```
STOP




*****************************************************.
***  Longitudinal Data Analysis for Social Science Researchers
**
**
** ESRC Researcher Development Initiative training programme:
**
**
**   Training materials lab 1:
**     INTRODUCTORY LONGITUDINAL DATA ANALYSIS AND DATA MANAGEMENT -
**         5 APPROACHES TO QUANTITATIVE LONGITUDINAL DATA ANALYSIS .
**
**
**      www.longitudinal.stir.ac.uk
**      Paul Lambert / Vernon Gayle, 26 August 2007
*****************************************************.



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



*****************************************************.
*****************************************************.
** The file below covers introductory examples of five approaches to
**       quantitative longitudinal data analysis:
**
***    Section 1: Repeated cross-sectional survey data
***    Section 2: Panel survey data
***    Section 3: Cohort study survey data
***    Section 4: Event history survey data
***    Section 5: Time series statistical data
**
*****************************************************.



*****************************************************.
** 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) Downloadable from the LDA site :
* - gb91soc2000.dat (this is used during variable constructions for the LFS exercise)
*
*
** 2) Downloadable from the UK Data Archive:
*  - ssa02.dta, ssa01.dta, ssa00.dta and ssa99.dta
*   (Scottish social attitudes 2002, 2001, 2000, 1999,
*     stata datasets for study numbers 4808, 4804, 4503, 4346, Stata format files)
*
*  - f87511.dta, qlfsja96.dta, qlfsja01.dta
*     (Labour Force Surveys mid 1991, 96, 2001 respectively,
*       stata datasets from study numbers 2875, 3647, 4448)
*
*  -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, ~153 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)
*
*  - 2364a.dta (National Child Development Study teaching dataset 1958-1981,
*   UKDA study number 2364  , Stata data files from the zip archive 2364STATA6.ZIP)
*
*
** 3) EITHER supplied by the workshop session instructors (taught workshops)
*      OR as a product of carrying out the lab2 stata commands:
*
* bh1to14_long.dta
* bh1to14_wide.dta
*
* (derived BHPS data files obtained from merging source BHPS files in lab 2).
*  Note - these derived files are used in the latter half of this command file,
*   but they are only formally derived in the lab 2 command file.
*
*
*****************************************************.

** .


*****************************************************.
** 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 a folder where you have saved the
*   open access online data files mentioned above  -
*  see http://www.longitudinal.stir.ac.uk/workshop_materials.html)


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


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

global path3b "d:\data\bhps\derived\"
* (the location of a folder where you have saved the BHPS derived
*   data files bhps1to14_long.dta and bhps1to14_wide.dta 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 path4 "d:\data\ssa\"
* (the location of your copies of the SSA data files mentioned above)


global path5 "d:\data\lfs\lda\"
* (the location of your copies of the LFS data files mentioned above)


global path6 "d:\data\ncds\teaching\"
* (the location of your copies of the NCDS teaching dataset described above)



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

**
**
** ii) 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
* (clears any other data within Stata)

set more off
* (switches off the default setting whereby output is shown one page at a time only)
* (may need to enter this manually for it to stay persistent)

set memory 128M
* (expands the memory allocated to the stata session, usually necessary if using large files)

capture log close
capture log using $path1\log_lab1.txt, replace text
* (These commands close any previous log file, then set a new log file for this
*   lab, a plain text file where basic output is saved, located in the directory defined
*   as 'path1').
**
**
**
*******************************************************.



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


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

**

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


*** LAB 1:  REVIEWING LONGITUDINAL DATA FILE TYPES


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




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




******************************************************
*** SECTION 1) REPEATED CROSS-SECTIONAL STUDIES
******************************************************



************************
***** EXAMPLE A) SCOTTISH NATIONALISM 1999-2002

* Prepare a pooled dataset from the Scottish Social
*  attitudes surveys 1999, 2000, 2001, 2002


**** EXAMPLE A: REPEATED CROSS-SECTION : POOLING DIFFERENT YEARS
****             FROM THE SCOTTISH SOCIAL ATTITUDES SURVEY
```



```
** The commands below open up 4 years of the Scottish Social
**  Attitudes survey, save a selection of variables, then add
**  them all together

use serial rsex rage marstat scotpar2 ukintnat wtfactor using $path4\ssa02.dta , clear
gen year=2002
save $path9\m1.dta, replace

use serial rsex rage marstat scotpar2 ukintnat wtfactor using $path4\ssa01.dta , clear
gen year=2001
save $path9\m2.dta, replace

use serial rsex rage marstat scotpar2 ukintnat wtfactor using $path4\ssa00.dta , clear
gen year=2000
save $path9\m3.dta, replace

use serialno rsex rage marstat scotpar2 ukintnat wtfactor using $path4\ssa99.dta , clear
gen year=1999
gen serial=serialno
drop serialno
save $path9\m4.dta, replace

* Comment: all these variables are collected and equivalent between
*   the various survey years; beware: it usually takes quite a
*    lot of effort to pick out appropriate variables like this

** Add them all together :

use $path9\m1.dta, clear
append using $path9\m2.dta
append using $path9\m3.dta
append using $path9\m4.dta
tab year
summarize
save $path9\ssa1.dta , replace


** We've pooled data on variables that are equivalent in each ssa
*  sweep - there aren't many choices though - a lot of variables are
*  not harmonised between even just these 4 years of surveys



use $path9\ssa1.dta , clear
summarize
tab year

** Look at the data structure : check the 'data editor' window, then:

sort year
by year: summarize

** Simple data: every case is a different person, asked the same questions
**  but spread across four different years of the survey

numlabel marstat, add
tab marstat
tab marstat [aweight=wtfactor]
* (note: year specific weighting, shouldn't be used on aggregate sample]
numlabel scotpar2 ukintnat , add
tab scotpar2
tab ukintnat

** Do values on Scotland change between years?

tab scotpar2 year if (scotpar2 >= 1 & scotpar2 <=5), col chi V
graph bar (mean) scotpar2, over(year) over(rsex) title("Scotland should be part of UK / Europe")
* scotpar2: this is a slight association with time, but it has a variable pattern of change
tab scotpar2 year [aweight=wtfactor] if (scotpar2 >= 1 & scotpar2 <=5), col
* weight is valid here for within year percents, however, STATA won't
*  allow a Chi-square on a weighted dataset

tab ukintnat year if (ukintnat >= 1 & ukintnat <=4), col chi V
graph bar (mean) ukintnat, over(year) over(rsex) title("Don't trust the UK government to support
Scotland")
tab ukintnat year [aweight=wtfactor] if (ukintnat >= 1 & ukintnat <=4) , col
```

```
* ukintnat: is a slight association - trust seems to decline over time


*** A more robust analysis would have some multivariate controls
** Outcome variables
numlabel scotpar2 ukintnat, add
tab scotpar2
gen nat1=(scotpar2==1 | scotpar2==2) if (scotpar2 >=1 & scotpar2 <=5)
tab ukintnat
gen nat2=(ukintnat==3 | ukintnat==4) if (ukintnat >=1 & ukintnat <=4)
tab nat1 nat2
** Explanatory variables
gen y2000=(year==2000)
gen y2001=(year==2001)
gen y2002=(year==2002)
numlabel rage, add
tab rage
mvdecode rage, mv(98,99)
numlabel rsex, add
tab rsex
gen fem=(rsex==2)
tab marstat
gen cohab=(marstat==1 | marstat==2) if (marstat >= 1 & marstat <= 5)
summarize nat1 nat2 y2000 y2001 y2002 rage fem cohab
corr nat1 nat2 y2000 y2001 y2002 rage fem cohab
** Total valid cases will be 5981 (listwise deletion)

logistic nat1 y2000 y2001 y2002 rage fem cohab
logit
** nat1 is marginally higher in 2000 and 2002, but not strongly
logistic nat2 y2000 y2001 y2002 rage fem cohab
logit
** nat2 is definitely higher in 2001, then 02, then 00
logistic nat2 year rage fem cohab
logit
* this shows that a linear trend for year is ok, though not as strong
*  as dummies

** Further tests for time effects would involve interacting time
**  with other variables
**  (..though there's no strong effects in this eg)
** Also, more interesting might be to try using control variables
**  that could have changed over the time period
** Lastly, query nature of 'time' : annual surveys but fieldwork
**  is listed as 'June-November' - when single year differences are
**  being studied, we really should return to original data and get
**  month and/or day of interview as well


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


************************************************
***** EXAMPLE B) EMPLOYMENT AND HIGHER EDUCATIONAL QUALIFICATIONS IN 1990'S


**** EXAMPLE B: REPEATED CROSS-SECTIONS : POOLING DIFFERENT
****             YEARS FROM THE UK LABOUR FORCE SURVEY



** The commands below opens up 3 time points from the UK Labour
**  Force Survey survey, saving a selection of variables and
**  adding them all together  (source files are extracts from LFS)
* The original LFS files can be obtained from the UK Data Archive


** Open the source file from mid-1991, edit variables and save .

use sex age sclass qualsm1 using $path5\f87511.dta, clear
summarize sex age sclass qualsm1
gen highdeg=qualsm1
recode highdeg 1=1 *=0
label variable highdeg "Has higher degree"
gen prof=sclass
recode prof 1=1 2/7=2 *=-999
mvdecode prof, mv(-999)
```



```
label define profl 1 "Professional occupation" 2 "Other occupation"
label values prof profl
sort sex
by sex: tab prof highdeg, row
gen year=1991
keep year sex age prof highdeg
save $path9\mtch1.dta , replace


** Open the source file from mid-1996, edit variables and save .
use sex age soclasm quals00 degree using $path5\qlfsja96.dta, clear
summarize sex age soclasm quals00 degree
gen highdeg=degree
recode highdeg 1=1 *=0
label variable highdeg "Has higher degree"
gen prof=soclasm
recode prof 1=1 2/7=2 *=-999
mvdecode prof, mv(-999)
label define profl 1 "Professional occupation" 2 "Other occupation"
label values prof profl
sort sex
by sex: tab prof highdeg, row
gen year=1996
keep year sex age prof highdeg
save $path9\mtch2.dta , replace


** Open the source file from mid-2001, edit variables and save .

use sex age nsecmmj  sc2kmmj soc2km hiqual quals01 degree ///
        using $path5\qlfsja01.dta , clear
summarize sex age nsecmmj  sc2kmmj hiqual quals01 degree

* Problem : occupational categorisations have changed.
* Use an index file available from http://www.camsis.stir.ac.uk/versions.html#Britain
* in order to match ns_sec to rgsc (see that website for further relevant instructions) .
sort soc2km
save $path9\m1.dta, replace
insheet using $path2\gb91soc2000.dat , clear
* (this is a database of infomration on occupations))
summarize
keep if (ukempst==0)
gen soc2km=soc2000
sort soc2km
merge soc2km using $path9\m1.dta
tab _merge
summarize sex soc2km rgsc
drop if (_merge==1)
keep soc2km rgsc sex age nsecmmj  sc2kmmj hiqual quals01 degree
summarize soc2km rgsc sex age nsecmmj  sc2kmmj hiqual quals01 degree
gen highdeg=degree
recode highdeg 1=1 *=0
label variable highdeg "Has higher degree"
gen prof=rgsc
recode prof 1=1 2/5=2 *=-999
mvdecode prof, mv(-999)
label define profl 1 "Professional occupation" 2 "Other occupation"
label values prof profl
sort sex
by sex: tab prof highdeg, row
gen year=2001
keep year sex age prof highdeg
save $path9\mtch3.dta , replace


** Match the three sets of files by appending them (adding) :.
use $path9\mtch1.dta , clear
append using $path9\mtch2.dta
append using $path9\mtch3.dta
tab year
summarize

save $path9\lfsextr1.dta , replace


** Note: the LFS is mainly a repeated cross-sectional survey
** and the respondents from these three years are different people.
```

```
* However the LFS does additionally have a limited panel aspect,
* with repsondents interviewed several times over a 15 month period
*  before leaving the survey

** Some problems with harmonising LFS data:
**   - are the samples selected by the same methods in each survey?
**   - how to apply survey weights?
**   - do the variables have the same meaning each year - eg occupational class boundaries?
**  Comment: the LFS is a bit more complicated in this regard than most other repeated x-sectional
**   surveys, for instance the variable names chosen for the SPSS files vary between surveys
**   unsystematically, although the questions behind them are usually more or less constant
**   (above we used intermittent variable manipulations to overcome this).


use $path9\lfsextr1.dta, clear


tab year
summarize
*LFS data from 3 years with pre-harmonised variables (see section 4.1B)
* Each case is a different person


*** Some analysis examples :

*** i) Time as a group :.
* (the combined file has 451213 cases, but only 199183 of them with vaild occupational
*  classification  - this isn't surprising as the sample covers all age ranges) .
sort year sex
by year sex: tab prof highdeg , row
* There seems to be a trend - see also the **graph below
* (Note - we've not bothered with weights here - we should do really)
graph bar (mean) highdeg , over(prof) over(year) over(sex) ///
  title("Proportion with higher degree by sex, occupation and time")

*** ii) Time as a variable :.
** Log Regression model with categorical vars expressed as dummies :.
* Effects of time can be used in several ways : .
* dummies for time can just control for structural differences over the period.
* interactions between time and other vars show _changing influences over time_.
* Rather artificially, here use time in years as if a continuous var for the interactions.
gen time=year - 1996
gen y1991=(year==1991)
gen y2001=(year==2001)
gen profd=prof if (prof >= 1)
recode profd 1=1 2=0
gen age10=age / 10 if (age >= 20)
gen age102=(age^2) / 1000  if (age >= 20)
gen fem=(sex==2)
gen timedeg=time*highdeg
summarize profd highdeg fem age10 age102 y1991 y2001 timedeg
corr profd highdeg fem age10 age102 y1991 y2001 timedeg
* leaves 187331 cases for analysis after listwise deletions.
* These are 20+ year olds who are currently working

logit profd highdeg fem age10 age102 y1991 y2001 timedeg
logistic profd highdeg fem age10 age102 y1991 y2001 timedeg

* See how the model puts a different emphasis to the table: model interaction shows that
* the benefit of a higher degree on chances of workers being 'professional' is
* actually less in later years (odds ratio for highdeg is <1)
*  ie the expansion of proportions in professional sector with higher degrees has actually
* been a bit less than the overall expansion of proportions with higher degrees.
* (though higher degrees are very influential) .

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



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



************************************************************
************************************************************
```



```
************************************************************
*** SECTION 2) PANEL  DATA
************************************************************.



*** Here we briefly illustrate two alternative formats;
**  Panel data structures will also be covered in lab 2 (using the BHPS) .



************************************************
***i) Panel Format 1: Multiple records per case ('long format'):
************************************************




** Open a pre-prepared file
**  (created in the lab 2 session)



use $path3b\bh1to15_long.dta , clear
summarize
tab year
* This is 15 years of BHPS data pooled together :
* one record per person*contact time combination
* (Created in lab 2).

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

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

xtdes, i(pid) t(wave)

** This is very informative :
** Of 30588 cases, 4661 are balanced panel (present all 15 waves)
** The next 3 largest groups are all present in waves 9-15; 11-15;
**   and 7-11 : The bulk of these cases will be from the 3 BHPS
**  boost samples (Scotland + Wales for 9-15; N. Ireland for 11-15;
**   and ECHP for 7-11).
** After that, there's any number of alternative sample inclusions.

* Get the full gorey details :

xtdes, patterns(1000) i(pid) t(wave)
* (there's probably too much data here for your output display..)


****



** A few analyses :.

summarize
numlabel _all, add
tab zvote
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
* This distribution seems interesting, but what of the panel cluster?
xttab vote5, i(pid)
* This gives us a bit more: the 'between' is the number of people
*  who at some (any) point vote for the given party;
*  the 'within' is the proportion of times voters who ever favour each party
```

```
*  have voted for it out of the total number of contacts: labour voters
*   are the most loyal.
** Note the difference with fixed-in-time info:
tab zsex
xttab zsex
* (because the between and total n's are equal, we see that everyone is stable gender)



** Regresson models :

summarize zjbcssm zsex zage  zhlghq1
mvdecode zjbcssm zage zhlghq1 zage , mv(-9/0)
gen fem=(zsex==2)
gen age2=zage^2
tab zqfedhi
gen hied=(zqfedhi >= 1 & zqfedhi <=4) if zqfedhi > 0
gen noed=(zqfedhi == 12 ) if zqfedhi > 0
summarize zjbcssm fem zage age2 hied noed

regress zjbcssm fem zage age2 hied noed
** But this model doesn't acknowledge clustering effect

xtreg  zjbcssm fem zage age2 hied noed , i(pid)
** This is a random effects panel model - acknowledging clustering



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




**********************************************
*** 2.2ii) Panel data format 2: one case per person ('wide' format)
***  (info from multiple years is matched onto the same person)
**********************************************



** Access a pre-prepared dataset
**  (created in the lab 2 sessions).



** Some summaries of this data:

use $path3b\bh1to15_wide.dta, clear

summarize

list pid sex avote bvote4 cvote dvote evote lvote in 1/10
list pid avote bvote cvote kvote lvote in 2000/2020

** A typical wide format panel dataset : one case per person
*   and records on voting for people over multiple time points


*** Example data manipulation :

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

summarize
egen numvote=neqany(*vote),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 (in the 'vote' vars) for
*  each of them: 6205 have only one; 2810 only two, etc.

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



```
**** Example analysis on a wide file: looking at transitions / sequences

** What about wave 1 (1991) to wave 15 (2005) voting differences?

summarize avote ovote
*(note - non-valid responses are currently excluded; to
*  get them back, it'd be best to reopen the data)
numlabel avote ovote, add
tab avote ovote

** Make some recodes to keep simple :

gen avote4=avote
gen ovote4=ovote
recode avote4 1=1 2=2 3/8=3 10 11=4
recode ovote4 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 ovote4 vote4l
mvdecode avote4 ovote4, mv(-9/-1)
tab avote4 ovote4 , row V
* The row percents show what happened to people in 2005 given their 1991 view

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




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


************************************************************
*** SECTION 3) COHORT STUDY DATA
************************************************************.


** The LDA materials do not examine in great detail any of the major
**  cohort studies' datasets (there are other training materials giving
*   this provision, see eg http://www.cls.ioe.ac.uk/ )

** The defining features of cohort studies are equivalent to those of panel
*   studies, and all the data management and data analysis issues which
*   apply to panel studies are, technically, the same as those that apply
*   to cohort data

** However  it is useful to appreciate there there tend to be some practical
*   differences between the use of cohort and panel study micro-social survey data.

** The most significant two issues are:

**  (1) Due to the generally longer time gaps between cohort studies' observations,
*        data management tends to be more complex, because there is higher attrition
*        and there is greater difficulty in harmonising survey variables between time
*         points


** (2) Due to the more focussed structure of the cohort sample, substantive interest
*       tends to be directed far more towards past influences on future behaviours
*       and lifecourse trajectories (whereas panel studies are often concerned more with
*       total sample propensities and the prevalence of general social processes)


** Most often, cohort study datasets are arranged in a 'wide' format
*   (though there can be a great deal of work required to construct such data)


** Example: NCDS subsample (teaching dataset)

use $path6\2364a.dta, clear
summarize

* Note that the end of the variable name indicates which year the data comes from
*  - this study is a cohort of those born in 1958, who were interviewed at ages
```

```
*     0,7,11,16 and 23. This file includes selected variables from each interview


** Example analyses:

numlabel _all, add
tab highqual
gen degdip=(highqual >= 1 & highqual <= 5)
label variable degdip "Has degree or diploma by age 23"
label define degdipl 0 "No degree/diploma" 1 "Has degree / diploma"
label values degdip degdipl
tab degdip

tab pasc0
tab read7
tab likes16
mvdecode pasc0, mv(-1)
tab pasc0 degdip, row V
mvdecode read7, mv(-1)
correlate read7 degdip
mvdecode likes16, mv(-1)
correlate likes16 degdip
gen one=1
graph bar (count) one, over(degdip) over(likes16) asyvar ///
     title("Dislikes school at age 16, by education at age 23")


logit degdip pasc0
est store father
logit degdip pasc0 read7
est store reading7
logit degdip pasc0 read7 likes16
est store school16

est table father reading7 school16, star stats(N r2_p)

** comment - all three have independent main effects


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


************************************************************
*** SECTION 4) EVENT HISTORY DATA
************************************************************.



** In this example we give a quick illustration of some event history
**  techniques using the BHPS life history files.
**  These files will also be discussed in the BHPS teaching sessions,
*    so these preliminary illustrations are somewhat optional .



**** See the BHPS's 'Combined Life History' files

use $path3d\ljempe.dta , clear
summarize
sort pid date
list pid date stemp duration enddate in 1/30

** This file has a series of life events sequential for each person
** and classified by employment activity plus additional employment
** details


histogram duration
* most events are short

numlabel _all, add
tab stemp
tab nextemp
tab stemp nextemp if (stemp > 0 & nextemp > 0)
```



```
****  Calculate 'real' dates

gen syear=floor(date/12) + 1900
gen smonth=date - (floor(date/12))*12
list pid date syear smonth stemp duration in 1/30

* Note: This particular derived file has not been updated since the wave 10 BHPS.
*       (some of the derived life history files have been updated more recently)

*** .



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


************************************************************
*** SECTION 5) TIME SERIES DATA
************************************************************.

** We do not spend much time in the LDA project looking at macro-social Time Series
**   datasets, our focus rather being toward micro-social survey projects.

** For illustrative purposes, here we create a short Time Series database, and
*   describe simple ways in which is may be analysed .



** Data construction: BHPS voting and occupational statistics by region.


use $path3b\bh1to15_long.dta , clear
numlabel _all, add
summarize
tab year
tab zvote
tab zregion
mvdecode zjbcssm, mv(-9/0)
histogram zjbcssm
histogram zfimn
histogram zage
keep if zregion >= 1 & zregion <= 18
gen reg3=zregion
recode reg3 1/16=1 17=2 18=3
tab reg3
tab zsex
gen convot=(zvote==1)
tab convot
tab zjbrgsc
gen working=(zjbrgsc >= 1 & zjbrgsc <= 7)
tab working

sort zsex reg3
table zsex reg3, c(mean convot mean working mean zjbcssm mean zfimn)

sort year zsex reg3
table year zsex reg3, c(mean convot mean working mean zjbcssm mean zfimn)

sav $path9\micro1.dta, replace

** Our main data will be statistics on voting, working, job situation, income,
*  by year, gender and region


collapse (mean) convot (mean) working (mean) zjbcssm (mean) zfimn, by(year zsex reg3)
summarize
list

** This will be our dataset:

label variable zsex "Gender"
label variable year "Year"
label variable reg3 "Region in Britain"
label variable convot "Percentage conservative support"
label variable working "Percentage working"
```

```
label variable zjbcssm "Mean occupational advantage score of employed"
label variable zfimn "Mean income of all adults"
summarize
sav $path9\part1.dta, replace


** Comment: Stata's 'statsby' command is another good way to generate
*   summary statistics like this; an advantage of it is that estimation results
*   from more complex models can be added to these outputs, rather than just the
*   descriptive statistics above.

use $path9\micro1.dta, clear
regress zfimn zjbcssm if working==1
statsby "regress zfimn zjbcssm if working==1" astat=e(r2), by(year zsex reg3)
saving($path9\part2.dta) replace
* this produces a new file with records for the r2 for these regressions for each combination
* of year, gender and region

use $path9\part1.dta, clear
sort year zsex reg3
merge year zsex reg3 using $path9\part2.dta, _merge(regressions)
tab regressions
* (ie there was one combinatin where a regression couldn't be run)
label variable astat "R2 for income predicted by Cambridge scale score"
drop regressions
summarize


save $path1\bhps_time_series.dta, replace




** Some illustrative time series data analysis

*****************************
*** i) Descriptive analysis

* Often researchers are just interested in describing simple patterns by time:



use $path1\bhps_time_series.dta, clear
table year zsex if reg3==1, c(mean convot)
graph twoway (scatter convot year) (qfit convot year) if reg3==1 & (zsex==1 | zsex==2), by(zsex) ///
   title("Support for Conservatives") note("Source: BHPS 1991-2005, England only")


* More elaborate graph:
use $path1\bhps_time_series.dta, clear
keep if (reg==1 & zsex==1)
gen convot1=convot
label variable convot1 "England, males"
keep year convot1
sort year
sav $path9\mtch1.dta, replace
use $path1\bhps_time_series.dta, clear
keep if (reg==1 & zsex==2)
gen convot2=convot
label variable convot2 "England, females"
keep year convot2
sort year
sav $path9\mtch2.dta, replace
*
use $path1\bhps_time_series.dta, clear
keep if (reg==2 & zsex==1)
gen convot3=convot
label variable convot3 "Wales, males"
keep year convot3
sort year
sav $path9\mtch3.dta, replace
use $path1\bhps_time_series.dta, clear
keep if (reg==2 & zsex==2)
gen convot4=convot
label variable convot4 "Wales, females"
keep year convot4
```



```
sort year
sav $path9\mtch4.dta, replace
*
use $path1\bhps_time_series.dta, clear
keep if (reg==3 & zsex==1)
gen convot5=convot
label variable convot5 "Scotland, males"
keep year convot5
sort year
sav $path9\mtch5.dta, replace
use $path1\bhps_time_series.dta, clear
keep if (reg==3 & zsex==2)
gen convot6=convot
label variable convot6 "Scotland, females"
keep year convot6
sort year
sav $path9\mtch6.dta, replace

use $path9\mtch1.dta, clear
merge year using $path9\mtch2.dta
drop _merge
sort year
merge year using $path9\mtch3.dta
drop _merge
sort year
merge year using $path9\mtch4.dta
drop _merge
sort year
merge year using $path9\mtch5.dta
drop _merge
sort year
merge year using $path9\mtch6.dta
drop _merge
summarize

graph twoway (scatter convot1 year) (qfit convot1 year)  ///
    (scatter convot2 year) (qfit convot2 year, clpattern(shortdash)) ///
     (scatter convot3 year) (qfit convot3 year)  ///
   (scatter convot4 year) (qfit convot4 year, clpattern(shortdash))  ///
   (scatter convot5 year) (qfit convot5 year)  ///
 (scatter convot6 year) (qfit convot6 year, clpattern(shortdash)), ///
   title("Support for Conservatives") note("Source: BHPS 1991-2005")


*****************************
*** ii) Modelling

use $path1\bhps_time_series.dta, clear
summarize
gen fem=(zsex==2)
gen wal=(reg3==2)
gen scot=(reg3==3)

** Basic regressions could be used:
regress convot year
regress convot year fem wal scot
regress convot year fem wal scot zfimn zjbcssm

** Usually however some more complex structures to the data are modelled, these include:
*   - non-linear trends in time dependence
*   - the possible role of autocorrelations in the time series (ie, lag values)
*   - the interrelation between explanatory variables and lagged variables

sort zsex reg3 year
gen lagcv=convot[_n-1] if (zsex==zsex[_n-1] & reg3==reg3[_n-1])
summarize convot lagcv

regress convot year lagcv

** The science of studying Time Series structures is well developed in economics.
**  Stata has a number of specialist functions for this purpose.
**  For further training materials see, for example, http://www.bized.ac.uk/timeweb/ .



************************************************************
************************************************************
```

```
capture log close

**************************************************
**************************************************
**** EOF
```