Methods and software to analyze gene-environment interactions under a case-mother - control-mother design with partially missing child genotype

Online supporting information

Alexandre Bureau, Yuang Tian, Patrick Levallois, Yves Giguère, Jinbo Chen, Hong Zhang

September 27, 2022

1 Review of software

This section reviews R software packages and functions implementing analysis approaches for case-mother - control-mother data.

1.1 R Package SPmlficmcm

Makao and Bureau [3] described the R package SPmlficmcm implementing semi-parametric maximum likelihood estimation for the prospective likelihood when the data are complete and when offspring genotype is missing for a subset of subjects. The model fitting function Spmlficmcm has arguments similar to other such functions in R, with a formula object to specify the model and a data argument to specify the data frame containing the variables in the model. Users have the choice to provide either the cases and non-cases population totals (N_0, N_1) in the vector N or the disease prevalence f in the argument p if N_0 and N_1 are unavailable. The program then assumes

large values of N_0 and N_1 satisfying the prevalence constraint $\frac{N_1}{N_0+N_1} = f$. The matrix "MatR" of the estimates and their standard errors with model terms as row names is part of the object returned by Spmlficmcm. Makao and Bureau [3] provide additional details on Spmlficmcm arguments and the object it returns.

1.2 R Package Haplin

Gjerdevik et al. [2] described the R package Haplin which includes the haplinStrat function to fit log-linear models to case-triad data with stratification on an environmental covariate. Case-mother - control-mother designs are handled by treating the father genotype as missing data, leading to the model with the constraints described by Shi et al. [4] in each stratum of a discrete covariate X when maternal and child genotypes are observed. Likelihood maximization with an expectation-maximization algorithm deals with missing paternal and partially missing child genotypes. Since usage of the Haplin package is explained elsewhere [2], we focus on the options of the haplinStrat function required for analyzing case-mother - control-mother data:

design Must be set to "cc.triad" to reflect the presence of cases and controls with parents (the mother only in this case).

use.missing Must be set to TRUE to handle missing paternal and partially missing child genotypes.

maternal Must be set to TRUE if one wants to estimate effects of g^M.

ccvar Column number of the case-control status Y

strata Column number of the environmental covariate X on which to stratify.

response Setting to mult will estimate multiplicative (log-additive) effects which is the usual behavior of R regression functions used in all analyses in this article. The default free will instead estimate distinct effects for heterozygotes and homozygote rare genotypes.

The data frame read by the preprocessing genDataRead function must include variables for the paternal genotypes, which are set entirely to NA under the case-mother - control-mother design.

The gxe function applied to the object returned by haplinStrat will test the null hypothesis of homogeneity of the g^M and g^C effects across strata of X. When X is ordinal or quantitative, the 1 degree-of-freedom trend test is similar to testing $\beta_{p+3} = 0$ or $\beta_{p+4} = 0$ in model (1).

In addition to model fitting functions, Gjerdevik et al. [1] recently described functions for power and sample size calculations in Haplin.

1.3 R functions CCMO.na and CCMO.dep

Estimation by maximization of the modified retrospective likelihood is implemented in the R functions CCMO.na for the unspecified distribution of G^M, X, and CCMO.dep for the double-additive logistic model of the maternal genotype (4). The two functions share the following required argument

Y Disease outcome variable (1 = affected, 0=unaffected)

gm Variable containing the number of minor alleles of the mother

gc Variable containing the number of minor alleles of the child

Xm Matrix of covariates involved in product terms with the genotype of the mother (can be a single variable in a vector)

Xc Matrix of covariates involved in product terms with the genotype of the child (can be a single variable in a vector)

Xo Matrix of all non-genetic covariates in the model (including those involved in product terms with the genotype variables).

f Disease prevalence.

The function CCMO.dep has the additional Xgm argument: a matrix of covariates involved in the double-additive logistic model of the maternal genotype (4). The use of this model enables the handling of missing maternal genotype in gm through the likelihood by CCMO.dep. Both

CCMO.na and CCMO.dep allow missing values in the child genotype gc, but not in covariates.

The object returned by the functions CCMO.na and CCMO.dep contains the log-likelihood logL, the vector of parameter estimates est and their estimated standard errors sd, as well as the matrix of estimated variance-covariance of the parameter estimates Matv. Regression coefficients are in the following order: intercept, main effect of mother genotype, main effect of child genotype, main effect of variables interacting with genotypes, main effect of variables NOT interacting with genotypes, interaction terms involving the mother genotype, interaction terms involving the child genotype. The corresponding coefficient estimates est.log and standard errors sd.log from standard logistic regression are included for comparison.

2 Illustration of software implementations on simulated data

We now illustrate the application of the software on a simulated data set similar to the one used in the case study. First we load the empirical distribution in the control sample of a subset of variables from the case study data: rs3813867 C allele count in the child and in the mother, TTHM exposure in the fourth quartile (1=yes, 0=no), prepregnancy mother's BMI category ($<$ 19.8,19.8-25.9,26-29.9 and $>$29.9) and multi-parity (1=yes, 0=no).

```
> download.file("https://github.com/abureau/CMCMsim/raw/main/data/ctrl.RData","ctrl.RData")
> load("ctrl.RData")
```

As an illustration of the data, we assess the association between maternal rs3813867 C allele and TTHM exposure in the fourth quartile using all controls:

```
> library(epitools)
> tab = apply(ctrl.dat,c(3,1),sum)
> epitab(tab)

$tab

     0        p0  1   p1 oddsratio    lower    upper    p.value
```

```
  0 852 0.7559894 52 0.65  1.000000       NA       NA         NA
  1 275 0.2440106 28 0.35  1.668252 1.033205 2.693622 0.04448547

$measure
[1] "wald"

$conf.level
[1] 0.95

$pvalue
[1] "fisher.exact"
```

Next, we simulate a population with the same number of births as in the Quebec City area during the study period (14,630), assuming the control sample without missing child genotype is representative of that population. We use a matrix with all possible covariate levels to recreate a dataset with individual variable values.

```
> set.seed(100)
> n_ctrl_noNA = sum(ctrl_noNA.dat)
> ctrl_prop_noNA = ctrl_noNA.dat/n_ctrl_noNA
> pop.obs = rmultinom(1,14630,as.vector(ctrl_prop_noNA))
> lev_noNA.dat = cbind((0:63)%%2,(0:63)%/%2%%2,(0:63)%/%4%%2,(0:63)%/%8%%4,(0:63)%/%32)
> pop.dat = data.frame(lev_noNA.dat[rep(1:nrow(lev_noNA.dat),pop.obs),])
> names(pop.dat) = c("gm","gc","TTHM","BMI","multipar")
```

We manually recode BMI using three indicator variables, as not all analysis functions handle R factors.

```
> pop.dat$BMI0 = ifelse(pop.dat$BMI==0,1,0)
> pop.dat$BMI2 = ifelse(pop.dat$BMI==2,1,0)
> pop.dat$BMI3 = ifelse(pop.dat$BMI==3,1,0)
> pop.dat$BMI = NULL
> names(pop.dat)
```

```
[1] "gm"       "gc"       "TTHM"     "multipar" "BMI0"     "BMI2"     "BMI3"
```

It is now time to simulate the outcome in the entire population. We use the estimates from the modified retrospective likelihood with unspecified distribution of G^M, X on the actual data as coefficents to simulate the outcome.

```
> download.file("https://github.com/abureau/CMCMsim/raw/main/data/CCMOcov.RS3813867complet.RData",
+               "CCMOcov.RS3813867complet.RData")
> load("CCMOcov.RS3813867complet.RData")
> # Constructing the design matrix
> Xmat = as.matrix(cbind(1,pop.dat[,c(1:3,7:4)],pop.dat$gm*pop.dat$TTHM,pop.dat$gc*pop.dat$TTHM))
> # Linear predictor
> tmp <- Xmat%*%CCMOcov.RS3813867complet$est[1:10]
> # Case probability
> p <- 1/(1+exp(-tmp))
> r <- runif(14630)
> pop.dat$outc <- ifelse(r<p,1,0)
> N=table(pop.dat$outc)
> N

    0     1
13878   752
```

Then, we load the software code.

```
> library("SPmlficmcm")
> source("https://github.com/yatian20/CCMO.na/blob/main/R/CCMO.na.R?raw=true")
> source("https://github.com/yatian20/CCMO.na/blob/main/R/CCMO.dep.R?raw=true")
```

We are now ready to draw a case-control sample from the simulated population.

```
> pop.dat$obs=1:nrow(pop.dat)
> n0 <- 1207; n1 <- 321
> study.dat <- SeltcEch("outc", n1, n0, "obs", pop.dat)
```

Next, we delete 10% of case genotypes and 6% of control genotypes, the proportion observed in the case study.

```
> ## Creation of missing data on the offspring genotype
> r = runif(n1)
> study.dat$gc[study.dat$outc==1][r<0.1] = NA
> r = runif(n0)
> study.dat$gc[study.dat$outc==0][r<0.06] = NA
```

Finally, we estimate the correctly-specified model using the Spmlficmcm, CCMO.na and CCMO.dep functions. Haplin was not applied to this simulated dataset and the case study dataset since it does not allow to adjust for covariates such as X2. Notice that for CCMO.na and CCMO.dep, variable names must be specified with the dataframe where they are found. One way to avoid this would be to attach the dataframe study.dat to the R search path. In the code below, estimates of the model coefficients and their estimated standard errors are gathered in a data frame. Columns with headers `"Estimate"` and `"Std.Error"` contain results from the prospective likelihood, columns with the `"RLNP"` prefix contain results from the retrospective likelihood with the unspecified distribution of G^M, X columns with the `"RLDA"` prefix contain results from the retrospective likelihood with the double additive model and columns with the `"LR"` prefix contain results from logistic regression. As can be seen from the display of this data frame below, estimates from the two estimation methods exploiting the constraints of the genetic and environmental variables relationships are close to the true values. Both methods also produced similar estimated standard errors, generally smaller than those from logistic regression, which discards entirely the observations with missing child genotype. The good behaviour of the prospective likelihood contrasts with the high proportion of failures in the simulation study reported in Table 2This is due to the different distribution of the covariates. With the discrete covariates simulated here, maximization failures or outlying estimates were rare (4%) and estimates had satisfactory statistical properties (see Table 2 in [3]).

```
> fl <- outc ~ gm + gc + TTHM + BMI3 + BMI2 + BMI0 + multipar + TTHM:gm + TTHM:gc
> ## Estimation of the parameters (with missing data)
> ## Prospective likelihood
> Rswm <- Spmlficmcm(fl, N, "gm", "gc", study.dat, 2)
> ## Retrospective likelihood, unspecified distribution of  G^M,X (function CCMO.na)
```

```
> prev  <-  N[2] / sum(N)
> Rpwm <- CCMO.na(Y=study.dat$outc,gm=study.dat$gm,gc=study.dat$gc,Xm=study.dat$TTHM,
+                 Xc=study.dat$TTHM,Xo=as.matrix(study.dat[,c("TTHM","BMI3", "BMI2", "BMI0", "multipar")]),
+                 f=prev,ind=FALSE)
> ## Retrospective likelihood, double-additive model for Pr(G^M | X) (function CCMO.dep)
> Rpda <- CCMO.dep(Y=study.dat$outc,gm=study.dat$gm,gc=study.dat$gc,Xm=study.dat$TTHM,
+                 Xc=study.dat$TTHM,Xgm=study.dat$TTHM,
+                 Xo=as.matrix(study.dat[,c("TTHM","BMI3", "BMI2", "BMI0", "multipar")]),
+                 f=prev,HWE=TRUE)
> ## Table of estimates and standard errors
> round(data.frame(Beta=CCMOcov.RS3813867complet$est[1:10],Rswm[["MatR"]][1:10,],
+                  RLNPest=Rpwm$est[1:10],RLNPsd=Rpwm$sd[1:10],RLDAest=Rpda$est[1:10],RLDAsd=Rpda$sd[1:10],
+                  LRest=Rpwm$est.log,LRsd=Rpwm$sd.log), digits = 3)
```

	Beta	Estimate	Std.Error	RLNPest	RLNPsd	RLDAest	RLDAsd	LRest	LRsd
Intercept	-2.576	-2.454	0.198	-2.434	0.206	-2.450	0.206	-0.876	0.214
gm	0.281	0.404	0.332	0.228	0.355	0.357	0.329	0.323	0.380
gc	-0.180	-0.120	0.367	-0.069	0.378	-0.162	0.366	-0.066	0.430
TTHM	0.236	0.097	0.154	0.099	0.154	0.103	0.154	0.070	0.161
BMI3	-0.158	-0.242	0.258	-0.248	0.259	-0.233	0.259	-0.238	0.269
BMI2	-0.255	-0.550	0.210	-0.560	0.212	-0.549	0.211	-0.544	0.220
BMI0	0.669	0.489	0.232	0.500	0.233	0.516	0.233	0.530	0.242
multipar	-0.792	-0.557	0.131	-0.567	0.132	-0.562	0.131	-0.582	0.137
gm:TTHM	-1.282	-1.228	0.775	-1.402	0.813	-1.448	0.791	-1.218	0.826
gc:TTHM	0.482	0.674	0.674	0.701	0.681	0.668	0.670	0.474	0.735

References

[1] Miriam Gjerdevik, Håkon K Gjessing, Julia Romanowska, Øystein A Haaland, Astanand Jugessur, Nikolai O Czajkowski, and Rolv T Lie. Design efficiency in genetic association studies. Statistics in Medicine, 39(9):1292–1310, 2020.

[2] Miriam Gjerdevik, Øystein A Haaland, Julia Romanowska, Rolv T Lie, Astanand Jugessur, and Håkon K Gjessing. Parent-of-origin-environment interactions in case-parent triads with or without independent controls. Annals Human Genetics, 82(2):60–73, 2018.

[3] Moliere Nguile-Makao and Alexandre Bureau. Semi-Parametric Maximum Likelihood Method for Interaction in Case-Mother Control-Mother Designs: Package SPmlficmcm. Journal of Statistical Software, 68(10):1–17, 2015.

[4] M. Shi, D. M. Umbach, S. H. Vermeulen, and C. R. Weinberg. Making the most of case-mother/control-mother studies. American Journal of Epidemiology, 168(5):541–547, 2008.