Modeling the Covariance

Biostatistics 653

Applied Statistics III: Longitudinal Analysis

Modeling the Covariance

- We simplify discussion by assuming $n_i = n$ for all subjects, so that $V(\epsilon_i) = \Sigma$. Later, we extend the discussion to consider the unbalanced case.

- We have previously considered three correlation structures: *independence*, *unstructured*, and *compound symmetry*.

Independence Structure

- The independence structure is rarely valid for longitudinal data but assumes

$$\Sigma = \sigma^2 \mathbf{I}_n = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Unstructured

- The unstructured covariance matrix is given by

$$\Sigma = \begin{pmatrix} \sigma^2_1 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma^2_2 & \sigma_{23} & \cdots & \sigma_{2n} \\ \vdots & & \ddots & & \vdots \\ \sigma_{1n} & \sigma_{2n} & \sigma_{3n} & \cdots & \sigma^2_n \end{pmatrix}$$

where we estimate n(n+1)/2 variance components.

- This structure is typically not preferred, due to the large number of parameters to be estimated and the fact that we can usually feel comfortable making some assumptions about the correlation structure over time.

Compound Symmetry

- The compound symmetry or exchangeable covariance matrix assumes all observations within a subject have the same correlation, regardless of their distance in time. It is given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \vdots & & \ddots & & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}$$

where we estimate 2 variance components.

- It is not appropriate if the timing between measurements in the study is thought to be related to their correlation.

Toeplitz

- The Toeplitz structure makes the assumption that any two responses that are equally spaced in time have the same correlation. It is given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}$$

in which we estimate n variance components. This structure does not assume that observations farther apart in time have lower correlation.

- This structure is not appropriate when measurements are not made at approximately fixed times.

- This structure may be simplified so that the correlation goes to zero for responses that are a given distance apart in time.

AR(1)

- The autoregressive of order 1, or AR(1), structure is a parsimonious structure that allows the correlation between two measurements to decrease as the time between them passes. Assuming the measures are made at approximately equal intervals of time across subjects, the structure is given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{pmatrix}$$

where we estimate 2 variance components.

Exponential

- The Markov or exponential structure is a generalization of the AR(1) structure to times that are unequally spaced. The powers of are taken to be the distances in time between the observations; that is

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \cdots & \rho^{|t_1-t_n|} \\ \rho^{|t_1-t_2|} & 1 & \rho^{|t_2-t_3|} & \cdots & \rho^{|t_2-t_{n-1}|} \\ \vdots & & \ddots & & \vdots \\ \rho^{|t_1-t_n|} & \rho^{|t_2-t_n|} & \rho^{|t_3-t_n|} & \cdots & 1 \end{pmatrix}$$

where we estimate 2 variance components.

Heterogeneous Variances

- We can relax the assumption of constant variances across measurement times by allowing heterogeneous variances. For example, for the AR(1) structure, a heterogeneous autoregressive covariance model is given by

$$\Sigma = \begin{pmatrix} \sigma^2_1 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \cdots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma^2_2 & \rho\sigma_2\sigma_3 & \cdots & \rho^{n-2}\sigma_2\sigma_n \\ \vdots & & \ddots & & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \cdots & \sigma^2_n \end{pmatrix}$$

where we estimate n+1 variance components.

Different Observation Times

- If the set of observation times is different for each subject (that is, $n_i \neq n$ for all i), then we should be careful in assuming certain covariance structures.

- Unstructured: must estimate large number of parameters (Σ_i could involve different covariances and variances for all i)

- Compound symmetry: does not incorporate actual values of times and so is still appropriate, though it has other drawbacks

- AR(1): can be nonsensical, as distance in time for adjacent observations for subject i might be longer than corresponding distance in time for subject i + 1

- Exponential: can assume $Cov(Y_{ij}, Y_{ik}) = \sigma^2 \rho^{|t_{ij} - t_{ik}|}$

Other Choices

- Other structures could be used as appropriate. For example, one potential drawback of the exponential covariance structure is that the correlation decays to zero as the time separation increases. However, one could assume that the covariance is a sum of two matrices, one with has an exponential structure, and one that has the compound symmetric structure. This allows us to place a lower bound on the correlation as measurements are further apart in time, consistent with the belief that some correlation in response is due to repeated measures on a single study subject but that measures taken closer in time will be more highly correlated than measures further apart in time. That is, we assume

Other Choices

$$\Sigma_i = \Sigma_{i1} + \Sigma_{i2}$$

$$\Sigma_{i1} = \sigma^2{}_1 \begin{pmatrix} 1 & \rho_1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \rho_1 & \cdots & 1 \end{pmatrix}$$

$$\Sigma_{i2} = \sigma^2{}_2 \begin{pmatrix} 1 & \rho_2^{|t_{i1}-t_{i2}|} & \rho_2^{|t_{i1}-t_{i3}|} & \cdots & \rho_2^{|t_{i1}-t_{in}|} \\ \rho_2^{|t_{i1}-t_{i2}|} & 1 & \rho_2^{|t_{i2}-t_{i3}|} & \cdots & \rho_2^{|t_{i2}-t_{in}|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_2^{|t_{i1}-t_{in}|} & \rho_2^{|t_{i2}-t_{in}|} & \rho_2^{|t_{i3}-t_{in}|} & \cdots & 1 \end{pmatrix}$$

Sum of Covariances

- Using this model, the correlation between replicate measurements on an individual at the same occasion would be $\frac{\rho\sigma_1^2+\sigma_2^2}{\sigma_1^1+\sigma_2^2}$, and as the time between measures on an individual gets quite large, the minimum correlation is $\frac{\rho\sigma_1^2}{\sigma_1^1+\sigma_2^2}$.

Covariance Model Selection

- Because model selection criteria for the mean depend on a correctly-specified covariance model, it often makes sense to select a covariance model first.

- However, we note that a covariance model depends on the assumed mean model.

- Because of this, we attempt to use a "maximal" model for the mean when choosing a covariance model.

Covariance Model Selection

After a maximal model for the mean has been selected, a number of different covariance models may be fit to the data. Covariance model selection may be done using (if appropriate)

- (restricted) likelihood ratio tests,

- AIC/BIC, or

- more complex (usually more valid!) procedures (this is an open research area).

Covariance Model Selection

- Likelihood ratio tests (using the REML likelihood) can be conducted for nested covariance models when the null does not involve setting a variance or covariance parameter equal to some value on the boundary of the parameter space (for example, $H_0: \sigma_B^2 = 0$ in a compound symmetry model). Two covariance models are nested when the reduced model is a special case of the full model. For example, the compound symmetry model is nested within the Toeplitz structure.

Covariance Model Selection

- The REML likelihood is preferred here because it reduces finite sample bias in covariance estimation. The likelihood ratio test is obtained by taking twice the difference in the maximized REML log-likelihoods, $T_{LRT} = 2(\hat{l}_{full} - \hat{l}_{reduced})$ and comparing the statistic to a chi-squared distribution with degrees of freedom equal to the difference between the number of covariance parameters in the full and reduced models.

Covariance Model Selection

- For certain null hypotheses, the likelihood ratio test may not be valid. In particular, we have problems when the null is on the boundary of the parameter space. Recall that variances cannot be negative and thus are bounded below by 0. Thus a likelihood ratio test of the null that a variance is zero is testing a null hypothesis that is on the boundary of the parameter space. In this case, the usual null distribution for the likelihood ratio test is no longer valid (i.e., not the chi-squared distribution previously described). If you ignore this problem, the resulting p-value is often overestimated, which could lead to selection of an overly parsimonious covariance model.

Covariance Model Selection

- The correct distribution for the likelihood ratio can be determined with effort but is not available in standard software. For this reason, some authors recommend using $\alpha = 0.10$ as an ad-hoc correction when testing values on the boundary of the parameter space. Selection of the covariance structure in a longitudinal data model is an open research question.

Covariance Model Selection

- For non-nested models, uncalibrated criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) may be used. The "best" model minimizes the criterion (in some software, the "best" model maximizes the negative of the criterion). These do not involve the comparison of two models and typically take the form

$$-2(\log \text{ likelihood}) + q(\text{number of parameters})$$

where q is a penalty against over-parameterized models. (Without q, we would select large covariance models over reduced models too often.) The penalty used by AIC is $q = 2$, and the penalty used by BIC is $q = \log(N)$.

Covariance Model Selection

- The BIC is sometimes defined using N* = N - dimension of β in REML estimation.

- The main idea behind BIC is to select the model with the highest posterior probability; however, BIC is only an approximation of the posterior model probability, and it places a very large penalty on estimation of each additional covariance parameter. For this reason, AIC is usually preferred for covariance model selection.

Example: Dental Data

- Using the preliminary mean model (saturated model), we selected a preliminary covariance structure (unstructured), and found the best mean model (depends on criterion used, but we'll go with the saturated model for kicks).

- Now, we will look at various options for the covariance model.

Example: Dental Data

```
/* saturated, UN covariance */
title2 'saturated, UN covariance';
proc mixed data=proyuniv;

class newid time;
model dist=male time male*time /solution;
repeated/type=un subject=newid r rcorr;
run;
/* saturated, CS covariance */
title2 'saturated, CS covariance';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=cs subject=newid r rcorr;
run;
```

Example: Dental Data

```
/* saturated, CSH covariance */
title2 'saturated, CS covariance';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=cs subject=newid r rcorr;
run;
/* saturated term, AR(1) */
title2 'linear term, AR(1) covariance';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=ar(1) subject=newid r rcorr;
run;
```

Example: Dental Data

```
/* saturated term, one-dependent */
title2 'saturated, One-dependent covariance';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=toep(2) subject=newid r rcorr;
run;
proc sort data=proyuniv; by gender;
/* saturated, UN covariance, by group */
title2 'saturated, UN covariance, by gender';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=un subject=newid r rcorr group=gender;
run;
```

Example: Dental Data

```
/* saturated, CS covariance, by gender */
title2 'saturated, CS covariance, by gender';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=cs subject=newid r rcorr group=gender;
run;
/* saturated, CSH covariance, by gender */
title2 'saturated, CS covariance, by gender';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=csh subject=newid r rcorr group=gender;
run;
```

Example: Dental Data

```
/* saturated, AR(1), by gender */
title2 'saturated, AR(1) covariance, by gender';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=ar(1) subject=newid r rcorr group=gender;
run;
/* saturated, one-dependent by gender */
title2 'saturated, One-dependent covariance, by gender';
proc mixed data=proyuniv;
class newid time;
model dist=male time male*time /solution;
repeated/type=toep(2) subject=newid r rcorr group=gender;
run;
```

Example: Dental Data

Estimated correlation matrices:

- Unstructured

$$\hat{\mathbf{R}} = \begin{pmatrix} 1.00 & 0.57 & 0.66 & 0.52 \\ 0.57 & 1.00 & 0.56 & 0.73 \\ 0.66 & 0.56 & 1.00 & 0.73 \\ 0.52 & 0.73 & 0.73 & 1.00 \end{pmatrix}$$

- Compound Symmetry

$$\hat{\mathbf{R}} = \begin{pmatrix} 1.00 & 0.62 & 0.62 & 0.62 \\ 0.62 & 1.00 & 0.62 & 0.62 \\ 0.62 & 0.62 & 1.00 & 0.62 \\ 0.62 & 0.62 & 0.62 & 1.00 \end{pmatrix}$$

- Heterogeneous Compound Symmetry

$$\hat{\mathbf{R}} = \begin{pmatrix} 1.00 & 0.63 & 0.63 & 0.63 \\ 0.63 & 1.00 & 0.63 & 0.63 \\ 0.63 & 0.63 & 1.00 & 0.63 \\ 0.63 & 0.63 & 0.63 & 1.00 \end{pmatrix}$$

Example: Dental Data

Estimated correlation matrices:

- AR(1)

$$\widehat{\mathbf{R}} = \begin{pmatrix} 1.00 & 0.62 & 0.38 & 0.23 \\ 0.62 & 1.00 & 0.62 & 0.38 \\ 0.38 & 0.62 & 1.00 & 0.62 \\ 0.23 & 0.38 & 0.62 & 1.00 \end{pmatrix}$$

- One-Dependent

$$\widehat{\mathbf{R}} = \begin{pmatrix} 1.00 & 0.36 & 0.00 & 0.00 \\ 0.36 & 1.00 & 0.36 & 0.00 \\ 0.00 & 0.36 & 1.00 & 0.36 \\ 0.00 & 0.00 & 0.36 & 1.00 \end{pmatrix}$$

Example: Dental Data

Covariance model	# covariance parameters	$-2\hat{l}_{REML}$	AIC	BIC
Unstructured	10	414.0	434.0	447.0
Compound Symmetry	2	423.4	427.4	430.0
Heterogeneous CS	5	421.4	431.4	437.9
AR(1)	2	450.8	454.8	457.4
One-dependent	2	462.0	466.0	468.6
Unstructured (gender-specific)	20	392.9	432.9	458.8
CS (gender-specific)	4	406.4	414.4	419.5
HCS (gender-specific)	10	402.1	422.1	435.1
AR(1) (gender-specific)	4	412.5	420.5	425.7
One-dependent (gender-specific)	4	439.6	447.6	452.8