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.

With LME we have so far assumed the model

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{27}$$

with $\boldsymbol{b}_i \sim_{ind} \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{\epsilon}_i \sim_{ind} \mathrm{N}(\boldsymbol{0}, \boldsymbol{E}_i)$, 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):

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

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

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

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

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

$$\delta_{ij} = \rho\delta_{i,j-1} + u_{ij}, \tag{28}$$

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 (28) gives, for $k > 0$,

$$\delta_{ij} = u_{ij} + \rho u_{i,j-1} + \rho^2 u_{i,j-2} + ... + \rho^{k-1} u_{j-k+1} + \rho^k \delta_{i,j-k}. \tag{29}$$

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 (29)

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

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 (29),

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

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: Banded 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.

Generalized Estimating Equations

We now describe the GEE method of modeling/inference. GEE attempts to make minimal assumptions about the data-generating process, and is constructed to answer population-level, rather than individual-level, questions.

We assume

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

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

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

with $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = \boldsymbol{0}$ for $i \neq i'$.

To motivate GEE we begin by assuming that $\boldsymbol{W}_i$ is known, and does not depend on unknown parameters. In this case the GLS estimator minimizes

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

and is given by the solution to the estimating function

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

The solution 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 have $\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, and if the information about $\boldsymbol{\beta}$ grows with increasing m, then $\widehat{\boldsymbol{\beta}}$ is consistent.

$\widehat{\boldsymbol{\beta}}$ is a consistent estimator for *any* fixed $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{W}_1, ..., \boldsymbol{W}_m)$. The weighting of observations by the latter dictates the efficiency of the estimator and not its consistency.

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

$$\left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \mathrm{var}(\boldsymbol{Y}_i) \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1} \boldsymbol{x}_i \right)^{-1} \tag{30}$$

If the assumed variance-covariance matrix is substituted, i.e. $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i$, then we obtain the *model-based* variance:

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

A Gauss-Markov theorem shows that, in this case, the estimator is efficient amongst linear estimators, *if* the variance model is correct.

The novelty of GEE is that rather than depend on the variance model being correct, sandwich estimation is used to repair any deficiency in the working variance model.

The GEE Algorithm

We now suppose that $\text{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ are unknown parameters in the variance-covariance model. A common approach is to assume

$$\boldsymbol{W}_i = \alpha_1 \boldsymbol{R}_i(\boldsymbol{\alpha}_2),$$

where $\alpha_1 = \text{var}(Y_{ij})$ and $\boldsymbol{R}_i(\boldsymbol{\alpha}_2)$ is a working correlation matrix depending on parameters $\boldsymbol{\alpha}_2$.

There are a number of choices for $\boldsymbol{R}_i$ including independence, exchangeable, and AR(1) models.

For known $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\beta}}$ is the root of the estimating equation

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \boldsymbol{0}. \tag{31}$$

When $\boldsymbol{\alpha}$ is unknown we require an estimator $\widehat{\boldsymbol{\alpha}}$ that converges to "something" so that, informally speaking, we have a stable weighting matrix, $\boldsymbol{W}(\widehat{\boldsymbol{\alpha}})$ in the estimating function.

The sandwich variance estimator is given by

$$\widehat{\text{var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \text{var}(\boldsymbol{Y}_i) \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right) \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \widehat{\boldsymbol{W}}_i^{-1} \boldsymbol{x}_i\right)^{-1} \tag{32}$$

where $\widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i(\widehat{\boldsymbol{\alpha}})$, and with $\text{var}(\boldsymbol{Y}_i)$ estimated by

$$(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})(\boldsymbol{Y}_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^{\mathrm{T}}. \tag{33}$$

This produces a consistent estimate of $\text{var}(\widehat{\boldsymbol{\beta}})$, so long as we have independence between units, i.e. $\text{cov}(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}) = 0$ for $i \neq i'$.

It is the replication across units that produces consistency and so the approach cannot succeed if we have no replication.

For inference the asymptotic distribution

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

may be used, where we emphasize that the asymptotics are in the number of units, m.

The variance estimator is sometimes referred to as *robust*, but *empirical* is a more appropriate description since the form is not robust to sample size and could be highly unstable for small m.

In the most general case we may allow $\boldsymbol{W}_i(\boldsymbol{\alpha}, \boldsymbol{\beta})$ so that regression parameters are contained in $\boldsymbol{W}_i$, to allow mean-variance relationships.

Example

In a longitudinal setting we may have the variance depending on the mean, $\mu_{ij} = \text{E}[Y_{ij}]$, and an autoregressive model:

$$\begin{aligned} \text{var}(Y_{ij} &= \alpha_1 \mu_{ij}^2 \\ \text{cov}(Y_{ij}, Y_{ik} &= \alpha_1 \alpha_2^{|t_{ij}-t_{ik}|} \mu_{ij}\mu_{ik} \\ \text{cov}(Y_{ij}, Y_{i'k}) &= 0, \quad i \neq i' \end{aligned}$$

with $j = 1, ..., n_i$, $k = 1, ..., n_{i'}$, and where t_{ij} is the time associated with Y_{ij}. Here

- α_1 is the variance (which is assumed constant across time and across individuals),
- α_2 is the correlation between responses on the same individual (which is assumed to be the same across individuals), and
- $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$.

In general the roots of the estimating equation

$$\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{W}_i^{-1}(\boldsymbol{\alpha}, \boldsymbol{\beta})(\boldsymbol{Y}_i - \boldsymbol{x}_i\boldsymbol{\beta}) = \mathbf{0}. \tag{34}$$

are not available in closed form because $\boldsymbol{\beta}$ occurs in $\boldsymbol{W}$.

We can write the $(k+1) \times 1$ estimating function in a variety of forms, for example:

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

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

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \sum_{k=1}^{n_i} \boldsymbol{x}_{ij} W_i^{jk}(Y_{ik} - \boldsymbol{x}_{ik}\boldsymbol{\beta})$$

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.

Relationship to the LMEM:

- The GEE approach is constructed to carry out marginal inference, and so individual-level inference cannot be performed.
- For a linear model marginalizing a LMEM produces a marginal model identical to that used in a GEE approach.
- For the LMEM sandwich estimation may of course be applied to the MLE of $\boldsymbol{\beta}$.

So far as the choice of "working" correlation structure is concerned, the trade-off is between choosing a simple structure for which there are few elements in $\boldsymbol{\alpha}$ to estimate, and a more complex model that will provide more efficient estimation of $\boldsymbol{\beta}$ if the variance model is closer to the true data generating mechanism but more instability in estimation of $\boldsymbol{\alpha}$.

To summarize, the GEE approach to modeling/estimation consists of:

1. A mean model $\mathrm{E}[\boldsymbol{Y}_i] = \boldsymbol{x}_i\boldsymbol{\beta}$.
2. A working variance model $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{W}_i(\boldsymbol{\alpha})$.
3. From 1. and 2. an estimating function is constructed, and sandwich estimation is applied to the variance of the resultant estimator.

Estimation of Variance Parameters

To formalize the estimation of $\boldsymbol{\alpha}$, we may introduce a second estimating equation. In the context of data with $\mathrm{var}(Y_{ij}) \propto v(\mu_{ij})$, with $\mu_{ij} = \mathrm{E}[Y_{ij}]$, the pair of estimating equations are given by:

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

where

- the "data" in the second estimating equation are given by
$$\boldsymbol{T}_i^{\mathrm{T}} = (R_{i1}R_{i2}, ..., R_{in_i-1}R_{in_i}, R_{i1}^2, ..., R_{in_i}^2),$$
with $R_{ij} = \{Y_{ij} - \mu_{ij}\}/v(\mu_{ij})^{1/2}$,
- $\boldsymbol{\Sigma}_i(\boldsymbol{\alpha}) = \mathrm{E}[\boldsymbol{T}_i]$ is a model for the correlations and variances for these standardized residuals,
- $\boldsymbol{E}_i = \frac{\partial \Sigma_i}{\partial \alpha}$, and
- $\boldsymbol{H}_i = \mathrm{cov}(\boldsymbol{T}_i)$ is the working covariance model for the squared and cross residual terms.

Recall that $\dim(\boldsymbol{\beta}) = k + 1$ so that the estimating equation $\boldsymbol{G}_1$ is of dimension $(k + 1) \times 1$ with $k + 1 = \dim(\boldsymbol{\beta})$, and $\boldsymbol{G}_2$ is of dimension $a \times 1$ where $a = \dim(\boldsymbol{\alpha})$.

The vector $\boldsymbol{T}_i$ has $n_i(n_i - 1)/2 + n_i$ elements in general. It is not straightforward to specify a working covariance model $\boldsymbol{H}_i$ for $\boldsymbol{T}_i$ and independence is often assumed.

If $\boldsymbol{G}_2$ is correctly specified then there will be efficiency gains.

A further advantage of this method is that it is straightforward to incorporate a regression model for the variance-covariance parameters, i.e. $\boldsymbol{\alpha} = g(\boldsymbol{x})$, for some link function $g(\cdot)$.

If $\mathrm{E}[\boldsymbol{T}] \neq \boldsymbol{\Sigma}$ then we will not get a consistent estimate of the true variance model but, importantly, consistency of $\boldsymbol{\beta}$ through $\boldsymbol{G}_1$ is guaranteed, so long as $\widehat{\boldsymbol{\alpha}}$ converges to "something".

We reiterate that a consistent estimate of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$ is guaranteed through the use of sandwich estimation, so long as units are independent.

For general $\boldsymbol{H}$ we will require the estimation of fourth order statistics, i.e. $\mathrm{var}(\boldsymbol{T})$, which is a highly unstable endeavor unless we have an abundance of data.

For this reason, working independence, $\boldsymbol{H}_i = \boldsymbol{I}$, is often used.

Example

As an illustration of the approach assume for simplicity $n_i = n = 3$ so that

$$\boldsymbol{T}_i^{\mathrm{T}} = [R_{i1}R_{i2} \quad R_{i1}R_{i3} \quad R_{i2}R_{i3} \quad R_{i1}^2 \quad R_{i2}^2 \quad R_{i3}^2]$$

with an exchangeable variance model:

$$\boldsymbol{\Sigma}_i(\boldsymbol{\alpha})^{\mathrm{T}} = \mathrm{E}[\boldsymbol{T}_i^{\mathrm{T}}] = [\alpha_1\alpha_2 \quad \alpha_1\alpha_2 \quad \alpha_1\alpha_2 \quad \alpha_1 \quad \alpha_1 \quad \alpha_1]$$

so that α_1 is the marginal variance, and α_2 is the correlation on observations on the same unit.

With $\boldsymbol{H}_i = \boldsymbol{I}$, a working independence model for the variance parmaters, the estimating function for $\boldsymbol{\alpha}$ is

$$\boldsymbol{G}_2(\widehat{\boldsymbol{\beta}}, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \begin{bmatrix} \alpha_2 & \alpha_2 & \alpha_2 & 1 & 1 & 1 \\ \alpha_1 & \alpha_1 & \alpha_1 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} R_{i1}R_{i2} \\ R_{i1}R_{i3} \\ R_{i2}R_{i3} \\ R_{i1}^2 \\ R_{i2}^2 \\ R_{i3}^2 \end{bmatrix} - \begin{bmatrix} \alpha_1\alpha_2 \\ \alpha_1\alpha_2 \\ \alpha_1\alpha_2 \\ \alpha_1 \\ \alpha_1 \\ \alpha_1 \end{bmatrix} \right)$$

Hence we need to simultaneously solve the two equations:

$$\begin{aligned} \sum_{i=1}^{m} \widehat{\alpha}_2 \left[\sum_{j<k} R_{ij}R_{ik} - \widehat{\alpha}_1\widehat{\alpha}_2 \right] + \sum_{j=1}^{3} (R_{ij}^2 - \widehat{\alpha}_1) &= 0 \\ \sum_{i=1}^{m} \widehat{\alpha}_1 \left[\sum_{j<k} R_{ij}R_{ik} - \widehat{\alpha}_1\widehat{\alpha}_2 \right] &= 0 \end{aligned}$$

Dividing the second of these by $\widehat{\alpha}_1$ gives

$$\widehat{\alpha}_1\widehat{\alpha}_2 = \frac{1}{3m} \sum_{i=1}^{m} \sum_{j<k} R_{ij}R_{ik}$$

and substituting this into the first equation gives

$$\widehat{\alpha}_1 = \frac{1}{3m} \sum_{i=1}^{m} \sum_{j<k} R_{ij}^2$$

to give a pair of method-of-moment estimators.

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 $\boldsymbol{G}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}^{(t)}) = \mathbf{0}$, with $\boldsymbol{G}$ given by (31), to give $\widehat{\boldsymbol{\beta}}^{(t+1)}$,
2. Estimate $\widehat{\boldsymbol{\alpha}}^{(t+1)}$ based on $\widehat{\boldsymbol{\beta}}^{(t+1)}$.

Set $t \rightarrow t + 1$ and return to 1.

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 6 summaries are presented under likelihood, Bayesian and GEE analyses.

Two Bayesian models were fitted, a normal model:

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

and a Student t_4 model:

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

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 6: 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 7: 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: $\boldsymbol{D} \neq \boldsymbol{D}^\star$; $D_{00} = D^\star_{00} - 2\bar{t}D^\star_{01} + \bar{t}^2 D^\star_{11}$, $D_{01} = D^\star_{01} - \bar{t}D_{11}$, $D_{11} = D^\star_{11}$.

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 $(\widehat{\boldsymbol{\epsilon}}_i/\widehat{\sigma}_i)^2$ 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.

Assessing Adequacy of the Temporal Covariance Structure

An informal method for assessing whether there is residual temporal dependence is to plot residuals versus time, we now consider more formal tools such as the correlgram and the variogram.

We begin with some definitions.

Consider a stochastic process $Y(t)$ and let

$$\gamma(t,s) = \text{cov}\{Y(t), Y(s)\} = \text{E}[\{Y(t) - \mu(t)\}\{Y(s) - \mu(s)\}],$$

denote the *autocovariance function* of $Y(t)$.

The term *serial dependence* signifies that there is dependence between $Y(t)$ and $Y(s)$ for at least some pairs (s,t) with $s \neq t$.

We write

$$Y(t) = \mu(t) + e(t),$$

where $\mu(t)$ is the deterministic trend component.

Definition: A process $e(t)$ is second-order stationary if $\mathrm{E}[e(t)]$ is constant, for all t, and $\gamma(t, s)$ depends only on $|t - s|$. For a residual process any non-zero constant has been absorbed into $\mu(t)$.

Example: The simplest example of a stationary random sequence is *white noise* which consists of a sequence of mutually independent random variables, each with mean 0 and finite variance σ^2.

There is a fundamental difficulty with trying to decompose $Y(t)$ into the trend and the stochastic component in a single series because the two are unidentifiable without further assumptions.

Is it serial dependence in the residuals, or a high-order polynomial trend for example?

The Autocorrelation Function

For a second-order stationary random process, the autocovariance function is

$$\mathrm{cov}\{Y(t), Y(t+u)\} = \mathrm{cov}\{e(t), e(t+u)\},$$

so that $C(0)$ is the variance of $Y(t)$ for all t.

The autocorrelation function is defined as

$$\rho(u) = \frac{C(u)}{C(0)}.$$

For equally-spaced data we could fit a model and then examine the autocorrelation function (ACF) of the residuals,

$$e_t = \frac{y_t - \widehat{y}_t}{\widehat{\mathrm{var}}(Y_t)^{1/2}}.$$

Consider a stochastic process $e(t)$, and realizations e_t, $t = 1, ..., n$. The *emprical* autocorrelation is defined as

$$\widehat{\rho}(u) = \widehat{\text{corr}}\{e(t), e(t+u)\} = \frac{\sum_{t=1}^{n-u} e_t e_{t+u}/(n-u)}{\sum_{t=1}^{n} e_t^2/n},$$

for $u = 0, 1, ...$

A *correlogram* plot is $\widehat{\rho}(u)$ versus u. If the residuals are a white noise process, we have the asymptotic result

$$\sqrt{n}\ \widehat{\rho}(u) \ \rightarrow_d \ \text{N}(0,1),$$

for $u = 1, 2, ...,$ to give confidence bands $\pm 1.96/\sqrt{n}$.

The Variogram

For unequally-spaced data the ACF is not so convenient, unless we round the observations.

An alternative is provided by the *semi-variogram* which is defined, for a process e_t and $d \geq 0$.

$$\gamma(d) = \frac{1}{2}\text{var}\,(e_t - e_{t-d}) = \frac{1}{2}\text{E}\left[\{e_t - e_{t-d}\}^2\right].$$

Recall that for a second-order stationary process, $\text{E}[e_t] = \mu$ for all t and $\text{cov}(e_t, e_{t-d})$ only depends on the distance d (which implies constant variance).

A smooth process is L_2-continuous, i.e.

$$\text{E}\{(e_t - e_{t-d})^2\} \rightarrow 0$$

as $d \rightarrow 0$. For a second-order stationary smooth process

$$\begin{aligned}\gamma(d) &= \frac{1}{2}\left\{\text{E}[e_t^2] + \text{E}[e_{t-d}^2] - 2\text{E}[e_t e_{t-d}]\right\} \\ &= \sigma_e^2\{1 - \rho(d)\},\end{aligned}$$

where $\text{var}(e) = \sigma_e^2$.

The semi-variogram is also well-defined for an *intrinsically* stationary process for which $\mathrm{E}[e_t] = \mu$ and for which

$$\mathrm{E}[(e_t - e_{t-d})^2] = 2\gamma(d).$$

As d increases then for observatons far apart in time

$$\gamma(d) \rightarrow \mathrm{var}(e_t) = \sigma_e^2,$$

which (recall) is assumed constant.

Consider measurement error, ϵ_t with $\mathrm{E}[\epsilon_t] = 0$, $\mathrm{var}(\epsilon_t) = \sigma_\epsilon^2$, and

$$Y_t = \mu_t + e_t + \epsilon_t,$$

so that we no longer have a smooth process. Then

$$\gamma(d) = \frac{1}{2}\mathrm{E}\left[\{Y_t - Y_{t-d}\}^2\right] = \sigma_e^2\{1 - \rho(d)\} + \sigma_\epsilon^2,$$

and we have a "nugget" effect σ_ϵ^2.

The Variogram in Longitudinal Data Analysis

Define the semi-variogram of the population residuals, $e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta}$, as

$$\gamma_i(d_{ijk}) = \frac{1}{2}\mathrm{E}\left[\{e_{ij} - e_{ik}\}^2\right],$$

for $d_{ijk} = \mid t_{ij} - t_{ik} \mid \geq 0$. We emphasize that we are examining differences on the *same* individual.

The sample semi-variogram uses the empirical halved differences between pairs of population residuals

$$v_{ijk} = \frac{1}{2}(e_{ij} - e_{ik})^2,$$

along with the spacings $u_{ijk} = t_{ij} - t_{ik}$.

With highly-irregular sampling times the variogram can be estimated from the pairs (u_{ijk}, v_{ijk}), $i = 1, ..., m$, $j < k = 1, ..., n_i$, with the resultant plot being smoothed.

The marginal distribution of each v_{ijk} is χ^2_1, and this large variability can make the variogram difficult to interpret.

The total variance is estimated as the average of $\frac{1}{2}(e_{ij} - e_{lk})^2$, for $i \neq l$, since

$$\frac{1}{2}\mathrm{E}\left[(e_{ij} - e_{lk})^2\right] = \frac{1}{2}\left\{\mathrm{E}[e_{ij}^2] + \mathrm{E}[e_{lk}^2]\right\} = \sigma^2,$$

assuming that observations on different individuals are independent (and the variance is constant over time, and for different individuals).

Consider the interpretation of the variogram for the model

$$Y_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta} + b_i + \delta_{ij} + \epsilon_{ij},$$

where $b_i \sim_{ind} N(0, \sigma_0^2)$ (note, univariate), $\epsilon_{ij} \sim_{ind} N(0, \sigma_\epsilon^2)$, and δ_{ij} represent error terms with serial dependence.

A simple and commonly-used form for serial dependence is the AR(1) model given by

$$\mathrm{cov}(\delta_{ij}, \delta_{ik}) = \sigma_\delta^2 \rho^{|t_{ij} - t_{ik}|}.$$

Under this model

$$\mathrm{var}(Y_{ij}|\boldsymbol{\beta}) = \sigma^2 = \sigma_0^2 + \sigma_\delta^2 + \sigma_\epsilon^2.$$

Consider the theoretical variogram for the residuals

$$e_{ij} = Y_{ij} - \boldsymbol{x}_{ij}\boldsymbol{\beta} = b_i + \delta_{ij} + \epsilon_{ij},$$

$i = 1, ..., m; j = 1, ...n_i$, with the AR(1) model.

For differences in residuals on the same individual

$$e_{ij} - e_{ik} = b_i + \delta_{ij} + \epsilon_{ij} - b_i - \delta_{ik} - \epsilon_{ik} = \delta_{ij} + \epsilon_{ij} - \delta_{ik} - \epsilon_{ik},$$

and so

$$\gamma_i(d_{ijk}) = \frac{1}{2}\mathrm{E}\left[(e_{ij} - e_{ik})^2\right] = \sigma_\delta^2(1 - \rho^{d_{ijk}}) + \sigma_\epsilon^2. \tag{35}$$

As $d_{ijk} \to 0$, $\gamma_i(d_{ijk}) \to \sigma_\epsilon^2$ and b_i is the mean of e_{ij} and so its variance does not appear in (35).

Figure 9 shows the theoretical semi-variogram under this model and for the population residuals.

The variogram is limited in its use for *population* residuals for the LMEM, as we now illustrate.

Consider, the mixed effects model with random intercepts and independent random slopes:

$$b_{i0} \sim N(0, D_0), \quad b_{i1} \sim N(0, D_1)$$

leads to non-constant marginal variance

$$\text{var}(Y_{ij}|\boldsymbol{\beta}) = \sigma^2_\epsilon + D_0 + D_1 t^2_{ij},$$

so that we would not want to look at a variogram of population residuals because we do not have second-order stationarity. However, we could look at individual residuals after the random intercepts and slopes model has been fitted.

In my experience the variogram is often dominated by sampling variability (and there can be strong dependence in the plot since each residual contributes many points).

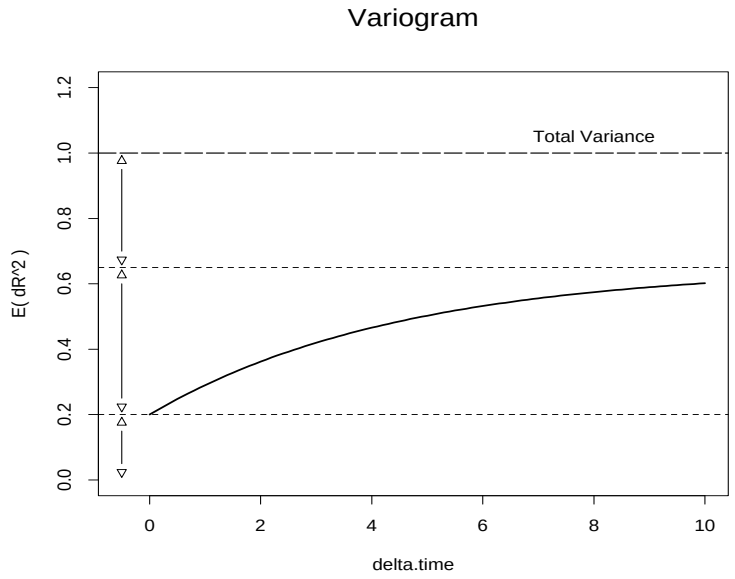

Figure 9: Theoretical variogram for a model with a random intercept, serial correlation, and measurement error.

Example: FEV1 over Time

Data Description

We examine data from an epidemiological study described by van der Lende (1981). We analyze a sample of 133 men and women. Study participants were followed over time to obtain information on the prevalence of, and risk factors for, chronic obstructive lung diseases.

The sample, initially aged 15-44, participated in follow-up surveys approximately every 3 years for up to 21 years.

At each survey, information on respiratory symptoms and smoking status was collected by questionnaire and spirometry was performed.

Pulmonary function was measured by spirometry and a measure of forced expiratory volume (FEV1) was obtained every three years for the first 15 years of the study, and also at year 19.

Each study participant was either a current or a former smoker, with current smoking defined as smoking at least one cigarette per day.

Missing Data

In this dataset FEV1 was not recorded for every subject at each of the planned measurement occasions so that the number of repeated measurements of FEV1 on each subject varied between 1 and 7.

Table 8 shows the numbers of observations available at each time point. There are 32 former smokers and 101 current smokers in total, and we see that the numbers with missing observations at each time point are not drastically different. Hopefully this means that the missingness does not depend on the unobserved FEV1 at these time points.

Time	Former smoker	Current smoker
0	3.52 (23)	3.23 (85)
3	3.58 (27)	3.12 (95)
6	3.26 (28)	3.09 (89)
9	3.17 (30)	2.87 (85)
12	3.14 (29)	2.80 (81)
15	2.87 (24)	2.68 (73)
19	2.91 (28)	2.50 (74)

Table 8: Mean FEV1 (and sample size) by smoking status and time.

Initial Plots

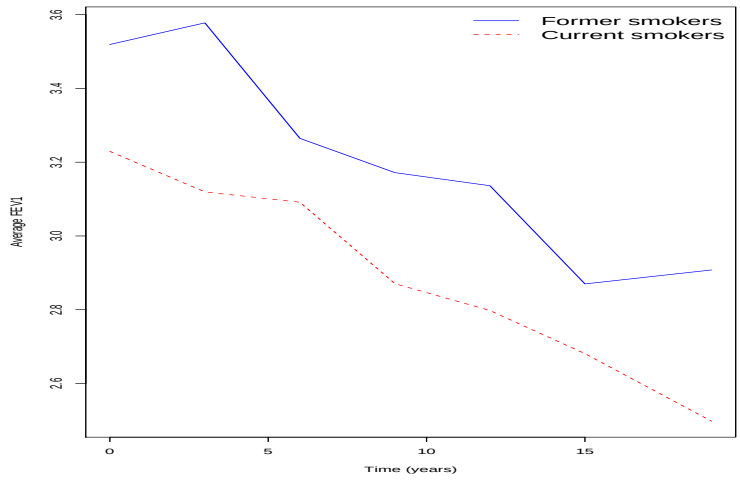

Figure 10: Mean FEV1 profiles versus time for 133 individuals: clear that there is a difference in the overall level, with former smokers having a higher level.

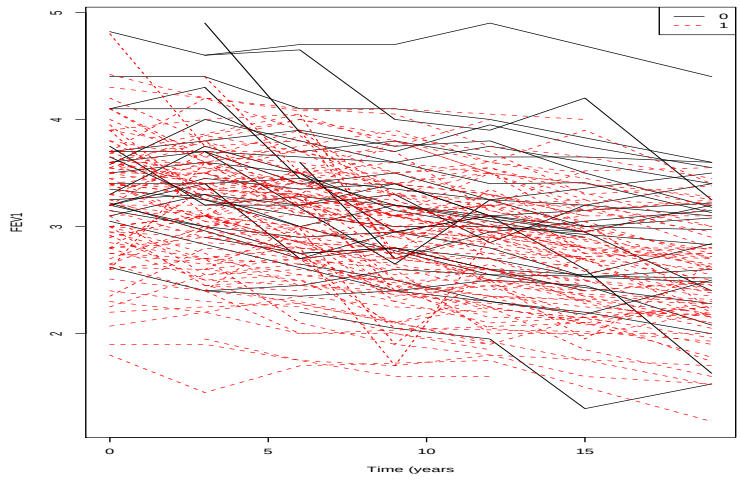

Figure 11: FEV1 versus time for 133 individuals, former smokers coded 0, current smokers 1. Large between-person variability in levels. It is clear that observations on the same individual will be correlated.

Models

Let Y_{ij} represent the FEV1 on individual i at time (from baseline) t_{ij} (in years), and S_i the smoking status with 0 representing former smoker and 1 current smoker.

We initially fit the following three models using REML:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_i + \epsilon_{ij} \quad (36)$$

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + b_i + \epsilon_{ij} \quad (37)$$

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 S_i + \beta_3 S_i \times t_{ij} + b_i + \epsilon_{ij} \quad (38)$$

with $b_i \sim_{iid} \mathrm{N}(0, \sigma_0^2)$ and $\epsilon_{ij} \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2)$, and with ϵ_{ij} and b_i independent, $i = 1, ..., m$, $j = 1, ..., n_i$.

Parameter Estimates

Compared to the equivalent LMEM the standard errors of the LS estimates corresponding to time-varying covariates (time and the interaction) are reduced in the LMEMs. This behavior occurs because within-person comparisons are more efficient in a longitudinal study.

Model	β_1 (Time)	s.e.	β_2 (Smoke)	s.e.	β_3 (Inter)	s.e.
LMEM TIME	-0.037	0.0013	–	–	–	–
LMEM TIME+SMOKE	-0.037	0.0013	-0.31	0.11	–	–
LMEM TIME × SMOKE	-0.034	0.0026	-0.27	0.11	-0.0046	0.0030
LS TIME × SMOKE	-0.038	0.0067	-0.31	0.085	-0.00041	0.0077

Table 9: Results of a least squares fit (LS) and various linear mixed effects model (LMEM) analyses.

Fixed Effect Testing

To illustrate how one may test between the three LMEMs in Table 9 we must use MLE for likelihood ratio tests since the data are not constant under the different models under REML (due to different $\widehat{\boldsymbol{\beta}}_G$).

For

$$H_0 : \text{TIME versus } H_1 : \text{TIME+SMOKE}$$

we have a statistic of 8.22 on 1 degree of freedom and a p-value of 0.0042. Hence there is strong evidence to reject the null and conclude that there are differences in intercepts for former and current smokers.

For

$$H_0 : \text{TIME+SMOKE versus } H_1 : \text{TIME+SMOKE+TIME}\times\text{SMOKE}$$

we have a statistic of 2.29 on 1 degree of freedom and a p-value of 0.13, hence there is no evidence to reject the null and conclude that the interaction is not needed.

Bayesian and GEE Analyses

We now report a Bayesian analysis of model (37) with improper flat priors on $\beta_0, \beta_1, \beta_2$, the improper prior $\sigma^2_\epsilon \propto \frac{1}{\sigma^2_\epsilon}$ and $\sigma_0^{-2} \sim \text{Ga}(0.5, 0.02)$. The latter prior gives 95% of its mass for σ_0, the between-individual slope, between 0.09 and 6.5.

The results are given in Table 10, and are very similar to the likelihood-based approach, which is reassuring.

We now fit the marginal model version of (37) using GEE. We use an exchangeable correlation structure, since clearly we have dependence between measurements on the same individual at different times, but the exact form of the correlation is not clear.

The results are given in Table 10, and again show good agreement for the regression coefficients.

In the exchangeable correlation structure there are two components to $\boldsymbol{\alpha}$ parameters, a marginal variance, α_1, and a common marginal correlation, α_2.

The exchangeable model may be compared to the random intercepts model in which we have marginal variance $\alpha_1 = \sigma_0^2 + \sigma_\epsilon^2$ and marginal correlation $\alpha_2 = \sigma_0^2/(\sigma_0^2 + \sigma_\epsilon^2)$.

We have $\widehat{\alpha}_1 = 0.31, \widehat{\alpha}_2 = 0.82$ which gives $\sqrt{\widehat{\alpha}_1 \times \widehat{\alpha}_2} = 0.50$, which is comparable to the estimates of σ_0 in Table 10.

Model	β_1 (Time)	s.e.	β_2 (Smoke)	s.e.	σ_0
Likelihood LMEM	-0.037	0.0013	-0.32	0.11	0.53
Bayes LMEM	-0.037	0.0013	-0.31	0.12	0.53
GEE	-0.037	0.0015	-0.31	0.11	–
Likelihood LMEM AR(1)	-0.037	0.0013	-0.31	0.11	0.53

Table 10: Results of linear mixed effects models (likelihood and Bayesian) and GEE analyses.

Assessment of Assumptions

We examine the assumptions for the linear model that includes time and smoking (but no interaction).

Figure 12 summarizes the Stage One residuals: the top left panel shows that the distribution of the errors is symmetric, but heavier tailed than normal — with such a large sample there is nothing troubling in this plot. No outlying points.

The next two figures plot the residuals against time and smoking status. We see no nonlinear behavior in the time plot, and no great divergence from constant variance in either plot.

A very important assumption in mixed effects modeling is that the random effects distribution do not depend on covariates. To examine this separate analyses were carried out for former and current smokers. The estimates for former smokers were $\widehat{\sigma}_\epsilon = 0.22, \widehat{\sigma}_0 = 0.58$, and for current smokers $\widehat{\sigma}_\epsilon = 0.21, \widehat{\sigma}_0 = 0.51$. The differences between the two groups are small and we conclude that there is no evidence against a common distribution.

The final plot is the absolute value of the residuals versus fitted values with a smoother. There is slight evidence of an increase but nothing to be too concerned about. These residual plots were based on residuals from the likelihood analysis, Bayesian versions were similar.

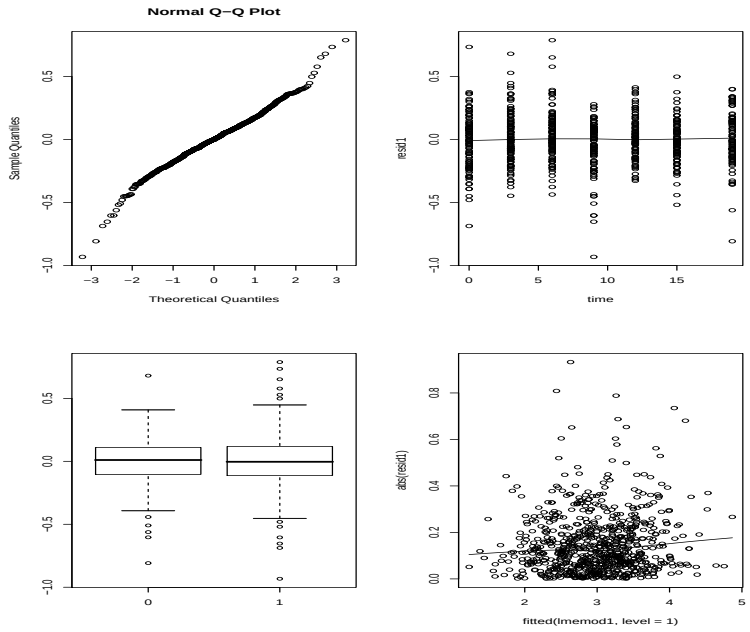

Figure 12: First stage residual plots.

For the 132 individuals who produce individual least squares fits, Figure 13 shows a QQ plot of the (a) intercepts, and (b) slopes, and (c) a bivariate scatter plot. The estimates look remarkably normal, and there are no outlying individuals.

Figure 14 gives boxplots of the LS estimates of (a) intercepts and (b) slopes versus smoking status. There is no evidence of a great difference in spread between former and current smokers.

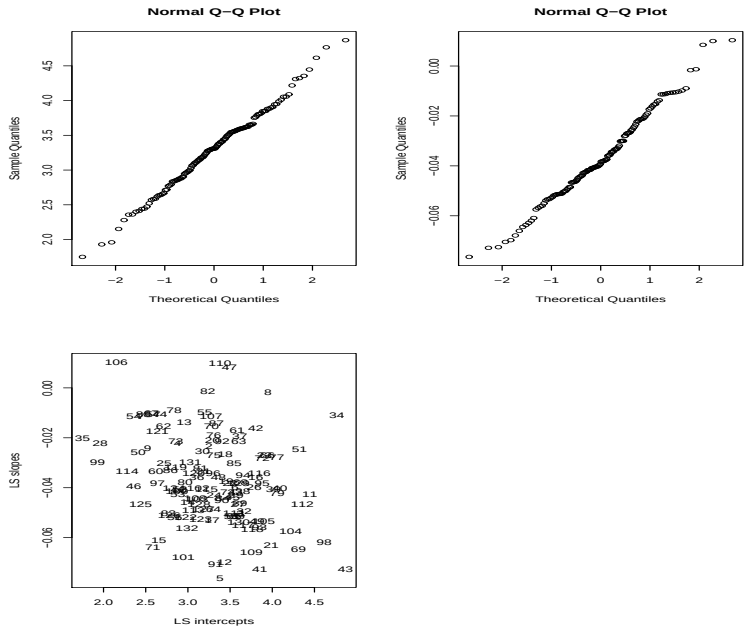

Figure 13: Second stage least squares estimate plots.

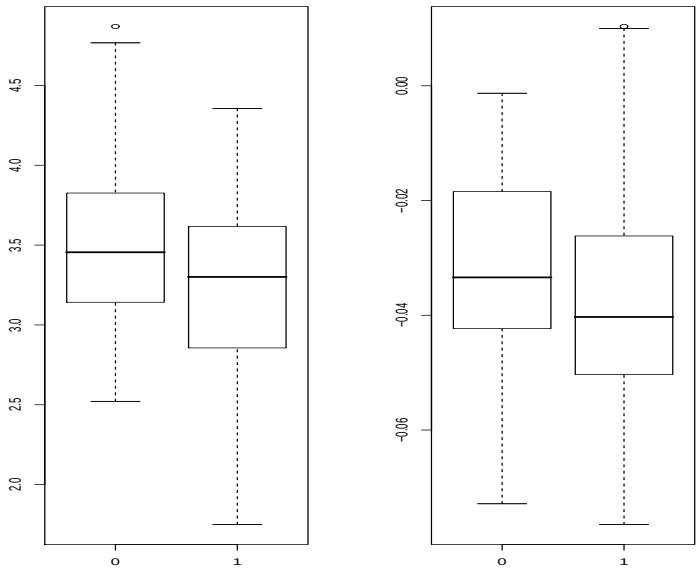

Figure 14: Least squares estimates of intercepts (left) and slopes (right) versus smoking status.

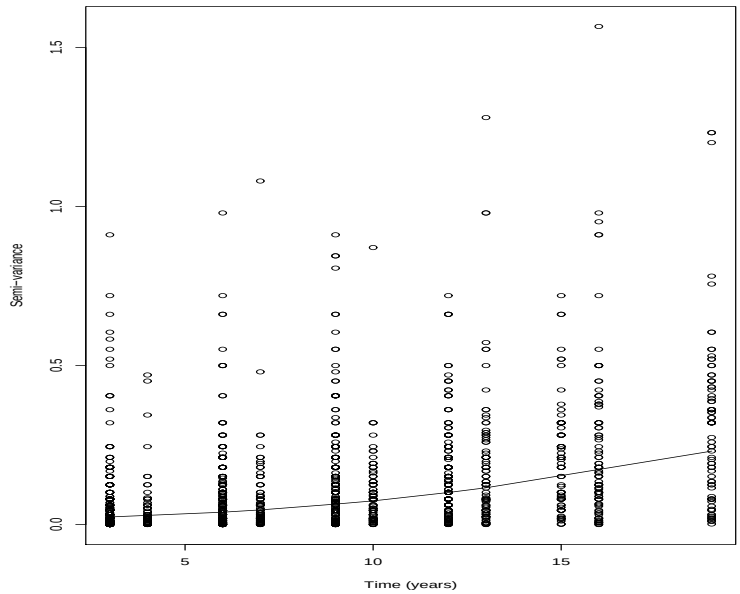

Figure 15: Semi-variogram of first stage residuals.

Finally we examine the residuals for serial correlation.

Figure 15 gives the semi-variogram of the first-stage residuals and indicates some dependence. Consequently, we fit an AR(1) model for the residuals using restricted likelihood and obtain the parameter estimates in the last row of Table 10.

This model is a significant improvement over the non-serial correlation model (as measured by a likelihood ratio test, $p = 0.0002$). However, there is virtually no change in the estimates/standard errors here since the AR correlation parameter is just 0.20, with an asymptotic 95% confidence interval of (0.087,0.30).

We may also examine whether random slopes are required. Fitting this model via restricted likelihood gave a standard deviation of $\widehat{\sigma}_1 = 0.0099$.

The likelihood ratio statistic testing for correlated random intercepts and slopes, versus random intercepts only is 10.9 which is significant at around the 0.0025 (where the distribution under the null is a mixture of χ^2_1 and χ^2_2 distributions).

Conclusions

Inference under either the random intercept or random intercepts and slopes models is relatively similar since although the random slopes model is a statistical improvement over the random intercepts only, the between-individual variability in slopes is small.

The population change in FEV1 over time is a drop of 0.037 litres per year, with a standard error of 0.0013–0.0015 depending on the model.

The median for the intra-person correlation is 0.84 with 95% interval (0.82,0.89) suggesting that the majority of the variability is between-person.

Stochastic Covariates

In some longitudinal situations, the response at time t on individual i may depend on not just the current covariates, but also previous values.

For example, in an investigation into the health effects of recent air pollution we may believe that the response depends on not just today's exposure, but also the preceeding days.

In such situations, obtaining the correct form of the model will in general be difficult, and instead we might decide to estimate the association for a simpler model.

As an example, suppose that we have a single covariate, and we decide to examine the *cross-sectional* association:

$$\mu_{ij} = \mathrm{E}[Y_{ij} \mid X_{ij}]. \tag{39}$$

In such a situation great care must be taken to obtain a consistent estimator.

We demonstrate with a GEE approach, though the pitfalls of estimation apply equally to likelihood and Bayesian approaches.

Example

Suppose the "true" model is given by:

$$\begin{aligned} \mathrm{E}[Y_{it}|X_{it}, X_{it-1}] &= \gamma_0 + \gamma_1 X_{it} + \gamma_2 X_{it-1} \\ X_{it} &= \rho X_{it-1} + \epsilon_{it} \end{aligned}$$

with $|\rho| < 1$. For example X_{it} may represent an air pollutant on day t, and Y_{it} a measure of an individual's lung function.

We may be interested in the cross-sectional effect of the pollutant, e.g. suppose we have data on X_{it} only. We have

$$\mathrm{E}[Y_{it}|X_{it}] = \beta_0 + \beta_1 X_{it}$$

where $\beta_0 = \gamma_0$ and $\beta_1 = \gamma_1 + \rho\gamma_2$.

Estimation for Stochastic Covariate Situations

The **full covariate conditional mean** (FCCM) condition is given by

$$\mu_{it} \equiv \mathrm{E}[Y_{it}|X_{it}] = \mathrm{E}[Y_{it}|X_{i1}, X_{i2}, ..., X_{iT}]$$

and if true gives an unbiased GEE estimating equation, as we now illustrate.

In the example we just described the FCCM condition was not satisfied.

With a GLM:

$$\eta_{ij} = g(\mu_{ij}) = \boldsymbol{x}_{ij}\boldsymbol{\beta},$$

and assume for simplicity $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$. The generalized estimating function is given by

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

which has second row

$$G_2(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left[\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} X_{ij} W^{\star}_{ijk} (Y_{ik} - \mu_{ik}) \right]$$

where (39) is the assumed model, i.e. $\mu_{ik} = \mathrm{E}[Y_{ij} \mid X_{ik}]$, and $W^{\star}_{ijk} = \frac{\partial \mu_{ij}}{\partial \eta_{ij}} W_i^{jk}$ with W_i^{jk} the (j,k)-th element of $\boldsymbol{W}_i^{-1}$.

To obtain consistency we require

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta})] = \mathbf{0}.$$

Previously we have seen that if the mean specification is correct then we obtain consistency of $\widehat{\boldsymbol{\beta}}$.

Since now the estimating function depends on the random variables $\boldsymbol{X} = (X_1, ..., X_m)^{\mathrm{T}}$ the expectation is with respect to $\boldsymbol{X}$ and $\boldsymbol{Y}$. Specifically

$$\mathrm{E}_{Y,X}[G_2(\boldsymbol{\beta})] = \sum_{i=1}^{m} \mathrm{E}_{Y_i,X_i} \left[\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} X_{ij} W^{\star}_{ijk} (Y_{ik} - \mu_{ik}) \right]$$

and

$$\begin{aligned} \mathrm{E}_{Y_i,X_i} \left[X_{ij} W^{\star}_{ijk} (Y_{ik} - \mu_{ik}) \right] &= \mathrm{E}_{X_i} \left\{ \mathrm{E}_{Y_i|X_i} [X_{ij} W^{\star}_{ijk} (Y_{ik} - \mu_{ik}) \right] \\ &= \mathrm{E}_{X_i} \left\{ X_{ij} W^{\star}_{ijk} (\mathrm{E} \left[Y_{ik} \mid X_{i1}, ..., X_{in_i} \right] - \mu_{ik}) \right\} \end{aligned}$$

Hence to ensure an unbiased estimating function, in general, and hence consistency of our estimator, we require the FCCM condition:

$$\mathrm{E} \left[Y_{ik} \mid X_{i1}, ..., X_{in_i} \right] = \mu_{ik} = \mathrm{E}[Y_{ik} \mid X_{ik}],$$

otherwise we have bias.

Suppose we assume working independence, the above simplifies to

$$G_2(\boldsymbol{\beta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij} W^{\star}_{ijj} (Y_{ij} - \mu_{ij}),$$

so that

$$\mathrm{E}[\boldsymbol{G}(\boldsymbol{\beta})] = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathrm{E}_{X_{ij}} \left[X_{ij} W^{\star}_{ijj} (\mathrm{E}[Y_{ij} \mid X_{ij}] - \mu_{ij}) \right] = \mathbf{0},$$

and we obtain a consistent estimator.

For more details see DHLZ, Section 12.3.1.

Cross-Sectional Versus Longitudinal Studies

Consider modeling $Y = \text{FEV}_1$ as a function of age. We might envisage that Y changes both with age within an individual, and that individuals may have different baseline levels of Y from which they begin, due to "cohort" effects. A birth cohort is a group of individuals who were born in the same year.

Cohort effects include the effects of environmental pollutants, and differences in lifestyle choices and medical treatment.

In a cross-sectional study a group of individuals are measured at a single time point. A great advantage of longitudinal studies, over cross-sectional studies is that both cohort and aging effects can be estimated.

As an illustration Figure 16 shows three hypothetical individuals outcome trajectory over calendar time — the starting positions are different due to cohort effects.

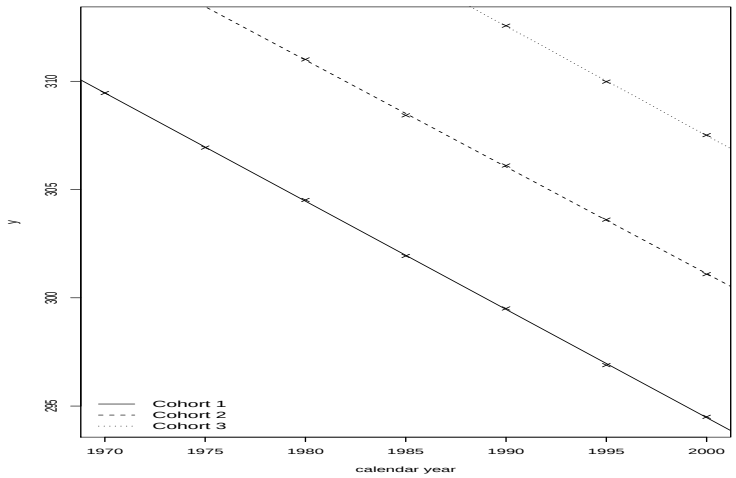

Figure 16: Three individual's trajectories over time.

Figure 17 shows the same individuals but with trajectories plotted versus age, and the cross-sectional association, which would resolve from observing the final measurement only, highlighted.

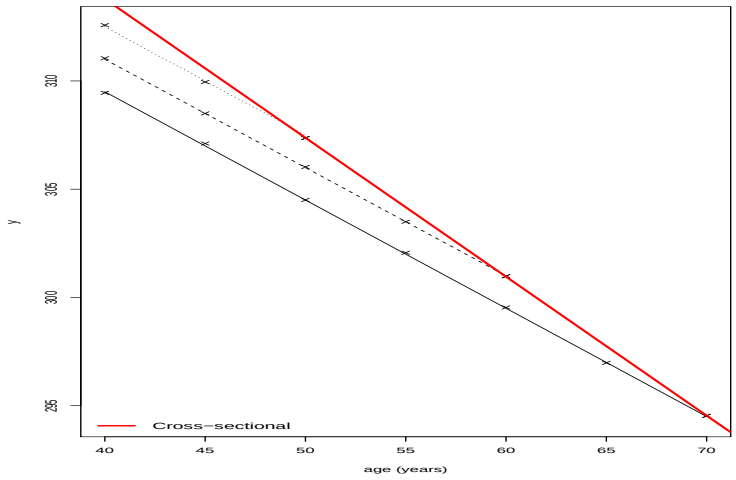

Figure 17: Relationship between cross-sectional and longitudinal effects in hypothetical example with three individuals.

To illustrate, consider the model:

$$\mathrm{E}[Y_{ij} \mid x_{ij}, x_{i1}] = \beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1})$$

where Y_{ij} is the j-th FEV_1 measurement on individual i and x_{ij} is the age of the individual when that measurement was taken, with x_{i1} begin the age on a certain day (so that all the individuals are comparable).

Parameter interpretation

We have

$$\mathrm{E}[Y_{i1} \mid x_{i1}] = \beta_0 + \beta_C x_{i1},$$

so that β_C is the average change in Y between two populations who differ by one unit in their baseline ages; said another way we are examining the differences in Y between two birth cohorts a year apart.

Also

$$\mathrm{E}[Y_{ij} \mid x_{ij}, x_{i1}] - \mathrm{E}[Y_{i1} \mid x_{i1}] = \beta_L (x_{ij} - x_{i1})$$

so that β_L is the longitudinal effect, that is the change in the average response between two populations who are in the same birth cohort, and whose ages differ by one year.

The usual cross-sectional model is given by:

$$\begin{aligned} \mathrm{E}[Y_{ij} \mid x_{ij}] &= \beta_0 + \beta_1 x_{ij} \\ &= \beta_0 + \beta_1 x_{i1} + \beta_1 (x_{ij} - x_{i1}) \end{aligned} \tag{40}$$

so that the model implicitly assumes equal longitudinal and cohort effects, i.e. $\beta_1 = \beta_L = \beta_C$.

In a cohort study with model (40) we have

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \overline{x})(Y_{ij} - \overline{Y})}{\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2}$$

with $\overline{x} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}$, $\overline{Y} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^{n_i} Y_{ij}$ with $N = \sum_{i=1}^m n_i$. The expected value of this estimator is

$$\mathrm{E}[\widehat{\beta}_1] = \beta_L + \frac{\sum_{i=1}^m n_i (x_{i1} - \overline{x}_1)(\overline{x}_i - \overline{x})}{\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2} (\beta_C - \beta_L)$$

so that the estimate if of a combination of cohort and longitudinal effects.

The cross-sectional regression model will give an unbiased estimate of the longitudinal association if $\beta_L = \beta_C$ or if $\{x_{i1}\}$ and $\{\overline{x}_i\}$ are orthogonal.

This illustrates that a benefit of a longitudinal study is the ability to estimate both cohort and longitudinal effects.

If we write $\beta_{0i} = \beta_0 + \beta_C x_{i1}$ then we could fit the model

$$\mathrm{E}[Y_{ij} \mid x_{ij}, x_{i1}] = \beta_{0i} + \beta_L (x_{ij} - x_{i1})$$

so that each individual has their own intercept, though this runs into problems with individuals with sparse data (can't use a random effects model since the intercepts are related to x_{i1}, invalidating an assumption of the model).