CATDATA: SOFTWARE FOR ANALYSIS OF CATEGORICAL DATA WITH COMPLETE OR MISSING RESPONSES

JULIO M. SINGER[1,†], FREDERICO Z. POLETO[1,‡], CARLOS DANIEL PAULINO[2,§]

[1]Instituto de Matemática e Estatística, Universidade de São Paulo, Brazil
[2]Instituto Superior Técnico, Universidade Técnica de Lisboa, Portugal
[†]jmsinger@ime.usp.br [‡]fred@poleto.com [§]dpaulino@math.ist.utl.pt

Abstract

We present a collection of computational routines written in the R language (R Development Core Team, 2007) for the analysis of categorical data with complete or missing responses under a product-multinomial scenario. For complete data or incomplete data generated by an ignorable missingness mechanism as defined in Little and Rubin (2002, Wiley), linear and log-linear models may be fitted via maximum likelihood (ML). Weighted least squares (WLS) methodology may as well be used to fit more general functional linear models for complete data or for incomplete data if a missing completely at random (MCAR) mechanism is assumed. The software also allows a hybrid approach, where ML is used in a first stage, and the estimated marginal probabilities of categorization and their covariance matrix are used in a second stage to fit the model via WLS, in the spirit of functional asymptotic regression methodology described by Imrey, Koch, Stokes *et al.* (1981, 1982, *International Statistical Review*) for complete data. The required computations are automatically conducted for complete data or for incomplete data when missing at random (MAR) or MCAR mechanisms are considered. For missing not at random (MNAR) mechanisms, the first step must be programmed by the user via one of the built-in optimization functions in the R software. Model formulation and use of the functions are similar to GENCAT, a program developed by Landis, Stanish, Freeman and Koch (1976, *Computer Programs in Biomedicine*), or by SAS' PROC CATMOD. We illustrate the procedures with three examples in the field of Biostatistics extracted from Paulino and Singer (2006, Blücher). The first involves fitting a regular log-linear model to a problem with complete data, the second deals with longitudinal data and the third is focused on incomplete data.

Keywords: discrete data, log-linear models, missing data, product-multinomial model.

1 Introduction

The `Catdata` package is a collection of computational routines written in the R language (R Development Core Team, 2007) for the analysis of categorical data with complete or missing responses under a product-multinomial scenario. In Figure 1 we present an outline of the available features of the library of functions. We intend to document all the functions and submit them as a contributed package to The Comprehensive R Archive Network (`http://cran.r-project.org`). Meanwhile, the source code for the functions may be loaded inside `R` using the command

`source("http://www.poleto.com/Catdata.r")`. It is also possible to download the file from this site and load it using the `source()` command, specifying where the file was saved and its label. A more detailed description of the functions is available in Poleto, Singer and Paulino (2007a) and Poleto (2007); in the former we show how to perform the analyses of the examples considered in Poleto, Singer and Paulino (2007b, 2007c) while in the latter we examine almost 40 examples discussed in Paulino and Singer (2006). The underlying theory is developed in Paulino and Singer (2006) with the exception of the special case of missing data under a product-multinomial setup, given in Poleto, Singer and Paulino (2007b).

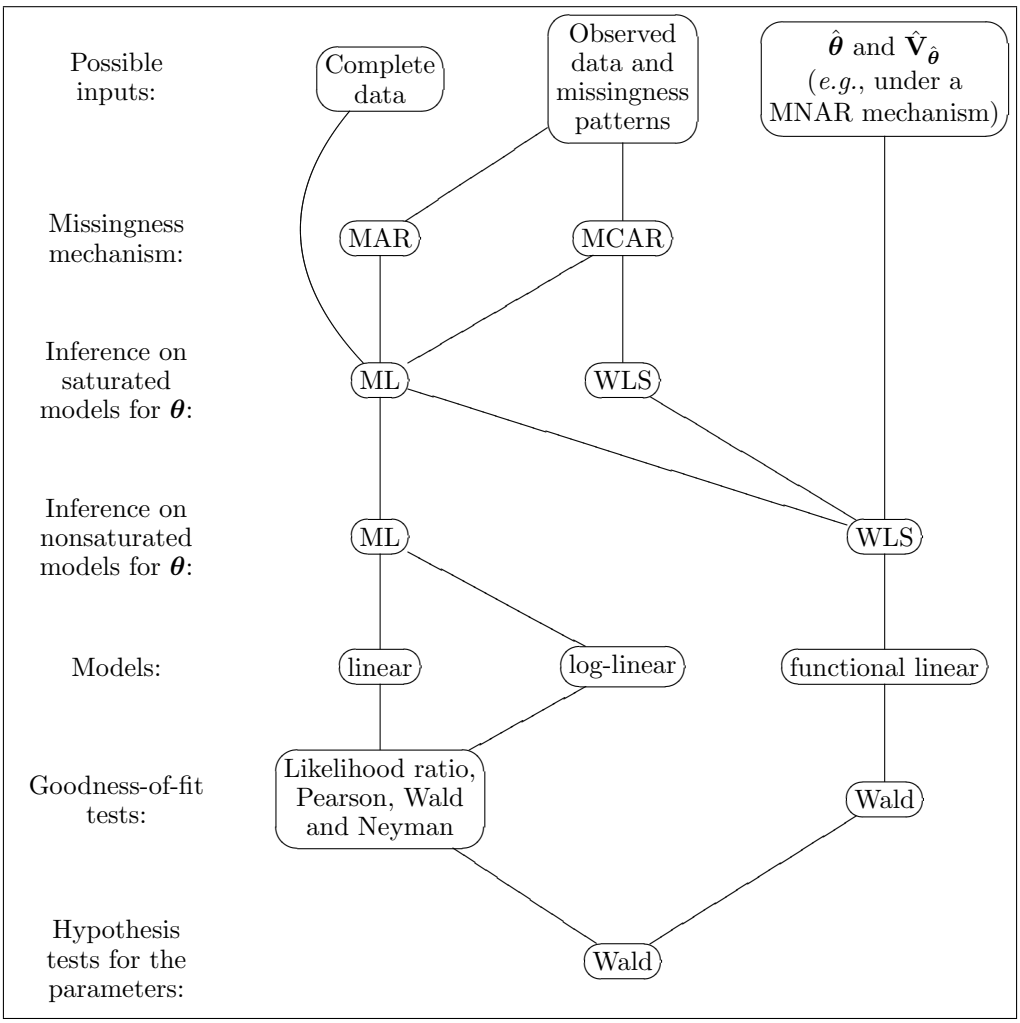

Figure 1: Types of analysis that can be conducted by the library of functions

In Figure 2 we present a flowchart representing the sequence in which the functions should be used for analysis. A brief description of each function follows.

- `readCatdata()` inputs the categorical data; it accommodates complete or missing data;
- `satMarML()` performs maximum likelihood (ML) analyses for saturated models under the missing at random (MAR) and missing completely at random (MCAR) mechanisms based on a `readCatdata()` object; it can only be used in the context of missing data;
- `satMcarWLS()` performs weighted least squares (WLS) analyses for saturated models under the MCAR mechanism based on a `readCatdata()` object; it can only be used in the context

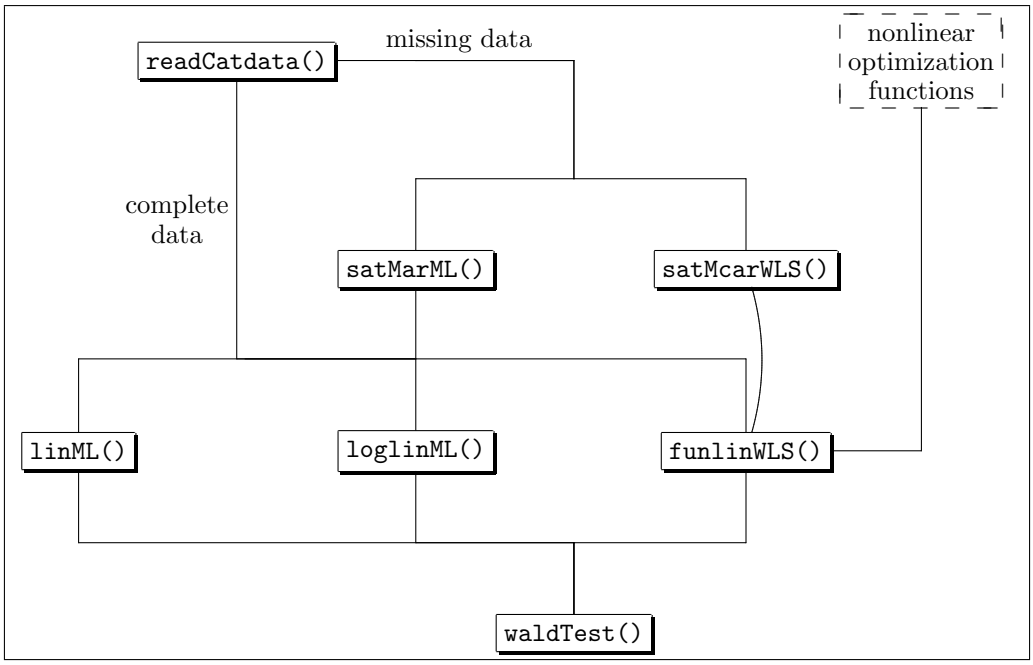

Figure 2: Sequence in which the functions are used for analysis

of missing data;

- `linML()` fits linear models by ML based on a `readCatdata()` object for complete data, or on a `satMarML()` object for missing data;
- `loglinML()` fits log-linear models by ML based on a `readCatdata()` object for complete data, or on a `satMarML()` object for missing data;
- `funlinWLS()` fits functional linear models by WLS based on a `readCatdata()` object for complete data, on `satMarML()` or `satMcarWLS()` objects for missing data or based on estimates of the probabilities of categorization ($\hat{\boldsymbol{\theta}}$) and a consistent estimate of its asymptotic covariance matrix ($\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$) obtained, for example, by one of the built-in nonlinear optimization functions of `R` under any missingness mechanism or even by other kinds of models for the categorization probabilities;
- `waldTest()` performs Wald tests on `linML()`, `loglinML()` and `funlinWLS()` objects, when the models are expressed in terms of freedom equations (Koch, Imrey, Singer, Atkinson and Stokes, 1985).

In Sections 2, 3 and 4, we illustrate the use of the routines with three examples in the field of Biostatistics extracted from Paulino and Singer (2006, Examples 9.12, 12.1 and 13.2). The first involves fitting a standard log-linear model to a problem with complete data, the second deals with longitudinal data and the third is focused on incomplete data. For a detailed analysis, the reader is referred to Paulino and Singer (2006). Other references on the underlying theory are Koch *et al.* (1985), Bishop, Fienberg and Holland (1975), Forthofer and Lehnen (1981), and Agresti (2002).

2 Log-linear model with complete data

The data in Table 1 are extracted from a study designed to evaluate the association between gender (A), age (B), categorized as $<$ or ≥ 55 years old, presence of hypertension (C), categorized as *yes* (systolic blood pressure $\geq$ 140mmHg and/or diastolic blood pressure $\geq$ 90mmHg) or *no*, and degree of obstructive coronary obstruction (D), categorized as $\geq 50\%$ or $< 50\%$.

Table 1: Observed frequencies of 1 448 cardiac patients

Sex (A)	Age (B)	Hypertension (C)	Degree of coronary obstruction (D)	
			$< 50\%$	$\geq 50\%$
female	< 55	no	31	17
		yes	42	27
	≥ 55	no	55	42
		yes	94	104
male	< 55	no	80	112
		yes	70	130
	≥ 55	no	74	188
		yes	68	314

The association among the four variables may be assessed by fitting log-linear models to the parameters of a multinomial distribution with $2^4 = 16$ categories. The probability that a randomly selected patient is classified in the category $(A = a, B = b, C = c, D = d)$, $a, b, c, d = 1, 2$ (following the order of appearance in Table 1) is $\{\theta_{abcd}\}$. The corresponding vector of parameters is $\boldsymbol{\theta} = (\theta_{1111}, \theta_{1112}, \ldots, \theta_{2222})'$. The log-linear model that includes only the first order interactions (AB, AC, AD, BC, BD, CD) may be expressed as

$$\ln(\theta_{abcd}) = \nu + u_a^A + u_b^B + u_c^C + u_d^D + u_{ab}^{AB} + u_{ac}^{AC} + u_{ad}^{AD} + u_{bc}^{BC} + u_{bd}^{BD} + u_{cd}^{CD} \qquad (1)$$

with some set of identifiability constraints. For example, if we use the cell $(A = 1, B = 1, C = 1, D = 1)$ as a reference, the identifiability constraints may be expressed as $u_1^A = u_1^B = u_1^C = u_1^D = u_{11}^{AB} = u_{12}^{AB} = u_{21}^{AB} = u_{11}^{AC} = u_{12}^{AC} = u_{21}^{AC} = u_{11}^{AD} = u_{12}^{AD} = u_{21}^{AD} = u_{11}^{BC} = u_{12}^{BC} = u_{21}^{BC} = u_{11}^{BD} = u_{12}^{BD} = u_{21}^{BD} = u_{11}^{CD} = u_{12}^{CD} = u_{21}^{CD} = 0$, leading to the following interpretation of the parameters:

- $\nu = \ln(\theta_{1111})$ is a component associated to the natural constraint of the multinomial distribution that will not be estimated by the R routines;
- $u_2^A = \ln(\theta_{2111}/\theta_{1111})$, $u_2^B = \ln(\theta_{1211}/\theta_{1111})$, $u_2^C = \ln(\theta_{1121}/\theta_{1111})$, $u_2^D = \ln(\theta_{1112}/\theta_{1111})$ are the marginal effects;
- $u_{22}^{AB} = \ln(\theta_{11cd}\theta_{22cd}/[\theta_{21cd}\theta_{12cd}])$, $c, d = 1, 2$, $u_{22}^{AC} = \ln(\theta_{1b1d}\theta_{2b2d}/[\theta_{2b1d}\theta_{1b2d}])$, $b, d = 1, 2$, $u_{22}^{AD} = \ln(\theta_{1bc1}\theta_{2bc2}/[\theta_{2bc1}\theta_{1bc2}])$, $b, c = 1, 2$, $u_{22}^{BC} = \ln(\theta_{a11d}\theta_{a22d}/[\theta_{a21d}\theta_{a12d}])$, $a, d = 1, 2$, $u_{22}^{BD} = \ln(\theta_{a1c1}\theta_{a2c2}/[\theta_{a2c1}\theta_{a1c2}])$, $a, c = 1, 2$, $u_{22}^{CD} = \ln(\theta_{ab11}\theta_{ab22}/[\theta_{ab21}\theta_{ab12}])$, $a, b = 1, 2$, correspond to the log odds ratios of each pair of variables, considered homogeneous conditionally on all the categories of the other variables.

The commands to fit this model by ML and WLS are as follows.

```
e912a.TF<-c(31,17,42,27,55,42,94,104,80,112,70,130,74,188,68,314)
e912a.catdata<-readCatdata(TF=e912a.TF)
e912a.X<-rbind(c(0,0,0,0),c(0,0,0,1),c(0,0,1,0),c(0,0,1,1),
               c(0,1,0,0),c(0,1,0,1),c(0,1,1,0),c(0,1,1,1),
               c(1,0,0,0),c(1,0,0,1),c(1,0,1,0),c(1,0,1,1),
               c(1,1,0,0),c(1,1,0,1),c(1,1,1,0),c(1,1,1,1))
e912a.X<-cbind(e912a.X,e912a.X[,1]*e912a.X[,2],e912a.X[,1]*e912a.X[,3]
 ,e912a.X[,1]*e912a.X[,4],e912a.X[,2]*e912a.X[,3],e912a.X[,2]*e912a.X[,4]
 ,e912a.X[,3]*e912a.X[,4]) #A,B,C,D, AB,AC,AD,BC,BD,CD
e912a.loglinml<-loglinML(e912a.catdata,X=e912a.X)
e912a.loglinwls<-funlinWLS(model=c("lin","log"),obj=e912a.catdata,X=e912a.X)
```

We may compare the results either by using the command `summary()` having the label of the object as argument or the brief version of `print()`, invoked by typing only the object label, as indicated below.

```
> e912a.loglinml

Call: loglinML(obj = e912a.catdata, X = e912a.X)

Maximum likelihood estimates of the parameters of the log-linear model:
       estimate  std.error  z-value  p-value
 [1,]   0.8603    0.1370     6.2780   0.0000
 [2,]   0.4630    0.1369     3.3810   0.0007
 [3,]   0.2306    0.1358     1.6977   0.0896
 [4,]  -0.9107    0.1524    -5.9770   0.0000
 [5,]  -0.5951    0.1341    -4.4379   0.0000
 [6,]  -0.4474    0.1268    -3.5298   0.0004
 [7,]   1.2314    0.1255     9.8082   0.0000
 [8,]   0.2806    0.1136     2.4698   0.0135
 [9,]   0.6729    0.1203     5.5953   0.0000
[10,]   0.4089    0.1173     3.4873   0.0005

Goodness of fit of the log-linear model (d.f.=5):
                  statistic  p-value
Likelihood ratio     3.0358   0.6945
Pearson              3.0666   0.6897
Neyman               2.9971   0.7004
Wald                 3.0546   0.6916

> e912a.loglinwls

Call: funlinWLS(model = c("lin", "log"), obj = e912a.catdata, X = e912a.X)

Weighted least squares estimates of the parameters of the model:
       estimate  std.error  z-value  p-value
 [1,]   0.8553    0.1373     6.2313   0.0000
 [2,]   0.4583    0.1372     3.3407   0.0008
 [3,]   0.2273    0.1358     1.6734   0.0942
 [4,]  -0.9038    0.1513    -5.9721   0.0000
 [5,]  -0.5894    0.1328    -4.4378   0.0000
 [6,]  -0.4419    0.1263    -3.4981   0.0005
 [7,]   1.2262    0.1252     9.7938   0.0000
 [8,]   0.2837    0.1135     2.5004   0.0124
 [9,]   0.6695    0.1197     5.5937   0.0000
[10,]   0.4050    0.1170     3.4604   0.0005

Wald goodness of fit statistic of the model (d.f.=5): 3.0546 (p-value=0.6916)
```

Paulino and Singer (2006) show all the steps of the forward selection procedure that lead to model (1). They also comment that from a clinical point of view it may be more appropriate to consider the marginal totals of each combination of the categories of A, B, and C as fixed, so these variables play the role of risk factors, the effects of which we want to investigate. We can do this either by including the terms $\{u_{abc}^{ABC}\}$ in model (1), or by assuming a product-multinomial distribution and fitting appropriate logistic models.

Under this last setting, we are interested in the probabilities $\{\theta_{d(abc)}\}$ that a patient be classified in the category $(D = d)$ conditionally on the values of the explanatory variables $(A = a, B = b, C = c)$, or alternatively, in the corresponding logit functions $\{\ln(\theta_{1(abc)}/\theta_{2(abc)})\}$. Letting $\boldsymbol{\theta} = (\theta_{1(111)}, \theta_{2(111)}, \ldots, \theta_{2(222)})'$, the functions of interest may be formulated as $\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta})$ with $\mathbf{A} = \mathbf{I}_8 \otimes (1, -1)$, where $\mathbf{I}_8$ indicates the identity matrix of order 8, $\otimes$ denotes the Kronecker product, and $\mathbf{ln}(\boldsymbol{\theta})$ is the vector (natural) logarithmic operator, the elements of which correspond to the natural logarithms of the elements of $\boldsymbol{\theta}$. The logistic model defined by (ABC, AD, BD, CD) may be expressed as

$$\ln\left(\frac{\theta_{1(abc)}}{\theta_{2(abc)}}\right) = \lambda + \alpha_a + \beta_b + \gamma_c \tag{2}$$

with the identifiability constraints $\alpha_2 = \beta_2 = \gamma_2 = 0$. The commands to fit this model by ML and WLS are as follows.

```
e912b.TF<-rbind(c(31, 17),c(42, 27),c(55, 42),c(94,104),
                c(80,112),c(70,130),c(74,188),c(68,314))
e912b.catdata<-readCatdata(TF=e912b.TF)
e912b.XL<-rbind(c(1,1,1,1),
                c(1,1,1,0),
                c(1,1,0,1),
                c(1,1,0,0),
                c(1,0,1,1),
                c(1,0,1,0),
                c(1,0,0,1),
                c(1,0,0,0))
e912b.loglinml<-loglinML(e912b.catdata,A=diag(8)%x%t(c(1,-1)),XL=e912b.XL)
e912b.loglinwls<-funlinWLS(model=c("lin","log"),obj=e912b.catdata,
                                        A1=diag(8)%x%t(c(1,-1)),XL=e912b.XL)
```

The detailed results obtained via ML are generated by the command

```
> summary(e912b.loglinml)

Call: loglinML(obj = e912b.catdata, A = diag(8) %x% t(c(1, -1)), XL = e912b.XL)


Maximum likelihood estimates of the probabilities under the log-linear model (LLM):
      [,1]    [,2]
[1,]  0.7131  0.2869
[2,]  0.6229  0.3771
[3,]  0.5592  0.4408
[4,]  0.4573  0.5427
[5,]  0.4205  0.5795
[6,]  0.3253  0.6747
[7,]  0.2702  0.7298
```

```
[8,]  0.1974  0.8026

Standard errors:
      [,1]    [,2]
[1,]  0.0312  0.0312
[2,]  0.0332  0.0332
[3,]  0.0325  0.0325
[4,]  0.0278  0.0278
[5,]  0.0273  0.0273
[6,]  0.0250  0.0250
[7,]  0.0214  0.0214
[8,]  0.0164  0.0164


Maximum likelihood estimates of the log-linear functions:
      observed  std.error  under the LLM  std.error
[1,]   0.6008    0.3018     0.9107         0.1524
[2,]   0.4418    0.2467     0.5018         0.1411
[3,]   0.2697    0.2049     0.2378         0.1316
[4,]  -0.1011    0.1423    -0.1711         0.1120
[5,]  -0.3365    0.1464    -0.3207         0.1121
[6,]  -0.6190    0.1482    -0.7296         0.1140
[7,]  -0.9324    0.1372    -0.9936         0.1086
[8,]  -1.5299    0.1338    -1.4025         0.1036


Maximum likelihood estimates of the parameters of the log-linear model:
      estimate  std.error  z-value   p-value
[1,]   -1.4025    0.1036   -13.5359    0.0000
[2,]    1.2314    0.1255     9.8081    0.0000
[3,]    0.6729    0.1203     5.5951    0.0000
[4,]    0.4089    0.1173     3.4870    0.0005


Fisher scoring attained the convergence criterion in 3 iterations.


Goodness of fit of the log-linear model (d.f.=4):
                  statistic  p-value
Likelihood ratio     3.0353   0.5519
Pearson              3.0641   0.5472
Neyman               3.0002   0.5578
Wald                 3.0506   0.5494


Estimated frequencies under log-linear model:
      [,1]    [,2]
[1,]   34.23   13.77
[2,]   42.98   26.02
[3,]   54.24   42.76
[4,]   90.55  107.45
[5,]   80.74  111.26
[6,]   65.06  134.94
[7,]   70.79  191.21
[8,]   75.42  306.58
```

Estimates for the odds ratios and the corresponding confidence intervals may be obtained from the results via the commands

```
> exp(e912b.loglinml$beta[2:4])
[1] 3.425991 1.959833 1.505168
> exp(e912b.loglinml$beta[2:4]-qnorm(0.975)*sqrt(diag(e912b.loglinml$Vbeta))[2:4])
[1] 2.678670 1.548299 1.196100
> exp(e912b.loglinml$beta[2:4]+qnorm(0.975)*sqrt(diag(e912b.loglinml$Vbeta))[2:4])
[1] 4.381806 2.480750 1.894098
```

The categorical data input function `readCatdata()` has the table of frequencies as the argument `TF`; when this argument is a vector, it assumes a multinomial distribution; when the argument is a matrix, with each row representing one subpopulation, it assumes a product-multinomial distribution. For each subpopulation, labeled $s = 1, \ldots, S$, there are R response categories, labeled $r = 1, \ldots, R$. Hence, the resulting table of frequencies is a $S \times R$ matrix.

The probability that a randomly selected unit from the subpopulation s is classified in the response category r is denoted $\{\theta_{r(s)}\}$. For subpopulation $s = 1, \ldots, S$, these probabilities are stacked in the vector $\boldsymbol{\theta}_s = \left(\theta_{r(s)}, r = 1, \ldots, R\right)'$, and then summarized in the vector $\boldsymbol{\theta} = \left(\boldsymbol{\theta}'_s, s = 1, \ldots, S\right)'$. All the (structural) models that we consider are expressed as functions of $\boldsymbol{\theta}$.

Model (1) is a special case of

$$\mathbf{ln}(\boldsymbol{\theta}) = [\mathbf{I}_S \otimes \mathbf{1}_R]\, \boldsymbol{\nu} + \mathbf{X}\boldsymbol{\beta}, \tag{3}$$

where $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_S)'$ is a vector with S components associated to the natural constraints, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector that embodies the $p \leq S(R-1)$ unknown parameters, and $\mathbf{X} = (\mathbf{X}'_1, \ldots, \mathbf{X}'_S)'$ is a $SR \times p$ matrix with each $R \times p$ submatrix $\mathbf{X}_s$ having its columns linearly independent from the vector $\mathbf{1}_R$ that defines the s-th natural constraint, $\mathbf{1}'_R\boldsymbol{\theta}_s = 1$, *i.e.*, $r(\mathbf{1}_R, \mathbf{X}_s) = 1 + r(\mathbf{X}_s)$, $s = 1, \ldots, S$, and $r(\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}) = S + p$, where $\mathbf{1}_R$ represents a $R \times 1$ vector with all elements equal to 1, and $r()$ is the rank operator.

Model (2) is a special case of a larger class of log-linear models expressed as

$$\mathbf{A}\, \mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}, \tag{4}$$

where $\mathbf{A}$ is a $u \times SR$ matrix with rank $r(\mathbf{A}) = u \leq S(R-1)$ such that $\mathbf{A}\,(\mathbf{I}_S \otimes \mathbf{1}_R) = \mathbf{0}_{u,S}$, where $\mathbf{0}_{u,S}$ denotes a $u \times S$ matrix with all elements equal to 0. The default choice of the routines for log-linear models is $\mathbf{A} = \mathbf{I}_S \otimes [\mathbf{I}_{R-1}, -\mathbf{1}_{R-1}]$; this generates logits with the baseline category R.

The freedom equation formulations (3) and (4) are respectively equivalent to the constraint formulations

$$\mathbf{U}\, \mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{S(R-1)-p}, \tag{5}$$

$$\mathbf{U}_L\, \mathbf{A}\, \mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}, \tag{6}$$

where $\mathbf{U}$ ($\mathbf{U}_L$) is a $[S\{R-1\}-p] \times SR$ ($[u-p] \times u$) is a full rank matrix defining the $S[R-1]-p$ ($u-p$) constraints such that $\mathbf{U}[\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}] = \mathbf{0}_{(SR-p),p}$ ($\mathbf{U}_L\mathbf{X}_L = \mathbf{0}_{(u-p),p}$), where $\mathbf{0}_{u-p}$ represents an $(u-p) \times 1$ vector with all null elements.

The functions `funlinWLS()` [in the cases with the argument `model=c("lin","log")`] and `loglinML()` are used to fit log-linear models by WLS and ML, respectively. The models may be specified by any of the formulations (3), (4), (5) or (6). The arguments correspond to the

label of the matrices used in the expressions, *i.e.*, `X`, `XL`, `U` and `UL`, respectively, for $\mathbf{X}$, $\mathbf{X}_L$, $\mathbf{U}$ and $\mathbf{U}_L$. The only exception is the matrix $\mathbf{A}$, which should be considered in the argument as `A1` for the WLS approach and as `A`, for the ML procedure.

3 Longitudinal data

The data in Table 2 are obtained from a study conducted with the objective of evaluating the efficacy of a treatment for urinary infection with respect to one of its symptoms. Fifty patients with this kind of infection were examined at three moments: right after the treatment was administered, and 14 and 21 days after the first assessment. The observed characteristic was the vaginal discharge level, classified as absent (0), light (1), moderate (2), or severe (3). Missing data, commonly obtained in this kind of problem, were imputed for illustration purposes, and are shown within parentheses; alternative approaches to accommodate the missing data must consider the techniques discussed in Section 4. More details are given in Paulino and Singer (2006).

Table 2: Vaginal discharge level in three assessments

Patient	Assessment			Patient	Assessment		
	initial	14 days	21 days		initial	14 days	21 days
1	1	0	0	26	2	0	0
2	2	0	(0)	27	2	3	(3)
3	1	0	0	28	3	0	1
4	2	0	0	29	2	2	1
5	2	1	(1)	30	2	0	0
6	2	(2)	(2)	31	3	(2)	0
7	2	(2)	(2)	32	0	(0)	0
8	1	1	1	33	1	1	0
9	3	0	0	34	1	0	0
10	2	1	2	35	1	0	0
11	2	1	3	36	1	1	0
12	1	1	0	37	0	0	1
13	2	0	0	38	0	0	1
14	2	0	0	39	1	0	0
15	2	1	1	40	2	0	0
16	2	1	1	41	1	1	0
17	2	1	0	42	1	0	0
18	1	0	0	43	2	(2)	(2)
19	1	0	0	44	2	2	(2)
20	2	0	0	45	2	0	1
21	1	1	0	46	2	(2)	0
22	3	1	0	47	3	1	0
23	3	0	0	48	3	0	0
24	2	1	1	49	2	1	1
25	2	0	0	50	3	0	0

Obs.: values shown within parentheses were imputed.

The questions of interest are: (i) to assess the temporal evolution of the response distribution,

i.e., if the frequencies of patients with vaginal discharge of higher intensity are smaller after 14 days of treatment; (ii) to evaluate whether the treatment can be interrupted after 14 days, *i.e.*, if some relevant characteristic *e.g.*, the proportion of patients with vaginal discharge moderate or severe remains unaltered after the second assessment.

Admitting that the patients correspond to a simple random sample of a (conceptual) population for which we would like to draw conclusions, we may adopt a multinomial distribution for inferential purposes. Observations on only 50 patients generate a sparse table for the $4^3 = 64$ response categories associated to the probabilities $\{\theta_{abc}\}$; here, θ_{abc} denotes the probability that a randomly selected patient is classified in the a-th, b-th and c-th vaginal discharge levels at the first, second and third assessments, respectively, where $a, b, c = 0, 1, 2, 3$. Despite the sparseness of the table, the analyses may be focused on functions of the first-order marginal distributions $\{\theta_{a\cdot\cdot} = \sum_{b,c} \theta_{abc}\}$, $\{\theta_{\cdot b\cdot} = \sum_{a,c} \theta_{abc}\}$, and $\{\theta_{\cdot\cdot c} = \sum_{a,b} \theta_{abc}\}$; these 12 probabilities $\boldsymbol{\theta}_{\cdot\cdot} = (\theta_{0\cdot\cdot}, \theta_{1\cdot\cdot}, \ldots, \theta_{\cdot\cdot 3})'$ may be obtained from $\boldsymbol{\theta} = (\theta_{000}, \theta_{001}, \ldots, \theta_{333})'$ by considering the linear function $\mathbf{A}\boldsymbol{\theta}$ with $\mathbf{A} = (\mathbf{I}_4 \otimes \mathbf{1}_{16}, \mathbf{1}_4 \otimes \mathbf{I}_4 \otimes \mathbf{1}_4, \mathbf{1}_{16} \otimes \mathbf{I}_4)'$.

We may also direct our attention to the expected proportion of patients with vaginal discharge moderate or severe. Here, the vector of functions of interest is

$$\mathbf{F}_1(\boldsymbol{\theta}) = \begin{pmatrix} \theta_{2\cdot\cdot} + \theta_{3\cdot\cdot} \\ \theta_{\cdot 2\cdot} + \theta_{\cdot 3\cdot} \\ \theta_{\cdot\cdot 2} + \theta_{\cdot\cdot 3} \end{pmatrix},$$

where $\mathbf{F}_1(\boldsymbol{\theta}) = \mathbf{A}_1\boldsymbol{\theta}$ with $\mathbf{A}_1 = ([0, 0, 1, 1]' \otimes \mathbf{1}_{16}, \mathbf{1}_4 \otimes [0, 0, 1, 1]' \otimes \mathbf{1}_4, \mathbf{1}_{16} \otimes [0, 0, 1, 1]')'$. Alternatively, if we assign scores (*e.g.*, absent=0, light=1, moderate=2, severe=3) to the response categories, we may compare the expected scores at each of the tree assessments. In this case, the vector of functions of interest is

$$\mathbf{F}_2(\boldsymbol{\theta}) = \begin{pmatrix} 0 \times \theta_{0\cdot\cdot} + 1 \times \theta_{1\cdot\cdot} + 2 \times \theta_{2\cdot\cdot} + 3 \times \theta_{3\cdot\cdot} \\ 0 \times \theta_{\cdot 0\cdot} + 1 \times \theta_{\cdot 1\cdot} + 2 \times \theta_{\cdot 2\cdot} + 3 \times \theta_{\cdot 3\cdot} \\ 0 \times \theta_{\cdot\cdot 0} + 1 \times \theta_{\cdot\cdot 1} + 2 \times \theta_{\cdot\cdot 2} + 3 \times \theta_{\cdot\cdot 3} \end{pmatrix},$$

and may be expressed as $\mathbf{F}_2(\boldsymbol{\theta}) = \mathbf{A}_2\boldsymbol{\theta}$ with $\mathbf{A}_2 = ([0, 1, 2, 3]' \otimes \mathbf{1}_{16}, \mathbf{1}_4 \otimes [0, 1, 2, 3]' \otimes \mathbf{1}_4, \mathbf{1}_{16} \otimes [0, 1, 2, 3]')'$. In both cases, a saturated model for the functions of interest may be specified in the form $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix},$$

so that the parameters β_2 and β_3 included in $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ may be interpreted, respectively, as the effects of the second and third assessments with regard to the expected proportion of patients with response moderate or severe $[\mathbf{F}_1(\boldsymbol{\theta})]$, or with regard to the expected scores $[\mathbf{F}_2(\boldsymbol{\theta})]$. These linear models may be fitted by WLS with the following commands.

```
e121.raw<-data.frame(inicial=c(1,2,1,2,2,2,2,1,3,2,2,1,2,2,2,2,2,1,1,2,1,3,3,2,2,
   2,2,3,2,2,3,0,1,1,1,1,0,0,1,2,1,1,2,2,2,2,3,3,2,3),
                     dias14 =c(0,0,0,0,1,2,2,1,0,1,1,1,0,0,1,1,1,0,0,0,1,1,0,1,0,
   0,3,0,2,0,2,0,1,0,0,1,0,0,0,0,1,0,2,2,0,2,1,0,1,0),
                     dias21 =c(0,0,0,0,1,2,2,1,0,2,3,0,0,0,1,1,0,0,0,0,0,0,0,1,0,
   0,3,1,1,0,0,0,0,0,0,0,1,1,0,0,0,0,2,2,1,0,0,0,1,0))
table(e121.raw[,1]);table(e121.raw[,2]);table(e121.raw[,3]) #marginal distributions
```

```
e121.catdata<-readCatdata(TF=c(table(e121.raw[,3:1]))) #joint distribution
e121.v1<-c(0,0,1,1);e121.v2<-c(0,1,2,3)
e121.A1<-rbind(e121.v1%x%rep(1,16),rep(1,4)%x%e121.v1%x%rep(1,4),rep(1,16)%x%e121.v1)
e121.A2<-rbind(e121.v2%x%rep(1,16),rep(1,4)%x%e121.v2%x%rep(1,4),rep(1,16)%x%e121.v2)
e121.X<-rbind(c(1,0,0),c(1,1,0),c(1,0,1))
e121.propwls<-funlinWLS(model="lin",obj=e121.catdata,A1=e121.A1,X=e121.X)
e121.scorwls<-funlinWLS(model="lin",obj=e121.catdata,A1=e121.A2,X=e121.X)
```

Details of the fitted models, followed by Wald tests of the hypotheses of no treatment effect ($\beta_2 = \beta_3 = 0$) and of equality of effects of the second and the third assessments ($\beta_2 = \beta_3$), may be examined by means of the commands indicated in the sequel.

```
> e121.propwls

Call: funlinWLS(model = "lin", obj = e121.catdata, A1 = e121.A1, X = e121.X)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]   0.6599    0.0670     9.8502   0.0000
[2,]  -0.4998    0.0707    -7.0655   0.0000
[3,]  -0.5198    0.0707    -7.3538   0.0000

Wald goodness of fit statistic of the model (d.f.=0): 0 (p-value=1)

> waldTest(e121.propwls,rbind(c(0,1,0),c(0,0,1)))

Call: waldTest(obj = e121.propwls, C = rbind(c(0, 1, 0), c(0, 0, 1)))

Wald statistic of the hypothesis (d.f.=2): 57.9457 (p-value=0)

> waldTest(e121.propwls,c(0,1,-1))

Call: waldTest(obj = e121.propwls, C = c(0, 1, -1))

Wald statistic of the hypothesis (d.f.=1): 0.2003 (p-value=0.6545)

> e121.scorwls

Call: funlinWLS(model = "lin", obj = e121.catdata, A1 = e121.A2, X = e121.X)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]   1.7599    0.1116    15.7704   0.0000
[2,]  -1.0996    0.1422    -7.7326   0.0000
[3,]  -1.2195    0.1557    -7.8309   0.0000

Wald goodness of fit statistic of the model (d.f.=0): 0 (p-value=1)

> waldTest(e121.scorwls,rbind(c(0,1,0),c(0,0,1)))

Call: waldTest(obj = e121.scorwls, C = rbind(c(0, 1, 0), c(0, 0, 1)))

Wald statistic of the hypothesis (d.f.=2): 68.2345 (p-value=0)

> waldTest(e121.scorwls,c(0,1,-1))

Call: waldTest(obj = e121.scorwls, C = c(0, 1, -1))
```

```
Wald statistic of the hypothesis (d.f.=1): 1.4209 (p-value=0.2333)
```

The results suggest that the treatment has an effective impact on the selected characteristics of the response distribution during the first 14 days, but not after that period. To fit reduced models that incorporate these conclusions, it suffices to consider the specification matrix $\mathbf{X} = (\mathbf{1}_3, [0, \mathbf{1}_2']')$. The results (not shown) suggest that

- the proportion of patients with vaginal discharge moderate or severe at the first assessment is 66% $[CI(95\%) = (53\%, 79\%)]$ and this proportion decreases by 51% $[CI(95\%) = (38\%, 64\%)]$ at the second assessment, remaining at the same level at the third assessment;
- the average score associated to the vaginal discharge at the first assessment is 1.74 $[CI(95\%) = (1.53, 1.96)]$, and this average score decreases by 1.14 $[CI(95\%) = (0.86, 1.41)]$ at the second assessment, remaining unaltered at the third assessment.

We may conduct an alternative analysis based on models for the cumulative logits, where the marginal distributions are not summarized so drastically. The vector of functions of interest and the corresponding specification matrix for a proportional odds model are

$$\mathbf{F}_3(\boldsymbol{\theta}) = \mathbf{ln}\begin{pmatrix} \theta_{0..}/(\theta_{1..} + \theta_{2..} + \theta_{3..}) \\ (\theta_{0..} + \theta_{1..})/(\theta_{2..} + \theta_{3..}) \\ (\theta_{0..} + \theta_{1..} + \theta_{2..})/\theta_{3..} \\ \theta_{.0.}/(\theta_{.1.} + \theta_{.2.} + \theta_{.3.}) \\ (\theta_{.0.} + \theta_{.1.})/(\theta_{.2.} + \theta_{.3.}) \\ (\theta_{.0.} + \theta_{.1.} + \theta_{.2.})/\theta_{.3.} \\ \theta_{..0}/(\theta_{..1} + \theta_{..2} + \theta_{..3}) \\ (\theta_{..0} + \theta_{..1})/(\theta_{..2} + \theta_{..3}) \\ (\theta_{..0} + \theta_{..1} + \theta_{..2})/\theta_{..3} \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix},$$

where $\mathbf{F}_3(\boldsymbol{\theta}) = \mathbf{A}_{32}\,\mathbf{ln}(\mathbf{A}_{31}\boldsymbol{\theta})$ with $\mathbf{A}_{32} = \mathbf{I}_9 \otimes [1, -1]$, $\mathbf{A}_{31} = (\mathbf{A}_{30} \otimes \mathbf{1}_{16}, \mathbf{1}_4 \otimes \mathbf{A}_{30} \otimes \mathbf{1}_4, \mathbf{1}_{16} \otimes \mathbf{A}_{30})'$, and $\mathbf{A}_{30} = ([1,0,0,0]', [0,1,1,1]', [1,1,0,0]', [0,0,1,1]', [1,1,1,0]', [0,0,0,1]')'$. The first three elements of $\boldsymbol{\beta}$ are interpreted as the logarithms of the odds of (i) vaginal discharge absent *versus* vaginal discharge light, moderate or severe; (ii) vaginal discharge absent or light *versus* vaginal discharge moderate or severe; (iii) vaginal discharge absent, light or moderate *versus* vaginal discharge severe; all at the first assessment. The last two elements of $\boldsymbol{\beta}$ correspond to the logarithm of the effects of the second and third assessments relatively to the first assessment for the three odds simultaneously. We can fit this model by WLS using the commands

```
e121.A30<-rbind(c(1,0,0,0),c(0,1,1,1),c(1,1,0,0),c(0,0,1,1),c(1,1,1,0),c(0,0,0,1))
e121.A31<-rbind(e121.A30%x%t(rep(1,16)) , t(rep(1,4))%x%e121.A30%x%t(rep(1,4)) ,
 t(rep(1,16))%x%e121.A30)
e121.A32<-diag(9)%x%t(c(1,-1))
e121.X3<-cbind(rep(1,3)%x%diag(3),cbind(c(0,1,0),c(0,0,1))%x%rep(1,3))
e121.propoddswls<-funlinWLS(model=c("lin","log","lin"),obj=e121.catdata,
 A1=e121.A31,A2=e121.A32,X=e121.X3)
```

The output and Wald tests corresponding to hypotheses of interests may be obtained via the following commands

```
> e121.propoddswls

Call: funlinWLS(model = c("lin", "log", "lin"), obj = e121.catdata, ...)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]  -2.0636    0.3337    -6.1847   0.0000
[2,]  -0.5946    0.2893    -2.0551   0.0399
[3,]   1.4629    0.3544     4.1275   0.0000
[4,]   2.2650    0.3635     6.2306   0.0000
[5,]   2.5330    0.3840     6.5962   0.0000

Wald goodness of fit statistic of the model (d.f.=4): 5.248 (p-value=0.2628)

> waldTest(e121.propoddswls,rbind(c(0,0,0,1,0),c(0,0,0,0,1)))

Call: waldTest(obj = e121.propoddswls, C = rbind(c(0, 0, 0, 1, 0), c(0, 0, 0, 0, 1)))

Wald statistic of the hypothesis (d.f.=2): 45.9619 (p-value=0)

> waldTest(e121.propoddswls,c(0,0,0,1,-1))

Call: waldTest(obj = e121.propoddswls, C = c(0, 0, 0, 1, -1))

Wald statistic of the hypothesis (d.f.=1): 1.2974 (p-value=0.2547)
```

The results suggest that the proportional odds model is compatible with the data (p=0.26). Conclusions about the effects of the second and the third assessments are similar to those of the previous analyses. A reduced model that incorporates the restriction $\beta_4 = \beta_5$ may be fitted with the following commands

```
e121.X4<-cbind(rep(1,3)%x%diag(3),c(0,1,1)%x%rep(1,3))
e121.propoddswls2<-funlinWLS(model=c("lin","log","lin"),obj=e121.catdata,
 A1=e121.A31,A2=e121.A32,X=e121.X4)
```

To print the output, the following command is sufficient.

```
> e121.propoddswls2

Call: funlinWLS(model = c("lin", "log", "lin"), obj = e121.catdata, ...)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]  -2.0324    0.3325    -6.1119   0.0000
[2,]  -0.6029    0.2892    -2.0844   0.0371
[3,]   1.5193    0.3509     4.3292   0.0000
[4,]   2.3619    0.3534     6.6832   0.0000

Wald goodness of fit statistic of the model (d.f.=5): 6.5454 (p-value=0.2567).
```

The following commands allow us to obtain point and 95% interval estimates for the 3 odds of having lower *versus* higher vaginal discharge levels at the first assessment as well as quantifying the effect of the second and third assessments.

```
> exp(e121.propoddswls2$beta)
```

```
[1]  0.1310143  0.5472298  4.5690852 10.6113164
> exp(e121.propoddswls2$beta-qnorm(0.975)*sqrt(diag(e121.propoddswls2$Vbeta)))
[1] 0.06827484 0.31043196 2.29668402 5.30814268
> exp(e121.propoddswls2$beta+qnorm(0.975)*sqrt(diag(e121.propoddswls2$Vbeta)))
[1]  0.2514066  0.9646574  9.0898614 21.2126994
```

The results suggest that after 14 days the three odds are multiplied by $\exp(2.36) = 10.61$ $[CI(95\%) = (5.31, 21.21)]$ and that this remains unaltered at the third assessment.

Even though the results of the three analyses point to the same direction, the conclusions should be faced with an exploratory spirit, since the sample size may not be sufficiently large to warrant the asymptotic approximations required by the WLS approach.

Keeping this in mind, we may also focus on the estimation of the probability of change in the level of the symptom between two consecutive assessments. First, we need to obtain the second-order marginal distributions governed by the parameters $\{\theta_{ab\cdot} = \sum_c \theta_{abc}\}$, $\{\theta_{a\cdot c} = \sum_b \theta_{abc}\}$, and $\{\theta_{\cdot bc} = \sum_a \theta_{abc}\}$; these 48 probabilities grouped in the vector $\boldsymbol{\theta}_{\cdot} = (\theta_{00\cdot}, \theta_{01\cdot}, \ldots, \theta_{\cdot 33})'$ may be obtained from $\boldsymbol{\theta}$ via the linear function $\mathbf{A}_{41}\boldsymbol{\theta}$ with $\mathbf{A}_{41} = (\mathbf{I}_{16} \otimes \mathbf{1}_4, \mathbf{I}_4 \otimes \mathbf{1}_4 \otimes \mathbf{I}_4, \mathbf{1}_4 \otimes \mathbf{I}_{16})'$. Then, the vector $\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_{42}\,\mathbf{A}_{41}\,\boldsymbol{\theta} = (\sum_{a>b} \theta_{ab\cdot}, \sum_{b>c} \theta_{\cdot bc})'$ with

$$\mathbf{A}_{42} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

contains the probabilities of change between consecutive assessments. This, along with appropriate tests to compare the change from the first to the second assessments to that corresponding to second and third assessments may be obtained via the following commands.

```
> e121.A41<-rbind(diag(16)%x%t(rep(1,4)),diag(4)%x%t(rep(1,4))%x%diag(4),
+  t(rep(1,4))%x%diag(16))
> e121.A42<-rbind(c(1,0,0),c(0,0,1))%x%rbind(c(0,0,0,0,1,0,0,0,1,1,0,0,1,1,1,0))
> e121.improvprob<-funlinWLS(model=c("lin","lin"),obj=e121.catdata,A1=e121.A41,
+  A2=e121.A42,X=diag(2))
> e121.improvprob

Call: funlinWLS(model = c("lin", "lin"), obj = e121.catdata, A1 = e121.A41,...)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]   0.6799    0.0660    10.3052   0.0000
[2,]   0.2201    0.0586     3.7559   0.0002

Wald goodness of fit statistic of the model (d.f.=0): 0 (p-value=1)

> waldTest(e121.improvprob,c(1,-1))

Call: waldTest(obj = e121.improvprob, C = c(1, -1))

Wald statistic of the hypothesis (d.f.=1): 20.0042 (p-value=0)

> e121.improvprob$beta-qnorm(0.975)*sqrt(diag(e121.improvprob$Vbeta))
[1] 0.5505801 0.1052200
> e121.improvprob$beta+qnorm(0.975)*sqrt(diag(e121.improvprob$Vbeta))
[1] 0.8092000 0.3348812
> cov2cor(e121.improvprob$Vbeta)
           [,1]       [,2]
```

```
[1,]  1.0000000 -0.3602557
[2,] -0.3602557  1.0000000
```

The results suggest that the probability of change from the first to the second assessments, 68% [$CI(95\%) = (55\%, 81\%)$], is significantly higher (p<0.01) than the probability of change from the second to the third assessments, 22% [$CI(95\%) = (11\%, 33\%)$]. Note also that these estimates are negatively correlated, as a higher probability of change in the first 14 days naturally is associated to a lower probability of change from the 14^{th} to the 21^{st} day.

4 Missing data

A sample of 97 children was evaluated with two methods for assessing susceptibility to dental caries. The first, a standard method, is based on counts of *Lactobacillus* bacteria in salivary samples and the second, a simplified method, in the reaction of saliva with resarzurine. In both cases the children were classified as having high, medium, or low susceptibility to dental caries. The objective of the study is to compare both marginal distributions of susceptibility to dental caries and to evaluate the agreement between the classifications obtained with both methods. Given the subjective characteristic of the second method, it did not allow a complete classification of 46 children, highlighting the missing data nature of the response. The observed frequencies are displayed in Table 3.

Table 3: Observed frequencies of susceptibility to dental caries

Simplified method	Standard method		
	high	medium	low
high	7	11	2
medium	3	9	5
low	0	10	4
high / medium	8	7	3
medium / low	7	14	7

Firstly, we disregard the units with incomplete data and perform a complete case analysis (CCA) on the data of 51 children. We assume a multinomial distribution with θ_{ij} denoting the probability that a randomly selected child be classified in the i-th category with the first method and j-th category with the second. Here $i, j = 1$ (high), 2 (medium), 3 (low).

We can assess the homogeneity of the marginal distributions of susceptibility to dental caries obtained under both methods via the linear model $\mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$ where

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} [\mathbf{I}_2, \mathbf{0}_2] \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes [\mathbf{I}_2, \mathbf{0}_2] \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{1}_2 \otimes \mathbf{I}_2,$$

$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \ldots, \theta_{33})'$, and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. If there is no interest in estimating $\boldsymbol{\beta}$, we may use the equivalent constraint formulation $\mathbf{U}\mathbf{A}\boldsymbol{\theta} = \mathbf{0}_2$, with $\mathbf{U} = ([1, -1] \otimes \mathbf{I}_2)$. The commands required to fit this linear model (both formulations) by ML and WLS are

```
e132.ccadata<-readCatdata(TF=c(7,11,2,3,9,5,0,10,4))
e132.A<-rbind(cbind(diag(2),c(0,0))%x%t(rep(1,3)),t(rep(1,3))%x%cbind(diag(2),c(0,0)))
e132.X<-rep(1,2)%x%diag(2);e132.U<-t(c(1,-1)%x%diag(2))
e132.linmlcca<-linML(e132.ccadata,A=e132.A,X=e132.X)
e132.linwlscca<-funlinWLS(model="lin",obj=e132.ccadata,A1=e132.A,X=e132.X)
e132.linmlcca2<-linML(e132.ccadata,A=e132.A,U=e132.U)
e132.linwlscca2<-funlinWLS(model="lin",obj=e132.ccadata,A1=e132.A,U=e132.U)
```

The results are printed by simply typing the following commands.

```
> e132.linmlcca

Call: linML(obj = e132.ccadata, A = e132.A, X = e132.X)

Maximum likelihood estimates of the parameters of the linear model:
      estimate  std.error  z-value  p-value
[1,]   0.2975    0.0501     5.9339   0.0000
[2,]   0.4663    0.0449    10.3855   0.0000

Goodness of fit of the linear model (d.f.=2):
                  statistic  p-value
Likelihood ratio     7.6587   0.0217
Pearson              7.0594   0.0293
Neyman              11.6353   0.0030
Wald                 8.5581   0.0139

> e132.linwlscca

Call: funlinWLS(model = "lin", obj = e132.ccadata, A1 = e132.A, X = e132.X)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]   0.2741    0.0486     5.6382   0.0000
[2,]   0.4634    0.0453    10.2329   0.0000

Wald goodness of fit statistic of the model (d.f.=2): 8.5548 (p-value=0.0139)

> e132.linmlcca2

Call: linML(obj = e132.ccadata, A = e132.A, U = e132.U)

Goodness of fit of the linear model (d.f.=2):
                  statistic  p-value
Likelihood ratio     7.6587   0.0217
Pearson              7.0594   0.0293
Neyman              11.6353   0.0030
Wald                 8.5581   0.0139

> e132.linwlscca2

Call: funlinWLS(model = "lin", obj = e132.ccadata, A1 = e132.A, U = e132.U)

Wald goodness of fit statistic of the model (d.f.=2): 8.5548 (p-value=0.0139)
```

The results suggest that the marginal homogeneity hypothesis should be rejected ($p<0.03$).

In the context of the general structure with S subpopulations and R response categories

under a product-multinomial distribution, the linear model presented above is a special case of

$$\mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{A}$ is an $u \times SR$ matrix defining the u linear functions of interest with rank $r(\mathbf{A}) = u \leq S(R-1)$, $\mathbf{X}$ is a $u \times p$ model specification matrix of rank $r(\mathbf{X}) = p \leq u$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector that contains the unknown parameters. This model, may be expressed in the alternative constraint formulation

$$\mathbf{U}\mathbf{A}\boldsymbol{\theta} = \mathbf{0}_{u-p}, \tag{8}$$

where $\mathbf{U}$ is an $(u-p) \times u$ matrix containing the $u-p$ constraints with full rank and such that $\mathbf{U}\mathbf{X} = \mathbf{0}_{(u-p),p}$. In both cases, the rows of $\mathbf{A}$ must be linearly independent from the columns of the matrix $\mathbf{I}_S \otimes \mathbf{1}_R$ of natural constraints, *i.e.*, $r(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R) = u + S$. The default choice of the sub-routines is $\mathbf{A} = \mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]$; it selects the first $R-1$ components of each of the S multinomial distributions.

The functions `funlinWLS()` [with the argument `model="lin"`] and `linML()` fit linear models by WLS and ML, respectively. The models may be specified under the formulation (7) or (8). The arguments are the labels associated to the matrices used in the model expressions, *i.e.*, `X` and `U`, respectively, for $\mathbf{X}$ and $\mathbf{U}$. The exception lies in the matrix $\mathbf{A}$, which should be in the argument as `A1` for the WLS approach, and as `A`, for the ML procedure.

The agreement between the methods may be evaluated by the Cohen *kappa* index

$$\kappa = \frac{\sum_{i=1}^{3} \theta_{ii} - \sum_{i=1}^{3} \theta_{i\cdot}\theta_{\cdot i}}{1 - \sum_{i=1}^{3} \theta_{i\cdot}\theta_{\cdot i}},$$

where $\{\theta_{i\cdot} = \sum_j \theta_{ij}\}$ and $\{\theta_{\cdot j} = \sum_i \theta_{ij}\}$. This index of agreement may be written as a functional linear model $\mathbf{F}(\boldsymbol{\theta}) = \boldsymbol{\pi}_1 + \mathbf{exp}(\mathbf{A}_4\ \mathbf{ln}\{\mathbf{A}_3\ \mathbf{exp}[\mathbf{A}_2\ \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})]\}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{A}_1 = \begin{bmatrix} (\mathbf{1}_2' \otimes [1, \mathbf{0}_3'] \ , \ 1) \\ \mathbf{1}_9' \\ \mathbf{I}_3 \otimes \mathbf{1}_3' \\ \mathbf{1}_3' \otimes \mathbf{I}_3 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2,6} \\ \mathbf{0}_{3,2} & \mathbf{1}_2' \otimes \mathbf{I}_3 \end{bmatrix},$$

$$\mathbf{A}_3 = \left[(1,0)', \mathbf{1}_2, -(2,1)'\mathbf{1}_3'\right], \quad \mathbf{A}_4 = [1, -1], \quad \boldsymbol{\pi}_1 = -1,$$

$\mathbf{X} = 1$ and $\boldsymbol{\beta} = \kappa$, with $\mathbf{exp}(\mathbf{a})$ denoting the vector exponential operator, the elements of which correspond to the exponentials of the elements of $\mathbf{a}$. We may fit this model by WLS via the following commands.

```
e132.kA1<-rbind(c(diag(3)),rep(1,9),diag(3)%x%t(rep(1,3)),t(rep(1,3))%x%diag(3))
e132.kA2<-rbind(cbind(diag(2),matrix(0,2,6)),cbind(matrix(0,3,2),t(rep(1,2))%x%diag(3)))
e132.kA3<-cbind(c(1,0),c(1,1),-c(2,1)%*%t(rep(1,3)));e132.kA4<-t(c(1,-1))
e132.kappacca<-funlinWLS(model=c("add","exp","lin","log","lin","exp","lin","log","lin"),
 obj=e132.ccadata,A1=e132.kA1,A2=e132.kA2,A3=e132.kA3,A4=e132.kA4,PI1=-1,X=1)
```

The output may be printed by typing

```
> e132.kappacca

Call: funlinWLS(model = c("add", "exp", "lin", "log", "lin", "exp", "lin", "log", ...)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]  0.0898    0.0998     0.8995   0.3684

Wald goodness of fit statistic of the model (d.f.=0): 0 (p-value=1)
```

The results suggest that the agreement between the simplified and the standard methods does not appear to be better than what is expected by chance since the Wald test does not reject the hypothesis that $\kappa = 0$ (p=0.37).

For partially classified categorical data, the user must also inform which are the response categories associated to each of the response classes; this is accomplished via the arguments `Zp` and `Rp` in the `readCatdata()` function as follows.

It is assumed that every missingness pattern has response classes that jointly constitute a partition of the response categories. The dataset under study, for example, has two missingness patterns: in the first, there is no distinction between the high and medium categories obtained from the simplified method; in the second, there is no distinction between the medium and low categories obtained by the simplified method. From Table 3 we notice that there are 3 frequencies associated with the first missingness pattern: 8, associated to the parameters θ_{11} and θ_{21}, 7, to θ_{12} and θ_{22}, and 3, to θ_{13} and θ_{23}. As the 3 response classes do not constitute a partition of the response categories, we define a last class with a null frequency associated to the parameters θ_{31}, θ_{32} and θ_{33}. Following the order in $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$, the 4 response classes, with frequencies $(8, 7, 3, 0)$, are respectively informed via the response indicator vectors $(1,0,0,1,0,0,0,0,0)'$, $(0,1,0,0,1,0,0,0,0)'$, $(0,0,1,0,0,1,0,0,0)'$ and $(0,0,0,0,0,0,1,1,1)'$, where the components equal to 1 indicate that the corresponding response categories of $\boldsymbol{\theta}$ are in the response class being defined; otherwise, the components of the response indicator vectors are equal to 0. Similarly, the second missingness pattern has 4 response classes, with frequencies $(0, 7, 14, 7)$, defined by the response indicator vectors $(1,1,1,0,0,0,0,0,0)'$, $(0,0,0,1,0,0,1,0,0)'$, $(0,0,0,0,1,0,0,1,0)'$ and $(0,0,0,0,0,1,0,0,1)'$. These 8 vectors are stacked side by side, forming the following 9×8 matrix

$$\left[\begin{array}{cccc|cccc} 1&0&0&0&1&0&0&0\\ 0&1&0&0&1&0&0&0\\ 0&0&1&0&1&0&0&0\\ 1&0&0&0&0&1&0&0\\ 0&1&0&0&0&0&1&0\\ 0&0&1&0&0&0&0&1\\ 0&0&0&1&0&1&0&0\\ 0&0&0&1&0&0&1&0\\ 0&0&0&1&0&0&0&1 \end{array}\right] = \left[\begin{array}{cc|cc} \mathbf{1}_2 \otimes \mathbf{I}_3 & \mathbf{0}_6 & \mathbf{1}_3 & \mathbf{0}_{3,3} \\ \mathbf{0}_{3,3} & \mathbf{1}_3 & \mathbf{0}_6 & \mathbf{1}_2 \otimes \mathbf{I}_3 \end{array}\right],$$

which must be specified in the argument `Zp`. The argument `Rp` must receive a vector indicating the number of response classes in each missingness pattern. For our example, `Rp`=(4,4). The commands to input the incomplete categorical data are

```
e132.TF<-c(7,11,2,3,9,5,1e-5,10,4, 8,7,3,0, 0,7,14,7)
e132.Zp<-cbind(rbind( cbind(kronecker(rep(1,2),diag(3)),rep(0,6)),
                      cbind(matrix(0,3,3),rep(1,3)) ),
               rbind( cbind(rep(1,3),matrix(0,3,3)),
                      cbind(rep(0,6),kronecker(rep(1,2),diag(3))) ) )
e132.catdata<-readCatdata(TF=e132.TF,Zp=e132.Zp,Rp=c(4,4))
```

Note that the null frequency of the completed categorized pattern was substituted by a small value (10^{-5}) to avoid estimates on the boundary of the parameter space during the iteration process. The `summary()` function applied to the object produced by the function `readCatdata()` generates useful output to check if the function addresses the missingness patterns correctly.

```
> summary(e132.catdata)

Call: readCatdata(TF = e132.TF, Zp = e132.Zp, Rp = c(4, 4))


S=1 subpopulations  x  R=9 response categories  with MISSING data


Table of frequencies of the complete data:
[1]  7.0e+00  1.1e+01  2.0e+00  3.0e+00  9.0e+00  5.0e+00  1.0e-05  1.0e+01  4.0e+00

Proportions of the complete data:
      [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]    [,9]
[1,]  0.1373  0.2157  0.0392  0.0588  0.1765  0.0980  0.0000  0.1961  0.0784

Standard errors of the proportions of the complete data:
      [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]    [,9]
[1,]  0.0482  0.0576  0.0272  0.0329  0.0534  0.0416  0.0001  0.0556  0.0376


Missing data frequencies and associated column vectors indicating
the relation with the original set of R response categories:

             [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,]  8         1     0     0     1     0     0     0     0     0
[2,]  7         0     1     0     0     1     0     0     0     0
[3,]  3         0     0     1     0     0     1     0     0     0
[4,]  0         0     0     0     0     0     0     1     1     1

              [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,]   0         1     1     1     0     0     0     0     0     0
[2,]   7         0     0     0     1     0     0     1     0     0
[3,]  14         0     0     0     0     1     0     0     1     0
[4,]   7         0     0     0     0     0     1     0     0     1
```

The output produced by `readCatdata()` always exhibits the observed proportions for the complete data pattern. The function `satMarML()` is used to estimate $\boldsymbol{\theta}$ by ML under the MAR mechanism (default) or the MCAR mechanism (when the argument `missing="MCAR"`).

```
> e132.satmarml<-satMarML(e132.catdata)
> e132.satmarml

Call: satMarML(catdataobj = e132.catdata)
```

```
S=1 subpopulations  x  R=9 response categories

Maximum likelihood estimates of the probabilities:
      [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]    [,9]
[1,]  0.1061  0.1418  0.0260  0.1516  0.2188  0.1241  0.0000  0.1652  0.0664

Standard errors (MAR):
      [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]    [,9]
[1,]  0.0358  0.0385  0.0179  0.0404  0.0520  0.0384  0.0001  0.0450  0.0302

Goodness of fit statistics of MCAR given MAR assumption (d.f.=6)
                  statistic    p-value
Likelihood ratio    35.9325     0.0000
Pearson             24.4088     0.0004
Neyman            7854.1068     0.0000
```

The results suggest that the MCAR assumption is rejected ($p<0.01$). Comparing the last two outputs, we notice that the marginal proportions obtained from the complete data may be quite different from those obtained via the ML estimates of $\{\theta_{ij}\}$ using all the available data. The previous models for $\boldsymbol{\theta}$ may be fitted again on the object generated by the function `satMarML()`. The new fitted models will inherit the properties of the MAR mechanism from the informed object. The necessary commands follow.

```
> e132.linmlmar<-linML(e132.satmarml,A=e132.A,X=e132.X)
> e132.linmlmar

Call: linML(obj = e132.satmarml, A = e132.A, X = e132.X)

Maximum likelihood estimates of the parameters of the linear model under MAR:
      estimate  std.error  z-value  p-value
[1,]   0.2649    0.0361     7.3332   0.0000
[2,]   0.5135    0.0372    13.7931   0.0000

Goodness of fit of the linear model given MAR (d.f.=2):
                  statistic  p-value
Likelihood ratio     0.1287   0.9377
Pearson              0.1289   0.9376
Neyman               0.1287   0.9377
Wald                 0.1285   0.9378

Goodness of fit of the linear model and MCAR given MAR (d.f.=8):
                  statistic    p-value
Likelihood ratio    36.0612     0.0000
Pearson             24.7743     0.0017
Neyman            7327.0080     0.0000

> e132.linwlsmar<-funlinWLS(model="lin",obj=e132.satmarml,A1=e132.A,X=e132.X)
> e132.linwlsmar

Call: funlinWLS(model = "lin", obj = e132.satmarml, A1 = e132.A, X = e132.X)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]   0.2649    0.0361     7.3363   0.0000
[2,]   0.5135    0.0373    13.7855   0.0000
```

```
Wald goodness of fit statistic of the model (d.f.=2): 0.1285 (p-value=0.9378)

> e132.kappamar<-funlinWLS(model=c("add","exp","lin","log","lin","exp","lin","log","lin"),
+  obj=e132.satmarml,A1=e132.kA1,A2=e132.kA2,A3=e132.kA3,A4=e132.kA4,PI1=-1,X=1)
> e132.kappamar

Call: funlinWLS(model = c("add", "exp", "lin", "log", "lin", "exp", "lin", "log", ...)

Weighted least squares estimates of the parameters of the model:
      estimate  std.error  z-value  p-value
[1,]  0.0171    0.1015     0.1682   0.8664

Wald goodness of fit statistic of the model (d.f.=0): 0 (p-value=1)
```

In contrast to the results obtained under the CCA, there is no evidence against the marginal homogeneity of the distributions associated to both methods under the MAR mechanism (p>0.90). However, the previous conclusion about the κ index remains valid (p=0.87).

The reader may refer to Poleto (2007) for an example on how to fit MNAR mechanisms, and to Poleto *et al.* (2007a) for an illustration on how to input partially classified categorical data under a product-multinomial setting, where the subpopulations may have different missingness patterns.

References

AGRESTI, A. (2002). *Categorical data analysis.* 2nd ed. New York: John Wiley & Sons.

BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975). *Discrete multivariate analysis: theory and practice.* Cambridge: The MIT Press.

FORTHOFER, R.N. and LEHNEN, R.G. (1981). *Public program analysis: a new categorical data approach.* Belmont: Wadsworth.

IMREY, P.B., KOCH, G.G., STOKES, M.E. *et al.* (1981). Categorical data analysis: some reflections on the log linear model and logistic regression. Part I: historical and methodological overview. *International Statistical Review* **49**, 265-283.

IMREY, P.B., KOCH, G.G., STOKES, M.E. *et al.* (1982). Categorical data analysis: some reflections on the log linear model and logistic regression. Part II: data analysis. *International Statistical Review* **50**, 35-63.

KOCH, G.G., IMREY, P.B., SINGER, J.M., ATKINSON, S.S. and STOKES, M.E. (1985). *Analysis of categorical data.* Montréal: Les Presses de L'Université de Montréal.

LANDIS, J.R., STANISH, W.M., FREEMAN, J.L. and KOCH, G.G. (1976). A computer program for the generalized chi-square analysis of categorical data using weighted least squares (GENCAT). *Computer Programs in Biomedicine* **6**, 196-231.

LITTLE, R.J.A. and RUBIN, D.B. (2002). *Statistical analysis with missing data.* 2nd ed. New York: John Wiley & Sons.

PAULINO, C.D. and SINGER, J.M. (2006). *Analysis of categorical data* (in Portuguese). Edgard Blücher: São Paulo.

POLETO, F.Z. (2007). *Commands (in R) to reproduce the analyses of the examples of the book Analysis of Categorical Data by Paulino and Singer (2006)* (in Portuguese). Unpublished manuscript. Available at `http://www.poleto.com/missing.html`.

POLETO, F.Z., SINGER, J.M. and PAULINO, C.D. (2007a). *Analyzing categorical data with complete or missing responses using the Catdata package*. Unpublished vignette for the R package. Available at http://www.poleto.com/missing.html.

POLETO, F.Z., SINGER, J.M. and PAULINO, C.D. (2007b). A product-multinomial framework for categorical data analysis with missing responses. Submitted for publication. Available at `http://www.poleto.com/missing.html`.

POLETO, F.Z., SINGER, J.M. and PAULINO, C.D. (2007c). Comparing diagnostic tests with missing data. Submitted for publication. Available at `http://www.poleto.com/missing.html`.

R DEVELOPMENT CORE TEAM (2007). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. `http://www.r-project.org`.

Acknowledgements

This research received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil and Fundação para a Ciência e Tecnologia (FCT) through the research centre CEMAT-IST, Portugal.