DatXPlore

1.0.0

January 2014

A software for data exploration and visualization

LIXOFT
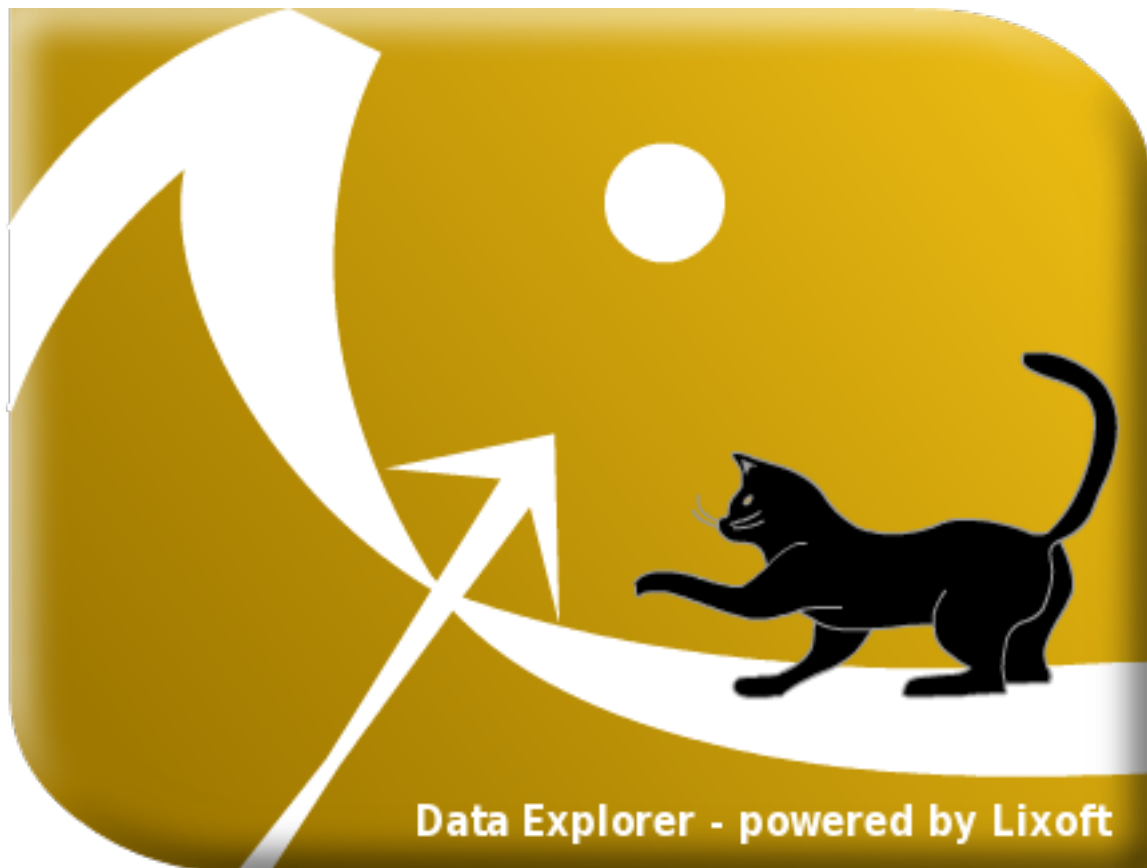
Contents

1   Introduction

DATXPLORE is a graphical and interactive software for the exploration and visualization of data. DATXPLORE provides various plots and graphics (box plots, histograms, survival curves...) to study the statistical properties of discrete and continuous data, and to analyze the behaviour depending on covariates, individuals, etc.

DATXPLORE can be used as a standalone software or be called directly from MONOLIX and share data with MONOLIX (see the MONOLIX documentation).

The examples studied in this guide are located into the directory :

```
<user home path>/lixoft/datxplore/datxplore100/demos/
```

2 Project

The project describes the information of the data set. DATXPLORE uses it to load and initialize the graphical user interface.

A DATXPLORE project describes the content of a data file and sets the types of the data set columns. The information about the file (i.e. the file path, the column delimiter and the header names) takes place into a FILEINFO section. The information about the data set column's type takes place into a CONTENT section.

Example of DATXPLORE **project**:

```
<DATAFILE>
  [FILEINFO]
    name = warfarinData ; project name
    file = 'D:/MLXTRAN/warfarin_data.txt'
    header = {id, time, amt, dv, dvid, wt, sex, age, xxx}
    delimiter = tab

  [CONTENT]
    id = {use=identifier}
    time = {use=time}
    dv = {
         use=observation,
         name={'concentration', 'pca'},
         type={continuous, continuous}
        }
    dvid = {
          use=observationType,
          value={1, 2}
         }
    sex = {
          use=covariate,
          type=categorical,
          name='gender',
          value={'H', 'F'},
          recode={'Man', 'Female'}
         }
    age = {
          use=covariate,
          type=continuous
         }
```

LIXOFT

2.1   File information

The [FILEINFO] section defines:

- The name of the project, by using the keyword `name`. The names's definition is optional.

- The path to the data file, by using the keyword `file`. This entry is mandatory.

- The header names, by using the keyword `header`. This entry is mandatory.

- The column's delimiter, by using the keyword `delimiter`, with the possible value `tab`, `comma`, `semicolumn`, `space`. This entry is mandatory.

Example:

```
[FILEINFO]
  name = warfarinData
  file = 'D:/MLXTRAN/warfarin_data.txt'
  header = {id, time, amt, dv, dvid, wt, sex, age, xxx}
  delimiter = tab
```

2.2   Data file content

The [CONTENT] section defines the column type: each column used in DatXPlore has to be typed.
DatXPlore accepts the following types:

- `identifier` : subject identifier
  Example:
  ```
          id = {use=identifier}
  ```

  `id` is the name of the column.

- time
  Example:
  ```
          time = {use=time}
  ```

  `time` is the name of the column.

- observation
  Example:

```
        dv = {
              use=observation,
              name={'concentration', 'pca'},
              type={continuous, continuous}
            }
```

`dv` is the name of the column.

The `name` item allows to redefine the name of the ouputs. `concentration` and `pca` are the name of the output: here, the column `dv` contains two possible outputs.

The `type` item allows to set the output type with the possible values : `categorical` and `continuous`. In the example, both outputs are `continuous`.

- observation type
  Example:

```
        dvid = {
              use=observationType,
              value={1, 2}
            }
```

`dvid` is the name of the column.

The `value` item allows to define the possible values of the observation type. Theses values are stored in the column association to the `observationType`.

- covariate
  Example:

```
        sex = {
              use=covariate,
              type=categorical,
              value={'H', 'F'},
            }
        ; 'age' is a continuous covariate
        age = {
              use=covariate,
              type=continuous
            }
```

`sex` and `age` are the names of the columns.

A covariate can be defined as continuous (e.g `age`) or categorical (e.g. `sex`). With a categorical, the modalities has to be indicated (in the example these modalities are `H` or `F`).

- regressor **Example:**

```
        xxx={use=regressor}
```

`xxx` is the name of the column.

3 Display

This section describes the graphical interface and the different graphics that can be visualized.

3.1 Overview

It is possible to divide the graphical interface in three parts in addition to the menu.
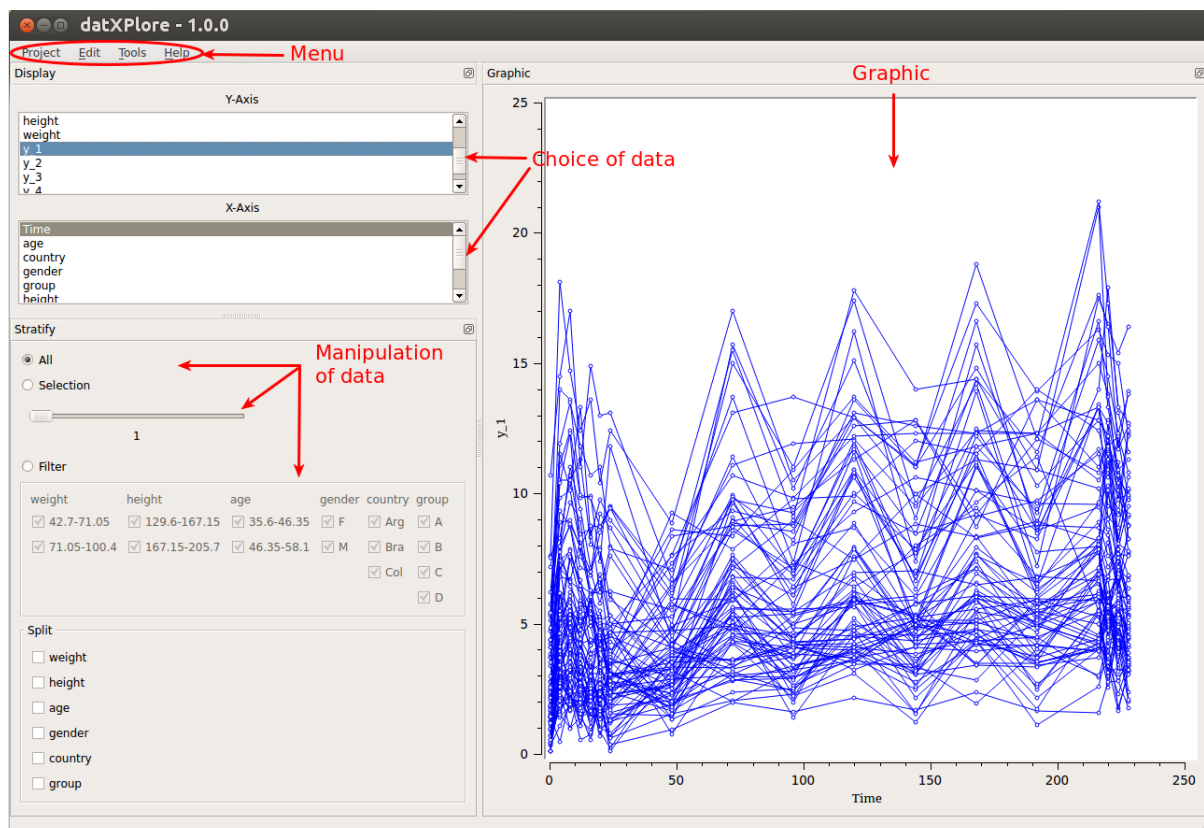
- Menu

  - Project
    Create, load or save a project. Export graphic(s) to .png, .bmp or .jpg format.
  - Edit
    Modify a project.
  - Tools
    Access to a window used to modify the covariates. Set data to log on Y-Axis if possible.

– Help

About. License.

- Graphic

  Set of graphics.

- Choice of Data

  Two lists take inventory of the data from the user dataset. According to the selected Y-Axis element, the X-Axis list is updated. Once an element of each list is selected, the graphics are created.

- Data manipulation

  Used to select, filter or split data.

  This part is described in details in another section (see section 4).

3.2 Choice of data
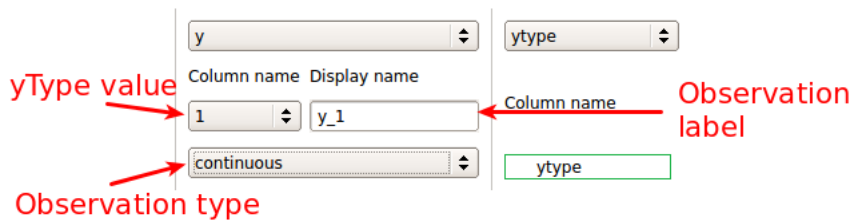
When the dataset is chosen in the project creation case or when the project is modified, the software opens a window to describe the data.
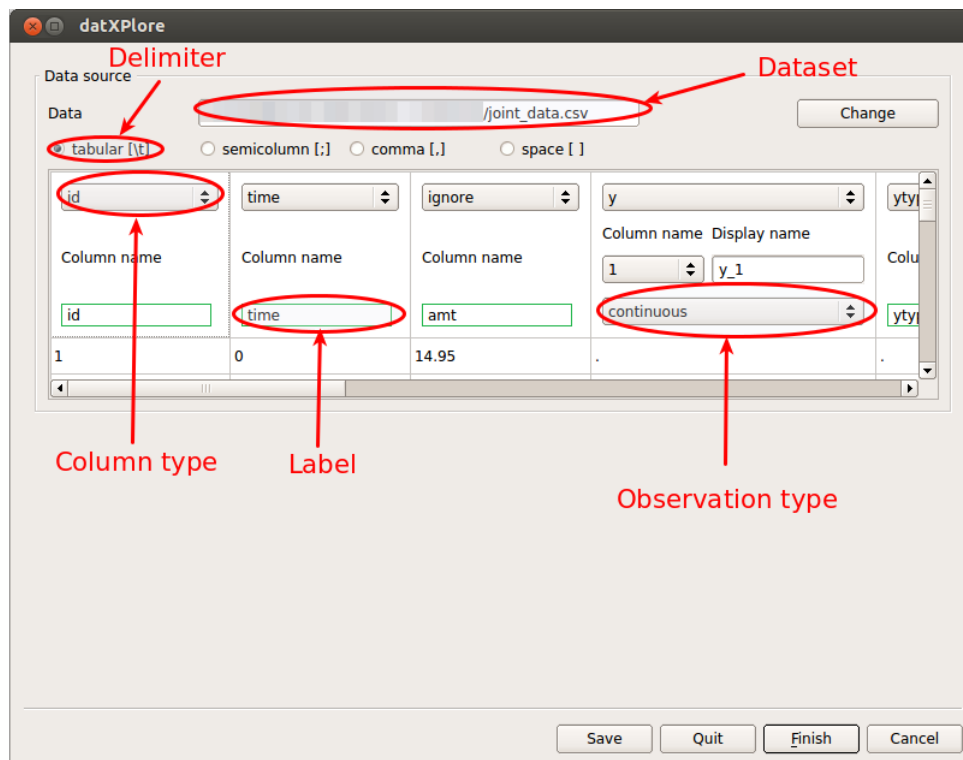
The types which can be selected are :

Type	Description
ignore	Column not used
id	Individuals
time	Time
y	Observation
yType	Used when several observations
cov	Continuous covariate
cat	Discrete covariate
reg	Regression
date	Date can be combined with time
mdv	Missing dependant values

When several observations exist in dataset, a yType column must be selected. Each observation model can take an observation type.

There is 3 types of observations

- continuous

- categorical

- event

3.3 Choice of graphics

According to the selected data in the list of Y-Axis and X-Axis, a kind of graphic is chosen.

There is 6 kinds of graphics :

![LIXOFT]

- Spaghetti : plots Y versus X by individual (continuous variable vs. continuous variable)
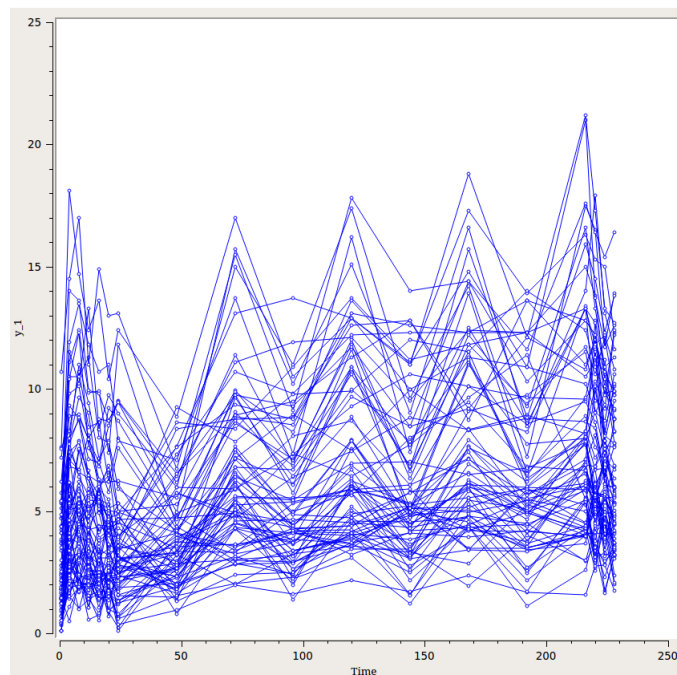
- Scatter : plots dots of Y versus X (continuous variable vs. continuous variable)
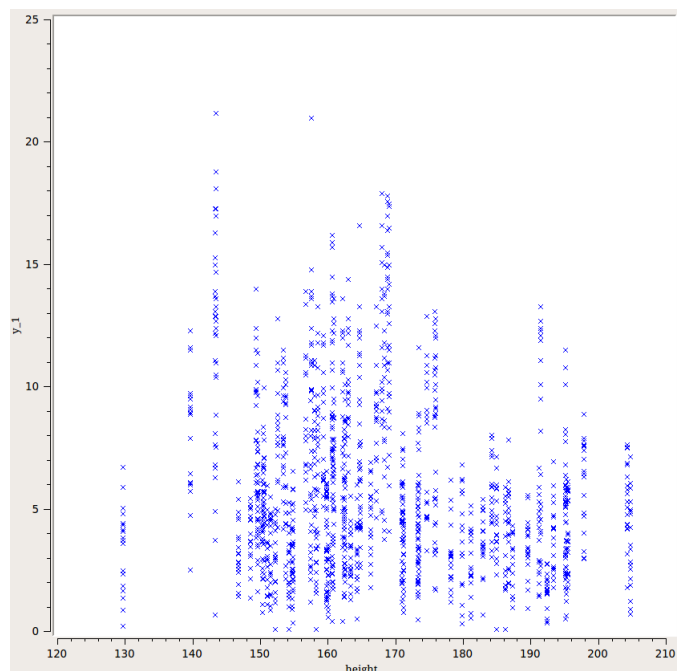
- Box & Whiskers Plot (continuous variable vs. discrete variable)
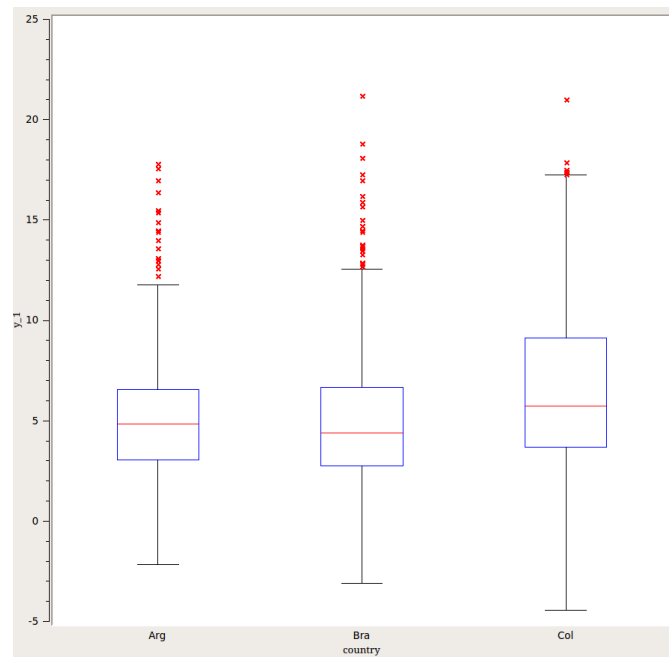
- Cumulative histogram (discrete variable vs. continuous variable)
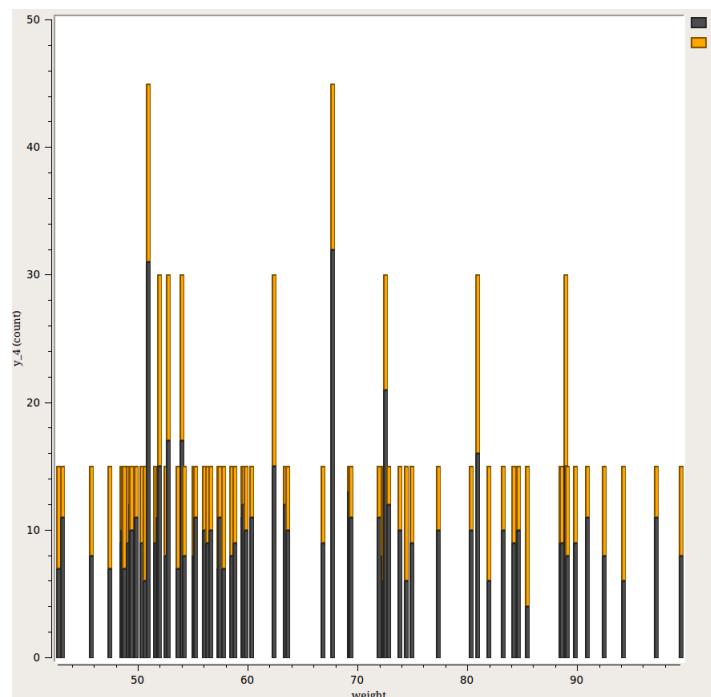
- Histogram by group (discrete variable vs. discrete variable)
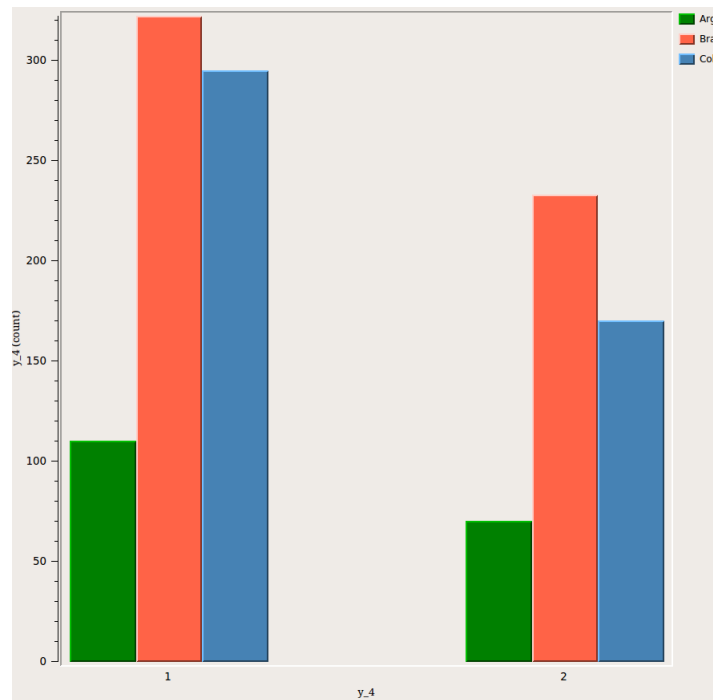
- Survival curve (Kaplan-Meier estimator and events mean)
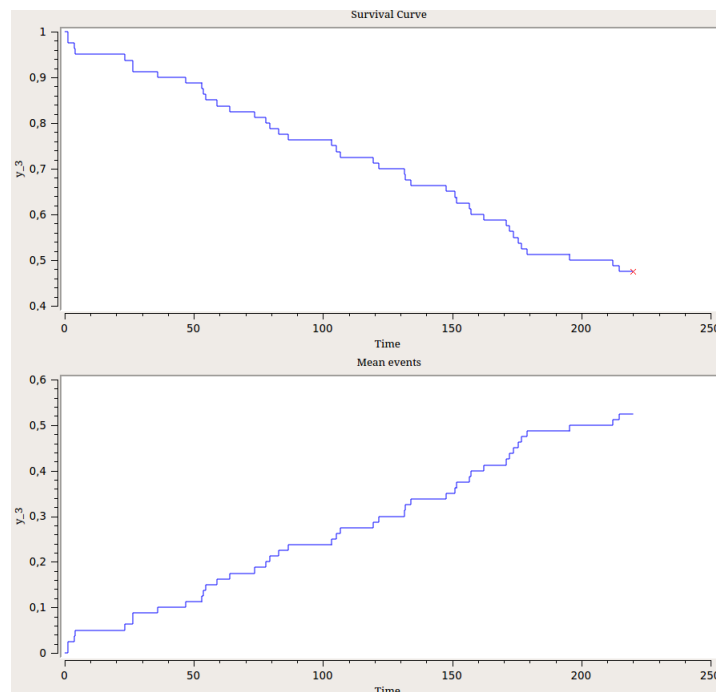
Links betweeen graphics and combinations of variables

Y	X	Graphic
Continuous observation	Time	Spaghetti
Continuous observation	Regression	Spaghetti
Continuous observation	Continuous observation	Spaghetti
Continuous observation	Continuous covariate	Scatter
Continuous observation	Discrete covariate	Box plot
Discrete observation	Time	Cumulative histo.
Discrete observation	Regression	Cumulative histo.
Discrete observation	Continuous covariate	Cumulative histo.
Discrete observation	Discrete covariate	Histo. by group
Continuous covariate	Continuous covariate	Scatter
Continuous covariate	Discrete covariate	Box plot
Discrete covariate	Continuous covariate	Cumulative histo.
Discrete covariate	Discrete covariate	Histo. by group
Event	Time	Survival curve

4 Data manipulation

There are various ways to visualize data. The data can be filtered, splitted or displayed by individual.

The next sections explain how to use each of these features.

4.1 Individual

- `All`

  When this feature is chosen, all individuals are displayed.

- `Selection`

  By enabling this feature, the software will display data only by individual.
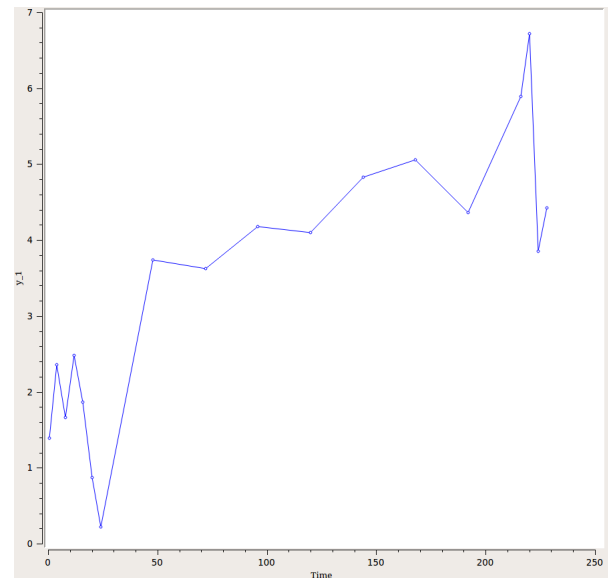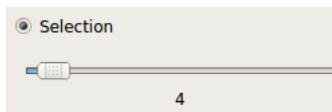
4.2 Filter

This feature is used to select only a part of one or several covariates (discrete or continuous).

Each group of continuous covariate or each modality of discrete covariate can be chosen to be used or not in graphics.

For example, the picture below shows a filter on the covariate `country`. The country `Col` is the only one to be used.

Next, the data are filtered on the `country` (the same way that above) and on the `Age` covariates (used values are greater than 46).
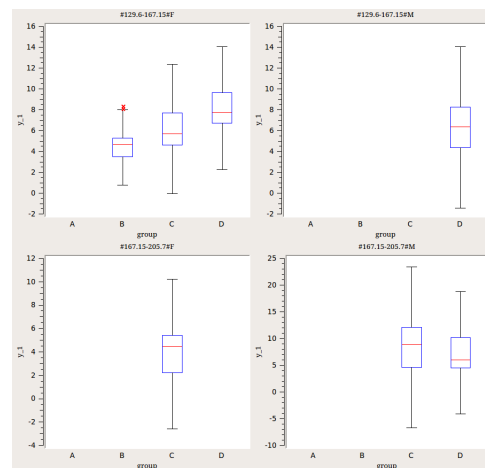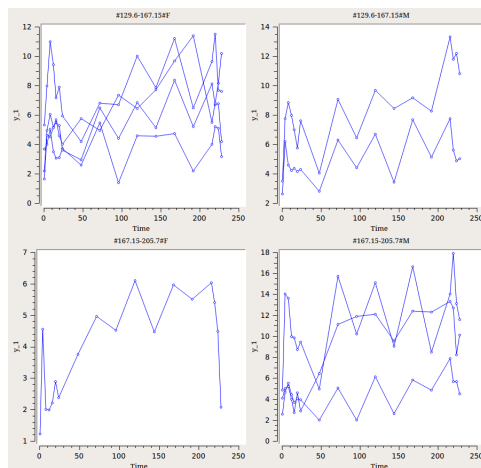
4.3 Split

This feature splits the graphic according to one or several covariates. The picture below is the result of a split on `gender` and on `height`
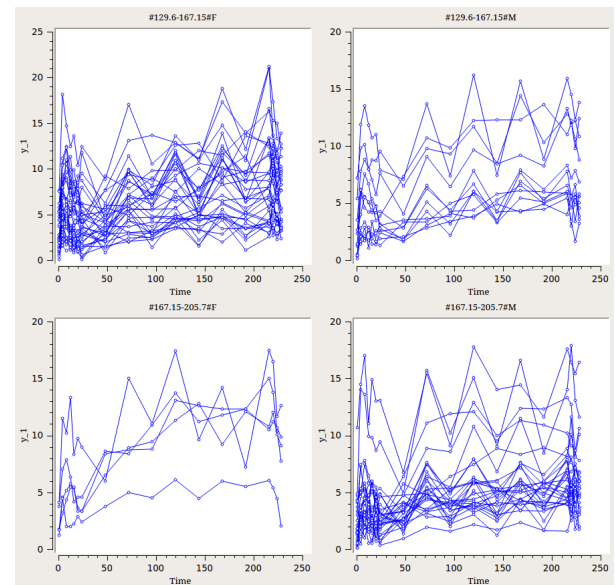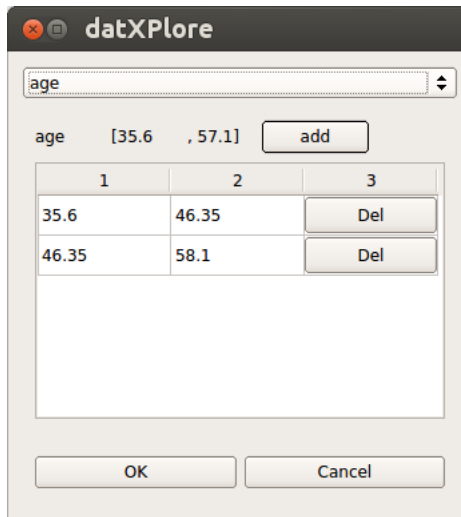

It is also possible to combine filter and split. Let's take the last filter and split. Spaghetti graphic is still used again but as every graphic could be chosen, box plot on `group` is created too.

4.4   Covariates transformation

If the groups of covariates used in filter and split are not adequate for the study, they can be modified by clicking on the menu bar `Tools->Covariate transformation`.

For continuous covariates, groups can be added or removed by clicking on the `Add` button. Then, values of bounds can be set.
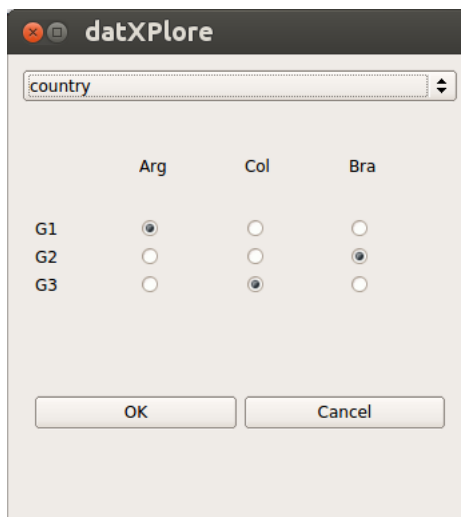
For discrete covariates, modalities can be formed into different groups which behave themselves as modalities.