Estimation of α's

- (For now) these are not of primary interest — nuisance
- Need to be estimated (at each iteration) in order to get $\widehat{\beta}$
- Moment based estimators can be derived, but needs to be done/implemented for each correlation structure
- Typically involves calculation from the Pearson residuals (for fixed $\widehat{\beta}$ [and possibly $\widehat{\phi}$]):

$$r_{ij} = \frac{Y_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\phi} var(\widehat{\mu}_{ij})}} = \frac{Y_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\phi} a_{ij}^{-1} v(\widehat{\mu}_{ij})}}$$

Estimation of $\widehat{\alpha}$ (1)

Independence

$$\rho(\alpha) = 0; \quad \mathbf{R}_i(\boldsymbol{\alpha}) = \mathbf{I}$$

so $\alpha = 0$

Exchangeable

$$\hat{\alpha} = \sum_{i=1}^{m} \sum_{j>j'} \hat{r}_{ij} \hat{r}_{ij'} / \{ \sum_{i=1}^{m} \frac{1}{2} n_i (n_i - 1) - p \}$$

$=$ average of product of all pairs of residuals.

Estimation of $\widehat{\alpha}$ (2)

AR(1)

$$E(\hat{r}_{ij}\hat{r}_{ij'}) \approx \alpha^{|t_{ij}-t_{ij'}|}$$

so we fit the regression model

$$\ln(\hat{r}_{ij}\hat{r}_{ij'}) = |t_{ij} - t_{ij'}| \ln \alpha.$$

Alternatively, some software set:

$$\widehat{\alpha} = \sum_{i=1}^{m} \sum_{j=1}^{n_i-1} \hat{r}_{ij}\hat{r}_{i,j+1} / \{\sum_{i=1}^{m}(n_i - 1) - p\}$$

$=$ average product of pairs of residuals that are next to each other

Estimation of $\widehat{\alpha}$ (3)

Unstructured

$$\widehat{\alpha}_{j,j'} = \sum_{i=1}^{m} \hat{r}_{ij}\hat{r}_{ij'}/(m-p)$$

General Remarks

- All of these are usually transparent to us: we are usually at the mercy of software.
- For more complicated situations, need to implement your own reasonable estimates.

Estimation of ϕ

If necessary:

$$\widehat{\phi} = \left(\sum_{i=1}^{m} n_i - p \right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} r_{ij}^2$$

which is the average of all squared residuals — the marginal variance

Efficiency Loss from Mis-specification of **R**

Table 1. *Asymptotic relative efficiency of $\hat{\beta}_I$ and $\hat{\beta}_G$ to generalized estimator with correlation matrix correctly specified for $\eta_{it} = \beta_0 + \beta_1 t/10$. Here, $\beta_0 = \beta_1 = 1$, $n_i = 10$. For upper entry $\alpha = 0{\cdot}3$; lower entry $\alpha = 0{\cdot}7$*

	Working R			
True R	Independence	1-dependence	Exchangeable	AR-1
1-Dependence	0·97	1·0	0·97	0·99
	0·74	1·0	0·74	0·81
Exchangeable	0·99	0·95	1·0	0·95
	0·99	0·23	1·0	0·72
AR-1	0·97	0·99	0·97	1·0
	0·88	0·75	0·88	1·0

Not much loss if α is small or if correlation is close to correct.

Efficiency Loss of GEE vs. MLE

Table 2. *Asymptotic relative efficiency of* $\hat{\beta}_I$ *and* $\hat{\beta}_G$ *assuming* AR 1 *correlation structure to the maximum likelihood estimate for first-order Markov chain with* $\theta_{it} = \beta_0 + \beta_1 x_i, x_i = 0$ *for Group* 0, $x_i = 1$ *for Group* 1. *Here* $\beta_0 = 0, \beta = 1$, *and for upper entry* $n_i = 10$, *lower entry* $n_i = 1, \ldots, 8$ *with equal probabilities*

	Correlation, α						
	0·0	0·1	0·2	0·3	0·5	0·7	0·9
$\hat{\beta}_I$	1·0	1·0	0·99	0·97	0·94	0·91	0·92
	1·0	1·0	0·98	0·96	0·92	0·86	0·81
$\hat{\beta}_G$ (AR 1)	1·0	1·0	0·99	0·99	0·98	0·97	0·98
	1·0	1·0	0·99	0·99	0·98	0·98	0·99

Reference is MLE, and AR(1) is close to Markov model, so not much loss here

What **R**? Other Experts Say...

- Liang and Zeger (1986): *little difference when correlation is moderate*
- McDonald (1993): Independence *may be recommended for practical purposes*
- Zhao, Prentice, Self (1992): assuming independence *can lead to important losses of efficiency*
- Fitzmaurice, Laird, Rotnitzky (1993): *important to obtain close approximation to* $Cov(\mathbf{Y}_i)$ *in order to achieve high efficiency*
- Mancl and Leroux (1996): *depends on covariate distribution, the cluster sizes, the response variable correlation, and the regression parameters... sensitive to between and within cluster variation of the covariates... efficiency losses for simple working correlation [...] can be large even for small to moderate correlation and cluster sizes*

Ah So.

Some options for what we can do:

- Utilize our knowledge on covariate distribution, correlation, regression parameters, between vs. within cluster variation of covariates?
- Pick a complicated correlation?
- Just fit a bunch of models and pick the best one?

None of these are very good solutions. Meh.

Example: Indonesian Infectious Disease Data

- 275 Indonesian children (1200 observations), each was followed for up to 6 consecutive quarters
- Outcome=respiratory infection (Y/N)
- Covariates=age, sex, xerophthalmia status, season, height
- Marginal logistic model:

$$logit(\mu_{ij}) = \mathbf{X}_{ij}^T \beta$$

- Initial working correlation matrix: exchangeable.

Example: Indonesian Infectious Disease Data - Input

```
> indon = read.table("indon1.dat", col.names =
    c("id", "season", "xero", "age", "sex", "height", "infect"))
> head(indon)
      id season xero age sex height infect
1 121013     -1    0  31   0     -3      0
2 121013      0    0  34   0     -3      0
3 121013      1    0  37   0     -2      0
4 121013      0    0  40   0     -2      0
5 121013     -1    0  43   0     -2      1
6 121013      0    0  46   0     -3      0
> library(gee)
> mod = gee(infect ~ season + xero + age + sex + height,
        id = as.factor(id), corstr = "exchangeable",
        family= binomial, data = indon)
> summary(mod)
```

Output (1)

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Call:
gee(formula = infect ~ season + xero + age + sex + height, id = id,
    data = indon, family = binomial, corstr = "exchangeable")

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-0.32924060 -0.11169648 -0.06702687 -0.03458892  0.98547600
```

Output (2)

```
Coefficients:
               Estimate  Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept) -2.35487490 0.162009071 -14.535451 0.163475902 -14.405028
season      -0.54151732 0.161194240  -3.359409 0.160329485  -3.377528
xero         0.61256297 0.458460732   1.336130 0.434898457   1.408520
age         -0.03126073 0.006835855  -4.573054 0.006274497  -4.982190
sex         -0.42197929 0.241454930  -1.747652 0.236398175  -1.785036
height      -0.05069619 0.021514270  -2.356398 0.024310463  -2.085365

Estimated Scale Parameter:  1.03001
Number of Iterations:  3
```

Output (3)

```
Working Correlation
           [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
[1,] 1.00000000 0.04466212 0.04466212 0.04466212 0.04466212 0.04466212
[2,] 0.04466212 1.00000000 0.04466212 0.04466212 0.04466212 0.04466212
[3,] 0.04466212 0.04466212 1.00000000 0.04466212 0.04466212 0.04466212
[4,] 0.04466212 0.04466212 0.04466212 1.00000000 0.04466212 0.04466212
[5,] 0.04466212 0.04466212 0.04466212 0.04466212 1.00000000 0.04466212
[6,] 0.04466212 0.04466212 0.04466212 0.04466212 0.04466212 1.00000000
```

Potential Models

Model 1 Logistic Regression: Ignore correlation

Model 2 GEE: Independence

Model 3 GEE: Exchangeable

Model 4 GEE: AR1

Model 5 GEE: Unstructured

```
> form = formula(infect ~ season + xero + age + sex + height)
> mod1 = glm(form, family = "binomial", data = indon)
> mod2 = gee(form, family = "binomial", id = id, data = indon, corstr = "independence")
> mod3 = gee(form, family = "binomial", id = id, data = indon, corstr = "exchangeable")
> mod4 = gee(form, family = "binomial", id = id, data = indon, corstr = "AR-M", Mv= 1) # fails!
> mod5 = gee(form, family = "binomial", id = id, data = indon, corstr = "unstructured")
```

Working Correlations (1)

```
> summary(mod2)$working  #independence$
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    0    0    0    0    0
[2,]    0    1    0    0    0    0
[3,]    0    0    1    0    0    0
[4,]    0    0    0    1    0    0
[5,]    0    0    0    0    1    0
[6,]    0    0    0    0    0    1
>
> round(summary(mod3)$working,3) # exchangeable$
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]
[1,] 1.000 0.045 0.045 0.045 0.045 0.045
[2,] 0.045 1.000 0.045 0.045 0.045 0.045
[3,] 0.045 0.045 1.000 0.045 0.045 0.045
[4,] 0.045 0.045 0.045 1.000 0.045 0.045
[5,] 0.045 0.045 0.045 0.045 1.000 0.045
[6,] 0.045 0.045 0.045 0.045 0.045 1.000
```

Working Correlations (2)

```
> round(summary(mod5)$working,3) # unstructured$
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,]  1.000  0.043  0.066  0.063 -0.004  0.003
[2,]  0.043  1.000  0.096  0.061  0.043 -0.030
[3,]  0.066  0.096  1.000  0.074  0.112 -0.046
[4,]  0.063  0.061  0.074  1.000 -0.031 -0.002
[5,] -0.004  0.043  0.112 -0.031  1.000  0.000
[6,]  0.003 -0.030 -0.046 -0.002  0.000  1.000
```

Coefficients and SEs

	GLM (SE)	Indep (SE)	Exch. (SE)	Unstr. (SE)
(Intercept)	-2.377 (0.15)	-2.377 (0.162)	-2.355 (0.163)	-2.349 (0.161)
season	-0.55 (0.161)	-0.55 (0.16)	-0.542 (0.16)	-0.519 (0.158)
xero	0.718 (0.435)	0.718 (0.42)	0.613 (0.435)	0.613 (0.424)
age	-0.032 (0.006)	-0.032 (0.006)	-0.031 (0.006)	-0.032 (0.006)
sex	-0.395 (0.22)	-0.395 (0.236)	-0.422 (0.236)	-0.386 (0.236)
height	-0.048 (0.02)	-0.048 (0.024)	-0.051 (0.024)	-0.053 (0.025)

Remarks

- Independence, Exchangeable correlation make sense. What about Unstructured?
- Another software package that is popular is geepack library:

```
library(geepack)
mod3_1 = geeglm(form, id = id, data = indon, family = binomial, corstr = "exch")
```

	gee	geepack
(Intercept)	-2.3549 (0.16)	-2.3548 (0.16)
season	-0.5415 (0.16)	-0.5415 (0.16)
xero	0.6126 (0.43)	0.6123 (0.43)
age	-0.0313 (0.01)	-0.0313 (0.01)
sex	-0.4220 (0.24)	-0.4220 (0.24)
height	-0.0507 (0.02)	-0.0507 (0.02)

gee vs. geepack

Let's try and AR1 structure, restricted to subjects with more than 1 observation:

```
subset = which(is.na(match(indon$id, names(which(table(indon$id) == 1)))))
mod_gee = gee(form, id = id, family = "binomial", corstr = "AR-M", Mv=1, data = indon, subset = subset)
mod_geepack = geeglm(form, id = id, family = "binomial", corstr = "ar1", data = indon, subset = subset)
```

Output:

	gee	geepack
(Intercept)	-2.3473 (0.161)	-2.3492 (0.161)
season	-0.5265 (0.160)	-0.5284 (0.160)
xero	0.6227 (0.441)	0.6342 (0.438)
age	-0.0304 (0.006)	-0.0304 (0.006)
sex	-0.4120 (0.239)	-0.4115 (0.239)
height	-0.0466 (0.024)	-0.0464 (0.024)

Estimates for α are 0.0633 and 0.0542.