=1995
summarize
save $path9\mtch5.dta, replace
use pid fhid fpno fsex fage fregion fqfedhi fvote ///
     using $path3\findresp.dta , clear
renpfix f z
gen year=1996
summarize
save $path9\mtch6.dta, replace
use pid ghid gpno gsex gage gregion gqfedhi gvote ///
     using $path3\gindresp.dta , clear
renpfix g z
gen year=1997
summarize
save $path9\mtch7.dta, replace
use pid hhid hpno hsex hage hregion hqfedhi hvote ///
     using $path3\hindresp.dta , clear
renpfix h z
gen year=1998
summarize
save $path9\mtch8.dta, replace
```



```
use pid ihid ipno isex iage iregion iqfedhi ivote ///
    using $path3\iindresp.dta , clear
renpfix i z
gen year=1999
summarize
save $path9\mtch9.dta, replace
use pid jhid jpno jsex jage jregion jqfedhi jvote ///
        using $path3\jindresp.dta , clear
renpfix j z
gen year=2000
summarize
save $path9\mtch10.dta, replace
use pid khid kpno ksex kage kregion kqfedhi kvote ///
      using $path3\kindresp.dta , clear
renpfix k z
gen year=2001
summarize
save $path9\mtch11.dta, replace
use pid lhid lpno lsex lage lregion lqfedhi lvote ///
      using $path3\lindresp.dta , clear
renpfix l z
gen year=2002
summarize
save $path9\mtch12.dta, replace
use pid mhid mpno msex mage mregion mqfedhi mvote ///
      using $path3\mindresp.dta , clear
renpfix m z
gen year=2003
summarize
save $path9\mtch13.dta, replace
use pid nhid npno nsex nage nregion nqfedhi nvote ///
      using $path3\nindresp.dta , clear
renpfix n z
gen year=2004
summarize
save $path9\mtch14.dta, replace
use pid ohid opno osex oage oregion oqfedhi ovote ///
      using $path3\oindresp.dta , clear
renpfix o z
gen year=2005
summarize
save $path9\mtch15.dta, replace


****
use $path9\mtch1.dta, clear
append using $path9\mtch2.dta
append using $path9\mtch3.dta
append using $path9\mtch4.dta
append using $path9\mtch5.dta
append using $path9\mtch6.dta
append using $path9\mtch7.dta
append using $path9\mtch8.dta
append using $path9\mtch9.dta
append using $path9\mtch10.dta
append using $path9\mtch11.dta
append using $path9\mtch12.dta
append using $path9\mtch13.dta
append using $path9\mtch14.dta
append using $path9\mtch15.dta

tab year
sort pid year
list pid  year zsex zage zvote in 1/25
xtdes, i(pid) t(year)



*****************************************************************.
*****************************************************************.
```

```
*************************************************************.
***  EXERCISE 3) DERIVING 'WIDE FORMAT' PANEL DATA FROM BHPS COMPONENT FILES   .
*************************************************************.




** 'Wide' format involves one case per person, with entries on that case
*    from multiple time points.
**  Because of this particular design, it is preferable to avoid having large
**  numbers of wave specific variables, as if so, the file will expand horizontally
**  to a point where it becomes difficult to manage .


** To prepare wide format data from 'first principles':.

use pid asex aage avote arace using $path3\aindresp.dta, clear
sort pid
sav $path9\mtch1.dta, replace
use pid bsex bage bvote4 brace using $path3\bindresp.dta , clear
sort pid
sav $path9\mtch2.dta, replace
use pid csex cage cvote crace using $path3\cindresp.dta , clear
sort pid
sav $path9\mtch3.dta, replace
use pid dsex dage dvote drace using $path3\dindresp.dta , clear
sort pid
sav $path9\mtch4.dta, replace
use pid esex eage evote erace using $path3\eindresp.dta , clear
sort pid
sav $path9\mtch5.dta, replace
use pid fsex fage fvote frace using $path3\findresp.dta , clear
sort pid
sav $path9\mtch6.dta, replace
use pid gsex gage gvote grace using $path3\gindresp.dta , clear
sort pid
sav $path9\mtch7.dta, replace
use pid hsex hage hvote hrace using $path3\hindresp.dta , clear
sort pid
sav $path9\mtch8.dta, replace
use pid isex iage ivote irace using $path3\iindresp.dta , clear
sort pid
sav $path9\mtch9.dta, replace
use pid jsex jage jvote jrace using $path3\jindresp.dta , clear
sort pid
sav $path9\mtch10.dta, replace
use pid ksex kage kvote krace using $path3\kindresp.dta , clear
sort pid
sav $path9\mtch11.dta, replace
use pid lsex lage lvote lrace using $path3\lindresp.dta , clear
sort pid
sav $path9\mtch12.dta, replace
use pid msex mage mvote mrace using $path3\mindresp.dta , clear
sort pid
sav $path9\mtch13.dta, replace
use pid nsex nage nvote nrace using $path3\nindresp.dta , clear
sort pid
sav $path9\mtch14.dta, replace
use pid osex oage ovote orace using $path3\oindresp.dta , clear
sort pid
sav $path9\mtch15.dta, replace

dir $path9\mtch*.dta



use $path9\mtch1.dta,  clear
merge pid using $path9\mtch2.dta, keep(bvote4 bsex bage brace) _merge(w2inf)
sort pid
merge pid using $path9\mtch3.dta, keep(cvote csex cage crace) _merge(w3inf)
sort pid
merge pid using $path9\mtch4.dta, keep(dvote dsex dage drace) _merge(w4inf)
sort pid
merge pid using $path9\mtch5.dta, keep(evote esex eage erace) _merge(w5inf)
```



```
sort pid
merge pid using $path9\mtch6.dta, keep(fvote fsex fage frace) _merge(w6inf)
sort pid
merge pid using $path9\mtch7.dta, keep(gvote gsex gage grace) _merge(w7inf)
sort pid
merge pid using $path9\mtch8.dta, keep(hvote hsex hage hrace) _merge(w8inf)
sort pid
merge pid using $path9\mtch9.dta, keep(ivote isex iage irace) _merge(w9inf)
sort pid
merge pid using $path9\mtch10.dta, keep(jvote jsex jage jrace) _merge(w10inf)
sort pid
merge pid using $path9\mtch11.dta, keep(kvote ksex kage krace) _merge(w11inf)
sort pid
merge pid using $path9\mtch12.dta, keep(lvote lsex lage lrace) _merge(w12inf)
sort pid
merge pid using $path9\mtch13.dta, keep(mvote msex mage mracel) _merge(w13inf)
sort pid
merge pid using $path9\mtch14.dta, keep(nvote nsex nage nracel) _merge(w14inf)
sort pid
merge pid using $path9\mtch15.dta, keep(ovote osex oage oracel) _merge(w15inf)

summarize
list in 27
drop w*inf
summarize
list pid asex avote bvote4  evote lvote nvote ovote in 200/210



** The main attraction of wide files is the study of transitions

numlabel _all, add
tab avote
tab ovote
label variable avote "Party support in 1991"
label variable ovote "Party support in 2005"
tab avote ovote if (avote >= 1 & avote <= 4 & ovote >= 1 & ovote <= 4), row V


*************



** Fixed in time data on wide files can be a little complicated, since
**  it is a common protocol not to collect the same fixed-in-time data every year.



***** Fixed in time (1): Gender

** Gender is fine
summarize *sex
gen sex=max(asex, bsex, csex, dsex, esex, fsex, gsex, hsex, isex, jsex, ksex, lsex, msex, nsex,
osex)
* (for gender, we can make the assumption that sex is always stable)
tab sex
correlate sex asex bsex csex dsex esex fsex gsex hsex isex jsex ksex lsex msex nsex osex
* (the correlations confirm stability of sex data over time)
* But - there is somebody in wave o with a different sex to other waves:
tab osex
tab osex asex
tab osex sex
* (In wave 15, there are 4 people for whom sex isn't recorded, but the max value from other waves
*    probably allocates them appropriately)



***** Fixed in time (2): Age

** Age is ok but note that the gap between waves isn't exactly one year.
summarize *age

list *age ///
  if (aage==bage) & (kage==lage)  & (aage >= 16 & aage <= 99 & kage >= 16 & kage <= 99) & avote==1
* (ie these are some people who have birthdays around the time of the BHPS interviews)
```

```
* For age data it is usually reasonable to subtract 1 year for a wave to wave difference
* (although remember that sometimes the gap between interviews is less or more).
gen age=-999
replace age=oage-15 if oage >= 15 & oage <= 102
replace age=nage-13 if nage >= 15 & nage <= 102
replace age=mage-12 if mage >= 15 & mage <= 102
replace age=lage-11 if lage >= 15 & lage <= 102
replace age=kage-10 if kage >= 15 & kage <= 102
replace age=jage-9 if jage >= 15 & jage <= 102
replace age=iage-8 if iage >= 15 & iage <= 102
replace age=hage-7 if hage >= 15 & hage <= 102
replace age=gage-6 if gage >= 15 & gage <= 102
replace age=fage-5 if fage >= 15 & fage <= 102
replace age=eage-4 if eage >= 15 & eage <= 102
replace age=dage-3 if dage >= 15 & dage <= 102
replace age=cage-2 if cage >= 15 & cage <= 102
replace age=bage-1 if bage >= 15 & bage <= 102
replace age=aage if aage >= 15 & aage <= 102
* (doing the replaces in this order gives priority to wave a)
label variable age "Age in 1991"
tab age
histogram age if age >= 1



***** Fixed in time (3): Ethnicity

** This is more problematic:
*   - administratively, ethnicity is only asked once
summarize *race*
list *race* in 1/2
list *race* in 30000/30002
*    - administratively, different ethnicity questions are asked at different times
tab arace
tab oracel
*    - conceptually, people could change their subjective identities, but the administrative
*       design doesn't allow for this.

*** Any solution to this data is likely to involve some substantive decisions, e.g.

gen race1=max(arace,brace,crace,drace,erace,frace,grace,hrace,irace,jrace,krace,lrace)
gen race2=max(mracel,nracel,oracel)
label values race1 arace
label values race2 mracel
tab race1
tab race2

gen carib=(race1==2 | race2==6 | race2==14)
tab carib

** Aside: Many of the most problematic fixed-in-time variables in the BHPS have been
*        reviewed by the data producers, with recommended values stored in the xwav* files.



***

** Make a wide format extract file:

sort pid
keep pid sex age ///
  avote bvote4 cvote dvote evote fvote gvote hvote ivote jvote kvote lvote mvote nvote ovote
summarize

save $path3b\bh1to15_wide.dta, replace

** NOTE THAT THIS FILE IS USED IN SEVERAL OTHER PARTS OF THESE LABS
```



```
************************************************************.
***  EXERCISE 4) TRANSLATING BETWEEN WIDE AND LONG FORMAT PANEL DATA   .
************************************************************.



*** A great feature of STATA is the ability to use 'reshape'
*** to switch quickly between wide and long format data
*** (rather than after writing screeds of programming..)

**** A)  Long to wide : .


use $path3b\bh1to15_long.dta , clear
summarize
xtdes, i(pid) t(wave)

** It's best to keep only a handful of variables for this :

keep pid year zsex zvote
summarize
reshape wide zsex zvote, i(pid) j(year)
summarize

** This has created a wide format file similar to that produced above
** New variables have been made by appending the year number to the orig varname



**** B) Wide to long :  .


use $path3b\bh1to15_wide.dta , clear
summarize
** Names for *vote are not of format that reshape command needs (must be standardised)
rename avote vote1991
rename bvote4 vote1992
rename cvote vote1993
rename dvote vote1994
rename evote vote1995
rename fvote vote1996
rename gvote vote1997
rename hvote vote1998
rename ivote vote1999
rename jvote vote2000
rename kvote vote2001
rename lvote vote2002
rename mvote vote2003
rename nvote vote2004
rename ovote vote2005


summarize
reshape long vote, i(pid) j(year)
summarize
tab vote year

* ie this has created something similar to the long format file we produced above.

************************************



************************************************************.
************************************************************.




************************************************************.
*** EXERCISE 5:  UNDERSTANDING THE BHPS: SAMPLE ORIGINS, AND USING WEIGHTS
```

```
*********************************************************.


** The complication with the BHPS is that different respondents
** have different status's as members of the sample - see the
**  associated handout, Table 2.



*******************.
** Firstly, an important issue is to take care about which
** weights should be used : .


* Most BHPS individual data is on the 'indresp' files for adult interviews in each wave.
use $path3\lindresp.dta, clear
tab lsex
tab lregion
* Note the unusual distribution: the BHPS includes regional 'booster' samples.
* If you want to do a nationally representative analysis within a particular
*  individual year, use an appropriate cross-sectional weight :.
tab lregion [aweight=lxrwtuk1]

** there are actually a lot of different weights - see the BHPS manual -
**  but the main relevant ones are:  .

/*
wxrwght : individual files, cross-sectional weight for BHPS main sample only
wxrwtuk1: individual files, cross-sectional weight incorporates BHPS boost sample data
                    (waves 9 -> only)
wxrwtuk2: individual files, cross-sectional weight for within boost regions
                    (waves 9 -> only, ie, to allow national level analysis on Wales, Scotland or N
Irel)
wxewght : individual files, cross-sectional weight for BHPS main sample for all ennumerated
                adults (not just for all interviewed adults)
wlrwght : individual files (except wave A), weighting for balanced panel longitudinal analysis
                (only if starting at wave A and following original adult respondents)
whhwght : household files, weighting for households to national population
wxhwtuk1 :  household files, weighting for households to national population + boost population
wxhwtuk2: household files, cross-sectional weight for within boost regions
                    (waves 9 -> only, ie, to allow national level analysis on Wales, Scotland or N
Irel)
*/

** Another example : compare Scotland, Wales  and N Ireland:.
use $path3\lindresp.dta, clear
numlabel _all, add
tab lsex
tab ljbrgsc
tab lregion
tab lregion [aweight= lxrwtuk2]
tab ljbrgsc lregion [aweight= lxrwtuk2] if (lregion >=17 & ljbrgsc >= 1), col
* (ie, these weight are representative in each region, but don't deflate regional sample sizes)



*******************
** The syntax below illustrates a way of deriving the component
*   information of the table used in the BHPS handout,
*    from the source BHPS data (for selected waves only):
***********.
*** Membership of the BHPS :.


use pid memorig using $path3\xwlsten.dta, clear
numlabel _all, add
tab memorig
sort pid
sav $path9\mtch1.dta, replace

use $path3\aindall.dta, clear
sort pid
merge pid using $path9\mtch1.dta
```



```
tab _merge
keep if (_merge==3)
gen vfo=aivfio
recode vfo 1 2=1 *=2
label define vfol 1 "Adult Interview" 2 "Other"
label values vfo vfol
tab vfo memorig
*.
use $path3\bindall.dta, clear
rename bsampst sampst
sort  pid
merge pid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
gen vfo=bivfio
recode vfo 1 2 3=1 *=2
label define vfol 1 "Adult Interview" 2 "Other"
label values vfo vfol
gen sampst2=sampst
recode sampst2 5=2 *=1
label define sampst2l 1 "OSM" 2 "TSM"
label values sampst2 sampst2l
tab vfo memorig
sort sampst2
by sampst2: tab vfo memorig
*

use $path3\lindall.dta, clear
rename lsampst sampst
sort pid
merge pid using $path9\mtch1.dta
tab _merge
keep if (_merge==3)
gen vfo=livfio
recode vfo 1 2 3=1  *=2
label define vfol 1 "Adult Interview" 2 "Other"
label values vfo vfol
gen sampst2=sampst
recode sampst2 5=2 *=1
label define sampst2l 1 "OSM" 2 "TSM"
label values sampst2 sampst2l
tab vfo memorig
sort sampst2
by sampst2: tab vfo memorig
*.


*****************************************************.
*****************************************************.




*********************************************************.
*** Exercise 6: DEALING WITH FIXED-IN-TIME BHPS DATA .
*********************************************************.



** A common problem in working with the BHPS is that many variables are
*    only asked once to respondents (on the assumption that they are fixed in time)
*    The result is that an individual's data on that variable may not be within the yearly
*     file that you are looking at.
** Increasingly, the BHPS data suppliers have been addressing this issue by collating
*    more and more fixed in time variables onto the 'xwav' data files from the core data
*    release. Nevertheless, there remain several example variables where fixed in time
*    data still needs to be merged manually.
** The manual solution involves knowing which variables may be affected in this
**  way (there will be a footnote in the documentation on the variable - high numbers of
**   'inapplicables' is the give-away) and treating those variables separately, as follows :.


*************************************************.
*** Example fixed-in-time variable: Father's occupation (RG Social Class).
```

```
* Task : get all available father's occupational info to the wave 15 individual file .
***********************************************.

** Naive mistake : .
use pid ojbrgsc opargsc using $path3\oindresp.dta, clear
numlabel _all, add
tab ojbrgsc
tab opargsc
* It seems odd that so few people have fathers occ info ...;
*  the reason is, for most people it was collected on an earlier wave.


*********************************************.

** Solution:  Collate all reports of fathers occ info between BHPS surveys :.
*             (note from the documentation it was only collected in waves 1, and 8 onwards).

use pid apargsc using $path3\aindresp.dta, clear
sort pid
save $path9\mtch1.dta, replace
use pid hpargsc using $path3\hindresp.dta, clear
sort pid
save $path9\mtch8.dta, replace
use pid ipargsc using $path3\iindresp.dta, clear
sort pid
save $path9\mtch9.dta, replace
use pid jpargsc using $path3\jindresp.dta, clear
sort pid
save $path9\mtch10.dta, replace
use pid kpargsc using $path3\kindresp.dta, clear
sort pid
save $path9\mtch11.dta, replace
use pid lpargsc using $path3\lindresp.dta, clear
sort pid
save $path9\mtch12.dta, replace
use pid mpargsc using $path3\mindresp.dta, clear
sort pid
save $path9\mtch13.dta, replace
use pid npargsc using $path3\nindresp.dta, clear
sort pid
save $path9\mtch14.dta, replace
use pid opargsc using $path3\oindresp.dta, clear
sort pid
save $path9\mtch15.dta, replace


use $path9\mtch1.dta, clear
merge pid using $path9\mtch8.dta, _merge(w8)
sort pid
merge pid using $path9\mtch9.dta, _merge(w9)
sort pid
merge pid using $path9\mtch10.dta, _merge(w10)
sort pid
merge pid using $path9\mtch11.dta, _merge(w11)
sort pid
merge pid using $path9\mtch12.dta, _merge(w12)
sort pid
merge pid using $path9\mtch13.dta, _merge(w13)
sort pid
merge pid using $path9\mtch14.dta, _merge(w14)
sort pid
merge pid using $path9\mtch15.dta, _merge(w15)

summarize
sort pid
egen pargsc= rmax(apargsc hpargsc ipargsc jpargsc kpargsc lpargsc mpargsc npargsc opargsc)
label list
label values pargsc apargsc
numlabel _all, add
tab pargsc
sort pid
save $path9\m1.dta, replace

** Now return to the individual file but match against the cross-wave data : .
use pid ojbrgsc opargsc using $path3\oindresp.dta, clear
tab ojbrgsc
tab opargsc
```



```
sort pid
merge pid using $path9\m1.dta
tab _merge
keep if ( _merge==3)
numlabel _all, add
tab ojbrgsc
tab opargsc
tab pargsc
* Much higher coverage on this variable now (though still not complete) .

mvdecode ojbrgsc opargsc pargsc, mv(-9,-8,-7)
tab ojbrgsc opargsc, col V
tab ojbrgsc pargsc, col V

************************************************************.

***********************************************************.
***********************************************************.
*** Exercise 7:  EXAMPLES USING THE BHPS YOUTH FILES
***********************************************************.
***********************************************************.

** One particular attraction of the BHPS youth files is the
**  potnetial to link youth responses with later adult entries :.

****   Exercise: Link data from the 1994 BHPS youth samples (11-15yrs)
****    with data from the 2005 adult response (22-26 yrs) .
*       {this exercise was also done in lab1}

use pid dypcomp dypsex using $path3\dyouth.dta , clear
sort pid
summarize
save $path9\mtch1.dta, replace


use pid oqfedhi ojbrgsc osex oage using $path3\oindresp.dta, clear
sort pid
merge pid using $path9\mtch1.dta
tab _merge

** The merge data here shows :
** 15,180 adults in wave l5 (2005) but not in d youth (1994)
** 326 kids in d youth (1994) but not in 2005 adult
** 447 kids in d youth (1994) who are also present in 2005 adult


tab dypsex osex
keep if (_merge==3)
tab oage

numlabel _all, add
tab oqfedhi
tab ojbrgsc
tab dypcomp
mvdecode oqfedhi , mv(-9,-7)
mvdecode ojbrgsc , mv(-9,-8)
mvdecode dypcomp, mv(-9)
recode dypcomp 3=2
tab dypcomp

tab oqfedhi dypcomp, col
* The quals profile of comp users seems higher - but not dramatically different
tab ojbrgsc dypcomp, col
gen profmang=(ojbrgsc==1 | ojbrgsc==2) if ojbrgsc >= 1 & ojbrgsc <= 7
```

```
label variable profmang "Advantaged current occupation"
tab profmang dypcomp, col chi V
graph bar (mean) profmang, over(osex) over(dypcomp) asyvar  ///
   title("Proportion 22-26 year olds with advantaged job in 2005") ///
    subtitle("by family access to a computer in 1994")
* The occupational position of comp users seems slightly better


***********************************************************.



**********************************************************
**********************************************************.




***********************************************************.
*** Exercise 8:  FURTHER EXAMPLES IN MATCHING PANEL DATA FILES
***********************************************************.



**********************************************************
*** Other materials : The latter sections of the lab 0 file
**  (introduction to Stata) included several data file matching
**  examples. One of those involved 'relationships between cases'
**  file matching on the  BHPS - reproduced below.
**  There is also below an additional note on STATA's 'joinby' command.

**********************************************************
***  i)   Relationships-between-cases Matching

**** Example : Link a previous record with later ones
****           Match a spouse's job to own case
****           Link info on every household sharer with every other one


*** Link spouse's job status to an individual record .

use ljbstat lsppid using $path3\lindresp.dta , clear
rename lsppid pid
rename ljbstat spjstat
label variable spjstat "Spouse's job status"
sort pid
summarize
save $path9\mtch1.sav , replace

use pid ljbstat lsppid  lmastat lsex using $path3\lindresp.dta, clear
sort pid
merge pid using $path9\mtch1.sav
tab _merge
** _merge=1 : 6825 cases on the current file without spouses (the same 6825)
** _merge=2 : 6825 incoming cases without spouses
** _merge=3 : 9772 cases with spouse's data matched

tab _merge lmastat
* (ie lmastat is only valid for merge=1 or 3

tab ljbstat spjstat if (_merge==3)
sort lsex
by lsex: tab ljbstat spjstat if (_merge==3)
by lsex: tab spjstat ljbstat if (_merge==3 & (ljbstat==2 | ljbstat==3)), col

** These tables show relation between own job status and spouse's



************************************************.
************************************************.
****** ii) Stata's 'joinby' command :

use $path3b\bh1to15_long.dta, clear
```



```
numlabel _all, add
summarize

** 'Joinby' is a very powerful command but one this it is easy to
**    go wrong with .
**  It identifies groups of cases then links all info on every
**  relevant variable to every case

keep pid zhid zage

sort zhid pid
save  $path9\mtch1.dta, replace
rename pid pidoth
rename zage zageoth
save  $path9\mtch2.dta, replace

use $path9\mtch1.dta, clear
summarize
joinby zhid using $path9\mtch2.dta
sort zhid pid
list in 20/40
drop if (pid==pidoth)
list in 20/40
summarize
* Joinby has made a new case for every combination of own pid
*  and household sharer's pid.


***********************************************************.



*****************************************************************
*****************************************************************




***********************************************************.
*** Exercise 9: LONG FORMAT DATA: DESCRIPTIVE ANALYSES .
***********************************************************.



************************************************
***  (i): Describing patterns within cases .

use $path3b\bh1to15_long.dta, clear
* (This file is a long format panel extract from the BHPS, derived above).

numlabel _all, add
summarize
xtdes, i(pid) t(wave)

sort pid wave year
list pid wave year zsex zvote in 1/50
* (the 'z''s are arbitrary - they just indicate the data comes from multiple observations
*   on the same people)



**************
** Some helpful data constructions for summarising the data structure:


** a) Generate a dummy for indicating that there is an earlier
**     record for this case (ie, a lag record)

sort pid wave
gen notfirst=(pid==pid[_n-1])
tab notfirst
list pid wave notfirst zsex zage in 1/25
```

```
** b) Identify only the first contact with each person :

sort pid wave
gen first=(pid~=pid[_n-1])
tab first
list pid wave notfirst first zsex zage in 1/25



** c) Generate a number listing the sequence number of contacts
**      per person

sort pid wave
gen contact=1
replace contact=contact[_n-1] +  1 if (pid==pid[_n-1])
tab contact
** See how 4661 cases have 15 contacts, ie, present every single wave
** Everyone must have at least one contact


******************
** Techniques for summarising variables in a panel data context:.


tab zvote
xttab zvote, i(pid)

tab zvote if zvote >= 1
xttab zvote if zvote>=1, i(pid)

* xttab: between figures show how many voters for parties are different people
*      (e.g., 8046 people in one year or another support conservative)
*        within figures show the stability of voting patterns: proportion of
*            records where a person votes for a party, if they do at least once
*      (e.g., of all people who ever support conservative, 63.6% of records are conservative)



** Stata also has a function for calculating 'transition probabilities':
gen vote5=zvote
recode vote5 1=1 2=2 3=3 4/8=4 10/11=5 *=.
label define vote5l 1 "Conservative" 2 "Labour" 3 "Liberal" 4 "Other" 5 "None"
label values vote5 vote5l
numlabel _all, add
tab vote5

xttrans vote5, i(pid) t(wave)

* The transition probabilities show the rates of movement between adjacent years by pids.

mvdecode zjbrgsc, mv(-9/0)
tab zjbrgsc
xttrans zjbrgsc, i(pid)

* Job mobility transitions.


******************



************************************************
***  (ii): Describing patterns, adjusting for individual clustering .

use $path3b\bh1to15_long.dta, clear
numlabel _all, add
summarize
gen lninc=ln(zfimn) if (zfimn >= 100)
summarize lninc
xtsum lninc, i(pid)
*   Between - deviation between people around each person's average log income
*   Within - deviation within people around their own average log income
*   (T-bar is the average number of valid records per person)
```



```
** Treat the repeated measures merely as noise:
ci lninc
svyset, psu(pid)
svydes
svy: mean lninc
** After clustering is accounted for, CI's are wider
*  (deff indicates magnitude of within-person clustering)


svyset, clear
tab zjbrgsc
keep if (zjbrgsc > 0)
sort zjbrgsc
by zjbrgsc: ci lninc
svyset, psu(pid)
svydes
svy: mean lninc, over(zjbrgsc)



*********************************.
** (iii): Using the data to get year by year summaries :

use $path3b\bh1to15_long.dta, clear
numlabel _all, add
summarize
tab zvote
* Do the age profiles of voters change over time?
keep if (zage >= 18 & (zvote==1 | zvote==2 | zvote==3))
sort zvote
graph box zage,  over(year, label(alternate) relabel(1 "91" 2 "92" 3 "93" 4 "94" 5 "95" 6 "96" 7
"97" 8 "98" 9 "99" 10 "00" 11 "01" 12 "02" 13 "03" 14 "04"  15 "05" )) ///
      over(zsex, relabel(1 "Male" 2 "Female"))
graph box zage,  over(year, label(alternate) relabel(1 "91" 2 "2" 3 "3" 4 "4" 5 "5" 6 "6" 7 "7" 8
"8" 9 "9" 10 "0" 11 "1" 12 "2" 13 "3" 14 "4" 15 "5" )) ///
      over(zvote, relabel(1 "Conservative" 2 "Labour" 3 "Liberal")) by(zsex)

summarize
collapse (mean) zage, by(zsex year zvote)
summarize
list year zsex zage zvote
* This has created a time series dataset - male and female average
*   voting ages for conservative voters
* Comment: (beware) - unlike SPSS's 'aggregate', Stata's 'collapse' command
*  _overwrites_ the current data



*********************************************
**  iv)Describing wide v's long format data :


*** Wide format is handy for the analysis of transitions and sequences :

use $path3b\bh1to15_wide.dta , clear
summarize
numlabel _all, add
tab avote
* Three national election years in period were 1992, 1997, 2001
tab bvote gvote if (bvote >= 1 & bvote <= 3 & gvote >= 1 & gvote <=3), row V
tab bvote kvote if (bvote >= 1 & bvote <= 3 & kvote >= 1 & kvote <=3), row V
tab gvote kvote if (gvote >= 1 & gvote <= 3 & kvote >= 1 & kvote <=3), row V

gen acon=(avote==1) if avote >= 1
gen bcon=(bvote==1) if bvote >= 1
gen ccon=(cvote==1) if cvote >= 1
gen dcon=(dvote==1) if dvote >= 1
gen econ=(evote==1) if evote >= 1
gen fcon=(fvote==1) if fvote >= 1
gen gcon=(gvote==1) if gvote >= 1
gen hcon=(hvote==1) if hvote >= 1
gen icon=(ivote==1) if ivote >= 1
gen jcon=(jvote==1) if jvote >= 1
gen kcon=(kvote==1) if kvote >= 1
```

```
gen lcon=(lvote==1) if lvote >= 1
gen mcon=(mvote==1) if mvote >= 1
gen ncon=(nvote==1) if nvote >= 1
gen ocon=(ovote==1) if ovote >= 1

egen propcons=rmean(acon - ocon)
egen ncons=rsum(acon - ocon)
summarize propcons
tab ncons



*** Some helpful wide-format data indicators:

*** i) Generate an indicator of number of waves present with voting info :

summarize
egen numvote=neqany(*vote bvote4), values(-9/-1,1/11)
tab numvote
* This shows that there are 30588 people in the datafile;
*  numvote is the number of times we have voting records for
*  each of them: 5775 have only one; 3011 only two, etc.
* (note we hade to use *vote and bvote4 to identify all vote variables)



*** ii) Pick out the first 'vote' response from each respondent

numlabel avote, add
tab avote
egen voteone=rfirst(avote bvote4 cvote dvote evote fvote gvote hvote ivote jvote kvote lvote mvote
nvote ovote)
describe avote
label values voteone avote
tab voteone


** iii) Pick out the first non-missing 'vote' response per person

summarize *vote
mvdecode *vote, mv(-9,-8,-7,-2,-1)
summarize *vote
egen votetwo=rfirst(avote bvote4 cvote dvote evote fvote gvote hvote ivote jvote kvote lvote mvote
nvote ovote)
label values votetwo avote
tab votetwo


******************



*** Long format is better suited to looking at adjacent transitions
**     or longer term within person stabilities

use $path3b\bh1to15_long.dta , clear
summarize
numlabel _all, add
gen convot=(zvote==1) if (zvote >= 1)
sort pid wave
gen lconvot=convot[_n-1] if pid==pid[_n-1]
tab lconvot convot, row V
* (Dummy tory voting: now compared to last year)

gen l6convot=convot[_n-6] if pid==pid[_n-6]
tab l6convot convot, row V
* (dummy tory voting now compared to 6 years ago)

* Long format data is well suited to adding further explanatory variables
* to an analyses through panel modelling techniques
************************************************************.


************************************************************.
```



```
************************************************************.
************************************************************.



************************************************************.
*** EXERCISE 10: PANEL DATA ANALYSIS: LOOKING AT TRANSITIONS / SEQUENCES .
************************************************************.



************************************************************
******** (10.1) wave 1 (1991) to wave 12 (2002) voting differences.

use pid asex aage avote using $path3\aindresp.dta, clear
sort pid
sav $path9\mtch1.dta, replace
use pid lsex lage lvote using $path3\lindresp.dta , clear
sort pid
sav $path9\mtch12.dta, replace

use $path9\mtch1.dta,  clear
merge pid using $path9\mtch12.dta, _merge(w12inf)
sort pid
summarize
tab w12inf
list pid asex lsex aage lage avote lvote in 110/130

drop w*inf
summarize


list pid asex avote lvote in 1/10
summarize avote lvote
*(note - non-valid responses are currently excluded; to
*  get them back, it'd be best to reopen the data)
numlabel avote lvote, add
tab avote lvote

** Make some recodes to keep simple :

gen avote4=avote
gen lvote4=lvote
recode avote4 1=1 2=2 3/8=3 10 11=4 *=.
recode lvote4 1=1 2=2 3/8=3 10 11=4 *=.
label define vote4l 1 "Conservative" 2 "Labour" 3 "Other party" 4 "None"
label values avote4 vote4l
label values lvote4 vote4l
tab avote4 lvote4 , row
* The row percents show what happened to people given their 1991 view

* Particular interest might be in 1991 conservatives who switch to
*  2002 labour, cf those who stay 2002 conservative

gen con2lab=0 if (avote==1 & (lvote==1 | lvote==2))
replace con2lab=1 if (avote==1 & lvote==2)
label define con2labl 0 "Stable conservative" 1 "Con 1991 to Labour 2002"
label values con2lab con2labl
tab con2lab

summarize aage
mean aage, over(con2lab)
tab asex
gen afem=(asex==2)
tab con2lab afem, col chi V


logistic con2lab afem aage
logit
** Suggests that Tories who switch votes are younger on average .
```

```
************************************************************
******** (10.2) : wave d youth to wave e youth

** Compare 11-15 year olds in 1994 and 1995 by household transitions, and check those
** against happiness measures .
** Sample of individuals from 1994 and 1995 :.

use $path3\dyouth.dta, clear
numlabel _all, add
tab dypsad
tab dyphlf
tab dypsex
tab dypsad dyphlf
gen dsad=((dypsad==3 | dypsad==4)  | (dyphlf >= 5) | (dypsad==2 & (dyphlf==3 | dyphlf==4)  ) )
tab dsad
keep pid dsad dypsex
sort pid
sav $path9\mtch1.dta, replace

use $path3\eyouth.dta, clear
tab eypsad
tab eyphlf
tab eypsex
tab eypsad eyphlf
gen esad=((eypsad==3 | eypsad==4)  | (eyphlf >= 5) )
sort pid
keep pid esad eypsex
sav $path9\mtch2.dta, replace

* indall files for age and gender .
use pid dhid dage using $path3\dindall.dta , clear
sort pid
sav $path9\mtch3.dta , replace
use pid ehid eage using $path3\eindall.dta , clear
sort pid
sav $path9\mtch4.dta , replace


** Match all these individual files together :.
use $path9\mtch1.dta, clear
sort pid
merge pid using $path9\mtch2.dta, _merge(wy1)
sort pid
merge pid using $path9\mtch3.dta, _merge(wy2)
sort pid
merge pid using $path9\mtch4.dta, _merge(wy3)
tab wy1
tab wy2
tab wy3
keep if wy1==3
*(selects only the 580 youths present in both waves d and e).
summarize
sav $path9\mtch5.dta , replace

** Match in household composition information on both waves.
use dhid dhhsize using $path3\dhhresp.dta, clear
sort dhid
sav $path9\mtch6.dta, replace
use ehid ehhsize using $path3\ehhresp.dta , clear
sort ehid
sav $path9\mtch7.dta, replace

use $path9\mtch5.dta, clear
sort dhid
merge dhid using $path9\mtch6.dta
drop _merge
sort ehid
merge ehid using $path9\mtch7.dta
summarize

tab dhhsize ehhsize
gen hchange=(dhhsize ~= ehhsize)
label variable hchange "Wave 5 household isn't exactly equiv people to wave 4"
```



```
tab hchange
* (most households are stable over 12 months, but 81 youths were in a changing household).
* (with more work it would be possible to know how and why the household is changing,
*   eg parental splits v's sibling moves).

** Do household transitors have different subjective welfare experiences? .

tab hchange dsad, row V
tab hchange esad, row V

** there is a trend though it's not quite significant.
gen sadchang=-999
replace sadchang=1 if (esad==0 & dsad==0)
replace sadchang=2 if (esad==0 & dsad==1)
replace sadchang=3 if (esad==1 & dsad==0)
replace sadchang=4 if (esad==1 & dsad==1)
label define sadcl 1 "Not miserable wave 5" 2 "Cheers up wave 5" ///
   3 "Becomes miserable wave 5" 4 "Stays miserable wave 5"
label values sadchan sadcl
mvdecode sadchang, mv(-999)
numlabel _all, add
tab sadchang
tab hchange sadchang, row V
* Again, there is a suggestion of a data pattern, but not confirmed as significant.

** Comment : lots of fairly arbitrary variables defined above - could have experimented a
**  bit more with them. Also, the data analysis here probably suffers from relatively low
** sample sizes  : the above could be repeated for later waves of the BHPS and a pooled
** sample would increase the analytical sample size (but would need to beware of multiple
** records from the same individuals).




************************************************************.
************************************************************.




************************************************************.
*** EXERCISE 11: LONG FORMAT DATA: PANEL DATA MODELS
************************************************************.



**************************************************
*********   Segment 11.1) Panel data models with metric outcomes
*******************************

**  Panel data regression models : .

use $path3b\bh1to15_long.dta, clear
numlabel _all, add
summarize
gen ghq=zhlghq1 if zhlghq1 >= 0
gen lninc=ln(zfimn) if (zfimn >= 100)
summarize ghq lninc
xtsum ghq lninc
histogram ghq
gen fem=(zsex==2)
gen age=zage if zage >= 16
tab zqfedhi
gen hied=(zqfedhi >= 1 & zqfedhi <= 4) if zqfedhi >= 1
gen noed=(zqfedhi >= 12) if zqfedhi >= 1
gen convot=(zvote==1) if (zvote >= 1)
gen labvot=(zvote==2) if (zvote >= 1)
summarize ghq lninc fem age hied noed convot labvot
correlate ghq lninc fem age hied noed convot labvot



******  Metric regression illustration:

** Linear regression on ghq : not appropriate because records aren't independent
```

```
regress ghq lninc fem age hied noed convot labvot

** Random effects panel on ghq - just controls for clustering

regress ghq
xtreg ghq , i(pid)

regress ghq lninc fem age hied noed convot labvot
xtreg ghq lninc fem age hied noed convot labvot, i(pid)

* (estimates don't change much from linear to random effects model, but the ci's widen,
* and rho, which represents within-person residual heterogeneity, is substantial)

**



*********************.
** Panel on ghq with lag effects
sort pid wave
gen lagghq=ghq[_n-1] if pid==pid[_n-1]
summarize ghq lagghq lninc fem age hied noed convot labvot
correlate ghq lagghq lninc fem age hied noed convot labvot
* (note that the number of cases is changed, n=121414, cf 145842 above
*     - can't have a lag for 1st record)

regress ghq lagghq lninc fem age hied noed convot labvot
xtreg ghq lagghq lninc fem age hied noed convot labvot, i(pid)
** Warning:  estimation with lag-dependent variables is controversial


**********************.

** Some variants on the panel data estimator:
*
regress ghq lninc fem age hied noed convot labvot
* (ignores panel data)
xtreg ghq lninc fem age hied noed convot labvot, pa robust i(pid)
* (population average panel cluster estimator)
regress ghq lninc fem age hied noed convot labvot, cluster(pid)
* (Huber-White cluster estimators)


** Fixed versus random effects with Hausman test
* (But: we have to remove fixed-in-time covariates, fem, hied, loed)
capture est clear
xtreg ghq lninc age convot labvot, fe i(pid)
est store fixed
xtreg ghq lninc age convot labvot, re i(pid)
hausman fixed
** This suggests that the random effects estimators are inconsistent,
*   ie, problematic
**   - but fixed effects estimates are often impractical as can't incorporate
*      fixed in time characteristics

**  A partial solution for fixed in time covariates: structural differences between groups:
sort zsex
by zsex: xtreg ghq lninc age  convot labvot, fe i(pid)


** Different estimators with lag effect models:

regress ghq lagghq lninc fem age hied noed convot labvot
xtreg ghq lagghq lninc fem age hied noed convot labvot, i(pid)
xtreg ghq lagghq lninc fem age hied noed convot labvot, pa robust i(pid)

xtreg ghq lagghq lninc age convot labvot, fe i(pid)
est store fixed
xtreg ghq lagghq lninc age convot labvot, re i(pid)
hausman fixed

* Again suggesting big differences in these estimations
```



```
*************************************************************.
*** Segment 11.2) Panel data models with binary outcomes .
*************************************************************.

use $path3b\bh1to15_long.dta, clear
numlabel _all, add
summarize


gen convot=(zvote==1) if (zvote >= 1)
tab zvote convot
gen labvot=(zvote==2) if (zvote >= 1)

gen lninc=ln(zfimn) if (zfimn >= 100)
summarize  lninc
histogram lninc
gen fem=(zsex==2)
gen age=zage if zage >= 16
tab zqfedhi
gen hied=(zqfedhi >= 1 & zqfedhi <= 4) if zqfedhi >= 1
gen noed=(zqfedhi >= 12) if zqfedhi >= 1
gen ed3=zqfedhi if zqfedhi >=1
recode ed3 1/4=3 5/11=2 12 13=1
label define feml 0 "Male" 1 "Female"
label define ed3l 1 "Lower education level" 2 "Intermediate" 3 "Higher education level"
label values fem feml
label values ed3 ed3l
graph bar (mean) convot, over(fem) over(ed3)


summarize convot lninc fem age hied noed
correlate convot lninc fem age hied noed

** Logistic regression on convot : not appropriate because records aren't independent

logit convot lninc fem age hied noed

** Random effects panel on convot - just controls for clustering

xtlogit convot lninc fem age hied noed, i(pid)

** Panel on conservative voting, with lag effects
sort pid wave
gen lagcvt=convot[_n-1] if pid==pid[_n-1]
summarize convot lagcvt lninc fem age hied noed
correlate convot lagcvt lninc fem age hied noed
 * (note that the number of cases is changed - can't have a lag for 1st record)
logit convot lagcvt lninc fem age hied noed
xtlogit convot lagcvt lninc fem age hied noed, i(pid)
** Warning:  estimation with lag-dependent variables is controversial



***********************************.
**  Alternative estimators for binary data :

logit convot  lninc age hied noed
est store convote1

logit convot  lninc age hied noed, cluster(pid) robust
est store convote2

svyset, psu(pid)
svy: logit convot  lninc age hied noed
est store convote3
svyset, clear

xtlogit convot  lninc age hied noed, re i(pid)
est store convote4

xtlogit convot  lninc age hied noed, pa i(pid)
est store convote4

xtlogit convot  lninc age hied noed, fe i(pid)
est store convote5
```

```
* Beware - the binary fixed effects estimator is not a full population model,
*  but a conditional logit for movers only

est table convote* , star stats(N ll bic)

*******************************************************




*******************************************************
*******************************************************




*******************************************************
** EXERCISE 12) PANEL DATA MODELS: LOOKING AT ESTIMATION ISSUES
***********************************************



use $path3b\bh1to15_long.dta, clear
numlabel _all, add
summarize
gen ghq=zhlghq1 if zhlghq1 >= 0
gen lninc=ln(zfimn) if (zfimn >= 100)
summarize ghq lninc
xtsum ghq lninc
histogram ghq
gen fem=(zsex==2)
gen age=zage if zage >= 16
tab zqfedhi
gen hied=(zqfedhi >= 1 & zqfedhi <= 4) if zqfedhi >= 1
gen noed=(zqfedhi >= 12) if zqfedhi >= 1
gen convot=(zvote==1) if (zvote >= 1)
gen labvot=(zvote==2) if (zvote >= 1)
summarize ghq lninc fem age hied noed convot labvot
correlate ghq lninc fem age hied noed convot labvot
gen fullinf= (ghq >= 0 & lninc >= 0 & fem >= 0 & age >= 0  ///
               & hied >= 0 & noed >= 0 & convot >= 0 & labvot >=0 )
tab fullinf



est clear

regress ghq
scalar ll_lin1=e(ll)
est store lin1
regress ghq, robust cluster(pid)
scalar ll_clus1=e(ll)
est store clus1
xtreg ghq ,  i(pid) be
scalar ll_be1=e(ll)/2
est store be1
xtreg ghq ,  i(pid) mle
scalar ll_re1=e(ll)
est store re1

regress ghq lninc fem age hied noed convot labvot
scalar ll_lin2=e(ll)
est store lin2
regress ghq lninc fem age hied noed convot labvot, robust cluster(pid)
scalar ll_clus2=e(ll)
est store clus2
xtreg ghq lninc fem age hied noed convot labvot, i(pid) be
scalar ll_be2=e(ll)/2
est store be2
xtreg ghq lninc fem age hied noed convot labvot, i(pid) mle
scalar ll_re2=e(ll)
est store re2


est stats
```



```
display ll_lin1
display ll_lin2
display ll_clus1
display ll_clus2

display ll_be1
display ll_be2
display ll_re1
display ll_re2

* The above show the six different model log-likelihoods

est table  lin2 clus2 be2  re2, star stats(N r2 r2_p ll)
* This shows that the interpretation of coefficient effects is broadly
*   robust across models

scalar ell_null=ll_lin1

scalar pr1=0
scalar pr2=1 - (ll_lin2 / ell_null)
scalar pr3=1 - (ll_clus1 / ell_null)
scalar pr4=1 - (ll_clus2 / ell_null)
scalar pr5=1 - (ll_be1 / ell_null)
scalar pr6=1 - (ll_be2 / ell_null)
scalar pr7=1 - (ll_re1 / ell_null)
scalar pr8=1 - (ll_re2 / ell_null)

display pr1
display pr2
display pr3
display pr4
display pr5
display pr6
display pr7
display pr8

* The above show approximations to r2 - proportional model improvements for each model

* (comment - the be estimators give deviance, not ll)



************************************************************.


capture log close


************************************************************.
************************************************************.




*********************************************************************




**** EOF .
```