The Dataset Project

Emmanuel Rousseaux

Institut d'études démographiques et du parcours de vie
Université de Genève
1211 Genève 4, Suisse

emmanuel.rousseaux@unige.ch

Outline

Plan

Teaching and Research Assistant at the Departement of Economics, SES, Unige

PhD directed by

- Gilbert Ritschard, iDemo, Unige
- Michel Beenan, ISS, Unige

Teaching and Research Assistant at the Departement of Economics, SES, Unige

PhD directed by

- Gilbert Ritschard, iDemo, Unige
- Michel Beenen, CS, Unige

Teaching and Research Assistant at the Departement of Economics, SES, Unige

PhD directed by

- Gilbert Ritschard, iDemo, Unige
- Michel Léonard, ISS, Unige

Teaching and Research Assistant at the Departement of Economics, SES, Unige

PhD directed by

- Gilbert Ritschard, iDemo, Unige
- Michel Léonard, ISS, Unige

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

(some) Research interests

- Decision Trees
- Association rules mining
- Nature-based optimization algorithms
- Health sociology
- Cognitive psychology

NCCR LIVES "Overcoming vulnerability: life course perspectives"

I work within the IP14 "Measuring life sequences and the disorder of lives" leaded by Gilbert Ritschard

Aims at providing ad hoc methods for life course analysis in order to have more insight about dynamics of vulnerability

NCCR LIVES "Overcoming vulnerability: life course perspectives"

I work within the IP14 "Measuring life sequences and the disorder of lives" leaded by Gilbert Ritschard

Aims at providing ad hoc methods for life course analysis in order to have more insight about dynamics of vulnerability

NCCR LIVES "Overcoming vulnerability: life course perspectives"

I work within the IP14 "Measuring life sequences and the disorder of lives" leaded by Gilbert Ritschard

Aims at providing ad hoc methods for life course analysis in order to have more insight about dynamics of vulnerability

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Applications

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multilevel Association Rules
- Apply these tools for getting new insight on poor health situations

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multilevel sequential pattern mining
- Apply these tools for getting new insight on poor health situations

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Apply these tools for getting new insight on poor health situations

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Apply these tools for getting new insight on poor health situations

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Apply these tools for getting new insight on poor health situations

Overview

- Providing a software framework for handling survey data in R
- Providing a software framework for handling life courses as a whole
- Providing new mining tools for rare events
  - Decision trees for the discovery of vulnerable profiles
  - Multi-channel association rules mining
- Apply these tools for getting new insight on poor health situations

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- Limited methods
- A lot of methods in the state-of-the-art are provided in R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- Limited number of methods
- A lot of methods in the state-of-the-art are provided on R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- Limited methods
- A lot of methods in the state-of-the-art are provided in R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- [illegible]
- A lot of methods in the state-of-the-art are provided in R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- [illegible]
- A lot of methods in the state-of-the-art are not available in R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- No real standard for sharing dataset
- A lot of methods in the state-of-the-art are provided on R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- No real standard for sharing dataset
- A lot of methods in the state-of-the-art are provided on R only

Motivation

- Currently no framework to handle survey data in R
- Possible in SPSS, SAS, Stata

However

- In theses commercial softwares, no rigourous consistency test
- No real standard for sharing dataset
- A lot of methods in the state-of-the-art are provided on R only

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- [illegible]
- [illegible]

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Goals

- store and manipulate life courses data
- sophisticated managment of missing values
- automatic consistency tests
- representativity of the initial population tests
- user-oriented functions
- automatic summaries

Representativity is central. Efforts have to be made for helping the user

- real structure for handling weights in the database
- representativity checks on each variable
- compute new weights to correctly balance a subdataset

Representativity is central. Efforts have to be made for helping the user

- real structure for handling weights in the database
- representativity checks on each variable
- compute new weights to correctly balance a subdataset

Representativity is central. Efforts have to be made for helping the user

- real structure for handling weights in the database
- representativity checks on each variable
- compute new weights to correctly balance a subdataset

Proposal: the 'Dataset' project

- 2 librairies in R:
  - [illegible]
- Full S4
- [illegible]

Proposal: the 'Dataset' project

- 2 librairies in R:
  - 'Dataset': for cross-sectional survey data
  - 'stDataset': for spatio-temporal survey data
- Full S4
- https://r-forge.r-project.org/projects/dataset/

Proposal: the 'Dataset' project

- 2 librairies in R:
  - 'Dataset': for cross-sectional survey data
  - 'stDataset': for spatio-temporal survey data
- Full S4
- https://r-forge.r-project.org/projects/dataset/

Proposal: the 'Dataset' project

- 2 librairies in R:
  - 'Dataset': for cross-sectional survey data
  - 'stDataset': for spatio-temporal survey data
- Full S4
- https://r-forge.r-project.org/projects/dataset/

Proposal: the 'Dataset' project

- 2 librairies in R:
  - 'Dataset': for cross-sectional survey data
  - 'stDataset': for spatio-temporal survey data
- Full S4
- https://r-forge.r-project.org/projects/dataset/

Proposal: the 'Dataset' project

- 2 librairies in R:
  - 'Dataset': for cross-sectional survey data
  - 'stDataset': for spatio-temporal survey data
- Full S4
- https://r-forge.r-project.org/projects/dataset/

Plan

In the Dataset package we mainly defined

- the Variable object: store one measure on individuals
- the Dataset object: store the full output of the survey

In the Dataset package we mainly defined

- the Variable object: store one measure on individuals
- the Dataset object: store the full output of the survey

In the Dataset package we mainly defined

- the Variable object: store one measure on individuals
- the Dataset object: store the full output of the survey

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

The Variable object is represented by

- codes: vector of codes for each individuals
- missings: vector specifying the coding of missings values
- values: vector specifying the coding of valid cases
- description: a label

Then the Variable is declined into different kind of measures

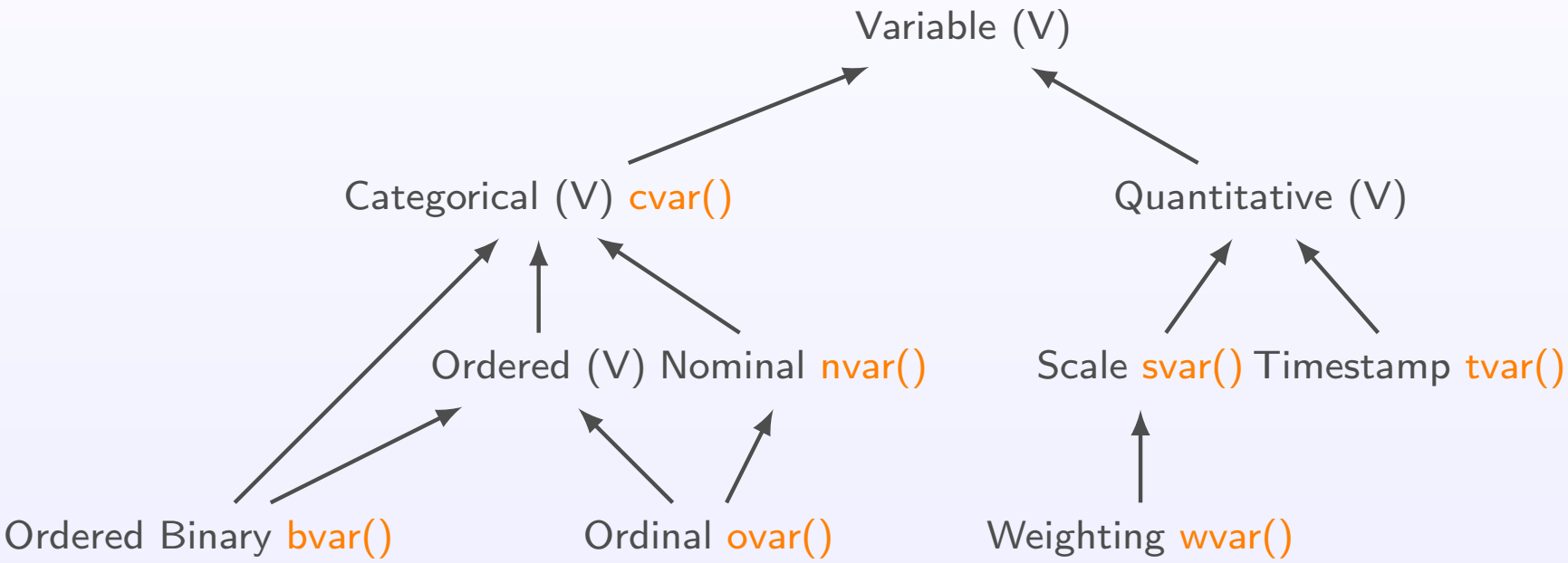

Figure: *Class diagramme of objects inheriting of the Variable class.*

Standard builders for Variables

- svar
- cvar
- ovar

Standard builders for Variables

- svar
- cvar
- ovar

Standard builders for Variables

- svar
- cvar
- ovar

Standard builders for Variables

- svar
- cvar
- ovar

Demo 1

Working with Variable objects

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information we want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

The Dataset object is represented by

- variables: list of variables
- name: name of the dataset
- description: a long label
- row.names: names for rows
- weights: a variable used for weighting
- control: some control variables
- infos: a list for storing other information the user want to share

How to create a Dataset object

- from an existing data.frame
- from a SPSS file
- by hand (as a list of Variable objects)

How to create a Dataset object

- from an existing data.frame
- from a SPSS file
- by hand (as a list of Variable objects)

How to create a Dataset object

- from an existing data.frame
- from a SPSS file
- by hand (as a list of Variable objects)

How to create a Dataset object

- from an existing data.frame
- from a SPSS file
- by hand (as a list of Variable objects)

Demo 2

Preprocessing step: starting from a data set in a spss file, we want get the data in R, et create our data set for our study

importing - recoding - exporting

Native operations on Dataset/Variable objects

- recoding Variables
- bivariate analysis
- logistic regression

Native operations on Dataset/Variable objects

- recoding Variables
- bivariate analysis
- logistic regression

Native operations on Dataset/Variable objects

- recoding Variables
- bivariate analysis
- logistic regression

Native operations on Dataset/Variable objects

- recoding Variables
- bivariate analysis
- logistic regression

Demo 3

Launching analysis

bivariate analysis - logistic regression

Plan

Finish up implementing

- Weights managment tools
- Basic spatial tools
- Easy data capture of a survey
- (Detection/correction of bugs)

Finish up implementing

- Weights managment tools
- Basic spatial tools
- Easy data capture of a survey
- (Detection/correction of bugs)

Finish up implementing

- Weights managment tools
- Basic spatial tools
- Easy data capture of a survey
- (Detection/correction of bugs)

Finish up implementing

- Weights managment tools
- Basic spatial tools
- Easy data capture of a survey
- (Detection/correction of bugs)

Finish up implementing

- Weights managment tools
- Basic spatial tools
- Easy data capture of a survey
- (Detection/correction of bugs)

Taking spatial data into account

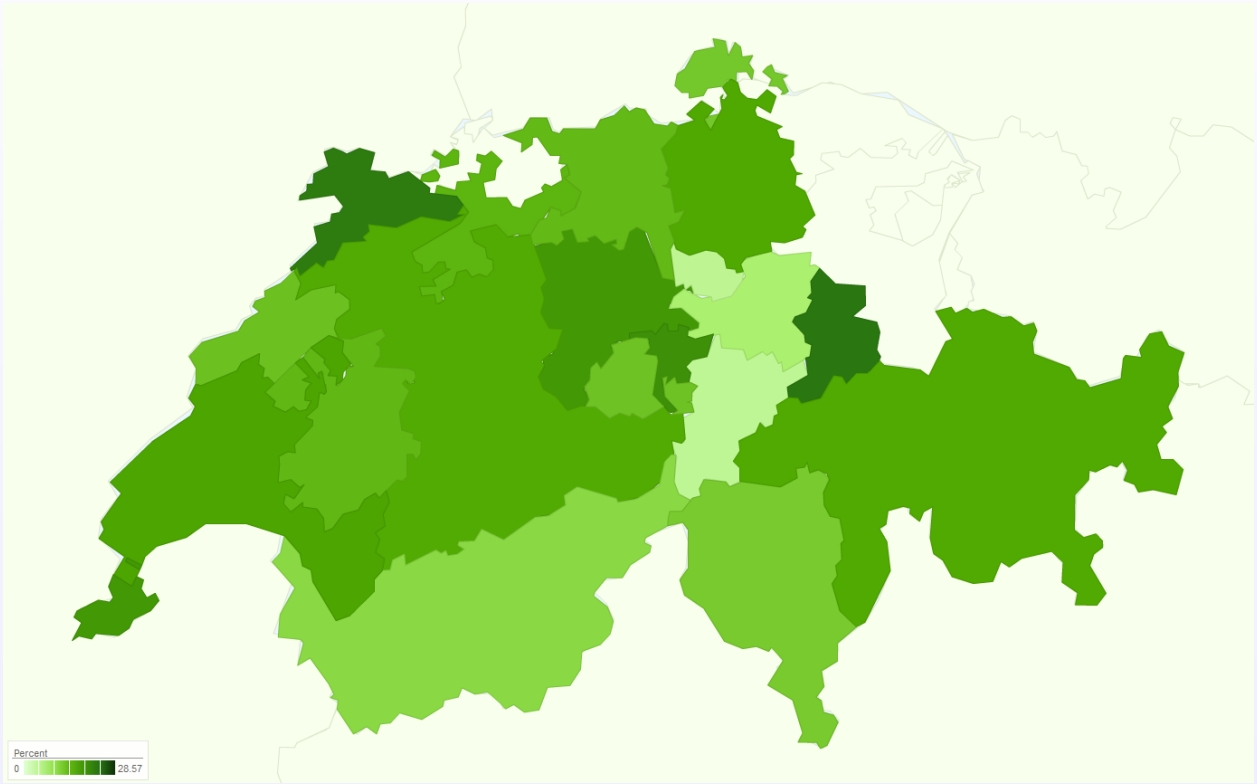

Figure: *Poor/Good SRH ratio. PSM 2011, wave 2010 (no weighted)*

stDataset: handling longitudinal survey data

- Same design as the Dataset package
- Longitudinal summary (in PDF)
- Manipulating trajectories directly
- Construction of an object "life course" ready for analysis

stDataset: handling longitudinal survey data

- Same design as the Dataset package
- Longitudinal summary (in PDF)
- Manipulating trajectories directly
- Construction of an object "life course" ready for analysis

stDataset: handling longitudinal survey data

- Same design as the Dataset package
- Longitudinal summary (in PDF)
- Manipulating trajectories directly
- Construction of an object "life course" ready for analysis

stDataset: handling longitudinal survey data

- Same design as the Dataset package
- Longitudinal summary (in PDF)
- Manipulating trajectories directly
- Construction of an object "life course" ready for analysis

stDataset: handling longitudinal survey data

- Same design as the Dataset package
- Longitudinal summary (in PDF)
- Manipulating trajectories directly
- Construction of an object "life course" ready for analysis

Presentations forthcoming

- (LACOSA) The 'Dataset' Project (poster)
- (SRSSS) Manipulating panel data with stDataset
- (SRSSS) Handling weights with Dataset and stDataset, illustration with the PSM

Presentations forthcoming

- (LACOSA) The 'Dataset' Project (poster)
- (SRSSS) Manipulating panel data with stDataset
- (SRSSS) Handling weights with Dataset and stDataset, illustration with the PSM

Presentations forthcoming

- (LACOSA) The 'Dataset' Project (poster)
- (SRSSS) Manipulating panel data with stDataset
- (SRSSS) Handling weights with Dataset and stDataset, illustration with the PSM

Presentations forthcoming

- (LACOSA) The 'Dataset' Project (poster)
- (SRSSS) Manipulating panel data with stDataset
- (SRSSS) Handling weights with Dataset and stDataset, illustration with the PSM

Elements

of bibliography

Elements of bibliography I

[Voorpostel et al.] Voorpostel, M., Tillmann, R, Lebert, F., Weaver, B., Kuhn, U., Lipps, O., Ryser, V.-A., Schmid, F., Rothenbühler, M., and Wernli, B. *Swiss Household Panel Userguide (1999-2010), Wave 12*. Lausanne: FORS.(October 2011).

Thank you for your attention

Any question?