Supplement to: "Effect of Breastfeeding on Gastrointestinal Infection in Infants: A Targeted Maximum Likelihood Approach for Clustered Longitudinal Data"

Mireille E. Schnitzer, Mark J. van der Laan, Erica E. M. Moodie
and Robert W. Platt

January 21, 2014

1 The influence curve for clustered data

Let D_i be the component of the influence curve calculated for each individual under the assumption of no clustering. For the parameter of interest, ψ, and the TMLE estimator $\hat{\psi}^{TMLE}$, we have that

$$\hat{\psi}^{TMLE} - \psi = \frac{1}{n}\sum_{i=1}^{n} D_i$$

up to a second-order term. Relabeling for each cluster $m = 1, ..., M$ and subjects within each cluster $j \in Z_m$,

$$\begin{aligned}\hat{\psi}^{TMLE} - \psi =& \frac{1}{n}\sum_{m=1}^{M}\sum_{j\in Z_m} D_{m,j}\\ =& \frac{1}{M}\sum_{m=1}^{M}\sum_{j\in Z_m} D_{m,j}\frac{M}{n}.\end{aligned}$$

This shows that the influence curve where cluster is the experimental unit can be written as

$$X_m = \sum_{j\in Z_m} D_{m,j}\frac{M}{n}.$$

Since independence is assumed between clusters, we have that the limit variance of $\sqrt{M}(\hat{\psi}^{TMLE} - \psi)$ is $Var(X)$, the covariance matrix between the clusters.

2 Details about the data generation in the simulation study

The data in the simulation study was generated in order to resemble the PROBIT data. The simulated data is longitudinal, and grouped in 31 clusters, each with 500 subjects. Each simulated individual has an observation with the structure
$O = (W, U, C_1, L_1, A_1, C_2, L_2, A_2, C_3, L_3)$ where exposure, $A_t, t = 1, 2$ is binary, $C_t, t = 1, 2, 3$ is the censoring indicator (and therefore also monotone), $L_t, t = 1, 2, 3$ is binary and $Y = \sum_{t=1}^{6} L_t$ is a count variable. W and U are one-dimensional Gaussian random variables, representing baseline confounders, generated using a cluster-specific mean. In Section 5 of the main text, we heuristically summarize the data generation. The major differences between the generated data and the PROBIT data are that the generated outcome is a summation over the binary intermediate variables, the sample size within each cluster is fixed at 550 (resulting in a similar overall sample size), and that the simulation only generates 3 time-points. The coefficients used in the data generating procedure were all informed by logistic regressions performed using the real data from the first three time points.

We used the following functions (written in R Statistical Software version 2.13.2, R Development Core Team 2011) to generate the data:

```
##########################################################
#Robust expit function (allowing for large values of x)
expit<-function(x){
z<-exp(x)/(1+exp(x))
z[is.na(z)]<-1
return(z) }
##########################################################
```

```
#########################################################
#DATA GENERATION

data_cluster<-function(i){
set.seed(i*5436)

#31 clusters, each of size 550
n<-31*550

W<-rep(NA,n)
U<-rep(NA,n)
HOSP<-rep(NA,n)

j<-1

#Generate the cluster-specific means for W and U
w_mean<-rnorm(n=31)/4
u_mean<-rnorm(n=31)

#Generate the baseline covariates W and U from w_mean and u_mean
for(c in 1:31){
#W in each cluster
W[j:(j+550-1)]<-rnorm(n=550)/4+w_mean[c]
#U in each cluster
U[j:(j+550-1)]<-rnorm(n=550)/4+u_mean[c]
#hosp
HOSP[j:(j+550-1)]<-c

j<-j+550

}

c1<-expit(-4.6+0.5*W+U) #because logit(0.01)=-4.59512 want about 1% censored
C1<-rbinom(n=n,size=1,prob=c1)

L1<-vector(length=n)
mu1<-expit(-4.5+0.5*W+U)
L1[C1==0]<-rbinom(n=(n-sum(C1==1)),prob=mu1[C1==0],size=1)

A1<-rep(0,n)
a1<-expit(2.2-W-0.5*U-0.8*L1) #logit(0.9)=~2.2
A1[C1==0]<-rbinom(n=(n-sum(C1==1)),prob=a1[C1==0],size=1)

C2<-rep(1,n)
c2<-expit(-5.3+0.5*W+U-0.5*A1+L1) #b/c logit(0.005)=-5.3 want about 1% censored
C2[C1==0]<-rbinom(n=(n-sum(C1==1)),size=1,prob=c2[C1==0])

L2<-vector(length=n)
mu2<-expit(-4+0.5*W+U+1.5*L1-0.5*A1) #logit(0.015)~-4.2
L2[C2==0]<-rbinom(n=(n-sum(C2==1)),prob=mu2[C2==0],size=1)

A2<-rep(0,n)
a2<-expit(1.4-W-0.5*U-0.5*L1-0.6*L2) #logit(0.8)=~1.4
A2[C2==0&A1==1]<-rbinom(n=sum(C2==0&A1==1),prob=a2[C2==0&A1==1],size=1)

C3<-rep(1,n)
c3<-expit(-5.3+0.5*W+U-0.5*A2+L2) #b/c logit(0.005)=-5.3 want about 1% censored
C3[C2==0]<-rbinom(n=(n-sum(C2==1)),size=1,prob=c3[C2==0])

L3<-vector(length=n)
mu3<-expit(-3.8+0.5*W+U+1.5*L1-0.5*A1+L2-0.5*A2) #logit(0.015)~-4.2 (modified empirically)
L3[C3==0]<-rbinom(n=(n-sum(C3==1)),prob=mu3[C3==0],size=1)

Y<-L1+L2+L3

L1[C1==1]<-NA
A1[C1==1]<-NA
L2[C2==1]<-NA
A2[C2==1]<-NA
L3[C3==1]<-NA
Y[C3==1]<-NA

return(as.data.frame(cbind(W,L1,L2,L3,Y,A1,A2,C1,C2,C3,U,HOSP)))


} #end data gen
```

In the fourth simulation scenario, the baseline variables W and U are transformed using two of the transformations in Kang and Schafer (2007). The data is generated as shown above, but it is supposed that the analyst receives the transformed variables

$$\begin{aligned} W_1 =& \frac{W}{1+\exp(U)} + 10 \\ U_1 =& \exp(U/2), \end{aligned}$$

instead of W and U.

References

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. ISBN 3-900051-07-0.