Overview

Marginal Models for Linear Models

For the linear model, general linear models which allow for a more general covariance structure $\boldsymbol{V}$,

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon, \quad \varepsilon \sim (\mathbf{0}, \boldsymbol{V})$$

are marginal models. They specify a model for the marginal mean $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\beta$, and for marginal variance and correlation structure, $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{V}$.

A model for the covariance structure $\boldsymbol{V}$ can be derived from the three components random effects, serial correlation, and additional independent variation ("measurement error").

Usually, normality of ε is also assumed. Note that the normal distribution is a special case: When we specify mean $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\beta$, as well as variance and correlation structure $\mathrm{Var}(\boldsymbol{Y}) = \boldsymbol{V}$, we have fully specified the distribution of $\boldsymbol{Y}$. Inference can then proceed based on the likelihood.

Marginal Models for Linear Models

Let $\boldsymbol{\theta}$ denote the unknown parameters in $\boldsymbol{V} \equiv \boldsymbol{V}(\boldsymbol{\theta})$, and $n = \sum_{i=1}^{I} n_i$.

The log-likelihood then is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Let $\boldsymbol{U}(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}}\boldsymbol{V}(\boldsymbol{\theta})$ and $\boldsymbol{U}_i(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}}\boldsymbol{V}_i(\boldsymbol{\theta})$. The score equations then are

$$\begin{aligned}
S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \boldsymbol{X}'\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \sum_{i=1}^{I}\boldsymbol{X}_i'\boldsymbol{V}_i(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{0} \\
S_{\boldsymbol{\theta}}(\boldsymbol{\beta}, \boldsymbol{\theta}) &= -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}(\theta)^{-1}\boldsymbol{U}(\boldsymbol{\theta})) + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0} \\
&= \sum_{i=1}^{I} -\frac{1}{2}\mathrm{tr}(\boldsymbol{V}_i(\theta)^{-1}\boldsymbol{U}_i(\boldsymbol{\theta})) + \frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{V}_i(\boldsymbol{\theta})^{-1}\boldsymbol{U}_i(\boldsymbol{\theta})\boldsymbol{V}_i(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})
\end{aligned}$$

using the matrix derivative rules $\frac{\partial \log|\boldsymbol{A}|}{\partial x} = \mathrm{tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\right)$ and $\frac{\partial \boldsymbol{A}^{-1}}{\partial x} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$.

Marginal Models for Linear Models

$S_{\beta}(\beta, \theta)$ and $S_{\theta}(\beta, \theta)$ are estimating equations for β and θ, both of the form

$$m(\mathbf{Y}; \beta, \theta) = \sum_{i=1}^{I} m(\mathbf{Y}_i; \beta, \theta) = \mathbf{0}.$$

The score equations in particular have nice properties. They are unbiased, $\mathrm{E}(S_{\beta}(\beta, \theta)) = \mathbf{0}$ and $\mathrm{E}(S_{\theta}(\beta, \theta)) = \mathbf{0}$, which is a general property of score equations and important as it ensures consistency of the estimators under regularity conditions.

However, note that this is actually true for $S_{\beta}(\beta, \theta)$ and $S_{\theta}(\beta, \theta)$ as estimating equations without the assumption $\mathbf{Y} \sim N(\mathbf{X}\beta, \mathbf{V})$, as long as the mean and variance are correctly specified.

$$\begin{aligned}
\mathrm{E}\{S_{\beta}(\beta, \theta)\} &= \mathrm{E}\{\mathbf{X}'\mathbf{V}(\theta)^{-1}(\mathbf{Y} - \mathbf{X}\beta)\} = \mathbf{X}'\mathbf{V}(\theta)^{-1}(\mathbf{X}\beta - \mathbf{X}\beta) = \mathbf{0}, \\
\mathrm{E}\{S_{\theta}(\beta, \theta))\} &= -\frac{1}{2}\mathrm{tr}(\mathbf{V}(\theta)^{-1}\mathbf{U}(\theta)) + \mathrm{E}\{\frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{V}(\theta)^{-1}\mathbf{U}(\theta)\mathbf{V}(\theta)^{-1}(\mathbf{Y} - \mathbf{X}\beta)\} \\
&= -\frac{1}{2}\mathrm{tr}(\mathbf{V}(\theta)^{-1}\mathbf{U}(\theta)) + \frac{1}{2}\mathrm{tr}(\mathbf{V}(\theta)^{-1}\mathbf{U}(\theta)\mathbf{V}(\theta)^{-1}\mathbf{V}(\theta)) + \mathbf{0} = \mathbf{0},
\end{aligned}$$

using $\mathrm{E}\{\mathbf{Z}'\mathbf{A}\mathbf{Z}\} = \mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$ for $\mathbf{Z} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Marginal Models for Linear Models

The score equations are also efficient. For $S(\boldsymbol{\beta},\boldsymbol{\theta}) = (S_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\theta})', S_{\boldsymbol{\theta}}(\boldsymbol{\beta},\boldsymbol{\theta})')'$,

$$\mathrm{Var}\{S(\boldsymbol{\beta},\boldsymbol{\theta})\} = \mathrm{E}\left\{-\frac{d}{d(\boldsymbol{\beta}',\boldsymbol{\theta}')}S(\boldsymbol{\beta},\boldsymbol{\theta})\right\} = \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\theta}),$$

the Fisher information matrix. It can be shown that for any unbiased estimating equation $m(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{0}$, the asymptotic variance of the resulting estimators is

$$\mathrm{E}\left\{-\frac{d}{d(\boldsymbol{\beta}',\boldsymbol{\theta}')}m(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta})\right\}^{-1}\mathrm{Var}\{m(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta})\}\,\mathrm{E}\left\{-\frac{d}{d(\boldsymbol{\beta}',\boldsymbol{\theta}')}m(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta})\right\}^{-1} \geq \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\theta})^{-1},$$

i.e. this is a lower bound analogous to the Cramér-Rao lower bound, which is achieved by the score function.

($\boldsymbol{A} \geq \boldsymbol{B}$ for two matrices $\boldsymbol{A}$ and $\boldsymbol{B} \Leftrightarrow \boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite.)

Marginal Models for Linear Models

So, the score equations are efficient, and they are unbiased even if the likelihood assuming normality is not correct, as long as mean and (co)variance are correctly specified.

What happens if the covariance matrix $\boldsymbol{V}$ is misspecified as $\boldsymbol{W}$ in our model (e.g. $\boldsymbol{W} = \boldsymbol{I_n}$)? We saw when talking about linear mixed models, that then $\widehat{\beta} = (\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Y}$ is still unbiased and consistent, but we lose efficiency $\rightarrow$ blackboard, as the covariance matrix now is

$$\text{Cov}(\widehat{\beta}) = (\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{V}\boldsymbol{W}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}.$$

However, we can estimate $\text{Cov}(\widehat{\beta})$ consistently (as $I \rightarrow \infty$) by

$$\left(\sum_{i=1}^{I}\boldsymbol{X}_i'\widehat{\boldsymbol{W}}_i^{-1}\boldsymbol{X}_i\right)^{-1}\left\{\sum_{i=1}^{I}\boldsymbol{X}_i'\widehat{\boldsymbol{W}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta})(\boldsymbol{Y}_i - \boldsymbol{X}_i\widehat{\beta})'\widehat{\boldsymbol{W}}_i^{-1}\boldsymbol{X}_i\right\}\left(\sum_{i=1}^{I}\boldsymbol{X}_i'\widehat{\boldsymbol{W}}_i^{-1}\boldsymbol{X}_i\right)^{-1},$$

and if the $\boldsymbol{W}$ is a reasonable approximation of $\boldsymbol{V}$, we don't lose too much.

Of course, if the $\boldsymbol{\theta}$ parameters are of main interest, this does not help us. But it is helpful if the mean parameters $\boldsymbol{\beta}$ are our main focus, as long as the mean is correctly specified.

Marginal Models in the Generalized Case

Let us now consider the generalized case. We might be interested, for example, in logistic models for binary data, or Poisson models for count data.
We observe repeated measurements $Y_{ij}, j = 1, \ldots, n_i$, on subjects $i = 1, \ldots, I$.
Marginally, observations on the same subject are not independent.

We have discussed mixed models as one way to model the correlation structure. However, these models also have drawbacks. First, the computational complexity involved in fitting these models, due to the integral over the random effects in the likelihood. Second, the assumptions might be too strong in some cases. In addition to the distributional assumptions on $\boldsymbol{Y}|\boldsymbol{b}$ and $\boldsymbol{b}$, one also assumes that Y_{ij} and Y_{ik}, $j \neq k$, are independent given $\boldsymbol{b}$ (no residual correlation here). Marginal models can be an alternative that requires less assumptions, and no numerical integration.

Marginal Models in the Generalized Case

Marginal models are appropriate if the focus is on population-average parameters. Remember that in a linear model, the interpretation of regression parameters β is the same in the marginal model (population-averaged) as in a mixed model (conditional on individual random effects). This is no longer the case in the generalized setting.

Example: Estimating how much vaccinating everybody would reduce the rate of a disease in the population, versus estimating the reduction in the probability of getting the disease by being vaccinated for an individual.

Also: marginal models typically aim at characterizing the mean of a response variable in terms of explanatory variables, while accommodating the longitudinal correlation structure. That is, they are most useful when the focus is on the mean rather than on the correlation structure itself.

Marginal Models in the Generalized Case

Remember, a marginal model specifies a (separate) model for marginal mean, marginal variance and marginal correlation

$$\begin{aligned} \mathrm{E}(Y_{ij}) = \mu_{ij} &= h(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \\ \mathrm{Var}(Y_{ij}) &= \phi v(\mu_{ij}) \\ \mathrm{Corr}(Y_{ij}, Y_{ik}) &= \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha}) \end{aligned} \tag{1}$$

where $h(\cdot)$ is a known link function, $v(\cdot)$ a known variance function and $\rho(\cdot)$ a known correlation function. ϕ is a scale parameter, which may need to be estimated, and $\boldsymbol{\alpha}$ are potentially additional parameters.

In the linear model, we could make use of the multivariate normal distribution. Specifying the first two moments together with normality gave us the likelihood. This is, however, not the case for other members of the GLM family. Here, there is no analogous class of multivariate models for the joint distribution of the $Y_{ij}, j = 1, \ldots, n_i$. To specify the likelihood, one has to make additional assumption on higher-order moments. It might be a) difficult to make reasonable assumptions on these higher-order moments and b) even then, the likelihood may be intractable, involving many nuisance parameters in addition to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and ϕ.

Generalized Estimating Equations (GEE)

In the absence of a full likelihood function, generalized estimating equations (GEE) (Liang & Zeger, Biometrika, 1986) are an alternative. These are the multivariate analogue of the quasi-score functions for GLMs (Wedderburn, Biometrika, 1974). As for quasi-likelihood, only mean and variance as a function of the mean (plus potentially overdispersion) are specified, but no probability distribution for the data is assumed. With notation as in (1),

$$S_{\beta}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\boldsymbol{W}_i$ contains the model-based values $\mathrm{Cov}(Y_{ij}, Y_{ik})$. $\boldsymbol{W}$ is also called the working covariance matrix to distinguish it from the true covariance matrix $\boldsymbol{V}$. ($\boldsymbol{W}$ is often chosen as a reasonable approximation to $\boldsymbol{V}$, and robust standard errors are then used for $\widehat{\boldsymbol{\beta}}$.)

This equation reduces to the score equation for $\boldsymbol{\beta}$ in the multivariate Gaussian case, and to the quasi-score equation for GLMs for $n_i = 1, i = 1, \ldots, I$.

M-estimators

Note the connection of GEE and M-estimators. M-estimators are the solutions to estimating equations (which replace the usual score equations), and are called M-estimators because they are "maximum-likelihood-like". (Alternative definitions define them as minimizers of certain functions, which then replace the usual negative log-likelihood.)

There is extensive research on M-estimators and a lot of classical statistics can be shown to be M-estimators, for example moment estimators. They are especially useful in the area of robust statistics, where the estimating functions are chosen to give good properties (usually in terms of bias and efficiency) when the data is truly generated from the assumed probability distribution, but still reasonable properties when true and assumed distribution differ (to some degree). Important work in this area is by Peter Huber, see for example Huber (1981).

The Sandwich estimator for the covariance matrix is also sometimes called the Huber-White estimator, see Huber (1967) and White (Econometrica, 1982).

Generalized Estimating Equations (GEE)

While the generalized estimating equations reduce to the quasi-score equations for $n_i = 1, i = 1, \ldots, I$, they have the additional complication for the multivariate case that $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ depends on the unknown $\boldsymbol{\alpha}$ as well as on $\boldsymbol{\beta}$.

Liang & Zeger, Biometrika, 1986 show that if

- $\boldsymbol{\alpha}$ is replaced by an estimator $\widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)$ that is $\sqrt{I}$-consistent for known $(\boldsymbol{\beta}, \phi)$ (i.e. $\sqrt{I}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = O_p(1) \Leftrightarrow \sqrt{I}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ is bounded in probability)
- ϕ is replaced by an estimator $\widehat{\phi}$ that is $\sqrt{I}$-consistent for known $\boldsymbol{\beta}$

the resulting estimator solving $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \widehat{\phi}(\boldsymbol{\beta}))) = \mathbf{0}$ is asymptotically as efficient as $\widehat{\boldsymbol{\beta}}$ would be for known $\boldsymbol{\alpha}$.

$\widehat{\boldsymbol{\beta}}$ then is consistent and asymptotically normal with mean $\boldsymbol{\beta}$ and covariance matrix

$$\left[\sum_{i=1}^{I} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \boldsymbol{W}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1} \left\{\sum_{i=1}^{I} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \boldsymbol{W}_i^{-1} \mathrm{Var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right\} \left[\sum_{i=1}^{I} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}' \boldsymbol{W}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right]^{-1}.$$

(Note that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ does not depend on the specific choice of $\sqrt{I}$-consistent estimators for $\boldsymbol{\alpha}$ and ϕ.)

Generalized Estimating Equations (GEE)

How to estimate ϕ (if not $\phi = 1$) and $\boldsymbol{\alpha}$ for given $\boldsymbol{\beta}$? Possibilities:

- Estimator for ϕ: $\widehat{\phi} = \frac{1}{n-p} \sum_{i=1}^{I} \sum_{j=1}^{n_i} R_{ij}^2$ using Pearson residuals $R_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$.

  ($\sqrt{I}$-consistent if fourth moments of Y_{ij} finite.)
- Common choices for working correlations and moment based estimators for $\boldsymbol{\alpha}$ (given ϕ), which borrow strength over the I subjects to estimate $\boldsymbol{\alpha}$ consistently:

Structure	$\text{Corr}(Y_{ij}, Y_{ik})$	Estimator for α
Independence	0	NA
Exchangeable	α	$\widehat{\alpha} = \phi \sum_{i=1}^{I} \frac{1}{n_i(n_i-1)} \sum_{j \neq k} R_{ij} R_{ik}$
AR(1)	$\alpha^{\lvert j-k \rvert}$	$\widehat{\alpha} = \phi \sum_{i=1}^{I} \frac{1}{n_i-1} \sum_{j \leq n_i - 1} R_{ij} R_{i,j+1}$
Unstructured	α_{jk}	$\widehat{\alpha}_{jk} = \phi \frac{1}{I} \sum_{i=1}^{I} R_{ij} R_{ik}$

Generalized Estimating Equations (GEE)

Implementation of GEE usually follows an iterative procedure if ϕ and $\boldsymbol{\alpha}$ are estimated by moment based estimators: $\boldsymbol{\beta}^{(0)}$ is initialized e.g. from a GLM assuming independence. Then iterate between

- Compute the Pearson residuals r_{ij}, and obtain moment based estimates for ϕ and $\boldsymbol{\alpha}$.
- Given current estimates $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\phi}$, $\widehat{\boldsymbol{\beta}}$ is updated using Fisher scoring

$$\widehat{\boldsymbol{\beta}}^{(\ell+1)} = \widehat{\boldsymbol{\beta}}^{(\ell)} - \left[\sum_{i=1}^{I}\left(\widehat{\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}}\right)'\widehat{\boldsymbol{W}}_i^{-1}\left(\widehat{\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}}\right)\right]^{-1}\left\{\sum_{i=1}^{I}\left(\widehat{\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}}\right)'\widehat{\boldsymbol{W}}_i^{-1}(\boldsymbol{y}_i - \widehat{\mu}_i)\right\}$$

where $\widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i(\widehat{\boldsymbol{\beta}}^{(\ell)}, \widehat{\boldsymbol{\alpha}}(\widehat{\boldsymbol{\beta}}^{(\ell)}, \widehat{\phi}(\widehat{\boldsymbol{\beta}}^{(\ell)})))$ and $\widehat{\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}}}$ is also evaluated at $\widehat{\boldsymbol{\beta}}^{(\ell)}$.

Generalized Estimating Equations (GEE)

Prentice (Biometrics, 1988) proposed an extension of GEE that estimates $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by simultaneously solving $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}$ and another unbiased estimating equation

$$S_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{I} \left(\frac{\partial \zeta_i}{\partial \boldsymbol{\alpha}} \right)' \boldsymbol{H}_i^{-1} (\boldsymbol{Z}_i - \zeta_i) = \mathbf{0},$$

where $\boldsymbol{Z}_i = (R_{i1}R_{12}, R_{i1}R_{13}, \ldots, R_{in_i-1}R_{1n_i}, R_{i1}^2, \ldots, R_{in_i}^2)'$ contains the products of Pearson residuals $R_{ij} = (Y_{ij} - \mu_{ij})/\sqrt{v(\mu_{ij})}$, and $\zeta_i = \mathrm{E}(\boldsymbol{Z}_i | \boldsymbol{\beta}, \boldsymbol{\alpha}))$.
For binary responses, the last n_i components of $\boldsymbol{Z}_i$ can be left out, as the variance is determined by the mean.

The most suitable weight matrix $\boldsymbol{H}_i$ would be $\mathrm{Var}(\boldsymbol{Z}_i)$, but this would require additional assumptions on the joint third and fourth moments of $\boldsymbol{Y}_i$. Alternatives for $\boldsymbol{H}_i$ will depend on the type of response. Diggle et al, 2002 suggest $\boldsymbol{H}_i = \mathrm{diag}\{\mathrm{Var}(R_{i1}R_{i2}), \ldots, \mathrm{Var}(R_{in_i-1}R_{in_i})\}$ for binary responses, which only depends on $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ (as $Y_{ij}^2 = Y_{ij}$), and the identity matrix for count data. The choice of $\boldsymbol{H}_i$ often does not have a large impact on inference for $\boldsymbol{\beta}$ when I is large.

Generalized Estimating Equations (GEE)

The solution $(\widehat{\beta}, \widehat{\alpha})$ of $S_{\beta}(\beta, \alpha) = \mathbf{0}$ and $S_{\alpha}(\beta, \alpha) = \mathbf{0}$ is asymptotically normally distributed. The covariance matrix can be consistently estimated for $I \to \infty$ by

$$\left(\sum_{i=1}^{I} \boldsymbol{C}_i' \boldsymbol{B}_i^{-1} \boldsymbol{D}_i \right)^{-1} \left\{ \sum_{i=1}^{I} \boldsymbol{C}_i' \boldsymbol{B}_i^{-1} \boldsymbol{V}_{0i} \boldsymbol{B}_i^{-1} \boldsymbol{C}_i \right\} \left(\sum_{i=1}^{I} \boldsymbol{D}_i' \boldsymbol{B}_i^{-1} \boldsymbol{C}_i \right)^{-1},$$

evaluated at $(\widehat{\beta}, \widehat{\alpha})$, where

$$\boldsymbol{C}_i = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} \boldsymbol{W}_i & \mathbf{0} \\ \mathbf{0} & \boldsymbol{H}_i \end{pmatrix}, \quad \boldsymbol{D}_i = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\alpha}} \\ \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\beta}} & \frac{\partial \boldsymbol{\zeta}_i}{\partial \boldsymbol{\alpha}} \end{pmatrix},$$

$$\boldsymbol{V}_{0i} = \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{z}_i - \boldsymbol{\zeta}_i \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{z}_i - \boldsymbol{\zeta}_i \end{pmatrix}'.$$

Idea of proof: Use unbiasedness of estimating equations and Taylor approximation similarly to the usual proof for the asymptotic distribution of MLEs. (Prentice, Biometrics, 1988)

Generalized Estimating Equations (GEE)

Properties of GEEs:

- $\widehat{\beta}$ is often nearly efficient relative to the maximum likelihood estimator if $\text{Var}(\boldsymbol{Y}_i)$ has been reasonably approximated. (For the multivariate Gaussian case, the GEE even reduces to the maximum likelihood score equation when $\text{Var}(\boldsymbol{Y}_i)$ has been correctly specified.)
- $\widehat{\beta}$ is consistent for $I \to \infty$ even when $\text{Var}(\boldsymbol{Y}_i)$ has been incorrectly specified.

Diggle et al, 2002 thus recommend: "When regression coefficients are the scientific focus [. . .] one should invest the lion's share of time in modelling the mean structure, while using a reasonable approximation to the covariance. The robustness of the inferences about β can be checked by fitting a final model using different covariance assumptions and comparing the two sets of estimates and their robust standard errors. If they differ substantially, a more careful treatment of the covariance matrix may be necessary."

A note on missing data

Remember the different missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) (missingness can depend on the observed values), not missing at random (NMAR) (missingness can also depend on the missing values).

- The sandwich variance estimator in general is not a consistent variance estimator under the MAR condition (only under MCAR).
- In general, GEE can lead to inconsistent inferences under MAR because we are not using a likelihood. Several authors (e.g., Robins, Rotnitzky, and others) have proposed a weighted GEE approach (with weights equal to the inverse of the probability of the response being observed) that can be used in this case.
- If there is informative missingness, there is generally no easy solution. More complicated methods (and assumptions about missingness) are required.

More about missing data later.

Inference in Marginal Models

- Inference for β can proceed using the asymptotic normality and the sandwich/robust covariance matrix. This only requires that the mean is correctly specified ($S_\beta(\beta, \alpha)$ is unbiased).
- Inference for α is not advisable, if a working correlation structure was chosen only as a tool to support the (more efficient) estimation of β. α can be interpreted with caution if the model-based and robust standard errors for β are close, as this is some indication that the working correlation has ben chosen close to the true correlation structure. However, even with unstructured correlation matrices (only advisable for smaller n_i), the covariance structure may further depend on covariates, i.e. closeness to the true structure is not guaranteed.
- Formal inference for α is possible using Prentice's extended GEE and the joint asymptotic normality of β and α if one is prepared to believe the specified covariance structure as well ($S_\alpha(\beta, \alpha)$ is unbiased).

Second-order Generalized Estimating Equations (GEE2)

While Prentice's method assumes independence between the two estimating equations $S_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\alpha})$ and $S_{\boldsymbol{\alpha}}(\boldsymbol{\beta},\boldsymbol{\alpha})$, a further extension, GEE2, specifies a single estimating equation.
$\boldsymbol{Y}_i$ and $\boldsymbol{Z}_i$ are combined in a single vector $\boldsymbol{T}_i$ of length $n_i + \binom{n_i}{2}$ with mean $\boldsymbol{\tau}_i = (\boldsymbol{\mu}_i, \boldsymbol{\zeta}_i)$. Let $\boldsymbol{\delta} = (\boldsymbol{\beta},\boldsymbol{\alpha})$. $\boldsymbol{\delta}$ can then be estimated by the solution to the second-order generalized estimating equations

$$S(\boldsymbol{\delta}) = \sum_{i=1}^{I} \left(\frac{\partial \boldsymbol{\tau}_i}{\partial \boldsymbol{\delta}}\right)' \boldsymbol{H}_i^{-1}(\boldsymbol{T}_i - \boldsymbol{\tau}_i) = \mathbf{0}.$$

$\boldsymbol{H}_i$ is chosen as $\mathrm{Var}(\boldsymbol{T}_i)$, which now contains third- and fourth-order moments. However, three-way and four-way correlations are set to zero. $\widehat{\boldsymbol{\delta}}$ can then be obtained using e.g. Fisher scoring.

If first and second order models have been correctly specified, $\widehat{\boldsymbol{\delta}}$ is consistent and asymptotically normal. The covariance matrix can again be estimated consistently using a sandwich estimator.

Higher-order Generalized Estimating Equations

Theoretically, one could use higher-order generalized estimating equations as well. However, this would require the specification of higher order moments, which GEE was supposed to avoid. Generally, either GEE1/GEE2 or full likelihood is thus used.

Marginal Models for Binary Responses

Now that we know how to estimate parameters for a marginal model using GEE, how do we specify sensible marginal models? Consider first the case of binary responses.

A natural simple model would be to add an exchangeable correlation structure to the usual logistic regression model,

- $\text{logit}(\mu_{ij}) = \log(\frac{\mu_{ij}}{1-\mu_{ij}}) = \log\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)}\right) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}$
- $\text{Var}(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$
- $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha, \;\; \forall i, j \neq k.$

Using $\boldsymbol{H}_i = \boldsymbol{I}_{\boldsymbol{n}_i}$ in $S_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}$ then gives

$$\widehat{\alpha} = \frac{1}{N_1} \sum_{i=1}^{I} \sum_{j<k} r_{ij} r_{ik}$$

with $N_1 = \sum_{i=1}^{I} \binom{n_i}{2}$ and $r_{ij} = (y_{ij} - \widehat{\mu}_{ij})/\sqrt{\widehat{\mu}_{ij}(1-\widehat{\mu}_{ij})}$.

Marginal Models for Binary Responses

However, note that the correlation for binary responses must satisfy complicated constraints in μ_{ij} and μ_{ik}:

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{P(Y_{ij}=1, Y_{ik}=1) - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1-\mu_{ij})\mu_{ik}(1-\mu_{ik})}}, \quad \text{where}$$

$$\max(0, \mu_{ij} + \mu_{ik} - 1) < P(Y_{ij}=1, Y_{ik}=1) < \min(\mu_{ij}, \mu_{ik}).$$

Thus, if we model the μ_{ij} as a function of covariates $\boldsymbol{x}$, it will not be reasonable to assume that the correlation is independent of $\boldsymbol{x}$, as we would like to do.

As an alternative, the association is often modeled using the unconstrained odds ratio

$$OR(Y_{ij}, Y_{ik}) = \frac{P(Y_{ij}=1, Y_{ik}=1)P(Y_{ij}=0, Y_{ik}=0)}{P(Y_{ij}=1, Y_{ik}=0)P(Y_{ij}=0, Y_{ik}=1)} \in (0, \infty),$$

where values > 1 indicate positive association.

Marginal Models for Binary Responses

Let $\alpha_{ijk} = \log(OR(Y_{ij}, Y_{ik}))$ (thus, $\alpha_{ijk} \in \mathbb{R}$) and $\mu_{ijk} = P(Y_{ij} = Y_{ik} = 1)$. Then,

$$\text{logit}\, P(Y_{ij} = 1 | Y_{ik} = y_{ik}) = \alpha_{ijk} y_{ik} + \log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right)$$

($\rightarrow$ blackboard). Now, if we assume the model $\alpha_{ijk} = \boldsymbol{d}'_{ijk}\boldsymbol{\alpha}$, where $\boldsymbol{d}_{ijk}$ is a vector of covariates, we can estimate $\boldsymbol{\alpha}$ from a logistic regression of Y_{ij} on the product $\boldsymbol{d}_{ijk} y_{ik}$. In particular, if we assume $\alpha_{ijk} \equiv \alpha$, we can estimate α by a logistic regression of Y_{ij} on $y_{ik}, 1 \le j < k \le n_i, i = 1, \ldots, I$.

The offset $\log\left(\frac{\mu_{ij} - \mu_{ijk}}{1 - \mu_{ij} - \mu_{ik} + \mu_{ijk}}\right)$ in this logistic regression depends on both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, and iteration between estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is thus necessary.

This approach to GEE for binary responses was proposed by Carey, Zeger & Diggle (Biometrika, 1993) and called alternating logistic regression (ALR). ALR has been implemented in the R package `orth` and in SAS `proc genmod`.

Marginal Models for Binary Responses

There are many other approaches to either model the joint probabilities $P(Y_{i1}, \ldots, Y_{in_i})$ and use the full likelihood, or to model the pairwise correlations and use GEE.

For examples, see Diggle et al (2002) and Prentice (Biometrics, 1988), and references therein.

Marginal Models for Counted Responses

Count data often shows overdispersion, i.e. the variance exceeds the mean, whereas a Poisson model would assume equality. A marginal models that accounts for overdispersion would be

- $\log \mathrm{E}(Y_{ij}) = \log(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta}$, where $\log(t_{ij})$ is an offset if not $t_{ij} = 1\ \forall i, j$
- $\mathrm{Var}(Y_{ij}) = \phi_{ij}\,\mathrm{E}(Y_{ij})$, where $\phi_{ij} = \phi(\boldsymbol{\alpha}_1)$
- $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\boldsymbol{\alpha}_2)$.

Here, the overdispersion parameter ϕ_{ij} is allowed to depend on covariates, e.g. $\phi_{ij} = \phi_1$ for one treatment group and $\phi_{ij} = \phi_2$ for the other, or ϕ_{ij} differing by time point.

Then, letting $\boldsymbol{\alpha} = (\boldsymbol{\alpha}'_1, \boldsymbol{\alpha}'_2)'$, $S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}$ and $S_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{0}$ can be solved simultaneously, where $\boldsymbol{Z}_i$ again includes the squared terms R^2_{ij} for estimation of the overdispersion parameters. Diggle et al, 2002 further suggest to use $\boldsymbol{H}_i = \boldsymbol{I}_{\boldsymbol{n}_i}$ to avoid dependence of $\boldsymbol{H}_i$ on higher-order parameters.

Marginal Models for Counted Responses

Again, there are many different approaches to model count data. For example, Solis-Trapala & Farewell (StatMed, 2005) suggest using a model where $\boldsymbol{Y}_i$ follows a multivariate negative binomial model for overdispersed correlated count data. This results in

- $\mathrm{E}(Y_{ij}) = \mu_{ij} = \exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})$
- $\mathrm{Var}(Y_{ij}) = \mu_{ij} + \phi\mu_{ij}^2$
- $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \phi\mu_{ij}\mu_{ik},\ j \neq k,$

and the authors again estimate the parameters using GEE and obtain standard errors for $\widehat{\boldsymbol{\beta}}$ via the sandwich estimator.

Marginal Models in R

See `examples_marginal_models.R` for example code for

- a marginal model for binary data fit via the ALR approach to GEE
- a marginal model for count data fit via GEE
- a comparison to a generalized linear mixed model for count data.

(This is the seizure data with counted responses revisited.)

Note: This example illustrates an important special case: In a log-linear mixed model, where $\boldsymbol{z}_{ij}$ contains a subset of variables in $\boldsymbol{x}_{ij}$, the β parameters for the variables in $\boldsymbol{x}_{ij}$ that are **not** in $\boldsymbol{z}_{ij}$ will have the same interpretation as in a marginal model. $\longrightarrow$ blackboard