Models for Correlated, Clustered Responses

Generalized Estimating Equations

Demetris Athienitis

Introduction

Correlated responses occur in several ways, including:

- Repeated measures/longitudinal studies: repeated observations on each subject
- Multiple, matched sets of subjects
  - Children in the same family
  - Children in the same elementary school class (children within class, class within school, school within district, etc)
  - Fetuses from the same litter

Usual model forms apply (e.g., logistic regression for binary response, cumulative logit for ordinal response), but model fitting must account for dependence (e.g., from repeated measures on subjects) in order to get appropriate standard errors and valid inferences.

Introduction

We will use two approaches to such data: Observations $(Y_1, Y_2, \ldots, Y_T)$

- (In this chapter) Generalized Estimating Equations (GEE) to simultaneously fit marginal models on each (marginal) $E(Y_t), t = 0, \ldots T$
- (In the next chapter) Generalized Linear Mixed Models (GLMM) to find random effect for the subject/block effect

GEE

Focusing on GEE for Repeated Measures:

- Specify model in usual way by deciding what the random, component, link function and systematic components are
- Select a *working correlation* matrix for best guess about correlation pattern between pairs of observations, within-cluster correlation

Example

For T repeated responses, *exchangeable* correlation matrix is

Time	1	2	$\cdots$	T
1	1	ρ	$\cdots$	ρ
2	ρ	1	$\cdots$	ρ
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T	ρ	ρ	$\cdots$	1

GEE

When there is positive within-cluster correlation (as often is the case):

- The standard errors for *between-cluster* effects (such as different treatment groups) and standard errors of estimated means within clusters tends to be larger than when independent
- The standard errors for *within-cluster* effects, such as a slope for a trend in the repeated measurements in a subject, tend to be smaller than when observations are independent

Fitting method gives estimates that are consistent even if correlation structure is miss-specified. Adjusts standard errors to reflect actual observed dependence. Therefore, overly complicated structures are not encouraged. For other structures the reader is encouraged to review the literature.

Example (Crossover Study: Drug vs Placebo continued)

		Placebo S	Placebo F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

Fit the model

$$\text{logit}\left[P(Y_t = 1)\right] = \alpha + \beta d, \qquad d = \begin{cases} 1 & \text{drug} \\ 0 & \text{placebo} \end{cases}$$

where $t = 1, 2$ represents the two time points, the two observations on each subject.

Example (continued)

But data in correct "long" format.

```
> head(crossm1)
  Subject   Treat Resp
1       1    Drug    1
2       1 Placebo    1
3       2    Drug    1
4       2 Placebo    1
5       3    Drug    1
6       3 Placebo    1
```

```
> tail(crossm1)
    Subject   Treat Resp
167      84    Drug    0
168      84 Placebo    0
169      85    Drug    0
170      85 Placebo    0
171      86    Drug    0
172      86 Placebo    0
```

Example (continued)

```
> cross.gee1=gee(Resp ~ Treat, id=Subject, data=crossm1,
+ family=binomial, corstr="exchangeable")
> summary(cross.gee1)
Coefficients:
             Estimate  Robust S.E.  Robust z
(Intercept) -1.067841   0.2471428  -4.320744
TreatDrug    1.959839   0.3772338   5.195289

Working Correlation
           [,1]       [,2]
[1,]  1.0000000 -0.2140746
[2,] -0.2140746  1.0000000
```

Odds of Success with drug is estimated to be $e^{1.96} = 7.1$ times odds with placebo. The 95% CI for odds ratio (for marginal probabilities) is

$$e^{1.96 \mp (1.96)(0.377)} \rightarrow (e^{1.22}, e^{2.70}) = (3.4, 14.9)$$

GEE: Remarks

- With $\hat{\rho} \approx 0$ it implies that there is no significant correlation between the "clustered" responses
- With cross-over designs it is important to allow enough time for the effects of the previous treatment not influence the results of the next treatment the unit will cross-over to
- With GEE approach, can also have "between-subject" explanatory variables. In the Drug vs Placebo, d was a variable monitored "within-subject" but we could have monitored "between-subject" gender and even order of treatment, e.g.

$$\text{sequence} = \begin{cases} 1 & \text{placebo then drug} \\ 2 & \text{drug then placebo} \end{cases}$$

GEE

GEE is a known as *quasi-likelihood* method.

- No particular form assumed for joint distribution of $(Y_1, Y_2, \ldots, Y_T)$
- Hence, no likelihood function, no LR inference (LR test, LR CI)
- For responses $(Y_1, Y_2, \ldots, Y_T)$ at T times, we consider *marginal model* that describes each Y_t in terms of explanatory variables

Example (Depression)

Response on mental depression (normal, abnormal) measured three times (after 1, 2, and 4 weeks of treatment) with two drug treatments (standard, new) and two severity of initial diagnosis groups (mild, severe).
Is the rate of improvement better with the new drug?

	Time	Response Pattern							
	0	A	A	A	A	N	N	N	N
	1	A	A	N	N	A	A	N	N
	2	A	N	A	N	A	N	A	N
Severity	Drug								
Mild	Std	6	15	4	14	3	9	13	16
	New	0	9	2	22	0	6	0	31
Severe	Std	28	27	15	9	9	8	2	2
	New	6	32	5	31	2	5	2	7

Example (continued)

- Y_t = response of randomly selected subject at time t (1 = normal, 0 = abnormal)
- s = severity of initial diagnosis (1 = severe, 0 = mild)
- d = drug (1 = new, 0 = std)
- t = time (0, 1, 2), which is log2(weeks of trt)

$$\log\left[\frac{P(Y_t=1)}{P(Y_t=0)}\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (dt)$$

so that

$$\log\left[\frac{P(Y_t=1)}{P(Y_t=0)}\right] = \begin{cases} \alpha + \beta_1 s + \beta_3 t & \text{if } d = 0 \text{(standard drug)} \\ \alpha + \beta_2 + \beta_1 s + (\beta_3 + \beta_4) t & \text{if } d = 1 \text{(new drug)} \end{cases}$$

Example (continued)

```
> dep.gee1=gee((response == "normal") ~ severity + drug*time,
+ id=subject, data=depression, family=binomial,
+ corstr="exchangeable")
> summary(dep.gee1)
Coefficients:
                  Estimate  Robust S.E.   Robust z
(Intercept)    -0.02809866   0.1741791  -0.1613205
severitysevere -1.31391033   0.1459630  -9.0016667
drug1          -0.05926689   0.2285569  -0.2593091
time            0.48246420   0.1199383   4.0226037
drug1:time      1.01719312   0.1877014   5.4192084

Working Correlation
              [,1]         [,2]         [,3]
[1,]  1.000000000 -0.003432729 -0.003432729
[2,] -0.003432729  1.000000000 -0.003432729
[3,] -0.003432729 -0.003432729  1.000000000
```

Example (continued)

- β_4 significant, strong evidence of faster improvement for new drug
- When initial diagnosis is severe, estimated odds of normal response are $e^{-1.31} = 0.27$ times estimated odds when initial diagnosis is mild, at each $d \times t$ combination
- $\hat{\beta}_2 = -0.06$ is drug effect only at $t = 0$. $e^{-0.06} = 0.94 \approx 1$, so essentially no drug effect at $t = 0$ (after 1 week). However, drug effect at end of study ($t = 2$) estimated to be $e^{\hat{\beta}_2 + 2\hat{\beta}_4} = 7.2$
- Estimated time effects are:
  - standard drug ($d = 0$): $\hat{\beta}_3 = 0.48$
  - new drug ($d = 1$): $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$
- Examined $s \times d$ and $s \times t$ interactions, but not statistically significant
- Started with exchangeable working correlation, but estimated $\rho \approx 0$

Missing Data

Remark

Missing data is not uncommon and can be very problematic unless missing completely at random (MCAR): missingness unrelated to response or any explanatory variables.

Missing at random (MAR) means missingness unrelated to response after controlling for explanatory variables. Methods exist to handle this and some other forms of missingness. Common solution involves the method of multiple imputations.

We learned

- GEE as a marginal (not case specific) model for correlated, clustered data
- Utilizes standard components
- Incorporates working correlation matrix that is robust to miss-specification