spm : an R-package that implements a Stochastic Process Model (SPM)

Ilya Y. Zhbannikov

2015-10-27

Overview

The R-package `spm` (https://github.com/izhbannikov/spm) is developed for modeling aging-related changes and it allows (1) data simulation and (2) estimating the process parameters using maximum likelihood estimation by optimizing parameters used in the model. Specifically, developed R-package spm allows (i) one-dimensional SPM; (ii) multiple dimensional SPM; (iii) data simulation for one- and multiple dimensions.

Introduction

General description of age-related changes.

Model description

What we propose

Data description

Data represents a typical clinical trait data and must be presented in form of two datasets: longitudinal dataset (follow-up studies), in which one record represents a single observation, and vital statistics, where one record represents all patient information. Longitudinal dataset must contain person ID (Identification number), status (dead(1)/alive(0)), time and measurements across the covariates. The `spm` can handle an infinite number of covariates but in practice, 5-7 covariates is enough.

Below there is an example of clinical data that can be used in `spm` and we will discuss the field later. Longitudinal studies:

```
##   X ID IndicatorDeath Age AgeNext      DBP      BMI
## 1 1  1              0  30      32 80.00000 25.00000
## 2 2  1              0  32      34 80.51659 26.61245
## 3 3  1              0  34      36 77.78412 29.16790
## 4 4  1              0  36      38 77.86665 32.40359
## 5 5  1              0  38      40 96.55673 31.92014
## 6 6  1              0  40      42 94.48616 32.89139
```

Vital statistics:

```
##   X ID IsDead   LSmort
## 1 1  1      1 85.34578
## 2 2  2      1 80.55053
## 3 3  3      1 98.07315
## 4 4  4      1 81.29779
## 5 5  5      1 89.89829
## 6 6  6      1 72.47687
```

Data fields description

Longitude studies

- ID - subject unique identificatin number.
- IndicatorDeath - 0/1, indicates death of a subject.
- Age - current age of subjects.
- AgeNext - next age of subject he will attend to the survey/exam.
- DBP, BMI - covariates, here "DBP" represents a diastolic blood pressure, "BMI" a body-mass index.

Vital statistics

- ID - subject's unique ID.
- IsDead - death indicator, 0 - alive, 1 - dead.
- LSmort - age at death of stopping observations.

Discrete and Continuous cases

There are two main SPM types in the package: discrete model and continuous model. Discrete model assumes equal intervals between follow-up observations. The example of discrete dataset is given below.

```r
library(spm)
data <- sim_discrete(N=10, ystart=c(80), k=1)
head(data)
```

```
##      id xi t1 t2   par1_1   par1_2
## [1,]  1  0 30 31 80.00000 68.53266
## [2,]  1  0 31 32 68.53266 72.18601
## [3,]  1  0 32 33 72.18601 74.32650
## [4,]  1  0 33 34 74.32650 76.48382
## [5,]  1  0 34 35 76.48382 69.13428
## [6,]  1  0 35 36 69.13428 72.80194
```

In this case there are equal intervals between t1 and t2 (Age and Age.next).

The opposite is continuous case, in which intervals between observations are not equal. The example of continuous case dataset is shown below:

```r
library(spm)
data <- simdata_cont(N=5,ystart = c(50))
head(data)
```

```
##   id xi       t1       t2        y   y.next
## 1  1  0 66.59706 67.09130 54.47705 51.61651
## 2  1  0 67.09130 67.72297 51.61651 53.73434
## 3  1  0 67.72297 69.17765 53.73434 55.36824
## 4  1  0 69.17765 70.65104 55.36824 54.92353
## 5  1  0 70.65104 71.37402 54.92353 56.00064
## 6  1  0 71.37402 72.86143 56.00064 60.61708
```

Discrete case

In discrete case, we use the following assumptions:

$$\bar{y}(t+1) = \bar{u} + \bar{R} \times \bar{y}(t) + \bar{\epsilon}$$

$$\mu(t) = \mu_0(t) + \bar{b}(t) \times \bar{y}(t) + \bar{Q} \times \bar{y}(t)^2$$

Where:

$$\mu_0(t) = \mu_0 e^{\theta t}$$
$$\bar{b}(t) = \bar{b} e^{\theta t}$$
$$\bar{Q}(t) = \bar{Q} e^{\theta t}$$

Continuous case

$$\mu(u) = \mu_0(u) + (\bar{m}(u) - \bar{f}(u)^* \times \bar{Q}(u) \times (\bar{m}(u) - \bar{f}(u)) + Tr(\bar{Q}(u) \times \bar{\gamma}(u))$$

$$dm(t)/dt = \bar{a}(t) \times (\bar{m}(t) - \bar{f}_1(t)) - 2\bar{\gamma}(t) \times \bar{Q}(t) \times (\bar{m}(t) - \bar{f}(t))$$
$$d\bar{\gamma}(t)/dt = \bar{a}(t) \times \bar{\gamma}(t) + \bar{\gamma}(t) \times \bar{a}(t)^* + \bar{b}(t) \times \bar{b}(t)^* - 2\bar{\gamma}t \times \bar{Q}(t) \times \bar{\gamma}(t)$$

Coefficient conversion between continuous and discrete cases

$$Q = Q$$
$$\bar{a} = \bar{R} - diag(k)$$
$$\bar{b} = \bar{\epsilon}$$
$$\bar{f1} = -1 \times \bar{u} \times \bar{a}^{-1}$$
$$\bar{f} = -0.5 \times \bar{b} \times \bar{Q}^{-1}$$
$$mu_0 = mu_0 - \bar{f} \times \bar{Q} \times t(\bar{f})$$
$$\theta = \theta$$

Case with time-dependent coefficients

In two previous cases, we assumed that coefficients is sort of time-dependant: we multiplied them on to

$$e^{\theta t}$$

. In general, this may not be the case. We extend this to a general case, i.e. (we consider one-dimensional case):

$$\bar{a(t)} = par_1 t + par_2$$

- linear function.

The corresponding equations will be equivalent to one-dimensional continuous case described above.

Simulation

We added one- and multi- dimensional simulation to be able to generate test data for hyphotesis testing. Data, which can be simulated can be discrete (equal intervals between observations) and continuous (with arbitrary intervals).

Discrete

For discrete case:

```
library(spm)
data <- sim_discrete(N=100, ystart=c(75, 94), k=2)
head(data)
```

```
##      id xi t1 t2   par1_1   par1_2   par2_1   par2_2
## [1,]  1  0 30 31 75.00000 72.59603 94.00000 85.63799
## [2,]  1  0 31 32 72.59603 64.76411 85.63799 85.30648
## [3,]  1  0 32 33 64.76411 61.73303 85.30648 88.10603
## [4,]  1  0 33 34 61.73303 56.71841 88.10603 84.04519
## [5,]  1  0 34 35 56.71841 55.05411 84.04519 81.14176
## [6,]  1  0 35 36 55.05411 57.28848 81.14176 74.64090
```

Continuous

For continuous case:

```
library(spm)
data <- simdata_cont(N=100)
head(data)
```

```
##   id xi       t1       t2        y   y.next
## 1  1  0 35.72877 36.36639 76.52788 77.06049
## 2  1  0 36.36639 36.90003 77.06049 74.88161
## 3  1  0 36.90003 38.30830 74.88161 73.24287
## 4  1  0 38.30830 38.38744 73.24287 73.62250
## 5  1  0 38.38744 38.74034 73.62250 74.00357
## 6  1  0 38.74034 39.33498 74.00357 75.84503
```

More Examples

[TODO]