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-HTR2A 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.

- We are primarily interested in modelling the association between the outcome and the specific candidate gene while controlling for the fact that the individuals in the study cannot be considered independent.
- That is, the within-twin association is considered a nuisance.
- Inference surrounding the within-twin association may, however, be of specific interest. For example, a stronger association for MZ than DZ twins suggests a genetic component to the disease etiology.
- If the difference in the within-twin association between MZ and DZ twins disappears (or becomes considerably smaller) after controlling for a genetic variant in the mean model then this suggests that the genetic variant is a meaningful predictor.

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.

Approach	Inference	
Mixed models	Conditional	Parametric
Bivariate logistic	Marginal	Parametric
GEE	Marginal	Semiparametric
Alternating logistic	Marginal	Semiparametric

1. Mixed models.
   - Conditional (subject-specific) rather than marginal (population-averaged) inference.
   - We will discuss approaches using mixed models in a separate seminar.
   - John Nelder is fitting hierarchical generalized linear models.
2. Bivariate logistic regression.
   - Full likelihood, but can be difficult to implement in practice using available software.
3. Generalised Estimating Equations (GEE).
   - Easy to implement but not easy (in practice) to allow different correlations for MZ and DZ.
4. Alternating logistic regression.
   - A special case of GEE1, 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

```
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;
```

Parameter		Estimate	Standard Error	Z	Pr > |Z|
Intercept		-1.1605	0.1042	-11.14	<.0001
G	A/A	0.3995	0.1841	2.17	0.0300
G	A/G	-0.1325	0.1377	-0.96	0.3357
G	G/G	0.0000	0.0000	.	.

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

- 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.

ALR assuming the same association for MZ and DZ

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

Parameter		Estimate	Standard Error	Z	Pr > |Z|	OR
Intercept		-1.1605	0.1041	-11.14	<.0001	
G	A/A	0.3983	0.1840	2.16	0.0304	
G	A/G	-0.1323	0.1376	-0.96	0.3363	
G	G/G	0.0000	0.0000	.	.	
Alpha1		1.3198	0.1944	6.79	<.0001	3.74

- `Alpha1` is the estimated log OR.

- The log OR is 1.3198 implying that the OR is $\exp(1.3198) = 3.74$.

$$OR = 3.74 = \frac{\text{odds(twin 2 depressed} \mid \text{twin 1 depressed)}}{\text{odds(twin 2 depressed} \mid \text{twin 1 not depressed)}}$$

- 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).

ALR allowing separate association for MZ and DZ

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

Parameter		Estimate	Standard Error	Z	Pr > |Z|	OR
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		1.7464	0.3425	5.10	<.0001	5.7
Alpha2		1.0903	0.2369	4.60	<.0001	3.0

- `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.

- 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, α_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.

```
proc sort data=mj.depress(keep=twin_id dz) out=zin nodupkey;
by twin_id;
run;

data zin;
set zin;
z1=1;
z2=1-dz;
run;
```

```
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	Standard 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

- α_2 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.
- α_1 is the log OR for MZ and α_2 the additional log OR for DZ.

		ALR		Bivariate logistic	
		Estimate	SE	Estimate	SE
Intercept		-1.1672	0.1042	-1.1672	0.1047
G	A/A	0.3932	0.1836	0.3932	0.1851
G	A/G	-0.1289	0.1377	-0.1289	0.1373
G	G/G	0.0000	0.0000	0.0000	0.0000
Alpha1		1.7464	0.3425	1.7464	0.3335
Alpha2		-0.6560	0.4161	-0.6560	0.4124

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

Beta		Naive	GEE(1)	ALR(1)	ALR(2)	BLR(2)
Intercept		-1.1609	-1.1605	-1.1605	-1.1672	-1.1672
G	A/A	0.3921	0.3995	0.3983	0.3932	0.3932
G	A/G	-0.1313	-0.1325	-0.1323	-0.1289	-0.1289
SE(beta)						
Intercept		0.0969	0.1042	0.1041	0.1042	0.1047
G	A/A	0.1740	0.1841	0.1840	0.1836	0.1851
G	A/G	0.1304	0.1377	0.1376	0.1377	0.1373

Naive: Logistic regression assuming independence
ALR: Alternating logistic regression
BLR: Bivariate logistic regression

Numbers in parentheses indicate the number of parameters used to model the within-twin association.

Estimation using mixed models

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} + \begin{pmatrix} x_1' \\ x_2' \end{pmatrix} \beta$$

where c_1 and c_2 are the shared environment random effects and g_1 and g_2 are the genetic random effects.

$$\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_c^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right)$$ for both MZ and DZ twins.

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_g^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right)$$ for MZ twins.

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_g^2 \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$$ for DZ twins.

$$\text{Heritability } h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_c^2 + 1}$$

References

[1] Jansson M, Gatz M, Berg S, Johansson B, Malmberg B, McClearn GE, *et al.*. Association between depressed mood in the elderly and a 5-HTR2A gene variant. *American Journal of Medical Genetics Part B Neuropsychiatric Genetics* 2003;**00**:00–00. (in press).

[2] Palmgren J. Approaches to modelling bivariate binary responses: An empirical adventure. In: *A Spectrum of Statistical Thought*, 1991, 1991; 201–212.

[3] Glynn RJ, Rosner B. Comparison of alternative regression models for paired binary data. *Statistics in Medicine* 1994;**13**:1023–36.

[4] Iachina M, Jorgensen B, Christensen K, Iachine I. Analysis of functional abilities for elderly Danish twins using GEE models. *Twin Res* 2002;**5**:289–93.

[5] Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. *Biometrika* 1993;**80**:517–526.