The TwoPhaseInd package: Estimation of gene-treatment interactions in randomized clinical trials exploiting gene-treatment independence

Xiaoyu Wang, James Y. Dai

December 16, 2015

1 Introduction

In randomized clinical trials, there are often ancillary studies that uses outcome-dependent sampling to identify baseline genetic markers that modify treatment effect. The TwoPhaseInd package assembles several functions to estimate gene-treatment interactions in randomized clinical trials exploiting gene-treatment independence in case-control sampling and case-cohort sampling. For case-control sampling, it computes two estimators- semiparametric maximum likelihood estimator exploiting (SPMLE) and maximum estimated likelihood estimator (MELE), exploiting the treatment-covariate independence resulted from randomization in two-phase randomized trials [1]. For case-cohort sampling, it has a function (acoarm) to estimate parameters in a cox regression model by a two-stage estimation procedure developed for augmented case-only designs [2].

2 SPMLE

We took a WHI biomarker study to illustrate our methods. The aforementioned 29 biomarkers were picked by WHI investigators as markers that are possibly associated with either stroke, venous thrombotic disease, or myocardial infarction. A comprehensive analysis of these samples was published by [3]. The results of this particular biomarker example were shown in [1]. The methodologies for estimating SPMLE and EMLE can be found in [1].

```
> data(whiBioMarker)
> dim(whiBioMarker)

[1] 16608    10

> str(whiBioMarker)
```

```
'data.frame':        16608 obs. of  10 variables:
 $ stroke : num  0 0 0 0 0 0 1 0 0 1 ...
 $ hrtdisp: num  1 1 0 1 1 1 1 1 0 1 ...
 $ papbl  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ age    : num  64 62 62 60 54 57 77 68 73 64 ...
 $ dias   : num  74 70 70 79 70 88 62 60 60 67 ...
 $ hyp    : Factor w/ 3 levels "Missing","No",..: 2 2 2 2 3 3 2 2 2 2 ...
 $ syst   : num  116 135 133 133 119 ...
 $ diabtrt: Factor w/ 3 levels "Missing","No",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ lmsepi : Factor w/ 5 levels "2 - <4 episodes per week",..: 5 4 1 4 1 5 5 4 2 2 ...
 $ phase  : num  1 1 1 1 1 1 1 1 1 1 ...

>
```

Here is an example of estimating SPMLE without exploiting independent and with no confounding factors:

```
> spmleNonIndNoExtra <- spmle(data=whiBioMarker,  ## dataset
+                  response="stroke",        ## response variable
+                  treatment="hrtdisp", ## treatment variable
+                  BaselineMarker="papbl",        ## environment variable
+                  extra=NULL,
+                  phase="phase",        ## phase indicator (1 and 2)
+                  ind=FALSE        ## independent or non-indepentent
+ )
> spmleNonIndNoExtra

                         beta  stder         pVal
(Intercept)           -4.4266 0.1147 0.0000000000
hrtdisp (Treatment)    0.3185 0.1473 0.0306369775
papbl (BaselineMarker) 2.2603 1.0152 0.0259794031
hrtdisp:papbl         -4.2536 1.2880 0.0009584502
```

Here is an example of SPMLE with exploiting independent and with no confounding factors:

```
> spmleIndNoExtra <- spmle(data=whiBioMarker,        ## dataset
+                response="stroke", ## response variable
+                treatment="hrtdisp",        ## treatment variable
+                BaselineMarker="papbl",                ## environment variable
+                extra=NULL,
+                phase="phase",        ## phase indicator
+                ind=TRUE        ## independent or non-indepentent
+ )
> spmleIndNoExtra
```

```
                          beta  stder         pVal
(Intercept)            -4.4198 0.1131 0.0000000000
hrtdisp (Treatment)     0.3077 0.1463 0.0354228673
papbl (BaselineMarker)  1.9063 0.9097 0.0361172352
hrtdisp:papbl          -3.9327 1.1533 0.0006499955
```

Here is an example of estimating SPMLE without exploiting independent and with confounding factors:

```
> spmleNonIndExtra <- spmle(data=whiBioMarker,        ## dataset
+                 response="stroke",        ## response variable
+                 treatment="hrtdisp",        ## treatment variable
+                 BaselineMarker="papbl",        ## environment variable
+                 extra=c(
+                         "age"                 ## age
+                         ## physical activity levels
+                          , "dias"         ## diabetes
+                          , "hyp"         ## hypertension
+                          , "syst"         ## systolic
+                          , "diabtrt"        ## diastolic BP
+                          , "lmsepi" ## waist:hip ratio
+                              ),        ## extra variable(s)
+                 phase="phase",        ## phase indicator
+                 ind=FALSE        ## independent or non-indepentent
+ )
> spmleNonIndExtra

                             beta  stder         pVal
(Intercept)               -3.9599 0.6756 4.602982e-09
hrtdisp (Treatment)        0.3698 0.1599 2.071078e-02
papbl (BaselineMarker)     2.3487 1.0565 2.620678e-02
hrtdisp:papbl             -4.1924 1.3313 1.637308e-03
age                        1.3736 1.1935 2.497868e-01
dias                      -0.8499 0.9990 3.949167e-01
hypNo                     -0.7751 0.6229 2.133320e-01
hypYes                    -0.7607 0.6288 2.263832e-01
syst                       3.3370 1.2286 6.603730e-03
diabtrtYes                 0.8811 0.3707 1.746453e-02
lmsepi4+ episodes per week 0.0022 0.3927 9.954563e-01
lmsepiMissing             -0.1904 0.6121 7.557121e-01
lmsepiNo activity          0.3231 0.4145 4.356103e-01
lmsepiSome activity        0.0659 0.3522 8.516191e-01
```

Here is an example of estimating SPMLE with exploiting independent and with confounding factors:

```
> spmleIndExtra <- spmle(data=whiBioMarker,        ## dataset
+             response="stroke",        ## response variable
+             treatment="hrtdisp",        ## treatment variable
+             BaselineMarker="papbl",        ## environment variable
+             extra=c(
+                "age"         ## age
+                              ## physical activity levels
+               , "dias"        ## diabetes
+               , "hyp" ## hypertension
+               , "syst" ## systolic
+               , "diabtrt"        ## diastolic BP
+               , "lmsepi" ## waist:hip ratio
+                 ),        ## extra variable(s)
+             phase="phase", ## phase indicator
+             ind=TRUE ## independent or non-indepentent
+ )
> spmleIndExtra
```

```
                              beta  stder         pVal
(Intercept)                -3.9647 0.6734 3.923841e-09
hrtdisp (Treatment)         0.3102 0.1467 3.440407e-02
papbl (BaselineMarker)      1.9058 0.9375 4.206696e-02
hrtdisp:papbl              -3.8688 1.1590 8.435226e-04
age                         1.7675 1.2051 1.424798e-01
dias                       -0.6402 0.9864 5.163627e-01
hypNo                      -0.8253 0.6189 1.823384e-01
hypYes                     -0.8161 0.6244 1.911675e-01
syst                        3.0481 1.2110 1.183349e-02
diabtrtYes                  0.9493 0.3715 1.060836e-02
lmsepi4+ episodes per week  0.1714 0.3879 6.586897e-01
lmsepiMissing              -0.1447 0.6089 8.121264e-01
lmsepiNo activity           0.3950 0.4085 3.336301e-01
lmsepiSome activity         0.1540 0.3488 6.588982e-01
```

3 MELE

Here is an example of MELE with exploiting independent and with no confounding factors:

```
> melIndNoExtra <- mele(data=whiBioMarker,        ## dataset
+             response="stroke", ## response variable
+             treatment="hrtdisp",        ## treatment variable
+             BaselineMarker="papbl",        ## environment variable
+             extra=NULL,
+             phase="phase",        ## variable for phase indicator
+             ind=TRUE ## independent or non-indepentent
+ )
> melIndNoExtra

                          beta  stder         pVal
(Intercept)            -4.4183 0.1128 0.0000000000
hrtdisp (Treatment)     0.3065 0.1460 0.0357817680
papbl (BaselineMarker)  1.8586 0.8981 0.0385007661
hrtdisp:papbl          -3.8660 1.1464 0.0007450917
```

Here is an example of MELE without exploiting independent with confounding factors:

```
> melNoIndNoExtra <- mele(data=whiBioMarker,        ## dataset
+               response="stroke",        ## response variable
+               treatment="hrtdisp",        ## treatment variable
+               BaselineMarker="papbl",        ## environment variable
+               extra=NULL,
+               phase="phase", ## phase indicator
+               ind=FALSE        ## independent or non-indepentent
+ )
> melNoIndNoExtra

                          beta  stder        pVal
(Intercept)            -4.4269 0.1148 0.000000000
hrtdisp (Treatment)     0.3202 0.1472 0.029616750
papbl (BaselineMarker)  2.2724 1.0199 0.025875181
hrtdisp:papbl          -4.2016 1.2855 0.001081392
```

Here is an example of MELE with exploiting independent and with confounding factors:

```
> melIndExtra <- mele(data=whiBioMarker,        ## dataset
+           response="stroke",        ## response variable
+           treatment="hrtdisp",                ## treatment variable
+           BaselineMarker="papbl",                ## environment variable
+           extra=c(
```

```
+               "age"         ## age
+                                  ## physical activity levels
+                , "dias"         ## diabetes
+                , "hyp" ## hypertension
+                , "syst"         ## systolic
+                , "diabtrt"        ## diastolic BP
+                , "lmsepi" ## waist:hip ratio
+                ),        ## extra variable(s)
+           phase="phase",        ## phase indicator
+           ind=TRUE        ## independent or non-indepentent
+ )
> melIndExtra

                              beta  stder         pVal
(Intercept)                -3.8846 0.7172 6.089906e-08
hrtdisp (Treatment)         0.3083 0.1463 3.511160e-02
papbl (BaselineMarker)      1.8662 0.9282 4.436775e-02
hrtdisp:papbl              -3.7931 1.1548 1.021672e-03
age                         1.7872 1.2034 1.375141e-01
dias                       -0.8270 1.0211 4.180127e-01
hypNo                      -0.8560 0.6636 1.971193e-01
hypYes                     -0.9329 0.6739 1.662278e-01
syst                        3.3869 1.2285 5.834062e-03
diabtrtYes                  0.9363 0.3711 1.164302e-02
lmsepi4+ episodes per week  0.1278 0.3903 7.434100e-01
lmsepiMissing              -0.2114 0.6500 7.450406e-01
lmsepiNo activity           0.4480 0.4086 2.729547e-01
lmsepiSome activity         0.1385 0.3515 6.935112e-01
```

Here is an example of MELE without exploiting independent and with confounding factors:

```
> melNoIndExtra <- mele(data=whiBioMarker,        ## dataset
+             response="stroke",        ## response variable
+             treatment="hrtdisp",        ## treatment variable
+             BaselineMarker="papbl",        ## environment variable
+             extra=c(
+                 "age"        ## age
+                                 ## physical activity levels
+                 , "dias"         ## diabetes
+                 , "hyp"         ## hypertension
+                 , "syst"         ## systolic
+                 , "diabtrt"        ## diastolic BP
```

```
+                  , "lmsepi" ## waist:hip ratio
+                  ),        ## extra variable(s)
+              phase="phase",        ## phase indicator
+              ind=FALSE        ## independent or non-indepentent
+ )
> melNoIndExtra
```

```
                              beta  stder         pVal
(Intercept)                -3.9227 0.7239 5.999024e-08
hrtdisp (Treatment)         0.3190 0.1587 4.441772e-02
papbl (BaselineMarker)      2.0377 1.0557 5.358469e-02
hrtdisp:papbl              -3.7559 1.3308 4.767720e-03
age                         1.8170 1.2290 1.392979e-01
dias                       -1.0119 1.0309 3.263064e-01
hypNo                      -0.7987 0.6694 2.328199e-01
hypYes                     -0.9390 0.6790 1.666968e-01
syst                        3.5970 1.2565 4.199987e-03
diabtrtYes                  0.7687 0.3844 4.551940e-02
lmsepi4+ episodes per week  0.1654 0.3953 6.756095e-01
lmsepiMissing              -0.2160 0.6578 7.426848e-01
lmsepiNo activity           0.4793 0.4148 2.478730e-01
lmsepiSome activity         0.1717 0.3586 6.319940e-01
```

4 ACOARM

```
> data(acodata)
> dim(acodata)
```

```
[1] 907  14
```

```
> str(acodata)
```

```
'data.frame':        907 obs. of  14 variables:
 $ vacc1_evinf      : int  1442 1489 913 920 1448 1465 377 1274 1472 1463 ...
 $ f_evinf          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ subcoh           : logi  TRUE FALSE FALSE FALSE TRUE FALSE ...
 $ ptid             : int  9601 9603 9605 9606 9607 9608 9609 9610 9613 9614 ...
 $ f_treat          : int  1 1 1 1 0 0 1 1 1 0 ...
 $ fcgr2a.3         : num  0 NA NA NA NA NA NA NA NA NA ...
 $ f_agele30        : int  0 0 0 0 1 0 0 0 1 1 ...
 $ f_hsv_2          : num  0 1 0 0 0 0 0 0 0 0 ...
 $ f_ad5gt18        : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
 $ f_crcm           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ any_drug         : num  1 1 1 0 0 0 0 1 0 0 ...
 $ num_male_part_cat: num  0 0 0 1 0 0 0 0 0 0 ...
 $ uias             : num  0 1 1 1 1 0 0 0 0 0 ...
 $ uras             : num  0 0 0 0 1 0 0 0 0 0 ...

>
```

For two-arm, placebo-controlled trials with rare failure time endpoints, we can augment the case-only design with random samples of controls from both arms, as in the classical case-cohort sampling scheme, or with a random sample of controls from the active treatment arm only. We show that these designs can identify all parameters in a Cox model and that the efficient case-only estimator can be incorporated in a two-step plug-in procedure[2]. A data example was shown in [2] that incorporating case-only estimators in the classical case-cohort design improves the precision of all estimated parameters; sampling controls only in the active treatment arm attains a similar level of efficiency. Here is an example of ACO using controls from the placebo arm.

```
> rfit0 <- acoarm(data=acodata,
+                 svtime="vacc1_evinf",
+                 event="f_evinf",
+                 treatment="f_treat",
+                 BaselineMarker="fcgr2a.3",
+                 id="ptid",
+                 subcohort="subcoh",
+                 esttype=1,
+                 augment=0,
+                 extra=c("f_agele30","f_hsv_2","f_ad5gt18","f_crcm","any_drug","num
> rfit0

                            beta  stder       pVal
fcgr2a.3 (BaselineMarker)  0.1784 0.3871 0.64494332
f_treat (Treatment)        0.6327 0.4938 0.20009115
interatcion               -0.2550 0.3821 0.50455514
f_agele30                  0.3637 0.6260 0.56120036
f_hsv_2                    1.6177 0.6588 0.01405904
f_ad5gt18                 -0.2784 0.6874 0.68553494
f_crcm                     0.5609 1.0100 0.57868519
any_drug                   0.9704 0.6623 0.14286876
num_male_part_cat         -1.7869 0.8573 0.03713643
uias                       0.7115 0.5203 0.17145652
uras                       0.9528 0.6391 0.13596896
```

Here is an example of ACO using controls from the active arm

```
> rfit1 <- acoarm(data=acodata,
+                  svtime="vacc1_evinf",
+                  event="f_evinf",
+                  treatment="f_treat",
+                  BaselineMarker="fcgr2a.3",
+                  id="ptid",
+                  subcohort="subcoh",
+                  esttype=1,
+                  augment=1,
+                  extra=c("f_agele30","f_hsv_2","f_ad5gt18","f_crcm","any_drug","num
> rfit1

                            beta  stder       pVal
fcgr2a.3 (BaselineMarker)  0.2360 0.3765 0.53070424
f_treat (Treatment)        0.6327 0.4938 0.20009115
interatcion               -0.2550 0.3821 0.50455514
f_agele30                  0.1902 0.5041 0.70593304
f_hsv_2                    0.8494 0.5389 0.11497257
f_ad5gt18                  0.3646 0.4553 0.42326823
f_crcm                    -0.1616 0.5843 0.78213299
any_drug                   1.0837 0.5540 0.05047852
num_male_part_cat          0.1792 0.6052 0.76711529
uias                       0.0663 0.4531 0.88368210
uras                       1.1437 0.4905 0.01972308
```

Here is an example of ACO using controls from both arms

```
> rfit2 <- acoarm(data=acodata,
+                  svtime="vacc1_evinf",
+                  event="f_evinf",
+                  treatment="f_treat",
+                  BaselineMarker="fcgr2a.3",
+                  id="ptid",
+                  subcohort="subcoh",
+                  esttype=1,
+                  augment=2,
+                  extra=c("f_agele30","f_hsv_2","f_ad5gt18","f_crcm","any_drug","num
> rfit2

                            beta  stder         pVal
fcgr2a.3 (BaselineMarker)  0.1904 0.3119 0.5414829219
f_treat (Treatment)        0.6327 0.4938 0.2000911539
interatcion               -0.2550 0.3821 0.5045551427
```

```
f_agele30                  0.0740 0.3436 0.8293998716
f_hsv_2                    1.2066 0.3981 0.0024400419
f_ad5gt18                  0.1039 0.3728 0.7804757254
f_crcm                     0.1086 0.4375 0.8039160862
any_drug                   1.1332 0.3709 0.0022464789
num_male_part_cat         -0.4866 0.4127 0.2383614665
uias                       0.2364 0.3324 0.4769372489
uras                       1.1534 0.3458 0.0008510074
```

5 Session Information

The version number of R and packages loaded for generating the vignette were:

```
R version 3.2.2 (2015-08-14)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.2 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] TwoPhaseInd_1.1.0

loaded via a namespace (and not attached):
[1] tools_3.2.2     survival_2.38-3 splines_3.2.2
```

References

[1] J. Y. Dai, M. LeBlanc, and C. Kooperberg. Semiparametric estimation exploiting covariate independence in two-phase randomized trials. *Biometrics*, 65(1):178–187, Mar 2009.

[2] J. Y. Dai, X. C. Zhang, C. Y. Wang, and C. Kooperberg. Augmented case-only designs for randomized clinical trials with failure time endpoints. *Biometrics*, 2015.

[3] C. Kooperberg, M. Cushman, J. Hsia, J. G. Robinson, A. K. Aragaki, J. K. Lynch, A. E. Baird, K. C. Johnson, L. H. Kuller, S. A. Beresford, and B. Rodriguez. Can biomarkers identify women at increased stroke risk? the women's health initiative hormone trials. *PLoS clinical trials*, 2(6):e28, Jun 15 2007.