Lecture 4: Covariance pattern models

Antonello Maruotti

Lecturer in Medical Statistics, S3RI and School of Mathematics
University of Southampton

Southampton, 13 June 2014

Summary

Covariance structure for repeated measurements

Modelling growth curves

Linear mixed models

- Statistical linear mixed models state that observed data consist of two parts
  - fixed effects
  - random effects
- Fixed effects define the expected values of the observations
- Random effects result from variation between subjects and from variation within subjects.

Linear mixed models

- Measures on the same subject at different times almost always are correlated, with measures taken close together in time being more highly correlated than measures take far apart in time
- Observations on different subjects are often assume independent
- Mixed linear models are used with repeated measures data to accommodate the fixed effects of covariates and the covariation between observations on the same subject at different times

Linear mixed models

- To model the mean structure in sufficient generality to ensure unbiasedness of the fixed effect estimates
- To specify a model for a covariance structure of the data
- Estimation methods are used to fit the mean portion of the model
- The fixed effects portion may be made more parsimonious
- Statistical inference are drawn base on fitting this final model

Model specification

- Let Y_{ijk} denote the value of the response measured at time k on subject j in group i
- $Y_{ijk} = (\beta_0 + \alpha_{0ij}) + (\beta_1 + \alpha_{1ij})x_{ij} + \epsilon_{ijk}$
- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{e}$
  $\mathbf{X}$ is a matrix of known covariates
  $\boldsymbol{\beta}$ is the vector of fixed parameters
  $\mathbf{Z}$ is a matrix collecting random effects
  $\boldsymbol{\alpha}$ is the vector of random parameters.

Covariance structure

- We assume that α and $\mathbf{e}$ are independent
- $\text{Var}(\mathbf{Y}) = \mathbf{ZGZ'}+\mathbf{R}$
  $\mathbf{G} = \text{Var}(\mathbf{U})$
  $\mathbf{R} = \text{Var}(\mathbf{e})$
- **ZGZ'** represents the between-subject portion of the covariance structure
- **R** represents the within-subject portion
- Modelling covariance structure refers to representing $\text{Var}(\mathbf{Y})$ as a function of a relatively small number of parameters.
- Functional specification of the covariance structure is done through **G** and **R**, often only in terms of $\mathbf{R}_{ij}$

Simple covariance structure

- Simple structure specifies that
  - the observations are independent
  - the observations have homogeneous variance σ^2
- The correlation function is equal to zero
- Simple structure is not realistic because it specifies that the observations on the same patient are independent
- $\mathbf{G}$=0 and $\mathbf{R}_{ij} = \sigma^2\mathbf{I}$

Compound Symmetric or Exchangeable

- Exchangeable structure specifies that observations on the same subject have homogeneous covariance σ_1 and homogeneous variance σ^2
- The correlation does not depend on the value of the lag, i.e. the correlations between two observations are equal for all pairs of observations on the same subject
- Exchangeable structure is often called *variance components* structure ,where σ_1 and $sigma^2$ represent between-subject and within-subject variances, respectively
- It can be specified in two ways through **G** and **R**
  - $\mathbf{G} = \sigma_1 \mathbf{I}, \quad \mathbf{R} = \sigma^2 \mathbf{I}$
  - $\mathbf{G} = \mathbf{0}, \quad \mathbf{R}_{ij} = \sigma^2 \mathbf{I} + \sigma_1 \mathbf{J}$

Autoregressive

- Autoregressive covariance structure specifies homogeneous variance, σ^2
- It specifies that covariances between observations on the same patient are not equal, but decrease towards zero with increasing lag.
- Autoregressive structure is entirely defined in terms of $\mathbf{R}$

$$\mathbf{G} = \mathbf{0}, \quad \mathbf{R}_{ij} = \sigma^2 \rho^{|k-l|}$$

Toeplitz

- Toeplitz structure specifies that covariance depends only on lag, but not as a mathematical function with a smaller number of parameters.
- Toeplitz structure is given with $\mathbf{G} = \mathbf{0}$. The elements of the main diagonal of $\mathbf{R} = \sigma^2$. All elements in a sub-diagonal $|k - l| = lag$ are $\sigma_{|k-l|}$

Unstructured

- The unstructured structure specifies no patterns in the covariance matrix
- The covariance matrix is completely general
- A very large number of parameters need to be estimated

Example

- Study reported in Pothoff and Roy (1964) and analysed further by Jennrich and Schluchter (1986)
- Growth measurements for 11 girls and 16 boys at 8, 10 12 and 14 years old

person:	Coded 1, 2,. . . , 27
occasion:	Coded 1, 2, 3, 4
gender:	Coded 1 for male and 2 for female (a person-level variable)
y:	Distance (mm) from the centre of the pituitary to the pterygomaxillary fissure
age:	Age in years (an occasion-level variable)

Statistical Modelling

How might we model growth?

- Assume independence between subjects, but expect repeat measurements on the same individual to be correlated
- Regard occasions as nested within subject
- Could formulate a linear mixed model utilising a random coefficients model
- An alternative formulation of the linear mixed model is to focus on the:
  - Marginal mean structure $\Rightarrow$ Modelled using fixed effects
  - Marginal covariance structure $\Rightarrow$ Requiring a model for the pattern of the covariance matrix associated with random variation between individual subject response vectors
- Known as a marginal model, or covariance pattern model $\Rightarrow$ response $=$ fixed effects model for mean $+$ random error

Covariance pattern models

A possible covariance pattern model for the growth study data, assumes homogeneous average growth rates, is as follows. Let y_{ijk} be the distance measurement on the jth subject in the ith gender group at the kth time point (occasion).

$$y_{ijk} = \beta_0 + g_i + \beta_1 age + \epsilon_{ijk}$$

- ϵ_{ijk} is a normal random error term
  - Mean 0
  - Correlated within subjects
  - Independent across subjects
- $Var(\epsilon_{ijk}) = \Sigma$, where ϵ_{ij} is the error vector associated with the jth subject in the ith group

xtmixed

independent

$$\begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma^2 \end{bmatrix}$$

- This then represents a linear model
- Not usually of interest in covariance pattern models
- Default in Stata

xtmixed

exchangeable

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \cdots & \sigma_1 & \sigma^2 \end{bmatrix}$$

- This is the implied marginal covariance structure from a random intercept model
- But with a covariance pattern model σ_1 is not a variance and hence may be negative

xtmixed

unstructured

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \cdots & \sigma_{K,K-1} & \sigma_K^2 \end{bmatrix}$$

- For a p-dimensional covariance matrix, $p(p+1)/2$ parameters are required, becoming large very rapidly as p increases

xtmixed

toeplitz 1

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & 0 \\ 0 & \sigma_{32} & \sigma_3^2 & \sigma_{33} \\ 0 & 0 & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

- Generally more appropriate for equally, or near equally, spaced time points

xtmixed

ar 1

$$\begin{bmatrix} \sigma_1^2 & \rho & \rho^2 & \rho^3 \\ \rho & \sigma_2^2 & \rho & \rho^2 \\ \rho^2 & \rho & \sigma_3^2 & \rho \\ \rho^3 & \rho^2 & \rho & \sigma_4^2 \end{bmatrix}$$

- $|\rho| < 1$
- Generally more appropriate for equally, or near equally, spaced time points

Output

```
                                                Wald chi2(2)      =     100.00
Log restricted-likelihood = -214.34726          Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
           1 |   2.045173   .7361411     2.78   0.005     .6023626    3.487983
           2 |           0  (base)
             |
         age |   .6746507   .0702289     9.61   0.000     .5370045    .8122969
       _cons |   15.37242   .9218572    16.68   0.000     13.56562    17.17923
------------------------------------------------------------------------------
```

Hence, the estimated mean growth profiles are:

$$\text{Females}\quad E(Y) = 15.37 + 0.67 \times \text{ age}$$

$$\text{Males}\quad E(Y) = 15.37 + 2.05 + 0.67 \times \text{ age}$$

Interpretation of fixed effects

- Age effect
  - Wald z-test: z=9.61, 1df, p-vaue¡0.001
    - Highly significant age effect
    - Not that surprising - why?
  - For a fixed gender estimated mean growth rate = 0.67 mm year (95% CI = [0.54, 0.81])
- Gender effect
  - Wald z-test: z = 2.78, 1df, p-value=0.005
    - Significant gender effect
  - For a fixed ages, estimated difference (M-F) = 2.05 mm (95% CI = [0.60, 3.49])

Output

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
person:               (empty) |
-----------------------------+------------------------------------------------
Residual: Unstructured       |
                     var(e1) |   5.374651   1.509931       3.098971    9.321441
                     var(e2) |   4.215106   1.202134       2.410176    7.371712
                     var(e3) |   6.335558    1.76549       3.669322    10.93916
                     var(e4) |   5.376425   1.609696       2.989818    9.668131
                  cov(e1,e2) |   2.786985   1.112032       .6074414    4.966528
                  cov(e1,e3) |   3.807089   1.371279       1.119431    6.494746
                  cov(e1,e4) |     2.6284   1.208944        .258913    4.997886
                  cov(e2,e3) |   2.909681   1.176418       .6039444    5.215418
                  cov(e2,e4) |     3.1684   1.153559        .907465    5.429334
                  cov(e3,e4) |   4.301478   1.486588       1.387819    7.215137
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(9) =    55.83   Prob > chi2 = 0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the
      boundary of the parameter space.
      If this is not true, then the reported test is conservative.
```

```
. ssc install xtmixed_corr, replace
. xtmixed_corr
Standard deviations and correlations for person = 1:
Standard Deviations:
    occasion |      1       2       3       4
-------------+--------------------------------
          sd |  2.318   2.053   2.517   2.319

Correlations:
    occasion |      1       2       3       4
-------------+--------------------------------
           1 |  1.000
           2 |  0.586   1.000
           3 |  0.652   0.563   1.000
           4 |  0.489   0.666   0.737   1.000
```

Heterogeneous average growth rates

- Is there a significant interaction between gender and age?
- Add an interaction term into the previous model

$$y_{ijk} = \beta_0 + g_i + \beta_{1i} age + \epsilon_{ijk}$$

Output

```
                                                Wald chi2(3)      =     131.93
Log restricted-likelihood =  -212.2734          Prob > chi2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |
          1  |  -1.583079   1.523314    -1.04   0.299    -4.568718    1.402561
          2  |           0  (base)
             |
         age |   .4763647   .0991584     4.80   0.000     .2820178    .6707116
             |
gender#c.age |
          1  |   .3504386   .1288105     2.72   0.007     .0979746    .6029026
          2  |           0  (base)
             |
       _cons |   17.42537   1.172647    14.86   0.000     15.12702    19.72372
------------------------------------------------------------------------------
```

Output

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
person: (empty)				
Residual: Unstructured				
var(e1)	5.425229	1.533941	3.117077	9.442537
var(e2)	4.190608	1.185948	2.4065	7.297403
var(e3)	6.263177	1.730251	3.644548	10.76331
var(e4)	4.986176	1.410422	2.86412	8.680484
cov(e1,e2)	2.709224	1.094638	.563772	4.854676
cov(e1,e3)	3.841122	1.376514	1.143204	6.53904
cov(e1,e4)	2.715146	1.1749	.4123844	5.017908
cov(e2,e3)	2.97451	1.17745	.6667494	5.28227
cov(e2,e4)	3.313683	1.129372	1.100154	5.527212
cov(e3,e4)	4.133223	1.377576	1.433223	6.833223

LR test vs. linear regression: chi2(9) = 59.01 Prob > chi2 = 0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space.
If this is not true, then the reported test is conservative.

```
. xtmixed_corr

Standard deviations and correlations for person = 1:

Standard Deviations:

    occasion |      1      2      3      4
-------------+--------------------------------
          sd |  2.329   2.047   2.503   2.233

Correlations:

    occasion |      1      2      3      4
-------------+--------------------------------
           1 |  1.000
           2 |  0.568   1.000
           3 |  0.659   0.581   1.000
           4 |  0.522   0.725   0.740   1.000
```

Choice of covariance structure

- The previous analysis assumed an unstructured covariance matrix for ϵ_{ij}
- From inspecting the estimated covariance matrix, $\hat{\Sigma}$, might there be a simpler covariance structure which is plausible?
- A more parsimonious covariance, if valid, is desirable for improved precision and power
- An incorrect choice of covariance structure may lead to erroneous conclusions
- An unstructured covariance matrix is not necessarily correct. Why?

Choice of covariance structure

Different covariance structures may be compared:

- Descriptively
- Using a hypothesis test the likelihood ratio test, assuming $\Rightarrow$ Nested models
- Using information theoretic criteria AIC, BIC$\Rightarrow$ Useful for comparing non-nested models

An example

Growth study with heterogeneous average growth rates

To compare different covariance structures for the same fixed effects model, use change in 2REML logL for an approximate LR test:

Covariance Structure	-2*log L	parameters
H0: CS	433.76	2
H1: UN	424.55	10

- Change in 2* logL $= 9.21$ on 8 df. Compare with upper percentage points of χ^2_8 gives p-value $= 0.32$
- No evidence against H0. Reasonable to use compound symmetry covariance structure

Final remarks

- Covariance pattern models are useful when primary interest is in modelling the mean structure
- Offers a flexible modelling approach
  - Different covariance structures are permissible
  - Allows for subjects with missing values
  - Time dependent explanatory variables
- Covariance pattern models are essentially linear models, allowing for correlated errors and heterogeneous variance