Restructuring the Millennium Cohort Study data sets for longitudinal analysis in R: An Open Science

project

Lydia Gabriela Speyer & Hildigunnur Anna Hall

Department of Psychology, University of Edinburgh

Millennium Cohort Study (MCS)

- Longitudinal study of around 19,000 children born in the UK between September 2000 and January 2002.
- To date, 7 sweeps of data have been collected
 - Ages: 9 months, 3, 5, 7, 11, 14, and 17 years
- Data are hosted by the Centre for Longitudinal studies and available to download for free from the UK Data Service

CENTRE FOR

STUDIES

LONGITUDINAL

Packages

- haven
 - to read spss files into R
- tidyverse/dplyr
 - to restructure

Data Structure

Instrument/respondent Identifier:

- m = main respondent;
- p = partner respondent;
- x = proxy respondent;
- h = household module;

d = derived variable

Cohort child identifier: indicates which child the variable refers to: a = child one in family; b = child two in family; c = child three in family; 0 = family level variable (i.e. answer applies to all children in the family, e.g. ethnicity of main respondent)

Sweep identifier: a = MCS1; b = MCS2; c = MCS3; d = MCS4;e = MCS5; f = MCS6

Multicode identifier: indicates whether there are multiple answers to the same question: e.g. a = type of illness during pregnancy a, b = type of illness during pregnancy b; 0 = only one answer

Variable identifier: 4 letter combination indicating specific item

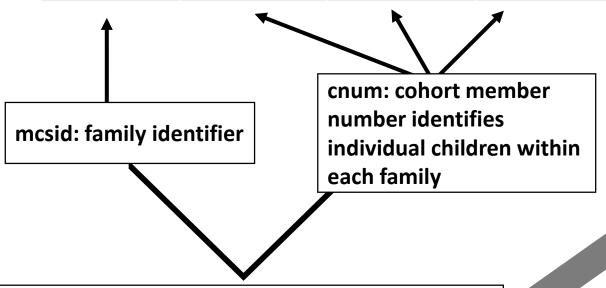
ahcnuma0

Restructuring Data in R

CHILD OF THE NEW CENTURY

JK Data Service

mcsid	ahcnuma0	ahcnumb0	ahcnumc0	amethn00	amilwm0a	amilwm0b	amhpwmaa	amhpwmab	amhpwmba	amhpwmbb	amhpwmca	amhpwmcb
M11323H	1	NA	NA	Caucasian	Infection	Diabetes	None	NA	NA	NA	NA	NA
M11413J	1	2	3	Caucasian	Diabetes	NA	Colic	NA	None	NA	Colic	Eczema
M23453R	1	2	NA	Caucasian	None	NA	Eczema	Cold	Jaundice	NA	NA	NA
M21153S	1	NA	NA	Asian	Anaemia	NA	Colds	NA	NA	NA	NA	NA



1. Extract a vector of column indices of variables that are specific to individual children, i.e. have a, b, or c as the second to last character. Use grep() to filter for regular expression "(.....)(a)(.)" or "(.....)(b)(.)" "(.....)(c)(.)"

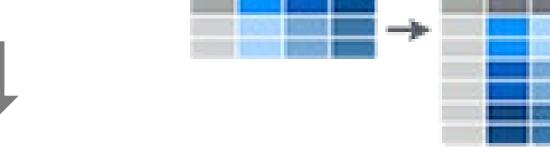


Data Structured as one line per family:

To analyse differences between individual children, data needs to be structured as one line per cohort member



2. Use gather() to collapse those columns into key value pairs, key = columnname, value = values



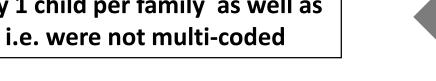
3. Use extract() to split key into separate components of column name based on regular expression "(....)(.)(.)" = "variable" "cnum" "multi"



4. Use unite() to recombine column names excluding "cnum"

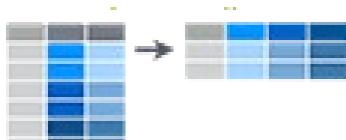


6. Remove rows that represent nobody as there was only 1 child per family as well as rows for variables that didn't have multiple possibilities, i.e. were not multi-coded





5. Use *spread()* to get multi-coded columns back into wide format





mcsid	ahcnum_0	amethn00	amilwm0a	amilwm0b	amhpwm_a	amhpwm_b
M11323H	1	Caucasian	Infection	Diabetes	None	NA
M11413J	1	Caucasian	Diabetes	NA	Colic	NA
M11413J	2	Caucasian	Diabetes	NA	None	NA
M11413J	3	Caucasian	Diabetes	NA	Colic	Eczema
M23453R	1	Caucasian	None	NA	Eczema	Cold
M23453R	2	Caucasian	None	NA	Jaundice	NA
M21153S	1	Asian	Anaemia	NA	Colds	NA



Datasets can now be merged based on mcsid and cnum

Acknowledgements

Many thanks to the University of Edinburgh, Economic and Social Research Council and Scottish **Graduate School for Social** Sciences and for funding our work.

Access our Code



Notes on Analysis

- Analyses need to correct for:
 - Sampling Weights and Non-**Response Bias**
 - **Finite Population Correction Factor**
 - Clustering
 - Stratification

survey package:

- descriptive statistics
- linear regression

lavaan.survey package:

- Structural equation modelling
- Growth curve modelling
- Factor analysis