Modelling bivariate binary responses with application to twin data

Paul Dickman, Annica Dominicus, Juni Palmgren
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
15 May 2003

Outline

• Background: Twin studies of mood disorders have indicated that the concordance of MZ twins differs markedly from that of DZ twins. The serotonergic system has been implicated in mood disorders.

• Aim: Model the association between a binary outcome (depressed mood) and a candidate gene (5-HT2A gene promoter polymorphism) in a cross-sectional study of elderly individuals [1].

• Complication: The study subjects are twins so cannot be considered independent.

• We’ll present this example involving one explanatory variable. It is simple to extend the models to the situation where we wish to control for additional confounders or study interactions.

• The exposures do not necessarily have to be genotypic.

Participants

• 1583 individuals from the Swedish Twin Registry (OCTO-twin, SATSA, Gender).
  – 221 pairs of MZ twins
  – 486 pairs of DZ twins
  – 169 singles (all DZ)

• Outcome defined as CES-D score of 16 or higher.

• The published paper was based on 1594 individuals – we are working with an earlier version of the data.

Approaches to modelling bivariate binary responses

• Juni Palmgren has worked extensively on these issues [2]. Annica Dominicus is also working in this area.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixed models</td>
<td>Conditional</td>
</tr>
<tr>
<td>Bivariate logistic</td>
<td>Marginal</td>
</tr>
<tr>
<td>GEE</td>
<td>Marginal</td>
</tr>
<tr>
<td>Alternating logistic</td>
<td>Marginal</td>
</tr>
</tbody>
</table>

Approach Inference

1. Mixed models.
   • Conditional (subject-specific) rather than marginal (population-averaged) inference.

2. Bivariate logistic regression.
   • Full likelihood, but can be difficult to implement in practice using available software.

   • Easy to implement but not easy (in practice) to allow different correlations for MZ and DZ.

4. Alternating logistic regression.
   • A special case of GEE, applicable only to binary outcomes, where the within twin pair association is modelled using odds ratios.

Generalised Estimating Equations (GEE)

• An attractive approach for estimating population averaged effects with correlated data.

• Instead of specifying a full distribution for the multivariate binary response we make assumptions about the mean, variance, and correlation structure.

• Generally only a small efficiency loss compared to a full likelihood approach.

• The estimated regression parameters are consistent (even if the covariance structure has been misspecified) and asymptotically normal.

• Robust standard errors.

• There exists a range of approaches for estimating marginal models using GEE – we will work with what is known as GEE1.

• In a comparison of various approaches for estimating the marginal model for bivariate binary responses, Glynn and Rosner [3] concluded that ‘none was uniformly superior to the others’.

• GEE2 could be used to obtain more efficient estimates of the twin associations, but at the price of losing robustness in the estimation of the marginal logits (Garrett Fitzmaurice).
### Estimating the marginal model in SAS

```sas
proc genmod data=mj.depress;
  class twin_id g;
  model d=g / type3 error=bin link=logit;
  repeated subject=twin_id / type=exch corrw;
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1605</td>
<td>0.1042</td>
<td>-11.14</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3995</td>
<td>0.1841</td>
<td>2.17</td>
<td>0.0300</td>
<td></td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1325</td>
<td>0.1377</td>
<td>-0.96</td>
<td>0.3367</td>
<td></td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

- Estimated within-twinpair correlation is 0.267.
- This is the model we used in the published paper [1].

### ALR assuming the same association for MZ and DZ

```sas
proc genmod data=mj.depress;
  class twin_id g;
  model d=g / type3 error=bin link=logit;
  repeated subject=twin_id / logor=exch;
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1605</td>
<td>0.1041</td>
<td>-11.14</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3983</td>
<td>0.1840</td>
<td>2.16</td>
<td>0.0304</td>
<td></td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1323</td>
<td>0.1376</td>
<td>-0.96</td>
<td>0.3363</td>
<td></td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

- Alpha1 is the estimated log OR.

### ALR allowing separate association for MZ and DZ

```sas
proc genmod data=mj.depress;
  class twin_id g dz;
  model d=g / type3 error=bin link=logit;
  repeated subject=twin_id / logor=exch;
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1672</td>
<td>0.1042</td>
<td>-11.21</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3932</td>
<td>0.1836</td>
<td>2.14</td>
<td>0.0322</td>
<td></td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1323</td>
<td>0.1376</td>
<td>-0.96</td>
<td>0.3363</td>
<td></td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Alpha1</td>
<td>1.3198</td>
<td>0.1944</td>
<td>6.79</td>
<td>&lt;.0001</td>
<td>3.74</td>
</tr>
<tr>
<td>Alpha2</td>
<td>1.0903</td>
<td>0.2369</td>
<td>4.60</td>
<td>&lt;.0001</td>
<td>3.0</td>
</tr>
</tbody>
</table>

- Alpha1 is the estimated log OR for MZ twins and Alpha2 is the estimated log OR for DZ twins.

### We modelled the within-twinpair association as a correlation and assumed that the association was the same for MZ and DZ twins.
- We knew that the correlation structure was misspecified but argued that the parameter estimates were consistent and the standard errors valid.
- There is no theoretical reason that we cannot estimate a single model for the mean structure but allow the association to differ for MZ and DZ twins. This is not, however, supported in standard software.
- Iachina et al. [4] presented a SAS IML macro for estimating the marginal model using GEE while allowing a model for the correlation structure, but it is not especially user-friendly.
- Carey et al. [5] proposed a related model, applicable only to multivariate binary data, where the within-twinpair association is modelled as an odds ratio rather than a correlation. This model is known as alternating logistic regression (ALR) and is implemented in SAS.

### It so happens that with ALR in SAS, one can estimate a single model for the mean while allowing separate within-twinpair associations for MZ and DZ twins.
- I’ll start by fitting the model where we assume the within-twinpair association is the same for MZ and DZ twins.

### The log OR is 1.3198 implying that the OR is exp(1.3198) = 3.74.

\[
OR = \frac{\text{odds(twin 2 depressed | twin 1 depressed)}}{\text{odds(twin 2 depressed | twin 1 not depressed))}}
\]

- Expressing the within-twinpair associations as ORs rather than correlations has very little effect on the parameter estimates.
- There are reasons to prefer the odds ratio.
- The estimated correlation coefficient is often restricted to a range other than [-1, 1] depending on the marginal probabilities.
- Estimates of the mean parameters are more robust to misspecification of the residual twin pair association when odds ratios are used rather than correlations (Juni has a nice example).

### We see that specifying a more appropriate structure for the twinpair associations has not substantially changed the parameter estimates or standard errors. In other words, the published estimates were reasonable.
- The within-twinpair associations are, as expected, stronger for MZ twins than DZ twins indicating that there remains an unexplained genetic component.
- If a genetic variant is a meaningful predictor then we should see a reduced difference in the within twinpair association between MZ and DZ twins.
- It is therefore of interest to be able to estimate this difference and to, for example, test whether the difference is statistically significant.
- SAS allows us to fit the same model, but with a different parameterisation for the within-twinpair associations.

\[
\log OR(Y_1, Y_2) = \alpha_1 + \alpha_2 MZ
\]

where MZ is an indicator for MZ.

- That is, \(\alpha_1\) is the estimated log OR for DZ twins and \(\alpha_1 + \alpha_2\) is the estimated log OR for MZ twins.
- We first must construct a design matrix (in the form of a SAS data set) specifying the model for the within-twinpair association. Should have one row for each twin pair.
- The first column of this matrix is all 1’s whereas the second column is 1 for MZ and 0 for DZ.

```sas
proc sort data=mj.depress(keep=twin_id dz) out=zin nodupkey;
  by twin_id;
  z2=1-dz;
  data zin;
  z1=1;
  set zin;
  z1=1-dz;
run;
```

### That is not, however, supported in standard software.
- There is no theoretical reason that we cannot estimate a single model for the mean structure but allow the association to differ for MZ and DZ twins. This is not, however, supported in standard software.
- Iachina et al. [4] presented a SAS IML macro for estimating the marginal model using GEE while allowing a model for the correlation structure, but it is not especially user-friendly.
- Carey et al. [5] proposed a related model, applicable only to multivariate binary data, where the within-twinpair association is modelled as an odds ratio rather than a correlation. This model is known as alternating logistic regression (ALR) and is implemented in SAS.

### ALR allowing separate association for MZ and DZ

```sas
proc genmod data=mj.depress;
  class twin_id g dz;
  model d=g / type3 error=bin link=logit;
  repeated subject=twin_id / logor=exch varz(dz);
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1672</td>
<td>0.1042</td>
<td>-11.21</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3932</td>
<td>0.1836</td>
<td>2.14</td>
<td>0.0322</td>
<td></td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1289</td>
<td>0.1377</td>
<td>-0.94</td>
<td>0.3491</td>
<td></td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Alpha1</td>
<td>1.7464</td>
<td>0.3425</td>
<td>5.10</td>
<td>&lt;.0001</td>
<td>5.7</td>
</tr>
<tr>
<td>Alpha2</td>
<td>1.0903</td>
<td>0.2369</td>
<td>4.60</td>
<td>&lt;.0001</td>
<td>3.0</td>
</tr>
</tbody>
</table>

- Alpha1 is the estimated log OR for MZ twins and Alpha2 is the estimated log OR for DZ twins. In the model without genotype these estimates were 1.778 and 1.068.

### ALR assuming the same association for MZ and DZ

```sas
proc genmod data=mj.depress;
  class twin_id g;
  model d=g / type3 error=bin link=logit;
  repeated subject=twin_id / logor=exch;
run;
```

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Error</th>
<th>Z</th>
<th>Pr &gt;</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1605</td>
<td>0.1041</td>
<td>-11.14</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3993</td>
<td>0.1840</td>
<td>2.17</td>
<td>0.0304</td>
<td></td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1325</td>
<td>0.1377</td>
<td>-0.96</td>
<td>0.3367</td>
<td></td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

- Alpha1 is the estimated log OR.
proc genmod data=mj.depress;
class twin_id g;
model d=g / type3 error=bin link=logit;
repeated subject=twin_id / logor=zfull zdata=zin zrow=(z1-z2);
run;

Parameter Estimate Error Z Pr > |Z|
Intercept -1.1672 0.1042 -11.21 <.0001
G A/A 0.3932 0.1836 2.14 0.0322
G A/G -0.1289 0.1377 -0.94 0.3491
G G/G 0.0000 0.0000 . .
Alpha1 1.0903 0.2369 4.60 <.0001
Alpha2 0.6560 0.4161 1.58 0.1149

• α₂ represents the additional association among MZ compared to DZ twins.
We cannot reject the null hypothesis that this additional association is zero.

Comparison with bivariate logistic regression
• Using Annica’s Stata macro to estimate the bivariate logistic regression model.
• α₁ is the log OR for MZ and α₂ the additional log OR for DZ.

ALR Bivariate logistic
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>SE</th>
<th>Estimate</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1672</td>
<td>0.1042</td>
<td>-1.1672</td>
<td>0.1047</td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3932</td>
<td>0.1836</td>
<td>0.3932</td>
<td>0.1851</td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1289</td>
<td>0.1377</td>
<td>-0.1289</td>
<td>0.1373</td>
</tr>
<tr>
<td>G G/G</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Alpha1</td>
<td>1.7464</td>
<td>0.3425</td>
<td>1.7464</td>
<td>0.3338</td>
</tr>
<tr>
<td>Alpha2</td>
<td>-0.6560</td>
<td>0.4161</td>
<td>-0.6560</td>
<td>0.4124</td>
</tr>
</tbody>
</table>

Comparison of the estimated regression parameters and standard errors from the various models

<table>
<thead>
<tr>
<th>Beta</th>
<th>Naive</th>
<th>GEE(1)</th>
<th>ALR(1)</th>
<th>ALR(2)</th>
<th>BLR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.1609</td>
<td>-1.1605</td>
<td>-1.1605</td>
<td>-1.1672</td>
<td>-1.1672</td>
</tr>
<tr>
<td>G A/A</td>
<td>0.3921</td>
<td>0.3995</td>
<td>0.3983</td>
<td>0.3932</td>
<td>0.3932</td>
</tr>
<tr>
<td>G A/G</td>
<td>-0.1313</td>
<td>-0.1325</td>
<td>-0.1323</td>
<td>-0.1289</td>
<td>-0.1289</td>
</tr>
<tr>
<td>Alpha1</td>
<td>1.7464</td>
<td>0.3425</td>
<td>1.7464</td>
<td>0.3338</td>
<td></td>
</tr>
<tr>
<td>Alpha2</td>
<td>-0.6560</td>
<td>0.4161</td>
<td>-0.6560</td>
<td>0.4124</td>
<td></td>
</tr>
</tbody>
</table>

Estimation using mixed models
\[
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
\sim
\begin{pmatrix}
c_1 \\
c_2
\end{pmatrix}
+ \begin{pmatrix}
g_1 \\
g_2
\end{pmatrix} \beta
\]
where \(c_1\) and \(c_2\) are the shared environment random effects and \(g_1\) and \(g_2\) are the genetic random effects.
\[
\begin{pmatrix}
c_1 \\
c_2
\end{pmatrix}
\sim
N\left(\begin{pmatrix}
0 \\
0
\end{pmatrix}, \sigma^2_c \begin{pmatrix}
1 & 1 \\
1 & 1
\end{pmatrix}\right)
\]
for both MZ and DZ twins.

Heritability
\[
h^2 = \frac{\sigma^2_g}{\sigma^2_c + \sigma^2_g + 1}
\]

References