Modeling the Mean and Covariance

Suppose we have a model of mean response as a product of time on a continuous value:

$$E(Y_{ij}) = \begin{cases} \beta_0 + \beta_1\, Time_{ij} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, Time_{ij} \end{cases}$$

The top is control and bottom is treatment. By allowing an index for the jth measurement we allow participants to have measurement in unequal time periods.

Consider the 3 hypotheses which could be tested:

- H0: $\beta_3 = 0$ for testing parallelism between groups
- H0: $\beta_1 = 0$ for testing flatness of change over time
- H0: $\beta_2 = 0$ for testing differences between groups

Quadratic Trends

Higher order terms should be tested (and removed from the model if appropriate) before the lower order terms

$$E(Y_{ij}) = \begin{cases} \beta_0 + \beta_1\, Time_{ij} + \beta_2\, Time_{ij}^2 \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4)\, Time_{ij} \\ \qquad + (\beta_2 + \beta_5)\, Time_{ij}^2 \end{cases}$$

In the above we would first test $\beta_2 = 0$ and if this fails to be rejected we can remove the term and test $\beta_1 = 0$

To avoid potential problems with collinearity we *center the variable of time around their mean before squaring it* in order to include it as a quadratic term

Linear Splines

There are application in which the mean response cannot be modeled accurately using a polynomial, like when the mean response increases rapidly for some duration and then more slowly thereafter. In such a case a linear spline model would be appropriate.

Assume the mean response follows a linear trend, **but the parameters change at a known time point t***; We can write the following linear spline model with a **knot** at time t*

$$\text{E}(Y_{ij}) = \beta_0 + \beta_1 Time_{ij} + \beta_2 Group_i + \beta_3 Time_{ij} \times Group_i + \\ \beta_4(Time_{ij} - t^*)_+ + \beta_5(Time_{ij} - t^*)_+ \times Group_i$$

where

$$(Time_{ij} - t^*)_+ = \begin{cases} Time_{ij} - t^* & \text{when } Time_{ij} > t^* \\ 0 & \text{when } Time_{ij} \leq t^* \end{cases}$$

We can then write the model separately for each group before and after t*

$$E(Y_{ij}) = \begin{cases} \beta_0 + \beta_1 Time_{ij} & Time_{ij} \leq t^* \\ (\beta_0 - \beta_4 t^*) + (\beta_1 + \beta_4) Time_{ij} & Time_{ij} > t^* \end{cases} \quad \text{Control}$$

$$E(Y_{ij}) = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) Time_{ij} & Time_{ij} \leq t^* \\ \{(\beta_0 + \beta_2) - (\beta_4 + \beta_5)t^*\} \\ + (\beta_1 + \beta_3 + \beta_4 + \beta_5) Time_{ij} & Time_{ij} > t^* \end{cases} \quad \text{Treatment}$$

We can test the following hypotheses:

- H0: $\beta_3 = \beta_5 = 0$ for testing differences in patterns of change
- H0: $\beta_3 = 0$ for testing group differences in patterns of change prior to t*
- H0: $\beta_4 = \beta_5 = 0$ for testing changes in the linear trend model after t*

Comparing Nested Models

When comparing nested models with respect to mean response, we can compute a likeliood ratio test (LTR) by running an ML (not REML) under the two models; full and reduced.

For example:

$$\text{Full model}: \quad y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \epsilon_i$$
$$\text{Reduced model}: \quad y_i = X_{1i}\beta_1 + \epsilon_i$$

testing for $\beta_2 = 0$ is given by:
LTR = 2 $(\ell(\beta_1, \beta_2) - \ell(\beta_2))$

For non-nested models we can use AIC or BIC:
AIC = -2(ℓ - c)
BIC = -2(ℓ - log(n)*c)
Where ℓ is the maximized log-likelihood, c is the number of regression parameters and n is the number of subjects. Select the model that minimized AIC or BIC. The disadvantage of this method is they are empirically/evidence based and we cannot use them to perform any inference.

Covariance Pattern Models

One of the most important aspects in modeling correlated data is estimating the covariance structure between measurement occasions. There is no single way of modeling the covariance structure, and a balance needs to be found when imposing structure.

- Too little structure may lead to many parameters for limited data, adversely effecting precision of β
- Too much structure improves β precision but increases the risk of mis-specification that could result in misleading inferences

Unstructured Covariance

For p measurement occasions:

$$\text{Cov}(Y_i) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

The covariance matrix must always be symmetric and positive-definite. The latter condition ensures that even though the repeated measurements can be highly correlated, none of them can be expressed as a linear combination of the others.

+ No explicit structure required for the covariance among the repeated measures
+ Can handle non-constant variance
- Unstructured covariance has (p*(p+1))/2 parameters. Even for a moderate number of repeated measurements the number of parameters needed to be estimated could be large
- Requires all individuals have the same measurement occasions. No irregular intervals

Compound Symmetry

For p measurement occasions, the compound symmetry model has the following form:

$$\text{Cov}(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

+ Very parsimonious, with only two parameters to be estimated regardless of the measurement occasions
+ Very appealing for certain designs, such as cluster sampling
- Requires strong assumption that the correlation between any pair of measurements to be the same no matter how large the time interval between the measurement is
- Assumes constant variance over time

Heterogeneous Compound Symmetry

For p measurement occasions, the heterogeneous compound symmetry model has the following form

$$\text{Cov}(Y_i) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \cdots & \rho\sigma_1\sigma_p \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_p\sigma_1 & \rho\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$$

This is a more flexible structure than CS, since it allows non-constant variance but with a smaller number of parameters to be estimated compared to unstructured covariance

First-Order Autoregressive Structure

For p measurement occasions the unstructured covariance has the following form:

$$\text{Cov}(Y_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{bmatrix}$$

+ Very parsimonious, with only two parameters to be estimated regardless of the measurement occasions
+ Appealing for longitudinal studies since the correlations decline over time as the distance between pairs of measurements increases
- Requires all measurements to be made at equal intervals of time
- Assumes constant variance over time

Heterogenous First-Order Autoregressive Structure

The constant variance assumption can be relaxed and we can have a heterogeneous AR(1) covariance pattern model:

$$\text{Cov}(Y_i) = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \cdots & \rho^{p-1}\sigma_1\sigma_p \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho^{p-2}\sigma_2\sigma_p \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1}\sigma_p\sigma_1 & \rho^{p-2}\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$$

Exponential Structure

When measurements are not equally spaced over time the AR(1) covariance structure can be reformulated as an exponential structure. Let {ti1 ... tip} denote the observation time for subject i, and assume variance is constant across measurement occasions

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho^{|t_{ij} - t_{ik}|}$$

The correlations declines exponentially over time as the distance between pairs of measurements increase, same as in the AR(1) case.
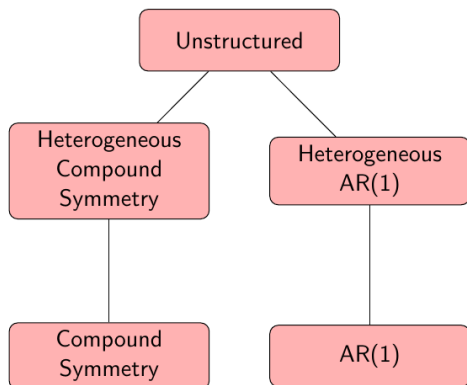
Choosing a Covariance Pattern Model

When the covariance model A with theta_a estimated parameters can be written as different covariance structure B (with theta_b > theta_a parameters) with certain restrictions imposed, then we say that A is in nested in B.

Ex. the compound symmetry model is nested within an unstructured covariance model. In this case we can test which covariance structure fits the data better by using a likelihood ratio test:

$$LR = 2(\ell_B - \ell_A)$$

where ℓ_B and ℓ_A are the maximized REML log-likelihoods. LR follows a $chi^2_{theta\_B - theta\_a}$

In many cases, we would like to compare non-nested models. For example, we would like to compare the fit between compound symmetry and AR(1) covariance structures it is not possible to use the LR test but we can use measures of fit for each model:
AIC = -2(ℓ - c)
BIC = -2(ℓ - log(n - p)*c)
where ℓ is the maximized log-likelihood, c is the number of covariance parameters and n is the number of subjects.