Generalized Estimating Equations (gee)
for glm–type data

Søren Højsgaard

mailto:sorenh@agrsci.dk

Biometry Research Unit

Danish Institute of Agricultural Sciences

January 23, 2006

Printed: January 23, 2006 File: R-mixed-geeglm-Lecture.tex

Contents

1 Preliminaries **3**

2 Working example – respiratory illness **4**

3 Correlated Pearson–residuals **9**

4 Marginal vs. conditional models **12**

5 Marginal models for glm–type data **14**

6 Estimating equations for gee–type data **16**
- 6.1 Specifications needed for GEEs . . . 18
- 6.2 Deriving and solving GEEs . . . 20
- 6.3 Newton–iteration . . . 27
- 6.4 Estimation of the covariance of $\hat{\beta}$ . . . 29
  - 6.4.1 Model based estimate . . . 29
  - 6.4.2 Emperical estimate – sandwich estimate . . . 30
- 6.5 The working correlation matrix . . . 31

7 Exploring different working correlations **33**

8 Comparison of the parameter estimates **37**
- 8.1 When do GEEs work? . . . 39
- 8.2 What to do – geeglm vs. lmer . . . 40

1 Preliminaries

These notes deal with fitting models for responses of type often dealt with with generalized linear models (glm) but with the complicating aspect that there may be repeated measurements on the same unit.

The approach here is generalized estimating equations (gee).

There are two packages for this purpose in R: geepack and gee. We focus on the former and note in passing that the latter does not seem to undergo any further development.

The geepack package is described in the paper by Halekoh, Højsgaard and Yun in Journal of Statistical Software, www.jstatsoft.org, January 2006.

2 Working example – respiratory illness

Example 1 The data are from a clinical trial of patients with respiratory illness, where 111 patients from two different clinics were randomized to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined.

- The recorded variables are:

  Center (1,2), ID, Treatment (A=Active, P=Placebo), Gender (M=Male,F=Female), Age (in years at baseline), Baseline Response.

- The response variables are:

  Visit 1 Response, Visit 2 Response, Visit 3 Response, Visit 4 Response.

Data for 8 patients are shown in Table 1.

	center	id	treat	sex	age	baseline	visit1	visit2	visit3	visit4
1	1	1	P	M	46	0	0	0	0	0
2	1	2	P	M	28	0	0	0	0	0
3	1	3	A	M	23	1	1	1	1	1
4	1	4	P	M	44	1	1	1	1	0
5	1	5	P	F	13	1	1	1	1	1
6	1	6	A	M	34	0	0	0	0	0
7	1	7	P	M	43	0	1	0	1	1
8	1	8	A	M	28	0	0	0	0	0

Table 1: Respiratory data for eight individuals. Measurements on the same individual tend to be alike.

Interest is in comparing the treatments, but also to include center, age, gender and baseline response in the model.

From Table 1 it is clear, that there is a dependency among the response measurements on the same person – measurements on the same person tend to be alike.

This dependency must be accounted for in the modelling. □

Example 2 A first approach is to ignore the dependency. This approach is *not* appropriate but illustrative.

Let y_{iv} denote the response measured on the ith person at visit v, where $v = 1, \ldots, 4$ Since the response outcomes are binary, $y_{iv} \in \{0, 1\}$, it is tempting to consider the binomial distribution as basis for the modelling. That is, to assume that $y_{iv} \sim bin(1, \mu_{iv})$ and that all y_{iv} are independent.

As specification of μ_{iv} we consider in the following the linear predictor

$$logit(\mu_{iv}) = \mu + \alpha_{center(i)} + \beta_{treat(i)} + \gamma \cdot age_i + \delta \cdot baseline_i$$

Note that the expression for $logit(\mu_{iv})$ does not include the vist v. We will write this briefly as

$$logit(\mu) = center + treat + age + baseline$$

Other linear predictors can clearly be considered.

Table 2 contains the parameter estimates under the model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	−0.20	0.52	−0.39	0.70
center	1.07	0.22	4.94	0.00
sexM	0.11	0.27	0.41	0.68
age	−0.02	0.01	−2.62	0.01
treatP	−1.01	0.21	−4.79	0.00

Table 2: Parameter estimates when assuming independence

A more elaborate model is to allow for over/underdispersion. That is, to make a quasi likelihood model where the variance function is

$$\phi\mu_{iv}(1-\mu_{iv})$$

The dispersion parameter ϕ is estimated as $\hat{\phi} = 1.009$, i.e. there is no indication of over/under dispersion. □

3 Correlated Pearson–residuals

Based on the fitted (wrong) independence model we can calculate the Pearson residuals

$$e_{iv} = \frac{y_{iv} - \hat{\mu}_{iv}}{\sqrt{\mu_{iv}(1 - \mu_{iv})}}, \quad i = 1, \ldots, N, v = 1, \ldots, 4$$

which under the model approximately have mean 0 and variance 1.

From these we can estimate the covariance matrix $\hat{\Sigma}$ in Table 3 which shows covariances between measurements on the same individual.

	1	2	3	4
1	1.01	0.45	0.38	0.45
2	0.45	1.01	0.53	0.45
3	0.38	0.53	0.98	0.52
4	0.45	0.45	0.52	1.00

Table 3: Covariance matrix based on Pearson residuals.

Since the elements on the diagonal in Table 3 are about 1, the matrix can also be regarded as a correlation matrix.

If the observations *were* independent then the true (i.e. theoretical) correlations should be zero.

The estimated correlations in Table 3 suggest that there is a positive correlation.

The task in the following is to account for the correlation between measurements on the same individual.

4 Marginal vs. conditional models

Linear mixed models of the type

$$y = X\beta + Zu + e$$

can be described as conditional models. Suppose that $\mathbb{V}\mathrm{ar}(e) = \sigma^2 I$. Then the conditional distribution of y given u is

$$y|u \sim N(X\beta + Zu, \sigma^2 I)$$

so in the conditional distribution, the components of y are independent.

Generally, we impose a structure on u in terms of $\mathbb{V}\mathrm{ar}(u) = G$. The marginal distribution of y is

$$y \sim N(X\beta, ZGZ + \sigma^2 I)$$

so marginally the components of y are dependent with the structure given in $V = ZGZ' + \sigma^2 I$.

So this way, one can see the linear mixed model formula as a way of building up a model in which the responses are correlated.

An alternative approach is to construct a marginal model directly, e.g.

$$y \sim N(X\beta, V)$$

by specifying directly a structure on V.

5 Marginal models for glm–type data

A way of dealing with correlated glm–type observations is to create a "marginal model" directly. That is, to create a model for e.g. a 4–dimensional response vector which consist of binary variables.

We follow the approach by Liang and Zeger (1986).

The term "marginal model" is quoted, because formally we do not specify a proper statistical model (in terms of making distributional assumptions).

All we do is to to specify

1. how the mean $\mathbb{E}(y)$ depends on the covariates through a link function $g(\mathbb{E}(y)) = X\beta$ and
2. how the variance $\mathbb{V}\mathrm{ar}(y)$ varies as a function of the mean (the variance function), i.e. $\mathbb{V}\mathrm{ar}(y) = v(\mathbb{E}(y))$.

With these specifications, one can derive a system of estimating equations by which an estimate $\hat{\beta}$ and $\mathbb{V}\mathrm{ar}(\hat{\beta})$ can be obtained. Asymptotically $\hat{\beta} \sim N(\beta, ...)$.

So we make fewer assumptions than if we specify a full statistical model. This extra flexibility comes at a price:

- The estimate $\hat{\beta}$ may not be the best possible.
- Hypothesis testing is based on Wald tests (since, as there is no distribution and hence no likelihood).
- Model checking is difficult.

One can regard GEEs as a "quick and dirty" method.

6 Estimating equations for gee–type data

For correlated glm–type data, estimating equations have in the litterature become known as generalised estimating equations (GEEs).

- GEEs can, in connection with correlated glm–type data, be regarded as an extension of the esimation methods (score equations) used GLMs/QLs. This justifies the term "generalized".
- On the other hand, the estimating equations used in connection with correlated glm–type data are are rather specialized type of estimating equations. As such, the term "generalized" is a little misleading.

For this reason the function for dealing with these types of data in the geepack package is called geeglm().

With GEEs for GLM–type data

- the emphasis is on modeling the expectation of the dependent variable in relation to the covariates (just like with GLMs), whereas
- the correlation structure is considered to be a *nuisance* (not of interest in itself), which is accounted for by the method.

6.1 Specifications needed for GEEs

The setting is as follows: On each of $i = 1, \ldots, N$ subjects, there are made n_i measurements $y_i = (y_{i1}, \ldots, y_{in_i})$.

- Measurements on different subjects are assumed to be independent
- Measurements on the same subject are allowed to be correlated.

The model formulation is similar to that of a GLM:

Systematic part: Relate the expectation $\mathbb{E}(y_{it}) = \mu_{it}$ to the linear predictor via the link function

$$h(\mu_{it}) = \eta_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta}$$

Random part: Specify how the variance $\mathbb{V}\mathrm{ar}(y_{it})$ is related to the mean $\mathbb{E}(y_{it})$ by specifying a variance function $V(\mu_{it})$ such that $\mathbb{V}\mathrm{ar}(y_{it}) = \phi V(\mu_{it})$.

The correlation part: In addition to these "GLM" steps we need to impose a correlation structure for observations on the same unit. This is done by specifying a working correlation matrix.

6.2 Deriving and solving GEEs

Consider observations $y_1, \ldots, y_n$ with common mean θ. The least squares criterion for estimating μ is to minimize

$$\Psi(\theta, y) = \sum_i (y_i - \theta)^2$$

This is achieved by setting the derivative to zero:

$$\psi(\theta; y) = \Psi'(\theta; y) = 2 \sum_i (y_i - \theta) = 0$$

We say that $\psi(\theta) = \psi(\theta; y)$ is an estimating function and $\psi(\theta; y) = 0$ is an estimating equation.

Consider data $y = (y_1, \ldots, y_n)$ and a model $p(y; \theta)$ where $\theta \in \Theta \subseteq R^p$. The idea behind estimating functions is to find a function $\psi(\theta)$ which immitates the score function

$$U(\theta) = \frac{d}{d\theta} \log p(y; \theta).$$

Let

$$\psi(\theta) \quad = \quad \psi(\theta, y)$$

be such a function denoted an estimating function. Solve (usually by iteration) the estimating equations

$$\psi(\theta) = 0 \text{ giving } \hat{\theta} = \hat{\theta}(y)$$

If $E_\theta(\psi(\theta)) = 0$ for all θ (which holds for the score function), then ψ is said to be unbiased. (Note that unbiasedness is a property of the estimation function rather than of the estimate $\hat{\theta}$.)

In practice we frequently we consider weighted sums of estimating functions of the form

$$\sum_i a_i(y_i - \mu_i(\theta))$$

which consequently is unbiased.

Unbiasedness of the estimation function implies that $\hat{\theta}$ is asymptotically consistent. Another property is that $\hat{\theta}$ is asymptotically normal.

- The sensitivity of the estimating function is the $p \times p$ matrix

$$S_\psi(\theta) = E_\theta(\frac{\partial \psi}{\partial \theta})$$

  which indicates how "steep" ψ is "on average". So a large value of S is good.

- The variability of an estimating function is

$$V_\psi(\theta) = Var_\theta(\psi(\theta)) = E_\theta(\psi(\theta)\psi(\theta)^\top)$$

  A small value of V is good because that indicates that different samples give almost the same $\hat{\theta}$.

- The Godambe information matrix is defined as

$$J_\psi(\theta) = S_\psi^\top(\theta) V_\psi(\theta)^{-1} S_\psi(\theta)$$

- It then holds that

$$\hat{\theta} \sim_{approx} N(\theta, J_\psi(\theta)^{-1})$$

Note: If ψ is the score function (arising from a likelihood) then $S_\psi(\theta) = -I(\theta)$ and $V_\psi(\theta) = I(\theta)$ and hence $J_\psi(\theta) = I(\theta)$.

If $\psi(\theta)$ has the form $\psi = X^\top(y - \mu)$ then

$$S_\psi(\theta) = X^\top[\frac{\partial \mu}{\partial \theta_1} : \cdots : \frac{\partial \mu}{\partial \theta_p}].$$

Moreover,

$$V_\psi(\theta) = E(\psi(\theta))^2 = E(X^\top(y - \mu)(y - \mu)^\top X)$$

which is not so easy to calculate. In principle, however, this quantity can be estimated using

$$\hat{V}_\psi(\theta) = X^\top(y - \hat{\mu})(y - \hat{\mu})^\top X$$

In practice, this may cause problem since $\hat{V}_\psi$ in this case may not be invertible.

The GEE by Liang and Zeeger (1986) for estimating a p vector β is given by

$$\psi(\beta) = \sum_i \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i(\beta)) \tag{1}$$

Since $g(\mu_{ij}) = x_{ij}'\beta$, the deriavative $\frac{\partial \mu_i'}{\partial \beta}$ has as its kjth entry

$$[\frac{\partial \mu_i'}{\partial \beta}]_{kj} = \frac{x_{ikj}}{g'(\mu_{ik})}$$

The variance is

$$\Sigma_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2}$$

where A_i is diagonal with the $v(\mu_{ij})$s on the diagonal and $R(\alpha)$ is the correlation matrix.

The correlation matrix is generally unknown, so therefore one specifies a "working correlation matrix", e.g. with an autoregressive in a repeated measurements problem.

In absence of a good guess of $R(\alpha)$ the identity matrix is often a good choice.

Typically $R(\alpha)$ is estimated from data iteratively by using the current estimate of β to calculate a function of the Pearson residuals

$$e_{ij} = \frac{y_{ij} - \mu_{ij}(\beta)}{\sqrt{v(\mu_{ij}(\beta))}} \tag{2}$$

The dispersion parameter is often estimated as

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{N} \sum_{j=1}^{n_i} e_{ij}^2 \tag{3}$$

6.3 Newton–iteration

The fitting algorithm then becomes

1. Compute an initial estimate of β from a glm (i.e. by assuming independence)

2. Compute an estimate $R(\alpha)$ of the working correlation on the basis of the current Pearson residuals and the current estimate of β

3. Compute an estimate of the variance as

$$\Sigma_i = \phi A_i^{1/2} \hat{R}(\alpha) A_i^{1/2}$$

4. Compute an updated estimate of β based on the Newton–step

$$\beta := \beta + [\sum_i \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta'}]^{-1} [\sum_i \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1} (y_i - \mu_i(\beta))]$$

Iterate through 2–4 until convergence. Note that ϕ needs not to be estimated until the last iteration.

The GEE estimate $\hat{\beta}$ of β is often very similar to the estimate obtained if observations were treated as being independent. In other words, the estimate $\hat{\beta}$ for GEEs is often very similar to the estimate obtained by fitting a QL–model to the data.

6.4 Estimation of the covariance of $\hat{\beta}$

There are two classical ways of estimating the covariance $\mathbb{C}\mathrm{ov}(\hat{\beta})$.

6.4.1 Model based estimate

$$\mathbb{C}\mathrm{ov}(\hat{\beta})_m = I_0^{-1}, \quad I_0 = \sum_i \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta'}$$

This is the "GEE–version" of inverse Fisher information used when esitmating $\mathbb{C}\mathrm{ov}(\beta)$ in a glm. Here $\mathbb{C}\mathrm{ov}(\hat{\beta})_m$ consistently estimates $\mathbb{C}\mathrm{ov}(\hat{\beta})$ if i) the mean model and ii) the working correlation are correct.

6.4.2 Emperical estimate – sandwich estimate

$$\mathbb{C}\mathrm{ov}(\hat{\beta})_e = I_0^{-1} I_1 I_0^{-1}$$

where

$$I_1 = \sum_i \frac{\partial \mu_i'}{\partial \beta} \Sigma_i^{-1} \, \mathbb{C}\mathrm{ov}(y_i) \Sigma_i^{-1} \frac{\partial \mu_i}{\partial \beta'}$$

Here $\mathbb{C}\mathrm{ov}(\hat{\beta})_e$ is a consistent estimate of $\mathbb{C}\mathrm{ov}(\hat{\beta})$ even if the working correlation is misspecified, i.e. if $\mathbb{C}\mathrm{ov}(y_i) \neq \Sigma_i$.

In practice, $\mathbb{C}\mathrm{ov}(y_i)$ is replaced by $(y_i - \mu_i(\beta))(y_i - \mu_i(\beta))'$.

6.5 The working correlation matrix

The notion of working correlation matrices can be introduced through Example 1 on the respiratory data.

Recall that the measurements on the ith person are $y_i = (y_{i1}, \ldots, y_{i4})$.

It is natural do imagine that the correlation between two measurements on the same unit decrease as the time between the measurements increase. A particular simple way of modelling this as

$$\mathrm{Corr}(y_{it}, y_{ik}) = \alpha^{|t-k|}$$

If $|\alpha| < 1$ then $\alpha^{|t-k|}$ tends to 0 as $|t-k|$ increase which is what we wanted. This correlation structure is called an *autoregression of order 1*, briefly written ar(1).

The ar(1) correlation structure can be written in matrix form as a correlation matrix:

$$R(\alpha) = \begin{bmatrix} 1 & \alpha^1 & \alpha^2 & \alpha^3 \\ \alpha^1 & 1 & \alpha^1 & \alpha^2 \\ \alpha^2 & \alpha^1 & 1 & \alpha^1 \\ \alpha^3 & \alpha^2 & \alpha^1 & 1 \end{bmatrix}$$

Note that R depends on a single unknown parameter α which must be estimated from data.

Some additional classical working correlation matrices are presented in Section 7.

7 Exploring different working correlations

Example 3 With the autoregressive working correlation structure the correlation parameter is estimated as $\hat{\alpha} = 0.61$ which implies that the correlation matrix is

	1	2	3	4
1	1.00	0.61	0.37	0.23
2	0.61	1.00	0.61	0.37
3	0.37	0.61	1.00	0.61
4	0.23	0.37	0.61	1.00

It is noted that this correlation matrix *does not* fit very closely to the empirical matrix in Table 3. □

Example 4 Table 3 actually suggests that the correlation is about the same no matter how far apart the measurements are in time. This corresponds to the exchangable working correlation matrix given by

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \alpha \\ \alpha & 1 & \alpha & \alpha \\ \alpha & \alpha & 1 & \alpha \\ \alpha & \alpha & \alpha & 1 \end{bmatrix}$$

(Exchangable means that one can swap the order of any two measurements without changing the correlation structure).

The estimated value for α is now $\hat{\alpha} = 0.46$ which fits much closer to the empirical matrix in Table 3. $\square$

Example 5 The unstructured working correlation matrix is:

$$R(\alpha) = \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} \\ \alpha_{14} & \alpha_{24} & \alpha_{34} & 1 \end{bmatrix}$$

The ustructured working correlation matrix should be used *only with great care*: If there are n measurements per unit there are $n(n+1)/2$ parameters to estimate. This number becomes very large even for moderate n. In practice this means that the correlation parameters can be poorly estimated or that the statistical program fails to produce a result. $\square$

Example 6 The independence working correlation matrix is particularly simple:

$$R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

With the indpendence working correlation matrix, the actual dependence between the observations is incorporated in the model *only* through the emperical covariance (which informally can be thought of as Table 3).

Note that R does not depend on any unknown parameter.

In absence of important prior knowledge, the independence working correlation is often a good choice. □

8 Comparison of the parameter estimates

So far, five models have been considered: 1) observations are assumed independent 2) a ar(1) working correlation 3) an exchangable correlation 4) the unstructured working correlation and 5) the independence working correlation.

Table 4 shows the parameter estimates/ standard errors.

The following general results are illustrated from this table

- The independence and exchangable working correlation structures produce exactly the same parameter estimates as one obtains if a GLM is fitted to data. That is always true.
- The standard errors produced by the GEEs are all of a comparable size (they are practically identical). The standard errors of the GEEs are about 50 % larger than the GLM standard errors.

	Est.GLM	SE.GLM	Est.Ar1	SE.Ar1	Est.Exc	SE.Exc
(Intercept)	−0.20	0.52	−0.49	0.82	−0.20	0.82
center	1.07	0.22	1.18	0.34	1.07	0.33
sexM	0.11	0.27	0.13	0.42	0.11	0.41
age	−0.02	0.01	−0.02	0.01	−0.02	0.01
treatP	−1.01	0.21	−0.91	0.33	−1.01	0.32

	Est.GLM	SE.GLM	Est.Un	SE.Un	Est.Ind	SE.Ind
(Intercept)	−0.20	0.52	−0.26	0.81	−0.20	0.82
center	1.07	0.22	1.08	0.33	1.07	0.33
sexM	0.11	0.27	0.13	0.41	0.11	0.41
age	−0.02	0.01	−0.02	0.01	−0.02	0.01
treatP	−1.01	0.21	−0.99	0.32	−1.01	0.32

Table 4: Comparison of parameter estimates.

8.1 When do GEEs work?

GEE works best if

- the number of observations per subject is small and the number of subjects is large
- in longitudinal studies (e.g. growth curves) the measurements are taken at the same times for all subjects

8.2 What to do – geeglm vs. lmer

Consider this example: On several subjects i, a binary response y_{it} has been measured on $t = 1, \ldots, 4$ occasions. Suppose the subjects have been given either control or treatment.

Marginal models – geeglm()

- The starting point for a marginal model with geeglm() is to model the parameter $p_{it} = Pr(y_{it} = 1)$ of interest, e.g.

$$logit\ p_{it} = \alpha_{treat(i)}$$

- We specify a variance function, e.g. $\mathbb{V}\mathrm{ar}(y_{it}) = p_{it}(1 - p_{it})$.
- Then we specify a corelation structure which should "capture" that $y_{i1}, \ldots, y_{i4}$ are dependent, e.g. an ar(1) structure.
- After fitting such a "model" we get a parameter estimate $\hat{\alpha}_1, \hat{\alpha}_2$ together with a variance estimate $\mathbb{V}\mathrm{ar}(\hat{\alpha}_k)$.

- We do not get a (usable) measure of how "correlated" measurements on the same unit are.

Conditional models – lmer()

- For a conditional model with lmer(), the starting point is to assume that $y_{it} \sim bern(p_{it})$ and that

$$logit\ p_{it} = \alpha_{treat(i)} + U_i, \quad U_i \sim N(0, \sigma_U^2)$$

  Hence, conditional on U_i, $y_{i1}, \ldots, y_{i4}$ are independent unconditionally they are not.

- After fitting such a "model" we get a parameter estimate $\hat{\alpha}_1, \hat{\alpha}_2$ together with a variance estimate $\mathbb{V}\mathrm{ar}(\hat{\alpha}_k)$ – and an estimate of σ_U^2.

- It is not clear how to obtain a (usable) measure of how "correlated" measurements on the same unit are – at least not on an interpretable scale.

- Moreover, we have

$$p_{it} = \frac{\exp(\alpha_{treat(i)} + U_i)}{1 + \exp(\alpha_{treat(i)} + U_i)}$$

and as such, U_i are difficult to interpret.