BINARY DATA MODELS

We devote an entire chapter to binary data since such data are challenging, both in terms of modeling the dependence, and parameter interpretation.

We again consider mixed effects models approaches (with inference from likelihood or Bayes perspectives), and GEE.

Motivating Data - Indonesian Child Health

DHLZ describe a dataset in which the response is the absence/presence of respiratory illness in 275 children, with mutiple measurements being collected over time – quarterly measurements were taken for up to six consecutive quarters. Age also recorded.

Question of interest:

Is the prevalence of respiratory infection higher amongst children who suffer from xerophthalmia, an ocular manifestation of chronic vitamin A deficiency?

Figure 36 shows the infection indicator versus time for the first 49 children.
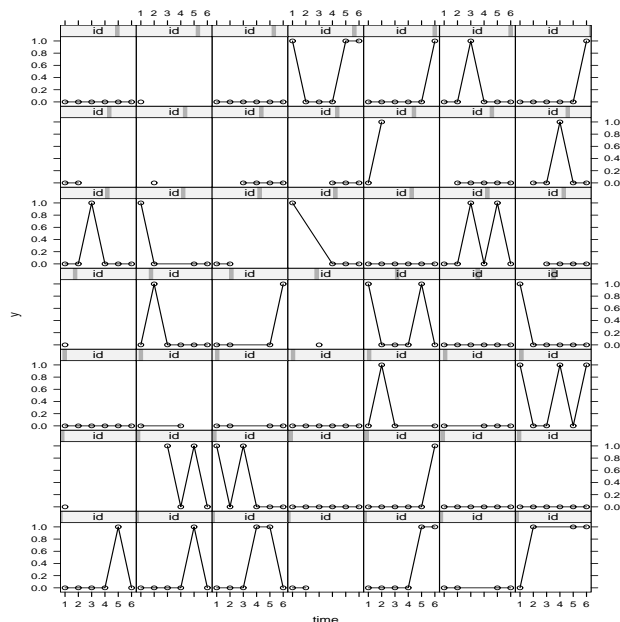
329

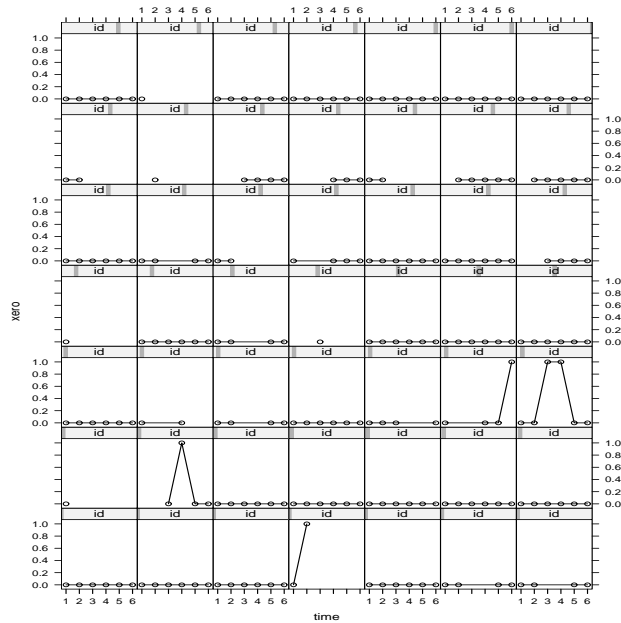Figure 36: Respiratory infection versus time for the first 49 children.

Figure 37: Xerophthalmia versus time for the first 49 children.

Exploratory Summaries

With binary responses and covariates, plots are not so informative - instead
look at tables.

```
> table(n) # no of responses per child (ie 22 children had 1),1200 in total
  1   2   3   4   5   6
 22  32  29  55  15 122
> xtabs(~xerocross+ycross) # cross-sectional comparison of respiratory
        ycross             # event (ycross) and xerophthalmia (xerocross)
xerocross   0   1          # at first time point
        0 234  31
        1   9   1
> 234/(9*31)
[1] 0.8387097
> summary(glm(cbind(ycross,1-ycross) ~ xerocross, family="binomial"))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0213     0.1911 -10.576   <2e-16 ***
xerocross    -0.1759     1.0712  -0.164     0.87
> exp(-.1759)
[1] 0.8387018
```

Now add in age (potential confounder).

```
> modglm <- glm(cbind(ycross,1-ycross) ~ agecross+xerocross, family="binomial")
> summary(modglm) # Cross-sectional analyses
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.17017    0.21785  -9.962   <2e-16 ***
agecross    -0.02395    0.01059  -2.262   0.0237 *
xerocross    0.22630    1.09367   0.207   0.8361
    Null deviance: 197.79  on 274  degrees of freedom
Residual deviance: 192.17  on 272  degrees of freedom
> modglmq <- glm(cbind(ycross,1-ycross) ~ agecross+xerocross,family=quasibinomial())
> summary(modglmq)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.17017    0.21493 -10.097   <2e-16 ***
agecross    -0.02395    0.01045  -2.293   0.0226 *
xerocross    0.22630    1.07900   0.210   0.8340
(Dispersion parameter for quasibinomial family taken to be 0.9733635)
    Null deviance: 197.79  on 274  degrees of freedom
Residual deviance: 192.17  on 272  degrees of freedom
```

Overview of Modeling Clustered Binary Data

For a single binary variable, Y, all moments are determined by $p = \mathrm{E}[Y]$.
Specifically $\mathrm{E}[Y^r] = p$ for $r \geq 1$.

Consider n observations from a single individual $\boldsymbol{Y} = (Y_1, ..., Y_n)^{\mathrm{T}}$.

The saturated model has $2^n - 1$ parameters (in contrast to n means, n
variances and $n(n-1)/2$ correlations for the saturated normal model).

There is no obvious or natural manner by which multivariate binary data can
be specified. We begin with a mixed effects formulation that induces
correlations.

Mixed Effects Models: Conjugate Model

$$Y \mid p \quad \sim \quad \text{Binomial}(n, p)$$
$$p \quad \sim \quad \text{Beta}(a, b)$$

gives

$$\Pr(Y = y) = \binom{n}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+b+n)}, \quad y = 0, 1, ..., n.$$

Marginal moments:

$$\text{E}[Y] = n\text{E}[p] = n\frac{a}{a+b}, \quad \text{var}(Y) = n\text{E}[p](1 - \text{E}[p])\frac{a+b+n}{a+b+1}$$

Limitations: likelihood is complex, extension to covariates increases complexity, and no tractable way to allow random effect slopes.

Mixed Effects Models: Logistic Model

A GLMM for binary data takes the binomial exponential family, with the canonical link being logistic.

We have

Stage 1: $Y_{ij} \mid p_{ij} \sim_{ind} \text{Binomial}(n_{ij}, p_{ij})$ with

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{z}_{ij}\boldsymbol{b}_i$$

Stage 2: $\boldsymbol{b}_i \mid \boldsymbol{D} \sim_{iid} \text{N}(\boldsymbol{0}, \boldsymbol{D})$.

Marginal moments are not available in closed form.

Likelihood Inference

Likelihood inference proceeds in the usual asymptotic fashion, following integration of the random effects.

A Bayesian analysis adds priors for the fixed effects and variance components.

Model checking is very difficult for binary data...

```
> library(lme4)
> memmod1 <- lmer(cbind(y,1-y) ~ age+xero+(1|id),family=binomial,method="PQL")
> summary(memmod1)
Random effects:
     Groups        Name     Variance     Std.Dev.
        id (Intercept)       3.7334       1.9322
Fixed effects:
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) -2.6920622  0.1795893 -14.9901 < 2.2e-16 ***
age         -0.0229009  0.0088506  -2.5875  0.009667 **
xero         0.1219975  0.5379760   0.2268  0.820602
memmod2 <- lmer(cbind(y,1-y) ~ age+xero+(1|id),family=binomial,method="Laplace")
> summary(memmod2)
Random effects:
     Groups        Name     Variance     Std.Dev.
        id (Intercept)       3.7334       1.9322
Fixed effects:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.6920622  0.1810916 -14.8658  < 2e-16 ***
age         -0.0229009  0.0089756  -2.5515  0.01073 *
xero         0.1219975  0.5380319   0.2267  0.82062
```

Likelihoods for Multivariate Binary Data

Log-Linear Model

We have $2^n - 1$ distinct probabilities, but we wish to consider formulations that allow more parsimonious descriptions as a function of covariates.

One choice is the log-linear model:

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}) \exp \left(\sum_j \theta_j^{(1)} y_j + \sum_{j_1 < j_2} \theta_{j_1 j_2}^{(2)} y_{j_1} y_{j_2} + ... + \theta_{12...n}^{(n)} y_1 ... y_n \right),$$

with $2^n - 1$ parameters

$$\boldsymbol{\theta} = (\theta_1^{(1)}, ..., \theta_n^{(1)}, \theta_{12}^{(2)}, ..., \theta_{n-1\,n}^{(2)}, ..., \theta_{12...n}^{(n)})^{\mathrm{T}},$$

and where $c(\boldsymbol{\theta})$ is the normalizing constant.

This formulation allows calculation of cell probabilities (but is less useful for describing $\Pr(\boldsymbol{Y} = \boldsymbol{y})$ as a function of $\boldsymbol{x}$).

Note that we have $2^n - 1$ parameters and we have two aims: reduce this number, and introduce a regression model.

Example: $n = 2$.

We have

$$\Pr(Y_1 = y_1, Y_2 = y_2) = c(\boldsymbol{\theta}) \exp \left(\theta_1^{(1)} y_1 + \theta_2^{(1)} y_2 + \theta_{12}^{(2)} y_1 y_2 \right),$$

where $\boldsymbol{\theta} = (\theta_1^{(1)}, \theta_2^{(1)}, \theta_{12}^{(2)})^{\mathrm{T}}$ and

$$c(\boldsymbol{\theta})^{-1} = \sum_{y_1=0}^1 \sum_{y_2=0}^1 \exp \left(\theta_1^{(1)} y_1 + \theta_2^{(1)} y_2 + \theta_{12}^{(2)} y_1 y_2 \right)$$

y_1	y_2	$\Pr(Y_1 = y_1, Y_2 = y_2)$
0	0	$c(\boldsymbol{\theta})$
1	0	$c(\boldsymbol{\theta}) \exp(\theta_1^{(1)})$
0	1	$c(\boldsymbol{\theta}) \exp(\theta_2^{(1)})$
1	1	$c(\boldsymbol{\theta}) \exp(\theta_1^{(1)} + \theta_2^{(1)} + \theta_{12}^{(2)})$

Hence we have (conditional) interpretations:

$$
\begin{aligned}
\exp(\theta_1^{(1)}) &= \frac{\Pr(Y_1 = 1, Y_2 = 0)}{\Pr(Y_1 = 0, Y_2 = 0)} \\
&= \frac{\Pr(Y_1 = 1|Y_2 = 0)}{\Pr(Y_1 = 0|Y_2 = 0)}
\end{aligned}
$$

the odds of an event at trial 1, given no event at trial 2;

$$
\begin{aligned}
\exp(\theta_2^{(1)}) &= \frac{\Pr(Y_1 = 0, Y_2 = 1)}{\Pr(Y_1 = 0, Y_2 = 0)} \\
&= \frac{\Pr(Y_2 = 1|Y_1 = 0)}{\Pr(Y_2 = 0|Y_1 = 0)}
\end{aligned}
$$

the odds of an event at trial 2, given an event at trial 1;

$$
\begin{aligned}
\exp(\theta_{12}^{(12)}) &= \frac{\Pr(Y_1 = 1, Y_2 = 1)\Pr(Y_1 = 0, Y_2 = 0)}{\Pr(Y_1 = 1, Y_2 = 0)\Pr(Y_1 = 0, Y_2 = 1)} \\
&= \frac{\Pr(Y_2 = 1|Y_1 = 1)/\Pr(Y_2 = 0|Y_1 = 1)}{\Pr(Y_2 = 1|Y_1 = 0)/\Pr(Y_2 = 0|Y_1 = 0)}
\end{aligned}
$$

the ratio of odds of an event at trial 2 given an event at trial 1, divided by the odds of an event at trial 2 given no event at trial 1. Hence if this parameter is larger than 1 we have positive dependence.

341

Quadratic Exponential Log-Linear Model

We describe three approaches to modeling binary data: conditional odds ratios, correlations, marginal odds ratios.

Zhao and Prentice (1990) consider the log-linear model with third and higher-order terms set to zero, so that

$$
\Pr(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}) \exp\left(\sum_j \theta_j^{(1)} y_j + \sum_{j<k} \theta_{jk}^{(2)} y_j y_k\right).
$$

For this model

$$
\frac{\Pr(Y_j = 1|Y_k = y_k, Y_l = 0, l \neq j, k)}{\Pr(Y_j = 0|Y_k = y_k, Y_l = 0, l \neq j, k)} = \exp(\theta_j^{(1)} + \theta_{jk}^{(2)} y_k).
$$

Interpretation:

- $\exp(\theta_j^{(1)})$ is the odds of a success, given all other responses are zero.

- $\exp(\theta_{jk}^{(2)})$ is the odds ratio describing the association between Y_j and Y_k, given all other responses are fixed and equal to zero.

342

Issues:

1. Suppose we now wish to model $\boldsymbol{\theta}$ as a function of $\boldsymbol{x}$.

   Example: Y respiratory infection, x mother's smoking (no/yes). Then we could let the parameters $\boldsymbol{\theta}$ depend on x, i.e. $\boldsymbol{\theta}(x)$. But the difference between $\theta_j^{(1)}(x=1)$ and $\theta_j^{(1)}(x=0)$ represent the effect of smoking on the *conditional* probability of respiratory infection at visit j, given that there was no infection at any other visits. Since smoking is likely to influence the response at all visits a part of the smoking effect will be conditioned away. Difficult to interpret, and we would rather model the *marginal* probability.

2. The interpretation of the $\boldsymbol{\theta}$ parameters depends on the number of responses n – particularly a problem in a longitudinal setting with different n_i.

A Related Model

Fitzmaurice and Laird (1993) parameterize in terms of $\mu_j = \Pr(Y_j = 1)$ (the marginal mean) and the second and higher-order parameters, $\theta_{12}^{(2)}, ..., \theta_{n-1n}^{(2)}, ..., \theta_{12...n}^{(n)}$. If we let $\boldsymbol{\theta}_1 = (\theta_1^{(1)}, ..., \theta_n^{(1)})$ and $\boldsymbol{\theta}_2 = (\theta_{12}^{(12)}, \theta_{13}^{(2)}, ..., \theta_{n-1n}^{(2)}, ..., \theta_{1...n}^{(n)})$ and $\boldsymbol{w} = (y_1 y_2, y_1 y_3, ..., y_{n-1n}, y_{123}, ..., y_1 y_2...y_n)$, then the loglinear model can be written as

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = c(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \exp(\boldsymbol{y}^{\mathrm{T}} \boldsymbol{\theta}_1 + \boldsymbol{w}^{\mathrm{T}} \boldsymbol{\theta}_2)$$

The parameters of interest are $\boldsymbol{\mu} = (\mu_1, ..., \mu_n) = \boldsymbol{\mu}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ so that we can reparameterize from $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ to $(\boldsymbol{\mu}, \boldsymbol{\theta}_2)$.

GEE used for inference (with a logistic mean model, for example).

The limitiation of this parameterizatio is that $\boldsymbol{\theta}_2$ as a conditional odds depends on the number of other responses. So most useful for cases in which we have the same number of observations per cluster.

If the association between responses is of interest then not so easy to interpret.

Bahadur Representation

Another approach to parameterizing a multivariate binary model was proposed by Bahadur (1961) who used marginal means, as well as second-order moments specified in terms of correlations.

Let

$$
\begin{aligned}
R_j &= \frac{Y_j - \mu_j}{[\mu_j(1-\mu_j)]^{1/2}} \\
\rho_{jk} &= \mathrm{corr}(Y_j, Y_k) = \mathrm{E}[R_j R_k] \\
\rho_{jkl} &= \mathrm{E}[R_j R_k R_l] \\
&\quad \ldots \quad \cdot \quad \ldots \\
\rho_{1,\ldots,n} &= \mathrm{E}[R_1 \ldots R_n]
\end{aligned}
$$

Then we can write

$$
\mathrm{Pr}(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{j=1}^{n} \mu_j^{y_j} (1-\mu_j)^{1-y_j}
$$

$$
\times \left(1 + \sum_{j<k} \rho_{jk} r_j r_k + \sum_{j<k<l} \rho_{jkl} r_j r_k r_l + \ldots + \rho_{1,\ldots,n} r_1 r_2 \ldots r_n \right)
$$

Appealing because we have the marginal means μ_j and "nuisance" parameters.

345

Limitations:

Unfortunately, the correlations are constrained in complicated ways by the marginal means.

Example: consider measurements on a single individual, Y_1 and Y_2, with means μ_1 and μ_2. We have

$$
\mathrm{corr}(Y_1, Y_2) = \frac{\mathrm{Pr}(Y_1 = 1, Y_2 = 1) - \mu_1 \mu_2}{\{\mu_1(1-\mu_1)\mu_2(1-\mu_2)\}^{1/2}}
$$

and

$$
\max(0, \mu_1 + \mu_2 - 1) \leq \mathrm{Pr}(Y_1 = 1, Y_2 = 1) \leq \min(\mu_1, \mu_2),
$$

which implies complex constraints on the correlation.

For example, if $\mu_1 = 0.8$ and $\mu_2 = 0.2$ then $0 \leq \mathrm{corr}(Y_1, Y_2) \leq 0.25$.

Marginal Odds Ratios

An alternative is to parameterize in terms of the marginal means and the marginal odds ratios defined by

$$
\begin{aligned}
\gamma_{jk} &= \frac{\Pr(Y_j = 1, Y_k = 1)\Pr(Y_j = 0, Y_k = 0)}{Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)} \\
&= \frac{\Pr(Y_j = 1 \mid Y_k = 1)/\Pr(Y_j = 0 \mid Y_k = 1)}{\Pr(Y_j = 1 \mid Y_k = 0)/\Pr(Y_j = 0 \mid Y_k = 0)}
\end{aligned}
$$

which is the odds that the j-th observation is a 1, given the k-th observation is a 1, divided by the odds that the j-th observation is a 1, given the k-th observation is a 0.

Hence if $\gamma_{jk} > 1$ we have positive dependence between outcomes j and k.

It is then possible to obtain the joint distribution in terms of the means $\boldsymbol{\mu}$, where $\mu_j = \Pr(Y_j = 1)$ the odds ratios $\boldsymbol{\gamma} = (\gamma_{12}, ..., \gamma_{n-1,n})$ and contrasts of odds ratios

347

We need to find $\mathrm{E}[Y_j Y_k] = \mu_{jk}$, so that we can write down the likelihood function, or an estimating function.

For the case of $n = 2$ we have

$$
\gamma_{12} = \frac{\Pr(Y_1 = 1, Y_2 = 1)\Pr(Y_1 = 0, Y_2 = 0)}{Pr(Y_1 = 1, Y_2 = 0)\Pr(Y_1 = 0, Y_2 = 1)} = \frac{\mu_{12}(1 - \mu_1 - \mu_2 + \mu_{12})}{(\mu_1 - \mu_{12})(\mu_2 - \mu_{12})},
$$

and so

$$
\mu_{12}^2(\gamma_{12} - 1) + \mu_{12}b + \gamma_{12}\mu_1\mu_2 = 0,
$$

where $b = (\mu_1 + \mu_2)(1 - \gamma_{12}) - 1$, to give

$$
\mu_{12} = \frac{-b \pm \sqrt{b^2 - 4(\gamma_{12} - 1)\mu_1\mu_2}}{2(\gamma_{12} - 1)}.
$$

		Y_2		
		0	1	
Y_1	0			$1 - \mu_1$
	1		μ_{12}	μ_1
		$1 - \mu_2$	μ_2	

348

Limitations

In a longitudinal setting (we add an i subscript to denote individuals), finding the μ_{ijk} terms is computationally complex (see Qaqish and Ivanova, 2006, Biometrika).

Large numbers of nuisance odds ratios if n_i's are large – assumptions such as $\gamma_{ijk} = \gamma$ for all i, j, k may be made.

Another possibility is to take

$$\log \gamma_{ijk} = \alpha_0 + \alpha_1 |t_{ij} - t_{ik}|^{-1},$$

so that the degree of association is inversely proportional to the time between observations.

349

Modeling Multivariate Binary Data Using GEE

For a marginal Bernoulli outcome we have

$$\Pr(Y_{ij} = y_{ij} | \boldsymbol{x}_{ij}) = \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}} = \exp(y_{ij}\theta_{ij} - \log\{1 + e^{\theta_{ij}}\}),$$

where $\theta_{ij} = \log(\mu_{ij}/(1 - \mu_{ij}))$, an exponential family representation.

For independent responses we therefore have the likelihood

$$\Pr(\boldsymbol{Y}|\boldsymbol{x}) = \exp\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} y_{ij}\theta_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n_i} \log\{1 + e^{\theta_{ij}}\}\right] = \exp\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} l_{ij}\right].$$

To find the MLEs we consider the score equation:

$$\boldsymbol{G}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\partial l_{ij}}{\partial \theta_{ij}} \frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}(y_{ij} - \mu_{ij}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{y}_i - \boldsymbol{\mu}_i).$$

So GEE with working independence can be implemented with standard software, though we need to "fix-up" the standard errors via sandwich estimation.

350

Non-independence GEE

Assuming working correlation matrices: $\boldsymbol{R}_i(\boldsymbol{\alpha})$ and estimating equation

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{W}_i = \boldsymbol{\Delta}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{\Delta}_i^{1/2}$.

Here $\boldsymbol{\alpha}$ are parameters that we need a consistent estimator of (Newey 1990, shows that the choice of estimator for $\boldsymbol{\alpha}$ has no effect on the asymptotic efficiency).

Define a set of $n_i(n_i - 1)/2$ empirical covariances

$$R_{ijk} = (Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})$$

We can then define a set of moment-based estimating equations to obtain estimates of $\boldsymbol{\alpha}$.

First Extension to GEE

Rather than have a method of moments estimator for $\boldsymbol{\alpha}$, Prentice (1988) proposed using a second set of estimating equations for $\boldsymbol{\alpha}$. In the context of data with $\mathrm{var}(Y_{ij}) = v(\mu_{ij})$ recall:

$$\begin{aligned} \boldsymbol{G}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \\ \boldsymbol{G}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^{m} \boldsymbol{E}_i^{\mathrm{T}} \boldsymbol{H}_i^{-1} (\boldsymbol{T}_i - \boldsymbol{\Sigma}_i) \end{aligned}$$

where $R_{ij} = \{Y_{ij} - \mu_{ij}\}$, to give "data"

$$\boldsymbol{T}_i^{\mathrm{T}} = (R_{i1}R_{i2}, ..., R_{in_i-1}R_{in_i}, R_{i1}^2, ..., R_{in_i}^2),$$

$\boldsymbol{\Sigma}_i(\boldsymbol{\alpha}) = \mathrm{E}[\boldsymbol{T}_i]$ is a model for the correlations and variances of the standardized residuals, $\boldsymbol{E}_i = \frac{\partial \boldsymbol{\Sigma}_i}{\partial \alpha}$, and $\boldsymbol{H}_i = \mathrm{cov}(\boldsymbol{T}_i)$ is the working covariance model.

The vector $\boldsymbol{T}_i$ has $n_i(n_i - 1)/2 + n_i$ elements in general — the working covariance model $\boldsymbol{H}_i$ is in general complex.

In the context of binary data Prentice (1988) the variances are determined by the mean and so the last n_i terms of $\boldsymbol{T}_i$ are dropped; he also suggests taking a diagonal working covariance model, $\boldsymbol{H}_i$, i.e. ignoring the covariances. The theoetical variances ae given by

$$\text{var}(R_{ij}R_{ik}) = 1 + (1-2p_{ij})(1-2p_{ik})\{p_{ij}(1-p_{ij})p_{ik}(1-p_{ik})\}^{-1/2}\Sigma(\boldsymbol{\alpha})_{ijk} - \Sigma(\boldsymbol{\alpha})^2_{ijk}$$

which depend on the assumed correlation model $\boldsymbol{\Sigma}$ — these may be taken as the diagonal elements of $\boldsymbol{H}_i$.

Application of GEE Extension to Marginal Odds Model

We have the marginal mean model

$$\text{logit } \text{E}[Y_{ij} \mid \boldsymbol{X}_{ij}] = \boldsymbol{\beta}\boldsymbol{X}_{ij}.$$

Suppose we specify a model for the associations in terms of the marginal log odds ratios:

$$\alpha_{ijk} = \log\left\{\frac{\Pr(Y_{ij}=1, Y_{ik}=1)\Pr(Y_{ij}=0, Y_{ik}=0)}{\Pr(Y_{ij}=1, Y_{ik}=0)\Pr(Y_{ij}=0, Y_{ik}=1)}\right\}.$$

These are nuisance parameters, but how do we estimate them?

Carey et al. (1992) suggest the following approach fo estimating $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let

$$
\begin{aligned}
\mu_{ij} &= \Pr(Y_{ij}=1) \\
\mu_{ik} &= \Pr(Y_{ik}=1) \\
\mu_{ijk} &= \Pr(Y_{ij}=1, Y_{ik}=1)
\end{aligned}
$$

It is easy to show that

$$
\frac{\Pr(Y_{ij} = 1 \mid Y_{ik} = y_{ik})}{\Pr(Y_{ij} = 0 \mid Y_{ik} = y_{ik})} = \exp(y_{ik}\alpha_{ijk})\frac{\Pr(Y_{ij} = 1, Y_{ik} = 0)}{\Pr(Y_{ij} = 0, Y_{ik} = 0)}
$$

$$
= \exp(y_{ik}\alpha_{ijk})\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right)
$$

which can be written as a logistic regression model in terms of conditional probabilities:

$$
\text{logit E}[Y_{ij} \mid Y_{ik}] = \log\left(\frac{\Pr(Y_{ij} = 1 \mid Y_{ik} = y_{ik})}{\Pr(Y_{ij} = 0 \mid Y_{ik} = y_{ik})}\right)
$$

$$
= y_{ik}\alpha_{ijk} + \log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right)
$$

where the term on the right is a known offset (the μ's are a function of $\boldsymbol{\beta}$ only). Suppose for simplicity that $\alpha_{ijk} = \alpha$ then given current estimates of $\boldsymbol{\beta}, \alpha$, we can fit a logistic regression model by regressing Y_{ij} on Y_{ik} for $1 \leq j < k \leq n_i$, to estimate α — this can then be used within the working correlation model.

Carey et al. (1992) refer to this method as **alternating logistic regressions**.

Indonesian Children's Health Example

```
> summary(geese(y ~ xero+age,corstr="independence",id=id,family="binomial"))
Mean Model:
 Mean Link:                    logit
 Variance to Mean Relation: binomial
 Coefficients:
               estimate        san.se         wald            p
(Intercept) -2.38479528 0.117676276 410.699689 0.000000e+00
age          -0.02605769 0.005306513  24.113112 9.083967e-07
xero          0.72015485 0.419718477   2.943985 8.619783e-02
Scale Model:
 Scale Link:                   identity
 Estimated Scale Parameters:
           estimate    san.se     wald            p
(Intercept) 0.977505 0.2766052 12.48871 0.0004094196
Correlation Model:
 Correlation Structure:      independence
Number of clusters:   275   Maximum cluster size: 6
```

```
> summary(geese(y  ~ age+xero, corstr="exchangeable", id=id,family="binomial"))
                estimate      san.se        wald              p
(Intercept) -2.37015400 0.117210489 408.902887 0.000000e+00
age          -0.02532507 0.005271204  23.082429 1.552026e-06
xero          0.58758892 0.449818037   1.706371 1.914569e-01
 Estimated Scale Parameters:
              estimate     san.se      wald              p
(Intercept) 0.9681312 0.2618218 13.67278 0.0002175859
 Estimated Correlation Parameters:
        estimate      san.se      wald          p
alpha 0.04423924 0.03222984 1.884079 0.1698713
> summary(geese(y  ~ age+xero, corstr="ar1", id=id,family="binomial"))
 Coefficients:
                estimate      san.se        wald              p
(Intercept) -2.37470963 0.11733291 409.620156 0.000000e+00
age          -0.02597886 0.00528451  24.167452 8.831225e-07
xero          0.63692645 0.44374132   2.060245 1.511859e-01
 Estimated Scale Parameters:
              estimate     san.se      wald              p
(Intercept) 0.9715817 0.2694961 12.99732 0.0003119374
 Estimated Correlation Parameters:
        estimate      san.se      wald          p
alpha 0.05844094 0.04528613 1.665344 0.1968834
```

357

Conditional Likelihood: Binary Longitudinal Data

Recall that conditional likelihood is a technique for eliminating nuisance parameters, here what we have previously modeled as random effects.

Consider individual i with binary observations $y_{i1}, ..., y_{in_i}$ and assume the random intercepts model $Y_{ij} \mid \gamma_i, \boldsymbol{\beta} \sim \text{Bernoulli}(p_{ij})$, where

$$\log \left(\frac{p_{ij}}{1 - p_{ij}} \right) = \boldsymbol{x}_{ij} \boldsymbol{\beta} + \gamma_i$$

where $\gamma_i = \boldsymbol{x}_i \boldsymbol{\beta} + b_i$ and $\boldsymbol{x}_{ij}$ (a slight change from our usual notation), are those covariates which change within an individual.

We have

$$
\begin{aligned}
\Pr(y_{i1}, ..., y_{in_i} \mid \gamma_i, \boldsymbol{\beta}) &= \prod_{j=1}^{n_i} \frac{\exp\left(\gamma_i y_{ij} + \boldsymbol{x}_{ij}\boldsymbol{\beta} y_{ij}\right)}{1 + \exp\left(\gamma_i + \boldsymbol{x}_{ij}\boldsymbol{\beta}\right)} \\
&= \frac{\exp\left(\gamma_i \sum_{j=1}^{n_i} y_{ij} + \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij}\boldsymbol{\beta}\right)}{\prod_{j=1}^{n_i}\left[1 + \exp\left(\gamma_i + \boldsymbol{x}_{ij}\boldsymbol{\beta}\right)\right]} \\
&= \frac{\exp\left(\gamma_i t_{2i} + \boldsymbol{t}_{1i}\boldsymbol{\beta}\right)}{\prod_{j=1}^{n_i}\left[1 + \exp\left(\gamma_i + \boldsymbol{x}_{ij}\boldsymbol{\beta}\right)\right]} \\
&= \frac{\exp\left(\gamma_i t_{2i} + \boldsymbol{t}_{1i}\boldsymbol{\beta}\right)}{k(\gamma_i, \boldsymbol{\beta})} \\
&= p(t_{1i}, \boldsymbol{t}_{2i} \mid \gamma_i, \boldsymbol{\beta})
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{t}_{1i} &= \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij}, \quad t_{2i} = \sum_{j=1}^{n_i} y_{ij} \\
k(\gamma_i, \boldsymbol{\beta}) &= \prod_{j=1}^{n_i}\left[1 + \exp\left(\gamma_i + \boldsymbol{x}_{ij}\boldsymbol{\beta}\right)\right].
\end{aligned}
$$

We have

$$
L_c(\boldsymbol{\beta}) = \prod_{i=1}^{m} p(\boldsymbol{t}_{1i} \mid t_{2i}, \boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{p(\boldsymbol{t}_{1i}, t_{2i} \mid \gamma_i, \boldsymbol{\beta})}{p(t_{2i} \mid \gamma_i, \boldsymbol{\beta})}
$$

where

$$
p(\boldsymbol{t}_{2i} \mid \gamma_i, \boldsymbol{\beta}) = \frac{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\gamma_i \sum_{j=1}^{n_i} y_{ij} + \sum_{k=1}^{n_i} \boldsymbol{x}_{ik} y_{ik}^l \boldsymbol{\beta}\right)}{k(\gamma_i, \boldsymbol{\beta})},
$$

where the summation is over the $\binom{n_i}{y_{i+}}$ ways of choosing y_{i+} ones out of n_i, and $\boldsymbol{y}_i^l = (y_{i1}^l, ..., y_{in_i}^l)$, $l = 1, ..., \binom{n_i}{y_{i+}}$ is the collection of these ways.
Hence

$$
\begin{aligned}
L_c(\boldsymbol{\beta}) &= \prod_{i=1}^{m} \frac{\exp\left(\gamma_i \sum_{j=1}^{n_i} y_{ij} + \sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij}\boldsymbol{\beta}\right)}{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\gamma_i \sum_{j=1}^{n_i} y_{ij} + \sum_{k=1}^{n_i} \boldsymbol{x}_{ik} y_{ik}^l \boldsymbol{\beta}\right)} \\
&= \prod_{i=1}^{m} \frac{\exp\left(\sum_{j=1}^{n_i} \boldsymbol{x}_{ij} y_{ij}\boldsymbol{\beta}\right)}{\sum_{l=1}^{\binom{n_i}{y_{i+}}} \exp\left(\sum_{k=1}^{n_i} \boldsymbol{x}_{ik} y_{ik}^l \boldsymbol{\beta}\right)}
\end{aligned}
$$

Notes

- Can be computationally expensive to evaluate likelihood if n_i is large, e.g. if $n_i = 20$ and $y_{i+} = 10$, $\binom{n_i}{y_{i+}} = 184,756$.

- There is no contribution to the conditional likelihood from individuals:

  - With $n_i = 1$.

  - With $y_{i+} = 0$ or $y_{i+} = n_i$.

  - For those covariates with $x_{i1} = ... = x_{in_i} = x_i$. The conditional likelihood estimates β's that are associated with within-individual covariates. If a covariate only varies between individuals, then it cannot be estimated using conditional likelihood.
    For covariates that vary both between and within individuals, only the within-individual contrasts are used.

- The similarity to Cox's partial likelihood may be exploited to carry out computation.

- We have not made a distributional assumption for the γ_i's!

Examples:

If $n_i = 3$ and $\boldsymbol{y}_i = (0, 0, 1)$ so that $y_{i+} = 1$ then

$$\boldsymbol{y}_i^1 = (1, 0, 0), \quad \boldsymbol{y}_i^2 = (0, 1, 0), \quad \boldsymbol{y}_i^3 = (0, 0, 1),$$

and the contribution to the conditional likelihood is

$$\frac{\exp(\boldsymbol{x}_{i3}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{i2}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{i3}\boldsymbol{\beta})}.$$

If $n_i = 3$ and $\boldsymbol{y}_i = (1, 0, 1)$ so that $y_{i+} = 2$ then

$$\boldsymbol{y}_i^1 = (1, 1, 0), \quad \boldsymbol{y}_i^2 = (1, 0, 1), \quad \boldsymbol{y}_i^3 = (0, 1, 1),$$

and the contribution to the conditional likelihood is

$$\frac{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta} + \boldsymbol{x}_{i3}\boldsymbol{\beta})}{\exp(\boldsymbol{x}_{i1}\boldsymbol{\beta} + \boldsymbol{x}_{i2}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{i1}\boldsymbol{\beta} + \boldsymbol{x}_{i3}\boldsymbol{\beta}) + \exp(\boldsymbol{x}_{i2}\boldsymbol{\beta} + \boldsymbol{x}_{i3}\boldsymbol{\beta})}.$$