```
*****************************************************.
***  Longitudinal Data Analysis for Social Science Researchers
**
**
** ESRC Researcher Development Initiative training programme:
**
**
**   Training materials lab 0:
**     INTRODUCTORY DATA ANALYSIS AND DATA MANAGEMENT IN SPSS  .
**
**
**      www.longitudinal.stir.ac.uk
**      Paul Lambert / Vernon Gayle, 26 August 2007
*****************************************************.




*********************************************************.
*****************************************************.
** The file below covers sixteen exercises:
**
***    Exercise 1:  Opening and processing SPSS
***    Exercise 2:  Handling SPSS data files and output
***    Exercise 3:  Using an SPSS Template file
***    Exercise 4:  Illustrating a variety of SPSS data analysis techniques
***    Exercise 5:  Getting data into SPSS
***    Exercise 6:  Univariate descriptive statistics
***    Exercise 7:  Core data management techniques
***    Exercise 8:  Bivariate descriptive techniques
***    Exercise 9:  Univariate inferential techniques on the GHS95 file
***    Exercise 10: Bivariate inferential techniques on the GHS95 file
***    Exercise 11: Multivariate inferential techniques on the GHS95 file
***    Exercise 12: Multivariate comparisons  on the GHS95 file
***    Exercise 13: Multivariate modelling techniques on the GHS95 file
***    Exercise 14: Multivariate modelling Extensions
***    Exercise 15: Using macros and include files
***    Exercise 16: Handling data: matching information between data files
***    Exercise 17: Selected issues in analysing complex survey data.
**
*********************************************************.




*****************************************************.
** GENERAL INSTRUCTIONS ON THESE FILES
**
** Work through this file in the SPSS syntax window, replicating
**  the SPSS do-file commands. Further help on working with SPSS is
**  available from the LDA web site.
**
**
*** This file assumes you have a number of files downloaded to your
**    machine. You will need the following:
*
** Data files downloadable from the LDA site :
* -ghs95.sav
* -cleandat.sav
* -wemp.dat
* -workhours.xls
* -itskills_entry.sps
*
** SPSS example macros ('include files') downloadable from the LDA site :
* -ghs95_prepare.sps
* -makedummyvars.sps
* -regressions.sps
* -seglabelsv1.sps
* -seglabelsv2.sps
* -varlabstonew.sps
*
** Data files 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)
*
* -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;
*      note that most of the exercise below does not use the BHPS data, so you
*      may prefer to proceed without downloading this study)
*
*
*************.
**
** These examples are written for SPSS version 14.0 for Windows (2006)
**  - most commands should be compatible with earlier versions of SPSS and other
**        operating systems
**
*****************************************************.

** .


*****************************************************.
** NOTIFICATION OF FILE LOCATIONS / DIRECTORIES
**
**
**
** i) File location declarations:
*** For the commands below to work, you should begin by running the following
**   macros, which tell SPSS where to look for the relevant data files (mentioned
**   above) on your machine : .
** NOTE: EDIT THE PATHS BELOW TO EQUIVALENTS APPROPRIATE TO 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 BHPS data
*   file(s) mentioned above) .


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



define !path8 () 'd:\lda\macros\' !enddefine.
* (the location of your copies of the demonstration SPSS macros (*.sps files)
*   downloadable from the website ).

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

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




*********************************************************.
*********************************************************.
*** SOME COMMENTS ON USING SPSS INTERACTIVE SYNTAX FILES  : .
*      (see also exercise 3).
*
*
*
* Comment: This is an interactive command file ('syntax file'), you can run
*          it from the syntax editor window in SPSS,
```

```
*          by highlighting the command line or lines which you wish to invoke
*          and either 'ctrl-r' or the 'run current' icon
*
* Comment: On an interactive command file, you write out the textual SPSS commands
*          and invoke them, typically in groups of a few sequential commands at a time.
*
* Comment: Gradually learning the textual SPSS commands is the valuable skill in learning
*          SPSS. The best way is to begin working through example files such as this one.
*          In time you will start to remember a few common commands, but additionally, you
*          should learn how and where to find information on the necessary SPSS commands
*          - i.e. from the SPSS manuals and help system, the numerous online SPSS resources,
*            or by using the 'paste' command from the GUI interface
*
* Comment: Notes on the example files below give some explanation of what each command is
*          doing, but they don't cover everything. Best is to try to work out yourself
*          what each line is doing, but ask the instructors for clarification if needed.
*
* Comment: On an interactive command file, a new command is indicated by a new line after
*          the last full stop (ie full stops are the 'command line delimiter').
*          BEWARE THAT FORGETTING FULL STOPS IS THE MOST COMMON SYNTAX PROGRAMMING MISTAKE!!.
*          Comments are added by begining the line with a *, which means that SPSS will ignore
*          the rest of the line - comments are useful to use to make notes for yourself.
*          Any new command, including a comment, is interpreted by SPSS as unfinished until
*          the next full stop is reached - beware that it is easy to inadvertently comment out
*          an active command by forgetting to close a previous comment with a full stop.
*
*          [advanced: if you run the whole file at once, as a 'batch file', it is also possible
*          to specify other symbols or line breaks as the command line delimiter]
*
*
* Comment: If running a group of command lines in a go in SPSS, if any of your
*          command lines include an error, SPSS will immediately move on to the next command
*          line and run from that point onwards (this is different from the way Stata
*          processes command lines). Beware that this can sometimes mean a group of lines
*          run and give the appearance of being complete, when actually an earlier error
*          took place
*
* Comment: SPSS processes command syntax by running the syntax on whatever current datafile
*          it has in memory, then displaying the outut generated by the relevant commands
*          in a third window known as the 'Output window'. To get the most out of an SPSS
*          session it IS STRONGLY RECOMMENDED THAT YOU CHANGE THE SETTINGS ON YOU OUTPUT
*          WINDOW FROM THE SPSS DEFAULT TO SETTINGS WHICH REVEAL MORE ABOUT THE UNDERLYING
*          COMMAND PROCESSING - SEE THE LDA WEBSITE FOR INSTRUCTIONS ON CHANGING OUTPUT FILE
*          SETTINGS.
*
*
* Comment: For most social scientists, processing SPSS through the syntax file editor is
*          appropriate for most purposes. The next step up is learning to program in SPSS,
*          which concerns writing programmes (macros) and batch files (whole files
*          processed in one go), to run complex tasks
*
*
*   SEE ALSO : http://www.longitudinal.stir.ac.uk/SPSS_support.html
*
*
*
*
************************************************************
************************************************************


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




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


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



```
*****************************************************************.
****** EXERCISE 1 - OPENING AND PROCESSING SPSS .
*****************************************************************.


get file=!path2+"ghs95.sav".
* (Opens the file from the relevant location).
fre var=typaccm .
* (Runs a frequency table for that file).



***** Extension : looking out tor a common mistake :.

***** The first fre command below will work; the second won't.
** (1) Show a frequency distribution for the variable sex.
fre var=sex.
** (2) Show another frequency distribition for a different variable:
fre var=genhlth.
** Q: What's the problem in (2)?.
** A: It lies with the full stop character. You need to end every command AND every
**   comment with a full stop. If not, SPSS thinks you're still on the command
**  that began in the previous line. So, in (2), SPSS thinks that the line 'fre var=sex.' is
** actually the second half of the comment, and ignores it.

** ** (2) corrected : Show the same frequency distribution again  .
fre var=genhlth.


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




*****************************************************************.
****** EXERCISE 2 - HANDLING SPSS DATA FILES AND OUTPUT .
*****************************************************************.

***************************************.
*** Segment 2.1: opening a data file and running some example analyses.

get file=!path2+"ghs95.sav".
* (Opens the SPSS format data file ghs95.sav).

fre var=typaccm.
* (Produces a frequency table of the values of the categories of the variable 'typaccm').

descriptives var=workhrs.
* (Gives some summary statistics on the values in the variable 'workhrs').

graph /histogram=workhrs.
* (Produces a 'histogram' showing the distribution of the metric values of 'workhrs').

examine variables=workhrs by sex /plot=boxplot  /statistics=extreme .
* (among other things, this produces a 'box and whisker plot' for the distribution of
*  the metric values of 'workhrs' by the categories of 'sex').

cro tables=hohscle by genhlth /cells=count row  .
* (a bivariate 'crosstabulation' of two categorical variables).


***************************************.
*** Segment 2.2: options for saving data.

get file=!path2+"ghs95.sav".

sav out=!path9+"ghs95_v2.sav".
sav out=!path9+"ghs95_v3.sav"
    /keep=typaccm edlev earnings soclase ecstaa centheat to washmach .

get file=!path9+"ghs95_v2.sav".
```

```
descriptives var=all.
* (just an exact copy of the original data).
get file=!path9+"ghs95_v3.sav".
descriptives var=all.
* (only includes a selection of variables).

get file=!path2+"ghs95.sav" /keep= typaccm edlev earnings soclase ecstaa .
descriptives var=all.
* (the 'keep' command can also be used to select variables).

export out=!path9+"ghs95_v4.por" .
import file=!path9+"ghs95_v4.por".
descriptives var=all.
* (these two commands save out, then open, a file in 'SPSS portable' format.

get file=!path2+"ghs95.sav" /keep= typaccm edlev earnings soclase ecstaa .
descriptives var=all.
write out=!path9+"ghs95_v5.dat"
         / typaccm ' ' edlev ' ' earnings ' '  soclase ' ' ecstaa .
execute.
* 'write out' exports the data in a 'plain text' format - open the file ghs95_v5.dat in
*  a plain text editor like 'Notepad' to see what this looks like .
* Note that the 'write out' command needs a few subcommands like the ' symbols, and
*  it needs to be followed by the 'execute' command.

data list file=!path9+"ghs95_v5.dat" free
      / typaccm edlev earnings class ecstatus .
descriptives var=all.
fre var=typaccm class .
* This command reads in the plain text data.
* Note that SPSS formatting like 'value labels' has been lost.
* Note how you're free to change the variable names during input should you wish to.

** Comment: there are a few other options for saving data, for instance alternatives in
*  the 'save translate' command allow SPSS data to be exported in certain other formats.

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

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




*****************************************************************.
******* EXERCISE 3 - USING AN SPSS TEMPLATE FILE.
*****************************************************************.



***************************************.
*** Segment 3.1 : .
*** Template: open a data file and run some example analyses (as in part 2).

get file=!path2+"ghs95.sav".

fre var=typaccm sex.
descriptives var=workhrs.
graph /histogram=workhrs.
examine variables=workhrs by sex /nototal /plot=boxplot  /statistics=descriptives extreme .
cro tables=hohscle by genhlth /cells=count row  .

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

**** Segment 3.2 : .
** Paste the lines above in the space below, and change the variable names.







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



```
***** Segment 3.3 : .
**** Paste the lines above in the space below, and change the file to 'cleandat.sav'
**    as well as (necessarily) the variable names.



*** Example answer to segment 3.3 :  .

get file=!path2+"cleandat.sav".

fre var=var00002 var00003 .
descriptives var=var00004.
graph /histogram=var00004.
examine variables=var00004 by var00002 /nototal /plot=boxplot /statistics=descriptives extreme  .
cro tables=var00015 by var00003 /cells=count row  .

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









*****************************************************************.
******* EXERCISE 4 - ILLUSTRATING A VARIETY OF SPSS DATA ANALYSIS TECHNIQUES .
*****************************************************************.



** (Preview - some univariate and bivariate techniques only - these
*             are also covered in later exercises).

** All of these examples use the example file from the GHS 1995 .

get file=!path2+"ghs95.sav".
descriptives var=all.

*********************************************************.
** Technique by : number of variables; level of measurement of variables; purpose of analysis .
***********************************.
** One variable, categorical, descriptive :.
fre var=typaccm.
graph /bar=typaccm.
************************************.
** One variable, metric, descriptive :.
graph /histogram=workhrs .
examine variables=workhrs /plot=boxplot  /statistics=extreme .
************************************.
** One variable, metric, inferential :.
examine variables=workhrs /plot=none /cinterval 95 .
************************************.
** Two variables, both categorical, descriptive  :.
cro tables=soclase by genhlth /cells=count row.
graph /bar=soclase by genhlth .
************************************.
** Two variables, both categorical, inferential  :.
fre var=soclase genhlth.
missing values soclase (-9,7).
cro tables=soclase by genhlth /cells=count row /statistics=chi phi .
************************************.
** Two variables, both metric, descriptive  :.
graph /scatterplot=workhrs with earnings.
compute lnearn=-999.
if (earnings ge 10 & earnings le 5000) lnearn=ln(earnings).
missing values lnearn (-999).
graph /scatterplot=workhrs with lnearn.
************************************.
** Two variables, both metric, inferential  :.
correlate var=workhrs lnearn .
************************************.
** Two variables, one metric, one categorical, descriptive  :.
examine variables=lnearn by soclase /nototal /plot=boxplot .
************************************.
** Two variables, one metric, one categorical, inferential  :.
```

```
means tables=lnearn by soclase /statistics=anova.
examine variables=lnearn by soclase /nototal /plot=boxplot .
graph /errorbar(CI 99)=lnearn by soclase .
graph /errorbar(CI 99)=lnearn by region .

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




*****************************************************************.
******* EXERCISE 5  - GETTING DATA INTO SPSS  .
*****************************************************************.



*** Segment 5.0 : adding data file information :.

** Clear your SPSS session.
new file.


** Tap in the 4 by 4 table of data below into the blank data editor grid.
*
1       17      1.73    A
1       18      1.85    B
2       17      1.60    C
2       18      1.69    A
*.

* Then run: .

rename variables (var00001=sex).
rename variables (var00002 var00003 var00004= age height grade).
variable labels sex "Student's gender" .
variable labels age "Age in years" .
variable labels  height "Height in metres" /grade "Higher Grade English result" .
add value labels sex 1 "Male" 2 "Female" .
fre var=all.
sav out=!path1+"varbycase1.sav".

* comment : on the rename variables and variable labels commands, we've shown some
* commands which are just one variable at a time, but others which do several variables at once.

************************************************.
**** (5.1) Opening a pre-formatted SPSS file :.

get file=!path2+"ghs95.sav".
descriptives var=all.

get file=!path2+"ghs95.sav" /keep=npersons typaccm .
descriptives var=all.

************************************************.
**** (5.2) Opening an SPSS portable file :.
get file=!path2+"ghs95.sav".
export out=!path9+"ghs95_port.por".

import file=!path9+"ghs95_port.por".
descriptives var=all.

************************************************.
*** (5.3)  Reading a plain text file :.

data list file=!path2+"wemp.dat" free
      / case femp mune time und1 und5 age .

* (This is a synthetic panel dataset where each record gives the
*   current employment status of husband-wife couples at different years).

descriptives var=all.
* (note: no variable or value labels - they need to be added).
```



```
variable labels case "Individual identifier"
   /femp "Wife's employment status"
   /mune "Husband's employment status"
   /time "Calendar time (year - 1975)"
   /und1 "Any children aged less than 1 year"
   /und5 "Any children age less than 5 years"
   /age "Wife's age"   .
add value labels
     femp 1 "Employed" 0 "Unemployed"
     /mune 1 "Unemployed" 0 "Employed"
     /und1 und5 1 "Yes" 0 "No" .
descriptives var=all.
fre var=femp mune und1 und5 .


** .


************************************************.
*** (5.4)  Importing from Excel :.

** Option 2  : reading in direct from an MS Excel file.
get data /type=xls /file=!path2+"workhours.xls"
          /sheet=name 'workhours data' /cellrange=full
          /readnames=on .
descriptives var=all.

* (note that the readnames subcommand asks the top row of data to be read
* as SPSS variable names - in SPSS versions earlier than 12, the names will be truncated
*  to a maximum width of 8 digits ).


************************************************.
*** (5.5)  Reading in data through a syntax command file :.

** Either, open the syntax file 'itskills_entry.sps' and run its commands.

** Or, if you're feeling clever - use the 'include' syntax command below to run
** the whole file as a 'batch' job.

include file=!path2+"itskills_entry.sps".

descriptives var=all.

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




*****************************************************************.
******* EXERCISE 6 - UNIVARIATE DESCRIPTIVE STATISTICS
*****************************************************************.



*****************************************************.
*** Segment 6.1) Open the file and look over it's contents.
*****************************************************.

get file=!path2+"ghs95.sav".

** Two similar SPSS univariate summary functions :.
descriptives var=all.
summarize var=all.
** These summary functions are best suited to continuous (metric)
** data, but they are also convenient ways to get to a quick review of the
**  contents of a data file.
* Note the use of 'all' to ask SPSS to run the analysis on all variables; we could alternatively
*  have put some specific variable names in.

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



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

```
*** Segment 6.2) Summarizing single categorical variables - nominal level .
*****************************************************.

get file=!path2+"ghs95.sav".

**********************************.
** Seg6.2(i) Nominal level statistics  :.

** The main way of seeing a distribution is through the 'frequencies' command (fre).
fre var=typaccm.
fre var=typaccm /format=dfreq .
fre var=typaccm /statistics=min max range mode  .

** It is often handy to run univariate frequencies on more than one variable at the same time : .
fre var=typaccm bedstndb hhtypf1 /statistics=min max range mode  .


** Another way of seeing a categorical distribution is through the tables command.
*   Although the 'tables' command ultimately gives better displays, it's syntax is more complicated
.
tables /table=typaccm .
tables /format blank missing ('.') /ftotal=ftot1 "Total"
  /tables (labels) + ftot1 by  (typaccm)
  /statistics count ((F5.0) ' Cases ').
* (we'll cover the 'tables' command a little bit more in the part3 syntax).


**********************************.
*** Seg6.2(ii) Nominal level graphics :.

** For most purposes, the graphs available under the 'graph' command are adequate.

graph /bar=count by typaccm /title="Accommodation type for GHS 1995 adults" .
graph /bar=pct by typaccm /title="Accommodation type for GHS 1995 adults" .

graph /pie=count by typaccm /title="Accommodation type for GHS 1995 adults" .
graph /pie=pct by typaccm /title="Accommodation type for GHS 1995 adults" .

*****.
*** Comment : Quite a few manipulations to the initial appearance of SPSS graphics
***   are possible after the initial generation - to perform these, you open up the graph editor
***   and change various options.
*****.


** Several graphs are also available as options under the 'frequencies' command, eg :.
fre var=typaccm /barchart=freq /pie=freq .


** More complicated graphics option can be invoked with the 'igraph' command -
*    here's an example, though we won't look into them in depth here.
igraph /viewname="Simple Pie Chart" /summaryvar= $count  /style=var(typaccm)
  /title="Type of accommodation, GHS 1995" /subtitle="Nationally representative sample: adults in
UK"
        /caption="Pie chart of accomodation types" /x1length=5.0 /ylength=3.0 /x2length=3.0
          /chartlook="none" /catorder var(typaccm) (ascending counts omitempty)
         /pie key=on start 90 cw  slice=numin label pct n .



**********************************.
** Seg6.2(iii) Ordinal level variable - statistics  :.

* The commands are much the same, except that the median and percentiles are more relevant .
fre var=hohscle .
fre var=hohscle  /percentiles=1 5 10 50 90 95 99 /statistics=min max range mode median sum.
tables /table=hohscle  .

* Threshold values can also be important in ordinal data -
*    see an example of using them in the part 2 syntax.


**********************************.
** Seg6.2(iv) Ordinal level variable - Graphics  :.

* There's no significant departure from nominal level data, eg : .
```



```
graph /bar=count by hohscle.
fre var=hohscle /barchart=freq /pie=freq .

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


*****************************************************.
*** Segment 6.3) Summarizing single metric level variables .
*****************************************************.


**********************************.
*** Seg6.3(i) Statistical summaries of a metric variable distribution :.

descriptives var=workhrs  .
descriptives var=workhrs /statistics=all.
summarize var=workhrs /cells=all.

** Comment: if you're unsure what a particular statistic means, one helpful SPSS function is
**  to open up the output table (right click then 'spss pivot table option' then open), then
**  right-click the mouse on the text of the relevant statistic name, and click 'what's this'.

** The 'descriptives' and 'summarize' commands can review more than one variable at a time:.
descriptives var=all.
summarize var=workhrs hohx earnings .


**Another way of getting some of these and other statistics:.
fre var=workhrs /format=notable /percentiles=5 10 50 90 95 /statistics=all.
fre var=workhrs /format=notable /percentiles=5 10 90 95 /ntiles=4 /statistics=all.


** The examine variables command gives many of these statistics
* (as well as graphical displays which in the first instance are suppressed below):.

examine variables=workhrs earnings  /plot=none /percentiles(1 5 10 25 50 75 90 95 99)
         /statistics descriptives extreme(4) .

* Comment: the 'extreme' table suggests that the data collection used a 'cropping' upper limit of
*  97 hours for this variable.


***************************.
** Seg6.3(ii) Graphical summaries :.


** Several options can generate much the same output :.
graph /histogram=workhrs .
graph /line=workhrs.
examine variables=workhrs /plot=boxplot stemleaf histogram  /statistics=all extreme .
fre var=workhrs /format=notable /histogram=freq  .

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




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




************************************************************.
****** EXERCISE 7 - CORE DATA MANAGEMENT TECHNIQUES  .
************************************************************.


*****************************************************.
*** Segment 7.1)  Open the GHS file and look at the distributions of a handful of variables .
```

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

get file=!path2+"ghs95.sav".

fre var=typaccm nadults origin edlev  .
descriptives var=nadults age  earnings .
graph /histogram= age.
graph /histogram=edlev .

** These are a selection of 'raw' variables - below we'll look at how we might adjust them
**  for the purposes of analysis.



*****************************************************.
*** Segment 7.2)  Variable managment on categorical variables .
*****************************************************.


get file=!path2+"ghs95.sav".

**********************************.
** Seg 7.2(i): 'Type of accomodation' variable.

fre var=typaccm /statistics=min max range mode  .
* Of the 10 different categories, three of them are big but 4 of them are quite small.
* A typical distributionally influenced decision would be to concentrate only on the
*  3 largest categories and merge the fourth and fifth (which appear similar).
** We do this by recoding the data but on a new variable (so as not to overwrite the original
**    information.

compute typaccm2=typaccm.
recode typaccm2 (1=1) (2=2) (3=3) (4,5=4) (else=-999).
add value labels typaccm2 1 "Detached house" 2 "Semi-detached house"
           3 "Terraced house" 4 "Flat" -999 "Exclude".
* (It is a convention to code values that we wish to exclude to high negative values).
variable label typaccm2 "Type of accomodation (reduced form)".
missing values typaccm2 (-999).
* (This command (see segment 3 below) tells SPSS that in most analyses, values
*    of -999 on this variable are to be ignored from processing).

fre var=typaccm2.
graph /pie=pct by typaccm2 .

** Comment: under this treatment, we'd ignore some cases from analysis just because
*    they came from a relatively rare category.
*    A common alternative is to use a residual 'other' category for small categories - we
*     can do that quickly by repeating the above but with a slight adjustment :.

compute typaccm3=typaccm.
recode typaccm3 (1=1) (2=2) (3=3) (4,5=4) (6,7,8,9,10=5) (else=-999).
add value labels typaccm3 1 "Detached house" 2 "Semi-detached house"
           3 "Terraced house" 4 "Flat" 5 "Other" -999 "Exclude".
variable label typaccm3 "Type of accomodation (reduced form)".
missing values typaccm3 (-999).

fre var=typaccm3.
graph /pie=pct by typaccm3 .



**********************************.
** Seg 7.2(ii): 'Ethnic origin' variable.

fre var=origin /statistics=min max range mode  .

* This distribution is a classic problem for sociologist - in nationally representative samples,
*  relatively few sample members are likely to identify with a minority ethnic group, and
*  although we're very interested in the differences between finer points of ethnic identity,
*  the sample data is unlikely to sustain such analyses.
*  Most researchers make a compromise, somewhere between the full sample data and no
*   information at all. Below a selection of commonly used possibilities is shown.
*  Note that best practice in these circumstances is to use a recoding that has some
*   justification in previous literature - eg one that was used in an earlier relevant application.

compute origin2=origin.
compute origin3=origin.
```



```
compute origin4=origin.
recode origin2 (1=1) (2,3,4,5,6,7,8=2) (else=-999).
recode origin3 (1=1) (2,3=2) (4,5,6=3) (7,8=4) (else=-999).
recode origin4 (1=1) (2=2) (4=3) (3,5,6,7,8=4) (else=-999).
add value labels origin2 1 "White" 2 "Non-white" .
add value labels origin3 1 "White" 2 "Black" 3 "South Asian" 4 "Other" .
add value labels origin4 1 "White" 2 "Black-Caribbean" 3 "Indian" 4 "Other"  .
missing values origin2 origin3 origin4 (-999).
fre var=origin2 origin3 origin4.


**********************************.
** Seg 7.2(iii): 'Education level' variable.

fre var=edlev .
** This variable has a lot of information, but simply has too many categories to be
*    easily analysed in some circumstances.
*   A typical treatment is to make a substantive decision about thresholds between different
*    categories, and derive a putatively ordinal revised measure.
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
fre var=edlev2 .



**********************************.
** Seg 7.2(iv): 'Dummy coding' of categorical data .

** For the purposes of certain analytical methods, it is often desirable to code categorical data
*   into dichotomous 'dummy variable' indicators. The values involved should normally be set
*   to 0 and 1, since this is the most convenient for many different purposes. For example : .

fre var=edlev2.
** These three categories could also be represented by three different dichotomous variables :.
compute hied=(edlev2=1).
compute meded=(edlev2=2).
compute loed=(edlev2=3).
fre var=edlev2 hied meded loed.

** Actually, dummy variables can be created in a few different ways.
*    The technique below is easier to follow, though requires more typing per variable .
compute hied2=edlev2.
recode hied2 (1=1) (2,3=0).
compute meded2=edlev2.
recode meded2 (2=1) (1,3=0).
compute loed2=edlev2.
recode loed2 (3=1) (1,2=0).
fre var=edlev2 hied meded loed hied2 meded2 loed2 .

** Comment: beware of the role of missing / other categories when dummy coding .


** It is often handy to run univariate frequencies on more than one variable at the same time : .
fre var=typaccm bedstndb hhtypf1 /statistics=min max range mode  .


*****************************************************.
*** Segment 7.3)  Variable managment on metric variables .
*****************************************************.

get file=!path2+"ghs95.sav".

**********************************.
** Seg 8.3(i): 'Earnings' variable.

descriptives var=earnings /statistics=all .
graph /histogram(normal)=earnings.
examine variables=earnings  /plot=boxplot histogram /percentiles(1 5 10 25 50 75 90 95 99)
         /statistics descriptives extreme(10) .

* This is a very typical 'raw' income distribution - it has a very wide spread (large range
*  and standard deviation), yet most cases are concentrated within a much smaller range of values.
* Also, the 'extremes' table shows that it is only a few cases at the ends of the distribution
```

```
*  which are responsible for a lot of the spread.

** One typical reaction is to 'crop' the analysis and only consider incomes in a certain
**  (substantively justified) range. See below: .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
** (this is an example of a 'conditional' recoding. 'le' means 'less than or equal to', and 'gt'
**   means 'greater than').
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.


** Another option is to make a linear transformation of the data in such a way that de-emphasises
** extremely high values - a logarithmic distribution can do this. See below : .
compute earn3=-999.
if (earnings gt 1) earn3=ln(earnings).
* This computes earn3 to be the natural log of the 'earnings' value. We only do this for
*  values of earnings greater than 1, as log transformations don't work for negative values or
*   values up to 1.
missing values earn3 (-999).
graph /histogram(normal)=earn3.

* (You could also combine the two, eg logged income only if over 50gbp per week).

* A strength of a logarithmic distribution is that it often makes for a much nicer 'shape'
*   to the histogram. A weakness though is that your values are now in 'logs'.
* For a note on such transformations, see Buckingham and Saunders 2004 p188-9.



**********************************.
** Seg 7.3(ii): 'Age' variable.

graph /histogram=age.

** Two treatments of age are common : .


** (1) Categorise it, if relevant, around socially significant values for the application you're
*        working on.

** Eg, say we were interested in car drivers, insurance companies typically use the following:.

compute age2=age.
recode age2 (17 thru 21=1) (22 thru 28=2) (29 thru 55=3) (56 thru 65=4) (66 thru hi=5) (else=-999).
add value labels age2 1 "Highest risk (17-21)" 2 "Medium risk (22-28)"
             3 "Lowest risk (29-55)" 4 "Modest risk (56-70)" 5 "High risk (over 70)".
fre var=age2.


** (2) Transform it by a linear expression.

** Say we wanted to make differences at higher ages more important :.

compute agesq1=-999.
if (age ge 16) agesq1=age**2.
* (the '**2' symbol means 'squared').
graph /scatterplot=age with agesq1.
* (we'll introduce the scatterplot command in part 3).

** A slightly cleverer transformation might give greater influence to both the oldest and youngest
:.
compute agetemp=age - 50.
compute agesq2=-999.
if (age ge 16) agesq2=agetemp**2.
graph /scatterplot=age with agesq2.


**********************************.
** Seg 7.3(iii): 'nadults' variable.

fre var=nadults.
** This is an interesting example of an interval level variable: it is a count measure, with a
**  very limited range of different categories.

graph /histogram(normal)=nadults.
```



```
descriptives var=nadults /statistics=mean stddev skew kurtosis .
** The distribution is a bit of a way from being 'normal'.

** Although this data could be treated as metric, it is more typical to 'downgrade' it
**  a level to ordinal, eg :.

compute nadults2=nadults.
recode nadults2 (1=1) (2=2) (3=3) (4=4) (5 thru hi=5) (else=-999).
add value labels nadults2 1 "1" 2 "2" 3 "3" 4 "4" 5 "5 of more" .
variable label nadults2 "Number of adults in household" .
missing values nadults2 (-999).
fre var=nadults2.


*********************************************************.
*** Segment 7.4)  Case selection: dealing with missing data .
*********************************************************.

** We have already made several uses of the 'missing data' command above.
** Missing data occurs when information on a particular variable for a particular
** case is not available, either because it was not collected successfully (missing)
** or perhaps because it is not relevant to the analysis here ('inapplicable').

** Conventionally, missing and inapplicable cases are coded with explicit values,
** usually negative numbers such as -999, -9,-7, -1, which remind an analyst that
** these cases are special cases.

** SPSS has the capacity to treat specified values as 'missing' for the purposes of analysis;
**  however some of the techniques often cause confusion or complications.

** Some examples :.

get file=!path2+"ghs95.sav".

**********************************.
***** Seg 7.4(i): Using 'missing' to treat certain numeric categories as missing.

fre var=soclase .
* In the predefined dataset, the value -9 has already been declared as missing, and is ignored.
* We can add the value 7 also :.
missing values soclase (-9,7).
fre var=soclase.
* We can also turn the missing value treatment from all categories :.
missing values soclase ().
fre var=soclase.
* Finally go back to the original :.
missing values soclase (-9).
fre var=soclase.
** In fact, most of the pre-defined variables in this example file have some missing categories
declared.
descriptives var=all.
** The 'N' column in the output shows the number of non-missing cases on each variable.
*  The total 'N' (should equal 3) shows that only {3} cases have non-missing data on every variable.


******************************.
***** Seg7.4(ii): SPSS definitions of 'user missing' and 'system missing' .

* 'user missing' = one or more numeric values treated as missing in analysis.
* 'system missing' = the relevant cell in the variable by case matrix doesn't have any numeric data.

* In fact, the GHS example dataset  only has user missing data; user missing data is usually
* safer, as it is easier to keep track of.

** System missing data can be generated, though, by defining new variables according to
**    variables  with missing data on them.

** For illustration:.
fre var=sex.
missing values sex (1).
compute sex2=sex*5.
fre var=sex sex2.
* note the different treatments of the male cases on the two variables.
* If you have a look at the right hand end of your dataset, in the 'data view' pane of the
* 'data editor' window, you'll see some examples of system missing cells (just full stops in them).

* Remember to change the original variable back to it's original form.
```

```
missing values sex ().
fre var=sex sex2.



*******************************************************.
*** Segment 7.5)  Case selection: conditional on values of other variables.
*******************************************************.

get file=!path2+"ghs95.sav".
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
variable label earn2 "Average weekly earnings (range 51-2000 only)".
missing values edlev2 earn2 (-999).


fre var=sex edlev2 .
examine variables=earn2  /plot=histogram  /statistics descriptives .

** In typical application, we might be interested only in looking at the average weekly
*    earnings of certain groups - eg men, or those with higher educational levels.

** There are several options .

** (i) Defining then using  'filter' variables.
compute males=(sex=1).
compute malhied=(sex=1 & edlev2=1).
fre var=sex males edlev2 malhied.
* A 'filter' variable _must_ be coded 1 for values you wish to use, and 0 otherwise.

examine variables=earn2  /plot=histogram  /statistics descriptives .
filter by males.
examine variables=earn2  /plot=histogram  /statistics descriptives .
filter off.
filter by malhied.
examine variables=earn2  /plot=histogram  /statistics descriptives .
filter off.

** (ii) Using 'split files'.
** The 'split files' command runs every process once for each of the 'splits' defined between
**   different cases).
** (Note an odd precursor: to run successfully, a 'split files' command has to be preceded
*    by a 'sorting' of the dataset in terms of the variables to be split by. Sorting involves
rearranging
*    the order of cases in the variable by case matrix (see session 5 lab)).


sort cases by sex edlev2.
split files by sex.
examine variables=earn2  /plot=histogram  /statistics descriptives .
split files by sex edlev2.
examine variables=earn2  /plot=histogram  /statistics descriptives .
split files off.
* (note that the split files command doesn't exclude user missing values on the split variables).


** (iii) Using the 'select if' option.

** This is the quickest immediate way - but it can be cumbersome to keep repeating.
** 'Select if' permanently deletes cases from the active file; usually we use a 'temp'
* command immediately before it, so that it applies to the next command only.

fre var=sex edlev2 .
examine variables=earn2  /plot=histogram  /statistics descriptives .

temp.
select if (sex=1).
examine variables=earn2  /plot=histogram  /statistics descriptives .

temp.
select if (sex=1 & edlev2=1).
```



```
examine variables=earn2  /plot=histogram  /statistics descriptives .


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




*******************************************************.
******* EXERCISE 8 - BIVARIATE DESCRIPTIVE TECHNIQUES .
*******************************************************.


*******************************************************.
** This file uses the example data file from the General Household Survey 1995
**   (available from the LDA WebCT site).
*****************************************.

*******************************************************.
*** Segment 8.1) Open the file and preliminary data management .
*******************************************************.


get file=!path2+"ghs95.sav".

descriptives var=all.

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in earlier example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".

fre var=edlev2 soclase .


summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.

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

*******************************************************.
*** Segment 8.2) Summarising Categorical to categorical relationships .
*******************************************************.



**********************************.
****** Seg8.2(i) Displaying the data values :.

** The most common technique is the 'crosstabs' command :.

cro edlev2 by sex.
cro edlev2 by sex /cells=count col .

cro edlev2 soclase by sex /cells=count col total .

cro edlev2 by soclase /cells=count col row .

** Summaries can also be made via a 'table' command :.
```

```
tables /format blank missing ('.') /ftotal=ftot1 "Total"
  /tables (soclase) + ftot1 by  (sex)
  /statistics count ((F5.0) ' Cases ').
* (There are many complex permutations to the tables command - there is a whole manual
*   describing there use! The reason people use 'tables' is because it offers more control
*   over output display ).


** Lastly, we could also use 'split files' commands to a similar end : .

sort cases by sex.
split files by sex.
fre var=edlev2 soclase.
split files off.



***********************************.
****** Seg8.2(ii) Graphically displaying the data distributions :.


** Clustered bar charts are the main approach : .

** In terms of the total number of cases :.
graph /bar=count by soclase by sex .
** In terms of  percentages - heights show the percentage of men and of women.
graph /bar=pct by soclase by sex .
** (Note the graphs show slightly different patterns: eg, more women in total are in social class II
**     than men; but _relatively_ more men are in it than women -  see the table %'s to confirm).


** Bar charts can also come in a 'stacked' format :.
graph /bar(stacked)=pct by soclase by sex .
graph /bar(stacked)=pct by soclase by edlev2 .


** Line graphs can also be used for categorical by categorical comparisons :.
graph /line(multiple)=pct by soclase by edlev2 .


***********************************.
****** Seg8.2(iii) Association statistics for categorical data :.

** Most statistics that you could need are available under the 'cro' command.


cro edlev2  by sex /cells=count col /statistics=phi .
cro edlev2  by soclase /cells=count col /statistics=phi .

** Note : Phi for nominal-by-nominal 2-by-2; Cramers V for all other nominal-by-nominal.

cro edlev2 soclase by sex /cells=count col /statistics=phi .

cro edlev2  by soclase /cells=count col /statistics=all .
** Remember : open up the output and right click on the text to get SPSS description of the
statistic.

** Comment: SPSS gives you far too many statistics.
** The main ones to remember for categorical data are :
**     Cramers V (or Phi) for nominal-by-nominal
**    Gamma for ordinal-by-ordinal (as the above table is).


*******************************************************.
*** Segment 8.3) Summarizing categorical to metric relationships .
*******************************************************.


***********************************.
****** Seg8.3(i) Looking at the data values :.

** Categorical - by - metric descriptions = tables of means.

fre var=edlev2.
descriptives var=earn2 workhrs  .
means tables=earn2 by edlev2 .
```



```
means tables=earn2 workhrs by edlev2 /cells=all  .

* Also the 'examine variables' command for the same things  .

examine variables=earn2 by edlev2 /nototal /percentiles(1 5 10 50 90 95 99)
        /plot=none  /statistics=descriptives extreme .


** Tables-of-mean statistics can also be computed by calculating univariate summaries under
* the split files command :.

sort cases by edlev2.
split files by edlev2.
descriptive var=earn2 workhrs /statistics=all .
fre var=earn2  /format=notable /percentiles=5 10 90 95 /ntiles=4 /statistics=all.
split files off.


***********************************.
****** Seg8.3(ii) Graphical images :.

** The most used option is probably the box and whisker plot  .

examine variables = earn2  by edlev2     /plot=boxplot .
examine variables = earn2  by edlev     /plot=boxplot .

* The box shows the median (central line) and the 25 and 75 percentiles (edge of box).


** Then there are errorbar (point) plots :.

graph /errorbar=earn2 by edlev2 .
graph /errorbar=earn2 by edlev.
* These show the arithmetic mean plus confidence intervals (see session 3).


** And there are bar charts :.

graph /bar=mean(earn2) by edlev2 .
graph /bar=mean(earn2) median(earn2) by edlev2 .
graph /bar=mean(earn2) median(earn2) by edlev .
* The height of the bars shows arithmetic mean of earnings; or median; other functions are
*   also available.


** There are also line charts - though only useful if you have a low number of different values.

fre var=tea.
missing values tea (-9,14).
graph /line(multiple)=count by tea by sex .
graph /line(multiple)=count by tea by soclase .


** And of course you can repeat univariate graphs :.

examine variables = earn2  by edlev2     /plot= stemleaf histogram .


** Finally, if you have a lot of categories, scatterplots can sometimes be used
*   (but not strictly appropriate).

fre var=edlev.
graph /scatterplot=edlev with earn2 .
* Although edlev is categorical, the scatterplot illustrates spread in different categories, and
* can also be used to pick out  outliers (box and whisker plots can be used also).



***********************************.
****** Seg8.3(iii) Categorical by metric assocation statistics


**** The most used statsitic is 'eta' .

means tables=earn2  by edlev2    /statistics=anova .
means tables=earn2 workhrs by edlev2 sex   /statistics=anova .
```

```
** Eg, education is weakly associated with working hours, but gender is strongly associated.


*********************************************************.
*** Segment 8.4) Summarizing metric to metric relationships .
*********************************************************.


***********************************.
****** Seg8.4(i) Graphical displays .

descriptives var=earn2 workhrs.


graph /scatterplot=workhrs with earn2  .
graph /scatterplot=age with earn2  .

* Note: scatterplots are easier to understand when they have fewer cases:.

temp.
select if (age ge 25 & age le 40 & sex=1).
graph /scatterplot=workhrs with earn2 /title="Hours of work by earnings, men age 25-40 only".

*  You can edit the scatterplot display by opening up the output window and
*    trying out different options.


** Multiple scatterplots can be displayed at once - though it's not always that helpful!.

graph /scatterplot(matrix)= workhrs earn2 age .

temp.
select if (sex=1 & edlev2=1).
graph /scatterplot(matrix)= workhrs earn2 age /title="Males with higher level education only" .



***********************************.
****** Seg8.4(ii) Association statistics for metric-metric relations .


* Correlations and regressions indicate strength of assocations
*  (correlations and regression results are equilvalent in the bivariate case).


correlate var=workhrs earn2 age .



regression var=workhrs earn2
     /dependent=earn2 /method=enter .
regression var=age earn2
     /dependent=earn2  /method=enter .

** Comment - there is a pattern of association, but it's weak - only explains a very small
**   proportion of the variance.


** Extension: quadratic function of age:  .
compute agesq=-999.
if (age ge 16) agesq=age**2.
missing values agesq (-999).
descriptives var=age agesq earn2 .
regression var=age agesq earn2
     /dependent=earn2  /method=enter .
** This makes sense - earnings rise then fall off again as age goes up.



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


*********************************************************.
*** Segment 8.5) Some extensions in bivariate analysis  .
*********************************************************.
```



```
***************************************.
*** Seg8.5(i) : An introductory illustration of case weighting .

fre var=sex.
** The sample split = males 48%, females 52%.
** But say we know a priori that the split should be 50/50.
** We could generate a new 'weight' variable which inflates the influence of men
** slightly, and deflates that of women - use the proportion (0.5) , the sample n (4633)
**  and the subsample ns (2218 / 2415)  .
compute sexwt=-999.
if (sex=1) sexwt=(0.5*4633)/2218 .
if (sex=2) sexwt=(0.5*4633)/2415 .
means tables=sexwt by sex.

fre var=sex  .
means tables=earn2 by soclase.
weight  by sexwt.
fre var=sex soclase .
means tables=earn2 by soclase .
weight off.
** Note how even very minor weights like this can have a small impact on other results .
** Properly derived sample weights typically take account of several demographic
**  factors at once, eg gender, age, ethnicity, .
** -> you should always at least try out weighting factors.

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


***********************************.
** Seg 8.5(ii): Tranforming between levels of measurement .

graph /histogram=tea.
fre var=tea.

** Note the value labels in the frequencies command: this is a categorical variable,
**  though it masquerades as an interval level variable!  .

** Data like this is usually best treated by categorising it :.
compute tea2=tea.
recode tea2 (1,2=1) (3,4=2) (5 thru hi=3) (else=-999).
add value labels tea2 1 "Left educ 15-16" 2 "Left educ 17-18" 3 "Left educ later" .
missing values tea2 (-999).
fre var=tea2.

** However, empirical experience will suggest that treating 'tea' as a metric variable
**  will not dramatically misrepresent it's information - so long as you control for the
**   category of current students (14).

** Eg :  .
compute tea3=tea.
recode tea3 (1,2=1) (3,4=2) (5 thru 13=3) (else=-999).
add value labels tea3 1 "Left educ 15-16" 2 "Left educ 17-18" 3 "Left educ later" .
missing values tea3 (-999).
compute tea4=tea.
missing values tea4 (-9,14).
fre var=tea tea2 tea3 tea4.

means tables=earn2 by tea2 tea3 /statistics=anova.
correlate var=earn2 tea tea4.

** The raw tea variable has a very different correlation to income than does the categorical
**   transformations - but the correlations from tea3 and tea4, ie categorical and metric, to
**  earnings are very close .


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




*************************************************************.
******* EXERCISE 9 - UNIVARIATE INFERENTIAL TECHNIQUES ON THE GHS95 FILE .
```

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



*******************************************************.
*** Segment 9.1) Open the ghs95 file and run some variable constructions.
*******************************************************.

get file=!path2+"ghs95.sav".

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in part 2 syntax example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".

fre var=edlev2 soclase .

summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.

** .


*******************************************************.
*** Segment 9.2) Some univariate metric inferential statistics and graphics  .
*******************************************************.


************************************.
****Seg9.2(i):  Confidence intervals around the mean .

descriptives var=workhrs.
graph /histogram=workhrs.

** 95% Confidence intervals around the mean for all the data :.
examine variables=workhrs   /plot=histogram  /percentiles(1 5 10 25 50 75 90 95 99)
        /statistics descriptives extreme(4) .
examine variables=workhrs   /plot=none       /statistics descriptives .

summarize var=workhrs /cells=all.
* (Summarize doesn't give you the CI's, but it does calculate the standard error,
*    from which CI's may be deduced ).

descriptives var=workhrs /statistics=all.
* ('descriptives' doesn't give you the CI's or the standard error, but it does
*  give you the standard deviation and n, from which the standard error then
*   in turn the CI's, can be deduced).


** 99% Confidence intervals :.
examine variables=workhrs   /plot=none
        /statistics descriptives /cinterval 99 .
** 90% Confidence intervals :.
examine variables=workhrs   /plot=none
        /statistics descriptives /cinterval 90 .


** Illustration : how CI's vary by sample size and level.
compute ag1630=(age ge 16 & age le 30) .
compute ag2830=(age ge 28 & age le 30).
compute ag30m=(age=30 & sex=1).
fre var=ag1630 ag2830 ag30m.

examine variables=workhrs   /plot=none /statistics descriptives /cinterval 95 .
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 99 .
```



```
filter by ag1630.
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 95 .
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 99 .
filter by ag2830.
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 95 .
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 99 .
filter by ag30m.
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 95 .
examine variables=workhrs   /plot=none /statistics descriptives /cinterval 99 .
filter off.
* (output from this analysis is in lecture 3a, slide 30).


************************************.
****Seg9.2(ii):  Graphics for confidence intervals  .


graph /errorbar(CI 95)=workhrs /title="Average hours per week, all GHS adults (n=2571)".
filter by ag30m.
graph /errorbar(CI 95)=workhrs  /title="Average hours per week, GHS males aged 30 (n=30)".
filter off.
* Comment: errorbars don't do much for a single variable.
*  They're usually better for comparing means from more than one group .
* To change the scale on the y-axis, open up the chart object in SPSS and
*   double click on the y-axis line, click on the 'scale' tab, and then uncheck the
*    range (min and max) boxes and change the values to a wider range.

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



*******************************************************.
*** Segment 9.3) Some categorical inferential statistics and graphics  .
*******************************************************.


*** Nominal level : confidence interval for a proportion :.

fre var=edlev2 .
** SPSS doesn't have a computation function for confidence intervals for a proportion.
** Need to do this manually or with MS excel :.
* Formula : see Blaikie 2003:p173; or de Vaus 2002:p232.
* Excel calculator: see near top of : http://staff.stir.ac.uk/paul.lambert/downloads.html .

**** N = 3611 .
**** Proportions: High = 9.8; Intermediate=30.3; Low=37.9 .
**  SE High = sqrt(0.098*(1-0.098) / 3611 ) = 0.0049 = 0.5% .
** => 95% CI high = 8.8 - 10.8%  .
**  SE Intermediate = sqrt(0.303*(1-0.303) / 3611 ) = 0.0076  = 0.8% .
** => 95% CI high = 28.7 - 31. 9 %  .
**  SE High = sqrt(0.379*(1-0.379) / 3611 ) = 0.0081 =  0.8% .
** => 95% CI high = 36.3 - 39.5 %  .


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



************************************************************.
******* EXERCISE 10 - BIVARIATE INFERENTIAL TECHNIQUES ON THE GHS95 FILE .
************************************************************.



*******************************************************.
*** Segment 10.1) Open the file and prepare a few variables.
*******************************************************.

get file=!path2+"ghs95.sav".

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in part 2 syntax example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
```

```
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".

fre var=edlev2 soclase .

summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.

** .

*******************************************************.
*** Segment 10.2) Cross-tabulations: Chi-squared inferential statistics .
*******************************************************.


**************************.
** Seg10.2(i) Pensions by gender for young adults .

compute ag1630=(age ge 16 & age le 30).
fre var=ag1630.
filter by ag1630.
fre var=sex perspens .
cro sex by perspens /cells=count row.
cro sex by perspens /cells=count row /statistics=chisq phi .
filter off.



**************************.
** Seg10.2(ii) Gender by consulting a doctor in last 2 weeks and age .

compute age2=age.
recode age2 (16 thru 30=1) (31 thru 50=2) (else=-999).
add value labels age2 1 "16-30 years" 2 "31-50 years".
missing values age2 (-999).
fre var=sex age2 doctalk .

** Generate what's probably the most sensible table of these three :.

cro doctalk by sex by age2  /cells=count row /statistics=chisq phi.

** Some further permutations for different bivariate significance statistics :.
cro sex by age2 by doctalk  /cells=count row /statistics=chisq phi.
cro age2 by doctalk by sex  /cells=count row /statistics=chisq phi.


**************************.
** Seg10.2(iii) Ordinal data illustration :.

fre var=soclase edlev2.
* both of these variables might be considered ordinal .
* We still do a crosstab, but now we might also use an ordinal data statistic
*   such as 'gamma'.
cro soclase by edlev2   /cells=count row /statistics=chisq phi gamma .
** Gamma is bigger than the Cramer's V - there is _more_ pattern if you look at orders.
** The p-values here are much the same;  it is _unusual_ for a nominal cf ordinal  categorical
**  association statistics to have different significance levels - but it's not impossible.

compute ag35=(age=35).
compute ag35m=(age=35 & sex=1).
filter by ag35.
cro soclase by edlev2   /cells=count row /statistics=chisq phi gamma .
filter by ag35m.
cro soclase by edlev2   /cells=count row /statistics=chisq phi gamma .
filter off.
* Comment : when the sample size goes down, the assocs and significances change -
*  but note that the sparse table is problematic anyway (low cell counts).
```



```
*** Comment: ordinal associations v correlations :
** There are some 'ordinal correlation' statistics available (esp 'Spearman's Rho', segment 4
below).
**  However these aren't really suited to ordinal data where there are relatively
**   few categories involved - they are intended for measures which are much
**  closer to an interval level.



*******************************************************.
*** Segment 10.3) Metric-categorical  statistics .
*******************************************************.

fre var=sex soclase.
graph /histogram= workhrs .


*********************************.
*** Seg10.3(i) Metric-categorical  graphics .

** We use confidence intervals :.
graph /errorbar(CI 95)=workhrs by soclase by sex  /title="Mean working hours with 95% CI's".
graph /errorbar(CI 95)=workhrs by sex by soclase /title="Mean working hours with 95% CI's".
* (note how the order of variables influences the presentation).

graph /errorbar(CI 99)=workhrs by soclase by sex  /title="Mean working hours with 99% CI's".
graph /errorbar(CI 90)=workhrs by soclase by sex  /title="Mean working hours with 90% CI's".


** You can also get the CI's as numbers, for example with :.
examine variables= workhrs by soclase by sex /nototal /plot=boxplot /statistics descriptives.



*********************************.
*** Seg10.3(ii) Metric-categorical association statistics ('anova') .

** There are several variations  :.

** 'means tables' is the easiest to begin with.
means tables=workhrs by sex by soclase /statistics=anova.
means tables=workhrs by soclase by sex /statistics=anova.
** NOTE: the anova significance is _bivariate only_ with the means procedure .
*   -> use the 'split files' command :.
sort cases by sex.
split files by sex.
means tables=workhrs by soclase /statistics=anova.
split files off.



** Many stats texts describe 'T-tests', which are interential tests for
*     the difference between the means of 2 groups (_only_ 2 groups).
t-test groups=sex(1,2) /variables=workhrs.
t-test groups=soclase(1,4) /variables=workhrs.



** T-tests are a subgroup of 'anova' methods, which can be applied to more
**   than two groups, eg :.
oneway workhrs by soclase .
means tables=workhrs by soclase / statistics=anova.
** (Note that the 'means tables' procedure use simple anova statistics).

** Comment: anova tests are covered at length in a great many textbooks, for two reasons:
*     - for some reason, they are thought to be particularly intuitive and thus pedagogically sound
*     - for some other reason, they are enormously popular in certain fields of the social sciences
*         eg, introductory level data analysis teaching in psychology is dominated by 'anova'
designs.
*     From a sociological analytical perspective, anova methods are usually a waste of time -
*       they are poor subsets of the more powerful 'statistical modelling' framework (session 4)
*        and they also require unusually strict parametric assumptions for the data in question.


*******************************************************.
*** Segment 10.4) Metric-metric inferential statistics .
```

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

descriptives var=workhrs earn2 .
graph /scatterplot=workhrs with earn2.

*********************************.
***  Seg10.4(i) Correlation statistics :.
correlate var=workhrs with earn2.

*********************************.
***  Seg10.4(ii) Bivariate regression model :.
regression var=workhrs earn2
     /dependent=earn2  /method=enter .


*********************************.
*** Seg10.4(iii) Non-parametric associations :.

** Example : working hours if between 30-50, by social class :.

compute wk3050f=(workhrs ge 30 & workhrs le 50 & sex=2).
fre var=wk3050f soclase.
filter by wk3050f.
graph /scatterplot=workhrs with soclase.
* (comment: to see patterns, use the 'bins' option on the scatterplot - open the scatterplot
*   as an output option, then 'edit' and 'properties' then 'point bins' and 'bins' ).
correlate var=workhrs with soclase.
* A non-parametric association : .
nonpar corr var=workhrs with soclase.
filter off.

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


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



```
*****************************************************.
******* EXERCISE 11 - MULTIVARIATE INFERENTIAL TECHNIQUES ON THE GHS95 FILE .
*****************************************************.



*****************************************************.
*** Segment 11.1) Open the file and prepare a few variables.
*****************************************************.

get file=!path2+"ghs95.sav".

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in part 2 syntax example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".

fre var=edlev2 soclase .

summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.
```



```
** .


*****************************************************.
*** Segment 11.2) Multivariate comparisons and inference.
*****************************************************.


** As one example :
*** Age and General health, by education and gender .

graph /histogram=age.
fre var=genhlth edlev2 sex .


** Graphically : .
sort cases by sex.
split file by sex.
graph /errorbar(CI 95)=age by genhlth by edlev2 .
split file off.
* Comment: not too easy to take in all of this info.

** In tables :.
sort cases by sex edlev2.
split files by sex edlev2.
means tables=age by genhlth /statistics=anova.
split files off.

** The tables of assoc statistics and anovas tells us whether age and health are significantly
**   associated within each group.
* (Note how 'split files' doesn't exclude missing values).
*   Lecture 3B has a format from this table.


*****************************************************.
*** Segment 11.3) Multivariate Regression model format.
*****************************************************.


** Outcome variable : earnings (cropped).

graph /histogram=earn2.

** Explanatory variables .

examine variables age workhrs /plot=boxplot .
compute age2=age**2.
graph /scatterplot=age with age2.

fre var=sex edlev2 region.
** Comment: categorical data needs to be expressed through 'dummy' variables
*    to go into a regression (minimum (c-1) variables for c categories)..

compute female=(sex=2).

compute hied=(edlev2=1).
compute loed=(edlev2=3).

compute london=(region=11).
compute outlond=(region=12 | region=13).
compute scotland=(region ge 18 & region le 22).
compute wales=(region=16 | region=17).

fre var=female hied loed london outlond scotland wales.


****** Regression model with all factors included :.

regression var=earn2 age age2 workhrs female hied loed london outlond scotland wales
     /dependent=earn2  /method=enter .


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

```
*********************************************************************.
******* EXERCISE 12 - MULTIVARIATE COMPARISONS  ON THE GHS95 FILE .
*********************************************************************.



*****************************************************.
*** Segment 12.1) Open the ghs95 file and run some variable constructions.
*****************************************************.

get file=!path2+"ghs95.sav".

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in previous syntax examples : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".
fre var=edlev2 soclase .

** Region and tenure :.
fre var= tenure.
compute tenure2=tenure.
recode tenure2 (1,2,3=1) (4,5,9,10,11=3) (6,7=2) (else=-999).
add value labels tenure2 1 "Own or buying" 2 "Social housing" 3 "Private renting".
missing values tenure2 (-999).
fre var=tenure2.
fre var=region.
* comment - there's more than one way the regional data could be recoded - you
*   could take greater acount of metropolitan status .
compute region2=region.
recode region2 (1,2,3,4=2) (5,6=1) (7,8,9=3) (10,14,15=4) (11=5) (12,13=6)
             (16,17=7) (18,19,20,21,22=8).
add value labels region2 1 "N West" 2 "North and Yorks" 3 "W and E Midlands"
         4 "South" 5 "Inner London" 6 "Outer London" 7 "Scotland" 8 "Wales" .
fre var=region2 .

** Earnings, working hours, and age.
summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.

** .

*****************************************************.
*** Segment 12.2) Multivariate comparisons where all variables are categorical .
*****************************************************.
**  .

***********.
* Seg12.2(i) Relationship between visits to the dentist and highest educational level.
***********.


fre var=dntstwhn edlev2 .
* These distributions suit analysis now - no further data management needed :.

cro tables=edlev2 by dntstwhn /cells=count row /statistics=phi chisq gamma .
graph /bar=pct by edlev2 by dntstwhn .
```



```
** These associations show a modest relation between the two variables -
*      people with higher levels of education are more likely to have regular check-ups .
*   (Both these variables could be regarded as ordinal : use the 'gamma' measure.

** Multivariate 1 : the association the same for men and women?.

cro tables=edlev2 by dntstwhn by sex /cells=count row /statistics=phi chisq gamma .

** A third (or further) variable in a crosstab is usually called the 'layer'
**   (the first two variables are the row and column) .


** In fact, controlling for a 3rd variable can also be achieved by several other data management
*    controls - including 'select', 'filter', or using 'split files' .

temp.
select if (sex=1).
cro table=edlev2 by dntstwhn /cells=count row .

** Comment: when 'temp' precedes 'select', the selection only operates on the next SPSS process.
**   Warning: if 'temp' doesn't precede 'select', then 'select' acts permanently on the file,
**      dropping all non-selected cases for good.

compute male=(sex=1).
filter by male.
cro edlev2 by dntstwhn /cells=count row .
show filter.
filter off.
cro edlev2 by dntstwhn /cells=count row.
show filter.

** Comment: when 'filter' is on, all analyses are restricted to cases where the filter variable=1.
**   The non-essential 'show filter' command is just used to confirm whether is filter
*       variable is currently in use - use it at any time to check your data.

sort cases by sex.
split file by sex.
cro edlev2 by dntstwhn /cells=count row.
graph /bar=pct by edlev2 by dntstwhn .
split file off.

** Comment: 'split files' runs all analyses on every permutation of the split variable.
**  Note that the data must be sorted by the split variable.


** Summary from above :.
cro tables=edlev2 by dntstwhn by sex /cells=count row /statistics=phi chisq gamma .
sort cases by sex.
split file by sex.
graph /bar=pct by edlev2 by dntstwhn .
split file off.

* The patterns shows a couple of things : women generally have higher rates of visiting the
*   dentist; both m and f rates are influenced by educational level in the same direction;
*    but men's rates are a little more influenced by educational level than are women's.



** Multivariate 2 : is the relation the same for men and women by their general health levels.
fre var=genhlth.
compute genhlth2=genhlth.
recode genhlth2 (1=1) (2,3=2) .
add value labels genhlth2 1 "Good health" 2 "Average / poor" .
fre var=genhlth2.

cro tables=edlev2 by dntstwhn by genhlth2 by sex
       /cells=count row /statistics=phi chisq gamma .
sort cases by sex.
split files by sex.
cro tables=edlev2 by dntstwhn by genhlth2 /cells=count row /statistics=phi chisq gamma .
graph /bar=mean(dntstwhn) by edlev2 by genhlth2 .
split files off.
* [note how we've used the mean of a categorical variable in the graphic
*   to give us an _approximate_ image of frequency of visits - the mean is not wholly
*  appropriate but serves to give a basic guide].
```

```
** Comment: some of the things that this shows are:
**  Male's use of dentists and assocaitions from educational level to use of dentists,
**   don't seem to vary according to whether in 'good health' or not.
**  Female's use of dentists and assocaitions from educational level to use of dentists,
**   do seem to vary slightly according to whether in 'good health' or not -  the association
*      is higher for those not in good health .
**  Note - overall, there's lots of information here - multiple comparisons soon become
**   quite clumsy.

** Problem 1: Too much disaggregation is not possible on limited size samples - quickly
*     start to have sparse tables and empty cells, which is ill suited to inferential conclusions.
*     This is particularly pronounced if some variables have skewed distributions where some
*      categories have very few cases, eg say we were interested in those in _poor_ health.


** Problem 2: Very complex tables and categorical comparisons are difficult to interpret.




**** Extension : fooling SPSS graphs .
** The graphs above involve 4 concepts, though SPSS bar charts only really allow 2 or
*  three variables.
fre var=edlev2 dntstwhn sex genhlth2 .
* We can 'cheat' the graph procedure to get a nicer display, by 'transforming'
*   data and 'merging' variables.
** Transform the dnsstwhn variable to a 'metric' (as above).
compute dnt2=dntstwhn.
recode dnt2 (1=4) (2=3) (3=2) (4=1).
variable label dnt2 "Frequency of visiting dentist" .
** 'Merge' gender and health into one variable.
compute sexhlth=(sex*10) + genhlth2.
add value labels sexhlth 11 "Male, good health" 12 "Male, average/poor health"
          21 "Female, good health" 22 "Female, average/poor health" .
** Now get a single graph :.
graph /bar=mean(dnt2) by sexhlth by edlev2  .
*****.



**************.
* Seg12.2(ii) Relationship between region and housing tenure.
**************.

fre var=tenure2 region2 .
* These distributions suit analysis now - no further data management needed :.

cro tables=region2 by tenure2 /cells=count row /statistics=phi chisq  .
graph /bar=pct by region2 by tenure2 .

** These associations show a modest relation between the two variables .
** However note that the association stats just show that there is some
**  association somewhere in the table - they don't show where exactly it lies.
** By inspection, inner London and Wales have different profiles to others,
**   with Wales having unusually high social housing, and IL high private renting.
** These can also be explored by making more specific contrasts, eg : .

temp.
select if (tenure2=1 | tenure2=3).
cro region2 by tenure2 /cells=count row /statistics=phi chisq .
temp.
select if ((tenure2=1 | tenure2=3) & region2 ne 5).
cro region2 by tenure2 /cells=count row /statistics=phi chisq .

** Inner London nearly entirely responsible for the regional patterns in owning v's renting
**    - if inner london is excluded, there's v little other difference.



**************.
* Seg12.2(iii) Further illustrations of multiple cross-classifications .
**************.

********.
*  'Tables' command gives us more control over presenting 3+ way crosstabulations, eg :.

tables /format blank missing ('.') /ftotal=ftot1 "Total"
```



```
  /tables dntstwhn + ftot1 by  sex > (edlev2)
  /statistics cpct ((pct4) ' ' :sex edlev2)   count ((F5.0) ' Cases ')
     /title="Visits to the dentist, by gender and educational level"  .


*******.
*  .


******************************************************.
*** Segment 12.3) Multivariate comparisons where one variable is metric, all others are categorical
.
******************************************************.
**  .

************.
* Seg12.3(i) Relationship between working hours, social class and gender .
************.

graph /histogram=workhrs .
fre var=sex soclase.


* Main descriptive method = nested analysis of means by groups (tables or graphs).

** Bivariate relationships :.
means tables=workhrs by sex soclase /cells=mean stddev count min max  /statistics=anova .
examine variables=workhrs by sex soclase /nototal /percentiles(1 5 10 50 90 95 99)
        /plot=boxplot /statistics=descriptives extreme .

** Multivariate relationships :.
examine variables=workhrs by soclase by sex /nototal /percentiles(1 5 10 50 90 95 99)
        /plot=boxplot /statistics=descriptives extreme .
means tables=workhrs by soclase by sex /cells=mean stddev count min max  /statistics=anova .
means tables=workhrs by sex by soclase  /cells=mean stddev count /statistics=anova .
sort cases by sex.
split files by sex.
means tables=workhrs by soclase /cells=mean stddev count /statistics=anova .
split files off.
* Note - for an anova statistics to be split between groups under 'means', using 'split files'
*   is the easiest way.
graph /bar=mean(workhrs) by sex by soclase.
graph /bar=median(workhrs) by sex by soclase.
graph /bar=mean(workhrs)  by soclase by sex .
graph /bar=median(workhrs)  by soclase by sex .
* (a different picture between means and medians).

graph /errorbar(CI 95)=workhrs by soclase by sex.

** Comment - there's lots of alternative presentational options here..! .

** Inference statistics: overlapping error bars show specific contrasts between category means,
**   whereas anova statistics show if there are _any_ contrasts anywhere in the data.


************.
* Seg12.3(ii) Relationship between income, education, region and gender .
************.


graph /histogram=earn2.
means tables=earn2 by edlev2 region2 sex /statistics=anova.

** All three factors are related to earnings on their own.

cro region2 by edlev2 /cells=count row /statistics=phi .
cro sex by edlev2 /cells=count row /statistics=phi .
cro region2 by sex /cells=count row /statistics=phi .
** region-educ, and sex-educ are both related, but region-sex are not.


** Now try looking at everything .

** (a) Reviewing : .
means tables=earn2 by region2 by edlev2 by sex .
sort cases by sex.
split files by sex.
```

```
graph /bar=mean(earn2) by region2 by edlev2  .
graph /bar=median(earn2) by region2 by edlev2  .
graph /bar=median(earn2) by edlev2 by region2 .
split files off.
examine variables=earn2 by sex by edlev2 by region2  /nototal /plot=boxplot .


** (b) Summarising with association statstics :.
sort cases by sex region2.
split files by sex region2.
means tables=earn2 by edlev2 /statistics=anova.
split files off.

** (c) Looking at the inferential significance :.

sort cases by sex region2.
split files by sex region2.
means tables=earn2 by edlev2 /statistics=anova.
split files off.

sort cases by sex .
split files by sex.
graph /errorbar(CI 95)=earn2 by edlev2 by region2 .
split files off.
* or to get them on the same graph :.
fre var=sex edlev2.
compute sexeduc=-999.
if (sex ge 1 & edlev2 ge 1) sexeduc=(sex*10) + edlev2.
add value labels sexeduc 11 "Male, higher education" 12 "Male, intermediate education" 13 "Male,
lower education"
                21 "Female, higher education" 22 "Female, intermediate education" 23 "Female, lower
education"  .
fre var=sexeduc.
missing values sexeduc (-999).
graph /errorbar(CI 95)=earn2 by sexeduc by region2.


*** Anova and Manova also undertake significance tests of associations plus their interactions:.

anova variables=earn2 by edlev2 (1,3) region2 (1,8) sex(1,2)  .

** Comment: the earnings variable here is continuous - a linear transformation
*   of it could change the story a bit :.

compute lnearn2=ln(earn2).
graph /scatterplot=earn2 with lnearn2.
* The log transform takes the emphasis away from high positive values.

anova variables=lnearn2 by edlev2 (1,3) region2 (1,8) sex(1,2)  .
** This gives different patterns of influence.

** However - anova / manova aren't really descriptive techniques - they're more like models.

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


************.
* Seg12.3(iii) Multivariate comparisons where two variables are metric, all others are categorical .
************.

** main methods = nested tables of bivariate correlations, and scatterplots .

graph /histogram=earn2 .
graph /histogram=workhrs .
fre var=sex edlev2 .



**  Nested bivariate correlations .

correlate var=earn2 with workhrs .
sort cases by sex edlev2 .
split files by sex edlev2 .
correlate var=earn2 with workhrs  .
split files off.
```



```
** Graphics :.

graph /scatterplot=workhrs with earn2 by sex .

graph /scatterplot=workhrs with earn2 by sexeduc.
*(assumes 'sexeduc' is created as above).

** Comment : you can adjust the scatterplot display by opening the graphic
*    and changing the point identifiers, for instance icons and using 'point bins'.
*   However, scatterplots like this don't look especially good with lots and lots of
*     cases - they can be more informative on smaller samples and/or categorical variables.



************************************************************.
*** Segment 12.4) Multivariate comparisons where 2+ variables are metric .
************************************************************.
**  .


**************.
** Seg12.4(i) All metric variables .
**************.

** this particular situation suits 'partial correlation coefficients' (which are best
*    compared immediately with bivariate correlations).

compute age2=age**2 .
descriptives var=earn2 workhrs age age2 .


** Bivariate comparisons :.
correlate var=earn2 workhrs age age2 .

** Multivariate comparison :.
* Earnings to working hours, adjusting for linear age .
partial correlation var=earn2 workhrs by age  .
* Earnings to working hours, adjusting for quadrtic age .
partial correlation var=earn2 workhrs by age age2 .
* Earnings to age, adjusting for working hours.
partial correlation var=earn2 age by workhrs .
* Working hours to age, adjusting for earnings  (though doesn't make much theoretical sense).
partial correlation var=workhrs age by earn2 .



** These compare to multiple regression coefficients (see next section):.
regression var=earn2 workhrs age age2
     /dependent=earn2 /method=enter .


**************.
** Seg12.4(ii) 2+ metric plus some categorical variables .
**************.

** Easiest strategy is to reduce the data so that only 1 variable is treated as metric, all others
**   are treated as categorical .

descriptives var=earn2 age .
fre var=sex region2.

compute age6=age.
recode age6 (16 thru 25=1) (26 thru 35=2) (36 thru 45=3) (46 thru 55=4)
                (56 thru 65=5) (66 thru hi=6).
add value labels age6 1 "16-25yrs" 2 "26-35yrs" 3 "36-45yrs" 4 "46-55yrs" 5 "56-65yrs" 6 "66+ yrs".
fre var=age6.

examine variables=earn2 by sex by region2 by age6   /nototal /plot=boxplot .
sort cases by sex.
split files by sex.
graph /errorbar(CI 95)=earn2 by age6 by region2.
split files off.



************************************************************.
*** Segment 12.5) Multivariate comparisons: extensions .
```

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


****************.
** Seg12.5(i) Using (sampling) weights .
***************.

** The GHS95 dataset doesn't have sample weights on it (because it's an extract).
** But most datasets will do so .

** Example : the BHPS 2002 data file.
get file=!path3+"lindresp.sav".

descriptives var=lxrwght .
fre var=lsex  lvote lqfedhi .
* Firstly some data manipulation :.
recode lvote (-9,-7,10,11=-999) (4 thru hi=4).
add value labels lvote 4 "Other" .
compute educ3=lqfedhi.
recode educ3 (1,2,3,4=1) (5,6,8,10=2) (7,9,11,12,13=3) (else=-999).
add value labels educ3 1 "Higher / post-school" 2 "Vocational or higher school" 3 "Low school or
below" .
missing values educ3 lvote (-999).

fre var=lsex lvote educ3 .

cro educ3 by lvote by lsex /cells=count row /statistics=phi .
** This analysis is unweighted.

** What about if we put the basic sample weights on: .
weight by lxrwght.
fre var=lsex lvote educ3 .
cro educ3 by lvote by lsex /cells=count row /statistics=phi .
weight off.

** All analyses after a 'weight by' command and until a 'weight off' command,
*     are run using the specified weighting variable.

** Note on weights : this example is fairly typical: if you look at the specific
*     univariate valid percents or table row percents, you see slight differences
*     between the weighted and unweighted proportions.
*   However the net conclusions from a multivariate analysis are not much
*    altered by using the weights.
*   This is normally the case - but not always, and the important thing is that you
*     don't know if weights will make a difference until after you've tried them out.

** Note also that most but not all SPSS procedures react to weights (certainly
**   all the basic ones); however some are unaffected by weights .



****************.
** Seg12.5(ii) Multivariate comparisons extension: missing values  .
***************.

** Use the GHS 95 example file.

get file=!path2+"ghs95.sav".

fre var=sex genhlth dntstwhn  .

missing values genhlth dntstwhn (-777).
*(this just means all values are used in data) .
cro sex by dntstwhn by genhlth /cells=count row .
* Observe - quite a lot of missing data on these questions.

** 4 common solutions :.

** (a) Most used (by far): listwise deletion :.
missing values genhlth dntstwhn (-9).
fre var=sex genhlth dntstwhn  .
cro sex by dntstwhn by genhlth /cells=count row /statistics=phi .


** (b) Modal imputation (occasional usage).
missing values genhlth dntstwhn (-777).
fre var=sex genhlth dntstwhn  .
```



```
* Modal genhlth = 'good' ; model dntstwhn='regular checkup'.
compute genhlth2=genhlth.
compute dnt2=dntstwhn.
recode genhlth2 (-9=1).
recode dnt2 (-9=1).
add value labels genhlth2 1 "Good" 2 "Fairly good" 3 "Poor".
add value labels dnt2 1 "Regular checkup" 2 "Occassional checkup"
          3 "Having trouble" 4 "Never go to dentist" .
fre var=genhlth2 dnt2.
cro sex by dnt2 by genhlth2 /cells=count row /statistics=phi .



** (c) a variation on 'modal imputation' is 'Null Imputation' -
*   code all the missings to the 'middle' or 'inextreme' category according to an a priori
*   judgement over the role of the categories. Eg, the 'null' category for genhlth
*   could be 'fairly good', and the 'null' for dntstwhn could be 'having trouble'.
missing values genhlth dntstwhn (-777).
fre var=sex genhlth dntstwhn  .
compute genhlth3=genhlth.
compute dnt3=dntstwhn.
recode genhlth3 (-9=2).
recode dnt3 (-9=3).
add value labels genhlth3 1 "Good" 2 "Fairly good" 3 "Poor".
add value labels dnt3 1 "Regular checkup" 2 "Occassional checkup"
          3 "Having trouble" 4 "Never go to dentist" .
fre var=genhlth3 dnt3.
cro sex by dnt3 by genhlth3 /cells=count row /statistics=phi .

** (d) Informed imputation .

* Here we use some other information about the cases with missing values, to make a 'best
*  guess' at what the likely value would have been.
** Statistical routines are available to do that for you, though they are still in development
*     and SPSS's options only allow for certain treatments.
** It is also possible to make informed imputations by simpler calculations,
**   which, as below, typically involve both informed and modal imputation : .


missing values genhlth dntstwhn (-9).
cro genhlth by longill /cells=count row  .
cro dntstwhn by teeth /cells=count row .
missing values genhlth dntstwhn (-777).


* For genhlth : if longill=3, genhlth is very likely to be 1; if longill=1, genhlth likely to be 3.
* For dntstwhn : anyone with no teeth can be assumed for these purposes not to visit dentist.

compute genhlth4=genhlth.
if (genhlth=-9) genhlth4=1.
if (genhlth=-9 & longill=3) genhlth4 =1.
if (genhlth=-9 & longill=1) genhlth4 =3.
*(so long as the ifs are run in this order, then the latter ones get priority).

compute dnt4=dntstwhn.
if (dntstwhn=-9) dnt4=1.
if (teeth=2) dnt4=1.
* (again, the ifs should be run in order, though in this case its not so important).

add value labels genhlth4 1 "Good" 2 "Fairly good" 3 "Poor".
add value labels dnt4 1 "Regular checkup" 2 "Occassional checkup"
          3 "Having trouble" 4 "Never go to dentist" .
fre var=genhlth4 dnt4.
cro sex by dnt4 by genhlth4 /cells=count row /statistics=phi .



*** Comment on missing value imputations: all these options are possibilities and
*    none are strictly 'right' or 'wrong' - depending on the situation, missing data
*   treatments can have quite a big impact upon our analysis.
*   To justify ignoring them, you have to show that some alternative treatments
*   don't make a huge difference to the outcomes.


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

```
*****************************************************************.
******* EXERCISE 13 - MULTIVARIATE MODELLING TECHNIQUES ON THE GHS95 FILE .
*****************************************************************.



******************************************************.
*** Segment 13.1) Open the file and prepare a few variables.
******************************************************.

get file=!path2+"ghs95.sav".


fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in part 2 syntax example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".
fre var=edlev2 soclase .

fre var=region.
compute region2=region.
recode region2 (1,2,3,4=2) (5,6=1) (7,8,9=3) (10,14,15=4) (11=5) (12,13=6)
              (16,17=7) (18,19,20,21,22=8).
add value labels region2 1 "N West" 2 "North and Yorks" 3 "W and E Midlands"
          4 "South" 5 "Inner London" 6 "Outer London" 7 "Scotland" 8 "Wales" .
fre var=region2 .

summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.

** .


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



******************************************************.
*** Segment 13.2) Example multiple regression models.
******************************************************.


****************.
*** Seg13.2(i) Prediction of earnings.
****************.

** Outcome variable : earnings (cropped).
graph /histogram=earn2.

** Explanatory variables .
examine variables age workhrs /plot=boxplot .
compute age2=age**2.
graph /scatterplot=age with age2.
fre var=sex edlev2 region.

** Comment: categorical data needs to be expressed through 'dummy' variables
*    to go into a regression (minimum (c-1) variables for c categories)..
compute female=(sex=2).
compute hied=(edlev2=1).
compute loed=(edlev2=3).
```



```
compute london=(region=11).
compute outlond=(region=12 | region=13).
compute scotland=(region ge 18 & region le 22).
compute wales=(region=16 | region=17).
fre var=female hied loed london outlond scotland wales.


****** Regression model with all factors included :.

regression var=earn2 age age2 workhrs female hied loed london outlond scotland wales
     /dependent=earn2  /method=enter .

** Interpretation: see the sign and significance of coefficient estimates.

** [Comment: to interpret a quadratic function of age, see the website:
** http://staff.stir.ac.uk/paul.lambert/polynomials.xls ] .



****** Use an algorithm for only including significant terms:.

regression var=earn2 age age2 workhrs female hied loed london outlond scotland wales
     /dependent=earn2  /method=stepwise .

* SPSS automatically excludes non-significant explanatory factors.
* Cramer 2004 makes a lot of the model selection technique: chapters 5 and 6
*  both cover multiple regression, differing only in terms of the model selection protocol.



****************.
*** Seg13.2(ii) Prediction of working hours.
****************.

** Dependent variables :.
graph /histogram=workhrs .

** Explanatory variables : try the same as earnings model  :.
descriptives var=age age2.
fre var=female hied loed london outlond scotland wales


*** Regression model :.

regression var=workhrs age age2 female hied loed london outlond scotland wales
     /dependent=workhrs  /method=enter .


** Comment: the relation between the models of seg 2(i) and (ii) are sometimes used to
**   formulate 'path analysis' diagrams.


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


******************************************************.
*** Segment 13.3) Example categorical regression models (1) :
****                              Logistic regression .
******************************************************.


****************.
*** Seg13.3(i) Prediction of probability of having a computer in the household .
****************.

** Dependent variable :.
fre var=computer.
* (remember that this data is from 1995).
compute comput=computer.
recode comput (1=1) (2=0) (else=-999).
missing values comput (-999).
fre var=comput.


** Explanatory variables (i) (these all defined above in segment 2):.
```

```
descriptives var=age age2 earn2 .
fre var=female hied loed london outlond scotland wales .

** Logistic regression model :.

logistic regression var= comput
       /method=enter  age age2 earn2 female hied loed london outlond scotland wales .

** Explanatory variables (ii) - look at household level well-being .

graph /histogram=hohx.
compute lnhohx=-999.
if (hohx ge 50 & hohx le 5000) lnhohx = ln(hohx).
missing values lnhohx (-999).
graph /histogram=lnhohx.

fre var=hohscle.
compute hohnm=hohscle.
recode hohnm (1,2,3=1) (4,5,6=0) (else=-999).
variable label hohnm "Head of household managerial, professional or skilled non-manual" .
missing values hohnm (-999).
fre var=hohnm.


logistic regression var= comput
       /method=enter  age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .

* (this explains patterns much better).

** And now try stepwise :.

logistic regression var= comput
       /method=stepwise
           age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .

* This leads to quite a pithy model.




*****************.
*** Seg13.3(ii) Prediction of probability of an 'advantaged' tenure situation .
*****************.

** Dependent variable :.
fre var=tenure.
** It is quite common to use logistic regression to devise a dichotomous contrast from
**  source data of a different form - for instance, here we allocate tenure categories to
** 'advantaged' and 'disadvantaged' situations, then model that.
compute ten2=tenure.
recode ten2 (1,2,3,9,10=1) (4,5,6,7,8,11,12=0) (else=-999).
missing value ten2 (-999).
variable label ten2 "Advantaged tenure position".
fre var=ten2.

** Explanatory variables (as in earlier segments):.
descriptives var=age age2 earn2 lnhohx .
fre var=female hied loed london outlond scotland wales hohnm  .


logistic regression var= ten2
        /method=enter
           age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .

logistic regression var= ten2
        /method=stepwise
           age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .


* Here for example, Scotland but not wales has less advantaged average tenure after
**   controlling for other factors.



*****************************************************.
*** Segment 13.4) Example categorical regression models (2) :
```



```
****                              Multi-category models and loglinear models  .
*****************************************************.

** There are several model formats .
** Here try one outcome: tenure in 4 categories .
fre var=tenure.
compute ten3=tenure.
recode ten3 (2=1) (1,3=2) (9,10,11=3) (4,5,6,7,8,12=4) (else=-999).
add value labels ten3 1 "Owner outright" 2 "Buying" 3 "Private renting" 4 "Social and/or
unfavourable renting" .
missing values ten3 (-999).
fre var=ten3.

* Comment: reducing to 4 categories is not essential but serves 2 purposes:
*   it is cognitively easier; and it increase the number of cases per categories.


** Explanatory variables as above  : .
descriptives var=earn2 lnhohx .
fre var=female hied loed london outlond scotland wales hohnm  .



*****************.
*** Seg13.4(i) Multinomial logistic regression  .
*****************.

fre var=ten3 .
** The mlogit is strictly nominal and contrasts all categories with each other .


nomreg ten3  with
      earn2 lnhohx female hied loed london outlond scotland wales hohnm
   /criteria = cin(95) delta(0) mxiter(100) mxstep(5) lconverge(0) pconverge(1.0E-6) singular(1.0E-
8)
  /model  /intercept = include
  /print= parameter summary lrt.

* This model currently contrasts against 'unfavourable' renting - to change that,
*  use the 'base' subcommand  - the below contrasts with 'buying' .

nomreg ten3 (base=2) with
      earn2 lnhohx female hied loed london outlond scotland wales hohnm
   /criteria = cin(95) delta(0) mxiter(100) mxstep(5) lconverge(0) pconverge(1.0E-6) singular(1.0E-
8)
  /model  /intercept = include
  /print= parameter summary lrt.


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



*****************.
*** Seg13.4(ii) Ordered logistic regression  .
*****************.

** If you feel the categorical data is ordinal, an ordered logit can be used.
** In terms of 'advantage', it might be plausible to treat ten3 as ordered from more to less.
fre var=ten3 .


plum ten3 with
        earn2 lnhohx female hied loed london outlond scotland wales hohnm
  /criteria = cin(95) delta(0) mxiter(100) mxstep(5) lconverge(0)
  pconverge(1.0E-6) singular(1.0E-8)
  /link = logit   /print = fit parameter summary .

** A test of  the value of the ordered logit would be little loss of (pseudo)-r2 compared to an
mlogit :.
* (mlogit will always explain more, but ordered logit may be more parsimonious).
** Here, there's quite a drop, suggesting the ordered logit isn't that helpful .



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

```
*** Seg13.4(iii) Loglinear modelling  .
*****************.

** Loglinear models primarily only apply when all the variables involved are categorical
* (though, stirctly, they can be adapted to incorporate metric variables).
**  They are also classically used when the variables invovled don't have a strong
*     dependent var v's explanatory vars structure.

** Loglinear models have been, historically, very widely used in sociology.
* The best description of them is in Gilbert 1993; also see the appendix in Buckingham and Saunders
2004.
** However : loglinear models are to some degree an outdated technology - now that
**   categorical regression models are more easily estimated, many people no longer
**   bother with loglinear models.
** Another weakness is that (although they can be extended considerably to pick up more
**  complex patterns), in their basic form, they don't distinguish between the effects of specific
**   category values in determining a pattern, but rather simply ascribe the influence to the
**   overall categorical variable.

fre var=ten3.
fre var=edlev2 hohscle region2.


** An example loglinear model .

* The SPSS loglinear modelling for multiple categorical relationships involves trying out several
* models in succession, and contrasting which ones fit better .


** i) Main and all 2 way effects only :.
genlog  ten3 edlev2 hohscle region2
  /model=poisson   /print=freq resid dev adjresid
  /plot=resid( adjresid) normprob( adjresid)
  /criteria = cin(95) iterate(50) convergence(0.001) delta(0.5)
  /design   ten3 edlev2 hohscle region2
        ten3*edlev2 ten3*hohscle ten3*region2 edlev2*hohscle edlev2*region2 hohscle*region2  .


*
* => LR chi-square = 625 on 540, p=0.007 , ie still a significant gap from adequate fit .

** ii) Main and all 2 way and selected  three way :.
genlog  ten3 edlev2 hohscle region2
     /model=poisson   /print=freq resid dev adjresid
  /plot=resid( adjresid) normprob( adjresid)
  /criteria = cin(95) iterate(50) convergence(0.001) delta(0.5)
  /design   ten3 edlev2 hohscle region2
        ten3*edlev2 ten3*hohscle ten3*region2 edlev2*hohscle edlev2*region2 hohscle*region2
           ten3*hohscle*region2  .
* => LR chi-square = 341 on 414 df, p=0.996  :
*   This model could be accepted : there is no longer a significant lack of fit between the
**    data and the model .
** Full procedure: try more permutations of variable effects, to locate the most parsimonious
** model that is also a good fit (see Gilbert 1993) .


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




*********************************************************************.
******* EXERCISE 14 - MULTIVARIATE MODELLING EXTENSIONS  .
*********************************************************************.


*****************************************************.
*** Segment 14.1) Open the GHS file and run some preliminary data management .
*****************************************************.


get file=!path2+"ghs95.sav".
```



```
descriptives var=all.

fre var=sex edlev soclase .

** Treat the edlev and soclase variables, as in part 2 syntax example : .
compute edlev2=edlev.
recode edlev2 (1,2,3=1) (4,5,6,7,8,13,15,16=2) (9,10,11,12,17=3) (else=-999).
variable label edlev2 "Highest educational qualification, 3 categories".
add value labels edlev2 1 "Higher level qualification" 2 "Intermediate" 3 "Low school level or no
qualification" .
missing values edlev2 (-999).
missing values soclase (-9,7).
variable label soclase "Registrar General's Social Class (based on current or last occupation)".

fre var=edlev2 soclase .

summarize var=earnings workhrs age /cells=count mean stddev min max .
compute earn2=earnings.
if (earnings le 50 | earnings gt 2000) earn2=-999.
missing values earn2 (-999).
variable label earn2 "Average weekly earnings (range 51-2000 only)".
graph /histogram(normal)=earn2.
graph /histogram(normal)=workhrs.
** .

examine variables age workhrs /plot=boxplot .
compute age2=age**2.
graph /scatterplot=age with age2.

compute female=(sex=2).
compute hied=(edlev2=1).
compute loed=(edlev2=3).
compute london=(region=11).
compute outlond=(region=12 | region=13).
compute scotland=(region ge 18 & region le 22).
compute wales=(region=16 | region=17).
fre var=female hied loed london outlond scotland wales.




*****************************************************.
*** Segment 14.2) Checking for Multicollinearity .
*************************************************************.

***** Check for multicollinearity in the full model :.

regression var=earn2 age age2 workhrs female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol collin  /dependent=earn2   /method=enter .

** Problem: age and age-2 are collinear, and large Standardised coeffs suggest they are problematic.
**  Comment: normally, age and age-2 can go in together ok, but in this example, they are
**  confounding the maths of the estimation.

** A simple solution : piecewise age  .

fre var=age.
compute age5=age.
recode age5 (16 thru 24=1) (25 thru 35=2) (36 thru 55=3) (56 thru 65=4) (66 thru hi=5).
add value labels age5 1 "16-24 years" 2 "25-35 years" 3 "36-55 years"
             4 "56-65 years" 5 "66+ years" .
fre var=age5.
compute ag1624=(age5=1).
compute ag2535=(age5=2).
compute ag3655=(age5=3).
compute ag5665=(age5=4).
compute ag66ab=(age5=5).


regression var=earn2 ag1624 ag3655 ag5665 ag66ab
      workhrs female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol collin  /dependent=earn2   /method=enter .

* Here, we've lost the neater quadratic, but the estimates will be more stable.
* Comment: collinearity only matters insofar as there's a danger in misrepresenting
*  the net contributions of different vars - in this example, it was important, because
```

```
*   the stdzd beta around age is much less after controlling for its collinearity.

logistic regression var= comput
       /method=enter  age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .

** this is another way of getting round the same problem .


**** Multicollinearity in a logistic or categorical regression model :.

** (Data management from part 2 example) .
fre var=computer.
compute comput=computer.
recode comput (1=1) (2=0) (else=-999).
missing values comput (-999).
fre var=comput.
graph /histogram=hohx.
compute lnhohx=-999.
if (hohx ge 50 & hohx le 5000) lnhohx = ln(hohx).
missing values lnhohx (-999).
graph /histogram=lnhohx.
fre var=hohscle.
compute hohnm=hohscle.
recode hohnm (1,2,3=1) (4,5,6=0) (else=-999).
variable label hohnm "Head of household managerial, professional or skilled non-manual" .
missing values hohnm (-999).
fre var=hohnm.


logistic regression var= comput
       /method=enter  age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm .

** there's no automated collinearity test in any of the categorical regression models .
* BUT - collinearity tests are independent of the functional form of the dependent variable -
*   so it's fine to just do a multiple regression and check the collinearity statistics for it.


regression var=comput age age2 earn2 female hied loed london outlond scotland wales lnhohx hohnm
    /statistics=R coeffs anova tol collin  /dependent=comput   /method=enter .

* (comment : possible collinearity problems on age, buty not on income).
* (This illustrates that multiple regression can be a reasonable summary for binary outcomes also).

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


*****************************************************.
*** Segment 14.3) Interaction terms  .
************************************************************.

** Interaction effects are, substantively, some of the most interesting features of statistical
models.


regression var=earn2 age ag66ab
       workhrs female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol   /dependent=earn2   /method=enter .

* Hypothesis : is the negative main effect of being female mediated by any of the other factors? .

** We construct two-way interaction terms just by multiplying :.
* (we could do the same for 3-way effects and so on).

compute femage=female*age.
compute femwhrs=female*workhrs .
compute femhied=female*hied.
compute femloed=female*loed.

compute femlond=female*london.
compute femolon=female*outlond.
compute femscot=female*scotland.
compute femwal=female*wales.

descriptives var=femage femwhrs femhied femloed femlond femolon femscot femwal .

** We can just chuck the effects in :.
** (Comment: multicollinearity isn't usually exacerbated too much by interaction effects, though
```



```
*   they will influence the collinearity diagnostic statstics if used).

regression var=earn2 age ag66ab
       workhrs female hied loed london outlond scotland wales
     femage femwhrs femhied femloed femlond femolon femscot femwal
    /statistics=R coeffs anova    /dependent=earn2   /method=enter .

* The inclusion of these options has increased the r2 slightly, but not that much.
* Note how interaction effects can change the main effects of previous variables -
* Esp: The main effect of gender has largely disappeared after accounting for these differences.

** It's sensible to remove again some of the uninfluential interactions :.
regression var=earn2 age ag66ab
       workhrs female hied loed london outlond scotland wales
     femage  femhied femloed  femolon femscot
    /statistics=R coeffs anova    /dependent=earn2   /method=enter .


*** When a lot of ionteractions with one term are possible, it is often preferred to treat
*   the term as a source of 'Structural breaks' in the data, and estimate different models
**   for each subcategory .
** Common examples of this are : different models for men and women; different
*     models for data from equivalent surveys in different countries or years.


** Example :.

sort cases by sex.
split files by sex.
regression var=earn2 age ag66ab
       workhrs hied loed london outlond scotland wales
         /statistics=R coeffs anova  tol /dependent=earn2   /method=enter .
split files off.
* The different r2's and differences in the coefficient estiamtes suggests it is useful
*  here to think of women as having structurally different earnings determinants to men.



************************************************************.
*** Segment 14.4) Checking model assumptions  .
************************************************************.



***** Export the predicted values and the residuals, and check their diagnostics :.

** Model 1: straight down the line.
regression var=workhrs age age2 female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol   /dependent=workhrs /method=enter
      /save pred (m1pred) zpred (m1zpred) resid (m1resid) zresid (m1zresid) .


** Model 2: a bit more problematic, mainly because of skewed earnings data.
regression var=earn2 age ag66ab
     workhrs female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol  /dependent=earn2   /method=enter
      /save pred (m2pred) zpred (m2zpred) resid (m2resid) zresid (m2zresid) .
* (check the end of your data file - there should be some new variables showing
*   the predicted values and residuals).

** Following Allison 1999, the three most important assumptions to check are
** mean independence; homoscedasticity; and linearity (in that order).



************** Test for mean independence (and for normal disturbance)  :
**      residuals should have a mean of zero .

graph /histogram(normal)= m1resid  .
graph /histogram(normal)=m2resid .

* This assumption is reasonable for both model.


************* Test for homoscedasticity : variance of residuals should be constant, and
*    unrelated to the explanatory variables  <=> do expected values have an uneven dispersion
*    against predicted values (eg greater dispersion at higher or lower values).
```

```
**    this can be revealed by plotting observed v's residuals, or just a histogram
*     of the standardised residuals .

graph /scatterplot=workhrs with m1zresid  .
graph /scatterplot=earn2 with m2zresid  .

graph /histogram(normal)=m1zresid.
graph /histogram(normal)=m2zresid.

* A 'normal Q-Q plot' is also often used here (just a better way of comparing against a normal
curve).

** Both m1 and m2 seem ok in this regard .



****************** Test for linearity  : are residuals independent of predicted values? .
graph /scatterplot=m1zpred with m1zresid  .
graph /scatterplot=m2zpred with m2zresid  .


** There should be no pattern to the scatterplot (ie, residuals are unrelated to predicted values).
** Model 1 is ok, but model 2 isn't .
* (The 'lines' in m1 just reflect the low number of discrete hours covered).




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


** Simplification: the same diagnostics can be checked within a regression model,
**   using SPSS subcommands :.


** Model 1 .
regression var=workhrs age age2 female hied loed london outlond scotland wales
    /statistics=R coeffs anova tol   /dependent=workhrs /method=enter
       /residuals durbin histogram(resid) histogram(zresid) /casewise plot(zresid) outliers(3)
           /scatterplot (*zresid, workhrs) (*zresid, *zpred)  .



*****************************************************************************.
** Extension illustration : Small scale regression models :
**    the student questionnaire .
***************.


****************************************.
*** First, the example (fictional) data from the lda slides.

** Input the data :.
data list free/  sex age height grade .
begin data
1 17 1.73 1
1 18 1.85 2
2 17 1.60 3
2 18 1.69 1
2 19 1.66 2
end data.
descriptives var=all.
* Comment: 1 case added cf most of the class eg sheets, to simplify model ouputs.
* (note how the grade data is coded to numeric values for convenience ).
** .

* (Constuct a dummy variable for gender).
compute male=(sex=1).
fre var=male.

regression var=height male
     /dependent=height  /method=enter .

regression var=height male age
     /dependent=height  /method=enter .
```



```
** Comment - the model can be estimated, but if the cases are sampled, we can't
*    have any confidence in the inferential significance of the association.
** If the cases we a census (eg, a class of 5 pupils only), the regression would work
*    as a true desription of them. However, if the number of explanatory variables
*    exceeeds the number of cases, then the regression estimation

** Saving predicted values and errors :.
regression var=height male age
     /dependent=height  /method=enter /save pred (m1pred) resid(m1resid) .
fre var=m1pred m1resid.

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




*********************************************************************.
******* EXERCISE 15 - USING MACROS AND INCLUDE FILES .
*********************************************************************.



*****************************************************.
*** Segment 15.1) Using an include file to process commonly used commands .
*****************************************************.

** An include file just runs all the text in the file as SPSS syntax, in one go.
** A simple use is to run sections of commands that we keep repeating -
**  eg, in earlier sessions, we've repeatedly run a set of data managment operations
**   on the GHS file - we can do this more quickly by

get file=!path2+"ghs95.sav".
include file=!path8+"ghs95_prepare.sps".



** Comment: this sort of use of an include file can be handy when you have quite
*    complex data management exercises - eg defining lots and lots of dummy variables.




*****************************************************.
*** Segment 15.2) Using an include file to store complex variable information .
*****************************************************.

** Some variables have complex information associated with them - for instance,
** lots of surveys collect occupational classification information
**  according to standardised categorisations, such as the 'SEG' (socio-economic group)
**   system, which has 21 standard categories. .

** A convenient technique is to store such value labels in separate syntax files,
**   and run those files as include files within your analytical session (rather
**   then adding too much text to your main file).

** The file seglabelsv1 allows you to add value labels to a variable called 'seg' -
** in the GHS example file, you could use this to change the value labelling slightly .

get file=!path2+"ghs95.sav".
fre var=segead.
* (the old value labels).
compute seg=segead.
fre var=seg.
include file=!path8+"seglabelsv1.sps" .
fre var=seg.



*****************************************************.
*** Segment 15.3) Defining some macros .
*****************************************************.
```

```
** Macros are explicit instructions that tell SPSS that, when it encounters a certain
**  text, it is actually to treat it as another bit of text.
** The precise format for macros is documented in the SPSS manuals, but
**  it is easiest to learn from examples .


get file=!path1+"ghs95.sav".

fre var=centheat video freezer washmach drier dishwash microwve phone cdplyer computer .


************.
** Macro 1 - just defines a group of variables .

define !comods () centheat video freezer washmach drier dishwash microwve phone cdplyer computer
!enddefine.

** This just tells SPSS that when it encounters !comods, it should replace it with the
** text centheat video freezer washmach drier dishwash microwve phone cdplyer computer .

fre var= !comods .

add value labels !comods 1 "Has this in household" 2 "Doesn't have this" .
fre var=!comods .


********** .
** Macro2 - define a relatively long command more easily :.

examine variables age workhrs /nototal /plot=boxplot /statistics descriptives extreme .

define exvars (vars=!enclose('{','}')) .
examine variables !vars /nototal /plot=boxplot /statistics descriptives extreme .
!enddefine.

** This tells SPSS that when it encounters 'exvars' it should read it as the command above,
**   with the specified codes of vars being substituted for the vars 'argument'.

exvars vars={age workhrs}.
exvars vars={age}.
exvars vars={tea earnings} .

** Comment - this might be used for relatively complicated commands, eg regression models.


** Additional comment: you can use 'set mprint' to show macro details :.

exvars vars={age}.
set mprint on.
exvars vars={age}.
set mprint off.
exvars vars={age}.

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


**************************************************.
*** Segment 15.4) Using an include file to define macros (1) - for complex variable information .
**************************************************.

** As in the earlier example (segment 4) it can be desirable to use include files to store
**  more complex variable information.

** However the information can be handled more easily by using the include
**   file to define a macro with the variable information, allowing the information to be
**   used on different variables rather than just a predefined name .
**   .


get file=!path2+"ghs95.sav".
fre var=segead.
* (the old value labels).
compute seg=segead.
compute seg2=segead.
fre var=seg seg2 .
include file=!path8+"seglabelsv2.sps" .
seglab occ={seg seg2} .
```



```
fre var=seg seg2 .

** Ie, by using the macro format, we can add the labels to any number of new variables in {}.


** Comment: many examples of this sort of include file are available over the internet - see
**  for instance : http://camsis.stir.ac.uk/occunits/distribution.html  .


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

**************************************************.
*** Segment 15.5) Using an include file to define macros (2) - for suites of commonly used commands
.
**************************************************.

** The file regressions.sps file has a selection of regression model options specified in it as
**  alternative macros (plus some relevant preliminary commands too).

** First some data definitons .
get file=!path2+"ghs95.sav".
compute female=(sex=1).
compute wk30p=workhrs.
recode wk30p (1 thru 30=0) (31 thru hi=1) (else=-999).
missing values wk30p (-999).
compute age2=age**2 .
descriptives var=workhrs wk30p age age2 female.

** Now the include file .

include file=!path8+"regressions.sps".

** Now some regressions - look at the contents of the include file to see what's going on.

regres1 dep=workhrs expl={female age age2} .

regres3 dep=wk30p expl={female age age2}.

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




**************************************************.
*** Segment 15.6) Very Useful macros (1) - generic Path names  .
**************************************************.


** This macro allows you to put a flexible 'path' definintion (file locations)
**   into your working files - a very useful macro if you change between different
**   computers from time to time.

** It proceeds in two steps :.

*** Step 1) - define the path that you want to use.  .

define !path7 () 'd:\data\lda\' !enddefine.

**** Step 2).

** You can now run SPSS commands substituting the path argument for the full text :.

get file=!path7+"ghs95.sav".
descriptives var=all.


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



**************************************************.
*** Segment 15.7) Very Useful macros (2) - defining dummy variables  .
**************************************************.

** This macro is a shortcut for computing new dummy variable names.
** it is defined within an include file.
```

```
include file=!path8+"makedummyvars.sps".

get file=!path2+"ghs95.sav".

fre var=marstat.

** Use the macro :.

factor marstat mst 7 .

** This will create 7 dummy variables for each category of marstat.

fre var=mst1 to mst7.
** (but you'll have to add value/variable labels yourself..).


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


*****************************************************.
*** Segment 15.8) Very Useful macros (3) - transferring value labels .
*****************************************************.


** This macro can be used to transfer value labels from one variable to a new one.
** Again, it's defined through an include file.
** It does require that you have write access to a directory called 'c:\temp'.



include file=!path8+"varlabstonew.sps" .


get file=!path2+"ghs95.sav".
fre var=marstat.
compute mar2=marstat.
fre var=mar2.

!applab marstat tovars mar2 .
fre var=mar2.
* (the online version of this particular macro won't work if you can't write to the folder
*    c:\temp\ - you will need to edit the folder in the macro to get this to work properly).

compute mar3=marstat.
compute mar4=marstat.
fre var=mar3 mar4.
!applab marstat tovars mar3 mar4.
fre var=mar3 mar4.

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



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




*****************************************************************.
******* EXERCISE 16 - HANDLING DATA: MATCHING INFORMATION BETWEEN DATA FILES .
*****************************************************************.


*****************************************************.
*****************************************************.
*** Segment 16.1) Adding files together .
*****************************************************.


** i) BHPS SMOKING DATA .
** This segment adds together data on cigarette smoking from 4 BHPS files ,
**   treating it in this case as if from unrelated repeated cross-sections .
```



```
get file=!path3+"iindresp.sav".
fre var=incigs .
missing values incigs (lo thru -1).
compute year=1999.
sav out=!path9+"mtch1.sav" /keep=year incigs /rename(incigs=ncigs) .
* (this saves out just the two variables from this individual level file in 1999,
**   to a temporary file in your 'temp' directory) .

** Repeat this on the other years :.
get file=!path3+"jindresp.sav".
fre var=jncigs .
missing values jncigs (lo thru -1).
compute year=2000.
sav out=!path9+"mtch2.sav" /keep=year jncigs /rename(jncigs=ncigs) .
get file=!path3+"kindresp.sav".
fre var=kncigs .
missing values kncigs (lo thru -1).
compute year=2001.
sav out=!path9+"mtch3.sav" /keep=year kncigs /rename(kncigs=ncigs) .
get file=!path3+"lindresp.sav".
fre var=lncigs .
missing values lncigs (lo thru -1).
compute year=2002.
sav out=!path9+"mtch4.sav" /keep=year lncigs /rename(lncigs=ncigs) .

** Now add all these files together :.
add files file=!path9+"mtch1.sav" /file=!path9+"mtch2.sav"
        /file=!path9+"mtch3.sav" /file=!path9+"mtch4.sav" .
descriptives var=all.
* We've now added the data together : we have one long file with records from
*   each year on top of each other  .

** Here's some illustrative analyses:.
fre var=year.
descriptives var=ncigs.
means tables=ncigs by year.
graph /errorbar(cI 95)=ncigs by year.


** Comment
* You'd typically want to control for other possible characteristics of the individuals involved
*  -  indeed, the BHPS includes an uneven new intake in 2002 (from N Ireland), so the apparent
*   trend at 2002 could be an artefact of this (check it simply by conrolling for region, age, etc).




** ii) SCOTTISH SOCIAL ATTITUDES SURVEY DATA .
** This segment adds together data on attitudes to Scotland from four years of
**   the Scottish Social Attitudes survey, 1999-2002.

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 : .
```

```
add files file=!path9+"m1.sav" /file=!path9+"m2.sav"
     /file=!path9+"m3.sav"  /file=!path9+"m4.sav"  .
fre var=year.
graph /bar=mean(scotpar2) by year by rsex
     /title="Opposition to independent / devolved Scotland" .
graph /errorbar=scotpar2 by year by rsex
    /title="Opposition to independent / devolved Scotland" .



** 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
** Often, an analyst would compute their own harmonised variables
*  using the same variable names across years.


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



******************************************************.
******************************************************.
*** Segment 16.2) Case to case matching  .
******************************************************.


***********************.
**** Seg16.2(i) - BHPS, connect 2002 occupation with 1999 occupation .
***********************.

** This example uses the BHPS 1999 and 2002 individual files, to match on a
** 1999 pattern to a 2002 pattern.

get file=!path3+"lindresp.sav".

fre var=lhlghq1 ljbsat  .
missing values lhlghq1 ljbsat (lo thru -1).
correlate var=lhlghq1 ljbsat .
** so job satisfaction has a modest correlation with sadness scale.

** Question - how much of 2002 sadness is due to 1999 sadness? .

** To match the 1999 data, we must :.
**     - extract the 1999 data to a temporary file
*      - identify and sort by a case identifier
*      - match it back onto the 2002 data.


get file=!path3+"iindresp.sav".

fre var=ihlghq1   .
missing values ihlghq1  (lo thru -1).
variable label ihlghq1 "Sadness in 1999".
sort cases by pid.
sav out=!path9+"mtch1.sav" /keep=pid ihlghq1 .

get file=!path3+"lindresp.sav".
sort cases by pid.
match files file=* /in=pres02 /file=!path9+"mtch1.sav" /in=pres99 /by=pid .

* (Comment: the  '*' here means 'this file', and the 'in' commands create new
*    variables which indicate sample presence.

descriptives var=all.
* Note that the 1999 data has been added to the end of the 02 file.
fre var=pres02 pres99.
select if (pres02=1).
descriptives var=all.
( (this restricts our analysis to the 16599 individuals sampled in 2002).

variable label lhlghq1 "Sadness in 2002".
fre var=ihlghq1 lhlghq1 ljbsat  .
missing values lhlghq1 ljbsat (lo thru -1).
correlate var=ihlghq1 lhlghq1 ljbsat .
```



```
* Partial correlation : job satisfaction by 2002 sadness, adjusted for 1999 sadness .
partial correlation var=lhlghq1 ljbsat by ihlghq1  .

** Concl: a correlation in 2002 of -0.27 between job satisfaction and sadness,
**  is _not mediated at all_ by any mutual correlation with 1999 sadness.


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


***********************.
**** Seg2(ii) - BHPS annual employment status records  .
***********************.


** 2002 employment status and employment class, matched onto 2001 file.

** 2002 info :.
get file=!path3+"lindresp.sav".
sort cases by pid.
fre var=ljbstat ljbstat .
sav out=!path9+"mtch1.sav" /keep=pid ljbstat ljbgold .
** Then open up 2001 file and match in 2002 temporary file.
get file=!path3+"kindresp.sav".
sort cases by pid.
match files file=* /in=pres01 /file=!path9+"mtch1.sav" /in=pres02 /by=pid.
fre var=pres01 pres02 .
cro pres01 by pres02.
* pres01 asnd pres02 indicate how many cases came in from the two years respectively.
* The panel variables: l prefix for wave 12 = 2002; k prefix for wave k = 11 = 2001 .
* 15450 present in both waves;
* 3417 of the 2001 respondents not present in 2002 (dropouts);
* 1147 of the 2002 respondents were not present in 2001 (new entries).

* Transitions in job status or job class?.
variable label ljbstat "Emploment status 2002" /kjbstat "Employment status 2001"
        /ljbgold "Job class 2002" /kjbgold "Job class 2001" .
cro kjbstat by ljbstat.
cro kjbgold by ljbgold.

* Note: when panel data is used, these sort of crosstabulations are often called 'transition
*   matrices'. One-to-one matching is also often used on cross-sectional data, eg if the original
*  data file was too big so different variables were stored separately on different files.


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



***********************.
**** Seg16.2(iii) - BHPS 4 waves match .
***********************.

** This example matches data from BHPS 1999-2003 on cigarette smoking to
**   _the same people_ over time.
**  Note this is a 'wide file' match rather than the long file match of segment 16.1.

** File peparation is similar, but:
**   - cig smoking info is kept in separate variables .
**   - we need to save the personal identifier variable 'pid'.
**   - the data in every individual file must be sorted by the identifier 'pid'
**   - we'll get an 'unbalanced' match - people aren't all in every year, and
**       we compute indicator variables in the 'match' command to show us this.


get file=!path3+"iindresp.sav".
fre var=incigs .
missing values incigs (lo thru -1).
sort cases by pid.
sav out=!path9+"mtch1.sav" /keep=pid incigs  .
get file=!path3+"jindresp.sav".
fre var=jncigs .
missing values jncigs (lo thru -1).
sort cases by pid.
sav out=!path9+"mtch2.sav" /keep=pid jncigs  .
get file=!path3+"kindresp.sav".
fre var=kncigs .
```

```
missing values kncigs (lo thru -1).
sort cases by pid.
sav out=!path9+"mtch3.sav" /keep=pid kncigs  .
get file=!path3+"lindresp.sav".
fre var=lncigs .
missing values lncigs (lo thru -1).
sort cases by pid.
sav out=!path9+"mtch4.sav" /keep=pid lncigs  .


** Now merge all these together.
match files file=!path9+"mtch1.sav" /in=w1 /file=!path9+"mtch2.sav" /in=w2
          /file=!path9+"mtch3.sav" /in=w3 /file=!path9+"mtch4.sav" /in=w4 /by=pid .
* (The 'in' command creates indicator variables to show if any cases came in at each wave).

descriptives var=all.
fre var=w1 w2 w3 w4 .
* (These n's show how many came in from 1999, 2000, 2001 and 2001 respectively).
* there are 22348 cases in total (in any of the 4 waves), but each year only about 16k .


** Evidence on trends (cf segment 1).
summarize variables= incigs jncigs kncigs lncigs /cells=mean median stddev count .
examine variables= incigs jncigs kncigs lncigs /plot=boxplot .
graph /bar=mean(incigs) mean(jncigs) mean(kncigs) mean(lncigs)
          median(incigs) median(jncigs) median(kncigs) median(lncigs) /missing=variable .
graph /line=mean(incigs jncigs kncigs lncigs ) /missing=variable .
graph /line=stddev(incigs jncigs kncigs lncigs ) /missing=variable .
* So arithmetic mean falls but stddev rises over period .

** Evidence on individual variations :.
correlate var=incigs jncigs kncigs lncigs .
* Seems to be evidence of individual's changing : weakening of correlation over time.

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




*******************************************************.
*******************************************************.
*** Segment 16.3) Table distribution ('one-to-many' matching)  .
*******************************************************.


** Example : BHPS household level file information linked to BHPS
**                individual level files .


** BHPS files from 2002 for individuals and households.

get file=!path3+"lindresp.sav".
sort cases by lhid pid .
descriptives var=pid lhid.
* 16597 adult individual interviews in the 2002 BHPS.
* Adults themselves are clustered into households identified by the
* wave specific wave 12 household identifier 'lhid'.
* ie, same lhid different pid = different people in the same household.
* (in fact, these individuals come from 9346 different households, though it's not
*   easy to instantly get that imformation out of SPSS).
sav out=!path9+"mtch1.sav" /keep=lhid pid lsex lage ljbstat lvote.


get file=!path3+"lhhresp.sav" .
descriptives var=lhid.
* This file is at the household level - each case corresponds to a single household in the
*   2002 BHPS (they come from one of the BHPS's source household level files, 'hhresp').
*  There are 9352 households with information in this sample.
* (Comment - there are 6 households on the household file with no entries on the individual
*    file, because all the relevant information for individuals in these units was missing).
fre var=lhhtype.
* (Just one example of the household level variables saved on this file).

* If we want to match that household level information to the individuals, we need
*  to save out using the household level identifier.
```



```
sort cases by lhid.
sav out=!path9+"mtch2.sav" /keep=lhid lhhtype.

match files file=!path9+"mtch1.sav"
          /table=!path9+"mtch2.sav" /by=lhid.
* (for a one-to-many match, you match a 'file' with many cases per category, against a 'table'
*   where only one case per category ).
descriptives var=all.
fre var=lhhtype.
* The hhtype information has been successfully distributed to all individuals
* in the individual level sample.

graph /bar=pct by lvote by lsex .
graph /bar(stacked)=pct by lvote by lsex .
graph /bar=mean(lvote) by ljbstat by lsex .


** Extension : number of units at different levels .
* The syntax below is one quick way to find the number of higher level units in
*    a lower level file, for instance the number of households in an indivual level file :.
sort cases by lhid pid.
compute hhld=1.
if (lag(lhid)=lhid) hhld=0.
fre var=hhld.
* values with hhld=1 are the 'first' individuals in each household - so there are 9346
*   household in the wave 12 individual level file.
***.



*******************************************************.
*******************************************************.
*** Segment 16.4) Aggregating (or 'many-to-one' summarising)  .
*******************************************************.


************************************.
*** Seg16.4(i) BHPS files from 2002 for individuals:
***              calculating variables for (i) regional average ages
*                                            and (ii) household oldest age .
************************.


get file=!path3+"lindresp.sav"
      /keep=pid lhid lage lsex lregion .
descriptives var=all.
* Select out only cases with valid age information and known region ('listwise deletion').
select if (lage ge 16 & lregion gt 0).
fre var=lregion lage.
* (sample is 16325 adults aged 16+).

means tables=lage by lregion /statistics=anova .
graph /errobar(CI 95)=lage by lregion.
* Summarises average adult ages different by regions - there are some modest differences .


** Now create a file containing mean ages by region.
* ('aggregate' calculates the value and 'match' brings it back into the data).
sort cases by lregion.
aggregate outfile=!path9+"agg1.sav" /break=lregion /regage=mean(lage).


** Also, create a file containing oldest ages by household.
* ('aggregate' calculates the value and 'match' brings it back into the data).
sort cases by lhid pid.
aggregate outfile=!path9+"agg2.sav" /break=lhid /hhage=max(lage).

** Check the contents of the files.
get file=!path9+"agg1.sav".
fre var=regage.
* (only a small file: each case is a different region).
get file=!path9+"agg2.sav".
descriptives var=hhage.
* (a much larger file: each case is a household).


** Re-open the original individual level file and match back information
```

```
*       from the aggregated statistics.
get file=!path3+"lindresp.sav"
       /keep=pid lhid lage lsex lregion .
select if (lage ge 16 & lregion gt 0).
descriptives var=all.

* Retrieve the regional average ages by a 'one-to-many' match.
sort cases by lregion.
match files file=* /table=!path9+"agg1.sav" /by=lregion.
variable label regage "Average age of adults in GOR region".
means table=lage regage by lregion.
* note means are identical, but standard dev of regage is zero.

** Highest age for household.
sort cases by lhid pid.
match files file=* /table=!path9+"agg2.sav" /by=lhid.
variable label hhage "Age of oldest person in household".
graph /scatterplot=lage with hhage.
means tables=lage hhage by lregion.
graph /errorbar(CI 95)=lage hhage by lregion.
* Observe - Pattern of regional differences is broadly the same whichever age we look at.


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



******************************************.
**** Seg 16.4(ii) - Highest income earner in a household .
******************************************.

get file=!path3+"lindresp.sav" .

** Say we're interested in the population of married women who's economic
**   activitiy is given as 'family care' - ie 'housewives'.

fre var=ljbstat lsex lmastat .
select if (lsex=2 & lmastat=1 &  ljbstat=6).
fre var=lsex ljbstat lvote .

** Income to education level :.
descriptives var=lfimn.
graph /histogram=lfimn.
compute lninc=-999.
if (lfimn gt 0) lninc=ln(lfimn).
missing values lninc (-999).
graph /histogram=lninc.
graph /errorbar(CI 95)=lninc by lqfedhi .

** We'd expect education to influence income, but we hardly see any pattern at all.

** Return to the household and get the sum of all incomes and the sum of _male_ incomes
*   in the household (recode missing data to '0' income) : .

get file=!path3+"lindresp.sav" .
descriptives var=lfimn .
recode lfimn (lo thru 0=0).
missing values lfimn (-777).
compute male=(lsex=1).
compute incmal=lfimn*male.

sort cases by lhid pid.
aggregate outfile=!path9+"mtch1.sav" /break=lhid
      / totinc=sum(lfimn) /malinc=sum(incmal).

* Now match this back to the original data :.

get file=!path3+"lindresp.sav" .

select if (lsex=2 & lmastat=1 & ljbstat=6).
fre var=lsex ljbstat lvote .

sort cases by lhid pid.
match files file=* /table=!path9+"mtch1.sav" /by=lhid.
descriptives var=all.
compute lninc=-999.
if (lfimn gt 0) lninc=ln(lfimn).
```



```
compute lnhhinc=-999.
if (totinc gt 0) lnhhinc=ln(totinc).
compute lnmalinc=-999.
if (malinc gt 0) lnmalinc=ln(malinc).
missing values lninc lnhhinc lnmalinc (-999).
graph /histogram=lninc.
graph /histogram=lnhhinc.
graph /histogram=lnmalinc.
missing values lqfedhi (lo thru 0).
graph /errorbar(CI 95)=lninc lnhhinc lnmalinc by lqfedhi .

** Comment - we see closer to the pattern we might expect for household
*     or for male income (though still a lot of mixture)  .



******************************************************.
******************************************************.
*** Segment 16.5) Relationships between cases matching   .
******************************************************.

** There are lots of circumstances where you might want to link together
** information from different cases in a sample .
** Sticking with the BHPS, this example shows how you could link information on
**  an individual with their spouse if present .
** Many surveys include information allowing you to link data between cases, whilst
*    many also include additional derived variables where some basic related cases information
*    has already been added in .



** BHPS files from 2002 for individuals .
** The extract2 file adds a variable isppid which is needed to link information
**   on different spouses.

get file=!path3+"lindresp.sav"
     /keep=pid lhid lsex lage ljbstat lsppid.
* The variable lsppid is the personal identifier number of the spouse, if present.
* Deal only with those who give a pid for a spouse.
* In addition, in order to distinguish male and female spouses, split by gender :.
select if (lsppid gt 0 & lsex=2).
descriptives var=all.
sort cases by lsppid.
sav out=!path9+"wif1.sav" /keep=lsppid lsex lage ljbstat
         /rename(lsppid lsex lage ljbstat=pid wifsex wifage wifjbst) .

* Match the women's data back into the original, for men only :.
get file=!path3+"lindresp.sav"
     /keep=pid lhid lsex lage ljbstat lsppid.
select if (lsppid gt 0 & lsex=1).
sort cases by pid.
match files file=* /in=hus /file=!path9+"wif1.sav" /in=wif /by=pid .
fre var=hus wif.
cro hus by wif.
** Most of the non 1-1 cases are males or females whose spouse didn't give
**   an interview in the individual file, for whatever reason .
* However, any homosexual spouses will be lost in this missing data here.
select if (hus=1 & wif=1).
descriptives var=all.


* Some illustrative analyses :.

variable label lsex "Husband sex" /wifsex "Wife's sex"
              /lage "Husband age" /wifage "Wife's age"
                /ljbstat "Husband economic activity" /wifjbst "Wife's economic activity" .

cro ljbstat by wifjbst.

graph /scatterplot=lage with wifage  .
* (some non-overlap suggests gender imbalance in age differences).
compute agedif=-999.
if (lage gt 16 & wifage gt 16) agedif=lage - wifage.
missing values agedif (-999).
```

```
graph / histogram=agedif .
* Average gap in ages between spouses is men 2.3 years older than their spouses.


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









*****************************************************************.
*******  EXERCISE 17 - SELECTED ISSUES IN ANALYSING COMPLEX SURVEY DATA.
*****************************************************************.


******************************************************.
*** Segment 17.1) Illustrating Multiple response questions .
******************************************************.

*** For this example, we'll use the Scottish Social Attitudes extract available from
**     the UK Data Archive (2002 version) .


import file=!path4+"ssa02.por".


********************.
**** Seg17(i) Multiple response groups :.
********************.


fre var=impnhs1 impnhs2.
** These two variables are interesting, but what if you wanted to know the total
*     number who mentioned each improvement in either question?.
** The solution is to treat the questions as a 'multiple response group' :.

** SPSS's inbuilt command :.
mult response groups=imprv (impnhs1 impnhs2 (1,9))
           /frequencies=imprv.
* Shows for instance that 43.0% of respondents mentioned category 5 as one of their two mentions.
* Often this format is more helpful when there are quite a lot of optional mentions.


** Extension : Crosstabulation by gender :.
mult response groups=imprv (impnhs1 impnhs2 (1,9))
            /variables=rsex(1,2)
           /tables=imprv by rsex.


** However, it is ususally more convenient to use available multiple response
**   options within the 'tables' command :.


tables /format blank missing ('.') /ftotal=ftot1 "Total" ftot2 "All"
    /mrgroup=mrvar " " impnhs1 impnhs2
     /tables  mrvar + ftot1
  /statistics count ((F5.0) ' ') cpct (mrvar (pct3.1) ' ' )
   /title= "Scottish Social Attitudes 2002:" "Improvements wanted for health service (multiple
responses)"    .



tables /format blank missing ('.') /ftotal=ftot1 "Total" ftot2 "All"
    /mrgroup=mrvar " " impnhs1 impnhs2
     /tables  mrvar + ftot1 by  rsex + ftot2
  /statistics count ((F5.0) ' ') cpct (mrvar (pct3.1) ' ':rsex)
   /title= "Scottish Social Attitudes 2002:" "Improvements wanted for health service (multiple
responses), by gender"    .

* (Note: you get greater control over the output format with the 'tables' command -
**  but it does have lots of confusing options).
```



```
********************.
**** Seg17.1(ii) Multiple response dichotomies :.
********************.

** A lot of data can be regarded as multiple response dichotomies, even if not originally
**  intended as such :.

fre var=j1sch j1work j1shops j1leis j1famfr j1pleas .
fre var=edqual1 to edqual16 .
* On both variables, the -1 inapplicable is actually more appropriate as a 'not mentioned'.
recode j1sch j1work j1shops j1leis j1famfr j1pleas edqual1 to edqual16
       (1 thru 40=1) (-1=0) (lo thru -2, 99=-999).
missing values j1sch j1work j1shops j1leis j1famfr j1pleas edqual1 to edqual16 (-999).
fre var=j1sch j1work j1shops j1leis j1famfr j1pleas  edqual1 to edqual16 .


** SPSS 'mult response' for multiple dichotomies :.
mult response groups=travel(j1sch j1work j1shops j1leis j1famfr j1pleas (1) )
      /fre=travel .
* The percentages show only the proportions from those giving a 1 anywhere:.
* Eg, 66% of people who sometimes drive, mentioned using their car to visit family / friends.


** To get a contrast with those not mentioning any, we need to use a 'none' indicator variable : .
compute j1notrv=max(j1sch, j1work, j1shops, j1leis, j1famfr, j1pleas).
recode j1notrv (1=0) (0=1).
variable label j1notrv "Doesn't use car for any of above".
fre var=j1notrv.
* (this computation has value 1 if all of the other j1's = 0, and 0 otherwise).
mult response groups=travel2(j1sch j1work j1shops j1leis j1famfr j1pleas j1notrv (1) )
      /fre=travel2 .
* So - 56% of all people use a car at some point for visiting relatives .

**** Same for educational qualifications :.
mult response groups=educ(edqual1 to edqual16 (1) )
      /fre=educ .
* Again, the percentages currently show only the proportions from those giving a 1 anywhere.
* Eg, 54% of people with any qualification have an o-level.

compute noqual=max(edqual1 to edqual16 ).
recode noqual (1=0) (0=1).
variable label noqual "Has none of above qualifications".
fre var=noqual.
* (this computation has value 1 if all of the other educ's = 0, and 0 otherwise).

mult response groups=educ2(edqual1 to edqual16 noqual (1) )
      /fre=educ2 .
** Eg, 38% of SSA adults have an O-level or equivalent pass.


**** As previously, we can also use the 'tables' command for multiple dichotomies :.
tables /format blank missing ('.') /ftotal=ftot1 "Total" ftot2 "All"
    /mdgroup=mdvar " " j1sch j1work j1shops j1leis j1famfr j1pleas j1notrv (1)
     /tables  mdvar + ftot1 by  rsex > urbanac
  /statistics count ((F5.0) ' ') cpct (mdvar (pct3.1) ' ':rsex,urbanac)
   /title= "Scottish Social Attitudes:" "Use of car for travelling (multiple response), by gender
and location"    .
*****.



******************************************************.
*** Segment 17.2) A large sample with lots of variables : heavily parameterised regressions.
******************************************************.

** Use the BHPS 2002 sample :.

get file=!path3+"lindresp.sav".

** Some data management to set up the first model :.


** Outcome variable :.
```

```
graph /histogram=lfimn .
compute lninc=-999.
if (lfimn ge 100 & lfimn le 50000) lninc=ln(lfimn).
missing values lninc (-999).
graph /histogram=lninc.


** Explanatory variables (plus reviews of the relations to income):.
fre var=lsex lage lmastat lrace lqfedhi ljbstat ljbrgsc ltenure lhlghq1 lregion  .

compute female=(lsex=2).
examine variables lninc by lsex /nototal /plot=boxplot .
graph /errorbar(CI 95)=lninc by lsex .

compute age2=lage**2.
graph /scatterplot=lage with lninc /title="BHPS 2002 : Age by average monthly income (logged)".
graph /scatterplot(matrix)=lninc lage age2 .
* (Tip : use 'properties' and 'point bins' to get an image of the relation from age to income).
* Another visualisation :.
sort cases by lsex.
split files by lsex.
examine variables lninc by lage /nototal /plot=boxplot /statistics=none.
split files off.
* (This gives a clustered graph but one that shows income by age.


compute cohab=(lmastat=1 | lmastat=2).
graph /errorbar=lninc by cohab by lsex .

graph /errorbar=lninc by lrace .
compute carib=(lrace=2).
compute indn=(lrace=5).
compute pkban=(lrace=6 | lrace=7).
compute otheth=(lrace=3 | lrace=4 | lrace=8 | lrace=9).

* Comment: on cohab and the ethnicity indicators, we use null imputation of missing values.
* Comment: a problem with the lrace data - its only present for new people to w12.

graph /errorbar=lninc by lqfedhi .
missing values lqfedhi (lo thru 0).
compute hied=(lqfedhi=1 | lqfedhi=2 | lqfedhi=3 | lqfedhi=4).
compute voced=(lqfedhi=5 | lqfedhi=8 | lqfedhi=10).
compute loed=(lqfedhi=9 | lqfedhi=10 | lqfedhi=13).
fre var=hied voced loed .

cro ljbstat by ljbrgsc .
** We can derive a categorical scheme from this which is more or less exhaustive :.
compute ecstjb=-999.
if (ljbrgsc=1) ecstjb=1.
if (ljbrgsc=2) ecstjb=2.
if (ljbrgsc=3) ecstjb=3.
if (ljbrgsc=4) ecstjb=4.
if ( (ljbrgsc=3 | ljbrgsc=4) & (ljbstat=1)) ecstjb=5.
if (ljbrgsc=5 | ljbrgsc=6 | ljbrgsc=7) ecstjb=6.
if (ljbstat=3 | ljbstat=9) ecstjb=7.
if (ljbstat=4) ecstjb=8.
if (ljbstat=6) ecstjb=9.
if (ljbstat=7) ecstjb=10.
if (ljbstat=8) ecstjb=11.
add value labels ecstjb
   1 "Professional" 2 "Managerial / technical" 3 "Skilled non-manual (employee)"
    4 "Skilled-manual (employee)" 5 "Self-employed skilled" 6 "Partly or unskilled or armed forces"
     7 "Unemployed" 8 "Retired" 9 "Housewife" 10 "Student" 11 "LT sick or disabled" .
missing values ecstjb (-999).
fre var=ecstjb .
graph /errorbar(CI 95)=lninc by ecstjb by lsex .
compute prof=(ecstjb=1).
compute manag=(ecstjb=2).
compute sknm=(ecstjb=3).
compute skmn=(ecstjb=4).
compute sksemp=(ecstjb=5).
compute punsk=(ecstjb=6).
compute unemp=(ecstjb=7).
compute retir=(ecstjb=8).
compute hswif=(ecstjb=9).
compute studnt=(ecstjb=10).
compute ltsdsb=(ecstjb=11).
```



```
fre var=prof manag sknm skmn sksemp punsk unemp retir hswif studnt ltsdsb .

fre var = ltenure.
compute hhown=(ltenure=1 | ltenure=2).
compute hhshs=(ltenure=3 | ltenure=4).
fre var=ltenure hhown hhshs .
* (ie, null imputation missing values).

missing values lhlghq1 (lo thru -1) .
graph /histogram=lhlghq1 .
correlate var=lhlghq1 with lninc .

graph /errorbar=lninc by lregion.
compute inlond=(lregion=1).
compute outlond=(lregion=2).
compute esouth=(lregion=3 | lregion=4 | lregion=5).
compute wales=(lregion=17).
compute scotl=(lregion=18).
compute nirel=(lregion=19).
fre var=inlond outlond esouth wales scotl nirel .


** All the available data :.
descriptives var= lninc
          female lage age2 cohab hied voced loed  carib indn pkban otheth lhlghq1
          prof manag sknm skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel  .



** Model 1: heavily parameterised main effects multiple regression :.

regression var=lninc
          female lage age2 cohab hied voced loed carib indn pkban otheth lhlghq1
          prof manag  skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel
    /statistics=R coeffs anova  /dependent=lninc   /method=enter .



** Model 2: main effects model with multicollinearity checked :.

regression var=lninc
          female lage age2 cohab hied voced loed carib indn pkban otheth lhlghq1
          prof manag  skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel
    /statistics=R coeffs anova tol collin  /dependent=lninc   /method=enter .
** The main possible issue concerns age, age2 and retired.
** Primary option would be to recode age to a categorical representation (cf lab 4)
**   but in this instance it's probably ok to carry on .



**********.
** Model 3: selected interaction effects tested :.

descriptives var= lninc
          female lage age2 cohab hied voced loed  carib indn pkban otheth lhlghq1
          prof manag sknm skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel  .


** Compute interactions simply by multiplying together terms :.

** Selected interactions with gender :.
compute femage=female*lage.
compute femcoh=female*cohab.
compute femhied=female*hied.
compute femvoced=female*voced.
compute femloed=female*loed.
compute femcar=female*carib.
compute femind=female*indn.
compute fempkb=female*pkban.
compute femothe=female*otheth.
compute femprof=female*prof.
compute femman=female*manag.
compute femsknm=female*sknm.
```

```
compute femskmn=female*skmn.
compute femsksp=female*sksemp.
compute feminl=female*inlond.
compute femoutl=female*outlond.
compute femeso=female*esouth.
compute femwal=female*wales.
compute femscot=female*scotl.
compute femnirl=female*nirel.

fre var=femage femcoh femhied femvoced femloed femcar femind fempkb femothe
         femprof femman femsknm femskmn femsksp feminl femoutl femeso femwal femscot femnirl.

** Selected interactions with education level :.
compute hiedinl=hied*inlond.
compute hiedolo=hied*outlond.
compute hiedeso=hied*esouth.
compute hiedwal=hied*wales.
compute hiedsco=hied*scotl.
compute hiednir=hied*nirel.
compute loedinl=loed*inlond.
compute loedolo=loed*outlond.
compute loedeso=loed*esouth.
compute loedwal=loed*wales.
compute loedsco=loed*scotl.
compute loednir=loed*nirel.


fre var=hiedinl hiedolo hiedeso hiedwal hiedsco hiednir
                loedinl loedolo loedeso loedwal loedsco loednir .

** Comment : adding interaction terms can also be quite long winded! .
* (see the lab 4 part 4 syntax for an SPSS macro shortcut on this).


** Regression with all of these :.

regression var=lninc
          female lage age2 cohab hied voced loed carib indn pkban otheth lhlghq1
          prof manag  skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel
           femage femcoh femhied femvoced femloed femcar femind fempkb femothe
         femprof femman femsknm femskmn femsksp feminl femoutl femeso femwal femscot femnirl
            hiedinl hiedolo hiedeso hiedwal hiedsco hiednir
                loedinl loedolo loedeso loedwal loedsco loednir
    /statistics=R coeffs anova tol  /dependent=lninc   /method=enter .


** Comment : model explanation has gone up a bit, though there's lots of insignificant terms
**    too - they would usually be removed.




** Also: there is a danger here of 'overfitting' the data, ie, estimating a model coefficient
*    for a variable that effectively only refers to a couple of cases (ie, you're not summarising
emprical
*    regularities, but case specific quirks).
* Solution is to not use interactions (or main effects) for very sparse variables.
descriptives var= lninc
          female lage age2 cohab hied voced loed  carib indn pkban otheth lhlghq1
          prof manag sknm skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel
         femage femcoh femhied femvoced femloed femcar femind fempkb femothe
         femprof femman femsknm femskmn femsksp feminl femoutl femeso femwal femscot femnirl
            hiedinl hiedolo hiedeso hiedwal hiedsco hiednir
                loedinl loedolo loedeso loedwal loedsco loednir   .

** eg, loedinl and loedolo should go.



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

** Model 4 : Moving to a more parsimonious structure :.
```



```
** It's more satisfactory to choose a parsimonious model interactively, but
**  you can also use an automated routine :.

regression var=lninc
          female lage age2 cohab hied voced loed carib indn pkban otheth lhlghq1
          prof manag  skmn sksemp punsk unemp retir hswif studnt ltsdsb
          hhown hhshs  inlond outlond esouth wales scotl nirel
           femage femcoh femhied femvoced femloed femcar femind fempkb femothe
         femprof femman femsknm femskmn femsksp feminl femoutl femeso femwal femscot femnirl
            hiedinl hiedolo hiedeso hiedwal hiedsco hiednir
                loedinl loedolo loedeso loedwal loedsco loednir
    /statistics=R coeffs anova tol  /dependent=lninc   /method=stepwise .

** Note the long and complex outputs - at the end of the 'variables included' table is the
automatically
** preferred model :.

** Here's an alternative chosen for substantive reasons :.
regression var=lninc
          female lage age2 cohab hied  loed carib indn pkban
          prof manag  sksemp punsk unemp retir hswif studnt ltsdsb
          inlond outlond esouth wales scotl nirel
           femage femcoh femhied  femcar
         femprof     femeso femwal femscot femnirl
             hiedwal hiedsco hiednir
                 loedwal loedsco loednir
    /statistics=R coeffs anova tol  /dependent=lninc   /method=enter .


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


****************************************.
**** Hierarchical data structure : .

******** .
** Model 5: Hierarchical fixed effects tested : region, and selected features of the household.

** One hiearchical fixed effect that we've already used is the region dummy variable indicators.
** Another would be specific information on the household :.
sort cases by lhid pid.
aggregate outfile=!path9+"mtch1.sav" /break=lhid /hhied=max(hied) /havage=mean(lage).
match files file=* /table=!path9+"mtch1.sav" /by=lhid.

regression var=lninc
          female lage age2 cohab hied  loed carib indn pkban
          prof manag  sksemp punsk unemp retir hswif studnt ltsdsb
          inlond outlond esouth wales scotl nirel
           femage femcoh femhied  femcar
         femprof     femeso femwal femscot femnirl
             hiedwal hiedsco hiednir
                 loedwal loedsco loednir
                  hhied havage
    /statistics=R coeffs anova tol  /dependent=lninc   /method=enter .

* Both the household effects, and some of the region effects, are making a difference.


****************************.
** Model 6: Hierarchical random effects tested : .

** The other sort of hierarchical effect to consider is a 'random' effect - basically this accounts
*   for undefinable 'similarity' between cases within a hierarchical cluster.
** The model options are ultimately quite complicated ('multilevel models'), but
** SPSS does allow for one relatively simple multilevel model regression specification :.


mixed lninc with female lage age2 cohab hied  loed carib indn pkban
          prof manag  sksemp punsk unemp retir hswif studnt ltsdsb
          inlond outlond esouth wales scotl nirel
           femage femcoh femhied  femcar
         femprof     femeso femwal femscot femnirl
             hiedwal hiedsco hiednir
                 loedwal loedsco loednir
                  hhied havage
        /criteria=cin(95) mxiter(100) mxstep(5) scoring(1) singular(0.000000000001)
        hconverge(0,absolute) lconverge(0,absolute) pconverge(0.000001, absolute)
```

```
          /fixed female lage age2 cohab hied  loed carib indn pkban
          prof manag  sksemp punsk unemp retir hswif studnt ltsdsb
          inlond outlond esouth wales scotl nirel
           femage femcoh femhied  femcar
         femprof     femeso femwal femscot femnirl
             hiedwal hiedsco hiednir
                 loedwal loedsco loednir
                  hhied havage
                   | sstype(3)
         /method=reml
         /print=corb solution  r
          /random=intercept | subject(lhid) covtype(ID)   .

** The main feature of this model is just that it 'corrects' the parameter estimates for any
*    hierarchical random effects clustering . It is only important to do this if the 'error
component'
*    associated with the clustering is of a substantial and significant magnitude - in fact,
*   in this example, it is not and thus the linear regression is probably adequate (see the
*    output 'estimates of covariance parameters : intercept').


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




*****************************************************************.
*******************************************************************************.
** Segment 17.3 : example applications of selected extension techniques :
*    -  Factor analysis and correspondence analysis
*****************************************************************************.


*****************.
*** Seg17.3(i) :  Factor analysis example .
*********************.


** Using the BHPS 1999 individual file :.

get file=!path3+"iindresp.sav".


** Factor analysis is used to look for structures in a number of metric variables.

* Some metrics in the data :.

compute csm2=ijbcssm.
recode csm2 (missing,sysmis,lo thru 0 = -999).
compute age2=iage.
recode age2 (missing,sysmis,lo thru 15=-999).
compute lninc=-999 .
if (ifimn gt 100 & ifimn lt 10000) lninc=ln(ifimn).
compute wght=ixrwght.
recode wght (missing,sysmis=-999).
* (wght can be used as a variable indicating sample under representation).
fre var=incigs.
compute cigs=incigs.
recode cigs (-9,-7=-999) (-8=0).
fre var=ihlghq1.
compute ghq=ihlghq1.
recode ghq (lo thru -1=-999) .

missing values csm2 age2 lninc wght cigs ghq (-999).
descriptives var=csm2 age2 lninc wght cigs ghq .

correlate var=csm2 age2 lninc wght cigs ghq .


** Factor analysis to explore all of these for the currently working population.

factor /variables csm2 age2 lninc wght cigs ghq  /missing listwise
        /analysis csm2 age2 lninc wght cigs ghq
       /print initial extraction /criteria mineigen(1) iterate(25)
```



```
       /extraction pc /rotation norotate /method=correlation   .

* Not a clear pattern but can interpret :.

** Factor 1 :  difference between those with higher camscore and income, higher age,
*   less smoking, under-represented in the sample , versus the opposite.

** Factor 2 : difference between those with lower camscore, older still, smoke more,
*   under-represented in sample, versus the opposite.


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

*****************.
*** Seg17.3(ii) :  Correspondence analysis example .
*********************.


get file=!path3+"iindresp.sav".


** Correspondence analysis relating social class with voting preference?.
fre var=ijbgold ivote.
missing values ijbgold ivote (lo thru 0).

tables /format blank missing ('.') /ftotal=ftot1 "Total" /missing=exclude
  /tables (ivote) + ftot1 by  (ijbgold)
  /statistics count ((F5.0) ' Cases ').

cro ivote by ijbgold  /cells=count col /statistics=phi.

correspondence table=ijbgold (1,11) by ivote (1,11)
  /dimensions=2 /measure=chisq
  /standardize=rcmean /normalization=symmetrical
  /print=table rpoint cpoint  /plot=ndim(1,max) biplot(20).

* Comment: the first dimension is v influential 61%, the second also strong (18%).
* The dimension scores illustrate structure more clearly, but the graph illustrates
*  how patterns of voting and social class go together in those first 2 dimensions.
* Suggests that the main patterns of association often involve the smaller classes (esp farmers)
*  and smaller voting patterns also.

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




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




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


** EOF .




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

```
** EOF .
```