Journal of Statistical Software

December 2015, Volume 68, Issue 10. doi: 10.18637/jss.v068.i10

Semi-Parametric Maximum Likelihood Method for Interaction in Case-Mother Control-Mother Designs: Package SPmlficmcm

Molière Nguile-Makao
Centre de Recherche de l'Institut
Universitaire en Santé Mentale de Québec

Alexandre Bureau
Université Laval

Abstract

The analysis of interaction effects involving genetic variants and environmental exposures on the risk of adverse obstetric and early-life outcomes is generally performed using standard logistic regression in the case-mother and control-mother design. However such an analysis is inefficient because it does not take into account the natural family-based constraints present in the parent-child relationship. Recently, a new approach based on semi-parametric maximum likelihood estimation was proposed. The advantage of this approach is that it takes into account the parental relationship between the mother and her child in estimation. But a package implementing this method has not been widely available. In this paper, we present **SPmlficmcm**, an R package implementing this new method and we propose an extension of the method to handle missing offspring genotype data by maximum likelihood estimation. Our choice to treat missing data of the offspring genotype was motivated by the fact that in genetic association studies where the genetic data of mother and child are available, there are usually more missing data on the genotype of the offspring than that of the mother. The package builds a non-linear system from the data and solves and computes the estimates from the gradient and the Hessian matrix of the log profile semi-parametric likelihood function. Finally, we analyze a simulated dataset to show the usefulness of the package.

Keywords: early-life outcome, genetic variants, logistic model, missing genotype data, mother-child pairs, R package.

1. Introduction

We focus on the problem of analyzing interaction effects involving genetic variants and environmental exposures on the risk of adverse obstetric and early-life outcomes such as premature birth and preeclampsia, and small-for-gestational-age (SGA) neonates. Obstetric and early-life outcomes involve the mother and her child. Several epidemiological studies suggest that the fetal susceptibility to environmental factors depends on his or her genotype and on the genotype of his or her mother (Infante-Rivard 2007). It thus appears important to include both genotypes as well as the environmental factors as predictors in the same model. Usually, standard logistic regression is used in the case-mother and control-mother design. However such analysis is inefficient here because it does not take into account the natural family-based constraints present in the parent-child relationship (Shi, Umbach, Vermeulen, and Weinberg 2008).

Recently, Chen, Lin, and Hochner (2012) proposed an extension of the approach based on semi-parametric maximum likelihood estimation. The extension proposed by Chen *et al.* (2012) in case-mother and control-mother design is interesting, because it takes into account both the genotype of the mother and of the child. The parental link between mother and her child is modeled through the parametric function of the joint distribution of the maternal and fetal genotype under the assumptions of random mating, Hardy-Weinberg equilibrium (HWE), and Mendelian inheritance. The distribution of the environmental variables given the maternal genotype and child genotype is considered as a nuisance parameter in the estimation. Furthermore, the authors make the assumption that the environmental factors are linked only to the genetic profile of the mother and not the one of her child, reducing the dimension of the nuisance parameter and permiting to simplify the likelihood function. In their approach, they also assume that the totals of case and control mother-child pairs eligible for recruitment (population totals) are available. The authors show by simulation studies the greater efficiency of their approach for estimating the association parameters compared to logistic regression.

Chen *et al.* (2012)'s method is actually appealing because of the wide array of studies of obstetric and early-life outcomes where it can be applied. For example, the study seeking to link SGA neonates to drinking water disinfection by-products conducted on case-mother and control-mother pairs from Quebec City (Canada) area (Levallois *et al.* 2012) or that of Infante-Rivard (2004), where they used logistic regression to study the modifying effect of genetic variants. A package implementing this method has not been widely available. In this paper, we present **SPmlficmcm** (Nguile-Makao and Bureau 2015), an R (R Core Team 2015) package implementing this method and an extension of the method to handle missing offspring genotype data by maximum likelihood (Allison 2001). The R package is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=SPmlficmcm`.

Our choice of how to treat missing offspring genotype data was motivated by the fact that in genetic association studies where the genetic data of mother and child are available, there are usually more missing data on the genotype of the offspring than that of the mother. Parents may be reluctant to consent to collect saliva and blood samples from their children. And even with the parent's consent, it is not easy to get a sufficient sample of saliva from the children. For example, in a sample of mother-child pairs from the Quebec City area (SGA cases and controls), among 1719 genotyped mother-child pairs we found on average 1.5% of missing maternal genotypes and on average 8% of missing offspring genotypes.

This article is structured in the following way: In Section 2, we describe Chen *et al.* (2012)'s method and the extension that we propose. In Section 3, we present a description of the package. In Section 4, we present an illustration on simulated data. And we finish by a discussion in Section 5.

2. Mathematical background

In this section, we will review briefly the semi-parametric maximum likelihood method (SML) proposed by Chen *et al.* (2012) and an extension of this method to the treatment of missing offspring genotype data that we propose. In the rest of the paper, we will denote this extension by SMLMD. Section 2.1 presents the notations, Section 2.2 gives the empirical log-likelihood function for the complete data with a new parametrization and Section 2.3 presents the empirical log-likelihood function for missing offspring genotype data.

2.1. Notations

Let Y denote the binary case-control status, G^M and G^C the respective maternal and offspring genotype, X the vector of environmental variables collected from the mother and G the set of possible values of the genotype. Data $(Y,\ X,\ G^M,\ G^C)$ is collected from n_1 case mother pairs and n_0 control mother pairs, which are sampled from N_1 ($N_1 > n_1$) case pairs and N_0 ($N_0 > n_0$) control pairs. Let n_{ijmc} denote the total number of pairs in the case-control sample with $Y = i$, $X = j$, $G^M = m$ and $G^C = c$ and $P_{ijmc}(\beta)$ their probability with $P_{ijmc}(\beta) = \mathsf{P}(Y = i \mid X = j,\ G^M = m,\ G^C = c;\ \beta)$ and θ the log odds of minor allele frequency (MAF) Under random mating, Hardy Weinberg equilibrium (HW) and Mendelien inheritance, the joint distribution, $\mathsf{P}(G^M = m,\ G^C = c) = P_{c|m}(\theta)P_m(\theta)$, $\forall(m,c) \in\ G^M \times G^C$ where $P_m(\theta) = \mathsf{P}(G^M = m;\ \theta)$ represents the mother genotype and $P_{c|m}(\theta) = \mathsf{P}(G^C = c|G^M = m;\ \theta)$ the conditional offspring genotype distribution given maternal genotype. We denote by $C_{jm} = \sum\limits_{i,\ c} n_{ijmc}$ the number of subjects in the case-control sample with $X = j$ and $G^M = m$.

2.2. General semi-parametric maximum likelihood estimation

Let first suppose the data collected on the mother-child pairs $(Y,\ X,\ G^M,\ G^C)$, has no missing data. For all subjects in the sample with $Y = i$, $X = j$, $G^M = m$ and $G^C = c$, we define the following functions: $h_{ijmc}(\beta,\ \theta) = P_{ijmc}(\beta)P_{c|m}(\theta)$ where $P_{ijmc}(\beta)$ is derived from a logistic regression model, and $P_{c|m}(\theta)$ is the conditional distribution of children genotype given maternal genotype. Let n_{ijmc} be the number of mother-child pairs having the status $(i,\ j,\ m,\ c)$. Chen *et al.* (2012) make the following assumption that the paternal allele of offspring genotype is independent from environmental factors, by consequence, $\mathsf{P}(X = j \mid G^M = m,\ G^C = c) = \mathsf{P}(X = j \mid G^M = m) = \delta_{jm}$ and δ_{jm} satisfies the constraint $\sum\limits_j \delta_{jm} = 1$. The empirical log-likelihood is written:

$$\begin{aligned}\ell(\beta,\ \theta,\ \delta_{jm}) = & \sum_{i,\ j,\ m,\ c} n_{ijmc}\log(h_{ijmc}(\beta,\ \theta)) + \sum_{j,\ m} C_{jm}\log(\delta_{jm}P_m(\theta)) \\ & + \sum_i \left[d_i \log\left(\sum_{(j,\ m,\ c)\in V_x\times G^2} \delta_{jm}h_{ijmc}(\beta,\ \theta)P_m(\theta) \right) \right],\end{aligned} \tag{1}$$

where V_x is the set of observed values of the vector X and the maternal and child genotype G^M and G^C in the data and $d_i = N_i - n_i$. For more detail, see Chen *et al.* (2012). The semi-parametric likelihood estimator is obtained in the following way:

Step 1: We estimate δ_{jm} using Lagrange multipliers for the constraints $\sum\limits_{j}\delta_{jm} = 1$ and we define

$$\forall i,\ m\ u_{im}(\beta,\ \theta) = 1\ -\ \frac{d_i P_m(\theta)}{N^*_m(\beta,\ \theta)\mathsf{P}(Y = i)}, \tag{2}$$

where

$$N^*_m(\beta,\ \theta) = \sum_j \frac{C_{jm}}{\sum\limits_{i,\ c} h_{ijmc} u_{im}(\beta,\ \theta)}.$$

Following Chen *et al.* (2012) we obtain a closed-form expression for $\widehat{\delta}_{jm}$ as

$$\widehat{\delta}_{jm} = \frac{C_{jm}}{N^*_m(\beta,\ \theta)\sum\limits_{i,\ c} h_{ijmc}(\beta,\ \theta)u_{im}(\beta,\ \theta)}, \qquad \forall j,\ m.$$

We note that the $u_{im}(\beta,\ \theta)$, $i = 0,\ 1$, $m \in G^M$ constitute a non-linear equation system. We have

$$\mathsf{P}(Y = i) = \sum_{j,\ m,\ c} h_{ijmc}(\beta,\ \theta)\delta_{jm}P_m(\theta), \tag{3}$$

and when we plug $\widehat{\delta}_{jm}$ in Equation 3 and we substitute the expression obtained for $\mathsf{P}(Y = i)$ in Equation 2, we have the following equation system $\forall i,\ m$:

$$u_{im}(\beta,\ \theta) = 1\ -\ \frac{d_i P_m(\theta)}{N^*_m(\beta,\ \theta)\sum\limits_{m' \in G}\frac{P_{m'}(\theta)}{N^*_{m'}(\beta,\ \theta)}\sum\limits_j \frac{C_{jm'}}{f_{jm'}(\beta,\ \theta;\ u_{m'})}\sum\limits_c h_{ijm'c}(\beta,\ \theta)}, \tag{4}$$

where $f_{jm'}(\beta,\ \theta; u_{m'}) = \sum\limits_{i,\ c} h_{ijm'c}(\beta,\ \theta)u_{im'}(\beta,\theta)$ and $u_{m'} = \{u_{im'};\ i = 1,\ 2\}$. We note that, for all $i,\ m$ the $u_{im}(\beta,\theta)$ are bounded. Chen *et al.* (2012) define, $N^*_m(\beta,\ \theta) = n_{++m+} + \sum\limits_i d_i\mathsf{P}(G^m = m \mid Y = i)$.

If we denote $n_{++m+} = \sum\limits_j C_{jm}$ and $\widehat{\rho}_i = \frac{\#\{u;\ Y_u = i\}}{n}$, an empirical estimator of $\mathsf{P}(Y = i)$, and $n = n_0\ +\ n_1$, then we have the inequalities:

$$1\ -\ \frac{d_i P_m(\theta)}{\widehat{\rho}_i n_{++m+}}\ \leq\ u_{im}(\beta,\ \theta)\ \leq\ 1\ -\ \frac{d_i P_m(\theta)}{\widehat{\rho}_i(n_{++m+}\ +\ N - n)}, \tag{5}$$

with $N - n = N_0\ +\ N_1\ -\ (n_0 + n_1)$.

Step 2: We solve the non-linear system in Equation 4. We denote by $\widehat{u}_{im}(\beta,\theta)$ the solution of the non-linear system, we plug $\widehat{\delta}_{jm}$ as well as $\widehat{u}_{im}(\beta,\theta)$ in Equation 1 to obtain the log profile likelihood:

$$\begin{aligned}\ell^p(\beta,\ \theta,\ \widehat{u}_{im}(\beta,\ \theta))\ =&\\ \sum_{i,\ j,\ m,\ c} n_{ijmc}\log(h_{ijmc}(\beta,\ \theta)) + \sum_{jm} C_{jm}\log&\left(\frac{P_m(\theta)}{N^*_m(\beta,\ \theta)f_{jm}(\beta,\ \theta,\ \widehat{u}_{im})}\right)\\ +\sum_i&\left[d_i\log\left\{\sum_m\frac{P_m(\theta)}{N^*_m(\beta,\ \theta)}\sum_j\frac{C_{jm}}{f_{jm}(\beta,\ \theta,\ \widehat{u}_m)}h_{ijm}(\beta,\ \theta)\right\}\right].\end{aligned} \tag{6}$$

Let $\widehat{\eta}_0 = (\widehat{\beta}_0,\ \widehat{\theta}_0)$ be initial values of the parameters $\eta = (\beta,\ \theta)$ obtained respectively by the modified logistic regression of Chen *et al.* (2012) and the following equation:

$$\sum_i \sum_u^{n_i} \frac{N_i}{n_i} \frac{\partial}{\partial \theta} \log \mathsf{P}(G^M = m_u,\ G^C = c_u;\ \theta) = 0. \tag{7}$$

Finally, an estimate of the parameter η is obtained by the following formula:

$$\widehat{\eta} = \widehat{\eta}_0 - \left\{ \frac{\partial^2 \ell^p}{\partial \eta \partial \eta}(\widehat{\eta}_0) \right\}^{-1} \frac{\partial \ell^p}{\partial \eta}(\widehat{\eta}_0). \tag{8}$$

Note that the variance of $\widehat{\eta}$ is estimated by the gradient and the inverse of the Hessian matrix of the log profile likelihood evaluated at point $\widehat{\eta}$. In numerical computations of the gradient, we have checked that keeping the $\widehat{u}_{im}(\beta,\ \theta)$ fixed to the solution for the specified values of β and θ or resolving them at each evaluation of ℓ^p leads to nearly identical values. We have therefore implemented the analytical gradient of the log profile likelihood function ℓ^p for fixed values of $\widehat{u}_{im}(\beta,\ \theta)$. The Hessian matrix is computed numerically from the gradient.

2.3. Generalization to missing offspring genotype data

Let $(Y,\ X,\ G^M,\ G^C)$ be the data collected on the mother-child pairs where we suppose that the genotype of a subset of children is missing at random (Little and Rubin 2002), as missingness is allowed to depend on case-control status and maternal genotype. We suppose that the missing offspring genotype data are completely at random. From the total sample, we constitute two sub-samples that we denote by S^1_{IJMC} for the complete data and S^2_{IJM} for the missing offspring genotype data. If we examine the non-linear system of the complete data in Equation 4, we notice that the system is completely defined through the information $(i,\ j,\ m)$. The offspring genotypes in the sample are not required because the system is written with a summation over all possible values of the offspring genotype compatible with the maternal genotype. Consequently, the non-linear system with missing offspring genotype data stays identical to the one for the complete data. Starting from the likelihood function of Equation 6 for the complete data, ℓ^p can be written the following way:

$$\ell^p(\beta,\ \theta,\ \widehat{u}_{mi}(\beta,\ \theta)) = \ell^p_1(\beta,\ \theta)\ +\ \ell^p_2(\beta,\ \theta, \widehat{u}_{im}(\beta,\ \theta)), \tag{9}$$

where

$$\ell^p_1(\beta,\ \theta) = \sum_{i,\ j,\ m,\ c} n_{ijmc} \log(h_{ijmc}(\beta,\ \theta))$$

and $\ell^p_2(\beta,\ \theta,\ \widehat{u}_{im}(\beta,\ \theta))$ is equal to the remaining terms of the right-hand side of Equation 6. The function $\ell^p_1(\beta,\ \theta)$ is completely defined only if we observe n_{ijmc} in the study sample. However, we only need the totals for each combination of $(i,\ j,\ m)$ in the study sample to define the function $\ell^p_2(\beta,\ \theta,\ \widehat{u}_{im}(\beta,\ \theta))$. As a consequence, the missing offspring genotype data modify the likelihood function of Equation 6 only within the function $\ell^p_1(\beta,\ \theta)$. For all $(i,\ j,\ m)$, we denote by $n_{ijm} = \sum\limits_c n_{ijmc}$ the number of mother-child pairs in the samples with $Y = i$, $X = j$, $G^M = m$ and where the offspring genotype is observed. We denote also for all $(i,\ j,\ m)$, $\overline{n}_{ijm}$ the number of mother-child pairs in the sample with $Y = i$, $X = j$,

$G^M = m$ and where the offspring genotype is missing. The log profile semi-parametric likelihood function with summation over missing offspring genotype is written:

$$\ell^p(\beta,\ \theta,\ \widehat{u}_{mi}(\beta,\ \theta)) = \ell_1^p(\beta,\ \theta)\ +\ \overline{\ell}_1^p(\beta,\ \theta)\ +\ \ell_2^p(\beta,\ \theta,\ \widehat{u}_{im}(\beta,\ \theta)), \tag{10}$$

where

$$\overline{\ell}_1^p(\beta,\ \theta) = \sum_{i,\ j,\ m} \overline{n}_{ijm} \log\left(\sum_{c\in\mathbf{G}} h_{ijmc}(\eta)\right).$$

This quantity represents the modification brought by the missing offspring genotype data to the likelihood function. The log profile semi-parametric likelihood is then written:

$$\begin{aligned} l^p(\beta,\ \theta,\ \widehat{u}_{im}(\eta)) = &\sum_{ijmc\in S^1_{IJMC}} n_{ijmc}\log(h_{ijmc}(\eta))\ +\ \sum_{ijm\in S^2_{IJM}} \overline{n}_{ijm}\log\left(\sum_{c\in\mathbf{G}^M} h_{ijmc}(\eta)\right) \\ &+\ \sum_{jm} C_{jm}\log\Big(\frac{P_m(\theta)}{N^*_m(\eta) f_{jm}(\eta,\ \widehat{u}_m(\eta))}\Big) \\ &+\ \sum_i \left[d_i \log\left\{\sum_m \frac{P_m(\theta)}{N^*_m(\eta)} \sum_j \frac{C_{jm}}{f_{jm}(\eta,\ \widehat{u}_m(\eta))} h_{ijm}(\eta)\right\}\right]. \end{aligned} \tag{11}$$

Remark: Tthe distribution of the offspring genotype is conditional on the mother genotype and not on the covariate values, and so is the missingness probability factoring out of the likelihood. We apply the same steps as in Section 2.2 to obtain the parameter estimates.

3. Package description

The R package **SPmlficmcm** (*semi-parametric maximum likelihood for interaction in case-mother control-mother*) implements the method of general semi-parametric maximum likelihood estimation for the complete data and data with missing offspring genotype. It contains one main function for the user: `Spmlficmcm()` performing the analysis for the complete data and data with missing offspring genotype. This function uses two auxiliary functions: `Est.Inpar()` to compute the initial values of the parameters (β, θ) and of the non-linear system, and `Nlsysteq()` to build the non-linear system. Finally the function `Spmlficmcm()` solves the non-linear system, builds the log profile likelihood and its gradient and computes the parameter estimates as well as the estimates of their standard errors. In this section, we describe the main functions and the estimation steps.

3.1. Main arguments of the functions

All functions use the main arguments described below and some functions use optional arguments. The main arguments are:

- `formula`: The model formula.
- `N`: A numeric vector of length two giving the number of eligible controls and cases in the population (`N` = (`N0`, `N1`)). If this information is unavailable, it is possible to specify

the disease population prevalence in the argument `p` instead of `N`. In that case, `N1` is set equal to 5 `n1`, in order to avoid observing `N1` < `n1` when prevalence is small. We then set `N0` $= \frac{1-p}{p}$`N1`.

- `gnma`: A character variable representing the name of the maternal genotype.
- `gnch`: A character variable representing the name of the child genotype.
- `start`: Vector of the initial values of the model parameters.
- `data`: A data frame in long format containing the following variables:
  - `id`: Identity of the mother-child pairs.
  - `outc`: Binary case-control status.
  - `gnma`: The maternal genotype.
  - `gnch`: The offspring genotype.

Remark: In the formula, the variables coding the maternal and child genotypes can differ from the `gnma` and `gnch` variables representing the number of minor alleles. This allows for instance to code dominant or recessive effects.

3.2. Description of the main function

`Spmlficmcm()` is the function used to estimate the regression parameters and the minor allele frequency (β, θ) via generalized semi-parametric maximum likelihood estimation. It uses the following steps:

Step 1: Obtaining initial values for the parameters

The main function calls `Est.Inpar()` to compute initial values for the parameters, which takes the optional argument `typ` to distinguish the data with missing offspring genotype (2) and the complete data (1). `Est.Inpar()` uses logistic regression (`glm()`) to estimate $\widehat{\beta}_0$ (initial value of β) and Equation 7 to estimate $\widehat{\theta}_0$ (initial value of θ). To resolve Equation 4 the `nleqslv()` function from the **nleqslv** package (Hasselman 2015) is called. We choose the Broyden method of global strategies such as line search and trust region. In particular we use the Broyden method because this method often shows superlinear convergence towards a solution (Dennis and Schnabel 1983). The same function computes the initial values of the relevant non-linear system using Equation 5.

Step 2: Construction of the non-linear system

`Spmlficmcm()` calls the function `Nlsysteq()` to build the non-linear system of Equation 4. It must be noted that the number of equations depends on the number of distinct maternal genotypes. The function uses the `nleqslv()` function of the **nleqslv** package and initial values to solve the non-linear system.

Step 3: Determining the log profile likelihood and estimation of parameters

In this last step, `Spmlficmcm()` determines and evaluates the log profile likelihood function of Equation 6 or 11 and its gradient, and computes the Hessian matrix numerically from the gradient. The method described in Equation 8 is then applied to evaluate the parameter estimates.

`Spmlficmcm()` returns an object containing as components the solution of the non-linear system, the matrix of the estimates with their standard errors, the variance-covariance matrix from the sample, the log likelihood function and the value of the log likelihood function assessed at the estimated parameters.

Remark: It is important to verify the following arguments: The data frame `data` must not contain missing values on the covariates such as environmental factors. The maternal and offspring genotypes must be coded as number of minor alleles carried by the individual. The missing offspring genotype data should be coded as `NA`.

4. Illustrations

In this section, we present an illustration of the use of the **SPmlficmcm** package on simulated data. We use the functions include in package **SPmlficmcm** to generate the data according to the model equation formula.

4.1. Simulation

Simulation parameters

We generate the data respecting the constraint between offspring genotype and maternal genotype. Firstly, we generate genotype data (G^M, G^C) for a cohort of $N_{sample} = 20000$ mother-child pairs and consider two continuous covariates $X1$ and $X2$, which are correlated with G^M. Given the values of (G^M, G^C, $X1$, $X2$), we generate a binary disease outcome *outc* from a logistic regression model with the following covariate effects: the log odds of the MAF 0.3, $\beta = c(-0.916,\ 0.857,\ 0.588,\ 0.405,\ -0.693,\ 0.488)$ corresponding respectively to the coefficients of the following terms $X1$, $X2$, gm, $gnch$, $X1 : gnch$, $X2 : gm$ and the intercept -2.23, in a simulation scenario describing the main and interaction effect of (G^M, G^C, $X1$, $X2$). We solve for the value of the intercept parameter so that the resultant phenotype prevalence, $\mathsf{P}(outc = 1)$, is around 6% in the simulation. We then sample $n_1 = 327$ case pairs and $n_0 = 1232$ control pairs from the cohort. We mainly report results when the MAF is 0.3 and both G^M and G^C are coded as the number of minor alleles. Secondly, we create another database from the first, introducing an average 9% of missing offspring genotype data and run the analysis summing over missing offspring genotypes. We repeat this process $B = 500$ times to create 500 complete samples and 500 samples containing the missing offspring genotypes data.

Simulated database

In the following code, we use the function `FtSmlrmCMCM()` to generate a dataset with environment factors that are continuous variables. `M` represents the size of the population, `rho` is the disease prevalence and `N` is a vector containing the number of affected and unaffected subjects in the population. On the last line, `n` mother-child pairs are sampled with the function `SeltcEch()`, where `n`= `n0` + `n1`.

```
R> library("SPmlficmcm")
R> set.seed(13200)
R> M <- 20000
R> fl <- outc ~ X1 + X2 + gm + gnch + X1 : gnch + X2 : gm
R> theta <- 0.3
R> beta <- c(-0.916, 0.857, 0.588, 0.405, -0.693, 0.488)
R> interc <- -2.23
R> vpo <- c(3, 4)
R> vprob <- c(0.35, 0.55)
R> vcorr <- c(2, 1)
R> Dataf <- FtSmlrmCMCM(fl, M, theta, beta, interc, vpo, vprob, vcorr)
R> rho <- table(Dataf["outc"])[2]/20000
R> N <- c(dim(Dataf[Dataf["outc"] == 0, ])[1],
+    dim(Dataf[Dataf["outc"] == 1, ])[1])
R> n0 <- 1232; n1 <- 327
R> DatfE1 <- SeltcEch("outc", n1, n0, "obs", Dataf)
```

Creation of the complete and missing data

The following code creates the sample with complete data (`DatfEmd`) and the full sample including subjects with missing offspring genotype (`DatfEcd`) from the full sample with complete data `DatfE1`.

```
R> DatfE <- DatfE1
R> DatfE[["gnch"]][sample(c(0, 1), dim(DatfE)[1], replace = TRUE,
+    prob = c(0.91, 0.09)) == 1] <- NA
R> DatfEcd <- na.omit(DatfE)
R> DatfEmd <- DatfE
```

Results for the data created

```
R> DatfEcd[26:30, ]

       obs outc X1 X2 gm gnch
18546   29    1  3  2  1    0
195251  30    1  2  1  1    0
9488    31    1  0  1  0    0
5429    32    1  0  0  0    1
5026    33    1  3  2  1    0

R> DatfEmd[26:30, ]

       obs outc X1 X2 gm gnch
16275   26    1  0  0  0   NA
16570   27    1  1  1  0    0
```

```
15788    28     1  1  1  0     0
18546    29     1  3  2  1     0
195251   30     1  2  1  1     0
```

We created 9% missing data with the `sample` function. Each database is a `data.frame` containing 6 columns. Column 1 represents the number of the mother-child pair; the next three columns represent the binary response variable and two continuous environmental variables. The last two variables represent respectively the maternal genotype and the offspring genotype. Both genotypes are coded as number of minor alleles. We have finally two databases, `DatfEmd` that contains the missing data on the offspring genotype and `DatfEcd` that contains only the pairs with complete data. Firstly, we show the use of the two mains functions of the package, thereafter we compare the results of three estimators obtained respectively by logistic regression, the SML approach and the SMLMD approach.

4.2. Estimation of parameters

Estimation of parameters without missing data

On the first line, we give the model equation, then we apply the function `Spmlficmcm()` to estimate the parameters on the sample with complete data.

```
R> fl <- outc ~ X1 + X2 + gm + gnch + X1 : gnch + X2 : gm
R> Rsnm <- Spmlficmcm(fl, N, "gm", "gnch", DatfEcd, 1)
R> round(Rsnm[["Uim"]], digits = 3)
```

The results shown below include first the solution U_{im} of the non-linear system, second the estimates of parameters, third the variance-covariance matrix, and fourth the value of the likelihood function evaluated at the parameter estimates.

```
[1] 0.063 0.063 0.060 0.146 0.147 0.144
```

```
R> round(Rsnm[["MatR"]], digits = 3)
```

```
          Estimate Std.Error
Intercept   -2.229     0.109
X1          -0.741     0.224
X2           0.749     0.163
gm           0.137     0.494
gnch         0.604     0.176
X1:gnch     -0.790     0.117
X2:gm        0.616     0.162
theta       -0.840     0.036
```

```
R> round(Rsnm[["Matv"]], digits = 5)
```

```
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
[1,]  0.01181 -0.00474 -0.00841  0.00102 -0.01204  0.00496  0.00377  0.00083
```

```
[2,] -0.00474  0.05016 -0.01003 -0.09002  0.00645 -0.00321  0.00229 -0.00008
[3,] -0.00841 -0.01003  0.02668  0.00899 -0.00176  0.00231 -0.01145  0.00003
[4,]  0.00102 -0.09002  0.00899  0.24450 -0.00631  0.00098 -0.03220 -0.00053
[5,] -0.01204  0.00645 -0.00176 -0.00631  0.03114 -0.01392  0.00488 -0.00079
[6,]  0.00496 -0.00321  0.00231  0.00099 -0.01392  0.01367 -0.00731  0.00021
[7,]  0.00377  0.00229 -0.01145 -0.03220  0.00488 -0.00731  0.02635 -0.00007
[8,]  0.00083 -0.00008  0.00003 -0.00053 -0.00079  0.00021 -0.00007  0.00130
```

```
R> Rsnm[["Value_loglikh"]]
```

```
[1] -17897.71
```

We also illustrate the use of the function with specification of the disease prevalence p, assuming N is unknown. Coefficient estimates and standard errors changed little. When varying N while keeping the disease prevalence p $= \frac{txpN1}{\mathtt{N}}$ constant, we observed that the coefficient estimates were insensitive to the value of N, but the standard errors of the coefficient estimates decreased slightly with N (data not shown). It is thus preferable to underestimate N and slightly overestimate standard errors than the opposite. Also, setting N too large leads to numerical errors.

```
R> prev <- N[2] / sum(N)
R> Rsnm2 <- Spmlficmcm(fl, gmname = "gm", gcname = "gnch", DatfE = DatfEcd,
+    typ = 1, p = prev)
R> round(Rsnm2$Uim, digits = 3)
```

```
[1] 0.086 0.087 0.082 0.200 0.201 0.197
```

```
R> round(Rsnm2[["MatR"]], digits = 3)
```

```
          Estimate Std.Error
Intercept   -2.229     0.110
X1          -0.741     0.224
X2           0.749     0.163
gm           0.137     0.495
gnch         0.604     0.176
X1:gnch     -0.790     0.117
X2:gm        0.616     0.162
theta       -0.840     0.036
```

```
R> Rsnm2$N
```

```
[1] 13099  1470
```

Estimation of the parameters (with missing data)

We use the function Spmlficmcm() with the option typ equal to 2 to estimate the model parameters on the missing data. The following code gives the solution U_{im} of the non-linear system, the parameter estimates, the variance-covariance matrix, the log-likelihood function and the log-likelihood function evaluated at the parameter estimates.

```
R> Rswm <- Spmlficmcm(fl, N, "gm", "gnch", DatfEmd, 2)
R> round(Rswm[["Uim"]], digits = 3)

[1] 0.063 0.063 0.060 0.146 0.147 0.144

R> round(Rswm[["MatR"]], digits = 3)

          Estimate Std.Error
Intercept   -2.223     0.105
X1          -0.815     0.219
X2           0.758     0.156
gm           0.165     0.482
gnch         0.613     0.176
X1:gnch     -0.797     0.116
X2:gm        0.680     0.154
theta       -0.847     0.035

R> round(Rswm[["Matv"]], digits = 5)

         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
[1,]  0.01092 -0.00427 -0.00743  0.00100 -0.01201  0.00491  0.00306  0.00076
[2,] -0.00427  0.04818 -0.00962 -0.08686  0.00634 -0.00313  0.00232 -0.00007
[3,] -0.00743 -0.00962  0.02444  0.00843 -0.00165  0.00227 -0.01025  0.00004
[4,]  0.00100 -0.08686  0.00843  0.23188 -0.00619  0.00095 -0.02872 -0.00050
[5,] -0.01201  0.00634 -0.00165 -0.00619  0.03098 -0.01372  0.00476 -0.00074
[6,]  0.00491 -0.00313  0.00227  0.00095 -0.01372  0.01349 -0.00726  0.00020
[7,]  0.00306  0.00232 -0.01025 -0.02872  0.00476 -0.00726  0.02357 -0.00008
[8,]  0.00076 -0.00007  0.00004 -0.00050 -0.00074  0.00020 -0.00008  0.00122

R> Rswm[["Value_loglikh"]]

[1] -19027.6
```

Here again disease prevalence can be specified when `N` is unknown, with little change to the coefficient estimates and standard errors.

```
R> Rswm2 <- Spmlficmcm(fl, gmname = "gm", gcname = "gnch", DatfE = DatfEmd,
+    typ = 2, p = prev)
R> round(Rswm2$Uim, digits = 3)

[1] 0.086 0.087 0.082 0.200 0.201 0.197

R> round(Rswm2[["MatR"]], digits = 3)
```

Model		Standard		SML		SMLMD	
Terms	True	Est.	Std.error	Est.	Std.error	Est.	Std.error
`Intercept`	−2.230	−1.367	0.126	−2.229	0.109	−2.223	0.105
`X1`	−0.916	−0.769	0.226	−0.741	0.224	−0.815	0.219
`X2`	0.857	0.740	0.164	0.749	0.163	0.758	0.156
`gm`	0.588	0.171	0.501	0.137	0.494	0.165	0.482
`gc`	0.405	0.565	0.183	0.604	0.176	0.613	0.176
`X1:gc`	−0.693	−0.788	0.121	−0.790	0.117	−0.797	0.116
`X2:gm`	0.488	0.640	0.167	0.616	0.162	0.680	0.154
`theta`	−0.847	–	–	−0.840	0.036	−0.847	0.035

Table 1: Comparison of results of three methods on the simulated data. Standard: refers to logistic regression. SML: semi-parametric maximum log-likelihood method (no missing data). SMLMD: semi-parametric maximum log-likelihood method (missing data).

```
          Estimate Std.Error
Intercept   -2.223     0.106
X1          -0.815     0.220
X2           0.758     0.156
gm           0.165     0.482
gnch         0.613     0.176
X1:gnch     -0.797     0.116
X2:gm        0.680     0.154
theta       -0.847     0.035
```

```
R> Rswm2$N
```

```
[1] 13099  1470
```

Using the same data, we applied logistic regression with the same equation, we computed the standard errors of the estimates and we compared them to the other two estimates. The results are reported in Table 1. When we compare the SML approach with the standard approach (logistic regression) on one simulated sample, we notice an amelioration i.e., a reduction of variance and the estimates are nearer to the true values. These results corroborate the results of Chen *et al.* (2012). Indeed Chen *et al.* (2012) showed that using the SML approach when the assumption is satisfied reduces the variance on average by 30%. When using the SMLMD approach, we again observed a reduction of the standard errors of the estimates. To confirm this error reduction and validate the properties of the semi-parametric maximum likelihood estimator, we assessed the following quantities: the bias, the mean square error (MSE), the empirical variance ($\mathbf{V}_{emp}$), the mean estimated variance and the confidence interval coverage for the two approaches (SML and SMLMD) on $B = 500$ replicates of the generated data. We removed 20 replicates where we observed negative variances and/or coefficient estimates beyond 3 median absolute deviations (scaled to estimate the standard deviation) from the median for both methods. The results are reported in Table 2. Tables 3 and 4 show the empirical covariances and the mean estimated covariances.

The MSE and empirical variance of the coefficient estimates for the mother genotype *gm* and the covariates $X1$ and $X2$ are slightly reduced when making use of the entire dataset, as

Model	SML					SMLMD				
Terms	Bias	MSE	$\mathbf{V}_{emp}(\widehat{\eta})$	$\overline{\mathbf{V}(\widehat{\eta})}$	$\mathsf{P}(\widehat{\eta} \in CI95)$	Bias	MSE	$\mathbf{V}_{emp}(\widehat{\eta})$	$\overline{\mathbf{V}(\widehat{\eta})}$	$\mathsf{P}(\widehat{\eta} \in CI95)$
`Intercept`	−0.007	0.013	0.013	0.011	0.923	−0.007	0.013	0.013	0.011	0.925
`X1`	0.008	0.063	0.063	0.053	0.938	0.009	0.060	0.061	0.049	0.925
`X2`	−0.015	0.034	0.034	0.027	0.925	−0.016	0.032	0.032	0.025	0.908
`gm`	0.014	0.269	0.269	0.256	0.938	0.010	0.261	0.262	0.232	0.946
`gnch`	0.018	0.039	0.039	0.032	0.935	0.017	0.039	0.038	0.032	0.933
`X1:gnch`	−0.019	0.017	0.017	0.014	0.933	−0.018	0.017	0.016	0.013	0.931
`X2:gm`	0.000	0.036	0.036	0.027	0.917	0.002	0.033	0.033	0.025	0.925
`theta`	0.000	0.001	0.001	0.001	0.938	0.000	0.001	0.001	0.001	0.927

Table 2: Comparison of results of two methods on the simulated data ($B = 480$ replicates where estimation succeeded). MSE: mean square error, $\mathbf{V}$: estimated variance, $\mathbf{V}_{emp}$: empirical variance, $\mathsf{P}(\widehat{\eta} \in CI95)$: 95% confidence interval coverage, SML: semi-parametric maximum likelihood method (no missing data), SMLMD: semi-parametric maximum likelihood method (missing data).

$$\mathsf{COV}(\widehat{\eta},\widehat{\eta}) = \begin{pmatrix} & -0.005 & -0.008 & 0.001 & -0.011 & 0.005 & 0.004 & 0.001 \\ -0.003 & & -0.010 & -0.097 & 0.006 & -0.003 & 0.002 & 0.000 \\ -0.010 & -0.016 & & 0.009 & -0.002 & 0.003 & -0.012 & 0.000 \\ -0.002 & -0.105 & 0.017 & & -0.006 & 0.002 & -0.032 & -0.001 \\ -0.014 & 0.003 & 0.000 & -0.004 & & -0.014 & 0.005 & -0.001 \\ 0.006 & -0.001 & 0.002 & -0.001 & -0.018 & & -0.007 & 0.000 \\ 0.004 & -0.001 & -0.014 & -0.035 & 0.008 & -0.010 & & 0.000 \\ 0.001 & 0.000 & 0.000 & -0.001 & -0.001 & 0.000 & 0.000 & \end{pmatrix}$$

Table 3: Comparison of the empirical covariance vs. mean covariance for the SML method computed on $B = 480$ samples where estimation succeeded. The empirical covariances are under the diagonal and mean estimated covariances are above.

$$\mathsf{COV}(\widehat{\eta},\widehat{\eta}) = \begin{pmatrix} & -0.004 & -0.008 & 0.001 & -0.011 & 0.005 & 0.003 & 0.001 \\ -0.002 & & -0.009 & -0.088 & 0.006 & -0.003 & 0.002 & 0.000 \\ -0.010 & -0.016 & & 0.008 & -0.002 & 0.003 & -0.011 & 0.000 \\ -0.002 & -0.102 & 0.017 & & -0.006 & 0.002 & -0.029 & -0.001 \\ -0.014 & 0.004 & 0.000 & -0.005 & & -0.014 & 0.005 & -0.001 \\ 0.006 & -0.002 & 0.002 & 0.000 & -0.017 & & -0.007 & 0.000 \\ 0.003 & 0.000 & -0.013 & -0.034 & 0.008 & -0.010 & & 0.000 \\ 0.001 & 0.000 & 0.000 & -0.001 & 0.000 & 0.000 & 0.001 & \end{pmatrix}$$

Table 4: Comparison of the empirical covariance vs. mean covariance for the SMLMD method computed on $B = 480$ samples where estimation succeeded. The empirical covariances are under the diagonal and mean estimated covariances are above.

Number	SML	SMLMD	SML	SMLMD
of terms	$n = 1441$	$n = 1559$	$n = 342$	$n = 393$
3	3.3	3.9	3.1	3.3
4	4.2	4.6	3.8	4.1
5	4.6	5.5	3.2	3.5
6	5.6	5.6	4.1	4.6

Table 5: Computing time expressed in seconds. n: sample size.

expected. The estimates were unbiased and the mean estimated variance of the coefficient estimates was slightly below the empirical variance, and confidence interval coverage was close to or slightly below the nominal 95% level. Some covariances are also estimated closer to 0 than the empirical value.

Computing times are reported in Table 5. In summary the computing time depends on the number of terms in the formula and the size of the sample.

5. Discussion

Package **SPmlficmcm** implements the semi-parametric maximum likelihood estimation for case-mother control-mother designs, allowing for missing offspring genotype. This method is important in studies where we want to determine the role of a polymorphism in interaction with the mother exposure to an environmental factor on obstetric and early-life outcome risk. Indeed these models permit to take into account the correlation between the maternal genotype and offspring genotype under the assumptions of Hardy-Weinberg equilibrium and Mendelian inheritance. The statistical properties of the estimates were satisfactory, although a slight underestimation of the empirical variance by the variance estimate led to a slight undercoverage of the confidence intervals. **SPmlficmcm** has been made available on CRAN. This package executes the computation relatively quickly. We hope that this package will encourage applied researchers to use this type of modeling, as it provides a relevant way to study gene-environment interactions in case-mother and control-mother designs.

Acknowledgments

This research was funded by the Canadian Institutes of health research (MOP-102708, PI Patrick Levallois, Université Laval). A. Bureau was supported by a research fellowship from the Fonds de recherche du Québec – Santé.

References

Allison P (2001). *Missing Data*, volume 136 of *Quantitative Applications in the Social Sciences*. Sage.

Chen J, Lin D, Hochner H (2012). "Semiparametric Maximum Likelihood Methods for Analyzing Genetic and Environmental Effects with Case-Control Mother-Case Child Pair Data." *Biometrics*, **68**(3), 869–877. doi:10.1111/j.1541-0420.2011.01728.x.

Dennis J, Schnabel R (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611971200.

Hasselman B (2015). ***nleqslv**: Solve Systems of Nonlinear Equations*. R package version 2.9, URL http://CRAN.R-project.org/package=nleqslv.

Infante-Rivard C (2004). "Drinking Water Contaminants, Gene Polymorphisms, and Fetal Growth." *Environmental Health Perspectives*, **112**(11), 1213. doi:10.1289/ehp.7003.

Infante-Rivard C (2007). "Studying Genetic Predisposition Among Small-for-Gestational-Age Newborns." In *Seminars in Perinatology*, volume 31, pp. 213–218.

Levallois P, Gingras S, Marcoux S, Legay C, Catto C, Rodriguez M, Tardif R (2012). "Maternal Exposure to Drinking-Water Chlorination By-Products and Small-for-Gestational-Age Neonates." *Epidemiology*, **23**(2), 267–276. doi:10.1097/ede.0b013e3182468569.

Little RJA, Rubin DB (2002). *Statistical Analysis with Missing Data.* 2nd edition. John Wiley & Sons, Hoboken. doi:10.1002/9781119013563.scard.

Nguile-Makao M, Bureau A (2015). ***SPmlficmcm**: Semiparametric Maximum Likelihood Method for Interactions Gene-Environment in Case-Mother Control-Mother Designs.* R package version 1.4, URL http://CRAN.R-project.org/package=SPmlficmcm.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Shi M, Umbach D, Vermeulen S, Weinberg C (2008). "Making the Most of Case-Mother/Control-Mother Studies." *American Journal of Epidemiology*, **168**(5), 541–547. doi:10.1093/aje/kwn149.

Affiliation:

Alexandre Bureau
Centre de Recherche de l'Institut Universitaire en Santé Mentale de Québec
and Département de Médecine Sociale et Préventive
Université Laval
1050 de la Médecine, room 2457
Québec, QC, G1V 0A6, Canada
Telephone: +1-418-656-2131, Ext: 3342
E-mail: alexandre.bureau@msp.ulaval.ca
URL: http://www.crulrg.ulaval.ca/pages_perso_chercheurs/bureau_a/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/

December 2015, Volume 68, Issue 10 *Submitted:* 2013-10-23
doi:10.18637/jss.v068.i10 *Accepted:* 2015-01-15