BIOS6643 Longitudinal

L8 Covariance

EJC

Department of Biostatistics   Informatics

Covariance Structure

Kronecker Product

Mt. K data

Fitting joint normal outcomes using mixed models

Some notes

Summary

Topics for today (Advanced covariance structures):

- Recap from last lecture: Mt. K data and fit with random effects; some unusual results: non-positive definite G; partial random effect results; model still usable?
- Quick review of $\boldsymbol{R}$ structures, up to Kronecker Product
- Modeling $\boldsymbol{G}$ and $\boldsymbol{R}$ simultaneously
- Another look at Mt. K data, several modeling approaches.
- Fitting a joint normal outcome using a mixed model!

Reading: Relevant sections from the LMM course notes.

Recap of R_i structures for, say, 4 measures

- Simple
- CS
- UN
- AR(1)
- Spatial power

Many others, but the above are ones I most commonly use. For 'CS', I usually just a random intercept, in which case you don't need to additionally specify CS for the $\boldsymbol{R}$ structure.

Kronecker Product

Modeling doubly repeated measures via the error covariance (R) matrix using the Direct Product (Kronecker) structure

- For some data sets, we may need to account for repeated measures over two dimensions.
- For example, say that strength measurements are taken on each leg for subjects over 3 time points. There are 2 'repeated measures' in space (i.e., body part) as well as 3 repeated measures over time.
- The covariance matrix to account for all repeated measures would then have size 6×6.
- Instead of dealing with this big messy matrix, it is easier to define it in pieces and then take the Kronecker product.

Example 1:

For the scenario described above, say that repeated measures over space can be modeled with the UN structure and repeated measures over time can be modeled with the AR(1) structure:

Structure for space:

$$R_{i1} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Structure for time:

$$R_{i2} = \sigma_\epsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix}$$

The combined (Kronecker product structure):

Example 1

$$R_i = R_{i1} \otimes R_{i2} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \otimes \sigma_\epsilon^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 R_{i2} & \sigma_{12} R_{i2} \\ \sigma_{12} R_{i2} & \sigma_2^2 R_{i2} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \\ \sigma_{12} \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} & \sigma_2^2 \begin{pmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{pmatrix} \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_1^2\phi & \sigma_1^2\phi^2 & \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 \\ \sigma_1^2\phi & \sigma_1^2 & \sigma_1^2\phi & \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi \\ \sigma_1^2\phi^2 & \sigma_1^2\phi & \sigma_1^2 & \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} \\ \sigma_{12} & \sigma_{12}\phi & \sigma_{12}\phi^2 & \sigma_2^2 & \sigma_2^2\phi & \sigma_2^2\phi^2 \\ \sigma_{12}\phi & \sigma_{12} & \sigma_{12}\phi & \sigma_2^2\phi & \sigma_2^2 & \sigma_2^2\phi \\ \sigma_{12}\phi^2 & \sigma_{12}\phi & \sigma_{12} & \sigma_2^2\phi^2 & \sigma_2^2\phi & \sigma_2^2 \end{pmatrix}$$

Note: the σ_ϵ^2 on the AR(1) structure is not included because it becomes redundant once we take the direct product, i.e., it is absorbed into parameters in the other matrix. Available Kronecker structures in SAS include: $UN@AR(1)$, $UN@CS$ and $UN@UN$. (The symbol '@' is used to denote '$\otimes$' since the latter is not on the keyboard!)

Example 2

Measuring cortisol 3 times a day, for 7 successive days.

Some things to remember about covariance matrices in mixed models:

- $Cov[\boldsymbol{Y}]$ or $Var[\boldsymbol{Y}] = \boldsymbol{V} = \boldsymbol{ZGZ}^\top + \boldsymbol{R}$
- If no random effects, then $Cov[\boldsymbol{Y}] = Cov[\boldsymbol{\epsilon}]$, i.e., $\boldsymbol{V} = \boldsymbol{R}$.
- Since the $Cov[\boldsymbol{Y}]$ is of primary interest, this formula suggests there are different ways to account for correlated data, through $\boldsymbol{R}$ (the covariance matrix of the errors), through $\boldsymbol{G}$ (The covariance matrix of random effects) or both.

Putting it together:

Specifying $\boldsymbol{G}$ and $\boldsymbol{R}$ in the same model

- So far we've discussed how you can either specify $\boldsymbol{G}$ or $\boldsymbol{R}$ in fitting a mixed model. However, you can actually do both, which may be advantageous for some data.
- Recall the Mt. Kilimanjaro data. By including up to quadratic random effects, we can actually develop an altitude-sensitive covariance structure that allows the correlation to decrease as altitude between measurements increases. (Altitude and time are closely related.)
- We can also directly model the repeated measures through $\boldsymbol{R}$. (Although repeated measures are not likely to be exactly equally spaced, we do not have exact times of measurement, so the AR(1) will have to do.) or, we can do both!

Quick aside: what do Mt. K data look like? $(x = altitude, km)$

id	RecNum	day	AM_PM	Oxygen_Stats	sum_diamox	x	diamox_ever
257	1	1	A	96	14	1.3	1
257	3	1	P	95	14	2.65	1
257	4	2	A	93	14	2.65	1
257	5	2	P	90	14	3.61	1
257	6	3	A	93	14	3.61	1
257	7	3	P	95	14	4.2	1
257	8	4	A	92	14	4.2	1
257	9	4	P	90	14	4.6	1
257	10	5	A	90	14	4.6	1
818	1	1	A	96	15	1.3	1
818	3	1	P	90	15	2.65	1
818	4	2	A	91	15	2.65	1
818	5	2	P	83	15	3.61	1
818	6	3	A	85	15	3.61	1
818	7	3	P	79	15	4.2	1
818	8	4	A	80	15	4.2	1
818	9	4	P	79	15	3.95	1

Modeling approaches for Mt. K data:

- Approach 1: random + simple $\boldsymbol{R}$ (i.e., $R = \sigma^2 \boldsymbol{I}$).
  - We did this already (see last slide set).
  - AIC=69599.3
    - Correlation parameter estimate is ~ 0.08 (estimated correlation between two errors 0.5 day apart, not responses).
- Approach 2: random + AR(1) structure for R.
- AIC=69545.2 (54 point drop). Both $\boldsymbol{G}$ and $\boldsymbol{R}$ contribute to the covariance structure: $\boldsymbol{V} = Var[\boldsymbol{Y}] = \boldsymbol{ZGZ}^\top + \boldsymbol{R}$.

```
proc mixed data=alldata; class id recnum;
model oxygen_sat= x x*x diamox_ever x*diamox_ever x*x*diamox_ever
      / outpm=outypm outp=outyp solution;
random intercept x x*x / subject=id v solution g type=un;
repeated recnum / type=ar(1) subject=id; run;
```

Some other possibilities:

- Approach 3: Remove RANDOM statement so that $\boldsymbol{V} = \boldsymbol{R}$.
  - The estimated correlation parameter (which now does represent the correlation between responses 0.5 day apart) in the structure increases to ~ 0.44.
  - This is because the random effects no longer contribute to the covariance between 2 responses, I.e., to get roughly the same covariance between 2 responses, the contribution from $\boldsymbol{R}$ needs to increase since there is no longer a contribution from $\boldsymbol{ZGZ}^\top$.
  - AIC increases A LOT.
- Approach 4: Up to linear random effects, simple R.
  - AIC also high
- Approach 5: Quadratic random effects plus Kronecker Product structure for errors.
  - Really complicated model! But AIC good.
  - Model makes intuitive sense.

AIC values for different covariance structure approaches.

Approach	# cov. parms*	AIC	Change
(1) Random effects only, up to quadratic, UN structure for G	6	69599.3	
(2) Up to quadratic random effects, UN structure for G; AR(1) on repeated measures over time (recnum)	7	69545.2	-54.1
(3) AR(1) for time; no random effects	2	71804.6	2205.3
(4) Random effects only, up to linear, UN structure for G	4	70117.4	518.1
(5) Up to quad random effects (as in 2), Kronecker Prod structure for errors**	9	**69382.5**	**-216.8**

*Does not include covariance parameters of '0'.

**See separate file for best model fit.

In comparing AIC, make sure same number of records used! For these fits, $n = 13,368$ used (1 missing value due to loss of info in am_pm and day variables)

For applications in which the random effects are defined on time rather some other variable (such as altitude, above), including a non-simple structure for time via $\boldsymbol{R}$ may still improve the model fit.

- For example, an outcome for which there is substantial between-subject heterogeneity (not accounted for in the predictors), but with repeated measures over time might require a random intercept plus an AR(1) structure for $\boldsymbol{R}$.
- Generally, it is recommended to first narrow the list of possible covariance structures, followed by a comparison of goodness-of-fit values for these possibilities.

Fitting joint normal outcomes using mixed models

- Let's say we have 2 outcomes with different units, and measurements over time for each of these. Can we use mixed models in this case? Sure, but proceed with caution!
- For simplicity, consider the COPDGene data, which has 2 measurements taken on subjects with COPD (GOLD groups 2 through 4) that are about 5 years apart. The 2 outcomes to be considered are FEV1 (a pulmonary function measure) and distance walked (an exercise ability measure).
- The original data set has FEV1 and distance walked in separate columns. In order to fit the model, we need to create a new, composite variable, call it y, and then create a variable, call it type, to identify when y is FEV1 and when y is distance walked (dist for short)

ID	y	visit	type	other var's…
1	1.8	1	FEV1	…
1	1.75	2	FEV1	…
1	1150	1	dist	…
1	1046	2	dist	…

- FEV1 is measured in liters, and is usally around 1 to 3, while distance walked is in feet, for a 6-minte time frame and is usually over 1000.

The SAS code to fit the model:

```
proc sort data=toget; by phase1_gold sid type visitnum; run;
proc mixed data=toget; where phase1_gold>1;
class sid type visitnum gender race smoking_status;
model y = type visitnum age_enroll age_enroll*age_enroll
      height_cm height_cm*height_cm gender race
      smoking_status
      type*visitnum
      type*age_enroll   type*age_enroll*age_enroll
      type*height_cm   /*type*height_cm*height_cm*/
      /type*gender*/
      type*race
      type*smoking_status
/ cl solution;
repeated type visitnum / type=un@un subject=sid r rcorr;
lsmeans type*visitnum; run;
```

- Ideally the list of covariates (i.e., terms other than type or visitnum) should include variables that are important to either FEV1 or distance walked. Also, I started with a model that included all possible interactions between type and other covariates; I later dropped $type \times height^2$ and $type \times gender$ due to very weak significance. The $visitnum \times type$ term is important to keep in the model, regardless of significance.
- It is important to remember that if we drop a particular $type \times covariate$ interaction, then we're assuming the relationship between that covariate and FEV1 can share the same slope as for the covariate and distance walked. When the units between the outcome types differ, this may be a tall order, and is why I start with a model that includes all possible interactions.

Selected output:

Subjects	569
Number of Observations Used	2263

Note that the UN(1,1) parameter for visitnum is set to 1 due to identifiability issues. The consequence of this is that the correlation between time points is constrained to be the same for both outcome types (highlighted to left).

Estimated R Matrix for sid 1

Row	Col1	Col2	Col3	Col4
1	141718	107520	44.8643	34.0379
2	107520	174629	34.0379	55.2829
3	44.8643	34.0379	0.1783	0.1353
4	34.0379	55.2829	0.1353	0.2197

Estimated R Correlation Matrix for sid 1

Row	Col1	Col2	Col3	Col4
1	1.0000	0.6835	0.2822	0.1929
2	0.6835	1.0000	0.1929	0.2822
3	0.2822	0.1929	1.0000	0.6835
4	0.1929	0.2822	0.6835	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
type UN(1,1)	sid	141718
UN(2,1)	sid	44.8643
UN(2,2)	sid	0.1783
visitnum UN(1,1)	sid	1.0000
UN(2,1)	sid	0.7587
UN(2,2)	sid	1.2322

Least Squares Means

Effect	type	visitnum	Estimate	Standard Error	DF	t Value	Pr > |t|
type*visitnum	dist	1	1197.12	22.7460	555	52.63	<.0001
type*visitnum	dist	2	1064.14	21.0384	555	50.58	<.0001
type*visitnum	fev1	1	1.4796	0.02554	555	57.93	<.0001
type*visitnum	fev1	2	1.3560	0.02362	555	57.40	<.0001

- If we fit FEV1 and distance walked in separate model, the correlation between time points for FEV1 is 0.82 and for dist is 0.55; this explains the 'pooled' correlation of 0.68 when modeling the outcomes jointly.
- Least-squares means estimates and SE's are similar when running individually as when running jointly. The joint model has slightly higher SE's for distance walked and slightly lower SE's for FEV1 due to model constraints.

Joint model approach				FEV1 alone			Distance walked alone		
Time	**type**	**Estimate**	**SE**	**Time**	**Estimate**	**SE**	**Time**	**Estimate**	**SE**
V1	**dist**	1197.12	22.7460	**V1**	1.4776	0.02905	**V1**	1196.93	20.1309
V2	**dist**	1064.14	21.0384	**V2**	1.3534	0.02594	**V2**	1060.99	19.7650
V1	**fev1**	1.4796	0.02554						
V2	**fev1**	1.3560	0.02362						

Notes

- The major advantage of running models with FEV1 and distance walked together is to better understand correlation between them, while simultaneously conducting inference for effects of interest, which is how the outcomes change over time.
- If the two outcomes being measured actually have the same units, then it may simplify the model, with respect to both the necessary predictors (including interaction terms) and the suitable covariance structure.
- For outcomes that are measured in different units, an alternative to modeling them in raw units is to standardize the outcomes in some fashion, which might simplify the necessary predictors and interaction terms. For example, FEV1 is sometimes put into 'percent of predicted' terms based on age, height gender and race (in some cases using squared terms for continuous variables). If a similar approach is used for distance walked, then the list of predictors reduces to type and visit (and their interaction).
- We could try the UN structure for the complete list of repeated measures. This would introduce 10 covariance parameters; a very flexible matrix but harder to fit and possibly introduces more parameters than necessary. In the case of the data above, it did not converge.

Summary