The TwoPhaseInd package: Semiparametric Estimation Exploiting Covariate Independence in Two-Phase Randomized Trials

Ting-Yuan Liu and James Y. Dai

March 23, 2011

1 Introduction

The package TwoPhaseInd provides the functions to compute the semiparametric maximum likelihood estimate (SPMLE) and the maximum estimated likelihood estimator (MELE) defined in Dai et al. (2009). In a two-phase sampling study for genetics or biomarkers, the treatment and the baseline biomakers or genotypes are independent by design. Exploiting the independence between a randomized treatment and the baseline markers yields substantial power gain in estimating their interactions, as shown by Dai et al. (2009). The technical details can be found in Dai et al. (2009). This manual provides brief description of the functions provided in the TwoPhaseInd package.

2 WHI Biomarker Study

The data is originally published in Kooperberg et al. (2007) and used as an illustration example in Dai et al. (2009). Only partial data are provided in this package for demonstration purpose.

```
> data(whiBioMarker)
> ls()

[1] "whiBioMarker"
```

3 Semiparametric Maximum Likelihood Estimate (SPMLE)

Based on the profile likelihood, Dai et al. (2009) developed a Newton-Raphson algorithm to compute the semiparametric maximum likelihood estimate (SPMLE). The function `spmle` can be used to perform this algorithm with or without exploiting independent.

In a two-phase sampling scheme, we record the response variable `stroke` and the treatment `hrtdisp` for everyone; The biomarker PAP (plasmin-antiplasmin complex), `papbl`, is only recorded in the second phase case-control sample. Several important clinical characteristics in the second-phase data are included to eliminate potential confounding. The data is stored as `whiBioMarker`.

3.1 SPMLE Without Exploiting Independent

Here is an example of SPMLE without exploiting independent:

```
> spmleNonIndExtra <- spmle(data = whiBioMarker, response = "stroke",
+     treatment = "hrtdisp", BaselineMarker = "papbl", extra = c("age",
+         "dias", "syst", "diabtrt"), phase = "phase", ind = FALSE)
> spmleNonIndExtra
```

```
                          beta  stder   pVal
(Intercept)            -4.5953 0.1310 0.0000
hrtdisp (Treatment)     0.3682 0.1561 0.0183
papbl (BaselineMarker)  2.3399 1.0341 0.0237
hrtdisp:papbl          -4.2106 1.3123 0.0013
age                     1.3178 1.1699 0.2600
dias                   -0.7693 0.9918 0.4379
syst                    3.1134 0.9819 0.0015
diabtrtYes              0.8627 0.3604 0.0167
```

3.2 SPMLE With Exploiting Independent

Here is an example of SPMLE with exploiting independent:

```
> spmleIndExtra <- spmle(data = whiBioMarker, response = "stroke",
+     treatment = "hrtdisp", BaselineMarker = "papbl", extra = c("age",
+         "dias", "syst", "diabtrt"), phase = "phase", ind = TRUE)
> spmleIndExtra
```

```
                          beta  stder   pVal
(Intercept)            -4.5632 0.1246 0.0000
hrtdisp (Treatment)     0.3106 0.1467 0.0343
papbl (BaselineMarker)  1.9625 0.9288 0.0346
hrtdisp:papbl          -3.9056 1.1600 0.0008
age                     1.6406 1.1767 0.1632
dias                   -0.5812 0.9785 0.5526
syst                    2.8240 0.9695 0.0036
diabtrtYes              0.9201 0.3605 0.0107
```

4 Maximum Estimated Likelihood Estimator (MELE)

The profile information matrix is computed explicitly via numerical differentiation. In certain situations where computing the SPMLE is slow, we propose a maximum estimated likelihood estimator (MELE), which is also capable of incorporating the covariate independence.

The following examples use the same response, treatment, and biomarker as used in the SPMLE examples to illustrate the usage of MELE algorithm by the function `mele`.

4.1 MELE Without Exploiting Independent

Here is an example of MELE without exploiting independent:

```
> melNonIndExtra <- mele(data = whiBioMarker, response = "stroke",
+     treatment = "hrtdisp", BaselineMarker = "papbl", extra = c("age",
+         "dias", "syst", "diabtrt"), phase = "phase", ind = FALSE)
> melNonIndExtra
```

```
                          beta  stder   pVal
(Intercept)            -4.5625 0.1247 0.0000
hrtdisp (Treatment)     0.3094 0.1464 0.0346
papbl (BaselineMarker)  1.9320 0.9250 0.0367
hrtdisp:papbl          -3.8344 1.1571 0.0009
age                     1.6791 1.1703 0.1514
dias                   -0.6711 1.0103 0.5065
syst                    2.8814 0.9827 0.0034
diabtrtYes              0.9164 0.3690 0.0130
```

4.2 MELE with Exploiting Independent

Here is an example of MELE with exploiting independent:

```
> melIndExtra <- mele(data = whiBioMarker, response = "stroke",
+     treatment = "hrtdisp", BaselineMarker = "papbl", extra = c("age",
+         "dias", "syst", "diabtrt"), phase = "phase", ind = TRUE)
> melIndExtra
```

```
                          beta  stder   pVal
(Intercept)            -4.5604 0.1277 0.0000
hrtdisp (Treatment)     0.3400 0.1564 0.0297
papbl (BaselineMarker)  2.0236 1.0461 0.0531
hrtdisp:papbl          -3.7069 1.3232 0.0051
```

```
age                      1.7316 1.1932 0.1467
dias                    -0.8453 1.0218 0.4081
syst                     2.9314 1.0025 0.0035
diabtrtYes               0.7432 0.3823 0.0519
```

Generally speaking, the estimators exploiting independence would yield smaller stander error; the SPMLE would be more efficient than the MELE.

5 Session Information

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

```
R version 2.12.2 (2011-02-25)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
[1] C

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

other attached packages:
[1] TwoPhaseInd_1.0.0

loaded via a namespace (and not attached):
[1] tools_2.12.2
```

References

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.

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.