LIVES
Swiss National Centre of Competence in Research

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
Institut d'études démographiques et du parcours de vie

Dataset: An efficient and secure software framework for handling survey data in R

Emmanuel Rousseaux[♮], Gilbert Ritschard[♮], Danilo Bolano[♮]

[♮]NCCR LIVES *Overcoming vulnerability, Life-course perspectives*, Switzerland

Storing and documenting complex survey data

Storing Life Course data

- ♢ Cross-sectional data
- ♢ Panel data
- ♢ Network data

Merging data and user manual

- ♢ Describing the design of the survey
- ♢ Labels for describing variables and values
- ♢ User-defined missing values
- ♢ Native weights handling
- ♢ Copyright information: author(s), license, etc.
- ♢ Data consistency checks at each stage
- ♢ Importing metadata from a DDI 3.1 XML file

Generation of codebooks

- ♢ Computed directly on the database
- ♢ Statistical summaries of data
- ♢ helpful for detecting abnormal values
- ♢ Useful when sharing data with others

Summary of the SHP all MP dataset

Generated by the R Dataset package
version 0.2.41

January 25, 2013

Overview

- Name: SHP all MP
- Description: Swiss Household Panel, release October 2013, Master personal database
- Number of variables: 72 (1 binaries, 0 ordinals, 40 nominals, 31 scales, 0 timestamps, 0 weightings)
- Number of individuals: 22976 (for 22976 rows)
- Percent of missing values: 60.65 %
- Weighting variable: none.
- Control variable(s): none.
- Spatial variable: none.
- Author(s):
- Contact e-mail:
- License:
- Release date:
- Citation:
- Website:
- Population:

Figure 1: Example of codebook generation, page 1.

Variable summary

Binary variables

	Variable	Description	N	NA (%)	Distribution (%)
7	sex	Sex	22976	0	woman (50.71), man (49.29)

Table 1: Binary variables summary

Nominal variables

	Variable	Description	N	NA (%)	Classes	Distribution (%)
1	filter11	Identification of the survey	22976	0	4	SHP_I (sample 1999) (67.73), SHP_II (sample 2004) (32.27), ...
9	status99	Type of interviews completed: grid, proxy, personal	12931	43.7	3	individual questionnaire (33.94), proxy questionnaire (11.48), grid only (10.85)
10	rnp99	Reason for not responding to ind. Questionnaire	12885	43.9	13	Interviewed (33.94), PROXY (11.48), Person cannot be reached (2.16), Refusal: not interested (1.93), Refusal: no time (1.83), Refusal: opposed to surveys as a matter- (1.24), No time immediately, appointment made (0.89), Age or health related problems (0.73), Refusal: other motives (0.71), Language problem (doesn't speak neither- (0.54), ...
11	rxa99	Reason for proxy	85	99.6	12	...
14	status00	Type of interviews completed: grid, proxy, personal	11678	49.2	3	individual questionnaire (30.78), proxy questionnaire (10.36), grid only (9.68)
15	rnp00	Reason for not responding to ind. Questionnaire	11548	49.7	13	Interviewed (30.78), PROXY (10.36), Refusal: not interested (2.52), Person cannot be reached (1.51), Refusal: opposed to surveys as a matter- (1.43), Refusal: no time (0.86), Refusal: other motives (0.80), Age or health related problems (0.61), Language problem (doesn't speak neither- (0.51), ...
16	rxa00	Reason for proxy	119	99.5	12	...
19	status01	Type of interviews completed: grid, proxy, personal	11116	51.6	3	individual questionnaire (28.73), grid only (10.19), proxy questionnaire (9.46)
20	rnp01	Reason for not responding to ind. Questionnaire	10326	55.1	13	Interviewed (28.73), PROXY (9.46), Refusal: not interested (1.83), Person cannot be reached (1.05), Refusal: no time (0.66), Refusal: other motives (0.61), Age or health related problems (0.61), Person is absent or phone isn't answeri- (0.60), Refusal: opposed to surveys as a matter- (0.53), ...
21	rxa01	Reason for proxy	93	99.6	12	...
24	status02	Type of interviews completed: grid, proxy, personal	9537	58.5	3	individual questionnaire (24.81), proxy questionnaire (8.64), grid only (8.06)
25	rnp02	Reason for not responding to ind. Questionnaire	8936	61.1	13	Interviewed (24.81), PROXY (8.64), Refusal: not interested (1.19), No time immediately, appointment made (0.82), Refusal: other motives (0.71), Refusal: no time (0.62), Person cannot be reached (0.57), ...
26	rxa02	Reason for proxy	163	99.3	12	...

Figure 2: Example of codebook generation, page 4.

Efficient tools for preparing data for analysis

Increase a scientist's time available to concentrate on her research question

1 – Retrieving relevant variables

- ♢ Explore data and search accross huge databases
- ♢ Several searching modes, use of keywords

```
health.var <- contains("health", shp)
##                                                  Description
## h11i76a                          Financial help: health insurance
## p11c01                                          Health status
## p11c02                             Satisfaction with health status
## p11c03                        Improvement in health: Last 12 months
## p11c04a                Health problems: Back problems: Last 4 weeks
## p11c05a          Health problems: Weakness, weariness: Last 4 weeks
## p11c06a            Health problems: Sleeping problems: Last 4 weeks
## p11c07a                   Health problems: Headaches: Last 4 weeks
## p11c08          Health impediment in everyday activities: Extension
## p11c19a                  Chronic illness or long-term health problem
## p11c11  Number of days affected by health problems: Last 12 months
## p11p54                                       Public expenses: Health
## x11c05                                    Assessment of health status
## x11c06                                 Suffering from health problems
## x11c07                                    Cause of health problems
## x11c09                 Days of suffering from health problems: Days
```

2 – Setting weights

```
weighting(study) <- "weights"
```

3 – Computing detailed frequencies

```
frequencies("health", study)
```

Coding	Missing	Label	N	N total	Percent	Percent (all)	Percent total
1		very well	1428		19.16	19.15	
2		well	4811		64.52	64.50	
3		so, so	1037		13.92	13.91	
4		not very well	157		2.11	2.11	
5		poor	21	7456	0.29	0.29	99.97
-2	x	no answer	2		100.00	0.03	
-1	x	does not know	0		0.00	0.00	
-3	x	inapplicable	0		0.00	0.00	
-7	x	filter error	0		0.00	0.00	
-8	x	other error	0	2	0.00	0.00	0.03
				7459			100

4 – Recoding variables

```
study$health.3 <- recode(
  study$health,
  'well' = 1:2,
  'poor' = 3,
  'very poor' = 4:5
)
```

5 – Verifying data

Export the summary of the database in a PDF file

```
exportPDF(study)
```

Comfortable front-ends for classical statistical analysis methods

Representativeness checks

- ♢ Performed automatically
- ♢ Useful when interpreting results

```
checkvars(study) <- c("sex", "work.stat")
shp.association <- subset(study, association == "Active member")

## => control on sex:  warning, p-value < 0.05
## man are oversampled
## woman are undersampled
## => control on work.stat:  warning, p-value < 0.05
## active occupied are oversampled
## unemployed, not in labor force are undersampled
```

Bivariate analysis methods

```
bivan(
  health.2 ~ sex + age.3 + association + work.stat,
  study
)
```

	chi2	cramer.v	gk.tau.sqrt	somer.d
sex	23.83 ***	0.06 ***	0.06 ***	0.04 ***
age.3	273.95 ***	0.19 ***	0.19 ***	0.13 ***
association	85.84 ***	0.11 ***	0.11 ***	0.08 ***
work.stat	232.88 ***	0.18 ***	0.18 ***	0.14 ***

Table 1: Bivariate analysis with the self-reported health as dependend variable. Legend: *** < 0.001, ** < 0.01, * < 0.05, + < 0.1

Logistic regression

```
reglog(
  formula = health.2 ~ sex + age.3,
  imbric = list( # optional argument
    . ~ association,
    . ~ work.stat
  ),
  target = 'poor',
  reference = list( # optional argument
    'association' = 'Not a member',
    'age.3' = '[0,30]'
  ),
  data = study
)
```

Export in an easy-to-read PDF file settings used, estimated odds ratios, quality measures, etc.

	Model 1	Model 2	Model 3
sexwoman	1.321 ***	1.254 ***	1.145 *
age.3(30,65]	3.113 ***	2.994 ***	3.468 ***
age.3(65,97]	5.946 ***	5.635 ***	3.598 ***
associationActive member		0.579 ***	0.595 ***
associationPassive member		0.618 ***	0.619 ***
work.statunemployed			2.460 ***
work.statnot in labor force			2.393 ***
(Intercept)	0.056 ***	0.071 ***	0.054 ***

Table 2: Estimated odds ratios , *** < 0.001, ** < 0.01, * < 0.05, + < 0.1, '' = NA

Rendering spatial data

```
spatial.country(study) <- "CH"
spatial.variable(shp) <- "canton11.short"
gchart.map("age", study)
```

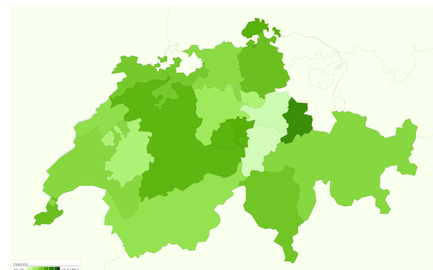

Figure 3: Mean age by canton in Switzerland. PSM 2012, wave 2011 [1].

Conclusion and Future work

- ♢ Efficient and secure framework for handling complex survey data
- ♢ Available on the R-Forge platform
- ♢ Compatibility with the DDI 3.1 specification
- ♢ Front-ends for other popular methods: linear models, clustering, survival analysis and structural equation modeling

[1] Voorpostel, M.; Tillmann, R.; Lebert, F.; Kuhn, U.; Lipps, O.; Ryser, V.-A.; Schmid, F.; Rothenbühler, M. Wernli, B. (2012). *Swiss Household Panel Userguide (1999-2011), Wave 13, October 2012*. Lausanne: FORS.

FNSNF
Swiss National Science Foundation
The National Centres of Competence in Research (NCCR)
are a research instrument of the Swiss National Science Foundation

Contact: emmanuel.rousseaux@unige.ch
IP 14 – Measuring life sequences and the disorder of lives