BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 9

Modelling the Covariance

Modelling the Covariance

Longitudinal studies present two aspects of the data that require modelling: mean response over time and covariance.

Although these two aspects of the data can be modelled separately, the models for the mean response and covariance are interdependent.

Recall: Covariance between any pair of residuals, say $[Y_{ij} - \mu_{ij}(\beta)]$ and $[Y_{ik} - \mu_{ik}(\beta)]$, depends on the model for the mean, i.e., depends on β.

A model for the covariance must be chosen on the basis of some assumed model for the mean response.

Modelling the Covariance

Three broad approaches can be distinguished:

(1) "unstructured" or arbitrary pattern of covariance

(2) covariance pattern models

(3) random effects covariance structure

Unstructured Covariance

Appropriate when design is "balanced" and number of measurement occasions is relatively small.

No explicit structure is assumed other than homogeneity of covariance across different individuals, $\text{Cov}(Y_i) = \Sigma_i = \Sigma$.

Chief advantage: no assumptions made about the patterns of variances and covariances.

Major drawback:

Number of covariance parameters grows rapidly with the number of measurement occasions.

Recall: With n measurement occasions, "unstructured" covariance matrix has $\frac{n\times(n+1)}{2}$ parameters:

the n variances and $n \times (n-1)/2$ pairwise covariances (or correlations),

$$\mathrm{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix}.$$

For $n = 3$ number of covariance parameters is 6

For $n = 5$ number of covariance parameters is 15

For $n = 10$ number of covariance parameters is 55

When number of covariance parameters is large, relative to sample size, estimation is likely to be very unstable.

Use of an unstructured covariance is appealing only when N is large relative to $\frac{n\times(n+1)}{2}$.

Additional major drawback: Unstructured covariance is problematic when there are mistimed measurements.

Covariance Pattern Models

When attempting to impose some structure on the covariance, a subtle balance needs to be struck.

With too little structure there may be too many parameters to be estimated with limited amount of data.

With too much structure, potential risk of model misspecification and misleading inferences concerning β.

Classic tradeoff between bias and variance.

Covariance pattern models have their basis in models for serial correlation originally developed for time series data.

Compound Symmetry

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,k}) = \rho$ for all j and k.

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

Parsimonious: two parameters regardless of number of measurement occasions.

Strong assumptions about variance and correlation are usually not valid with longitudinal data.

Toeplitz

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho_k$ for all j and k.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}.$$

Assumes correlation among responses at adjacent measurement occasions is constant, ρ_1.

Toeplitz only appropriate when measurements are made at equal (or approximately equal) intervals of time.

Toeplitz covariance has n parameters (1 variance parameter, and $n - 1$ correlation parameters).

A special case of the Toeplitz covariance is the (first-order) autoregressive covariance.

Autoregressive

Assumes variance is constant across occasions, say σ^2, and $\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho^k$ for all j and k, and $\rho \geq 0$.

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}.$$

Parsimonious: only 2 parameters, regardless of number of measurement occasions.

Only appropriate when the measurements are made at equal (or approximately equal) intervals of time.

Compound symmetry, Toeplitz and autoregressive covariances assume variances are constant across time.

This assumption can be relaxed by considering versions of these models with heterogeneous variances, $\text{Var}(Y_{ij}) = \sigma_j^2$.

A heterogeneous autoregressive covariance pattern:

$$\text{Cov}(Y_i) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \dots & \rho^{n-1}\sigma_1\sigma_n \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \dots & \rho^{n-2}\sigma_2\sigma_n \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 & \dots & \rho^{n-3}\sigma_3\sigma_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1}\sigma_1\sigma_n & \rho^{n-2}\sigma_2\sigma_n & \rho^{n-3}\sigma_3\sigma_n & \dots & \sigma_n^2 \end{pmatrix},$$

and has $n + 1$ parameters (n variance parameters and 1 correlation parameter).

Banded

Assumes correlation is zero beyond some specified interval.

For example, a banded covariance pattern with a band size of 3 assumes that $\text{Corr}(Y_{ij}, Y_{i,j+k}) = 0$ for $k \geq 3$.

It is possible to apply a banded pattern to any of the covariance pattern models considered so far.

A banded Toeplitz covariance pattern with a band size of 2 is given by,

$$\text{Cov}(Y_i) = \sigma^2 \begin{pmatrix} 1 & \rho_1 & 0 & \dots & 0 \\ \rho_1 & 1 & \rho_1 & \dots & 0 \\ 0 & \rho_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix},$$

where $\rho_2 = \rho_3 = \cdots = \rho_{n-1} = 0$.

Banding makes very strong assumption about how quickly the correlation decays to zero with increasing time separation.

Exponential

When measurement occasions are not equally-spaced over time, autoregressive model can be generalized as follows.

Let $\{t_{i1}, ..., t_{in}\}$ denote the observation times for the i^{th} individual and assume that the variance is constant across all measurement occasions, say σ^2, and

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

for $\rho \geq 0$.

Correlation between any pair of repeated measures decreases exponentially with the time separations between them.

Referred to as "exponential" covariance because it can be re-expressed as

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \sigma^2 \rho^{|t_{ij} - t_{ik}|} \\ &= \sigma^2 \exp\left(-\theta \left|t_{ij} - t_{ik}\right|\right), \end{aligned}$$

where $\theta = -\log(\rho)$ or $\rho = \exp(-\theta)$ for $\theta \geq 0$.

Exponential covariance model is invariant under linear transformation of the time scale.

If we replace t_{ij} by $(a + bt_{ij})$ (e.g., if we replace time measured in "weeks" by time measured in "days"), the same form for the covariance matrix holds.

Choice among Covariance Pattern Models

Choice of models for covariance and mean are interdependent.

Choice of model for covariance should be based on a "**maximal**" model for the mean that minimizes any potential misspecification.

With balanced designs and a very small number of discrete covariates, choose "saturated model" for the mean response.

Saturated model: includes main effects of time (regarded as a within-subject factor) and all other main effects, in addition to their two- and higher-way interactions.

Maximal model should be in a certain sense the most elaborate model for the mean response that we would consider from a subject-matter point of view.

Once maximal model has been chosen, residual variation and covariation can be used to select appropriate model for covariance.

For nested covariance pattern models, a likelihood ratio test statistic can be constructed that compares "full" and "reduced" models.

Recall: two models are said to be nested when the "reduced" model is a special case of the "full" model.

For example, compound symmetry model is nested within the Toeplitz model, since if the former holds the latter must necessarily hold, with $\rho_1 = \rho_2 = \cdots = \rho_{n-1}$.

Likelihood ratio test is obtained by taking twice the difference in the respective maximized REML log-likelihoods,

$$G^2 = 2(\widehat{l}_{\text{full}} - \widehat{l}_{\text{red}}),$$

and comparing statistic to a chi-squared distribution with df equal to difference between the number of covariance parameters in full and reduced models. [Note: this assumes that a parameter is not being set to a boundary value, e.g. a variance is not being set to zero].

To compare non-nested model, an alternative approach is the Akaike Information Criterion (AIC).

According to the AIC, given a set of competing models for the covariance, one should select the model that minimizes

$$\begin{aligned} \text{AIC} &= -2(\text{maximized log-likelihood}) + 2(\text{number of parameters}) \\ &= -2(\widehat{l} - c), \end{aligned}$$

where $\widehat{l}$ is the maximized REML log-likelihood and c is the number of covariance parameters.

Example: Exercise Therapy Trial

- Subjects were assigned to one of two weightlifting programs to increase muscle strength.
- Treatment 1: number of repetitions of the exercises was increased as subjects became stronger.
- Treatment 2: number of repetitions was held constant but amount of weight was increased as subjects became stronger.
- Measurements of body strength were taken at baseline and on days 2, 4, 6, 8, 10, and 12 (data are in 'stren.txt'), though there were a few missing measurements.
- For illustration, we focus first only on measures of strength obtained at baseline (or day 0) and on days 4, 8, and 12; so equally spaced.

Before considering models for the covariance, it is necessary to choose a maximal model for the mean response.

We chose maximal model to be the saturated model for the mean.

First, we consider an unstructured covariance matrix.

Note that the variance appears to be larger by the end of the study when compared to the variance at baseline.

Correlations decrease as the time separation between the repeated measures increases.

Table 1: Estimated unstructured covariance matrix for the strength data at baseline, day 4, day 8, and day 12.

Day	0	4	8	12
0	9.668	10.186	9.698	9.536
4	10.186	12.606	12.374	12.069
8	9.698	12.374	13.426	13.007
12	9.536	12.069	13.007	14.095

Table 2: Estimated unstructured correlation matrix for the strength data at baseline, day 4, day 8, and day 12.

Day	0	4	8	12
0	1.0000	0.9227	0.8512	0.8169
4	0.9227	1.0000	0.9511	0.9054
8	0.8512	0.9511	1.0000	0.9455
12	0.8169	0.9054	0.9455	1.0000

Despite apparent increase in variance over time, we consider an autoregressive model for the correlation.

Assume variance is constant across occasions, say σ^2, and
$\text{Corr}(Y_{ij}, Y_{i,j+k}) = \rho^k$ for all j and k, and $\rho \geq 0$.

This results in the following estimates of the variance and correlation parameters, $\widehat{\sigma}^2 = 11.91$ and $\widehat{\rho} = 0.9340$.

Table 3: Estimated autoregressive covariance matrix for the strength data at baseline, day 4, day 8, and day 12.

Day	0	4	8	12
0	11.908	11.121	10.387	9.702
4	11.121	11.908	11.121	10.387
8	10.387	11.121	11.908	11.121
12	9.702	10.387	11.121	11.908

Table 4: Estimated autoregressive correlation matrix for the strength data at baseline, day 4, day 8, and day 12.

Day	0	4	8	12
0	1.0000	0.9340	0.8723	0.8147
4	0.9340	1.0000	0.9340	0.8723
8	0.8723	0.9340	1.0000	0.9340
12	0.8147	0.8723	0.9340	1.0000

There is a hierarchy among the models: autoregressive is nested within unstructured.

LRT comparing autoregressive and unstructured covariance,

$$G^2 = 516.7 - 509.0 = 7.7,$$

and can be compared to a chi-squared distribution with 8 (i.e., 10 - 2) degrees of freedom (p close to 0.5).

Thus, based on statistical test, the autoregessive model provides an adequate fit to the covariance (but could be other rationale for choosing a more complex model).

Now, let's consider analysis if we focus on measures of strength obtained at baseline (or day 0), day 2 and days 4, 8 and 12; so now have unequally spaced measurements.

Table 5: Estimated unstructured covariance matrix for the strength data at baseline, day 2, day 4, day 8, and day 12.

Day	0	2	4	8	12
0	9.668	9.908	10.177	9.690	9.501
2	9.908	10.851	10.850	10.236	9.964
4	10.177	10.850	12.559	12.341	12.016
8	9.690	10.236	12.341	13.431	12.993
12	9.501	9.964	12.016	12.993	14.066

Table 6: Estimated unstructured correlation matrix for the strength data at baseline, day 2, day 4, day 8, and day 12.

Day	0	2	4	8	12
0	1.0000	0.9673	0.9236	0.8503	0.8147
2	0.9673	1.0000	0.9295	0.8480	0.8065
4	0.9236	0.9295	1.0000	0.9503	0.9041
8	0.8503	0.8480	0.9503	1.0000	0.9453
12	0.8147	0.8065	0.9041	0.9453	1.0000

Consider exponential model for the covariance, where

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

for $t_i = (0, 2, 4, 8, 12)$ for all subjects.

Results: $\widehat{\sigma}^2 = 11.92$ and $\widehat{\rho} = 0.9794$.

Table 7: Estimated exponential correlation matrix for the strength data at baseline, day 2, day 4, day 8, and day 12.

Day	0	2	4	8	12
0	1.0000	0.9593	0.9202	0.8468	0.7792
2	0.9593	1.0000	0.9593	0.8827	0.8123
4	0.9202	0.9593	1.0000	0.9202	0.8468
8	0.8468	0.8827	0.9202	1.0000	0.9202
12	0.7792	0.8123	0.8468	0.9202	1.0000

There is a hierarchy among the models: exponential is nested within unstructured.

LRT comparing exponential and unstructured covariance, yields

$$G^2 = 612.7 - 592.5 = 20.2,$$

and when compared to a chi-squared distribution with 13 degrees of freedom (i.e., 15 - 2), $0.1 > p > 0.05$.

Exponential covariance provides a reasonable fit to the data (but could be other rationale for choosing a more complex model).

Strengths/Weaknesses of Covariance Pattern Models

Covariance pattern models attempt to characterize the covariance with a relatively small number of parameters.

However, many models (e.g., autoregressive, Toeplitz, and banded) appropriate only when repeated measurements are obtained at equal intervals and hence cannot handle irregularly timed measurements.

While there is a large selection of models for correlations, choice of models for variances is limited.

They are not well-suited for modelling data from inherently unbalanced longitudinal designs.

Table 8: Covariance pattern modelling options using PROC MIXED in SAS.

TYPE =	<pattern>	Specifies the covariance pattern
	UN	Unstructured
	CS	Compound symmetry
	AR(1)	First-order autoregressive
	TOEP	Toeplitz
	UN(n)	Banded unstructured, with n bands
	CSH	Heterogeneous compound symmetry
	ARH(1)	Heterogeneous first-order autoregressive

Table 9: Illustrative commands for an autoregressive model using PROC MIXED in SAS.

```
PROC MIXED;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=AR(1) SUBJECT=id R RCORR;
```

Table 10: Illustrative commands for an exponential model using PROC MIXED in SAS.

```
PROC MIXED;
  CLASS id group time;
  MODEL y=group time group*time /S CHISQ;
  REPEATED time / TYPE=SP(EXP)(ctime) SUBJECT=id R RCORR;
```