Quasi-Likelihood

We now describe a method for inference, generalized estimating equations, that attempts to make minimal assumptions about the data-generating process.

We begin with a recap of the related quasi-likelihood procedure, which is an alternative to MLE, when we do not wish to commit to specifying the full distribution of the data and we can assume independent data. The resultant estimators are known as quasi-MLE (QMLE).

The approach is based on specifying the first two moments of the data only, and assuming they take the form:

$$\begin{aligned} \mathrm{E}[\boldsymbol{Y} \mid \boldsymbol{\beta}] &= \boldsymbol{\mu}(\boldsymbol{\beta}) \\ \mathrm{cov}(\boldsymbol{Y} \mid \boldsymbol{\beta}) &= \alpha \boldsymbol{V}\{\boldsymbol{\mu}(\boldsymbol{\beta})\} \end{aligned}$$

where $\boldsymbol{\mu}(\boldsymbol{\beta}) = [\mu_1(\boldsymbol{\beta}), ..., \mu_n(\boldsymbol{\beta})]^{\mathrm{T}}$ represents the regression function and $\boldsymbol{V}$ is a diagonal matrix (so the observations are uncorrelated), with

$$\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V\{\mu_i(\boldsymbol{\beta})\},$$

and $\alpha > 0$ a scalar which is independent of $\boldsymbol{\beta}$.

Consider the sum of squares

$$(\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha, \tag{30}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ and $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\beta})$. To minimize this sum of squares there are two ways to proceed.

First approach: differentiate and obtain

$$-2\boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha + (\boldsymbol{Y} - \boldsymbol{\mu})^{\mathrm{T}} \frac{\partial \boldsymbol{V}^{-1}}{\partial \boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{\mu})/\alpha,$$

where $\boldsymbol{D}$ is the $n \times p$ matrix of derivatives with elements $\partial \mu_i / \partial \beta_j$, $i = 1, ..., n; j = 1, ..., p$. Unfortunately the expectation of this expression is not zero, and so an inconsistent estimator of $\boldsymbol{\beta}$ will result.

Second approach: pretend $\boldsymbol{V}$ is not a function of $\boldsymbol{\beta}$, so that $\widehat{\boldsymbol{\beta}}$ is the root of:

$$\boldsymbol{D}(\widehat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{V}(\widehat{\boldsymbol{\beta}})^{-1} \{\boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\}/\alpha = \boldsymbol{0}.$$

As shorthand we write this estimating function as

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\{\boldsymbol{Y}-\boldsymbol{\mu}\}/\alpha. \tag{31}$$

This estimating function is linear in the data and so its properties are straightforward to evaluate. In particular:

1. $\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta})] = \boldsymbol{0}$.
2. $\mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$.
3. $-\mathrm{E}\left[\frac{\partial U}{\partial \beta}\right] = \mathrm{cov}\{\boldsymbol{U}(\boldsymbol{\beta})\} = \boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D}/\alpha$.

Applying the earlier result on properties of estimators arising from estimating functions:

$$(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{1/2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \alpha\boldsymbol{I}_p),$$

where we have so far assumed that α is known.

Since the root of (31) does not depend on α, $\widehat{\boldsymbol{\beta}}$ is consistent regardless. For appropriate standard errors we require an estimator of α however.

Unknown α

Since

$$\mathrm{E}[(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})] = n\alpha,$$

an unbiased estimator of α would be

$$\widehat{\alpha} = (\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{\mu})(\boldsymbol{Y}-\boldsymbol{\mu})/n,$$

a degrees of freedom corrected (but not in general, unbiased) estimate is given by the Pearson statistic divided by its degrees of freedom:

$$\widehat{\alpha} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i-\widehat{\mu}_i)^2}{V(\widehat{\mu}_i)},$$

where $\widehat{\mu}_i = \widehat{\mu}_i(\widehat{\boldsymbol{\beta}})$.

The asymptotic distribution that is used in practice is therefore given by

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\widehat{\boldsymbol{D}}/\widehat{\alpha})^{1/2}(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \boldsymbol{I}_p),$$

In general we may use sandwich estimation with quasi-likelihood. We have

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\mathrm{var}(\boldsymbol{Y})\boldsymbol{V}^{-1}\boldsymbol{D}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}\alpha^2,$$

and $\mathrm{var}(\boldsymbol{Y})$ may be estimated by the diagonal matrix with elements $(Y_i-\widehat{\mu}_i)^2$.

Why “Quasi”

Integration of the quasi-score (31) gives

$$l(\mu, \alpha) = \int_y^\mu \frac{y - t}{\alpha V(t)} \mathrm{d}t$$

which, if it exists, behaves like a log-likelihood. As an example, for the model $\mathrm{E}[Y] = \mu$ and $\mathrm{var}(Y) = \alpha\mu$ we have

$$l(\mu, \alpha) = \int_y^\mu \frac{y - t}{\alpha t} \mathrm{d}t = \frac{1}{\alpha}[y \log \mu - \mu + c],$$

where $c = -y \log y - y$ and $y \log \mu - \mu$ is the log likelihood of a Poisson random variable.

The word “quasi” refers to the fact that the score may or not correspond to a probability function.

For example, the variance function $\mu^2(1 - \mu)^2$ does not correspond to a probability distribution.

Example: Air Pollution Data

We examine the association between daily mortality, Y_i, and the daily value of PM_{10} (particulate matter less than 10 micrometers, which is about 0.0004 inches, in diameter), x_i, with $i = 1, ..., 335$, indexing the 335 days on which there are no missing PM_{10} is to be investigated.

Figure 10 shows the association between log daily counts and PM_{10}.

Assume the model

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}] = \exp(\boldsymbol{x}_i \boldsymbol{\beta}), \quad \mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha E[Y_i \mid \boldsymbol{\beta}].$$

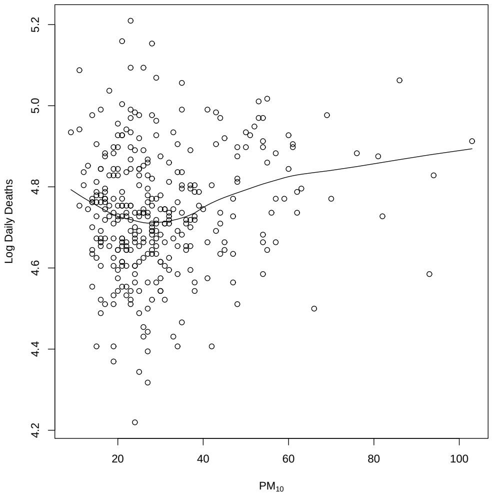

Figure 10: Log daily deaths versus PM_{10}.

Fitting the quasi-likelihood model yields $\widehat{\boldsymbol{\beta}} = (4.71, 0.0015)^{\text{T}}$ and $\widehat{\alpha} = 2.77$ so that the quasi-likelihood standard errors are $\sqrt{\widehat{\alpha}} = 1.67$ times larger than the Poisson model-based standard errors.

The variance-covariance matrix is given by

$$(\widehat{\boldsymbol{D}}^{\text{T}} \widehat{\boldsymbol{V}}^{-1} \widehat{\boldsymbol{D}})^{-1} \widehat{\alpha} = \begin{bmatrix} 0.019^2 & \star \\ -0.89 \times 0.019 \times 0.00056 & 0.00056^2 \end{bmatrix}.$$

Standard errors of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are 0.019 and 0.00056.

Asymptotic 95% confidence interval for β_1 is given by $(0.00040, 0.0026)$.

A more useful summary is a confidence interval for the relative risk associated with a 10-unit increase in PM_{10}, which is

$$(e^{0.00040 \times 10}, e^{0.026 \times 10}) = (1.004, 1.026)$$

so that the interval suggests that the increase in daily mortality associated with a 10-unit increase in PM_{10} is between 0.4% and 2.6%.

Extension to Quasi-Likelihood

Suppose we have

$$\begin{aligned} \mathrm{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \mathrm{var}(Y_i \mid \boldsymbol{\beta}) &= V_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{aligned}$$

where $\boldsymbol{\alpha}$ is a $k \times 1$ vector of parameters that appear only in the variance model.

Previously, in quasi-likelihood method, we had "separable" mean and variance models, that is, $\mathrm{var}(Y_i \mid \boldsymbol{\beta}) = \alpha V_i(\mu_i)$ (which is why we obtained a consistent estimator even if the form of the variance was wrong).

Let $\widehat{\boldsymbol{\alpha}}_n$ be a consistent estimator of $\boldsymbol{\alpha}$. We state without proof the following result. The estimator $\widehat{\boldsymbol{\beta}}_n$ that satisfies

$$\boldsymbol{G}(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n) = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)^{\mathrm{T}} \boldsymbol{V}^{-1}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n) \left\{ \boldsymbol{Y} - \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}}_n) \right\} \tag{32}$$

has asymptotic distribution

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}}^{1/2} \widehat{\boldsymbol{D}})^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ N_p(\boldsymbol{0}, \boldsymbol{I}_p) \tag{33}$$

where $\widehat{\boldsymbol{D}} = \boldsymbol{D}(\widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$. Sandwich estimation may be used to obtain empirical standard errors which are correct even if the variance model is wrong, so long as we have a consistent estimator of α.

Computation

Previously we assumed $\mathrm{var}(Y_i) = \alpha V_i(\mu_i)$, and the estimating function did not depend on α and so, correspondingly, $\widehat{\boldsymbol{\beta}}$ did not depend on α, though the standard errors did.

In general iteration is convenient to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate.

Then set $j = 0$ and iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(j)}) = \boldsymbol{0}$ to give $\widehat{\boldsymbol{\beta}}^{(j+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(j+1)}$ with $\widehat{\mu}_i = \mu_i \left(\widehat{\boldsymbol{\beta}}^{(j+1)} \right)$. Set $j \rightarrow j + 1$ and return to 1.

Example: Air Pollution Data

Consider the random effects formulation:

$$\text{E}[Y_i \mid \boldsymbol{\beta}, \theta_i] = \text{var}(Y_i \mid \boldsymbol{\beta}, \theta_i) = \mu_i(\boldsymbol{\beta})\theta_i \tag{34}$$

with

$$\text{E}[\theta_i] = 1, \quad \text{var}(\theta_i) = 1/\alpha. \tag{35}$$

Assuming $\theta_i \sim_{iid} \text{Ga}(\alpha, \alpha)$, we could derive the marginal distribution of the data (which is negative binomial) and proceed with likelihood.

As an alternative we consider the model

$$\begin{aligned} \text{E}[Y_i \mid \boldsymbol{\beta}] &= \mu_i(\boldsymbol{\beta}) \\ \text{var}(Y_i \mid \alpha, \boldsymbol{\beta}) &= \mu_i(\boldsymbol{\beta})\{1 + \mu_i(\boldsymbol{\beta})/\alpha\}. \end{aligned} \tag{36}$$

that are the marginal first two moments of the data given (34) and (35).

The form (36) suggests the estimating function for $\boldsymbol{\beta}$ (with α assumed known):

$$\sum_{i=1}^{n} \boldsymbol{D}(\boldsymbol{\beta})_i^{\text{T}} \boldsymbol{V}_i^{-1}(\alpha, \boldsymbol{\beta})\{y_i - \mu_i(\boldsymbol{\beta})\}$$

For a fixed α we can solve this estimating equation to obtain an estimator $\widehat{\boldsymbol{\beta}}$.

We describe a method-of-moments estimator for α for the *quadratic* variance model we have

$$\text{var}(Y_i \mid \boldsymbol{\beta}, \alpha) = \text{E}[(Y_i - \mu_i)^2] = \mu_i(1 + \mu_i/\alpha),$$

and so

$$\alpha^{-1} = \text{E}\left[\frac{(Y_i - \mu_i)^2 - \mu_i}{\mu_i^2}\right],$$

$i = 1, ..., n$, leading to the method-of-moments estimator

$$\widehat{\alpha} = \left\{ \frac{1}{n-p} \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2 - \widehat{\mu}_i}{\widehat{\mu}_i^2} \right\}^{-1}. \tag{37}$$

If we have a consistent estimator $\widehat{\alpha}$, and the mean is correctly specified then valid inference follows from

$$(\widehat{\boldsymbol{D}}^{\mathrm{T}}\widehat{\boldsymbol{V}}(\widehat{\alpha})^{-1}\widehat{\boldsymbol{D}})^{1/2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \rightarrow_d N(\boldsymbol{0},\boldsymbol{I}_p).$$

We fit this model to the air pollution data.

The estimates (standard errors) are $\widehat{\beta}_0 = 4.71$ (0.018) and $\widehat{\beta}_1 = 0.0014$ (0.00056).

The moment-based estimator is $\widehat{\alpha} = 65.20$.

This analysis therefore produces virtually identical inference with the quasi-likelihood approach in which the variance was a linear function of the mean.

In Figure 11 we plot the linear and quadratic variance functions (over the range of the mean for these data) and we see that they are very similar.

Examination of the residuals did not clearly indicate the superiority of either variance model; it is typically very difficult to distinguish between the two models, unless the mean of the data has a large spread.

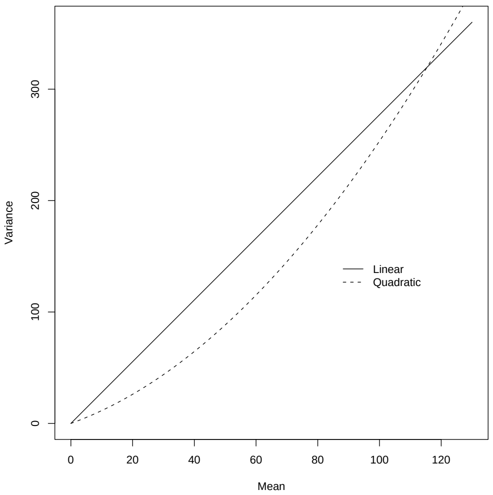

Figure 11: Linear and quadratic variance functions for the air pollution data.

Example: Rcode for Quasi-Poisson Regression

We run the Poisson regression and then evaluate the method-of-moments estimator of α “by hand”.

```
> mod1 <- glm(ynew˜x1new,family=poisson)
> summary(mod1)
Coefficients:
                  Value   Std. Error   t value
(Intercept) 4.705062304 0.0113962988 412.85880
      x1new 0.001458115 0.0003348748   4.35421
(Dispersion Parameter for Poisson family taken to be 1 )
    Null Deviance: 927.372 on 334 degrees of freedom
Residual Deviance: 908.6531 on 333 degrees of freedom
Number of Fisher Scoring Iterations: 3
Correlation of Coefficients:
      (Intercept)
x1new -0.8949913
> resid1 <- (ynew - mod1$fit)/sqrt(mod1$fit)
> alphahat <- sum(resid1 * resid1)/(length(ynew) - 2)
> alphahat
[1] 2.772861
```

We now fit the Quasi-Likelihood model with

$$\mathrm{E}[Y_i|\boldsymbol{\beta}] = \mu_i = \exp(\beta_0 + \beta_1 x_i)$$

and

$$\mathrm{var}(Y_i|\boldsymbol{\beta}) = \alpha\mu_i = \alpha \exp(\beta_0 + \beta_1 x_i).$$

```
> mod2 <- glm(ynew˜x1new,quasi(link=log,variance=mu))
> summary(mod2)
Coefficients:
                  Value  Std. Error    t value
(Intercept) 4.705062304 0.018976351 247.943468
      x1new 0.001458115 0.000557611   2.614932
(Dispersion Parameter for Quasi-likelihood
    family taken to be 2.772667 )
    Null Deviance: 927.372 on 334 degrees of freedom
Residual Deviance: 908.6531 on 333 degrees of freedom
Number of Fisher Scoring Iterations: 3
Correlation of Coefficients:
      (Intercept)
x1new -0.8949913
```

The standard errors are multiplied by $\sqrt{\widehat{\alpha}}$ (=1.67 here), but the estimates are unchanged.

Example: Rcode for Quadratic Variance Model

The `glm.nb` function carries out MLE for the negative binomial model (it is part of the `MASS` library).

We find the MLE of α, and then use this as a starting value for the iterative strategy in which a method-of-moments estimator is used.

```
> library(MASS)
> modnegbinmle <- glm.nb(y~x)
> summary(modnegbinmle)
Call:
glm.nb(formula = y ~ x, init.theta = 67.7145, link = log)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.7055974  0.0188269 249.941   <2e-16 ***
x           0.0014405  0.0005577   2.583   0.0098 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Correlation of Coefficients:
  (Intercept)
x -0.90
              Theta:  67.71
          Std. Err.:  8.27
> alphahat <- 67.61
```

Now iterate to a solution by estimating $\boldsymbol{\beta}$ for fixed α, and then re-estimating α.

```
> alphanew <- 0
> counter <- 0
> for (i in 1:5){
        fit <- glm(y~x,family=negative.binomial(alphahat))
        mu <- fit$fitted
        alphanew <- 1/(sum(((y-mu)^2-mu)/mu^2)/(length(y)-2))
        alphahat <- alphanew
        cat("Iteration ",i,alphahat,"\n")
}
Iteration  1 65.19642
Iteration  2 65.19649
Iteration  3 65.19649
Iteration  4 65.19649
Iteration  5 65.19649
> summary(fit)
Call:
glm(formula = y ~ x, family = negative.binomial(alphahat))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.705605   0.019071 246.747   <2e-16 ***
x           0.001440   0.000565   2.549   0.0112 *
(Dispersion parameter for Negative Binomial(65.1965)
family taken to be 1.001560)
```

Generalized Estimating Equations

Suppose we assume

$$\mathrm{E}[\boldsymbol{Y}_i \mid \boldsymbol{\beta}] = \boldsymbol{x}_i\boldsymbol{\beta},$$

and consider the $n_i \times n_i$ *working* variance-covariance matrix:

$$\mathrm{var}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) = \boldsymbol{W}_i.$$

To motivate GEE we begin by assuming that $\boldsymbol{W}_i$ is known. In this case the GLS estimator minimizes

$$\sum_{i=1}^m (\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

and is given by the solution to the estimating function

$$\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

which is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{Y}_i.$$

We now examine the properties of this estimator.

We have

$$\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{\beta},$$

so long as the mean is correctly specified.

If the information about $\boldsymbol{\beta}$ grows with increasing m, then $\widehat{\boldsymbol{\beta}}$ is consistent.

The variance, $\mathrm{var}(\widehat{\boldsymbol{\beta}})$, is given by

$$\left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}\left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\mathrm{var}(\boldsymbol{Y}_i)\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)\left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1}.$$

If the assumed variance-covariance matrix is correct, i.e. $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i$, then

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^m \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}\boldsymbol{x}_i\right)^{-1},$$

and a Gauss-Markov Theorem shows that, in this case, the estimator is efficient amongst linear estimators.

If m is large then a multivariate central limit theorem shows that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal.

We now suppose that $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is of so that $\boldsymbol{\alpha}$ are parameters in the variance-covariance model. The regression parameters are contained in $\boldsymbol{W}_i$ to allow, mean-variance relationships, e.g.

$$\begin{aligned} \text{var}(Y_{ij} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \alpha_1 \mu_{ij}^2 \\ \text{cov}(Y_{ij}, Y_{ik} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \alpha_1 \alpha_2^{|t_{ij} - t_{ik}|} \mu_{ij} \mu_{ik} \end{aligned}$$

where

- $\mu_{ij} = \boldsymbol{x}_{ij} \boldsymbol{\beta}$,
- α_1 is the variance (which is assumed constant across time and across individuals), and
- α_2 is the correlation (which is assumed to be the same for all individuals), and
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$.

For known $\boldsymbol{\alpha}$ we would minimize

$$\sum_{i=1}^{m} (\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta})^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}),$$

with solution given by the root of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}) = \boldsymbol{0}.$$

In general the roots of this equation are not available in closed form (because $\boldsymbol{\beta}$ occurs in $\boldsymbol{W}$).

However, if $\boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{W}_i(\boldsymbol{\alpha})$ we have

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{x}_i \right)^{-1} \sum_{i=1}^{m} \boldsymbol{x}_i^{\text{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}) \boldsymbol{Y}_i.$$

Finally, suppose that $\boldsymbol{\alpha}$ is unknown but we have a method by which a consistent estimator $\widehat{\boldsymbol{\alpha}}$ is produced (e.g. method of moments).

We then solve the estimator function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}).$$

In general iteration is needed to simultaneously estimate $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Let $\widehat{\boldsymbol{\alpha}}^{(0)}$ be an initial estimate, then set $t = 0$ and iterate between

1. Solve $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(t)}) = \boldsymbol{0}$ to give $\widehat{\boldsymbol{\beta}}^{(t+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(t+1)}$ with $\widehat{\mu}_i = \mu_i\left(\widehat{\boldsymbol{\beta}}^{(t+1)}\right)$. Set $t \rightarrow t+1$ and return to 1.

We have

$$\mathrm{var}(\widehat{\boldsymbol{\beta}})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathrm{N}_{k+1}\left(\boldsymbol{0}, \boldsymbol{I}\right),$$

where

$$\begin{aligned} \widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) &= \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right)^{-1} \\ &\times \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\mathrm{var}(\boldsymbol{Y}_i)\boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right) \\ &\times \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}})\boldsymbol{x}_i\right)^{-1}. \end{aligned}$$

We have assumed that $\mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = 0$ for $i \neq i'$, and this is required for the asymptotic distribution to be appropriate.

The final element of GEE is sandwich estimation of $\text{var}(\widehat{\boldsymbol{\beta}})$. In particular $\text{cov}(\boldsymbol{Y}_i)$ is estimated by

$$(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^{\mathrm{T}},$$

may be multiplied by $N/(N-p)$ to account for estimation of $\boldsymbol{\beta}$ ($N = \sum_i n_i$).

Empirical would be a better word than *robust* (which is sometimes used) for the estimator of the variance – not robust to sample size, in fact could be highly unstable.

We can write the $(k+1) \times 1$ estimating function as

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta})$$

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{W}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}[\boldsymbol{x}_{i1}\cdots\boldsymbol{x}_{in_i}]\begin{bmatrix} W_i^{11} & \cdots & W_i^{1n_i} \\ \cdots & \cdots & \cdots \\ W_i^{n_i1} & \cdots & W_i^{n_in_i}\end{bmatrix}\begin{bmatrix} Y_{i1} - \boldsymbol{x}_{i1}\boldsymbol{\beta} \\ \cdots \\ Y_{in_i} - \boldsymbol{x}_{in_i}\boldsymbol{\beta}\end{bmatrix}$$

where W_i^{ij} denotes entry (i,j) of the inverse $\boldsymbol{W}_i$. We use the middle form since this emphasizes that the basic unit of replication is indexed by i.

Example: Suppose for simplicity that we have a balanced design, with $n_i = n$ for all i, and assume a working variance-covariance matrix with

$$\begin{aligned} \text{var}(Y_{ij}) &= \mathrm{E}[(Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta})^2] = \mathrm{E}[\epsilon_{ij}^2] = \alpha_1 \\ \text{cov}(Y_{ij}, Y_{ik}) &= \mathrm{E}[(Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta})(Y_{ik} - \boldsymbol{x}_{ik}\boldsymbol{\beta})] = \mathrm{E}[\epsilon_{ij}\epsilon_{ik}] = \alpha_1\alpha_{2jk}, \end{aligned}$$

for $i = 1, ..., m; j, k = 1, ..., n; j \neq k$. Hence we have $n + n(n-1)/2$ elements of $\boldsymbol{\alpha}$.

Letting

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}},$$

method-of-moments estimators are given by

$$\widehat{\alpha}_1 = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}^2,$$

and

$$\widehat{\alpha}_1\widehat{\alpha}_{2jk} = \frac{1}{m}\sum_{i=1}^{m} e_{ij}e_{ik}.$$

Generalized Estimating Equation (GEE) Summary

We have:

- Regression parameters (of primary interest) $\boldsymbol{\beta}$ and,
- Variance-covariance parameters $\boldsymbol{\alpha}$.

We have considered the GEE

$$\boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0},$$

where

- $\mu_i = \mu_i(\boldsymbol{\beta}) = \boldsymbol{x}_i \boldsymbol{\beta}$.
- $\boldsymbol{D}_i = \boldsymbol{D}_i(\boldsymbol{\beta}) = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i^{\mathrm{T}}$,
- $\boldsymbol{W}_i = \boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the “working” covariance model,

Three important ideas:

1. Separate estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.
2. Sandwich estimation of $\text{var}(\widehat{\boldsymbol{\beta}})$.
3. Replication across units in order to estimate covariances – so we have assumed that observations on different units are independent.

Notes:

- We have seen the first and second ideas in independent data situations – e.g. estimation of the α parameter in the quadratic negative binomial model.
- We may use method of moments estimators for $\boldsymbol{\alpha}$ (or set up another estimating equation, see later).
- We could go with model-based standard errors:

$$\text{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{D}_i \right)^{-1}. \qquad (38)$$

If we have an independence working model ($\boldsymbol{W}_i = \boldsymbol{I}$) then no iteration necessary (since no $\boldsymbol{\alpha}$ in the GEE) – in this case we'd want to use sandwich estimation, however.

Dental Example

Look at various estimators of $\boldsymbol{\beta}$ for girls only. Note here that we might question the asymptotics for GEE since we only have replication across $m = 11$ units (girls) (check with simulation – see coursework).

Start with ordinary least squares – unbiased estimator for $\boldsymbol{\beta}$, but standard errors are wrong because independence is assumed.

```
> summary(lm(distance~age,data=Orthgirl))

Call:
lm(formula = distance ~ age, data = Orthgirl)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3727     1.6378  10.608 1.87e-13 ***
age           0.4795     0.1459   3.287  0.00205 **
Residual standard error: 2.164 on 42 degrees of freedom
Multiple R-Squared: 0.2046,     Adjusted R-squared: 0.1856
F-statistic:  10.8 on 1 and 42 DF,  p-value: 0.002053
```

Now implement GEE with working independence – the following is an `R` implementation.

```
> library(nlme); data(Orthodont); Orthgirl <- Orthodont[Orthodont$Sex=="Female",]
> install.packages("geepack")
> library(geepack)
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="independence"))
Call:
geese(formula = distance ~ age, id = Subject, data = Orthgirl,corstr = "independence")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate    san.se      wald            p
(Intercept) 17.3727273 0.7819784 493.56737 0.000000e+00
age          0.4795455 0.0666386  51.78547 6.190604e-13
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 4.470403 1.373115 10.59936 0.001131270
Correlation Model:
 Correlation Structure:     independence
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```

Next we examine an exchangeable correlation structure in which all pairs of observations on the same unit have a common correlation:

```
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="exchangeable"))
geese(formula = distance ~ age, id = Subject, data = Orthgirl,
    corstr = "exchangeable")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate    san.se      wald            p
(Intercept) 17.3727273 0.7819784 493.56737 0.000000e+00
age          0.4795455 0.0666386  51.78547 6.190604e-13
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 4.470403 1.373115 10.59936 0.001131270
Correlation Model:
 Correlation Structure:     exchangeable
 Correlation Link:          identity
 Estimated Correlation Parameters:
       estimate    san.se     wald            p
alpha 0.8680178 0.1139327 58.04444 2.564615e-14
Number of clusters:   11   Maximum cluster size: 4
```

Notes:

- Independence estimates are always identical to OLS because we have assumed working independence, which means that the estimating equation is the same as the normal equations.
- Standard error for β_1 is smaller with GEE because regressor (time) is changing within an individual.
- Here we obtain the same estimates for exchangeable as working independence but only because balanced and complete (i.e. no missing) data.

Finally we look at AR(1) and unstructured errors – this time we see slight differences in estimates and standard errors.

```
> summary(geese(distance~age,id=Subject,data=Orthgirl,corstr="ar1"))
geese(formula = distance ~ age, id = Subject, data = Orthgirl, corstr = "ar1")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate     san.se      wald            p
(Intercept) 17.3049830 0.85201953 412.51833 0.000000e+00
age          0.4848065 0.06881228  49.63692 1.849965e-12
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se   wald            p
(Intercept) 4.470639 1.341802 11.101 0.0008628115
Correlation Model:
 Correlation Structure:     ar1
 Correlation Link:          identity
 Estimated Correlation Parameters:
       estimate     san.se     wald p
alpha 0.9298023 0.07164198 168.4403 0
Number of clusters:   11   Maximum cluster size: 4
```

Now delete last two observations from girl 11 to illustrate that identical answers before were consequence of balance and completeness of data.

```
> Orthgirl2<-Orthgirl[1:42,]
> summary(lm(distance~age,data=Orthgirl2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.0713     1.5102  11.966 8.56e-15 ***
age           0.3963     0.1357   2.921  0.00571 **
Residual standard error: 1.964 on 40 degrees of freedom
> summary(geese(distance~age,id=Subject,data=Orthgirl2,
corstr="independence"))
Coefficients:
              estimate     san.se      wald            p
(Intercept) 18.0713312 0.82603439 478.61250 0.000000e+00
age          0.3962971 0.06934195  32.66253 1.096304e-08
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se     wald           p
(Intercept) 3.674926 1.317669 7.778294 0.005287771
Correlation Model:
 Correlation Structure:     independence
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```

```
> summary(geese(distance~age,id=Subject,data=Orthgirl2,corstr="exchangeable"))
Call:
geese(formula = distance ~ age, id = Subject, data = Orthgirl2,
    corstr = "exchangeable")
Mean Model:
 Mean Link:                 identity
 Variance to Mean Relation: gaussian
 Coefficients:
              estimate     san.se      wald            p
(Intercept) 17.6050097 0.79007168 496.52320 0.000000e+00
age          0.4510122 0.06641218  46.11913 1.112765e-11
Scale Model:
 Scale Link:                identity
 Estimated Scale Parameters:
            estimate   san.se    wald           p
(Intercept) 3.706854 1.320019 7.88589 0.004982194
Correlation Model:
 Correlation Structure:     exchangeable
 Correlation Link:          identity
 Estimated Correlation Parameters:
       estimate     san.se     wald p
alpha 0.7968515 0.09367467 72.36198 0
Returned Error Value:    0
Number of clusters:   11   Maximum cluster size: 4
```

Comparison of Analyses

In Table 7 summaries are presented under likelihood, Bayesian and GEE analyses.

Two Bayesian models were fitted, a normal model:

$$
\begin{aligned}
\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} &\sim_{iid} \quad \text{N}(\boldsymbol{\beta}, \boldsymbol{D}), \quad \text{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D}) = \boldsymbol{D} \\
\boldsymbol{D}^{-1} &\sim \quad \text{W}(r, \boldsymbol{R}^{-1}), \quad \text{E}[\text{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D})] = \frac{\boldsymbol{R}}{r-3} \\
\boldsymbol{R} &= \quad \begin{bmatrix} 1.0 & 0 \\ 0 & 0.1 \end{bmatrix}, \quad r = 4
\end{aligned}
$$

and a Student t_4 model:

$$
\begin{aligned}
\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D} &\sim_{iid} \quad \text{St}_4(\boldsymbol{\beta}, \boldsymbol{D}), \quad \text{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D}) = 2\boldsymbol{D} \\
\boldsymbol{D}^{-1} &\sim \quad \text{W}(r, \boldsymbol{R}_t^{-1}), \quad \text{E}[\text{var}(\boldsymbol{\beta}_i \mid \boldsymbol{\beta}, \boldsymbol{D})] = 2\frac{\boldsymbol{R}_t}{r-3} \\
\boldsymbol{R}_t &= \quad \begin{bmatrix} 0.5 & 0 \\ 0 & 0.05 \end{bmatrix}, \quad r = 4
\end{aligned}
$$

Approach	$\widehat{\beta}_0$	s.e.($\widehat{\beta}_0$)	$\widehat{\beta}_1$	s.e.($\widehat{\beta}_1$)
LMEM ML	22.65	0.62	0.480	0.065
LMEM REML	22.65	0.63	0.479	0.066
Bayes Normal	22.65	0.60	0.479	0.075
Bayes t_4	22.65	0.58	0.475	0.073
GEE Independence	22.65	0.55	0.480	0.067
GEE AR(1)	22.64	0.58	0.485	0.069

Table 7: Summaries for fixed effects.

- Overall, the analyses are in good correspondence.

Approach	$\widehat{\text{var}}(\beta_{0i})$	$\widehat{\text{var}}(\beta_{1i})$	$\widehat{\text{corr}}(\beta_{0i}, \beta_{1i})$	$\widehat{\sigma}_\epsilon$
LMEM ML	1.98	0.15	0.55	0.67
LMEM REML	2.08	0.16	0.53	0.67
Bayes Normal	1.93 (1.29,2.96)	0.18 (0.10,0.31)	0.39 (-0.32,0.85)	0.70 (0.52,0.93)
Bayes t_4	2.06 (1.18,3.46)	0.20 (0.11,0.35)	0.42 (-0.34,0.88)	0.71 (0.54,0.95)

Table 8: Summaries for variance components.

GEE with working independence gives $\alpha_1 = 4.47$.

GEE with working AR(1) gives $\alpha_1 = 4.47$, $\alpha_2 = 0.93$.

The parameterization adopted for the linear model changes the interpretation of $\boldsymbol{D}$. For example:

Model 1: $(\beta_0 + b_{0i}) + (\beta_1 + b_{1i})t_j$, $\boldsymbol{b}_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$.

Model 2: $(\gamma_0 + b^\star_{0i}) + (\gamma_1 + b^\star_{1i})(t_j - \bar{t})$, $\boldsymbol{b}^\star_i \sim \text{N}(\boldsymbol{0}, \boldsymbol{D}^\star)$.

Giving $\beta_0 = \gamma_0 - \gamma_1\bar{t}$, $\beta_1 = \gamma_1$.

$b_{0i} = b^\star_{0i} - \bar{t}b^\star_{1i}$, $b_{1i} = b^\star_{1i}$.

Moral: $\boldsymbol{D} \neq \boldsymbol{D}^\star$; $D_{00} = D^\star_{00} - 2\bar{t}D^\star_{01} + \bar{t}^2 D^\star_{11}$, $D_{01} = D^\star_{01} - \bar{t}D_{11}$, $D_{11} = D^\star_{11}$.

Covariance Models for Clustered Data

Whether we take a GEE or LME approach (with inference from the likelihood or from the posterior) we require flexible yet parsimonious covariance models.

In GEE we require a working covariance model

$$\text{cov}(\boldsymbol{Y}_i) = \boldsymbol{W}_i,$$

$i = 1, ..., m$.

With LME we have so far assumed the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{39}$$

with $\boldsymbol{b}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, with $\boldsymbol{E}_i = \boldsymbol{I}_{n_i}\sigma^2$.

With $\boldsymbol{z}_i \boldsymbol{b}_i = \boldsymbol{1}_{n_i} b_i$ we obtained an *exchangeable* (also known as compound symmetry):

$$\text{var}(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

This model is particularly appropriate for clustered data with no time ordering (e.g. ANOVA).

An obvious extension for longitudinal data is to assume

$$\boldsymbol{y}_i = \boldsymbol{x}_i \boldsymbol{\beta} + \boldsymbol{z}_i \boldsymbol{b}_i + \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i,$$

with:

- Random effects $\boldsymbol{b}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{D})$.
- Serial correlation $\boldsymbol{\delta}_i \sim_{ind} \text{N}(\boldsymbol{0}, \boldsymbol{R}_i \sigma_\delta^2)$, with $\boldsymbol{R}_i$ an $n_i \times n_i$ correlation matrix with elements

  $$R_{ijj'} = \text{corr}(Y_{ij}, Y_{ij'} | \boldsymbol{b}_i),$$

  $j, j' = 1, ..., n_i$.
- Measurement error $\boldsymbol{\epsilon}_i \sim_{ind} \text{N}(0, \boldsymbol{I}_{n_i} \sigma_\epsilon^2)$.

In general it is difficult to identify all three sources of variability – but the above provides a useful conceptual model.

See DHLZ, Chapter 5; Verbeke and Molenberghs, Chapter 10; Pinheiro and Bates, Chapter 5.

Within-Unit Covariance Models

Autoregressive errors

A widely-used time series model is the autoregressive, AR(1), process

$$\delta_{ij} = \rho\delta_{i,j-1} + u_{ij}, \tag{40}$$

for $j \geq 2$, $|\rho| \leq 1$ where $u_{ij} \sim_{iid} \mathrm{N}(0, \sigma_u^2)$ and are independent of δ_{ik}, $k > 0$. For LMEM we require a likelihood and hence the joint distribution of $\boldsymbol{\delta}_i$, for GEE the first two moments.

Repeated application of (40) gives, for $k > 0$,

$$\delta_{ij} = u_{ij} + \rho u_{i,j-1} + \rho^2 u_{i,j-2} + ... + \rho^{k-1} u_{j-k+1} + \rho^k \delta_{i,j-k}. \tag{41}$$

Assume the process has been running since $j = -\infty$ and that it is 'stable' so that $|\rho| < 1$ and the δ_{ij} all have the same distribution.

Then, from (41)

$$\mathrm{var}(\delta_{ij}) = \sigma_u^2(1 + \rho^2 + \rho^4 + ... + \rho^{2(k-1)}) + \rho^{2k}\mathrm{var}(\delta_{i,j-k}).$$

As $k \to \infty$, since $\sum_{l=1}^{\infty} x^{l-1} = 1/(1-x)$,

$$\mathrm{var}(\delta_{ij}) = \frac{\sigma_u^2}{(1-\rho^2)} = \sigma_\delta^2,$$

and, by substitution of (41),

$$\mathrm{cov}(\delta_{ij}, \delta_{i,j-k}) = \mathrm{E}[\delta_{ij}\delta_{i,j-k}] = \frac{\sigma_u^2 \rho^k}{(1-\rho^2)} = \sigma_\delta^2 \rho^k.$$

Hence under this model we have

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \rho & \rho^2 & ... & \rho^{n_i-1} \\ \rho & 1 & \rho & ... & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & ... & \rho^{n_i-3} \\ ... & ... & ... & ... & ... \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & ... & 1 \end{bmatrix}$$

as the correlation matrix for $\boldsymbol{\delta}_i$.

Often this model is written in the form

$$\mathrm{cov}(Y_{ij}, Y_{ik}) = \sigma_\delta^2 \exp(-\phi d_{ijk}),$$

$(\rho = \mathrm{e}^{\phi})$ with $d_{ijk} = |t_{ij} - t_{ik}|$ which is valid for unequally-spaced times also.

Toeplitz: Unstructured correlation:

$$\text{var}(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Heterogeneous versions with non-constant variance can also be fitted.

For example, the heterogenenous exchangeable model is given by:

$$\text{var}(\boldsymbol{Y}_i) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \rho\sigma_1\sigma_4 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho\sigma_2\sigma_4 \\ \rho\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 \\ \rho\sigma_4\sigma_1 & \rho\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 \end{bmatrix}$$

Note that we should be careful when specifying the covariance structure – identifiability problems may arise if we try to be too flexible.

Assessment of Assumptions

Each of the approaches to modeling that we have described depend upon assumptions concerning the structure of the data; to ensure that inference is appropriate we need to attempt to check that these assumptions are valid.

We first recap the assumptions:

GEE

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{e}_i,$$

with working covariance model $\text{var}(\boldsymbol{e}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$, $i = 1, ..., m$.

G1 Marginal model $\text{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

G2 m is sufficiently large for asymptotic inference to be appropriate.

G3 m is sufficiently large for robust estimation of standard errors.

G4 The working covariance $\boldsymbol{W}_i(\boldsymbol{\alpha})$ is not far from the "true" covariance structure; if this is the case then the analysis will be very inefficient (standard errors will be much bigger than they need to be).

LMEM via Likelihood Inference

Model:

$$\boldsymbol{Y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,$$

with $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ independent ($\boldsymbol{E}_i$ may have complex structure depending on both independent and dependent terms), $i = 1, ..., m$.

L1 Mean model for fixed effects $\boldsymbol{x}_i\boldsymbol{\beta}$ is appropriate.

L2 Mean model for random effects $\boldsymbol{z}_i\boldsymbol{b}_i$ is appropriate.

L3 Variance model for $\boldsymbol{\epsilon}_i$ is correct.

L4 Variance model for $\boldsymbol{b}_i$ is correct.

L5 Normality of $\boldsymbol{\epsilon}_i$.

L6 Normality of $\boldsymbol{b}_i$.

L7 m is sufficiently large for asymptotic inference to be appropriate.

LMEM via Bayesian Inference

Model as for LMEM, plus priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$.

Each of L1–L6 (asymptotic inference is not required if, for example, MCMC is used, though “appropriate” priors are needed).

Overall strategy

Before any formal modeling is carried out the data should be examined, in table and plot form, to see if the data have been correctly read in and to see if there are outliers.

For those individuals with sufficient data, individual-specific models should also be fitted, to allow examination of the appropriateness of initially hypothesized models in terms of the:

- linear component (which covariates, including transformations and interactions),
- and assumptions about the errors, such as constant variance and serial correlation.

Following fitting of marginal, mixed models, the assumptions should then be re-assessed, primarily through residual analysis.

Residual Analysis

Residuals may be defined with respect to different levels of the model.

A vector of unstandardized *population-level* (marginal) residuals is given by

$$\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}.$$

A vector of unstandardized *unit-level* (Stage One) residuals is given by

$$\boldsymbol{\epsilon}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta} - \boldsymbol{z}_i\boldsymbol{b}_i.$$

The vector of random effects, $\boldsymbol{b}_i$, is also a form of (Stage Two) residual.

Estimated versions of these residuals are given by

$$\begin{aligned} \widehat{\boldsymbol{e}}_i &= \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\epsilon}}_i &= \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i \end{aligned}$$

and $\widehat{\boldsymbol{b}}_i$, $i = 1, ..., m$.

Recall from consideration of the ordinary linear model that estimated residuals have dependencies induced by the estimation procedure; in the dependent data context the situation is much worse as the “true” residuals have dependencies due to the dependent error terms of the models used.

Hence standardization is essential to remove the dependence.

Standardized Population Residuals

If $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is the true error structure then

$$\text{var}(\boldsymbol{e}_i) = \boldsymbol{V}_i, \quad \text{and} \quad \text{var}(\widehat{\boldsymbol{e}}_i) \approx \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}}),$$

so that the residuals are dependent under the model, which means that it is not possible to check whether the covariance model is correctly specified (both form of the correlation structure and mean-variance model).

Plotting $\widehat{e}_{ij}$ versus x_{ij} may also be misleading due to the dependence within the residuals.

As an alternative, let $\widehat{\boldsymbol{V}}_i = \boldsymbol{L}_i\boldsymbol{L}_i^{\text{T}}$ be the Cholesky decomposition of $\widehat{\boldsymbol{V}}_i = \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})$, the estimated variance-covariance matrix.

We can use this decomposition to form

$$\widehat{\boldsymbol{e}}_i^\star = \boldsymbol{L}_i^{-1}\widehat{\boldsymbol{e}}_i = \boldsymbol{L}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}).$$

so that $\text{var}(\boldsymbol{e}_i^\star) \approx \boldsymbol{I}_{n_i}$. We have the model

$$Y_i^\star = \boldsymbol{x}_i^\star\boldsymbol{\beta} + \boldsymbol{e}_i^\star$$

where $\boldsymbol{Y}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{Y}_i$, $\boldsymbol{x}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{x}_i$, $\boldsymbol{e}_i^\star = \boldsymbol{L}_i^{-1}\boldsymbol{e}_i$.

Hence plots of $\widehat{e}_{ij}^\star$ against columns of $\boldsymbol{x}_{ij}^\star$ should not show systematic patterns, *if* the assumed form is correct.

QQ plots of $\widehat{e}_{ij}^\star$ versus the expected residuals from a normal distribution can be used to assess normality (normal residuals are not required for GEE, but will help asymptotics).

Unstandardized versions will still be normally distributed if the $\boldsymbol{e}_i$ are (since the $e_{ij}^\star$ are linear combinations of $\boldsymbol{e}_i$), though the variances may be non-constant, and there may be strong dependence between different points.

The correctness of the mean-variance relationship can be assessed via examination of $e_{ij}^{\star 2}$ versus $\widehat{\mu}_{ij}^\star = \boldsymbol{x}_{ij}^\star\widehat{\boldsymbol{\beta}}$.

Local smoothers can be added to plots to aid interpretation. Plotting symbols also useful – unit number, or observation number.

Stage One Residuals

If $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \sigma_i^2\boldsymbol{I}_{n_i})$ then residuals

$$\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}} - \boldsymbol{z}_i\widehat{\boldsymbol{b}}_i$$

may be formed. Standardized versions are given by $\widehat{\boldsymbol{\epsilon}}_i/\widehat{\sigma}_i$.

The standardized versions should be used if the σ_i are unequal across i. Some uses:

- Plot residuals against covariates. Departures may suggest adding in covariates, both to $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.
- To provide QQ plots – mean-variance relationship is more important to detect than lack of normality (so long as sample size is not small).
- assess constant variance assumption – one useful plot is versus $\widehat{\mu}_{ij} = \boldsymbol{x}_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}\widehat{\boldsymbol{b}}_i$.
- assess if serial correlation present in residuals

may be plotted against covariates to assess the form of the model, with QQ plots assessing normality of the measurement errors.

If $\boldsymbol{\epsilon}_i \sim \text{N}(\boldsymbol{0}, \sigma_\epsilon^2\boldsymbol{R}_i)$ with $\boldsymbol{R}_i$ a correlation matrix then the residuals should be standardized, as with population residuals.

Stage Two Residuals

Predictions of the random effects $\widehat{\boldsymbol{b}}_i$ may be used to assess assumptions associated with the random effects distribution, in particular:

- Are the random effects normally distributed?
- If we have assumed independence between random effects, does this appear reasonable?
- Is the variance of the random effects independent of covariates $\boldsymbol{x}_i$?

It should be born in mind that interpretation of random effects predictions is more difficult since they are functions of the data.

Recall that $\widehat{\boldsymbol{b}}_i$ are shrinkage estimators, and hence assumptions about $\boldsymbol{b}_i$ may not be reflected in $\widehat{\boldsymbol{b}}_i$.

We may fit curves for particular individuals with n_i large, and then check the assumptions from these.

For the LMEM it is better to examine first and second stage residuals – population residuals are a mixture so if something wrong not clear at which stage there is trouble.