The LMM we have described is remarkable general and flexible. In the following example, we explore the possibilities of this model on a longitudinal data set concerning growth.

Example – Potthoff&Roy Growth Data

Potthoff and Roy describe a study conducted at the University of North Carolina Dental School in two groups of children (16 boys and 11 girls). At ages 8, 10, 12, and 14, y =the distance (mm) from the center of the ptituitary gland to the pterygomaxillary fissure was measured. Change in this distance during growth is important in orthodontal therapy. The data are listed in Table 3.3 of Davis' book (p.58).

- In this example we'll fit several models to these data, all of which are of the LMM form given on p.97.

- Explanatory variables to be considered for the columns of $\mathbf{X}_i$ and $\mathbf{Z}_i$ include age (both as continuous variable and as a factor), gender, and their interaction. Let

$$x_i = \begin{cases} 0, & \text{if } i^{\text{th}} \text{ subject is male,} \\ 1, & \text{if female,} \end{cases} \qquad \mathbf{t}_i = \begin{pmatrix} 8 \\ 10 \\ 12 \\ 14 \end{pmatrix}, \quad \forall i$$

  and let m08, m10, m12, m14 be indicators for ages 8–14 for males, f08, f10, f12, f14 be corresponding indicators for females.

- In what follows we'll often replace i by hi to identify the i^{th} subject in the h^{th} gender group ($h = 1$ for boys, $h = 2$ for girls) rather than the i^{th} subject overall.

- There is more than one valid strategy for developing an appropriate LMM for the analysis of a given data set, but most frequently, the following procedure is followed:

    1. Specify a saturated, or "full" model for the mean structure*, and then try to identify a parsimonious but adequate variance-covariance structure through the specification of random effects and assumptions on $\mathbf{D}$ and $\mathbf{R}$. Selection of this var-cov structure is usually done with AIC, but LRTs and subject-matter considerations can also be helpful.

    2. Then based on the chosen var-cov structure, if reduction of the mean structure is desired,** do this using approximate F and t tests rather than Wald or LR tests (unless the sample size is quite large, in which case either can be used). The K-R method for F and t tests is recommended.

    3. While ML can be used to develop the model (to facilitate LR tests and model selection criteria comparisons across models with different mean structure), refit the final model with REML and use the REML-fitted model as the basis of all final inferences.

- Of course all model building is an iterative process, so this "final" model should be subjected to the same model diagnostics and process of revision as with fixed effect regression models.

- We illustrate this process using the Potthoff and Roy dental growth data. See dental.sas and its output, dental.lst.

- See Davis (§6.4), Verbeke & Molenberghs (2000, §17.4), and also Verbeke & Molenberghs (1997, §4.4) for alternative analyses of these data.

---

* in the absence of a clear choice of saturated model, use a model that is at least as complex as any that you are willing to consider.

** which is not always appropriate, for example with experimental data.

Model 1: Let y_{hij} = the response at measurement occasion j, for the i^{th} subject in gender group h. Model 1 includes a separate mean for each gender×time combination:

$$y_{hij} = \beta_{hj} + \varepsilon_{hij}$$

where we assume $\boldsymbol{\varepsilon}_{hi} = (\varepsilon_{hi1}, \ldots, \varepsilon_{hi1})^T \sim N(\mathbf{0}, \mathbf{R}(\boldsymbol{\theta}))\ \forall h, i$, where $\mathbf{R}$ is constant over h, i and is of the completely unstructured form $\Rightarrow \boldsymbol{\theta}$ has $4 + \binom{4}{2} = 10$ elements corresponding to the 4 diagonal elements of $\mathbf{R}$ and the 6 unique off-diagonal elements.

- The REPEATED statement in PROC MIXED determines the form of $\mathbf{R}$, the RANDOM statement specifies the random effects in the model (unlike PROC GLM, random effects appear only on the RANDOM statement, not on the MODEL statement) and their var-cov matrix $\mathbf{D}$.

- The S option of the MODEL statement prints the fixed effects estimates (S for "solution").

- type=un specifies the unstructured form for $\mathbf{R}$, subject=id specifies the cluster identifier, r=k and rcorr=k print the covariance matrix ($\mathbf{R}$) and correlation matrix, respectively, corresponding to subject k (if =k is omitted SAS assumes k=1).

Model 0: In model 1 we assumed that $\mathbf{R}$ was the same for boys and girls. We relax this assumption in model 0 with the group=sex option on the repeated statement. That is, we assume $\boldsymbol{\varepsilon}_{hi} = (\varepsilon_{hi1}, \ldots, \varepsilon_{hi1})^T \sim N(\mathbf{0}, \mathbf{R}_h(\boldsymbol{\theta}))$. Notice that r=1,12 now asks for the $\mathbf{R}_h$ matrix for subject 1 (the first girl) and for subject 12 (the first boy).

- There does appear to be differences across gender in $\mathbf{R}_h$. Both AIC and a LRT of model 0 versus model 1, support model 0 as more appropriate for these data despite its large increase in parameters.
  - The AIC for model 0 is 448.7 vs 452.5 for model 1.
  - The LRT statistic is 416.5-392.7=23.8 which is asymptotially $\chi^2(10)$, gving $p = .0081$.

Models 0a–0e: In models 0a–0e, we retain the same mean structure and fit several simpler variance-covariance models to these data by imposing some structure on $\mathbf{R}_h$, $h = 1, 2$, the error var-cov matrix for boys and girls.

- In models 0a and 0b, we fit heteroscedastic (SAS uses the term heterogeneous) and non-heteroscedastic versions of the **Toeplitz** structure for $\mathbf{R}_h$, $h = 1, 2$. A Toeplitz var-cov matrix is banded. In the non-heteroscedastic form (TYPE=TOEP),

$$\mathbf{R}(\boldsymbol{\theta}) = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & \theta_4 \\ & \theta_1 & \theta_2 & \theta_3 \\ & & \theta_1 & \theta_2 \\ & & & \theta_1 \end{pmatrix}$$

- In the heteroscastic form (TYPE=TOEPH), the diagonal elements of $\mathbf{R}$ above are allowed to differ, allowing for different variances at each age, and the correlation matrix $\text{corr}(\boldsymbol{\varepsilon}_{hi})$ is assumed to be banded.
  - It is very common for longitudinal data to exhibit heteroscedasticity over time, so heteroscedastic forms are always worth considering.

- However, in this case, the Toeplitz form fits best.

- In model 0c and 0d, we fit heteroscedastic and non-heteroscedastic compound symmetry forms (TYPE=CSH and TYPE=CS, respectively). The non-heteroscedastic form is the same as the Toeplitz form above, but with the restriction that $\theta_2 = \theta_3 = \theta_4$. Again, the heteroscedastic form allows the diagonal elements to differ.

  - Recall that the CS form is also induced by specifying $\mathbf{R} = \sigma^2\mathbf{I}$ and including random, subject-specific intercepts. We will return to this point later.

- Note that of all var-cov structures considered so far, the CS form with separate parameter values for boys and girls (model 0d) has the smallest AIC (431.4).

- Finally, we check whether we can assume $\mathbf{R}_1 = \mathbf{R}_2$ with a common CS form in model 0e by dropping the GROUP=SEX option on the REPEATED statement. This model yields a higher AIC (446.6) and fits worse according to a LRT (test statistic=426.6-407.4=19.2 on 2 d.f.).

Therefore, we settle on the CS structure with separate parameter values for boys and girls. Now we consider reducing the mean. This can be done by fitting simpler models and conducting LRTs.

However, (1) we should not do this unless we use ML rather than REML as the basis of the log-likelihood comparisons; and (2) unless the sample size is very large, it is better to do inference on the fixed effects via approximate F and t tests.

- Next we refit model 0d, using an alternative parameterization that yields main effect and interaction tests for sex and age. We see by the significant sex*age interaction ($F_{3,68.4} = 3.01, p = .0360$) that, while it may be possible to model effects of age as linear, we probably should not expect those linear effects to be the same for the two genders.

We next test $\mathrm{E}(y_{hij}) = \beta_{hj}$ (the mean specification in model 0d) versus $\mathrm{E}(y_{hij}) = \alpha_h + \beta_h \mathrm{age}_{hij}$. That is, we test whether the effects of age are linear, without assuming that these linear effects are the same in the two genders. The latter model can be considered a non-parallel slopes ancova (analysis of covariance) model.

- A LRT of this hypothesis is straight-forward to conduct as a test of nested models, but because F tests outperform LR tests for fixed effects in a LMM (unless the sample size is large), we conduct this test as a test of the general linear hypothesis $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ using orthogonal polynomial contrasts on the β_{hj}'s in model 0d.

- Specifically, we can test whether there is no linear trend with age for boys via the contrast

$$\psi_1 = -3\beta_{11} - \beta_{12} + \beta_{13} + 3\beta_{14} + 0\beta_{21} + 0\beta_{22} + 0\beta_{23} + 0\beta_{24}$$

  and we can test whether there is no linear trend with age for girls via the contrast

$$\psi_2 = 0\beta_{11} + 0\beta_{12} + 0\beta_{13} + 0\beta_{14} - 3\beta_{21} - \beta_{22} + \beta_{23} + 3\beta_{24}$$

- Quadratic and cubic patterns for boys are captured by the contrasts

$$\psi_3 = \beta_{11} - \beta_{12} - \beta_{13} + \beta_{14} + 0\beta_{21} + 0\beta_{22} + 0\beta_{23} + 0\beta_{24}$$
$$\psi_4 = -\beta_{11} + 3\beta_{12} - 3\beta_{13} + \beta_{14} + 0\beta_{21} + 0\beta_{22} + 0\beta_{23} + 0\beta_{24}$$

  and for girls, by the contrasts

$$\psi_5 = 0\beta_{11} + 0\beta_{12} + 0\beta_{13} + 0\beta_{14} + \beta_{21} - \beta_{22} - \beta_{23} + \beta_{24}$$
$$\psi_6 = 0\beta_{11} + 0\beta_{12} + 0\beta_{13} + 0\beta_{14} - \beta_{21} + 3\beta_{22} - 3\beta_{23} + \beta_{24}$$

- A joint test of $H_0 : \psi_3 = \psi_4 = \psi_5 = \psi_6 = 0$ tests whether there is no nonlinear pattern in the mean response for boys and no nonlinear pattern for girls, so this test is equivalent to the hypothesis that the non-parallel slopes ancova model holds.

– This test is conducted in PROC MIXED via the CONTRAST statement, yielding $F_{4,52.5} = 0.34, p = 0.8520$ so we fail to reject the hypothesis that the simpler model holds. The tests of no linear trend for boys ($F_{1,45} = 68.83, p < .0001$) and for girls ($F_{1,45} = 78.20, p < .0001$) indicate that there is a significant linear trend for both boys and girls, which is, by the 4 d.f. test, also not nonlinear.

Therefore, we adopt model 2, the non-parallel slopes ancova model,

$$y_{hij} = \alpha_h + \beta_h \text{age}_{hij} + \varepsilon_{hij}, \quad \text{where } \boldsymbol{\varepsilon}_{hi} \overset{ind}{\sim} N(\mathbf{0}, \mathbf{R}_h) \qquad \text{(Model 2)}$$

and $\mathbf{R}_h$ is of the CS form.

- Refitting this model using the parameterization

$$\text{E}(y_{hij}) = \beta_0 + \beta_1 \text{sex}_{hij} + \beta_2 \text{age}_{hij} + \beta_3 \text{sex}_{hij} \text{age}_{hij}$$

yields a test of equal slopes across gender via the t test on β_3 (t-test on sex*age in the SAS program), which gives $t_{70.9} = -2.83, p = .0060$, so we reject the hypothesis of parallel slopes.

The final model, therefore, is the non-parallel slopes ancova model, Model 2.

- It is worth noting that this purely fixed effect model is the marginal form implied by the hierarchical model

$$y_{hij} = (\alpha_h + b_{hi}) + \beta_h \text{age}_{hij} + \varepsilon_{hij}, \qquad \text{(Model 2alt)}$$

where

$$\varepsilon_{hij} \overset{iid}{\sim} N(0, \sigma_h^2) \quad \text{and} \quad b_{h1}, \ldots, b_{hn_h} \overset{iid}{\sim} N(0, \sigma_{bh}^2), \quad h = 1, 2.$$

– The random intercept model above implies compound symmetry for the variance-covariance structure for all observations that share the same random intercept.

- This random intercept model is fit as Model 2alt, and notice it gives identical results to those of Model 2.

  - Note that Model 2alt implies that the covariance between any pair of observations which share a random intercept is equal to the variance component of that random intercept and, therefore, is necessarily positive. However, one can have models of the form given by Model 2, where the covariance in the CS for for $\mathbf{R}$ is negative.

  - So, there is a subtle distinction between the marginal and hierarchical forms of the model: every model of the form in Model 2alt is necessarily of the form given in Model 2, but the converse is not true.

  - In practice, negative covariances/correlations in compound symmetric var-cov structures are extremely rare, so the models are, practically equivalent.

In this example we started by carefully choosing the variance-covariance structure for the data. The validity of model-based inferences for the fixed effects (which is what we are usually primarily interested in) depends crucially on the variance-covariance structure being modeled appropriately (not under-specified). However, if the variance-covariance structure is misspecified, valid inferences can still be salvaged, as long as we don't ignore that misspecification.

The solution is to use an estimator of $\text{var}(\hat{\boldsymbol{\beta}})$ which is robust to misspecification of $\text{var}(\mathbf{y})$.

The "Robust" or "Sandwich" Estimator of var($\hat{\boldsymbol{\beta}}$):

- For simplicity, consider the LMM where $\mathbf{X}$ is of full rank.

Recall that the ML and REML estimators of $\boldsymbol{\beta}$ in the LMM are given by

$$\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}} = \{\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{V}(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{y}. \qquad (\heartsuit)$$

We saw that for $\boldsymbol{\theta}$ known,

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \{\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\underbrace{\mathrm{var}(\mathbf{y})}_{=\mathbf{V}(\boldsymbol{\theta})}\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\{\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1} \\ &= \{\mathbf{X}^T\mathbf{V}(\boldsymbol{\theta})^{-1}\mathbf{X}\}^{-1}, \end{aligned} \qquad (\spadesuit)$$

and that asymptotically, it doesn't matter if $\boldsymbol{\theta}$ is known or unknown; either way, the asymptotic variance-covariance of $\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}}$ is still given by ($\spadesuit$).

In the clustered data case, these formulas simplify somewhat: ($\heartsuit$) becomes

$$\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}} = \left\{\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}_i\right\}^{-1}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{y}_i, \qquad (\heartsuit')$$

and ($\spadesuit$) becomes

$$\begin{aligned}&\mathrm{var}(\hat{\boldsymbol{\beta}}) \\ &= \left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right\}^{-1}\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathrm{var}(\mathbf{y}_i)\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right\}^{-1} \\ &= \left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right\}^{-1}. \end{aligned} \qquad (\spadesuit')$$

Note that if $\mathrm{var}(\mathbf{y}_i) \neq \mathbf{V}_i(\boldsymbol{\theta})$ (that is, if $\mathbf{V}_i$ is an incorrectly specified variance-covariance matrix for $\mathbf{y}_i$), then ($\heartsuit'$) is still a legitimate estimator of $\boldsymbol{\beta}$ (in fact it can still be proven to be a consistent estimator), but the simplification between lines 2 and 3 of ($\spadesuit'$) no longer holds!

That is, for $\mathbf{V}_i$ misspecified, the asymptotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}}$ is

$$\left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right\}^{-1}\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\underbrace{\mathrm{var}(\mathbf{y}_i)}_{\neq\mathbf{V}_i(\boldsymbol{\theta})}\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\boldsymbol{\theta})^{-1}\mathbf{X}_i\right\}^{-1}.$$

Of course, this quantity must be estimated to get $\hat{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}})$. Since $\mathrm{var}(\mathbf{y}_i) \neq \mathbf{V}_i(\boldsymbol{\theta})$ if $\mathbf{V}_i$ is misspecified, we would not want to estimate $\mathrm{var}(\mathbf{y}_i)$ by $\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})$.

Instead, we can estimate $\mathrm{var}(\mathbf{y}_i)$ from the residuals from the model. This leads to

$$\tilde{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}}) = \left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}_i\right\}^{-1}$$

$$\times\sum_i \mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{e}_i\mathbf{e}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}_i\left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}_i\right\}^{-1},$$

where

$$\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}}.$$

- $\tilde{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}})$ has been proposed by several authors including Huber (1967), White (1980), and Liang and Zeger (1986).

- $\tilde{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}})$ is often called a "sandwich estimator" because of its form, or a "robust estimator" because it is robust to misspecification of $\mathrm{var}(\mathbf{y}_i)$.

- It can be shown to be a consistent estimator of $\mathrm{var}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}})$ under misspecification of $\mathrm{var}(\mathbf{y}_i)$. However, when $\mathrm{var}(\mathbf{y}_i)$ is correctly specified $\tilde{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}})$ is a much less efficient estimator than the so-called model-based estimator,

$$\hat{\mathrm{av}}\mathrm{ar}(\hat{\boldsymbol{\beta}}_{(\mathrm{RE})\mathrm{ML}}) = \left\{\sum_i \mathbf{X}_i^T\mathbf{V}_i(\hat{\boldsymbol{\theta}}_{(\mathrm{RE})\mathrm{ML}})^{-1}\mathbf{X}_i\right\}^{-1}.$$

- The robust or sandwich estimator is available in PROC MIXED with the EMPIRICAL option on the PROC MIXED statement, and its use precludes the use of the small sample inference adjustments provided by the DDFM=SATTERTH and DDFM=KR options.

- The use of the sandwich var-cov estimator is recommended only when there is strong reason to suspect that $\text{var}(\mathbf{y}_i)$ is misspecified and/or when the number of clusters n is quite large.

Back to the Example:

- In the second to last call to PROC MIXED in dental2.sas, model 2 is refitted with $\mathbf{R}_{hi} = \sigma^2\mathbf{I}$ for all h, i. That is, we assume all of the data are independent and homoscedastic. In this setting where we have longitudinal data and where we have observed differences in variability between boys and girls, we can be fairly confident that this is an overly simplistic and incorrect var-cov structure for these data, so to salvage valid asymptotic inference, we can use the sandwich estimator by specifying the EMPIRICAL option on the PROC MIXED statement.

- For comparison purposes, the last call to PROC MIXED assumes independence without invoking the empirical option. Note the substantial differences in the inferences on fixed effects. Those from the independence model without the use of the sandwich estimator, should not be trusted.

Another Example — Methemoglobin in Sheep, Again:

The first time we analyzed these data, we used a RM-ANOVA approach. Recall that this approach fits a split plot model, which is a model with a subject specific intercept implying a compound symmetry var-cov structure.

To deal with departures from compound symmetry in the observed var-cov structure, G-G or H-F adjustments to hypothesis tests were done.

Rather than fitting the wrong var-cov structure and adusting the analysis for non-sphericity, a more appealing approach is to fit the right var-cov structure so that such adjustments are not necessary. This is now possible with the LMM machinery that we have learned.

- See sheep3.sas and sheep3.lst.

- In the first two calls to PROC MIXED, we simply refit the split-plot model in which a compound symmetry var-cov structure is assumed within each subject (sheep). This is done either by including a random sheep effect (first call to PROC MIXED) or by specifying $\mathbf{R}_{hi} = \mathbf{R}$ to have a CS form, common to all sheep (second call to PROC MIXED).

- In each case we use the REML estimation method.

- Notice that the results from these two approaches are identical (same restricted loglikelihoods, variance component estimates, F tests, etc.).

- Notice also that the REML variance component estimates here coincide with the ANOVA estimators obtained in sheep1.sas using METHOD=TYPE3.

- In the third call to PROC MIXED, we change the var-cov structure from compound symmetry (CS) to completely unstructured (UN). The unstrutured variance-covariance structure estimates a parameter for every element in the upper (or lower) triangle of $\mathbf{R}$. Thus, it imposes no structure on $\mathbf{R}$.

    - This is the same variance-covariance assumption as used in the multivariate approaches (profile analysis, growth curve analysis), and in fact the F tests on no2 and time are exactly the same as in the profile analysis done in sheep2.sas.

- The estimated $\mathbf{R}$ matrix is printed on p.5 as a covariance matrix and then again on p.6 as a correlation matrix. The form of this matrix can be helpful in suggesting a structured var-cov matrix for $\mathbf{y}_{hi}$ that is simpler than UN but fits better than CS.

    - The documentation for the REPEATED statement in SAS' PROC MIXED contains a nice summary and description of the structured variance-covariance matrices that can be fitted in that software. Here is a direct link to that documentation: http://tinyurl.com/a3sr4tm

- While we could try several of these structures and compare them with AIC to try to select an appropriate var-cov model, the form of $\mathbf{R}$ on pp.5–6 of sheep3.lst suggests that this approach may not be fruitful. It appears that observations at measurement occasions 2–6 are strongly positively correlated with each other and negatively correlated with measurement occasion 1. This suggests that there's something fundamentally different about measurement occasion 1.

- From the description of the experiment, recall that the treatment began between measurement occasions 1 and 2. That is, measurement occasion 1 was a baseline response level for each sheep. This suggests that perhaps it should be treated differently than the other measurement occasions in the analysis.

- So, before finishing this example we've got a couple of other issues to consider: 1) variance-covariance models, and 2) methods of handling baseline values.

Variance-covariance Models:

Before we consider a RM-ANCOVA model for the Sheep data, we need to discuss models for the variance-covariance structure.

Diggle *et al.* (2002) identify three qualitatively different sources of random variability in longitudinal data:

1. Grouping effects or shared-characteristics.

   – These are most appropriately modelled with random effects.

2. Serial correlation — repeated observations on a subject may be governed by a time-varying stochastic process operating within that subject. This results in observations from the same subject being correlated, where the correlation depends upon the time separation in the the observations. Typically, correlation decreases with time separation.

   – Serial correlation is most appropriately modelled through $\mathbf{R}_i$, the variance of $\boldsymbol{\varepsilon}_i$.

3. Measurement error — the individual data points themselves (i.e., the repeated measurements taken on each subject) may be subject to measurement error, which introduces additional variability into the reponse. E.g., if some lab assay is required to obtain the methemoglobin measurements at each of the 6 sampling times, this assay may introduce additional measurement error in trying to quantify methemoglobin in the sheep's bloodstream.

   – Measurement error can also be accounted for through $\mathbf{R}_i$.

These three sources of variability are illustrated in Figure 3.1 of Verbeke and Molenberghs (2000), reproduced below.

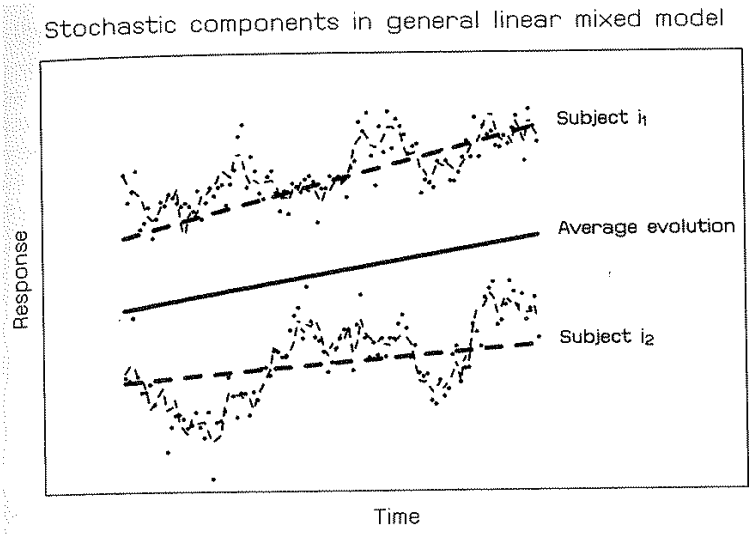


FIGURE 3.1. *Graphical representation of the three stochastic components in the general linear mixed model (3.11). The solid line represents the population-average evolution. The lines with long dashes show subject-specific evolutions for two subjects i_1 and i_2. The residual components of serial correlation and measurement error are indicated by short-dashed lines and dots, respectively.*

The inclusion of random effects in $\mathbf{b}_i$ in the LMM accounts for grouping effects. To account for serial correlation and measurement error, Diggle *et al.* (2002) suggest decomposing the error term $\boldsymbol{\varepsilon}_i$ as

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i},$$

where $\boldsymbol{\varepsilon}_{(1)i}$ accounts for measurement error and $\boldsymbol{\varepsilon}_{(2)i}$ accounts for serial correlation.

The resulting LMM can now be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_{(1)i} + \boldsymbol{\varepsilon}_{(2)i}, \quad i = 1, \ldots, n,$$

where

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}(\boldsymbol{\theta})), \quad \boldsymbol{\varepsilon}_{(1)i} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{t_i}), \quad \boldsymbol{\varepsilon}_{(2)i} \sim N(\mathbf{0}, \tau^2\mathbf{H}_i)$$

and $\mathbf{b}_1, \ldots, \mathbf{b}_n, \boldsymbol{\varepsilon}_{(1)1}, \ldots, \boldsymbol{\varepsilon}_{(1)n}, \boldsymbol{\varepsilon}_{(2)1}, \ldots, \boldsymbol{\varepsilon}_{(2)n}$ are independent.

- Note that this is just the same LMM as before but with $\mathbf{R}_i = \sigma^2\mathbf{I}_{t_i} + \tau^2\mathbf{H}_i$, which may or may not be assumed equal across all i.

To capture heteroscedasticity (non-constant variance) and correlation, it is useful to decompose $\mathbf{H}_i$ as

$$\mathbf{H}_i = \mathbf{B}_i^{1/2}\mathbf{C}_i\mathbf{B}_i^{1/2}$$

where

$$\mathbf{B}_i^{1/2} = \frac{1}{\tau}\begin{pmatrix} \sqrt{\text{var}(\varepsilon_{(2)i1})} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\text{var}(\varepsilon_{(2)i2})} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{\text{var}(\varepsilon_{(2)it_i})} \end{pmatrix}$$

and

$$\mathbf{C}_i = \text{corr}(\boldsymbol{\varepsilon}_{(2)i})$$

$$= \begin{pmatrix} 1 & \text{corr}(\varepsilon_{(2)i1}, \varepsilon_{(2)i2}) & \text{corr}(\varepsilon_{(2)i1}, \varepsilon_{(2)i3}) & \cdots & \text{corr}(\varepsilon_{(2)i1}, \varepsilon_{(2)it_i}) \\ \text{corr}(\varepsilon_{(2)i2}, \varepsilon_{(2)i1}) & 1 & \text{corr}(\varepsilon_{(2)i2}, \varepsilon_{(2)i3}) & \cdots & \text{corr}(\varepsilon_{(2)i2}, \varepsilon_{(2)it_i}) \\ \vdots & \vdots & \vdots & \ddots & \\ & \vdots & & & \\ \text{corr}(\varepsilon_{(2)it_i}, \varepsilon_{(2)i1}) & \text{corr}(\varepsilon_{(2)it_i}, \varepsilon_{(2)i2}) & \text{corr}(\varepsilon_{(2)it_i}, \varepsilon_{(2)i3}) & \cdots & 1 \end{pmatrix}$$

Pinheiro and Bates (2000) use this decomposition combined with a variance function to model heteroscedasticity and a correlation function to model serial correlation which are both quite flexible and yield a wide variety of variance-covariance structures to choose from.

Variance Modelling:

Heteroscedasticity can be captured by assuming non-constant variance in the elements of $\boldsymbol{\varepsilon}_{(2)i}$. Specifically, we assume

$$\text{var}(\varepsilon_{(2)ij}) = \tau^2 g^2(\mathbf{v}_{ij}, \boldsymbol{\delta})$$

where $\mathbf{v}_{ij}$ is a vector of *variance covariates*, $\boldsymbol{\delta}$ is a vector of *variance parameters* to be estimated (part of $\boldsymbol{\theta}$), and $g^2(\cdot)$ is a known *variance function.*

Correlation Modelling:

In general, our correlation model will be

$$\text{corr}(\varepsilon_{(2)ij}, \varepsilon_{(2)ik}) = h\{d(p_{ij}, p_{ik}), \boldsymbol{\rho}\}$$

where $\boldsymbol{\rho}$ is a vector of *correlation parameters*, $h(\cdot)$ is a known *correlation function*, p_{ij}, p_{ik} are the measurement times for observations y_{ij}, y_{ik}, and $d(\cdot, \cdot)$ is a known *distance function.*

- The correlation function $h(\cdot)$ is assumed continuous in $\boldsymbol{\rho}$, returning values in $[-1, +1]$. In addition, $h(0, \boldsymbol{\rho}) = 1$, so that observations that are 0 distance apart (identical observations) are perfectly correlated.

Combining these variance and correlation functions leads to a rich variety of variance-covariance structures possible for $\mathbf{R}_i$.

- Many of the possible combinations are implemented in PROC MIXED as different specifications of the TYPE= option on the REPEATED statement.

- In PROC MIXED, the measurement error term is omitted by default, but it can be included in $\mathbf{R}_i$ by using the LOCAL option on the REPEATED statement.

- Pinheiro and Bates' lme software for R and S-PLUS also implements a wide variety of structures for $\mathbf{R}_i$. See their book, or the STAT 8230 class notes on my web page (pp.144–152), for details

Methods for Handling Baseline Values:

Suppose that we have data $\mathbf{y}_{hi} = (y_{hi1}, y_{hi2}, \ldots, y_{hit_{hi}})^T$ on subjects $i = 1, \ldots, n_h$, in groups $h = 1, \ldots, s$, where y_{hi1} is the response value at measurement occasion 1, which occurs at the outset of the study.

- In a randomized study we assume that this value is measured *before* randomization to the treatment groups, or at least before any effect of the treatments could possibly occur.

- In a non-randomized study where groups are defined by pre-existing characteristics of the subjects (e.g., race, gender), of course group differences may be present at baseline.

Fitzmaurice, et al. (2011) discuss 4 methods for handling baseline response:

1. Treat the baseline value as part of the response vector making no assumptions about group differences in the mean response at baseline.

2. Treat the baseline value as part of the response vector but assume that the mean response is equal across groups at the baseline measurement occasion. I.e., assume $\mathrm{E}(y_{hi1}) = \mathrm{E}(y_{h'i'1})$ for all $, h, i, h', i'$.

3. Subtract the baseline response from each post-baseline measurement and analyze the resulting **change scores** (aka gain scores): $y_{hi2} - y_{hi1}, y_{hi3} - y_{hi1}, \ldots, y_{hit_{hi}} - y_{hi1}$.

4. Analyze only the post-baseline measurements, but treat the baseline response as a covariate. I.e., build a model of the form

$$y_{hij} = \gamma y_{hi1} + \mathbf{x}_{hij}^T \boldsymbol{\beta} + \mathbf{z}_{hi}^T \mathbf{b}_{hi} + \varepsilon_{hij}, \quad j = 2, \ldots, t_{hi}. \qquad (*)$$

Comments, comparisons and caveats about these approaches:

A. Strategies (2) and (4) are appropriate for randomized trials or other situations in which it is reasonable to assume that the mean response is equal across groups at baseline. These methods should not be used elsewhere.

   This restriction on the domain of application for these methods is somewhat obvious for strategy (2). For strategy (4) perhaps some additional explanation is required.

   In model (*), suppose the response is height, and we have an observational study in which we follow a sample of children — both boys and girls — over time. The ANCOVA strategy (strategy (4)) would model post-baseline heights conditional on height at baseline. The mean response in this model is a conditional mean:

$$\mathrm{E}(y_{hij}|y_{hi1}, \mathbf{x}_{hij}) = \gamma y_{hi1} + \mathbf{x}_{hij}^T \boldsymbol{\beta}.$$

   Thus, $\boldsymbol{\beta}$ has a conditional interpretation and a test of group×time interaction in this model addresses the question of whether height changes more (or less) for boys than girls given that the subject has a particular height at baseline.

   This is quite different than the question addressed by either strategy (1) or (3). In those approaches, the group×time interaction addresses the question of whether the mean height gain over time differs between genders.

   - The former question makes a comparison between boys and girls conditional on the fact that they started out at the same height. Conditional on that fact, we would expect boys to grow more than girls because if the boys and girls have the same height at baseline, then we are talking about the future growth of a baseline population consisting of tall girls and/or short boys.

   - The latter question involves no such conditioning.

B. Strategy (1) or (3) should be used for non-randomized groups that are not equivalent at baseline. The choice between these two approaches can be made based on practical grounds because they are essentially equivalent approaches.

   There are two main practical considerations:

   i. First, to implement strategy (3) it is necessary to form the change scores for each subject. This is impossible for subjects with missing responses at baseline, so these subjects must be omitted from the analysis. In contrast, all subjects may be analyzed in strategy (1).

   ii. Second, the interpretation of the main effects and interactions in the two analyses differs.

      In strategy (1), the test of group effects is not of scientific interest. Groups may differ marginally solely because of differences at baseline, not because of any treatment effect or other post-baseline group difference. Only the group×time interaction addresses the question of whether there are any groups differences other than pre-existing ones.

      In strategy (3), we model the mean change score, so the group main effect compares the mean change from baseline across groups, the time main effect examines whether the mean change from baseline is constant over post-baseline measurement times, and the group×time interaction addresses whether the mean change profiles over time are parallel across groups.

      These tests are of interest in and of themselves, but the group×time interaction from strategy (1) is the "usual" test of main interest. This test can be recovered from the strategy (3) analysis as a joint test of group and group×time.

C. When it is appropriate to assume equal means across groups at baseline, strategies (2) and (4) offer greater efficiency relative to strategies (1) and (3).

   – Intuitively, this can be seen fairly easily by comparing strategies (1) and (2). These approaches are essentially identical except that strategy (2) assumes equal means at baseline and strategy (1) does not. Roughly speaking, assumptions "buy" power and efficiency and "pay for" these advantages by sacrificing generality and robustness.

   – Our text demonstrates the efficiency gain of the ANCOVA (strategy (4)) over strategy (3) more formally. Another way to think about the difference between strategies (3) and (4) is as follows:

     The analysis of change score in strategy (3) fits a model of the form

     $$y_{hij} - y_{hi1} = \mathbf{x}_{hij}^T \boldsymbol{\beta} + \mathbf{z}_{hij}^T \mathbf{b}_{hi} + \varepsilon_{hij}.$$

     Note that this model can be re-written as

     $$y_{hij} = \gamma y_{hi1} + \mathbf{x}_{hij}^T \boldsymbol{\beta} + \mathbf{z}_{hij}^T \mathbf{b}_{hi} + \varepsilon_{hij}, \qquad (**)$$

     where $\gamma = 1$.

     The repeated measures ANCOVA model estimates γ rather than assuming $\gamma = 1$. Therefore, it subsumes the gain score model, and is guaranteed to fit at least as well as the gain score model at the cost of only one additional d.f.

D. In situations where strategies (2) and (4) are both appropriate, the choice between them can be based on practical grounds and personal preference.

   - Strategy (4) requires omission of subjects for whom the baseline response is missing.

   - Strategy (4) makes an implicit assumption that $\text{cov}(y_{hi1}, y_{hij})$ is constant for $j = 2, \ldots, t_{hi}$. However, this can, and typically should be relaxed by allowing γ to change over time by replacing γ with γ_j or even γ_{hj}.

- More generally, any covariate can be handled this way. We first include it and allow its slope to depend upon the treatment structure, then simplify this assumption .

More generally (that is, when controlling for any type of covariate, not necessarily a baseline value), the basic RM-ANCOVA model is as follows:

$$y_{hij} = \mu + \alpha_h + \beta_j + (\alpha\beta)_{hj} + \gamma_{hj} w_{hij} + b_{hi} + \varepsilon_{hij}$$

where $\{b_{hi}\} \overset{iid}{\sim} N(0, \sigma_b^2)$ independent of $\{\boldsymbol{\varepsilon}_{hi}\} \overset{ind}{\sim} N(\mathbf{0}, \mathbf{R}_{hi})$.

- In some cases, we might instead consider a random slope and intercept model where we replace b_{hi} above with $b_{1hi} + b_{2hi} w_{hij}$. This would make particularly good sense when the covariate w does not vary over time.

- The RM-ANCOVA model above does not assume parallel slopes across groups and time points in the relationship between y and w, but instead allows this slope to differ across both groups and time. Typically, this is more complexity than is needed, but we also want to avoid assuming parallel slopes without checking that assumption.

- Therefore, one reasonable strategy is to allow γ_{hj} to depend on h and j and then reduce this model by testing whether γ is constant over h, j or both.

Sheep Example (again) — Comparison of Baseline Strategies:

- See sheep4.sas. In this program we handle the baseline value with all four methods we have discussed. See the comments in the SAS program for description of the methods and relationships among them.

Although there are pros and cons to all 4 methods of handling baseline values, I tend to prefer the RM-ANCOVA approach for situations in which it is applicable (i.e., randomized assignment to groups, no missing baseline values). Here is another example.

Blood Pressure Example — RM-ANCOVA Model:

- In this example*, we consider a repeated measures study of the effects of a drug and exercise program on blood pressure.

A medical team designed an experiment to investigate the effects of a drug and an exercise program on a person's systolic blood pressure (bp). 32 subjects with marginal to high systolic bp were randomly assigned to one of four combinations of exercise and drug regimes (exercise=no, drug=no; exercise=no, drug=yes; exercise=yes, drug=no; and exercise=yes, drug=yes). Eacher person returned for a bp measurement each week for 6 weeks following the onset of treatment. In addition, a baseline (week 0) initial blood pressure (ibp) measurement was taken on each person, and it was believed that a person's ibp may affect how they respond to the treatments.

Let y_{hij} represent the bp at the j^{th} week for the i^{th} subject at the h^{th} combination of drug and exercise ($j = 1, \ldots, 6$, $h = 1, 2, 3, 4$), and let y_{hi0} represent ibp for the h, i^{th} subject.

- In bloodpress1.sas, the first call to PROC MIXED fits model (†) to these data using REML estimation and assuming $\mathbf{R}_{hi} = \sigma^2\mathbf{I}$ for all h, i. Call this model 1a.

* Taken from Milliken and Johnson, *Analysis of Messy Data, Vol. III*

- Note that while model (†) is less restrictive than model (**), it still makes the assumption that the relationship between bp and ibp is linear. It is possible that this relationship is nonlinear, so before fitting (†), plots of bp versus ibp are produced to check the adequacy of the linearity assumption. In this example, it seems justified.

- Next we fit the model

  $$y_{hij} = \mu_{hj} + \beta_{hj} y_{hi0} + \varepsilon_{hij}, \quad \text{var}(\boldsymbol{\varepsilon}_{hj}) = \theta_1 \mathbf{J} + \theta_2 \mathbf{I} \quad \text{(compound symmetry)}$$

  using the REPEATED statement rather than the RANDOM statement. This model we call model 1b.

  - Note that models 1a and 1b aren't exactly the same. Model 1a is a hierarchical model which implies a model of the form 1b, where $\theta_1, \theta_2 \geq 0$ (θ_1, θ_2 are variance components in model 1a). However, the marginally specified model 1b doesn't require $\theta_1, \theta_2 \geq 0$. It only requires $\text{var}(\boldsymbol{\varepsilon}_{hi})$ to be p.s.d., which translates into the condition:

    $$\frac{\theta_1}{\theta_1 + \theta_2} \geq -\frac{1}{\max_i(t_i) - 1}.$$

  - I.e., models 1a and 1b differ in that θ_1 is the non-negative variance of b_{hi} in 1a, but can be negative in 1b. In our example, θ_1 is estimated to be positive, so the two models give the same results.

- Next we fit several models with the same mean structure as in models 1a and 1b, but with various different var-cov structures. The var-cov structures considered and the resulting AIC and BIC values (in smaller is better form) for these structures are given below.

Model Number	Form of $\mathbf{R}$	Subject Effects?	AIC	BIC
1a	$\sigma^2\mathbf{I}$	Yes	900.9	903.9
1b	Compound Symmetry	No	900.9	903.9
2	Heterogeneous CS	No	907.1	917.3
3	AR(1)	Yes	870.0	874.4
4	Hetero AR(1)	Yes	863.4	875.1
5	ANTE(1)	No	870.7	886.8
6	Toeplitz	No	875.2	884.0
7	Hetero Toeplitz	No	880.0	896.1
8	Unstructured	No	878.0	908.8

- According to AIC, the best var-cov structure is provided by model 4. Therefore, we adopt the heterogeneous AR(1) structure with random subject effects and proceed to simplifying the mean structure.

- Because this is a designed experiment with relative few measurement occasions, we really don't want to impose much structure on the model for the mean response. In fact, if we did not have a covariate to control for, we'd probably adopt a full cell-means type mean structure: $\mathrm{E}(y_{hij}) = \mu_{hj}$.

- Therefore, all we really want to do with the mean structure is to simplify the portion of the model that controls for the covariate (the $\beta_{hj}y_{hi0}$ part of the mdoel).

    - If we were dealing with observational data, we might instead be trying to build the most parsimonious model that adequately explains the data. In that case, we might choose to simplify the non-covariate part of the model (the μ_{hj} part, in our example) as well.

- To determine how the covariate part of the model may be over-specified, we refit model 4 with $\beta_{hj}y_{hi0}$ broken apart into separate components corresponding to two, three, and four-way interactions between ibp and exercise, drug, and time. This does not change the model at all, but allows us to determine where the significant interactions are (i.e., does the slope of bp on ibp depend upon exercise, drug, time, and which combinations of these factors?).

- In reducing the mean strucure, it usually makes sense to restrict attention to hierarchical models. Hierarchical models are models in which interactions are included only if all lower-order interactions contained in that interaction are included as well.

    - An example of a non-hierarchical model is one in which the two-way interaction A*B is included, but main effects of A are not.

- Therefore, to reduce the mean of the model, we will eliminate insignificant terms one at a time, where terms with the highest $p-$values are removed first, but only when that elimination yields a hierarchical model.

- This leads to first removing the four-way interaction ibp*drug*exercise*time. Then the following terms are eliminated in order: ibp*exercise*time, ibp*exercise*drug, ibp*time*drug, ibp*exercise, ibp*drug.

- This yields model 4g, our final model, in which all remaining terms are significant:

$$y_{hij} = \mu_{hij} + \beta_j y_{hi0} + b_{hi} + \varepsilon_{hij},$$

  where var($\boldsymbol{\varepsilon}_{hj}$) is of a heterogeneous AR(1) form.

- In this model, there is a significant three-way interaction between drug, exercise and time, and there is a significant effect of the baseline value ibp that depends upon time. Because of the significant covariate, we can't simply estimate a mean response under a certain treatment by time combination. The mean response depends upon the value of ibp! So, we must estimate

$$\mathrm{E}(y_{hij}|y_{hi0} = c) = \mu_{hj} + \beta_j c$$

  for one or more values of c.

  - Note that estimation at $c = 0$ makes no sense at all.

- What values of c might we consider? One natural strategy is to estimate the mean response for subjects with average, low, and high values of ibp. We could do this by considering c equal to the mean ibp level observed in the study and also equal to the mean $\pm$ 1 s.d. Alternatively, we might set c equal to each of the quartiles of ibp.

  - This can be done with with the AT option on the LSMEANS statement. The resulting lsmeans are on pp.7–8 of bloodpress1.lst.

- Profile plots of the LSMEANS for each treatment over time estimated at each value of ibp show the nature of the three-way interaction between exercise, drug, and time (see last 3 pages of handout).

- The DIFF option on the LSMEANS statement allows pairwise and other types of comparisons between the treatment means. In the presence of a significant three-way interaction between exercise, drug and time, I decided that I would make pairwise comparisons between the exercise=drug="No" condition (we can think of this as a control condition) and each other combination of exercise and drug, made separately at each time point.

- To do this in SAS is tricky: it can be done using the AT option, but the AT option does not work for CLASS variables, so we need to refit our model, treating still treating time as a factor, but implementing it via dummy variables, rather than by using the CLASS statement. Then, the AT statement can be used to fix the time and value of ibp at which we want all pairwise comparisons with the control treatment. This generates three pairwise comparisons at each timepoint and ibp level.

  - To control for the multiple comparisons problem induced by conducting these three pairwise comparisons we can use the Dunnett procedure which controls the strong family-wise error rate for the family of the three pairwise comparisons with the control treatment at each time point by ibp level.

  - Alternatively, we may prefer to control the error rate for a larger family (e.g., all of the pairwise comparisons we are going to do in the entire analysis of these data). Doing so would be trickier still, but not impossible; e.g., it could be done using the Bonferroni method combined with Dunnet's approach, but we do not pursue this issue here.

Modeling the effect of time.

Thus far, most of the LMMs we have considered for longitudinal data have allowed the mean profile over time to vary arbitrarily.

- That is, we have often used saturated models where we've allowed a distinct mean response for every group by time combination.

Such a model makes no assumption about the relationship between the response and time. Therefore, it is quite generally valid.

However, there are several situations where modelling the effect of time in terms of some more parsimonious, often smooth, functional form is a better choice.

- When the number of measurement occasions t is moderate to large (e.g., more than 4 or 5).
  - In such situations allowing arbitrarily varying means at each time point is often unnecessary because there is an underlying more parsimonious pattern of change that may be modeled yielding greater insight into the effect of time.
  - With many measurement occasions we rarely are interested in analyzing differences between groups at every time point. A full time profile model becomes unnecessarily unwieldly to understand, summarize and work with.
  - The group$\times$time interaction test in such a model is an omnibus test that looks at whether group differences differ at any two time points. It is not directed at any particular hypothesis (e.g., the rate of increase through time is faster for one group than another) so it lacks power for specific forms of interaction, and typically requires subsequent more detailed hypothesis testing to understand the specific nature of the interaction.

- When the timing of the measurements is not the same for all subjects.
  - In this case it makes no sense to estimate the mean response for all subjects (or for a group of subjects) at a particular measurement occasion. That measurement occasion may be occasion number 3 (say) for all subjects, but it may have occurred at week 7 following treatment for some subjects and week 10 for others.
  - In this case, the mean response should be modeled as a function of the elapsed time since the beginning of the study or from some other reference point.
- In some situations all subjects do not have the same time origin and elapsed time since an origin common to all subjects is not the operative *metameter* for time.
  - This can occur when we are interested in modelling growth and subjects enter into the study at different ages, when we are interested in studying sexual development among adolescent girls since menarche (the time of first menstrual cycle), etc.

In these situations it is typically more appealing to model the effects of time (or age, or whatever the metameter is) through low order polynomials (linear or quadratics), or other simple functional forms.

Example — Six Cities Study of Air Pollution and Health *

The 6 Cities Study was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and the factors that influence such growth. A cohort of 13,379 first and second grade children from 6 US cities was enrolled and annual pulmonary function measurements obtained on each child until high school graduation or loss to follow-up. Among the outcomes was FEV1, forced expiratory (air) volume during the first second of a breathing task.

Here we analyze a random sample of 300 girls from one of the cities in the study. The data consist of measurements of FEV1, height and age on these children. Note that children were measured at different measurement occasions, were recruited into the study at different baseline ages, and have varying numbers of measurements through time (ranging from 1 to 12 observations). Therefore, an analysis of response profiles over time is inappropriate here. Instead, we use age as the metameter for time. In addition, we follow FLW and analyze log(FEV1) rather than FEV1 on its original scale.

- See handout FEVExample1. In this SAS program we first plot the data against age and against log(height). Individual profiles are connected for each girl. There appears to be an increasing trend with age and log(height) that may not be strictly linear.

- Next we plot log(FEV1) versus both baseline age and log baseline height. These plots suggest approximately linear relationships.

* See §8.8 of Fitzmaurice, Laird and Ware (our textbook). Henceforth I'll refer to these authors as FLW.

- Next we fit a model suggested by FLW:

$$\begin{aligned} y_{ij} = \beta_1 + \beta_2 \text{age}_{ij} + \beta_3 \log(\text{height})_{ij} + \beta_4 \text{age}_{i1} \\ + \beta_5 \log(\text{height})_{i1} + b_{1i} + b_{2i} \text{age}_{ij} + \varepsilon_{ij}, \end{aligned} \qquad (M1)$$

  where y_{ij} is the response at measurement occasion j for subject i and

$$\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \overset{iid}{\sim} N\left(\mathbf{0}, \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_3 \end{pmatrix}\right), \quad \text{indep. of} \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \theta_4)$$

  Here, notice that we have current age (age_{ij}) as well as baseline age (age_{i1}) in the model to capture both longitudinal and cross-sectional effects of age. The same is true for log(height).

- Notice that model M1 is not saturated in either the mean or the variance-covariance structure. In previous examples we have developed LMMs by first fitting maximally-complex models and simplifying. In this case, this approach is not so easily implemented.

  - Firstly, the set of measurement occasions is not common for all subjects, and age, height and their baseline values are all continuous covariates, so we could consider a maximal model by specifying high-order polynomials for all these covariates, but there is no saturated model in the mean.

  - Secondly, because the set of measurement occasions differs across subjects, the unstructured specification of $\mathbf{V}$ (e.g., through the omission of all random effects and the use of TYPE=UN to specify $\mathbf{R}$) is also not available. (This is because $\text{cov}(\varepsilon_{ij}, \varepsilon_{ik})$ does not have the same meaning for two different values of i).

Instead, we start with a relatively simple model for the mean suggested by the intial plots of the data, and a random intercept and slope specification to determine $\mathbf{V}$.

- We have seen that, for uncorrelated errors, random intercepts imply compound symmetry if the errors are homoscedastic and, if the errors are heteroscedastic, we still get equi-correlation among all pairs of observation that share the random intercept (e.g., all obs from the same subject).

  What does the random intercept and slope model imply?

(See §8.4 of McCulloch, Searle and Neuhaus) Consider a simple model of the form

$$y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})x_{ij} + \varepsilon_{ij},$$

where

$$\mathbf{D} = \text{var}(\mathbf{b}_i) = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix}, \quad \mathbf{R}_i = \text{var}(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{I}, \quad \forall i.$$

For this model, it can be shown that $\mathbf{V}_i = \text{var}(\mathbf{y}_i)$ has diagonal elements

$$V_{ijj} = \sigma_0^2 + 2\sigma_{01}x_{ij} + \sigma_1^2 x_{ij}^2 + \sigma^2,$$

(heteroscedastic, depending on x) and off-diagonal elements

$$V_{ijk} = \sigma_0^2 + \sigma_{01}(x_{ij} + x_{ik}) + \sigma_1^2 x_{ij}x_{ik}.$$

- If we re-scale V_{ijk} as a correlation rather than a covariance, we get $\text{corr}(y_{ij}, y_{ik}) = V_{ijk}/\sqrt{V_{ijj}V_{ikk}}$, which can be shown to

  i be monotonically decreasing in $|x_{ij} - x_{ik}|$, but not to 0, and

  ii yield a correlation of 1 in the limit as $x_{ij} \to \infty$ for fixed $x_{ij} - x_{ik}$. I.e., two observations that lie along a given subject-specific linear trend in x become perfectly correlated as you push that pair of observations further out to the right (make them both have bigger x values).

- Taking x to be time, property (i) says that serial correlation decays with the time lag between observations, but never goes away completely. Empirically, this pattern often holds for longitudinal data, whereas the decay of an AR(1) structure, say, which goes to 0, is often too rapid and extreme.

Model M1 contains both baseline values of the covariates and their current values.

To understand the difference between the cross-sectional and longitudinal effects of a covariate, consider the marginal mean implied by (M1):

$$\text{E}(y_{ij}) = \beta_1 + \beta_2 \text{age}_{ij} + \beta_3 \log(\text{height})_{ij} + \beta_4 \text{age}_{i1} + \beta_5 \log(\text{height})_{i1}$$

This model implies

$$\begin{aligned}\text{E}(y_{i1}) &= \beta_1 + \beta_2 \text{age}_{i1} + \beta_3 \log(\text{height})_{i1} + \beta_4 \text{age}_{i1} + \beta_5 \log(\text{height})_{i1} \\ &= \beta_1 + (\beta_2 + \beta_4)\text{age}_{i1} + (\beta_3 + \beta_5) \log(\text{height})_{i1}\end{aligned}$$

and

$$\begin{aligned}\text{E}(y_{ij} - y_{i1}) =& \beta_1 + \beta_2 \text{age}_{ij} + \beta_3 \log(\text{height})_{ij} + \beta_4 \text{age}_{i1} + \beta_5 \log(\text{height})_{i1} \\ &-\{\beta_1 + \beta_2 \text{age}_{i1} + \beta_3 \log(\text{height})_{i1} + \beta_4 \text{age}_{i1} + \beta_5 \log(\text{height})_{i1}\} \\ =& \beta_2(\text{age}_{ij} - \text{age}_{i1}) + \beta_3\{\log(\text{height})_{ij} - \log(\text{height})_{i1}\}.\end{aligned}$$

- From these equations it becomes clear that the coefficient on current age β_2 has an interpretation as the effect of a unit increase in age on the expected change in the response variable (for any given change in log(height)).

  - This is the longitudinal effect of age (conditional on whatever other variables are in the model — in this case log(height) and baseline log(height)).

- The cross-sectional effect of age (the effect of differences in age at baseline within the cohort of subjects) is $\beta_2 + \beta_4$. Therefore, β_4, the coefficient on baseline age, has an interpretation as the difference between the longitudinal and cross-sectional effects of age (again, conditional on whatever else is in the model).

Back to the example:

- For the moment let's assume that our variance-covariance and mean specification in this model are adequate and consider the interpretation of our fixed effect parameter estimates:

  - Notice that there appears to be a significant difference between the longitudinal and cross-sectional effects of age here (p=.0270), but not for height (p=.1340).

  - The cross-sectional effect of age is estimated to be $\hat{\beta}_2 + \hat{\beta}_4 = .02353 - .01651 = .00702$. This implies that for girls of any given baseline height, we can expect an average increase of .00703 in y for a 1 year difference in age. Since y is log(FEV), this corresponds to a .7% increase in FEV1 ($e^{.00702} = 1.007$).

  - The longitudinal effect of age is estimated to be $\hat{\beta}_2 = .02353$ on the log(FEV1) scale which translates to $e^{.02353} = 1.024$ or approximately 2.4% change in FEV1 for 1 year of aging for a given change in height over that time.

  - One way to interpret the coefficient on $\log(\text{height})_{ij}$ is to consider the effect of a 10% increase in height. This corresponds to a $\log(1.1) \approx .1$ change in log(height), therefore a 10% increase in height for a given change in age is associated with a .224 increase in E{log(FEV1)}, or a 25% increase in the median value of FEV1 ($e^{.224} = 1.25$).

- Note the option vcorr=35 requests the $\mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \mathbf{R}_i$ matrix, rescaled as a correlation matrix for subject number 35. This subject happened to have measurements at every integer age between 7 and 18, so we can see what the correlations through time fit by this model look like.

- This correlation matrix appears on pp.3–4 of the output. It is apparent that the correlations through time decay, but not to zero here (property (i)), and the correlations are largest near the bottom right corner of this matrix (property (ii)).

- FLW also consider one alternative model for these data in which we replace the random slope on age with a random slope on log(height). That is, they also consider the model

$$\begin{aligned} y_{ij} = \beta_1 + \beta_2 \text{age}_{ij} + \beta_3 \log(\text{height})_{ij} + \beta_4 \text{age}_{i1} \\ + \beta_5 \log(\text{height})_{i1} + b_{1i} + b_{2i} \log(\text{height})_{ij} + \varepsilon_{ij} \end{aligned}$$

- This model is fit in the second call to PROC MIXED. Since it has the same mean as the previous model, we can compare these two models via their AIC values, or equivalently (since they have the same number of parameters) via their restricted loglikelihoods. According to these criteria, model 2 is preferred (AIC of -4581.5 for model 2 versus -4559.5 for model 1).

- This is as far as FLW took this example. However, I was a bit curious that we ended up with a model that was linear in age and log(height) given the initial data plots, which seemed to be somewhat nonlinear. Therefore, I took a look at the residuals from model 1 and plotted them versus age.

- Notice that this plot looks very poor. There seems to be a distinct wave in the residual scatter. This implies mean misspecification, and the need for higher order terms in age, or some other means of modelling the longitudinal effect of age/time on the response.

- In Model 3, we expand the mean specification from model 1 by including quadratic, cubic and quartic terms in age. We also include a random quadratic effect of age. That is, model 3 becomes:

$$\begin{aligned} y_{ij} = \beta_1 + \beta_2 \text{age}_{ij} + \beta_3 \log(\text{height})_{ij} + \beta_4 \text{age}_{i1} + \beta_5 \log(\text{height})_{i1} \\ + \beta_6 \text{age}_{ij}^2 + \beta_7 \text{age}_{ij}^3 + \beta_8 \text{age}_{ij}^2 + b_{1i} + b_{2i} \text{age}_{ij} + b_{3i} \text{age}_{ij}^2 + \varepsilon_{ij} \end{aligned}$$

- In this model, quadratic through quartic terms in age are highly significant. In addition, the residuals from this model versus age now look much better. So, for the moment we accept this mean specification and consider alternative models for the covariance.

- Models 3a–3c fit various alternative variance-covariance structures. None of these offers any improvement over model3 according to the AIC criterion. Therefore, we settle on model 3.

In this example we end with a model in which the effects of age/time cannot be modelled adequately via low order polynomials. When it is necessary to include cubics, quartics, etc., model interpretability and parsimony suffers. In such situations it may be preferrable to adopt a different (non-polynomial-based) strategy for modelling changes through time.

- In this case, it may be that a nonlinear model — e.g., an asymptic regression model or sigmoidal growth curve — may fit the data better, more parsimoniously, and be more interpretable.

- Alternatively, we might consider other linear models that account for time effects more flexibly. For example, spline models may be a good choice here. We will return to this example later to illustrate this approach.

Flexible Modeling of the Mean in LMMs via Splines

As compared with ordinary polynomials, considerably greater flexibility can be achieved by modelling the effects of time (or any other variable) via splines.

- **Splines** are curves that are formed by joining or tieing together several low order polynomials. The curves that are joined together are all of the same order and ar typically linear, quadratic, or at most cubic.

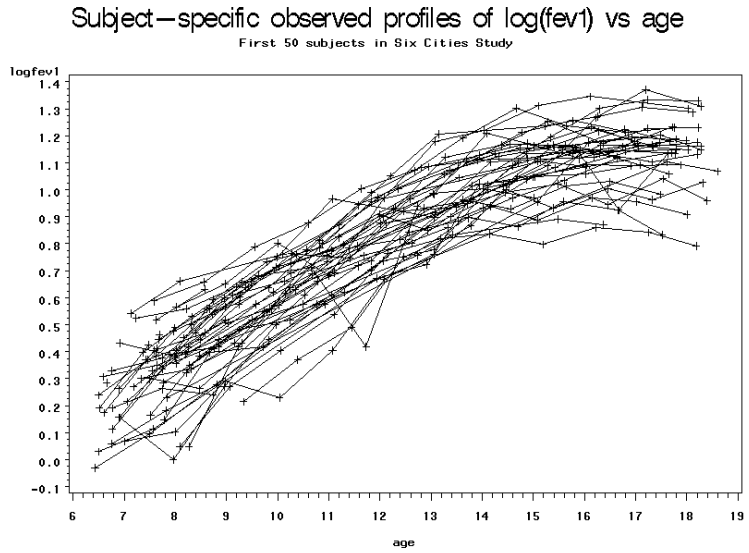
- The locations at which the component curves are joined to form the spline are known as **knots.**

- The number and location of the knots can be taken to be fixed (known) or unknown. Predictably, things are easier when these quantities are known and we will concentrate on this case. In practice, unless only a few knots are used, the choices of knot location and number are not terribly crucial and simple rules of thumb about knot specification often work well.

Linear Splines:

The simplest type of spline pieces together straight lines to form a piecewise linear curve.

For example, recall the FEV1 data from the Six Cities Study of Air Pollution. Some of the raw data from this study appear below:



Here, log(FEV1) appears to increase linearly until about age 14, leveling off thereafter. A model for the mean response that reflects this hypothesized relationship is

$$\mathrm{E}(y_{ij}) = \beta_1 + \beta_2 x_{ij} + \beta_3 (x_{ij} - \kappa)_+ \qquad (*)$$

where y_{ij} is the log(FEV1) measurement at the jth measurement occasion for the ith subject, x_{ij} is the corresponding age, κ is a single knot location and

$$(w)_+ = \begin{cases} w, & \text{if } w > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that this model implies linearity in x both before and after the knot with distinct slopes and intercepts:

$$\mathrm{E}(y_{ij}) = \begin{cases} \beta_1 + \beta_2 x_{ij} & \text{if } x_{ij} \leq \kappa, \\ (\beta_1 - \beta_3\kappa) + (\beta_2 + \beta_3)x_{ij} & \text{if } x_{ij} \geq \kappa, \end{cases}$$

with continuity at the knot (plug in $x_{ij} = \kappa$ in each case).

Such a model can be extended to account for group structure in the data as well. Suppose (y_{hij}, x_{hij}) is the (response, age) pair at the jth measurement occasion, for the ith subject in the hth group, $h = 1, \ldots, s$. Then we can extend model (*) as follows:

$$\mathrm{E}(y_{hij}) = \beta_{1h} + \beta_{2h}x_{hij} + \beta_{3h}(x_{hij} - \kappa)_+ \qquad (**)$$

- Based on model (**), the null hypothesis of no group differences in patterns of change over time is given by

$$H_0 : \{\beta_{21} = \cdots = \beta_{2s} \quad \text{and} \quad \beta_{31} = \cdots = \beta_{3s}\}$$

The two-piece linear or "broken-stick" model can be extended by the inclusion of 2 or more knots, $\kappa_1, \ldots, \kappa_K$. Such a model with K knots consists of $K + 1$ joined line segments.

- In practice, unless the goal is a nonparametric estimate of the functional relationship between y and x (i.e., "smoothing"), it is often adequate (and best) to restrict attention to $K = 1$ or $K = 2$ well-chosen knots.

- The exact choice of knot location is ideally done via a combination of subject-matter considerations and objective statistical criteria. In some cases, the analysis may not be particularly sensitive to knot location. In such cases, arbitrary knot location choices can be defended by reporting the results of a sensitivity analysis.

- For $K = 1$, ML estimation of κ can done by *profiling* this parameter. That is, the model can be fit via ML over a grid of κ values and the κ that produces the largest maximized loglikelihood overall will be the MLE. This can extended to higher K as well, although it quickly becomes computationally infeasible.

Example: Six Cities Study

- See the handout labelled FEVExample2.sas. In this SAS program we first plot the data from the first 50 subjects before fitting a series of broken stick models.

- In Model 1, we assume the response y_{ij}, the log-transformed FEV1 value at the jth measurement occasion for the i subject, follows the model

$$\text{Model 1:} \quad y_{ij} = (\beta_1 + b_i) + \beta_2 x_{ij} + \beta_3 (x_{ij} - \kappa)_+ + \varepsilon_{ij}$$

  where $b_1, \ldots, b_n \overset{iid}{\sim} N(0, \sigma_b^2)$, the ε_{ij}'s are $\overset{iid}{\sim} N(0, \sigma^2)$, x_{ij} is age, and $\kappa = 14$.

  - This model accounts for within subject-correlation through the inclusion of a random subject-specific intercept b_i.

  - The knot location $\kappa = 14$ years of age was chosen subjectively "by eye". However, it is reasonable to expect that a growth curve like this might level off near the age of 14 when many girls are reaching their adult physical size.

- The fit of the predicted responses $\hat{y}_{ij}$ from Model 1 versus the raw data from three subjects (subject IDs 16, 18, and 35) is examined in the second–fifth plots produced in FEVExample2.sas.

- These plots look fairly good, but the random intercept model only allows them to be shifted up or down from subject to subject. That is, they are parallel for all subjects. We relax the assumption of parallelism in Model 2:

  Model 2: $y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})x_{ij} + (\beta_3 + b_{3i})(x_{ij} - \kappa)_+ + \varepsilon_{ij}$

  where $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i})^T \sim N(\mathbf{0}, \mathbf{D})$, $\mathbf{D}$ a symmetric positive definite matrix.

- This model allows random intercepts and slopes both before and after the knot location which vary from subject to subject.

- According to AIC, this model fits much better than Model 1 (-3976.2 for Model 1, -4190.7 for Model 2). In addition, the fitted curves for subjects 16, 18 and 35 look better (the 6th plot generated by the SAS program).

- Next, to investigate the knot location $\kappa = 14$ we refit model 2 with $\kappa = 13, 13.5, 14, 14.5, 15$ and compare the maximized likelihoods, L. Actually, we examine $-2 \log L$ and find that this quantity takes its minimum value at $\kappa = 14$.

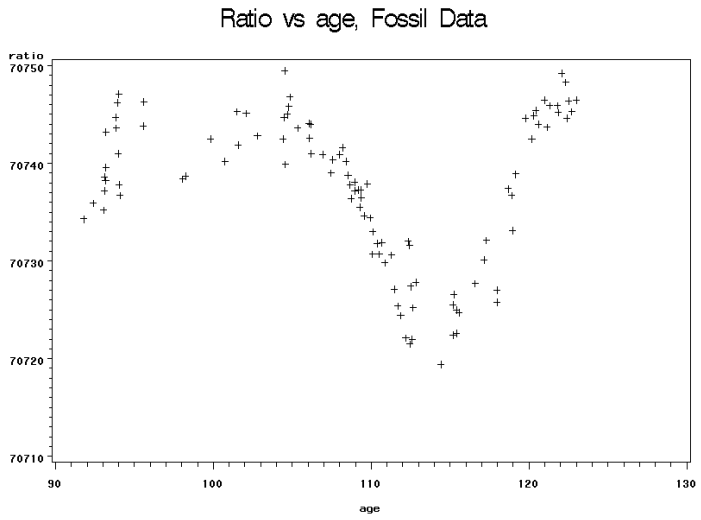
  - We could take this process further and obtain the MLE for κ by refining our grid of κ values. That is, we could examine more values between 13.5 and 14.5 to find the minimizer of $-2 \log L$. However, 14 is an appealing round number and is close to optimal, so we adopt it as our knot value.

- In Model 3 we add in the covariates baseline age, log(height) and log(baseline height) that were used in the models considered in FEVExample1. All of these covariates have t statistics that are highly significant.

- Finally, in Model 4 we consider a richer variance-covariance structure in which we assume a Gaussian spatial exponential residual covariance structure. Model 4 further reduces the AIC for the model to -4699.1 versus -4634.8 for Model 3.

Smoothing Via Splines and LMMs:

One way to achieve more flexibility in a spline model is to simply add knots. Consider the following data from a study in which strontium ratios were measured on 106 fossils of known ages:

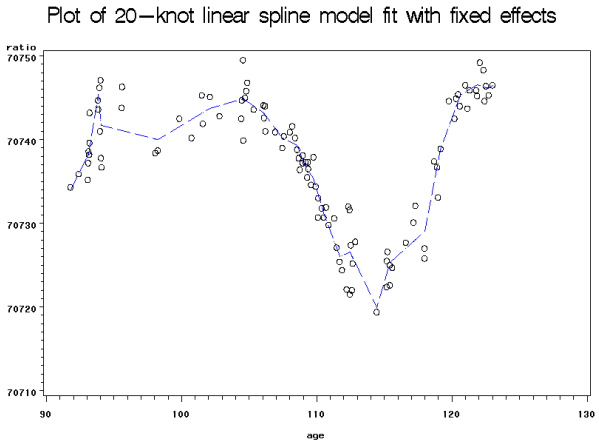


Clearly, there is a complex nonlinear relationship between the strontium ratio and fossil age. One way to achieve a flexible fit to these data is to use a spline with a relatively large number of knots.

The next plot gives the fitted curve corresponding to the $K = 20$ knot model:

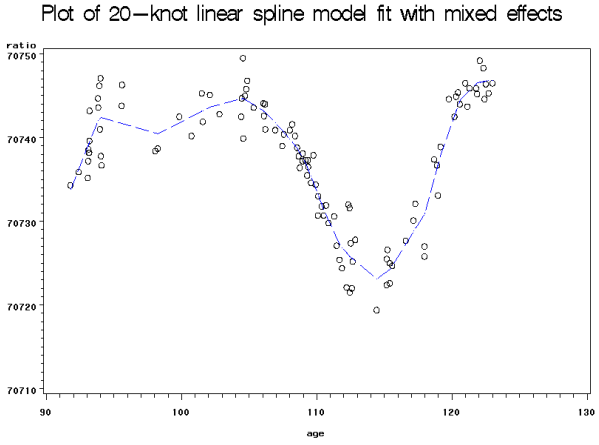
$$y_i = \alpha_1 + \alpha_2 x_i + \sum_{j=1}^{K} \beta_j (x_i - \kappa_j)_+ + \varepsilon_i \qquad (\heartsuit)$$

where (y_i, x_i) is the (strontium ratio, fossil age) pair for the ith fossil, $\alpha_1, \alpha_2, \beta_1, \ldots, \beta_K$ are fixed effects, and $\varepsilon_1, \ldots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$.



The next plot gives the fitted curve corresponding to exactly the same model, except that the fixed effects $\beta_1, \ldots, \beta_K$ are replaced by random effects $b_1, \ldots, b_K \overset{iid}{\sim} N(0, \sigma_b^2)$:

$$y_i = \alpha_1 + \alpha_2 x_i + \sum_{j=1}^{K} b_j (x_i - \kappa_j)_+ + \varepsilon_i \qquad (\clubsuit)$$



- By including a fairly large number of knots relative to the sample size, model ($\heartsuit$) achieves a flexible fit to the data.

- However, by treating the coefficients on the truncated line basis functions $(x_i - \kappa_j)_+$, $j = 1, \ldots, K$, as random rather than fixed, model ($\clubsuit$) achieves a much smoother, but still flexible, fit to the data than model ($\heartsuit$).

What's going on here? How do mixed models achieve this smoothing effect?

The answer lies in the connection between penalized least squares and (RE)ML/empirical BLUP fitting of the linear mixed model.

As we've seen previously, the (RE)ML estimator of $\boldsymbol{\beta}$ and EBLUP predictor of $\mathbf{b}$ in the LMM are given by the solution of the "Mixed Model Equations" (*cf* equation (*) on p.67):

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{X} & \mathbf{D}(\hat{\boldsymbol{\theta}})^{-1} + \mathbf{Z}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{y} \end{pmatrix},$$

where $\hat{\boldsymbol{\theta}}$ is either the ML or REML estimator of the variance-covariance parameter vector $\boldsymbol{\theta}$.

We've seen previously that the solutions of these equations yield the MLE of $\boldsymbol{\beta}$ and the empirical BLUP of $\mathbf{b}$. However, the mixed model equations can also be "derived" as the normal equations in the least-squares problem:

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta},\mathbf{b}} \begin{pmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b} \end{pmatrix}^T \begin{pmatrix} \mathbf{D}(\hat{\boldsymbol{\theta}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\hat{\boldsymbol{\theta}}) \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{b} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b} \end{pmatrix} \\ &\quad = \arg\min_{\boldsymbol{\beta},\mathbf{b}} \left\{ \mathbf{b}^T\mathbf{D}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{b} + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})^T\mathbf{R}(\hat{\boldsymbol{\theta}})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) \right\} \end{aligned}$$

- Such a problem is sometimes called a **penalized least squares** problem.

It might be obvious why this terminology is appropriate, but if not consider the special case in which $\mathbf{b}$ contains elements $b_1, \ldots, b_q \overset{iid}{\sim} N(0, \sigma_b^2)$ and $\mathbf{R} = \sigma^2\mathbf{I}$. In this case, the minimization problem becomes

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta},\mathbf{b}} \left\{ \frac{1}{\hat{\sigma}^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2 + \frac{1}{\hat{\sigma}_b^2}\|\mathbf{b}\|^2 \right\} \\ &\quad = \arg\min_{\boldsymbol{\beta},\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2 + \hat{\alpha}\|\mathbf{b}\|^2 \right\} \end{aligned}$$

where $\hat{\alpha} = \hat{\sigma}^2/\hat{\sigma}_b^2$ and $\|\mathbf{v}\|$ denotes the norm $\|\mathbf{v}\| = \sqrt{\mathbf{v}^T\mathbf{v}}$.

- Here, minimization of the least squares criterion $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2$ is subject to the penalty $\hat{\alpha}\|\mathbf{b}\|^2$ being imposed on the coefficients $b_1, \ldots, b_q$.

- Note that the first term $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}\|^2$ is the objective function in the fixed effect version of the model in which $\mathbf{b}$ is treated as fixed. Treating $\mathbf{b}$ as random results in coefficients $\mathbf{b}$ that tend to be shrunk toward zero, relative to those that would be obtained if $\mathbf{b}$ were fixed.

  - For this reason, EBLUPs are often called shrinkage estimates or shrinkage predictors.*

- The parameter $\alpha = \sigma^2/\sigma_b^2$ controls the degree of shrinkage in the $\mathbf{b}$ coefficients and it plays the role of a "bandwidth" in the smoothing done by the linear mixed model. The estimation of α (the bandwidth) is done automatically via REML or ML estimation in the fitting of the mixed model.

- In the two smoothed curves for the fossil data on p.144, it is clear that in the spline fit with the LMM (the second plot), the individual line sements that make up the piecewise linear curve have slopes that tend to be shrunk toward zero relative to those in the fixed effect model (the first plot). This results in a smoother fit to the data.

* See also §8.7 of FLW.

Truncated Polynomial Splines:

Linear splines have the advantage of simplicity. They are easy to understand, explain and interpret. However, they can be somewhat rough because they have sharp corners.

Often, smoother fits can be obtained with higher order truncated polynomial splines. For example, while a linear spline with K knots is generated by linear combinations of the basis functions

$$1, x, (x - \kappa_1)_+, \ldots, (x - \kappa_K)_+$$

a quadratic spline with K knots is generated by linear combinations of the basis functions

$$1, x, x^2, (x - \kappa_1)_+^2, \ldots, (x - \kappa_K)_+^2$$

- Such spline models are less jagged because the basis functions $(x - \kappa_1)_+^2, \ldots, (x - \kappa_K)_+^2$ each have continuous first derivatives (they are smooth), whereas $(x - \kappa_1)_+, \ldots, (x - \kappa_K)_+$ do not.

More generally, a pth order polynomial spline can be generated by the truncated power basis of degree p:

$$1, x, x^2, \ldots, x^p, (x - \kappa_1)_+^p, \ldots, (x - \kappa_K)_+^p$$

- In practice, people very rarely go beyond $p = 3$.

For a pth order polynomial spline, the truncated power basis is only one of many bases that can be used to generate the spline. Other commonly used spline bases are B-splines, radial bases, and Demmler-Reinsch bases.

- All of these bases are mathematically equivalent. The difference between them is in their computational properties. We will avoid discussion of these computational issues and restrict attention to the truncated power series bases.

Choice of Knot Number and Location:

To accomplish a smooth fit, the important thing about knot specification is that knots be dense in areas of high curvature. This can often be accomplished by fairly simple rules of thumb. For example, Wand (2002) recommends using $K = \min(n/4, 35)$ knots for a data set of size n with knots located as follows:

$$\kappa_k = \left(\frac{k+1}{K+2}\right)^{\text{th}} \text{ sample quantile of the unique } x\text{'s}, \quad k = 1, \ldots, K.$$

Fossil Example

- See handout Fossil1.sas. In this SAS program we use Wand's rule of thumb for knot number and location and fit several spline models to illustrate the difference between fixed and mixed versions of the spline model and different choices of basis functions.

- After plotting the data, Model 1 is fit. This corresponds to the linear spline model (♡) on p.141 in which all coefficients are fixed effects. This model has $K = 21$ knots located according to Wand's rule of thumb.

- Next, Model 2 corresponds to model (♣) on p.142. Again it is a linear spline model with the same 21 knots, but with mixed effect coefficients.

- Model 3 uses the same knots as Model 2, but changes to a quadratic spline model. Notice that this model is considerably smoother than the linear spline model with the same knots.

- Finally, in Model 4, we use functionality in SAS's PROC GLIMMIX to fit a spline model with radial basis functions. This approximates a thin-plate spline. For details see Ruppert, Wand and Carroll's book, Semiparametric Regression (2003, Ch.13) for details.

  - Note that PROC GLIMMIX is currently available as a stand-alone add-on to SAS that can be downloaded, along with its documentation, from SAS's web site (www.sas.com).

Smoothing the Effect of Time in Longitudinal Data Analysis:

One of the nice features of doing smoothing with mixed models is that within the LMM framework it is easy to extend this methodology to account for more complexities in the data by extending the model. To illustrate we return to the analysis of forced expiratory volume in the six cities study.

Six Cities Study Example — Round 3:

Although the one-knot linear spline models we fit in FEVExample2.sas fit the data fairly well and were fairly easily interpretted, we now return to this data set to see whether we can achieve a more flexible and closer fit to these data with penalized splines.

- See FEVExample3.sas. In this program we fit a series of linear spline models with $K = 35$ knots chosen via Wand's rule of thumb.

- In Model 1, we fit the model

$$\text{Model 1:} \quad y_{ij} = \beta_1 + \beta_2 x_i + \sum_{k=1}^{35} b_k (x_{ij} - \kappa_k)_+ + \varepsilon_{ij}$$

where $b_1, \ldots, b_K \overset{iid}{\sim} N(0, \sigma_b^2)$, independent of $\varepsilon_{11}, \ldots, \varepsilon_{nt_i} \overset{iid}{\sim} N(0, \sigma^2)$. In addition, (y_{ij}, x_{ij}) is the (log(FEV1), age) pair for the ith subject at the jth measurement occasion.

  - Although it captures the mean trend with age fairly well, this model takes no account of the heterogeneity among subjects or within-subject correlation. Therefore, it is clearly inadequate for modeling longitudinal data.

- In Model 2, a random subject-specific intercept is added:

$$\text{Model 2:} \quad y_{ij} = (\beta_1 + c_i) + \beta_2 x_i + \sum_{k=1}^{35} b_k (x_{ij} - \kappa_k)_+ + \varepsilon_{ij}$$

where $c_1, \ldots, c_n \overset{iid}{\sim} N(0, \sigma_c^2)$ independent of the b_k's and ε_{ij}'s which are defined as in model 1.

- In Model 3, subject-specific random intercepts and slopes are included:

$$\text{Model 3:} \quad y_{ij} = (\beta_1 + c_{1i}) + (\beta_2 + c_{2i}) x_i + \sum_{k=1}^{35} b_k (x_{ij} - \kappa_k)_+ + \varepsilon_{ij}$$

where $\mathbf{c}_i = \begin{pmatrix} c_{1i} \\ c_{2i} \end{pmatrix}$, $i = 1, \ldots, n$, are $\overset{iid}{\sim} N(\mathbf{0}, \mathbf{D})$, $\mathbf{D}$ a symmetric, positive definite (unspecified) matrix.

- Finally, Model 4 adds a spatial exponential variance-covariance structure:

$$\text{Model 4:} \quad \text{same as Model 3, with} \quad \boldsymbol{\varepsilon}_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{R}_i)$$

where $\mathbf{R}_i$ has ℓ, mth element $\sigma^2 e^{-d_{\ell m}^2/\rho^2}$ where $d_{\ell m}$ is the Euclidean distance between $x_{i\ell}$ and x_{im}.

- Models 1-4 are all fit with REML. They have the same fixed effects specifications so their likelihoods and AICs are comparable. Models 1-4 have AICs -1998.2, -3977.4, -4115.3, -4275.8, respectively. Of these four models, Model 4 has the lowest AIC.

- Fitted curves from this model for subjects 16, 18, and 35 are plotted at the end of FEVExample3.sas. These curves appear to fit well, although it is not clear that the extra flexibility allowed by this 35 knot penalized linear spline model results in a dramatically different or improved fit relative to the simple two-piece linear spline models fit in FEVExample2.sas.

Model Diagnostics and Remediation

- Read Ch.10 of our text (FLW). See also section 4.3 and chapter 5 of the book by Pinheiro and Bates.

In R, there are several tools available for fitting LMMs. I will discuss the nlme package, which was written by Pinheiro and Bates and discussed by these authors in their book (on reserve). The main tool in this package for fitting LMM is the lme() function.

- I find model diagnostics and residual analysis much easier in R with the nlme package, so we will switch our focus from SAS to R now.

Variance Functions Available in R (in the nlme package):

- Variance functions in the nlme software are described in §5.2.1 in Pinheiro and Bates (2000). Here, we give only brief descriptions.

1. varFixed. The varFixed variance function is $g^2(v_{ij}) = v_{ij}$. That is,

$$\text{var}(\varepsilon_{ij}) = \sigma^2 v_{ij},$$

   the error variance is proportional to the value of a covariate. This is the common weighted least squares form.

2. varIdent. This variance specification corresponds to different variances at each level of some stratification variable s. That is, suppose s_{ij} takes values in the set $\{1, 2, \ldots, S\}$ corresponding to S different groups (strata) of observations. Then we assume that observations in stratum 1 have variance σ^2, observations in stratum 2 have variance $\sigma^2\delta_1$, ..., and observations in startum S have variance $\sigma^2\delta_S$.

   That is,

$$\text{var}(\varepsilon_{ij}) = \sigma^2\delta_{s_{ij}}, \qquad \text{so that } g^2(s_{ij}, \boldsymbol{\delta}) = \delta_{s_{ij}}$$

   where, for identifiability we take $\delta_1 = 1$.

3. varPower. This generalizes the varFixed function so that the error variance can be a to-be-estimated power of the magnitude of a variance covariate:

$$\text{var}(\varepsilon_{ij}) = \sigma^2 |v_{ij}|^{2\delta} \qquad \text{so that } g^2(v_{ij}, \delta) = |v_{ij}|^{2\delta}.$$

   The power is taken to be 2δ rather than δ so that s.d.$(\varepsilon_{ij}) = \sigma |v_{ij}|^{\delta}$.

   A very useful specification is to take the variance covariate to be the mean response. That is,

$$\text{var}(e_{ij}) = |\mu_{ij}|^{2\delta}$$

   However, this corresponds to $g^2(\mu_{ij}, \delta) = |\mu_{ij}|^{2\delta}$ depending upon the mean. Such a model is fit with a variant of the ML estimation algorithm. However, this technique is not maximum likelihood, and indeed ML estimation is not recommended for such a model. Instead the method is what is known as *pseudo likelihood estimation*.

4. varConstPower. The idea behind this specification is that varPower can often be unrealistic when the variance covariate takes values close to 0. The varConstPower model specifies

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \left(\delta_1 + |v_{ij}|^{\delta_2}\right)^2, \quad \delta_1 > 0.$$

   That is, for $\delta_2 > 0$ (as is usual), the variance function is approximately constant and equal to δ_1^2 for values of the variance covariate close to 0, and then it increases as a power of $|v_{ij}|$ as v_{ij} increases in magnitude away from 0.

5. varExp. The variance model for varExp is

$$\text{var}(\varepsilon_{ij}) = \sigma^2 \exp(2\delta v_{ij})$$

6. varComb. Finally, the varComb class allows the preceding variance classes to be combined so that the variance function of the model is a product of two or more component variance functions.

Correlation Structures Available in R (in the nlme package):

- The nlme package includes correlation structures to account for time dependence (serial correlation structures) and spatial dependence (spatial correlation structures). It also has a couple of generally applicable correlation structures.

- Correlation structures in the nlme software are described in §5.3 in Pinheiro and Bates (2000). Here, we give brief descriptions of the serial and general correlation structures.

Serial Correlation Structures:

The work-horse class of models in time-series analysis is the class of **Autoregressive-Moving Average** (ARMA) models.

We will apply these models to the errors, ε_{ij}, but for notational convenience let's index ε by t to indicate time and drop the subject index i.

In an Autoregressive (AR) model, we assume the current observation ε_t is a linear function of previous observations plus "white noise" (a mean zero, constant variance error term):

$$\varepsilon_t = \phi_1\varepsilon_{t-1} + \cdots + \phi_p\varepsilon_{t-p} + a_t, \quad \mathrm{E}(a_t) = 0, \mathrm{var}(a_t) = \sigma^2.$$

- The number of previous observations on which ε_t depends, p, is called the order of the process and we write $AR(p)$.

- The simplest, and practically most useful, AR model is an AR(1):

$$\varepsilon_t = \phi\varepsilon_{t-1} + a_t, \quad \text{where } -1 < \phi < +1.$$

- For an AR(1) model,

$$\mathrm{corr}(\varepsilon_t, \varepsilon_s) = \phi^{|t-s|}$$

  and ϕ represents the correlation between two observations one time unit apart.

A Moving-Average model is one in which the observation ε_t at time t is a linear combination (weighted average, in some sense) of past independent and identically distributed white noise error terms plus a current time white noise error:

$$\varepsilon_t = \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} + a_t$$

- The number of past errors on which ε_t depends is the *order* of the process, so above we have an MA(q) process.

- Again, an order 1, in this case MA(1), process is often useful. For an MA(1),

$$\text{corr}(\varepsilon_t, \varepsilon_s) = \begin{cases} 1, & \text{if } s = t; \\ \theta_1/(1+\theta_1^2) & \text{if } |s-t| = 1; \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

- In general, MA(q) processes have nonzero correlation for observations $\leq q$ time units apart and 0 correlation for observations $> q$ time units apart.

Combining an AR(p) process with a MA(q) process we get an ARMA(p, q) process:

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + \sum_{j=1}^{q} \theta_j a_{t-j} + a_t.$$

- It is always possible to model any autocovariance structure to an arbitrarily small level of precision with a high enough order AR or MA process. Often, we will find that a very low order AR, MA, or ARMA model will suffice.

1. corAR1. This correlation structure is specified as corAR1(value, form = *one-sided formula*), where value specifies an (optional) inital value for estimating the AR(1) parameter ϕ and *one-sided formula* is a formula of the form:

   $$\sim \text{covariate}|\text{Groupingvariable}$$

   Here, the covariate is an integer-valued time index and |Groupingvariable is an optional group specification. Groups are specified to be units of observations on which repeated measurements through time are taken.

2. corCAR1. This correlation structure is a continuous-time version of an AR(1) correlation structure. The specification is the same as in corAR1, but now the covariate indexing time can take any non-negative non-repeated value and we restrict $\phi \geq 0$.

3. corARMA. This correlation structure corresponds to an ARMA(p, q) model. AR(p) and MA(q) models can be specified with this function, but keep in mind that the corAR1 specification is more efficient than specifying corARMA with $p = 1$ and $q = 0$.

   We can specify an ARMA(1,1) model with initial values of $\phi = .8, \theta = .4$ via corARMA(value = c(.8,.4), form = $\sim$ covariate|Groupingvariable, p=1, q=1).

General Correlation Structures:

1. corCompSymm. In this structure,

$$\text{corr}(\varepsilon_i, \varepsilon_j) = \begin{cases} 1 & \text{if } i = j; \text{ and} \\ \rho & \text{if } i \neq j. \end{cases}$$

   That is, the correlation between any two distinct observations is the same. Like many of the correlation structures, this structure is often useful within groups.

2. corSymm. Specifies a completely general correlation structure with a separate parameter for every non-redundant correlation. E.g., for an example with cluster size 5 corSymm(form $= \sim 1 \mid$ Cluster) specifies the correlation matrix

$$\mathbf{C} = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ & 1 & \rho_5 & \rho_6 & \rho_7 \\ & & 1 & \rho_8 & \rho_9 \\ & & & 1 & \rho_{10} \\ & & & & 1 \end{pmatrix}$$

   where initial values for $\boldsymbol{\rho}$ can be supplied with an optional value= specification.

Spatial Correlation Structures:

- A classic reference on spatial statistics is Cressie, *Statistics for Spatial Data.* The following material is based on Pinheiro and Bates (2000, §5.3), who base their treatment on material in Cressie's book.

Let $\varepsilon_{\mathbf{p}}$ denote the observation (error term in our nonlinear model) corresponding to position $\mathbf{p} = (p_1, p_2, \ldots, p_r)^T$.

- In a two-dimensional spatial context, $r = 2$ and $\mathbf{p} = (p_1, p_2)^T$ gives two dimensional coordinates. In a temporal context, $r = 1$ and $\mathbf{p}$ is a scalar equal to the time at which the measurement was taken.

Time series correlation structures are typically described by their autocorrelation function (which we've denoted $h(\cdot)$ above). Spatial correlation structures are usually described by their *semivariogram.*

For a given distance function $d(\cdot)$, the semivariogram is a function γ of the distance between two points $\varepsilon_{\mathbf{p}}$ and $\varepsilon_{\mathbf{q}}$ say, and a parameter $\boldsymbol{\rho}$, that measures the association between two points that distance apart:

$$\gamma\{d(\varepsilon_{\mathbf{p}}, \varepsilon_{\mathbf{q}}), \boldsymbol{\rho}\} = \frac{1}{2}\mathrm{var}(\varepsilon_{\mathbf{p}} - \varepsilon_{\mathbf{q}})$$

We assume the observations have been standardized to have $\mathrm{E}(\varepsilon_{\mathbf{p}}) = 0$ and $\mathrm{var}(\varepsilon_{\mathbf{p}}) = 1$ for all $\mathbf{p}$. Such a standardization does not alter the correlation structure.

In that case, it is easy to see the relationship between the semivariogram $\gamma(\cdot)$ and the autocorrelation function $h(\cdot)$:

$$\gamma(s, \boldsymbol{\rho}) = 1 - h(s, \boldsymbol{\rho}).$$

From this relationship it is clear that observations 0 distance apart have $h(0, \boldsymbol{\rho}) = 1$ and thus $\gamma(0, \boldsymbol{\rho}) = 0$. The autocorrelation function h increases continuously to 1 as the distance decreases to 0. Hence the semivariogram increases continuously to 0 as distance decreases to 0.

In some applications it is useful to violate this by introducing a *nugget effect* into the definition of the semivariogram. This nugget effect is a parameter c_0 that forces $\gamma(0, \boldsymbol{\rho}) = c_0$ where $0 < c_0 < 1$ rather than $\gamma(0, \boldsymbol{\rho}) = 0$ when the distance between the observations is 0.

The following spatial correlation structures are implemented in the nlme software in S-PLUS and R. All have a scalar-valued correlation parameter ρ. This parameter is known as the *range* in the spatial literature.

1. corExp. (Exponential) This structure corresponds to the semivariogram
$$\gamma(s, \rho) = 1 - \exp(-s/\rho)$$
and the autocorrelation function $h(s, \rho) = \exp(-s/\rho)$.

2. corGauss. (Gaussian) This structure corresponds to the semivariogram
$$\gamma(s, \rho) = 1 - \exp\{-(s/\rho)^2\}$$
and the autocorrelation function $h(s, \rho) = \exp\{-(s/\rho)^2\}$.

3. corLinear. (Linear) This structure corresponds to the semivariogram
$$\gamma(s, \rho) = 1 - (1 - s/\rho)1_{\{s<\rho\}}$$
and the autocorrelation function $h(s, \rho) = (1 - s/\rho)1_{\{s<\rho\}}$. Here $1_{\{A\}}$ represents the indicator variable that equals 1 when condition A is true, 0 otherwise.

4. corRatio. (Rational Quadratic) This structure corresponds to the semivariogram
$$\gamma(s, \rho) = \frac{(s/\rho)^2}{1 + (s/\rho)^2}$$
and the autocorrelation function $h(s, \rho) = \{1 + (s/\rho)^2\}^{-1}$.

5. corSpher. (Spherical) This structure corresponds to the semivariogram
$$\gamma(s, \rho) = 1 - \{1 - 1.5(s/\rho) + .5(s/\rho)^3\}1_{\{s<\rho\}}.$$

- A nugget effect can be added to any of these structures. With a nugget effect c_0, the semivariogram with the nugget effect $\gamma_{\text{nugg}}(\cdot)$ is defined in terms of the semivariogram without the nugget effect $\gamma(\cdot)$ as follows:

$$\gamma_{\text{nugg}}(s, c_0, \rho) = \begin{cases} c_0 + (1 - c_0)\gamma(s, \rho), & \text{if } s > 0; \text{ and} \\ 0, & \text{otherwise.} \end{cases}$$

- When using the above spatial correlation structures, the user can choose between distance metrics. Currently implemented distance metrics are *Euclidean distance*, $d(\varepsilon_{\mathbf{p}}, \varepsilon_{\mathbf{q}}) = \|\mathbf{p}-\mathbf{q}\| = \sqrt{\sum_{i=1}^{r}(p_i - q_i)^2}$, *Manhattan distance*, $d(\varepsilon_{\mathbf{p}}, \varepsilon_{\mathbf{q}}) = \sum_{i=1}^{r} |p_i - q_i|$, and *maximum distance*, $d(\varepsilon_{\mathbf{p}}, \varepsilon_{\mathbf{q}}) = \max_{i=1,\ldots,r} |p_i - q_i|$.

- One can get a feel for these various semivariogram models by examining them as functions of distance for different choices of the range parameter ρ and the nugget effect c_0. The 5 semivariograms listed above are plotted below for $\rho = 1$, $c_0 = .1$.

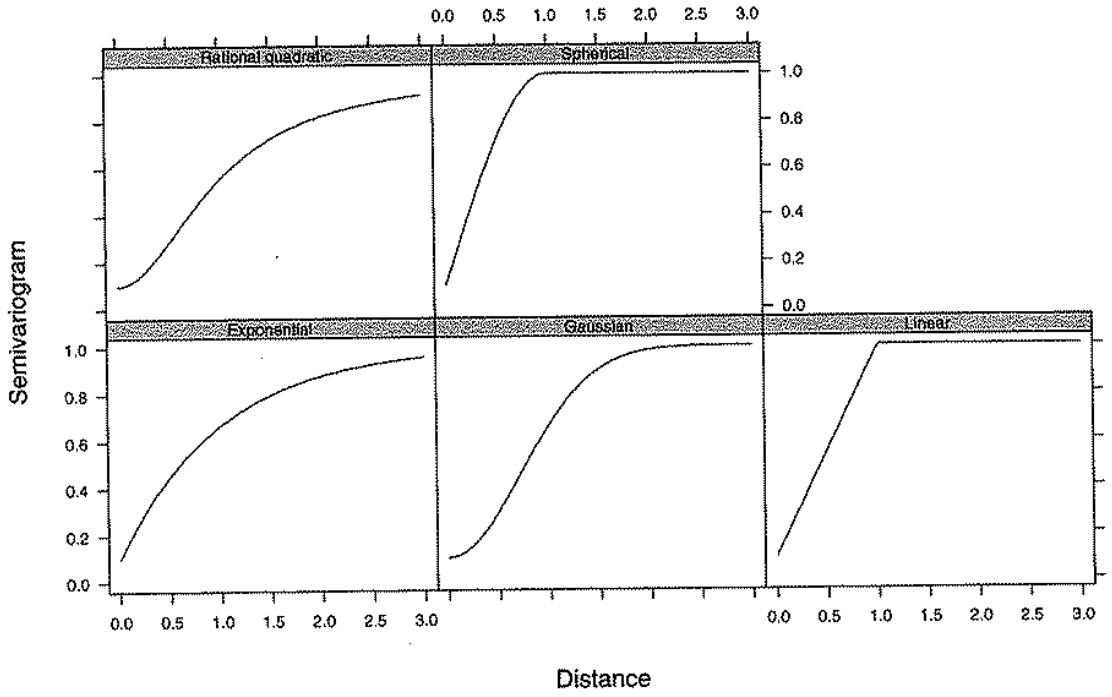


FIGURE 5.9. Plots of semivariogram versus distance for the isotropic spatial correlation models in Table 5.2 with range = 1 and nugget effect = 0.1.

Q: *How do we choose a correlation structure?*

A: This is a hard question that would be the focus of several weeks worth of study in a time series or spatial statistics course.

In a regression context, inference on the regression parameters is the primary interest. We need to account for a correlation structure if one exists to get those inferences right, but we're typically not interested in the correlation structure in and of itself. Therefore, we opt for simple correlation structures that capture "most of the correlation" without getting caught up in extensive correlation modeling.

In a time-dependence context, AR(1) models are often sufficient.

If we are willing to consider other ARMA models, two tools that are useful in selecting the right ARMA model are the *sample autocorrelation function* (ACF) and the *sample partial autocorrelation function* (PACF).

Let

$$r_{ij} = \frac{y_{ij} - \hat{y}_{ij}}{\sqrt{\hat{\text{var}}(\varepsilon_{ij})}}$$

denote the standardized residuals from a fitted LMM.

- Here $\hat{y}_{ij}$ are predicted values of y_{ij}. That is, they are elements of $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$.

The sample autocorrelation at lag ℓ is defined as

$$\hat{\rho}(\ell) = \frac{\sum_i^n \sum_{k=1}^{t_i - \ell} r_{ik} r_{i,k+\ell} / N(\ell)}{\sum_i^n \sum_{k=1}^n r_{ik}^2 / N(0)}, \qquad \ell = 1, 2, \ldots,$$

where $N(\ell)$ ($N(0)$) is the number of residual pairs being summed in the numerator (denominator).

The sample partial autocorrelation at lag ℓ is a sample estimate of the correlation between observation ε_{ik} and $\varepsilon_{i,k+\ell}$ after removing the effects of $\varepsilon_{i,k+1}, \ldots, \varepsilon_{i,k+\ell-1}$. The estimate is obtained by a recursive formula not worth reproducing here.

- AR(p) models have PACFs that are non-zero for lags $\leq p$ and 0 for lags $> p$. Therefore, we can look at the magnitude of the sample PACF to try to identify the order of an AR process that will fit the data. The number of "significant" partial autocorrelations is a good guess at the order of an appropriate AR process.

- MA(q) models have ACFs that are nonzero for lags $\leq q$ and 0 for lags $> q$. Again, we can look at the sample ACF to choose q.

- ARMA(p, q) models will have sample ACFs and PACFs that don't fit these simple rules. The following table (from Seber & Wild, Ch. 6, p.320) describes the general behavior of the ACF and PACF functions for various ARMA models.

Table 6.3 Properties of the ACF and the PACF for Various ARMA Models[a]

Model	ACF	PACF
AR(1)	Exponential or oscillatory decay	$\phi_{kk} = 0$ for $k > 1$
AR(2)	Exponential or sine wave decay	$\phi_{kk} = 0$ for $k > 2$
AR(q_1)	Exponential and/or sine-wave decay	$\phi_{kk} = 0$ for $k > q_1$
MA(1)	$\rho_k = 0$ for $k > 1$	Dominated by damped exponential
MA(2)	$\rho_k = 0$ for $k > 2$	Dominated by damped exponential or sine wave
MA(q_2)	$\rho_k = 0$ for $k > q_2$	Dominated by linear combination of damped exponentials and/or sine waves
ARMA(1, 1)	Tails off. Exponential decay from lag 1	Tails off. Dominated by exponential decay from lag 1
ARMA(q_1, q_2)	Tails off after $q_2 - q_1$ lags Exponential and/or sine wave decay after $q_2 - q_1$ lags	Tails off after $q_1 - q_2$ lags Dominated by damped exponentials and/or sine waves after $q_1 - q_2$ lags

[a]From Abraham and Ledolter [1983] with permission from John Wiley and Sons.

Similarly, in a spatial setting, we can estimate and plot the semivariogram to help choose an appropriate spatial correlation model. The Variogram() function in the nlme library will compute either one of two semivariogram estimators: the classical estimator, and a robust estimator that reduces the influence of outliers. See Pinheiro and Bates (2000, p.231) for formulas.

Example — Percent Body Fat Pre- and Post-Menarche

This example is taken from our text and concerns an analysis of a subset of data collected as part of the MIT Growth and Development Study. See FLW p.273 for details of the study design. Here we follow the analysis presented in FLW with some minor modifications.

- See fat.R, an R script that fits several models to these data and illustrates model diagnostic techniques and the model updating process. You will have to run this script on your own to generate the output.

Methods for Discrete and Categorical Longitudinal Data

Modeling Approaches for Longitudinal Data (GLMs and Extensions):

There are three main classes of GLMs (or extensions of GLMs) that are applied to longitudinal and other clustered data:

1. **Marginal models.** Model makes assumptions regarding the unconditional (marginal) distribution, or at least the first few moments, of the response variable. Typical assumptions:

   i. The marginal mean $\mu_{ij} = \mathrm{E}(y_{ij})$ for the j^{th} response on the i^{th} subject (cluster) is related to covariates via

   $$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}.$$

   ii. The marginal variance is related to the marginal mean via

   $$\mathrm{var}(y_{ij}) = \phi v(\mu_{ij}),$$

   iii. The marginal covariance between y_{ij} and $y_{ij'}$ is a known function of $\mu_{ij}, \mu_{ij'}$ and, possibly, a vector of unknown parameters $\boldsymbol{\alpha}$.

2. **Random effects or mixed effects models.** In random effects models, assumptions are made about the conditional distribution of the response variable given a vector of (usually cluster-specific) random effects.

- For non-normal data, the data are usually assumed to be independent given the random effects. That is, within-cluster correlation explained entirely through random effects.

  Typical assumptions:

  i. Given a vector $\mathbf{b}_i$ of cluster-specific random effects, the y_{ij}s are independent, with $y_{ij}|\mathbf{b}_i \sim$ ED with conditional mean $\mathrm{E}(y_{ij}|\mathbf{b}_i) \equiv \mu_{ij}^c$. Note this implies that $\mathrm{var}(y_{ij}|\mathbf{b}_i) = a_{ij}(\phi)v(\mu_{ij}^c)$ for v a variance function.

  ii. The conditional mean μ_{ij}^c is related to covariates via

  $$g(\mu_{ij}^c) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i.$$

  iii. $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are iid with some probability density function f, which is often assumed to be the normal density.

3. **Transition (or dynamic) models.** Here, the model is specified by making assumptins about the conditional distribution of y_{ij} given $y_{i,j-1}, y_{i,j-2}, \ldots, y_{i1}$ (the past), for each subject.

- Often, the conditional mean and variance of y_{ij} given the past is modeled via a GLM with past observations $y_{i1}, \ldots, y_{i,j-1}$ sometimes entering into the linear predictor.

Possible assumptions:

i. Conditional mean is related to covariates through a link function and linear predictor:

$$g(\mathrm{E}(y_{ij}|y_{i1}, \ldots, y_{i,j-1})) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \sum_{r=1}^{s} f_r(y_{i1}, \ldots, y_{i,j-1}; \boldsymbol{\alpha})$$

where $f_1, \ldots, f_s$ are known functions of the past observations and a vector of unknown parameters $\boldsymbol{\alpha}$.

ii. Conditional variance related to conditional mean:

$$\mathrm{var}(y_{ij}|y_{i1}, \ldots, y_{i,j-1}) = \phi v(\mathrm{E}(y_{ij}|y_{i1}, \ldots, y_{i,j-1})).$$

iii. Given the past, y_{ij}'s are independent. Might also assume some distribution in the ED family, say.

Comparison of the approaches:

- Transition models require large cluster sizes to be most effective. This means that they are less commonly used for typical longitudinal data sets than are the other model classes.

- Marginal models are typically partially specified models, and methods to fit them are usually related to quasilikelihood. Random-effects models are typically fully specified. Transition models can be fully or partially specified; examples of each exist.

- The three model types relate covariates to different quantities (different means). Thus, in general they have regression parameters with different interpretations.

  - Marginal models have parameters that describe a covariate's effect on the marginal mean response, where the mean is taken over the population of subjects from which the sample was drawn. Thus they have a **"population-averaged" interpretation.**

  - Random effects models describe a covariate's effect on the mean response of a subject with a particular value of the subject-specific random effect $\mathbf{b}_i$. Therefore, the parameter represents the effect on the expected response of an individual, not the mean response over a population. Thus, random effects models are said to yield **"subject-specific" interpretations.**

  - Transition models have parameters that describe the effect of a covariate on the mean response at a particular time given the past. This may yield both a population-averaged and subject-specific interpretation.

- In linear mixed models, regression parameters have both subject-specific and population-averaged interpretations because a conditionally specified LMM implies a corresponding marginal model with the same linear predictor (without the random effects part). However, this correspondence no longer holds for nonlinear link functions as arise in the modelling of discrete data.

- Therefore, marginal GLMs and mixed GLMs for discrete longitudinal data have parameters with different types of interpretations. This means that they have different domains of application.

- It is certainly true that all three types of models have significant drawbacks. E.g.,

  - Marginal models are not fully specified models, and may not be consistent with any valid probability generating mechanism.

  - Mixed models are limited by the assumption that within-cluster correlation is completely described by random effects. This limits the possible correlation structures that can be modelled (e.g., it is not as easy to model decaying correlation structures over time).

  However, it is a mistake to think in terms of one approach being better than the others. They all have their strengths, and they have distinct purposes and domains of application.

Generalized Estimating Equations

For repeated measures of a continuous outcome, many techniques of analysis are multivariate, based on an assumption of multivariate normality.

Let $y_{i1}, \ldots, y_{it_i}$ be the repeated measures taken on subject (or cluster) i.

- If $t_i = t$ for all i and the t measurement occasions are the same for all i, then it makes sense to regard each time-specific measurement as a separate variable.

- That is, we can think of $\mathbf{y}_i = (y_{i1}, \ldots, y_{it})^T$ as a $t-$dimensional vector-valued response, and we have a sample of such multivariate responses: $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

- If we add the assumption that $\mathbf{y}_i \overset{ind}{\sim} N_t(\mathbf{X}\boldsymbol{\beta}, \Sigma)$ then we can use the multivariate linear model and other normal-theory multivarite methods to analyze the data.

What if $\mathbf{y}_i$ is a vector of observations of a discrete or otherwise non-normal random response?

There are relatively few multivariate distributions other than the m'variate normal that are tractable.

- Some multivariate Poisson, multivariate Bernoulli, etc. distributions have been defined, but these are of limited usefulness for analysis.

However, if we're willing to specify less than the full multivariate discrete (non-normal) distribution (e.g., just the first two moments) then we can take a quasilikelihood approach. This is the idea behind generalized estimating equations (GEEs).

GEEs are

- An extension of quasilikelihood to longitudinal data.
- Method is semi-parametric in the sense that the estimating equations are derived without a full specification of the joint distribution of $\mathbf{y}_i$.
- Instead, we specify only
  - the mean and variance for the marginal (unconditional) response y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, t_i$.
  - a "working" covariance matrix for the vector of repeated measurements from each subject.
- GEEs yield consistent and asymptotically normal estimators of regression parameters, even with misspecification of the within-cluster correlation structure (i.e., even when working$\neq$true correlation matrix).
- Method avoids the need for a multivariate distribution.
- Presence of non-zero within-cluster covariance treated as a nuisance, so (original GEE) method is meant for situations in which regression parameters are of interest and covariance parameters are nuisance parameters.
- Clusters are assumed independent, and asymptotics are derived as number of clusters $\to \infty$.
- Even with incorrect working correlation structure, valid inference for regression parameter $\boldsymbol{\beta}$ is possible, because var($\hat{\boldsymbol{\beta}}$) can be consistently estimated.

The Method:

The model is specified marginally (unconditionally):

1. The marginal mean, $\mathrm{E}(y_{ij}) = \mu_{ij}$ for response in cluster i at measurement occasion j is related to explanatory variable through a GLM-type specification:
$$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$

2. The marginal variance $\mathrm{var}(y_{ij})$ is assumed to depend upon the marginal mean as in a GLM:
$$\mathrm{var}(y_{ij}) = \phi v(\mu_{ij})$$
where $v(\cdot)$ is the variance function, ϕ a possibly unknown scale (or dispersion) parameter.

3. Assume a "working" correlation structure for the vector of observations $\mathbf{y}_i = (y_{i1}, \ldots, y_{it_i})^T$, $i = 1, \ldots, n$:

   Let $\mathbf{R}(\boldsymbol{\alpha})$ be a $t_i \times t_i$ symmetric matrix that has the properties of a correlation matrix and is parameterized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s)^T$, an $s \times 1$ vector of correlation parameters.

   - $\mathbf{R}(\boldsymbol{\alpha})$ is the working correlation matrix.

   Let
$$\mathbf{A}_i = \mathrm{diag}\{\mathrm{var}(y_{i1}), \ldots, \mathrm{var}(y_{it_i})\} = \phi \begin{pmatrix} v(\mu_{i1}) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & v(\mu_{i1t_i}) \end{pmatrix}$$

   Now define
$$\mathbf{V}_i = \mathbf{V}_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}, \phi) = \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$$

- Note that we don't assume $\mathbf{R}(\boldsymbol{\alpha})$ is chosen correctly, but, if it is, $\mathbf{V}_i = \mathrm{var}(\mathbf{y}_i)$.

Recall the QL approach for the non-longitudinal (independent) data GLM. In that case $\mathbf{y}_i = y_i$ was univariate, and we obtained the QL regression parameter estimators by solving the quasi-score equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} \right)^T (y_i - \mu_i)/v(\mu_i) = \mathbf{0}$$

or equivalently

$$\mathbf{D}^T \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0},$$

where $\mathbf{D} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} \right)$ and $\mathbf{V} = \text{diag}\{v(\mu_1), \ldots, v(\mu_n)\}$.

In the longitudinal framework, we have a vector valued response $\mathbf{y}_i$ for each i, so if $\mathbf{R}(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{y}_i$ and if $\boldsymbol{\alpha}$ is known, the quasi-score equations become

$$\sum_{i=1}^{n} \underbrace{\mathbf{D}_i^T}_{p \times t_i} \underbrace{\mathbf{V}_i^{-1}(\boldsymbol{\mu}_i, \boldsymbol{\alpha}, \phi)}_{t_i \times t_i} \underbrace{(\mathbf{y}_i - \boldsymbol{\mu}_i)}_{t_i \times 1} = \mathbf{0}_{p \times 1}, \qquad (*)$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}$ and $\mathbf{V}_i$ is defined on the previous page.

Again, if $\mathbf{R}(\boldsymbol{\alpha})$ is the true correlation matrix of $\mathbf{y}_i$ and if $\boldsymbol{\alpha}$ is known, then the left-hand side of (*) is the quasi-score function and as an estimating function for $\boldsymbol{\beta}$, it even has an optimality property.

However, $\boldsymbol{\alpha}$ and ϕ will typically be unknown and must be estimated. What do we do in that case?

Suppose that, given $\boldsymbol{\beta}$, we have a $\sqrt{n}-$consistent estimator $\hat{\phi}(\boldsymbol{\beta})$; and, given $\boldsymbol{\beta}$ and ϕ, we have a $\sqrt{n}-$consistent estimator $\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \phi)$. Then substituting these estimators into the quasi-score equation (*), we have the estimating equation known as GEE:

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} \right)^T \mathbf{V}_i^{-1}(\boldsymbol{\mu}_i, \hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})), \hat{\phi}(\boldsymbol{\beta}))(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

- The solution to this equation, assuming it is unique, is the GEE regression parameter estimator, which we'll denote $\hat{\boldsymbol{\beta}}_{\text{GEE}}$.

<u>$\sqrt{n}-$consistent:</u>

$\hat{\theta}$ consistent for θ means that $\hat{\theta} = \theta + \text{o}_\text{p}(1)$. I.e., $\hat{\theta} - \theta \xrightarrow{\text{p}} 0$ as $n \to \infty$.

$\hat{\theta}$ $\sqrt{n}-$consistent for θ means that $\hat{\theta} = \theta + \text{O}_\text{p}(\frac{1}{\sqrt{n}})$. I.e., $\sqrt{n}(\hat{\theta} - \theta)$ is bounded in probability, or $\hat{\theta} - \theta$ goes to 0 at the same rate as does $\frac{1}{\sqrt{n}}$.

- consistency says nothing about the rate of convergence.
- $\sqrt{n}-$consistency implies consistency, but also that the estimator converges at rate $\frac{1}{\sqrt{n}}$.

Asymptotics:

Under regularity conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GEE}} - \boldsymbol{\beta}) \xrightarrow{\text{d}} N_p\left(\mathbf{0}, \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}\right)$$

where

$$\mathbf{I}_0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{D}_i$$

and

$$\mathbf{I}_1 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \left[\text{var}(\mathbf{y}_i)\right] \mathbf{V}_i^{-1} \mathbf{D}_i$$

- Notice that if $\text{var}(\mathbf{y}_i) = \mathbf{V}_i$, then $\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{GEE}}) = \mathbf{I}_0^{-1}$. This variance-covariance expression is known as the "model-based' or "naive" variance-covariance.

- However, $\text{var}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{GEE}}) = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$ regardless of whether or not $\text{var}(\mathbf{y}_i) = \mathbf{V}_i$. Therefore, $n\mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$ is a valid variance-covariance expression even if the working correlation matrix $\mathbf{R}(\boldsymbol{\alpha})$ is misspecified.

- Because of its form, $\mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$ is sometimes called a "sandwich variance."

Sandwich Variance Estimator:

In general, a Taylor series expansion of an unbiased estimating function U (unbiased here means an estimating function with expected value 0), a function of parameter $\boldsymbol{\beta}$, gives

$$0 = U(\hat{\boldsymbol{\beta}}) = U(\boldsymbol{\beta}) + U'(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \cdots$$

This implies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx -\left[n^{-1}U'(\boldsymbol{\beta})\right]^{-1}\left[n^{-1/2}U(\boldsymbol{\beta})\right]$$

so, if $U(\boldsymbol{\beta})$ is linear in $\mathbf{y}$ so that $U'(\boldsymbol{\beta})$ is non-random, we obtain

$$\text{var}\left(\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \approx \left[n^{-1}U'(\boldsymbol{\beta})\right]^{-1} n^{-1}\text{var}\left(U(\boldsymbol{\beta})\right)\left[n^{-1}U'(\boldsymbol{\beta})\right]^{-1}$$

In the GEE case, if we assume $\boldsymbol{\alpha}$ and ϕ are known, $U'(\boldsymbol{\beta})$ is as follows

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}^T}U(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}^T}\left\{\sum_i \mathbf{D}_i^T\mathbf{V}_i^{-1}(\boldsymbol{\mu}_i, \boldsymbol{\alpha}, \phi)(\mathbf{y}_i - \boldsymbol{\mu}_i)\right\} \\ &= \sum_i \left\{\frac{\partial}{\partial \boldsymbol{\beta}^T}\mathbf{D}_i^T\right\}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \sum_i \mathbf{D}_i^T\left\{\frac{\partial}{\partial \boldsymbol{\beta}^T}\mathbf{V}_i^{-1}\right\}(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ &\quad - \sum_i \mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\end{aligned}$$

As $n \to \infty$, the first two terms above go to their means, which in each case is $\mathbf{0}$, by the WLLN. In addition,

$$\text{var}\left(U(\boldsymbol{\beta})\right) = \sum_i \mathbf{D}_i^T\mathbf{V}_i^{-1}\underbrace{\text{var}(\mathbf{y}_i - \boldsymbol{\mu}_i)}_{=\text{var}(\mathbf{y}_i)}\mathbf{V}_i^{-1}\mathbf{D}_i$$

Putting these results together, we get the sandwich expression for the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{GEE}}$:

$$\text{avar}\left(\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{GEE}} - \boldsymbol{\beta})\right) = \mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$$

- Liang and Zeger ('86) showed that this result still holds if $\boldsymbol{\alpha}$, ϕ are replaced by their $\sqrt{n}-$consistent estimators.

- We can estimate $\text{avar}(\hat{\boldsymbol{\beta}}_{\text{GEE}}) = \frac{1}{n}\mathbf{I}_0^{-1}\mathbf{I}_1\mathbf{I}_0^{-1}$ by replacing $\text{var}(\mathbf{y}_i)$ in $\mathbf{I}_1$ by $(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T$ and evaluating all quantities at the estimates $\hat{\boldsymbol{\beta}}_{\text{GEE}}$, $\hat{\phi}(\hat{\boldsymbol{\beta}}_{\text{GEE}})$ and $\hat{\boldsymbol{\alpha}}(\hat{\boldsymbol{\beta}}_{\text{GEE}}, \hat{\phi}(\hat{\boldsymbol{\beta}}_{\text{GEE}}))$.

- Because the sandwich variance estimator is valid without the assumption that $\text{var}(\mathbf{y}_i)$ has been correctly specified, it is sometimes called a "robust" variance estimator.

Inference

Wald based inference based upon the model-based or sandwich var-cov estimator may be used to test hypotheses and form confidence intervals in the usual way.

That is, to test a hypothesis of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d},$$

where $\mathbf{C}$ is a $c \times p$ matrix of constants of full row rank and $\mathbf{d}$ a $c \times 1$ constant vector, we use the Wald test statistic

$$(\mathbf{C}\hat{\boldsymbol{\beta}}_{\text{GEE}} - \mathbf{d})^T[\mathbf{C}\hat{\text{var}}(\hat{\boldsymbol{\beta}}_{\text{GEE}})\mathbf{C}^T]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}_{\text{GEE}} - \mathbf{d}) \overset{a}{\sim} \chi^2(c), \quad \text{under } H_0.$$

A Wald-based approximate $100(1-\alpha)\%$ confidence interval on an estimable function $\mathbf{c}^T\boldsymbol{\beta}$ is given by

$$\mathbf{c}^T\hat{\boldsymbol{\beta}}_{\text{GEE}} \pm z_{1-\alpha/2}\sqrt{\mathbf{c}^T\hat{\text{var}}(\hat{\boldsymbol{\beta}}_{\text{GEE}})\mathbf{c}}.$$

- Alternatively, Rotnitsky and Jewell ('90), Boos ('92), and Hanfelt and Liang ('95) have discussed score-like tests and likelihood ratio-like tests. Since there is no likelihood function in a GEE model, these procedures are based either on

    a. replacing the likelihood function with a quasilikelihood function, or

    b. using the standard likelihood-based inference methods based upon the model in which the data are assumed independent, and then adjusting the inference somehow for non-independence (usually, by adjusting the d.f. of the limiting chi-square distribution).

- Approximate (generalized) score tests and Wald tests are implemented in PROC GENMOD. There is some evidence that the former are superior to the latter. Generalized LR tests are not implemented.

- Several authors have made suggestions for improving the sandwich variance-covariance estimator to improve inference methods in GEEs in small to moderate sample size situations (e.g., Pan, 2001). In addition, there have been suggestions for modifiying the Wald test to an F test for better inference in small samples.

Choice of Working Correlation Matrix:

- Should choose the form of $\mathbf{R}$ to be consistent with empirical correlations.

- $\hat{\boldsymbol{\beta}}_{\text{GEE}}$ is consistent regardless of whether or not $\mathbf{R}(\boldsymbol{\alpha})$ is specified correctly.

- However, choice of $\mathbf{R}(\boldsymbol{\alpha})$ and quality of the estimators of $\boldsymbol{\alpha}$ and ϕ will affect the efficiency of $\hat{\boldsymbol{\beta}}_{\text{GEE}}$.

- Efficiency consequences are more severe for small n, large cluster sizes, high intra-cluster correlations.

Possibilities:

1. Independence. In this case $\mathbf{R} = \mathbf{I}_{t_i}$ ($s = 0$, i.e., no α parameter).

- In this case, we fit a GLM to the data as if they were independent but obtain an appropriate variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{\text{GEE}}$ by using the sandwich variance estimator $\neq$ independence GLM variance estimator.

- If the number of clusters is large compared to the cluster size, and the intracluster correlation not too large, there may be little loss of efficiency with this approach.

- This structure has the advantage that it makes sense regardless of whether the measurement times and/or cluster sizes differ across subjects.

2. Equicorrelation. In this case $s = 1$ and $R_{jj'} = \alpha \; \forall j \neq j'$. I.e.,

$$\mathbf{R}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ & 1 & \alpha & \cdots & \alpha \\ & & \ddots & & \vdots \\ & & & 1 & \alpha \\ & & & & 1 \end{pmatrix}$$

- This correlation structure when combined with constant variance within a cluster, is what is known as **compound symmetry**. However, if the variance depends on the mean, and if the mean depends upon **time-varying covariates** then this structure will differ from compound symmetry.

- This correlation structure is often called the **exchangeable** correlation structure, because it means that the order of the within cluster observations doesn't matter (they are exchangeable).

- Also suitable for varying cluster sizes, and varying measurement occasions.

3. AR(1). In this case, $s = 1$ and $R_{jj'} = \alpha^{|j-j'|}\ \forall j, j'$. I.e.,

$$\mathbf{R}(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{t_i-1} \\ & 1 & \alpha & \cdots & \alpha^{t_i-2} \\ & & \ddots & & \vdots \\ & & & 1 & \alpha \\ & & & & 1 \end{pmatrix}$$

- For a continuous response, this is the correlation structure for a 1^{st}-order autoregressive process.

- Note that we are not assuming an AR(1) process for our data, only that our data have the same correlation structure as in an AR(1) process. Thus, measurement occasions need not be equally spaced, although this structure makes much better sense if they are.

4. Stationary $m-$dependent. In this case, $s = 1$ and

$$R_{jj'} = \begin{cases} \alpha^{|t_j - t_{j'}|}, & \text{if } |t_j - t_{j'}| \leq m; \\ 0, & \text{otherwise.}, \end{cases}$$

where t_j is the j^{th} observation time.

- For example, for $t_j = j$ and $m = 1$ (1-dependence) we have

$$\mathbf{R}(\alpha) = \begin{pmatrix} 1 & \alpha & 0 & 0 & \cdots & 0 \\ & 1 & \alpha & 0 & \cdots & 0 \\ & & \ddots & & & \vdots \\ & & & & 1 & \alpha \\ & & & & & 1 \end{pmatrix}$$

- For a continuous response, this is the correlation structure for a 1^{st}-order moving average process.

5. Completely Unspecified. In this case, $R_{jj'} = \alpha_k$ where $k = 1, \ldots, t_i(t_i - 1)/2$ indexes the unique combinations of j, j', subject to $j < j'$. I.e.,

$$\mathbf{R}(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_{t_i-1} \\ & 1 & \alpha_{t_i} & \cdots & \alpha_{2t_i-3} \\ & & \ddots & & \vdots \\ & & & 1 & \alpha_{t_i(t_i-1)/2} \\ & & & & 1 \end{pmatrix}$$

- Here, $s = t_i(t_i - 1)/2$.
- This structure is guaranteed to be correct, but may lead to inefficiency if a more parsimonious structure will suffice.
- With large cluster sizes (many observation times) this structure can "cost" too many nuisance parameters (in $\boldsymbol{\alpha}$).
- This structure makes sense only when the measurement times are the same in all clusters.
- Miller et al. ('93, *Biometrics*) established that under a completely unstructured working correlation matrix, GEE is an iterated version of WLS.

QIC Model Selection Criterion:

Pan ('01, *Biometrics*) introduced a generalization of the AIC criterion appropriate for the GEE context. This criterion can be useful for selecting a working correlation matrix and/or selecting covariates in the linear predictor.

Under the assumption of independence, a quasilikelihood can be formed as the sum of quasilikelihood contributions from each observations.

Let $\hat{\boldsymbol{\beta}}(\mathbf{R})$ be the GEE estimator of $\boldsymbol{\beta}$ under a given working matrix $\mathbf{R}$. Then, under independence the total quasilikelihood for a model with design matrix $\mathbf{X}$, response $\mathbf{y}$, and known dispersion parameter ϕ is given by

$$Q(\hat{\boldsymbol{\beta}}(\mathbf{I}), \phi) = \sum_{i=1}^{K} \sum_{j=1}^{n_i} Q(\hat{\boldsymbol{\beta}}(\mathbf{I}), \phi, y_{ij}, \mathbf{x}_{ij})$$

Pan suggested an AIC-like criterion, where this independence-based quasilikelihood, evaluated at $\hat{\boldsymbol{\beta}}(\mathbf{R})$, the GEE estimator fit under working matrix $\mathbf{R}$, is used in place of the loglikelihood in the AIC criterion, and a penalty is used that generalizes the AIC penalty. That is, Pan proposes the criterion

$$QIC(\mathbf{R}, \phi) = -2Q(\hat{\boldsymbol{\beta}}(\mathbf{I}), \phi) + 2\text{tr}(\hat{\Omega}_{\mathbf{I}} \hat{\mathbf{V}}_{\mathbf{R}})$$

where $\hat{\mathbf{V}}_{\mathbf{R}}$ is the robust (sandwich) estimator of $\text{var}(\hat{\boldsymbol{\beta}}(\mathbf{R}))$ and $\hat{\Omega}_{\mathbf{I}}$ is the inverse of the model-based estimate of $\text{var}(\hat{\boldsymbol{\beta}}(\mathbf{I}))$ evaluated at $\hat{\boldsymbol{\beta}}(\mathbf{R})$.

- Note that the penalty reduces to $2p$, the AIC penalty term, if $\mathbf{R} = \mathbf{I}$.

- In practice, we must plug in an estimator of ϕ. For comparison of models via QIC, the same estimate of ϕ should be used for the models to be compared. The estimate of ϕ should be taken from the most complex model to be considered for the data.

- QIC can be used to compare models with different working correlation matrices and/or different specifications of the mean. However, if we just want to compare models with different means (and same correlation structure), then we can use QIC_u, the same criterion with $2p$ used as the penalty term instead of $2\text{tr}(\hat{\Omega}_{\mathbf{I}} \hat{\mathbf{V}}_{\mathbf{R}})$. SAS prints out both QIC and QIC_u.

Estimating $\boldsymbol{\alpha}$ and ϕ:

M.O.M. Liang & Zeger (1986) in their original GEE approach recommended method of moment estimators based on residuals on a case-by-case basis.

Given $\hat{\boldsymbol{\beta}}_{\text{GEE}}$, calculate standardized (Pearson) residuals:

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

Then use r_{ij}'s to estimate ϕ and $\boldsymbol{\alpha}$:

$$\hat{\phi}(\hat{\boldsymbol{\beta}}_{\text{GEE}}) = \sum_{i=1}^{n} \sum_{j=1}^{t_i} \frac{r_{ij}^2}{N - p}$$

where $N = \sum_i t_i$ = total sample size, and $p = \dim(\boldsymbol{\beta})$.

The MOM estimator of $\boldsymbol{\alpha}$ depends upon the working correlation structure. E.g., exchangeable case:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} \sum_{j \neq j'} r_{ij} r_{ij'}}{\{\sum_i t_i(t_i - 1)\} - p} = \text{average obs'd correlation}$$

Fitting the model consists of iterating between solving the GEE for $\hat{\boldsymbol{\beta}}$ given $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ and updating $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ with the MOM estimators based on the current $\hat{\boldsymbol{\alpha}}$. This process is iterated to convergence.

- This approach, using MOM estimators for α and ϕ is what was originally proposed by Liang and Zeger ('86), and it is what is implemented in PROC GENMOD in SAS. However, several refinements have been proposed that are not yet broadly available.

Example – Epileptic Seizures:

A clinical trial was conducted to assess the efficacy of the anti-epileptic drug progabide. $n = 59$ epileptics counted the number of seizures that they experienced during an 8-week baseline period. Patients were then randomized to placebo or progabide and seizure counts were recorded over four consecutive two-week periods following the onset of treatment.

Data:

				Two-week Period			
Subject	Trt	Base	Age	1	2	3	4
1	0	11	31	5	3	3	3
2	0	11	30	3	5	3	3
⋮							
29	1	76	18	11	14	9	8
30	1	38	32	8	7	9	4
⋮							

Question: *Does progabide reduce the seizure rate?*

Data are counts, suggesting a Poisson regression model, with log link, $v(\mu) = \mu$ (identity) variance function.

However, seizure counts recorded on the same individual are likely to be correlated. In addition, there is substantial overdispersion in the data as can be seen by examining the variance to mean ratios by group:

	Two-week Period			
Group	1	2	3	4
Active	38.7	16.8	23.8	18.8
Placebo	10.8	7.5	24.5	7.3

- The fact that these ratios are >> 1 indicates substantial overdispersion. Some of this might be explainable with refined modeling of the mean (e.g., through modeling its dependence on the covariates), but all of it is unlikely to be accounted for in this way.

- Both correlation, overdispersion can be accounted for in a GEE model:

Data: Let y_{hij} be the response at time j ($j = 0, 1, 2, 3, 4$) for the i^{th} subject in treatment group h ($h = 1$ for placebo, $h = 2$ for progabide).

Model:

$$\log(\mu_{hij}) = \lambda_{hj} + \beta_{hj} \log(y_{hi0}), \quad \begin{array}{l} h = 1, 2 \\ i = 1, \ldots, n_h \\ j = 1, 2, 3, 4 \end{array} \qquad (\dagger)$$

In addition,

$$\text{var}(y_{hij}) = \phi \mu_{hij}$$

and we will assume a working correlation matrix for $\text{corr}(\mathbf{y}_{hi})$.

- This is an ANCOVA model where the baseline seizure count has been treated as a covariate. Within-subject correlation is taken care of with a working correlation matrix and by using the sandwich var-cov matrix rather than by using random subject effects as we might have done in a mixed model.

- See the handout titled epileps1.sas. In this program, we fit the model above with GEE, using several choices of working correlation matrix. Then the covariate part of the model $\beta_{hj} \log(y_{hi0})$ is simplified, leading to a final model from which inferences are drawn.

- We first plot the data against the covariate to see whether a linear relationship seems reasonable. In this case, it does.

- Next, we fit model (†) using independence (labeled pp.2–4), AR(1) (pp.5–6), and exchangeable (pp.7–8) working correlation structures.

- When fitting a GEE model, GENMOD first fits the corresponding univariate GLM (ignoring clustering) to obtain initial parameter estimates. These estimates are no different than the GEE estimates assuming working correlation matrix $\mathbf{R} = \mathbf{I}$, however the sandwich variance-covariance estimator adjusts the standard errors of the GEE regression parameter estimates substantially (compare empirical and model-based SEs).

- Prior to version 9.2, PROC GENMOD printed out all the default GLM output from the initial model (under independence), which included model fit statistics (e.g., the deviance, AIC, etc.) for that model. This caused considerable confusion, because those statistics don't apply when using GEEs (which is no a likelihood-based procedure). Fortunately, these results are no longer printed, but remember, such crieria are not appropriate in the GEE setting.

- The MOM estimators of the α parameter of the AR(1) and exchangeable working correlation matrices are given in the working correlation matrix on pp.5 and 7. Notice that substantial within-subject correlation is estimated to be present in these data: $\hat{\rho} = .51$ (AR(1)) or $\hat{\rho} = .40$ (exchangeable) depending on the working structure chosen.

- SAS estimates $\sqrt{\phi}$ rather than ϕ itself. Unfortunately, this estimate is only printed out if we specify the MODELSE option on the REPEATED statement. In this case, $\sqrt{\hat{\phi}} = 2.1608$ in the case of $\mathbf{R} = \mathbf{I}$ (independence). Similar results were obtained for the other correlation structures (suppressed in this output). This indicates substantial overdispersion relative to the Poisson variance for these data.

- Notice that the **R** =independence and **R** =AR(1) and **R** =exchangeable results are quite similar. However the QIC criterion favors (slightly) **R** =AR(1). Note that for comparing QIC values across models it is crucial to use the same value of ϕ in the models to be compared. This can be done with the SCALE= and NOSCALE options. If models are not being compared via QIC, it is better to let ϕ be a free parameter to be estimated (e.g., in the final model).

- Note that I was unable to get the unstructured working correlation structure to converge for this model, but we will consider this structure after first reducing the mean structure. It turns out that the unstructured model fits no better accoring to QIC.

- As in previous examples involving covariates, we next simplify the covariate part of the model. This is done by eliminating the non-significant terms involving logbase one at a time. At the end of this process we arrive at the final model:

$$\log(\mu_{hij}) = \lambda_{hj} + \beta \log(y_{hi0}).$$

- Based on this model, the treatment*time interaction and both main effects are not significant. We computed the lsmeans for each treatment and the difference between those means using the LSMEANS statement.

  - Note that these are estimates of the log means, not the means. If we want a point estimate of a mean, we need to exponentiate the results given by LSMEANS. If we want a confidence intervals for a contrast in the μ_{hij}'s, the best thing to do is form the confidence intervals for the corresponding contrast in $\boldsymbol{\beta}$ and exponentiate (apply the inverse link) to the endpoints.

- The ESTIMATE statement can also be used to estimate a contrast in $\boldsymbol{\beta}$. It has an EXP option to transform to the original scale.

GEE Extensions:

GEE-1 (Prentice, '88; Prentice and Zhao, '91; terminology due to Liang et al. '92):

In the original GEE paper, second moment (correlation and dispersion) parameters were estimated based on simple functions of the residuals. Different estimators were proposed in each working correlation model. This is an *ad hoc* approach.

- The dictionary definition of ad hoc is, "with respect to the particular purpose at hand and without consideration of wider application." I.e., a method that is not motivated by and does not fit into any general theory, but rather seems to be a reasonable thing to do for the problem at hand.

Prentice and Zhao generalized this approach by proposing estimating equations for the association (correlation) parameters which they referred to as "ad hoc estimating equations."

The GEE for $\boldsymbol{\beta}$ (first moment parameter) has the form

$$U_1 = \sum_{i=1}^{n} \underbrace{\mathbf{D}_{i11}^{T}}_{\text{deriv. matrix}} \overbrace{\mathbf{V}_{i11}^{-1}}^{\text{wt. matrix}} \underbrace{(\mathbf{y}_i - \boldsymbol{\mu}_i)}_{\text{elementary est. fn.}} = \mathbf{0}$$

P&Z suggested using a second estimating equation for second moment parameters ($\boldsymbol{\alpha}$):

$$U_2 = \sum_{i=1}^{n} \mathbf{D}_{i22}^T \mathbf{V}_{i22}^{-1}(\mathbf{s}_i - \boldsymbol{\sigma}_i) = \mathbf{0}$$

where $\boldsymbol{\sigma}_i = (\sigma_{i11}, \sigma_{i12}, \ldots, \sigma_{it_it_i})^T$ is the (working) variance-covariance matrix of $\mathbf{y}_i$ in vector form, and $\mathbf{s}_i = (s_{i11}, s_{i12}, \ldots, s_{it_it_i})^T$ is the sample version of $\boldsymbol{\sigma}_i$, consisting of the sample covariances of $\mathbf{y}_i$. I.e., $\mathbf{s}_i$ has expected value $\boldsymbol{\sigma}_i$ and consists of elements

$$s_{ijk} = (y_{ij} - \mu_{ij})(y_{ik} - \mu_{ik}).$$

In addition, $\mathbf{D}_{i22} = \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\alpha}^T}$ and $\mathbf{V}_{i22}$ is a "working" form for $\text{cov}(\mathbf{s}_i)$ (that is, it contains a working assumption for the third and fourth-order moment structure of $\mathbf{y}_i$).

- Notice that the form of U_2 is parallel to that of U_1.

- The MOM estimators can be expressed as solutions to estimating equations of the form given in U_2 for properly chosen $\mathbf{V}_{i22}$ matrices, so P&Z's approach generalizes the original L&Z approach.

Rather than iterating between solving U_1 and U_2, an equivalent approach is to stack U_1 and U_2 and solve the resulting larger estimating equation:

$$U = \sum_{i=1}^{n} \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{f}_i = \mathbf{0}$$

where

$$\mathbf{D}_i = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} & \mathbf{0} \\ \mathbf{0} & \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\alpha}^T} \end{pmatrix} = \begin{pmatrix} \mathbf{D}_{i11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{i22} \end{pmatrix},$$
$$\mathbf{V}_i = \begin{pmatrix} \mathbf{V}_{i11} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{i22} \end{pmatrix}, \qquad \mathbf{f}_i = \begin{pmatrix} \mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{s}_i - \boldsymbol{\sigma}_i \end{pmatrix}$$

- The GEE-1 parameter estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by solving $U = \mathbf{0}$ (by, e.g., Fisher scoring). ϕ is usually estimated separately using the MOM estimator on p.168. Alternatively, a better option would be to absorb ϕ into $\boldsymbol{\alpha}$ as the $(s+1)^{\text{th}}$ element of this second moment parameter, and $\hat{\phi}$ would then be obtained as part of the solution to $U = \mathbf{0}$.

- $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are treated as **orthogonal** to one another even though they are, in general, not. Parameter orthogonality here refers to uncorrelated score components ($\mathbf{V}_i$ and $\mathbf{D}_i$ are block-diagonal). See Cox and Reid (1987, *JRSS-B*) for more on this concept.

- The orthogonal treatment of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ has the benefit that $\hat{\boldsymbol{\beta}}$ is consistent under misspecification of $\text{cov}(\mathbf{y}_i)$ (wrong working correlation matrix).

GEE-2 (Prentice and Zhao, 1991):

In this approach $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are modelled and estimated simultaneously. $\boldsymbol{\alpha}$ is no longer considered to be a nuisance parameter.

- In ordinary GEE and GEE-1, the first moment parameter $\boldsymbol{\beta}$ is of interest and estimation for $\boldsymbol{\beta}$ involves second-order nuisance parameters ϕ and $\boldsymbol{\alpha}$.

- In GEE-2, the first and second-order monent parameters $\boldsymbol{\beta}$, ϕ and $\boldsymbol{\alpha}$ are all of interest and estimation for these parameters involves third and fourth-order nuisance parameters.

Consider the quadratic estimating function

$$\sum_i \mathbf{D}_i^T \mathbf{V}_i^{-1} \mathbf{f}_i \qquad (\heartsuit)$$

where now

$$\mathbf{D}_i = \mathrm{E}\left(\frac{\partial \mathbf{f}_i}{\partial \binom{\boldsymbol{\beta}}{\boldsymbol{\alpha}}^T}\right) = \frac{\partial \binom{\boldsymbol{\mu}_i}{\boldsymbol{\sigma}_i}}{\partial \binom{\boldsymbol{\beta}}{\boldsymbol{\alpha}}^T} = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} & \mathbf{0} \\ \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\beta}^T} & \frac{\partial \boldsymbol{\sigma}_i}{\partial \boldsymbol{\alpha}^T} \end{pmatrix}$$

and $\mathbf{f}_i$ is as in GEE-1.

If the weight matrix $\mathbf{V}_i$ equals $\mathrm{var}(\mathbf{f}_i)$ or

$$\mathbf{V}_i = \mathrm{var}\begin{pmatrix} \mathbf{y}_i \\ \mathbf{s}_i \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\mathbf{y}_i) & \mathrm{cov}(\mathbf{y}_i, \mathbf{s}_i) \\ \mathrm{cov}(\mathbf{s}_i, \mathbf{y}_i) & \mathrm{var}(\mathbf{s}_i) \end{pmatrix}$$

then $(\heartsuit)$ is optimal in the class of estimating functions for $\boldsymbol{\beta}, \alpha$ that are quadratic in $\mathbf{y}$.

- Essentially, $(\heartsuit)$ is the quasi-score function for $(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)$.

Unfortunately, the use of ($\heartsuit$) requires knowledge of $\text{cov}(\mathbf{y}_i, \mathbf{s}_i)$ and $\text{var}(\mathbf{s}_i)$; i.e., it requires knowledge of the third and fourth order moments of $\mathbf{y}_i$.

Since this knowledge is rarely(!) available, Prentice and Zhao suggested using working specifications for $\text{cov}(\mathbf{y}_i, \mathbf{s}_i)$ and $\text{var}(\mathbf{s}_i)$, parallel to the approach in the original GEE approach.

That is, the GEE-2 estimating equations are

$$\sum_i \mathbf{D}_i^T \tilde{\mathbf{V}}_i^{-1} \mathbf{f}_i = \mathbf{0} \qquad \text{(GEE} - 2\text{)}$$

where $\mathbf{D}_i$ and $\mathbf{f}_i$ are as before, and

$$\tilde{\mathbf{V}}_i = \begin{pmatrix} \text{var}(\mathbf{y}_i) & \tilde{\text{cov}}(\mathbf{y}_i, \mathbf{s}_i) \\ \tilde{\text{cov}}(\mathbf{s}_i, \mathbf{y}_i) & \tilde{\text{var}}(\mathbf{s}_i) \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{i11} & \mathbf{V}_{i12} \\ \mathbf{V}_{i12}^T & \mathbf{V}_{i22} \end{pmatrix}.$$

Here, tildes indicate working structures.

- Note that $\text{var}(\mathbf{y}_i)$ has no tilde. That is, in GEE-2 we assume that the second moment model is correctly specified.
  - If that assumption fails, then GEE-2 is not guaranteed to yield consistent estimators of $\boldsymbol{\beta}$.
  - However, if that assumption is met, then GEE-2 can lead to greater efficiency in the estimation of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ than GEE-1, and these parameters are consistently estimated under misspecification of the third and fourth order moment models.

EGEE (Hall and Severini, '98):

- Goal was to increase efficiency of $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ (and hopefully, as a consequence increase efficiency of $\hat{\boldsymbol{\beta}}$) without sacrificing consistency of $\hat{\boldsymbol{\beta}}$ under misspecification of working covariance structure.

- Ideas of extended quasilikelihood (Nelder and Pregibon '87) used to motivate the quasi-score estimating equation for $\boldsymbol{\beta}$ combined with a Gaussian (normal) score function for estimation of $\boldsymbol{\alpha}$ and ϕ.

- Turns out to be a special case of GEE-1.

- Efficiency of $\hat{\boldsymbol{\alpha}}$ and $\hat{\phi}$ better than GEE with MOM and GEE-1 with most other choices of $\mathbf{V}_{i22}$; similar efficiency to GEE-2 for these parameters.

- EGEE typically performs better than GEE-2 and similar to other GEE-1 methods with respect to $\boldsymbol{\beta}$.

- This approach has also been advocated by Crowder (2001) who called it Gaussian estimation and pointed out that it solves some of the theoretical ambiguities of GEE.

GEE-ALR (Carey et al. '93, *Biometrika*):

- A GEE-1 method appropriate to repeated binary responses.

- Within-cluster associations parameterized in terms of odds ratios rather than correlations. Model fit by solving a pair of estimating equations as in GEE-1, but solutions of each equation can be obtained by fitting a logistic regression.

- Hall ('98) used EGEE with odds ratio parameterization for binary data and found that it yielded parameter estimates with smaller MSEs than GEE-ALR.

Finally, EQL (Hall, 2001):

- Extension of EQL from univariate situation to longitudinal data context.

- For certain data types, substantial gains in efficiency can be obtained, but somewhat limitted domain of application and can be difficult to implement. EGEE more generally recommended.

- Hall (2001) also derived REML-like bias adjustments for estimation of second moment parameters using EGEE, EQL.

- Currently, only original GEE and GEE-ALR implemented in SAS' PROC GENMOD. Orginal GEE also available in S-PLUS, Stata, and SUDAAN software packages.

Pitfalls and Drawbacks to GEE

The fact that the marginal model underlying GEE is not a full statistical model specifying the full data generating mechanism, and the fact that estimates are obtained by solving an estimating equation rather than optimizing some objective function lead to some pitfalls and potential problems with GEE:

1. For a given data type, mean model and variance specification, a variety of different correlation structures can be assumed for $\mathbf{y}_i$. However, particular combinations of these assumptions may not be consistent with *any* legitimate data generating mechanism.

- E.g., suppose that for clustered count data I assume $\text{var}(y_{ij}) = \mu_{ij} = \exp(\eta_{ij})$ for some linear predictor η_{ij}.

  - Suppose the cluster size $t_i \geq 3$ for all i. Then there is no known method of generating such data if the 1-dependence correlation structure is used.

  - Existing models for multivariate Poisson data imply that the correlation depends upon the mean. So a particular correlation structure may not be consistent with the mean model. Furthermore, existing multivariate Poissons do not allow negative correlations.

2. Liang & Zeger's theory for GEE says that $\hat{\boldsymbol{\beta}}_{\text{GEE}}$ is consistent even if an incorrect working form for $\text{corr}(\mathbf{y}_i) = \mathbf{R}(\boldsymbol{\alpha})$ is used, if a $\sqrt{n}$-consistent estimator of $\boldsymbol{\alpha}$ is used. However, if $\mathbf{R}(\boldsymbol{\alpha})$ is an incorrect structure, then what does it mean for $\hat{\boldsymbol{\alpha}}$ to be consistent? Consistent for what?

- This conceptual problem with the notion of consistency when speaking of a parameter estimator for a parameter that may not exist under the true model, was pointed out by Crowder ('95, *Bio'ka*).

  - He indicated that even if the notion of consistency of $\hat{\boldsymbol{\alpha}}$ is problematic, the theory should still be valid as long as $\hat{\boldsymbol{\alpha}}$ tends stochastically to some limit as $n \to \infty$.

  - The convergence of $\hat{\boldsymbol{\alpha}}$ will not be a problem as long as the estimating equation for $\boldsymbol{\alpha}$ corresponds to the gradient of some objective function to be optimized (as in EGEE/Gaussian estimation).

  - However, for arbitrary estimating functions for $\boldsymbol{\alpha}$, it can be a problem and there are some situations in which it can be shown that Liang & Zeger's original MOM estimators of $\boldsymbol{\alpha}$ will not converge to a limit. Fortunately, these situations are somewhat unusual, and typically can be avoided by making sensible choices when using GEE (or by using EGEE/Gaussian estimation). However, this whole issue is problematic from a theoretical standpoint.

Extension of GEE to Longitudinal Categorical Data:

Suppose that for the i^{th} subject at the j^{th} time point, we have a categorical response variable z_{ij} taking on one of K possible values which we'll represent with the integers $1, \ldots, K$.

Let $\pi_{ijk} = \Pr(z_{ij} = K)$.

Then z_{ij} can be represented as a vector of $K-1$ indicator variables: $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ij,K-1})^T$, where

$$y_{ijk} = \begin{cases} 1 & \text{if } z_{ij} = k; \\ 0 & \text{otherwise.} \end{cases}$$

The entire data vector for the i^{th} subject can be written as

$$\begin{aligned} \mathbf{y}_i &= (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{it_i}^T)^T \\ &= (y_{i11}, \ldots, y_{i1,K-1}, \ldots, y_{it_i1}, \ldots, y_{it_i,K-1})^T. \end{aligned}$$

- Note that now the response at a particular measurement occasion for a particular subject $\mathbf{y}_{ij}$ is a vector, not a scalar, and the elements of this vector are correlated. This correlation is the correlation among the elements of a multinomial vector.

- In addition, the data are clustered, so repeated observations on the same subject are correlated because of shared characteristics and/or serial correlation.

Therefore, $\text{var}(\mathbf{y}_i) \equiv \mathbf{V}_i$ has a doubly-correlated structure, with typical elements given by

$$\text{cov}(y_{ijk}, y_{ij'k'}) = \begin{cases} \pi_{ijk}(1 - \pi_{ijk}) & \text{if } j = j', k = k', \\ -\pi_{ijk}\pi_{ijk'}, & \text{if } j = j', k \neq k', \\ \frac{\text{corr}(y_{ijk}, y_{ij'k'})}{\{\pi_{ijk}(1-\pi_{ijk})\pi_{ij'k'}(1-\pi_{ij'k'})\}^{1/2}} & \text{if } j \neq j'. \end{cases}$$

Suppose that we assume some link function g relating the π_{ijk}'s to covariates $\mathbf{x}_{ijk}$. For example, if the z_{ij}s were ordered categorical responses, then we might assume a cumulative logit link:

$$\text{logit}\{\Pr(z_{ij} \leq g)\} = \log\left\{\frac{\pi_{ij1} + \cdots + \pi_{ijg}}{\pi_{ij,g+1} + \cdots + \pi_{ijK}}\right\} = \mathbf{x}_{ijg}^T\boldsymbol{\beta}, \quad g = 1, \ldots, K-1.$$

The GEE for $\boldsymbol{\beta}$ is now

$$\sum_{i=1}^{n} \mathbf{D}_i^T \tilde{\mathbf{V}}_i^{-1}(\mathbf{y}_i - \boldsymbol{\pi}_i) = \mathbf{0},$$

where $\tilde{\mathbf{V}}_i$ is $\mathbf{V}_i$ with a working structure $\tilde{\text{corr}}(y_{ijk}, y_{ij'k'})$ used in place of $\text{corr}(y_{ijk}, y_{ij'k'})$, and $\mathbf{D}_i = \partial\boldsymbol{\pi}_i/(\partial\boldsymbol{\beta})$.

Example – Iowa 65+ Rural Health Survey:

1926 elderly individuals were followed over 6 yrs. Each subject was surveyed at years 0, 3, and 6.

Response: No. of friends reported.

- Response is ordinal: 1= no friends; 2= one or two friends; 3= three or more friends.

Question: *Is the distribution of reported number of friends changing over time?*

Data:

Year 0	Year 3	Year 6	Freq.
0	0	0	31
0	0	1–2	22
0	0	3+	54
0	1–2	0	15
⋮	⋮	⋮	⋮
3+	3+	3+	706
Total			1926

- See friends1.sas. In this program we fit multinomial response models using GEE to these data using PROC GENMOD. Note that SAS only implements the working independence structure and it only implements link functions for cumulative probabilities (e.g., cumulative logit links). Thus, GEE models for nominal categorical responses are not supported.

- The first model fit in friends1.sas is a cumulative logit model which assumes

$$\text{logit}\{\Pr(z_{ij} \leq 1)\} = \log\left\{\frac{\pi_{ij1}}{\pi_{ij2} + \pi_{ij3}}\right\} = \alpha_1 + \beta_j$$

$$\text{logit}\{\Pr(z_{ij} \leq 2)\} = \log\left\{\frac{\pi_{ij1} + \pi_{ij2}}{\pi_{ij3}}\right\} = \alpha_2 + \beta_j$$

  - For identifiability, SAS sets $\alpha_1 = 0$.

  - Thus, this model says that the log odds of have 0 friends is equal to β_1 in year 0, β_2 in year 3, and β_3 in year 6. The log odds of having 0–2 friends is $\alpha_2 + \beta_1$ in year0, $\alpha_2 + \beta_2$ in year 3, and $\alpha_2 + \beta_3$ in year 6.

- Such a model is known as a **proportional odds model,** because it implies that the ratio of the odds of having 0 friends for year 0 versus year 3 (say) is the same as the ratio of the odds of having 0–2 friends for year 0 versus year 3. The first of these ratios is given by

$$\frac{\exp(\alpha_1 + \beta_1)}{\exp(\alpha_1 + \beta_2)} = \exp(\beta_1 - \beta_2),$$

  and the second of these ratios is

$$\frac{\exp(\alpha_2 + \beta_1)}{\exp(\alpha_2 + \beta_2)} = \exp(\beta_1 - \beta_2).$$

- The PSCALE option introduces a dispersion parameter into the model to account for overdispersion relative to the multinomial variance. This dispersion parameter is estimated with the Pearson statistic. That is, the PSCALE option implies

$$\text{var}(\mathbf{y}_{ij}) = \phi(\text{diag}_{k=1}^{K-1}\{\pi_{ijk}\} - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}^T),$$

  where ϕ is to be estimated, whereas without the PSCALE option, $\phi = 1$.

- From the output on p.4 of friends1.lst we see that the estimated log odds of having 0 friends are going from -1.5034 to -1.8969, to -2.2987 over the 3 years in which subjects were surveyed. The estimated log odds of having 0–2 friends are going from 1.2696-1.5034 to 1.2696-1.8969, to 1.2696-2.2987 over the 3 years in which subjects were surveyed.

- A CONTRAST statement was used to determine whether the decrease in the log odds of having fewer friends is linear in time. The results strongly suggest that it is: linearity is significant (score test: $X^2 = 158.60$, $p < .0001$) and nonlinearity is not (score test: $X^2 = 0.01$, $p = .9296$).

- Therefore, we consider the linear trend model

$$\text{logit}\{\Pr(z_{ij} \leq 1)\} = \log\left\{\frac{\pi_{ij1}}{\pi_{ij2} + \pi_{ij3}}\right\} = \alpha_1 + \beta \text{year}_j$$

$$\text{logit}\{\Pr(z_{ij} \leq 2)\} = \log\left\{\frac{\pi_{ij1} + \pi_{ij2}}{\pi_{ij3}}\right\} = \alpha_2 + \beta \text{year}_j$$

- This model estimates a linear decrease in the log odds of having a small numbers of friends with a slope of $\hat{\beta} = -.1325$, which is significantly different from 0 (score test: $X^2 = 158.54$, $p < .0001$; Wald test: $Z = -12.99$, $p < .0001$).