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



**** SPSS 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.por (this is used during variable constructions for the LFS exercise)
*
*
** 2) Downloadable from the UK Data Archive:
*  - ssa02.por, ssa01.por, ssa00.por and ssa99.por
*   (Scottish social attitudes 2002, 2001, 2000, 1999,
*     SPSS datasets for study numbers 4808, 4804, 4503, 4346, Stata format files)
*
*  - f87511.por, qlfsja96.por, qlfsja01.por
*     (Labour Force Surveys mid 1991, 96, 2001 respectively,
*       SPSS datasets from study numbers 2875, 3647, 4448)
*
*  -All BHPS Waves 1-15 component files in SPSS format (UK Data Archive Study number
*   5151 (June 2007 release) (extracted from the zip file 5151SPSS.ZIP)
*      (warning - these are a large volume of files, ~152 different files, ~ 600MB)
*
* - All SPSS format 'episode' files from the BHPS Derived life history files
*     (UKDA study number 3954) (some cover waves 1-10 only) (you want to access
*     the 3 files data files on the top directory of the '3954.zip' archive,
*      called newpan.por, xlempe.por, and xljobe.por, plus the 3 files in the
```



```
*      'episode' subfolder of the '3954.zip' archive, called l*.por)*
*
*  - 2364a.por (National Child Development Study teaching dataset 1958-1981,
*   UKDA study number 2364  , SPSS data files from the zip archive 2364SPSS.ZIP)
*
*
** 3) EITHER supplied by the workshop session instructors (taught workshops)
*      OR as a product of carrying out the lab2 stata commands:
*
* bh1to15_long.sav
* bh1to15_wide.sav
*
* (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 : .


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


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


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

define !path3b () 'd:\data\bhps\derived\' !enddefine.
* (the location of a folder where you have saved the BHPS derived
*   data files bhps1to15_long.dta and bhps1to15_wide.dta mentioned above) .


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


define !path4 () 'd:\data\ssa\' !enddefine.
* (the location of your copies of the SSA data files mentioned above)


define !path5 () 'd:\data\lfs\lda\' !enddefine.
* (the location of your copies of the LFS data files mentioned above)


define !path6 () 'd:\data\ncds\teaching\' !enddefine.
* (the location of your copies of the NCDS teaching dataset described above)



define !path9 () 'd:\temp\' !enddefine.
* (a location of a temporary folder where you can save intermediate files) .


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

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



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

import file=!path4+"ssa02.por"
  /keep=serial rsex rage marstat scotpar2 ukintnat wtfactor .
compute year=2002 .
sav out=!path9+"m1.sav".

import file=!path4+"ssa01.por"
    /keep=serial rsex rage marstat scotpar2 ukintnat wtfactor .
compute year=2001 .
sav out=!path9+"m2.sav".

import file=!path4+"ssa00.por"
     /keep=serial rsex rage marstat scotpar2 ukintnat wtfactor .
compute year=2000.
sav out=!path9+"m3.sav".

import file=!path4+"ssa99.por"
      /keep=serialno rsex rage marstat scotpar2 ukintnat wtfactor .
compute year=1999 .
```



```
compute serial=serialno .
sav out=!path9+"m4.sav" /drop=serialno .

* 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 : .

get file=!path9+"m1.sav".
sort cases by serial.
add files file=* /file=!path9+"m2.sav" /file=!path9+"m3.sav"
     /file=!path9+"m4.sav" .
fre var= year.
descriptives var=all.
sav out=!path9+"ssa1.sav".


** 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 .


get file=!path9+"ssa1.sav".
descriptives var=all.
fre var=year.

** Look at the data structure : .
sort cases by year .
split files by year.
descriptives var=all.
split files off.

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

fre var=marstat .
weight by wtfactor.
fre var=marstat.
weight off.
* (note: year specific weighting, shouldn't be used on aggregate sample] .

fre var= scotpar2  ukintnat

** Do values on Scotland change between years?

temp.
select if (scotpar2 ge 1 & scotpar2 le 5) .
cro scotpar2 by year /cells=count col /statistics=chisq phi .
* scotpar2: is an association, but it has a variable pattern of change .

weight by wtfactor.
temp.
select if (scotpar2 ge 1 & scotpar2 le 5) .
cro scotpar2 by year /cells=count col /statistics=chisq phi .
weight off.

temp.
select if (ukintnat ge 1 & ukintnat le 4) .
cro ukintnat by year /cells=count col /statistics=chisq phi .



*** A more robust analysis would have some multivariate controls
** Outcome variables
fre var=scotpar2  .
compute nat1=scotpar2.
recode nat1 (1,2=1) (3,4,5=0) (else=-999).
fre var=ukintnat .
compute nat2=ukintnat.
recode nat2 (3,4=1) (1,2=0) (else=-999) .
cro nat1 by nat2 .

** Explanatory variables .
compute y2000=(year=2000).
compute y2001=(year=2001).
compute y2002=(year=2002).
```

```
fre var=rage .
missing values rage (98,99) .
fre var=rsex.
compute fem=(rsex=2).
fre var=marstat .
compute cohab=marstat.
recode cohab (1,2=1) (3,4,5=0) (else=-999) .
descriptives var= nat1 nat2 y2000 y2001 y2002 rage fem cohab .
missing values nat1 nat2 cohab (-999) .
descriptives var= nat1 nat2 y2000 y2001 y2002 rage fem cohab .
correlate var= nat1 nat2 y2000 y2001 y2002 rage fem cohab .

** Total valid cases will be 5981 (listwise deletion)  .

logistic regression var=nat1 /method=enter y2000 y2001 y2002 rage fem cohab
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).

** nat1 is marginally higher in 2000 and 2002, but not strongly .

logistic regression var=nat2 /method=enter y2000 y2001 y2002 rage fem cohab
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).

** nat2 is definitely higher in 2001, then 02, then 00 .

logistic regression var=nat2 /method=enter year rage fem cohab
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).

* 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 -see below for an alternative application ).
** Also, more interesting might be to try using control variables
**  that could have changed over the time period
** Lastly, query the nature of 'time' : these are 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 and extract relevant variables .
import file=!path5+"lfs91.exp".
fre var=sex age sclass qualsm1 .
compute highdeg=qualsm1.
recode highdeg (1=1) (else=0).
variable label highdeg "Has higher degree".
compute prof=sclass.
recode prof (1=1) (2 thru 7=2) (else=-999).
missing values prof (-999).
add value labels prof 1 "Professional occupation" 2 "Other occupation" .
sort cases by sex.
split files by sex.
cro prof by highdeg /cells=count row .
split files off.
```



```
compute year=1991.
sav out=!path9+"mtch1.sav" /keep=year sex age prof highdeg .

** Open the source file from mid-1996 and extract relevant variables .
import file=!path5+"personja96.exp".
fre var=sex age soclasm quals00 degree .
compute highdeg=degree.
recode highdeg (1=1) (else=0).
variable label highdeg "Has higher degree".
compute prof=soclasm.
recode prof (1=1) (2 thru 7=2) (else=-999).
missing values prof (-999).
add value labels prof 1 "Professional occupation" 2 "Other occupation" .
sort cases by sex.
split files by sex.
cro prof by highdeg /cells=count row .
split files off.
compute year=1996.
sav out=!path9+"mtch2.sav" /keep=year sex age prof highdeg .

** Open the source file from mid-2001 and extract relevant variables .
import file=!path5+"personja01.por".
fre var=sex age nsecmmj  sc2kmmj hiqual quals01 degree  .
compute highdeg=degree.
recode highdeg (1=1) (else=0).
variable label highdeg "Has higher degree".
cro nsecmmj by sc2kmmj .
* Problem : occupational categorisations have changed.
* Use an index file available from http://www.camsis.stir.ac.uk/occunits/distribution.html#UK
* in order to match ns_sec to rgsc (see that website for further relevant instructions) .
sort cases by soc2km .
sav out=!path9+"temp1.sav".
import file=!path2+"gb91soc2000.por".
select if (ukempst=0).
sort cases by soc2000.
rename var (soc2000 = soc2km).
match files file=!path9+"temp1.sav" /table=* /by=soc2km.
fre var=rgsc ns_sec .
cro rgsc ns_sec by nsecmmj .
* comment: the match is not perfect but it seems reasonable.
compute prof=rgsc.
recode prof (1=1) (2 thru 5=2) (else=-999).
missing values prof (-999).
add value labels prof 1 "Professional occupation" 2 "Other occupation" .
sort cases by sex.
split files by sex.
cro prof by highdeg /cells=count row .
split files off.
compute year=2001.
sav out=!path9+"mtch3.sav" /keep=year sex age prof highdeg .

** Data analysis on matched files :.

add files file=!path9+"mtch1.sav" /file=!path9+"mtch2.sav"
          /file=!path9+"mtch3.sav".
descriptives var=all.

*** 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 cases by year sex.
split files by year sex.
cro prof by highdeg /cells=count row .
split files off.
* This gives the data used in the table shown for 2.2.
compute sexprof=sex*10 + prof.
add value labels sexprof 11 "Male, Professional occupation" 12 "Male, other occupation"
        21 "Female, Professional occupation" 22 "Female, other occupation" .
graph /bar=mean(highdeg) by sexprof by year
  /title="Proportion with higher degree by sex, occupation and time" .



*** ii) Modelling approach: 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.
compute time=year - 1996  .
compute y1991=(year=1991).
compute y2001=(year=2001).
compute profd=(prof=1).
compute age10=age / 10.
compute age102=(age**2) / 1000 .
compute fem=(sex=2).
compute timedeg=time*highdeg .
descriptives var=profd highdeg fem age10 age102 y1991 y2001 timedeg /missing=listwise .
* leaves 199183 cases for analysis.
logistic regression var= profd
       /method=enter   highdeg fem age10 age102 y1991 y2001 timedeg
        /criteria pin(.05) pout(.10) iterate (20) cut (.5)  /missing=listwise .

* 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, 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).


** But some problems with pooled LFS analysis:
**   - 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).

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



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



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




************************************************************
*** 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'):
***********************************************


*** Example i : BHPS panel data model for GHQ (happiness) scale given current
** economic activity and education.

** Note, this illustrates a panel dataset, though the SPSS model shown isn't a particularly
** interesting one. For an outcome such as GHQ, which has relatively high variability within
** individual's, a random effects model can be illuminating.

** GHQ, labour force status and highest education over a five year spell 1994-1999.

* wave 4 .
get file=!path3+"dindresp.sav" /keep=pid djbstat dqfedhi dage dsex dhlghq1 .
```



```
* creates variables for use in harmonised file.
compute fem=(dsex=1).
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad=dhlghq1 .
sort cases by pid.
compute wave=4.
sav out=!path9+"mtch4.sav" /keep=pid wave ghqsad fem working unemp study age degdip .
* wave 5 .
get file=!path3+"eindresp.sav" /keep=pid ejbstat eqfedhi eage esex ehlghq1
  /rename (ejbstat eqfedhi eage esex ehlghq1 = djbstat dqfedhi dage dsex dhlghq1)  .
* creates variables for use in harmonised file.
compute fem=(dsex=1).
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad=dhlghq1 .
sort cases by pid.
compute wave=5.
sav out=!path9+"mtch5.sav" /keep=pid wave ghqsad fem working unemp study age degdip .
* wave 6 .
get file=!path3+"findresp.sav" /keep=pid fjbstat fqfedhi fage fsex fhlghq1
  /rename (fjbstat fqfedhi fage fsex fhlghq1 = djbstat dqfedhi dage dsex dhlghq1)  .
* creates variables for use in harmonised file.
compute fem=(dsex=1).
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad=dhlghq1 .
sort cases by pid.
compute wave=6.
sav out=!path9+"mtch6.sav" /keep=pid wave ghqsad fem working unemp study age degdip .
* wave 7 .
get file=!path3+"gindresp.sav" /keep=pid gjbstat gqfedhi gage gsex ghlghq1
  /rename (gjbstat gqfedhi gage gsex ghlghq1 = djbstat dqfedhi dage dsex dhlghq1)  .
* creates variables for use in harmonised file.
compute fem=(dsex=1).
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad=dhlghq1 .
sort cases by pid.
compute wave=7.
sav out=!path9+"mtch7.sav" /keep=pid wave ghqsad fem working unemp study age degdip .
* wave 8 .
get file=!path3+"hindresp.sav" /keep=pid hjbstat hqfedhi hage hsex hhlghq1
  /rename (hjbstat hqfedhi hage hsex hhlghq1 = djbstat dqfedhi dage dsex dhlghq1)  .
* creates variables for use in harmonised file.
compute fem=(dsex=1).
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad=dhlghq1 .
sort cases by pid.
compute wave=8.
sav out=!path9+"mtch8.sav" /keep=pid wave ghqsad fem working unemp study age degdip .

** Add all these together :.

add files file=!path9+"mtch4.sav" /file=!path9+"mtch5.sav"
        /file=!path9+"mtch6.sav" /file=!path9+"mtch7.sav" /file=!path9+"mtch8.sav" /by=pid wave.
descriptives var=all.
** (Note - take a look at the data window to see the structure of this file with multiple records
coming
**   from the same person).
```

```
** How many people :.
fre var=wave.
* This shows the total number of data points (50267) -
*    many of them are multiple contacts with the same people. .
sort cases by pid wave.
compute first=1.
if (pid=lag(pid)) first=0.
fre var=first.
* The number in first=1 is the number of different people : 13802 different people here.

** Note: The BHPS is an UNBALANCED PANEL - people drop in and out over time.


missing values all (lo thru -1).
descriptives var=all.
graph /histogram=ghqsad.

** A 'cross-sectional model' : ignoring panel data structure.

regression var=ghqsad
            fem working unemp study age degdip wave
                   /statistcs=r coeff anova outs collin toll
              /dependent=ghqsad /method=enter .

* (negative coefficients = less sad, positive = more sad).

** Panel data models then use various techniques to acknowledge the fact that the multiple records
** in the data file are related to each other, eg multiple records per person.
** The variance components (random effects, aka multilevel fixed slopes) model
** really only treats the multiple records from the same person as a nuisance, and tries to
respecify
** the model correctly after taking account of this : it is not a very interesting model in most
** circumstances (except when pooling records just to increase sample size), but it is about the
only
** panel model that SPSS can manage : .

mixed ghqsad with fem working unemp study age degdip wave
         /criteria=cin(95) mxiter(100) mxstep(5) scoring(1) singular(0.000000000001)
         hconverge(0,absolute) lconverge(0,absolute) pconverge(0.000001, absolute)
          /fixed fem working unemp study age degdip wave | sstype(3)
         /method=reml
         /print=corb solution  r
          /random=intercept | subject(pid) covtype(ID)   .


** In this case (as most of the time) the variance components panel model tells much
** the same model as the flawed cross-sectional model.
** As a point of interest, a correct cross-sectional model (ie, only one record per person)
** also tells the same story, for instance : .

temp.
select if (wave=4).
regression var=ghqsad
            fem working unemp study age degdip
                   /statistcs=r coeff anova outs collin toll
              /dependent=ghqsad /method=enter .

** The notable difference being that the effect of unemployment is not confirmed as significant,
**   primarily because of a low number of cases in unemployment.

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




***********************************************
*** ii) Panel Format 2: Multiple time point records per case ('wide format'): .
************************************************************.

*** Example 2.2ii : BHPS panel data model for GHQ (happiness) evolution over time  .
* wave 4 .
get file=!path3+"dindresp.sav" /keep=pid djbstat dqfedhi dage dsex dhlghq1 .
* creates variables for use in harmonised file.
compute fem=(dsex=1).
```



```
compute working=(djbstat=1 | djbstat=2 | djbstat=5).
compute unemp=(djbstat=3 | djbstat=9).
compute study=(djbstat=7).
compute age=dage .
compute degdip=(dqfedhi ge 1 & dqfedhi le 5).
compute ghqsad94=dhlghq1 .
sort cases by pid.
compute wave=4.
sav out=!path9+"mtch4.sav" /keep=pid wave ghqsad94 fem working unemp study age degdip .
* wave 5 .
get file=!path3+"eindresp.sav" /keep=pid ehlghq1
  /rename (ehlghq1 = ghqsad95)  .
sort cases by pid.
sav out=!path9+"mtch5.sav" /keep=pid ghqsad95  .
* wave 6 .
get file=!path3+"findresp.sav" /keep=pid fhlghq1
  /rename (fhlghq1 = ghqsad96)  .
sort cases by pid.
sav out=!path9+"mtch6.sav" /keep=pid ghqsad96  .
* wave 7 .
get file=!path3+"gindresp.sav" /keep=pid ghlghq1
  /rename (ghlghq1 = ghqsad97)  .
sort cases by pid.
sav out=!path9+"mtch7.sav" /keep=pid ghqsad97  .
* wave 8 .
get file=!path3+"hindresp.sav" /keep=pid hhlghq1
  /rename (hhlghq1 = ghqsad98)  .
sort cases by pid.
sav out=!path9+"mtch8.sav" /keep=pid ghqsad98  .


** Add all these together :.

match files file=!path9+"mtch4.sav" /in=w94 /file=!path9+"mtch5.sav"
       /file=!path9+"mtch6.sav" /file=!path9+"mtch7.sav" /file=!path9+"mtch8.sav" /by=pid .
fre var=w94.
descriptives var=all.
select if (w94=1).
descriptives var=all.
** (Note - take a look at the data window to see the structure of this file with
**   a single record per person, but information from multiple time points on it ).

descriptives var=ghqsad94 ghqsad95 ghqsad96 ghqsad97 ghqsad98 .
missing values ghqsad94 ghqsad95 ghqsad96 ghqsad97 ghqsad98 (lo thru -1) .
descriptives var=ghqsad94 ghqsad95 ghqsad96 ghqsad97 ghqsad98 .
correlate var=ghqsad94 ghqsad95 ghqsad96 ghqsad97 ghqsad98 .
graph /scatterplot=ghqsad94 with ghqsad98.

descriptives var=ghqsad98 ghqsad94 fem age working unemp study degdip .
regression /dependent=ghqsad98
    /method=enter ghqsad94 fem age working unemp study degdip .



** Comment: Panel data in SPSS : SPSS can deal with various panel data formats
*   - they are still just rectangular files - but it doesn't have any inbuilt commands
*    specifically catering to its nuances - users have to write their own commands to
*    do this. This compares badly to STATA, which has several appropriate in-built commands.


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




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

```
************************************************************
*** 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).

import file=!path6+"2364a.por".
descriptives var=all.

* 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: .

fre var=highqual.
compute degdip=(highqual >= 1 & highqual <= 5).
variable label degdip "Has degree or diploma by age 23".
add value label degdip 0 "No degree/diploma" 1 "Has degree / diploma" .
fre var=degdip.

fre var=pasc0.
fre var=read7.
fre var=likes16.
missing values pasc0 (-1).
cro pasc0 by degdip /cells=count row /statistics=phi.
missing values read7 (-1).
correlate read7 degdip.
missing values likes16 (-1).
correlate likes16 degdip.

graph /bar= pct by likes16 by degdip
     /title="Dislikes school at age 16, by education at age 23".


* Model 1.
logistic regression var=degdip /method=enter pasc0
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).
* Model 2.
logistic regression var=degdip /method=enter pasc0 read7
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).
* Model 3.
logistic regression var=degdip /method=enter pasc0 read7 likes16
           /criteria pin(.05) pout(.10) iterate (20) cut (.5).
```



```
** 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 .


import file=!path3d+"ljempe.por".
descriptives var=all.
sort cases by pid date.
fre var=stemp.
list /variables=pid date stemp duration enddate /cases= from 1 to 30.

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


graph /histogram=duration.
* most events are short.
temp.
select if (stemp=1 | stemp=2 | stemp=3 | stemp=6 | stemp=7).
examine variables=duration by stemp /nototal /plot=boxplot.

* Interest often focusses upon changes in state between current and next state.
fre var=stemp.
fre var=nextemp.
temp.
select if (stemp > 0 & nextemp > 0).
cro stemp by nextemp .


****  Calculate 'real' dates.
descriptives var=date.
compute syear=trunc(date/12) + 1900 .
compute smonth=date - (trunc(date/12))*12.
list /variables=pid date syear smonth stemp duration /cases= from 1 to 30 .

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



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


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


get file=!path3b+"bh1to15_long.sav".
descriptives var=all.
fre var=year zvote zregion  .
missing values zjbcssm (-9 thru 0).
graph /histogram=zjbcssm.
graph /histogram=zfimn.
graph /histogram=zage.
select if (zregion >= 1 & zregion <= 18).
compute reg3=zregion.
recode reg3 (1 thru 16=1) (17=2) (18=3).
fre var=reg3.
fre var=zsex.
compute convot=(zvote=1).
fre var=convot .
fre var=zjbrgsc.
compute working=(zjbrgsc >= 1 & zjbrgsc <= 7).
fre var=working.

means tables=convot working zjbcssm zfimn by zsex by reg3.

sort cases by reg3.
split files by reg3.
means tables=convot working zjbcssm zfimn by zsex by year  /cells=mean.
split files off.
** Our data will be statistics on voting, working, job situation, income,
*  by year, gender and region

sort cases by year zsex reg3.
aggregate outfile=!path9+"agg1.sav" /break=year zsex reg3
    /convot=mean(convot) /working=mean(working)
    /zjbcssm=mean(zjbcssm) /zfimn=mean(zfimn).
get file=!path9+"agg1.sav".
descriptives var=all.

** This will be our dataset:.
variable label zsex "Gender" .
variable label year "Year" .
variable label reg3 "Region in Britain" .
variable label convot "Percentage conservative support".
variable label working "Percentage working" .
variable label  zjbcssm "Mean occupational advantage score of employed" .
variable label zfimn "Mean income of all adults" .
descriptives var=all.
list.
sav out=!path1+"bhps_time_series.sav".

** Time series data analysis

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

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



get file=!path1+"bhps_time_series.sav".

temp.
select if (reg3=1).
means tables=convot by year by zsex /cells=mean.

temp.
select if (reg3=1).
graph /bar=mean(convot) by year by zsex
  /title="Support for Conservatives in England" /subtitle="Source: BHPS 1991-2005".
```



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


get file=!path1+"bhps_time_series.sav".
descriptives var=all.
compute fem=(zsex=2).
compute wal=(reg3=2).
compute scot=(reg3=3).

** Basic regressions could be used: .
regression var=convot year /dependent=convot /method=enter .
regression var=convot year fem wal scot /dependent=convot /method=enter .
regression var=convot year fem wal scot zfimn zjbcssm  /dependent=convot  /method=enter .


** 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 cases by zsex reg3 year.
compute lagcv= -999.
if (zsex=lag(zsex) & reg3=lag(reg3)) lagcv=lag(convot).
missing values lagcv (-999).
descriptives var=convot lagcv.

regression var=convot year lagcv /dependent=convot /method=enter .

** The science of studying Time Series structures is well developed in economics.
**  For further training materials see, for example, http://www.bized.ac.uk/timeweb/ .



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


************************************************
************************************************
**** EOF .
```