Analysis of multivariate binomial data: case control or ascertainment sampling

Klaus Holst & Thomas Scheike

March 3, 2020

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.

Case-Control Sampling

Sometimes, pairs are recruited after a case-proband is selected. This proband, can be either a

- case: must be representative of cases

  or a

- control: must be representative of controls

First thinking about pairs, we estimate parameters using the conditional likelihood given sampling wich for a binary 2 x 2 table can be written as

$$\frac{P(i,j)}{P(j)}$$

the probailty of seeing (i,j) for the pair, given that the proband was observed as (j).

We note that if the marginal is known, or possibly estimated from the full cohort. Then we can estimate dependence parameters using just the terms $P(i,j)$ for the pairs. We can thus ignore the special sampling for models with marginal specification. If the marginal can not be obtained from other sources we need to maximize the full-pairwise-likelihood in all parameters, that is both dependence and marginal parameters.

Similary, one can select a whole family based on having selected a proband, that is selected a representative member of either cases or controls. In this case we fit the models by using composited likelihoods, considering all pairs that involves the probands. This will give some lacking efficiency compared to looking at the full likelihood of the family given the proband.

Ascertainment Sampling

Similarly, in the setting of pairs we can select all pairs where there is at least one event of interest.

First thinking about pairs, we estimate parameters using the conditional likelihood given sampling wich for a binary 2 x 2 table can be written as

$$\frac{P(i,j)}{1-P(0,0)}$$

the probailty of seeing (i,j) for the pair, given that it is sampled.

If the marginal can estimated from a full sample we can then estimate the dependence parameter using the ascertainment likelihood.

Generally, when whole families are ascertained the computation of the true truncation probability can be hard to the fact that families are hard to define in the real world. Nevertheless, if a random sample of such family is at hand. We suggest to in these families take out all pairs that satisfies the ascertainment criterion. With a family, with given size n we have binary observations $(Y_1, ..., Y_n)$. The family is sampled or a random sample of families such that

$$\sum_{i=1}^{n} Y_i \geq 1.$$

We let the conditional distribution given sampling, be denoted as

$$P^O(\cdot) = P(\cdot | \sum_{i=1}^{n} Y_i \geq 1)$$

Now, we note that all pairs within these family that satisfies that

$Y_i + Y_j \geq 1$, will have distribution

$$\begin{aligned} P^O(Y_i = o_1, Y_j = o_2 | Y_i + Y_j \geq 1) &= \frac{P^O(o_1, o_2)}{P^O(Y_i + Y_j \geq 1)} \\ &= \frac{P(Y_i = o_1, Y_j = o_2, \sum_{i=1}^n Y_i \geq 1)}{P(Y_i + Y_j \geq 1, \sum_{i=1}^n Y_i \geq 1)} \\ &= \frac{P(Y_i = o_1, Y_j = o_2)}{P(Y_i + Y_j \geq 1)} = \frac{P(o_1, o_2)}{1 - P(0,0)} \end{aligned}$$

since we only consider the probabilities where $o_1 + o_2 \geq 1$. Also here we could condition on covariates.

So considering these pairs, or a random sample of them should yield valid inference. When standard errors are computed we need to rely on GEE type arguments. An advantage of this is that the ascertainment probability is much easier to get for the pairs. Again using the pairwise structure will lead to loss of efficiency compared to using the full likelihood of the ascertained families.

In addition we note that when looking at one pair that has been ascertained then

$$P(Y_i = o_1, Y_j = o_2 | Y_i + Y_j \geq 1) = \sum_{k=1}^{2} P(Y_i = o_1, Y_j = o_2 | Y_i + Y_j = k) P(Y_i + Y_j = k | Y_i + Y_j \geq 1).$$

where $o_1 + o_2 \geq 1$. Note that the dependence will affect the probabilities $P(Y_i + Y_j = 2)/(P(Y_i + Y_j = 2) + P(Y_i + Y_j = 1))$ and $P(Y_i + Y_j = 1)/(P(Y_i + Y_j = 2) + P(Y_i + Y_j = 1))$. In particular when the marginal parameters are known the dependence parameters can be estimated using the proportion of concordant pairs compared to the non-concordant pairs with respect to the outcome.

When considering the pairs with different responses we learn "only" (up to model specification) about covariate effects. For example when $\text{logit}(P(Y_i = 1|\alpha_k)) = \alpha_k + \beta X_i$ for $i = 1, 2$ with α_k a pair (cluster) specific effect and subject specific covariates X_i for $i = 1, 2$, then $P(Y_i = 1, Y_j = 0)/P(Y_i + Y_j = 1) = \text{expit}((X_i - X_j)\beta)$, and with the standard definitions $\text{logit}(p) = \log(p/(1-p))$ and $\text{expit}(x) = \exp(x)/(1 + \exp(x))$.

The twin-stutter data

We consider the twin-stutter where for pairs of twins that are either dizygotic or monozygotic we have recorded whether the twins are stuttering [1]

We here consider MZ and same sex DZ twins.

Looking at the data

```r
library(mets)
data(twinstut)
twinstut$binstut <- 1*(twinstut$stutter=="yes")
twinstut <- subset(twinstut,zyg%in%c("mz","dz"))
head(twinstut)
```

[1]

```
Loading required package: timereg
Loading required package: survival
Loading required package: lava
lava version 1.6
mets version 1.2.3

Attaching package: ‘mets’

The following object is masked _by_ ‘.GlobalEnv’:

    object.defined
   tvparnr zyg stutter    sex age nr binstut
1        1  mz      no female  71  1       0
2        1  mz      no female  71  2       0
3        2  dz      no female  71  1       0
8        5  mz      no female  71  1       0
9        5  mz      no female  71  2       0
11       7  dz      no   male  71  1       0
```

- First, we select an ascertaiment sample of the data, thus selecting a random sample of all ascertained pairs.
- Secondly, we select a case-control sample of this data to illustrate the use of the methods.

Ascertaiment Sampling

Selecting the ascertained pairs

```
library(mets)
data(twinstut)
twinstut$binstut <- 1*(twinstut$stutter=="yes")
twinstut <- subset(twinstut,zyg%in%c("mz","dz"))
dnumeric(twinstut) <- ~.
dfactor(twinstut,labels=c("DZ","MZ")) <- binzyg~zyg.n
ddrop(twinstut) <- ~"*.n"

twinstut <- dby(twinstut,binstut~tvparnr,stuttot=sum,nn=seq_along,n=length)
twina <- subset(twinstut,n==2 & stuttot>=1)
```

Selecting on the pairs where there is stuttering at taking a look at the tables of discordance and concordance for the twins.

```
twinda <- fast.reshape(twina,id="tvparnr")
twind <- fast.reshape(twinstut,id="tvparnr")
dtable(twind,"binst*"~I(stuttot1>=1))
dtable(twinda,~"binst*")
```

```
I(stuttot1 >= 1): FALSE

         binstut2    0
binstut1
0                 6632
------------------------------------------------------------
I(stuttot1 >= 1): TRUE

         binstut2   0   1
binstut1
0                   0 289
```

```
1                 281 111

         binstut2   0   1
binstut1
0                   0 289
1                 281 111
```

Now doing the analyses

Biprobit model

Looking at the full data for comparison. We estimate an unstructured probit model with different correlations for MZ and DZ twins.

```
b1 <- biprobit(binstut~sex,~-1+binzyg,data=twinstut,id="
     tvparnr")
summary(b1)
```

```
              Estimate    Std.Err          Z p-value
(Intercept)  -1.794821   0.023289 -77.066826  0.0000
sexmale       0.401430   0.030179  13.301756  0.0000
r:binzygDZ    0.132458   0.062516   2.118802  0.0341
r:binzygMZ    1.096915   0.073574  14.909085  0.0000

logLik: -4400.536  mean(score^2): 1.022e-06
    n pairs
21288  7313

Contrast:
        Dependence    [binzygDZ]
        Mean          [(Intercept)]

                        Estimate 2.5%    97.5%
Rel.Recur.Risk          1.77662  0.92746 2.62577
OR                      1.88752  1.09432 3.25566
Tetrachoric correlation 0.13169  0.00993 0.24960

Concordance             0.00235  0.00140 0.00393
Casewise Concordance    0.06456  0.03937 0.10413
Marginal                0.03634  0.03287 0.04016
```

Note, that the Casewise Concordance is a consistently estimated under complete ascertainment, i.e., when we consider a random sample of affected twins (at least on of the twins must have the event).

Odd-Ratio modelling

First looking at the marginal model based on the full data we find the overall level of stuttering and also that males have a much higher stuttering risk.

```
margbin <- glm(binstut~factor(sex),data=twinstut,family=
     binomial())
summary(margbin)
```

```
Call:
```

```
glm(formula = binstut ~ factor(sex), family = binomial(), data = twinstut)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4127  -0.4127  -0.2716  -0.2716   2.5763

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.28191    0.05000  -65.64   <2e-16 ***
factor(sex)male  0.86171    0.06211   13.87   <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: 9328.6  on 21287  degrees of freedom
Residual deviance: 9124.7  on 21286  degrees of freedom
AIC: 9128.7

Number of Fisher Scoring iterations: 6
```

First, fitting the OR model for MZ and DZ for the full data, we find that MZ have a much higher dependence than DZ twins.

```
theta.des <- model.matrix( ~-1+factor(zyg),data=twinstut)
bin <- binomial.twostage(margbin,data=twinstut,var.link=1,
             clusters=twinstut$tvparnr,theta.des=theta.des)
summary(bin)
```

```
Dependence parameter for Odds-Ratio (Plackett) model
With log-link
$estimates
                   theta        se
factor(zyg)dz 0.5238541 0.2400861
factor(zyg)mz 3.4930902 0.1865567

$or
              Estimate Std.Err   2.5%  97.5%   P-value
factor(zyg)dz    1.689  0.4054  0.894  2.483 3.111e-05
factor(zyg)mz   32.887  6.1354 20.862 44.913 8.308e-08

$type
[1] "plackett"

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

Now, using the overall marginal we look at the adjusted likelihood and find very similar results on the ascertained sample. Note, that the marginals are crucial for this analysis to give useful results.

```
theta.des <- model.matrix( ~-1+factor(zyg),data=twina)
bina <- binomial.twostage(margbin,data=twina,var.link=1,
        clusters=twina$tvparnr,theta.des=theta.des,
           pair.ascertained=1)
summary(bina)
```

```
Dependence parameter for Odds-Ratio (Plackett) model
With log-link
$estimates
                   theta        se
factor(zyg)dz 0.4874213 0.2472523
factor(zyg)mz 3.4753766 0.1985974
```

```
$or
              Estimate Std.Err    2.5%  97.5%   P-value
factor(zyg)dz    1.628  0.4026  0.8391  2.417 5.245e-05
factor(zyg)mz   32.310  6.4167 19.7335 44.886 4.771e-07

$type
[1] "plackett"

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

Additive gamma modelling

First, again for comparision fitting the full data for the AE model. We get the size of the genetic variance in this model.

```
out <- twin.polygen.design(twinstut,id="tvparnr",zygname="zyg",zyg="dz",type="ae")
bintwin <- binomial.twostage(margbin,data=twinstut,
     clusters=twinstut$tvparnr,detail=0,theta=c(0.1)/1,var.link=0,
     random.design=out$des.rv,theta.des=out$pardes)
summary(bintwin)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.9094847 0.09536268

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1   0.9095 0.09536 0.7226 1.096 1.469e-21

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

We first here take at the look at the marginal model for the ascertained sample, and note as expected that this sample give highly biased estimated for the marginal model.

```
outa <- twin.polygen.design(twina,id="tvparnr",zygname="zyg",zyg="dz",type="ae")
marga <- glm(binstut~sex,data=twina,family=binomial())
summary(marga)
```

```

Call:
glm(formula = binstut ~ sex, family = binomial(), data = twina)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.334  -1.298   1.028   1.028   1.061

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.27895    0.08739   3.192  0.00141 **
sexmale      0.08242    0.11237   0.733  0.46328
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1851.8  on 1361  degrees of freedom
Residual deviance: 1851.2  on 1360  degrees of freedom
AIC: 1855.2

Number of Fisher Scoring iterations: 4
```

Now, using the overall marginal model we look at the adjusted likelihood and find very similar results on the ascertained sample. Note, that the marginals are crucial for this analysis to give useful results.

```
abintwin1 <- binomial.twostage(margbin,data=twina,
          clusters=twina$tvparnr,detail=0,theta=c(0.1)/1,var.
              link=0,
          random.design=outa$des.rv,theta.des=outa$pardes,
              pair.ascertained=1)
summary(abintwin1)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.8920274 0.09732786

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1    0.892 0.09733 0.7013 1.083 4.946e-20

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

In fact for this model we can also do a full-MLE fitting jointly the dependence parameters and the marginal model. This is based on the twostage option (twostage=0 is MLE). Here the starting value is given at the marginal model for the ascertained model. This gives quite similar results to the previous analyses with a genetic variance around 1.

```
aabintwin1 <- binomial.twostage(marga,data=twina,
        clusters=twina$tvparnr,detail=0,theta=c(0.1)/1,var.
            link=0,
        random.design=outa$des.rv,theta.des=outa$pardes,pair.ascertained=1,twostage=0)
summary(aabintwin1)
coef(marga)
coef(margbin)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
               theta        se
dependence1 1.014398 0.1045593

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1    1.014  0.1046 0.8095 1.219 2.967e-22

attr(,"class")
[1] "summary.mets.twostage"
(Intercept)     sexmale 
  0.2789484   0.0824214 
    (Intercept) factor(sex)male 
     -3.2819072       0.8617053 
```

Case Control Sampling

First, taking out all cases and one control for each case, we establish the pairs of these probands. This is based on keeping track of the twin related to the proband. Here using some utility functions in the mets packages.

Then we write up the random design vectors and the parameter design for each pair using the kinship coefficient.

When specifying the pairs in the case-control setup the second column should be the probands.

```
library(mets)
data(twinstut)
twinstut$binstut <- 1*(twinstut$stutter=="yes")
twinstut <- subset(twinstut,zyg%in%c("mz","dz"))
dnumeric(twinstut) <- ~.
dfactor(twinstut,labels=c("DZ","MZ")) <- binzyg~zyg.n
ddrop(twinstut) <- ~"*.n"

twinstut <- dby(twinstut,binstut~tvparnr,stuttot=sum,nn=seq_along,n=length)
twinstut <- subset(twinstut,n==2)
twinstut <- dtransform(twinstut,nnrow=1:nrow(twinstut))
twinstut <- dby(twinstut,binstut~tvparnr,nnn=seq_along)
twinstut <- dby2(twinstut,nnrow~tvparnr,pairnr=rev)

```

```
cases <- which(twinstut$binstut==1)
controls <- sample(which(twinstut$binstut==0),1217)
rowsca <- with(twinstut,nnrow[cases])
rowsco <- with(twinstut,nnrow[controls])
rpairs <- c(rowsca,rowsco)
cc.pairs <- cbind( with(twinstut,pairnr.nnrow[rpairs]),
    rpairs)

ids <- sort(unique(c(cc.pairs)))

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

dataid <- dsort(twinstut[ids,],"tvparnr")
dataid=dtransform(dataid,kinship=0.5)
dataid=dtransform(dataid,kinship=1,binzyg=="MZ")
kinship <- dataid$kinship[pair.new[,2]]

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

```
     [,1] [,2]
[1,]    4    3
[2,]   16   15
[3,]   18   17
[4,]   32   31
[5,]   38   37
[6,]   44   43
[1] "random.design" "theta.des"     "ant.rvs"
     [,1] [,2] [,3]
[1,]    1    1    0
[2,]    1    0    1
[1] 0.5 0.5 0.5
```

Now doing the analyses, first with know marginals, that is marginals from the full data. For this analysis, since marginals do not contain dependence parameters we do not need to specify that this is case-control sampling. Having a correct is crucial for this to work, but this is certainly often possible in register based studies where a full cohort is also available.

```
cc <- binomial.twostage(margbin,data=dataid,clusters=dataid$
    tvparnr,pairs=pair.new,
       random.design=out$random.design,theta.des=out$theta.
           des,
        pairs.rvs=out$ant.rvs,case.control=0,twostage=1)
summary(cc)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.8791843 0.09707036

$type
[1] "clayton.oakes"
```

```
$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1   0.8792 0.09707 0.6889 1.069 1.339e-19

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

We now do the same analysis specifying the case-control sampling. This should result in the same dependence parameters as is also the case.

```
cc3 <- binomial.twostage(margbin,data=dataid,
             clusters=dataid$tvparnr,
        pairs=pair.new,
        random.design=out$random.design,
        theta.des=out$theta.des,
        pairs.rvs=out$ant.rvs,
        case.control=1,twostage=1)
summary(cc3)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.8791843 0.09707036

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1   0.8792 0.09707 0.6889 1.069 1.339e-19

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

This model can also be fitted using a full likelihood of both dependence parameters and marginal parameters. Here there is no need to have a correctly specified marginal. We here use the marginal fitting from the case-control data as as starting values. Again we find a genetic variance around 1. The marginal parameters are also consistent with the results from the full analyses for the marginal parameters.

```
marga <- glm(binstut~sex,data=dataid,family=binomial())
cc3 <- binomial.twostage(marga,data=dataid,
        clusters=dataid$tvparnr,
        pairs=pair.new,
        random.design=out$random.design,
        theta.des=out$theta.des,
        pairs.rvs=out$ant.rvs,
        case.control=1,twostage=0)
summary(cc3)
```

```
Dependence parameter for Clayton-Oakes model
Variance of Gamma distributed random effects
$estimates
                theta         se
dependence1 0.9222504 0.09729347

$type
[1] "clayton.oakes"

$h
            Estimate Std.Err 2.5% 97.5% P-value
dependence1        1       0    1     1       0

$vare
NULL

$vartot
   Estimate Std.Err   2.5% 97.5%   P-value
p1   0.9223 0.09729 0.7316 1.113 2.566e-21

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

When probands are related, here we may choose both case and controls from the same twin-pair then we need to adjust standard errors by grouping together contribution from related probands. This can be done using the se.cluster option that specifies how to cluster in the computation of the standard errors. In this case, however, this will be same as the clusters since these also are identical across pairs.

Combining Case Control and Ascertainment Sampling

When specifying such models based on the pairs, it is in fact possible to combine ascertained pairs with case-control sampling by specifing vectors as the case.control=c(1,0,1,0) and pair.ascertained=c(0,1,0,1) arguments. Here with two case-control pairs, and two ascertained pairs.