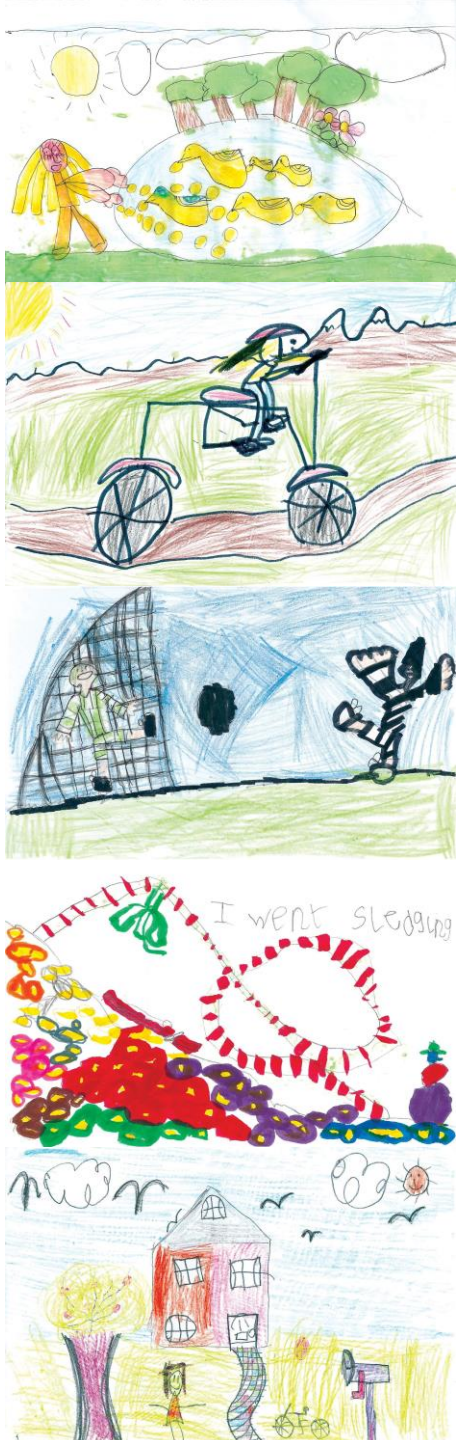
ScotCen
Social Research that works for society

Growing Up in Scotland

A practical overview of the data and documentation

Overview

- The datasets
- Common data manipulations
- Analysis using weights and stratification variables



The data

1.

GUS datasets

- Deposited as separate SPSS datasets for each sweep:
  - Sweep 1: Birth (BC1 & BC2) and Child cohorts
  - Sweep 2: Birth and Child cohorts
  - Sweep 3: Birth and Child cohorts
  - Sweep 4: Birth and Child cohorts
  - Sweep 5: Birth cohort only
  - Sweep 6: Birth cohort only
- Between 1100 and 2100 variables in each dataset including derived variables

ScotCen Social Research

Variable naming convention

1	2	3	4, 5 & 6	7 & 8
Source of data	Sweep	Key theme prefix	Sub-theme stem	Question / variable number
M (Main carer/adult interview) **P** (Partner interview) **D** (Derived variable) **DP** (derived from partner interview) **W** (Weights and heights) **AL** (Area level)	a = sweep 1 b = sweep 2 c = sweep 3 d = sweep 4 e = sweep 5	e.g. H = Health P = Parenting C = Childcare PS = Primary School	e.g. 'veg', 'SDQ', 'bed', 'lsi' = long-standing illness	

ScotCen Social Research

Variable name example

Variable name: McApha01 *(8 characters maximum)*		
Variable label: 'Mc - Child rode bicycle in last wk' *(shortened to 40 characters or less if possible)*		
Character 1	**M**	Indicates that the source of the data was the main carer interview
Character 2	**c**	Indicates that the data was collected at sweep 3
Character 3	**A**	Indicates that the variable concerns information around the general theme of 'Activities'
Character 4,5,6	**pha**	Reflects that the specific content of this variable relates to physical activity
Character 7,8	**01**	Denotes that this is the first question in this specific topic

ScotCen Social Research

Household Grid (HG)

This information is available for each household member (up to 15 people)

PersNo2	Mc - ID person 2
McHGsx2	Mc - Sex person 2
DcHGag2	Dc Age of person 2 at interview (years)
McHGmr2	Mc - Legal marital status person 2
McHGlv2	Mc - Whether living together as a couple - person 2
DcHGmr2	Dc - De facto marital status - person 2
McHGr21	Mc - Relationship of person 2 to study child
McHGr32	Mc - Relationship of person 3 to person 2
McHGr42	Mc - Relationship of person 4 to person 2
McHGr52	Mc - Relationship of person 5 to person 2

Household summary derived variables

DcHGnmad	Dc Number of adults (16 or over) in household
DcHGnmkd	Dc Number of children in household
DcHGnmsb	Dc - Number of siblings in household
DcHGhsiz	Dc Household size
DcHGrsp01	Dc - Whether respondent is natural mother
DcHGrsp02	Dc - Whether respondent is natural father
DcHGnp01	Dc - Number of natural parents in household
DcHGnp02	Dc - Natural mother in household
DcHGnp03	Dc - Natural father in household
DcHGnp04	Dc - Respondent living with spouse/partner
DcHGrsp07	Dc Who is the respondent in relation to the child
DcHGprim	Dc Whether child was mothers first-born
DcHGbord	Dc - Study child s birth order

Derived analysis variables

- Maternal age (d#hgmag2)
- Lone parent/Couple family (d#hgrsp04)
- Respondent NSSEC (d#msec01)
- Highest education level of respondent (d#medu01)
- Equivalised household income quintiles (d#eqv5)
- Ethnicity of respondent (d#meth07)

Derived variables syntax

- To summarise or combine other variables
- To replace questionnaire variables where risk to confidentiality (e.g. religion, ethnicity)

```
**BANDED VERSION OF MAIN CHILDCARE PROVIDER HOURS

Recode DcCman02 (1 thru 8=1) (9 thru 16=2) (17 thru 40 = 3) (41 thru hi=4) (Else = -3) into DcCman06.
Exe.
IF (DcCany02=2) DcCman06=-1.
Var labs DcCman06 'Dc Main ccare hours per week - Banded'.
Val labs DcCman06
 -3 'No information or less than an hr per wk'
 -1 'No childcare'
 1 'Up to 8 hours'
 2 'Between 9 and 16 hours'
 3 'Between 17 and 40 hours'
 4 'More than 40 hours'.
Missing values DcCman06 (-3,-1).
FORMATS DcCman06 (F2.0).
Execute.
```

Other useful variables

Area variables

ALeURin2	ALe - SG urban-rural classification
ALeSNim2	ALe - SIMD 2006 quintiles
ALeLow15	ALe - Flag lowest 15% datazones

Weighting variables

DcWTbrth	Dc Birth cohort Sw5 weight
DcWTchld	Dc Child cohort Sw3 weight
DecWTbth2	Dc Birth cohort Sw3 weight - longitudinal
DcWTchd2	Dc Child cohort Sw3 weight - longitudinal

Repeat and new data

Repeat data

- Same question asked at different sweeps
- To the same range (ex. 1) of cases or not (ex. 2)
- Example 1: Mb & McHgen01 asked to both cohorts
- Example 2: Sw3 McFesy01 asked to older cohort; Sw2 MbFesy01 asked to younger cohort
- **The detail of when specific variables were included can be determined from the variable list**

New data

- New questions introduced
- Example: Parent-child relationship (Pianta) questions at Sweep 5

Feed-forward data

Feed-forward data

- Information fed-forward from one sweep to the next
- Simply updated if:
  - Change of circumstances for same respondent, or
  - Different respondent

IF same respondent as last sweep [MeHGrsp03 = 1]

> **MeMedck1**
> Can I just check, have you gained any new qualifications since we last spoke to you in
> ^int_month last year?
> 1 Yes
> 2 No
>
> *IF gained new qualifications [MeMedck1 = 1]*
>>
>> **MeMedck2**
>> SHOWCARD L11 (card with list of school examinations)
>> Are any of those qualifications listed on this card?
>> 1 Yes
>> 2 No

Common data manipulations

2.

Merging datasets via menu

****See Handout Booklet****

•**Open dataset you want to merge new variables into: the '1st' dataset**
(example: Sweep 3 birth cohort)

2.Sort 1st dataset on 'IDnumber' in ascending order

3.Open dataset you want to extract the new variables from: the '2nd' dataset
(example: Sweep 2 BC to be added to Sweep 3 BC)

4.Sort 2nd dataset on 'IDnumber' in ascending order

5.Go back to 1st dataset and use menu ' Data / Merge Files / Add Variables'

6.Save merged dataset under a <u>new</u> name

ScotCen
Social Research

Recoding variables

Example: study child's general health at Sw3 (BC)

- **Check original variable frequencies: McHgen01**
- **Open a new syntax file via menu ('File' → 'New')**
- **Type simple 'Recode' syntax,** for example group the original variables into answer categories Good (1,2) / Fair (3) and Bad (4,5)

  Recode McHgen01 (1 thru 2=1) (3=2) (4 thru 5=3) (else=copy) INTO GenHbdS3.

  Exe.

- **Check frequencies, tidy up variable and value labels, output formats**

Note: if merging successive sweeps into the same dataset, there will be some system missings ('sysmis') for those cases which skipped one or more of the sweep(s)- you can use Recode to allocate them a missing value– example with Sweep 2 missing at Sweep 3:

RECODE McHgen01_banded (sysmis=-1).

Exe.

Computing a Derived Variable

Example: developmental milestones on Sw1 BC2

- **Check frequencies of original variables** MaDbab02 to MaDbab08
- **Create a new variable** 'Devlpt1' – scale variable measuring number of developmental milestones missed
- **Use Compute syntax to combine the five variables into one:**

```
RECODE MaDtbab02 To MaDbab08 (1 thru 2=0) (3=1)
(else=copy) INTO Temp1 to Temp6.

Exe.

MISSING VALUES Temp1 to Temp6 (-9 thru -1).

COMPUTE Devlpt1 =
SUM(Temp1,Temp2,Temp3,Temp4,Temp5,Temp6).

Exe.
```

- **Check frequencies, tidy up variable and value labels, output formats**

Creating manageable datasets

- The GUS datasets are very large – around 2000 variables in each
- But most analysis will only involve a very small proportion of these variables
- It is useful to create smaller analysis datasets with only the variables you need
- Two particularly good methods of doing this using SPSS syntax are the **KEEP** and **DROP** commands

The KEEP command

- The **KEEP** command allows you to open a large data file specifying which of the variables from that file you wish to ***INCLUDE*** in your working data file.
- The KEEP command can be appended to either the GET FILE or SAVE OUTFILE commands
- Both individual variables and ranges of variables can be specified

GET FILE='C:\temp\GUSSW3B_30.sav'

/Keep = idnumber, dcwinc01, dchgmag2 to dcmedu02.

SAVE OUTFILE='C:\temp\Keep Save As Test.sav'

/Keep = idnumber, dcwinc01, dchgmag2 .

The DROP command

- The **DROP** command allows you to open a large data file specifying which of the variables from that file you wish to ***REMOVE*** from your working data file.
- The DROP command can be appended to either the GET FILE or SAVE OUTFILE commands
- Both individual variables and ranges of variables can be specified

GET FILE='C:\temp\GUSSW3B_30.sav'

/Drop = samptype to dcwtchd2.

SAVE OUTFILE='C:\temp\Drop Save As Test.sav'

/Drop = dcurind1, dcurind2 .

ScotCen Social Research

Analysis using weights and stratification variables

3.

Using the GUS weights

- There are two weights for each cohort on all datasets plus a separate weight for analysis of the partner interview data
- Selection of the correct weight is dependent on the data you are using and analysis being undertaken

Which weight?

Cross-sectional weight

- Use for any cross-sectional analysis of SINGLE SWEEP DATA ONLY

Longitudinal weight

- Use for analysis of MORE THAN ONE SWEEP OF DATA
- Weight used should be from the LATEST sweep (i.e. if analysing sweep 3 and sweep 5 data, use sweep 5 longitudinal weight)

Sweep 2 Partner interview weight

- Use for any analysis of Partner interview data

Applying weights in SPSS

- All analysis should be undertaken on weighted data
- Weights can be applied via the SPSS menu but simpler to apply using syntax command: *Weight by…*
- E.g. to run a frequency on household income in the birth cohort at sweep 5:

```
weight by dewtbrth.
fre dewinc01.
exe.
```

- E.g. to run a crosstab on household income in birth cohort by family type:

```
weight by dewtbrth.
cross dewinc01 by dehgrsp04
/cells = count row
/count = truncate cell.
exe.
```

Note: 'count = truncate cell' prevents SPSS from rounding the weighted counts and simply truncates any decimals

Significance testing in complex samples

- It is common to undertake tests to explore the ‘statistical significance’ of differences between groups
- These tests allow us to estimate the extent to which the result presented by the data is a true reflection of the population status rather than a chance result
- Difficulty in that significance tests assume that the sample you are dealing with is a simple random sample.
- But GUS sample is clustered and stratified - each of these affect the amount of error in the data, which in turn affect the confidence intervals and thus the results of significance tests.
- Thus the complex sample design must be accounted for when testing for significance.

Using the complex samples module in SPSS

- The first step in accounting for sample design is the creation of a complex samples ‘plan file’
- The file incorporates the following variables:
  - Survey weight variable (d#wt####)
  - Stratification variable (d#strat)
  - Cluster variable (d#psu)
- Different plan files are thus necessary for different types of analysis involving:
  - Different sweeps of data
  - Longitudinal or cross-sectional analysis
- Once created, plan files can be saved and linked to in subsequent analysis, no need to re-create everytime.
- See ‘Coping with Complex Samples’ guides on GUS website

Visit our website and sign up to our newsletter:

www.growingupinscotland.org.uk

Follow us on twitter: @growingupinscot

Email us:

paul.bradshaw@scotcen.org.uk,

lesley.kelly@crfr.ed.ac.uk

For further information

ScotCen
Social Research