Analysing correlated discrete data

Christophe Lalanne

Outline

"Far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise." —John W Tukey (1915–2000)

An example
Another example
What are the statistical challenges?
What are the solutions?
The GEE approach
Illustrated code for GEE analysis
How about a GLMM approach?
More on GLMM
Bibliography

An example

A longitudinal study of the health effects of air pollution. (Ware et al., 1984; Fitzmaurice and Laird, 1993)

			Maternal smoking	
			Age 10	
Age 7	Age 8	Age 9	No	Yes
No	No	No	237/10	118/6
		Yes	15/4	8/2
	Yes	No	16/2	11/1
		Yes	7/3	6/4
Yes	No	No	24/3	7/3
		Yes	3/2	3/1
	Yes	No	6/2	4/2
		Yes	5/11	4/7

Table of counts of *wheezing status* (Yes/No) for $N = 537$ children aged 7 to 10 years, with following two factors: maternal smoking (`smoke`, considered fixed) and time (`age`).

What is the marginal expectation of children's response as a function of these covariates (including the interaction `smoke:time`)?

Another example

Responses to a questionaire on verbal aggression. (De Boeck and Wilson, 2004)

			Response		
Situation	Mode	Behavior	No	Perhaps	Yes
other	want	curse	158	207	267
		scold	244	179	209
		shout	312	183	137
	do	curse	200	205	227
		scold	298	189	145
		shout	446	121	65
self	want	curse	226	247	159
		scold	377	178	77
		shout	457	127	48
	do	curse	289	225	118
		scold	420	152	60
		shout	546	68	18

Table of responses for $N = 316$ subjects asked to respond on three-point Likert items describing aggressive verbal response depending on `situ`, `mode`, and `btype`.

Are the individual responses influenced by one of the three factors, or a combination thereof?

What are the statistical challenges?

- The response is not continuous (e.g., score); we need to resort on a Binomial or Multinomial distribution.
- Individual responses are correlated; this breaks standard independence assumption at the levels of the residuals.
- The choice of the modeling strategy will affect the conclusions that can be drawn from the study.

What are the solutions?

- Don't bother with the complications — this might work. . . sometimes.
- Use Generalized Estimating Equations (GEE) to estimate population-averaged effects, by assuming a working correlation matrix to account for within-unit correlation. (Liang and Zeger, 1986; Hanley et al., 2003)
- Use Generalized Linear Mixed Models (GLMM) to estimate subject-specific regression parameters, for some fixed and/or random effects, possibly with different correlation structure. (McCulloch et al., 2008; Molenberghs and Verbeke, 2005)

The GEE approach

Let's assume a given working correlation matrix, e.g., one of Independence (`"ind"`), Exchangeable (`"exch"`), Unstructured (`"uns"`), Autoregressive (`"ar1"`). This will be our variance model.

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,t} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,t} & \rho_{2,t} & \cdots & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & \rho & \cdots & \rho^{t-1} \\ \rho & 1 & \cdots & \rho^{t-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{t-1} & \rho^{t-2} & \cdots & 1 \end{pmatrix}$$

How to choose one?

- Unstructured: few number of units per cluster, balanced complete design;
- Exchangeable: no logical ordering for observations within a cluster;
- Autoregressive (or auto-regressive): to account for time-varying response;
- Independent: when the number of clusters is small (Diggle et al., 1994).

How to check whether it is the correct specification?

Sensitivity to mispecification will be reflected in standard error of parameter estimates. Variance estimators can be model-based (useful for small number of clusters, otherwise use a Jackknife estimator) or empirical-based (so-called sandwich estimator, asymptotically unbiased).

Importantly, a GEE model will give valid results even with a misspecified correlation structure provided the sandwich variance estimator is used.

We'll be fitting a model of the form

$$\text{logit}(\mu) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{smoke} + \beta_3 \text{age} \times \text{smoke}$$

This is is the mean model, which describes how the mean relates to the factors of interest.

In addition, we need to specify a variance function. In this case, we choose $\mathbb{V}(\mu) = \phi\mu \cdot (1 - \mu)$ (and here, let's take a scale parameter $\phi = 1$).

Note that we make no assumption for the distribution of observations.

More maths background can be found on this website:

http://gbi.agrsci.dk/statistics/courses/phd07/material/Day10/

Illustrated code for GEE analysis

ohio.R

We start by fitting a basic model where the working correlation is symmetric with correlation ρ (recall the *compound symmetry* hypothesis in ANOVA with repeated measurements).

Import what we need

```
library(geepack)
data(ohio)
```

Model specification

```
fm <- resp ~ age*smoke
```

Fit a GEE with exchangeable correlation structure

```
gee.fit <- geese(fm, id=id, data=ohio, family=binomial,
                 corstr="exch", scale.fix=TRUE)
```

Get model parameters

```
summary(gee.fit)
```

The regression coefficients indicate the effect of each predictor:

```
 Coefficients:
              estimate     san.se        wald          p
(Intercept) -1.90049529 0.11908698 254.6859841 0.00000000
age         -0.14123592 0.05820089   5.8888576 0.01523698
smoke        0.31382583 0.18575838   2.8541747 0.09113700
age:smoke    0.07083184 0.08852946   0.6401495 0.42365667
```

The within-cluster correlation is estimated at $\rho = 0.355$:

```
 Estimated Correlation Parameters:
      estimate     san.se     wald p
alpha 0.354531 0.03582698 97.92378 0
```

The results indicate that there is a significant decrease of wheeziness ($p < 0.001$), but no significant effect of mother's smoking status. An unstructured working correlation structure would yield quite the same results.

However, using model-based SEs (from the gee package) with an independance correlation structure would underestimate parameters variance of time-stationary effects:

```
 Coefficients:
            Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)  -1.9008     0.0887  -21.42      0.1191  -15.963
age          -0.1413     0.0695   -2.03      0.0582   -2.426
smoke         0.3140     0.1394    2.25      0.1878    1.671
age:smoke     0.0708     0.1107    0.64      0.0883    0.802
```

(Compare to Table 2 in Fitzmaurice and Laird, 1993.)

Note on computing Wald statistics manually

The robust and naive variance-covariance matrices for parameter estimates are stored in `gee.fit$vbeta` and `gee.fit$vbeta.naiv`, respectively. They are close one to the other, with exact same values up to the third figure, as can be seen with e.g., `summary(as.vector(gee.fit$vbeta-gee.fit$vbeta.naiv))`:

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.000278 -0.000256 -0.000110 -0.000051  0.000145  0.000293
```

The Wald z-test is computed as $\widehat{\beta}_j/\mathrm{SE}(\widehat{\beta}_j)$, and we can check that using $\mathrm{SE}(\widehat{\beta}_j)$ from the VC matrix would yield identical results, using the following:

`(gee.fit$beta/sqrt(diag(gee.fit$vbeta)))^2`.

Wald z-statistics are distributed as $\chi^2(1)$.

The `geepack` package also provides an `anova()` to compare nested models (Halekoh and Højsgaard, 2006).

What is the adjusted OR for age? What about the odds of a positive wheezing status at age 8 vs. 10 when mother smoke or not?

Remove the interaction term

```
gee.fit2 <- geeglm(update(fm, . ~ . -age:smoke), id=id,
                   data=ohio, family=binomial,
                   corstr="exch", scale.fix=TRUE)
```

Adjusted odds-ratio for age

```
exp(coef(gee.fit2)["age"])
```

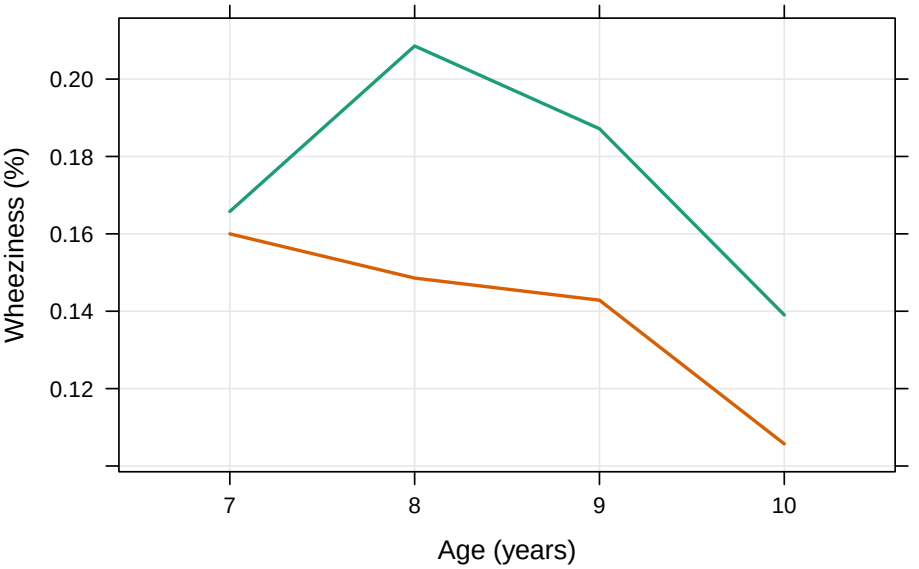
Refit a model where Age is treated as factor

```
gee.fit3 <- geeglm(resp ~ as.factor(age) + smoke, id=id,
                   data=ohio, family=binomial,
                   corstr="exch", scale.fix=TRUE)
```

Set up the corresponding contrast

```
if (require(doBy)) esticon(gee.fit3, c(0, -1, 0, 1, 1))
exp(.Last.value$Estimate)
```

Figure 1 ***Plot of the marginal distribution over years.***



How about a GLMM approach?

The basic syntax in R reads:

Import what we need	`library(lme4)`
Fit a basic GLMM	`fit.glmm <- lmer(resp ~ age+smoke+(1|id), data=ohio, family=binomial)`
Display fixed effects	`fixef(fit.glmm)`
Plot the distribution of random effects	`plot(ranef(fit.glmm))`

The results show again a significant effect of age:

```
Random effects:
 Groups Name        Variance Std.Dev.
 id     (Intercept) 5.49     2.34
Number of obs: 2148, groups: id, 537

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3740     0.1871  -18.03   <2e-16 ***
age          -0.1768     0.0699   -2.53    0.012 *
smoke         0.4147     0.2960    1.40    0.161
```

This could be confirmed with an LRT, like `anova(fit.glmm, update(fit.glmm, . ~ . - age))` which indicates a significant $\chi^2(1) = 6.86$ with $p = 0.009$.

What's the difference then?

Here, we are using a conditional approach, hence the need to specify a distribution for the random effects (here, only the intercept term). The model now looks like:

$$\text{logit}(\mu \mid \nu_i) = X\beta + \nu_i$$

where $\nu_i \sim N(0, \sigma_\nu^2)$ (random effects have zero mean on the logit scale). In other words, instead of modeling the population averaged log odds, the above random effects model will allow to model μ accounting for subject-specific variations.

The GEE and GLMM are only equivalent in the case of an identity link function (linear regression). Only the interpretation (and the appropriateness) of model coefficients change when using other link function. (Hubbard et al., 2010)

More on GLMM

In contrast to epidemiological cohort studies, with cross-sectional data or repeated measures collected on the same individual, like responses to items in a questionnaire (Case study 2), we are often more interested in working at the subject level.

Bibliography

1 Ware, J., Dockery, D., Spiro, A., Speizer, F. and Ferris, B. (1984). Passive smoking, gas cooking and respiratory health in children living in six cities. *American Review of Respiratory Diseases, 129*, 366–374. PMID: 6703495.

2 Fitzmaurice, G. and Laird, N. (1993). A likelihood-based method for analysing longitudinal binary response. *Biometrika, 80*(1), 141–151.

3 De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer.

4 Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22.

5 Hanley, J., Negassa, A., Edwardes, M. and Forrester, J. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology, 157*(4), 364–375. PMID: 12578807.

6 McCulloch, C., Searle, S. and Neuhaus, J. (2008). *Generalized, Linear, and Mixed Models*. 2nd edition Wiley Interscience.

7 Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer.

8 Diggle, P., Liang, K.-Y. and Zeger, S. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford Science.

9 Halekoh, U. and Højsgaard, S. (2006). The R package geepack for generalized estimating equations. *Journal of Statistical Software, 2*. Online version.

10 Hubbard, A., Ahern, J., Fleischer, N., Van der Laan, M. and Lippman, S. et al. (2010). To gee or not to gee: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology, 21*(4), 467–474. PMID: 20220526.

Made with ConTEXt version 2011.05.18 18:04, R version 2.13.1 (2011-07-08), gee_tutor.tex fa3ce81 on 2011/09/12