Analysis of multivariate binomial data: family analysis

Klaus Holst & Thomas Scheike

February 9, 2018

Overview

When looking at multivariate binomial data with the aim of learning about the dependence that is present, possibly after correcting for some covariates many models are available.

- Random-effects models logistic regression covered elsewhere (glmer in lme4).

  in the mets package you can fit the

- Pairwise odds ratio model
- Bivariate Probit model
  - With random effects
  - Special functionality for polygenic random effects modelling such as ACE, ADE ,AE and so forth.
- Additive gamma random effects model
  - Special functionality for polygenic random effects modelling such as ACE, ADE ,AE and so forth.

These last three models are all fitted in the mets package using composite likelihoods for pairs of data. The models can be fitted specifically based on specifying which pairs one wants to use for the composite score.

The models are described in futher details in the binomial-twin vignette.

Simulated family data

We start by simulating family data with and additive gamma structure on ACE form. Here 40000 families consisting of two parents and two children. The response is ybin and there is one covariate x.

```
library(mets)
set.seed(100)
data <- simbinClaytonOakes.family.ace(40000,2,1,beta=NULL,
    alpha=NULL)
data$number <- c(1,2,3,4)
data$child <- 1*(data$number==3)
head(data)
```

```
Loading required package: timereg
Loading required package: survival
Loading required package: lava
lava version 1.5.1
mets version 1.2.1.2

Attaching package: 'mets'

The following object is masked _by_ '.GlobalEnv':

    object.defined

Warning message:
failed to assign RegisteredNativeSymbol for cor to cor since cor is already defined in the 'mets' namespace
  ybin x   type cluster number child
1    1 0 mother       1      1     0
2    1 1 father       1      2     0
3    1 1  child       1      3     1
4    1 1  child       1      4     0
5    0 0 mother       2      1     0
6    1 1 father       2      2     0
```

We fit the marginal models, and here find a covariate effect at 0.3 for x. The marginals can be specified excatly as one wants.

```
aa <- margbin <- glm(ybin~x,data=data,family=binomial())
summary(aa)
```

```
Call:
glm(formula = ybin ~ x, family = binomial(), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5283  -1.3910   0.8632   0.9779   0.9779

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.489258   0.007291    67.1   <2e-16 ***
x           0.306070   0.010553    29.0   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 206272  on 159999  degrees of freedom
Residual deviance: 205428  on 159998  degrees of freedom
AIC: 205432

Number of Fisher Scoring iterations: 4
```

Additive gamma model

For the additive gamma of this type we set-up the random effects included in such a family to make the ACE valid using some special functions for this.

The model is constructe with one enviromental effect shared by all in the family and 8 genetic random effects with size (1/4) genetic variance. Looking at the first family we see that the mother and father both share half the genes with the children and that the two children also share half their genes with this specification. Below we also show an alternative specification of this model using all pairs.

```
# make ace random effects design
out <- ace.family.design(data,member="type",id="cluster")
out$pardes
head(out$des.rv,4)
```

```
     [,1] [,2]
 [1,] 0.25    0
 [2,] 0.25    0
 [3,] 0.25    0
 [4,] 0.25    0
 [5,] 0.25    0
 [6,] 0.25    0
 [7,] 0.25    0
 [8,] 0.25    0
 [9,] 0.00    1
     m1 m2 m3 m4 f1 f2 f3 f4 env
[1,]  1  1  1  1  0  0  0  0   1
[2,]  0  0  0  0  1  1  1  1   1
[3,]  1  1  0  0  1  1  0  0   1
[4,]  1  0  1  0  1  0  1  0   1
```

We can now fit the model calling the two-stage function

```
# fitting ace model for family structure
ts <- binomial.twostage(margbin,data=data,clusters=data$
    cluster,
theta=c(2,1)/9,
random.design=out$des.rv,theta.des=out$pardes)
summary(ts)
# true variance parameters
c(2,1)/9
# total variance
1/3
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.2425610 0.03747680
dependence2 0.1255742 0.01607478

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err  2.5% 97.5%  P-value
dependence1    0.659  0.0611 0.539 0.779 4.25e-27
dependence2    0.341  0.0611 0.221 0.461 2.39e-08

$vare
NULL

$vartot
   Estimate Std.Err  2.5% 97.5%  P-value
p1    0.368  0.0252 0.319 0.418 3.31e-48

attr(,"class")
[1] "summary.mets.twostage"
[1] 0.2222222 0.1111111
[1] 0.3333333
```

Pairwise fitting

We now specify the same model via extracting all pairs. The random effecs structure is simpler when just looking at pairs. A special function writes up all combinations of pairs. There are 6 pairs within each family, and we keep track of who belongs to the different families. We first simply give the pairs and we then should get the same result as before.

```
mm <- familycluster.index(data$cluster)
head(mm$familypairindex,n=20)
pairs <- mm$pairs
dim(pairs)
head(pairs,12)
```

```
 [1] 1 2 1 3 1 4 2 3 2 4 3 4 5 6 5 7 5 8 6 7
[1] 240000      2
      [,1] [,2]
 [1,]    1    2
 [2,]    1    3
 [3,]    1    4
 [4,]    2    3
 [5,]    2    4
 [6,]    3    4
 [7,]    5    6
 [8,]    5    7
 [9,]    5    8
[10,]    6    7
[11,]    6    8
[12,]    7    8
```

Now with the pairs we fit the model

```
tsp <- binomial.twostage(margbin,data=data,
            clusters=data$cluster,
            theta=c(2,1)/9,detail=0,
          random.design=out$des.rv,theta.des=out$pardes,pairs=
              pairs)
summary(tsp)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
Error in theta.des %*% theta : non-conformable arguments
```

Here a random sample of pairs are given instead and we get other estimates.

```
set.seed(100)
ssid <- sort(sample(1:nrow(pairs),nrow(pairs)/2))
tsd <- binomial.twostage(aa,data=data,clusters=data$cluster,
          theta=c(2,1)/9,step=1.0,
          random.design=out$des.rv,iid=1,Nit=10,
          theta.des=out$pardes,pairs=pairs[ssid,])
summary(tsd)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
Error in theta.des %*% theta : non-conformable arguments
```

To specify such a model when only the pairs are availble we show how to specify the model. We here use the same marginal "aa" to make the results comparable. The marginal can also be fitted based on available data.

We start by selecting the data related to the pairs, and sets up new id's and to start we specify the model using the full design with 9 random effects. Below we show how one can use with only the random effects needed for each pair, which is typically simpler.

```
head(pairs[ssid,])
ids <- sort(unique(c(pairs[ssid,])))

pairsids <- c(pairs[ssid,])
pair.new <- matrix(fast.approx(ids,c(pairs[ssid,])),ncol=2)
head(pair.new)

dataid <- dsort(data[ids,],"cluster")
outid <- ace.family.design(dataid,member="type",id="cluster"
    )
outid$pardes
head(outid$des.rv)
```

```
     [,1] [,2]
[1,]    1    2
[2,]    1    3
[3,]    2    4
[4,]    3    4
[5,]    5    6
[6,]    5    7
     [,1] [,2]
[1,]    1    2
[2,]    1    3
[3,]    2    4
[4,]    3    4
[5,]    5    6
[6,]    5    7
      [,1] [,2]
 [1,] 0.25    0
 [2,] 0.25    0
 [3,] 0.25    0
 [4,] 0.25    0
 [5,] 0.25    0
 [6,] 0.25    0
 [7,] 0.25    0
 [8,] 0.25    0
 [9,] 0.00    1
     m1 m2 m3 m4 f1 f2 f3 f4 env
[1,]  1  1  1  1  0  0  0  0   1
[2,]  0  0  0  0  1  1  1  1   1
[3,]  1  1  0  0  1  1  0  0   1
[4,]  1  0  1  0  1  0  1  0   1
[5,]  1  1  1  1  0  0  0  0   1
[6,]  0  0  0  0  1  1  1  1   1
```

Now fitting the model with the data set up

```
tsdid <- binomial.twostage(aa,data=dataid,clusters=dataid$
    cluster,
    theta=c(2,1)/9,
    random.design=outid$des.rv,theta.des=outid$pardes,pairs=
        pair.new)
summary(tsdid)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
Error in theta.des %*% theta : non-conformable arguments
```

We now specify the design specifically using the pairs. The random.design and design on the parameters are now given for each pair, as a 3 dimensional matrix. with a direct specification of random.design and the design on the parameters theta.design. In addition we need also to give the number of random effects for each pair. These basic things are constructed by certain functions for the ACE design.

```
pair.types <- matrix(dataid[c(t(pair.new)),"type"],byrow=T,
    ncol=2)
head(pair.new,7)
head(pair.types,7)

theta.des <- array(0,c(4,2,nrow(pair.new)))
random.des <- array(0,c(2,4,nrow(pair.new)))
# random variables in each pair
rvs <- c()
for (i in 1:nrow(pair.new))
{
   if (pair.types[i,1]=="mother" & pair.types[i,2]=="father"
       )
   {
   theta.des[,,i] <- rbind(c(1,0),c(1,0),c(0,1),c(0,0))
   random.des[,,i] <- rbind(c(1,0,1,0),c(0,1,1,0))
   rvs <- c(rvs,3)
   } else {
   theta.des[,,i] <- rbind(c(0.5,0),c(0.5,0),c(0.5,0),c(0,1)
       )
   random.des[,,i] <- rbind(c(1,1,0,1),c(1,0,1,1))
   rvs <- c(rvs,4)
   }
}
```

```
     [,1] [,2]
[1,]    1    2
[2,]    1    3
[3,]    2    4
[4,]    3    4
[5,]    5    6
[6,]    5    7
[7,]    5    8
     [,1]     [,2]
[1,] "mother" "father"
[2,] "mother" "child"
[3,] "father" "child"
[4,] "child"  "child"
[5,] "mother" "father"
[6,] "mother" "child"
[7,] "mother" "child"
```

For pair 1 that is a mother/farther pair, we see that they share 1 environmental random effect of size 1. There are also two genetic effects that are unshared between the two. So a total of 3 random effects are needed here. The theta.des relates the 3 random effects

to possible relationships in the parameters. Here the genetic effects are full and so is the environmental effect. In contrast we also consider a mother/child pair that share half the genes, now with random effects with (1/2) gene variance. We there need 4 random effects, 2 non-shared half-gene, 1 shared half-gene, and one shared full environmental effect.

```
# 3 rvs here
random.des[,,1]
theta.des[,,1]
# 4 rvs here
random.des[,,2]
theta.des[,,2]
head(rvs)
```

```
     [,1] [,2] [,3] [,4]
[1,]    1    0    1    0
[2,]    0    1    1    0
     [,1] [,2]
[1,]    1    0
[2,]    1    0
[3,]    0    1
[4,]    0    0
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    1
[2,]    1    0    1    1
     [,1] [,2]
[1,]  0.5    0
[2,]  0.5    0
[3,]  0.5    0
[4,]  0.0    1
[1] 3 4 4 4 3 4
```

Now fitting the model, and we see that it is a lot quicker due to the fewer random effects needed for pairs.

```
tsdid2 <- binomial.twostage(aa,data=dataid,clusters=dataid$
    cluster,
      theta=c(2,1)/9,
      random.design=random.des,
      theta.des=theta.des,pairs=pair.new,pairs.rvs=rvs)
summary(tsdid2)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
Error in theta.des %*% theta : non-conformable arguments
```

The same model can be specifed even simpler via the kinship coefficient. For this speicification there are 4 random effects for each pair, but some have variance 0. The mother-father pair, here shares a random effect with variance 0, and have two non-shared genetic effects with full variance, in addition to a fully shared environmental effect.

```
kinship <- c()
for (i in 1:nrow(pair.new))
{
if (pair.types[i,1]=="mother" & pair.types[i,2]=="father")
    pk1 <- 0 else pk1 <- 0.5
```

```
kinship <- c(kinship,pk1)
}
head(kinship,n=10)

out <- make.pairwise.design(pair.new,kinship,type="ace")
names(out)
out$random.des[,,1]
out$theta.des[,,1]
```

```
 [1] 0.0 0.5 0.5 0.5 0.0 0.5 0.5 0.5 0.5 0.5
[1] "random.design" "theta.des"     "ant.rvs"
     [,1] [,2] [,3] [,4]
[1,]    1    1    0    1
[2,]    1    0    1    1
     [,1] [,2]
[1,]    0    0
[2,]    1    0
[3,]    1    0
[4,]    0    1
```

Now, fitting the model we get the results from before.

```
tsdid3 <- binomial.twostage(aa,data=dataid,clusters=dataid$
    cluster,
       theta=c(2,1)/9,random.design=out$random.design,
       theta.des=out$theta.des,pairs=pair.new,pairs.rvs=out$
           ant.rvs)
summary(tsdid3)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
Error in theta.des %*% theta : non-conformable arguments
```

Pairwise odds ratio model

To fit the pairwise odds-ratio model in the case of a pair-specification there are two options for fitting the model.

1. One option is to set up some artificial data similar to twin data with
   - a pair-cluster-id (clusters)
   - with a cluster-id to get GEE type standard errors (se.cluster)
2. We can also use the specify the design via the theta.des that is also a matrix of dimension pairs x design with the design for POR model.

Starting by the second option. We need to start by specify the design of the odds-ratio of each pair. We set up the data and find all combinations within the pairs. Subsequently, we remove all the empty groups, by grouping together the factor levels 4:9, and then we construct the design.

```
tdp <-cbind( dataid[pair.new[,1],],dataid[pair.new[,2],])
names(tdp) <- c(paste(names(dataid),"1",sep=""),
       paste(names(dataid),"2",sep=""))
tdp <-transform(tdp,tt=interaction(type1,type2))
dlevel(tdp)
drelevel(tdp,newlevels=list(mother.father=4:9)) <- obs.types
    ~tt
dtable(tdp,~tt+obs.types)
tdp <- model.matrix(~-1+factor(obs.types),tdp)
```

```
type1 #levels=:3
[1] "child"  "father" "mother"
----------------------------------------
type2 #levels=:3
[1] "child"  "father" "mother"
----------------------------------------
tt #levels=:9
[1] "child.child"   "father.child"  "mother.child"  "child.father"
[5] "father.father" "mother.father" "child.mother"  "father.mother"
[9] "mother.mother"
----------------------------------------
```

tt	obs.types	mother.father	child.child	father.child	mother.child
child.child		0	19991	0	0
father.child		0	0	39837	0
mother.child		0	0	0	40212
child.father		0	0	0	0
father.father		0	0	0	0
mother.father		19960	0	0	0
child.mother		0	0	0	0
father.mother		0	0	0	0
mother.mother		0	0	0	0

We then can fit the pairwise model using the pairs and the pair-design for descrbing the OR. The results are consistent with the the ACE model as the mother-father have a lower dependence as is due only the environmental effects. All other combinations should have the same dependence as also seem to be the case.

To fit the OR model it is generally recommended to use the var.link to use the parmetrization with log-odd-ratio regression.

```
porpair <- binomial.twostage(aa,data=dataid,clusters=dataid$
    cluster,
      theta.des=tdp,pairs=pair.new,model="or",var.link=1)
summary(porpair)
```

```
Dependence parameter for Odds-Ratio (Plackett) model
With log-link
$estimates
                                  theta         se
factor(obs.types)mother.father 0.1269881 0.03132228
factor(obs.types)child.child   0.3819107 0.03108233
factor(obs.types)father.child  0.3046284 0.02239909
factor(obs.types)mother.child  0.3293741 0.02233648

$or
                         Estimate Std.Err 2.5% 97.5%   P-value
factor(obs.types)moth....     1.14  0.0356 1.07  1.21 1.16e-223
factor(obs.types)chil....     1.47  0.0455 1.38  1.55 4.26e-227
factor(obs.types)fath....     1.36  0.0304 1.30  1.42  0.00e+00
```

```
factor(obs.types)moth.....1     1.39  0.0310 1.33  1.45  0.00e+00

$type
[1] "or"

attr(,"class")
[1] "summary.mets.twostage"
```