FES 524: Natural Resources Data Analysis

Reading 6.2: Correlation structures

Contents

1 Correlation matrices

This reading delves a little deeper into *correlation structures*. Like with statistical models, the mathematical representation of correlations can be an efficient way to describe and understand correlation structures. You will only be asked to describe any correlations on assignments in words, not with mathematical notation, but the goal of introducing correlations as symbols is to help in your overall understanding of correlations.

A correlation matrix is one way to describe the within-subject correlations among errors. A correlation matrix is a square matrix, meaning there are as many rows as columns. Each element of the matrix stores the pairwise correlation between the errors of the repeated measures. For example, the value in the first row and third column stores the correlation between ϵ_{i1} and ϵ_{i3}, the errors for the first and third repeated measurements of subject i. Correlations do not depend on order, so the element in the first row and third column will be equal to the element of the third row and first column. We call a matrix like this *symmetric*, since the values are mirrored over the diagonal. We will denote the correlation matrix for the within-subject errors as $\mathbf{\Omega}_i$ and the covariance matrix $\mathbf{\Sigma}_i$, and individual elements as $\mathbf{\Omega}_{i,jk}$ for the element in the j^{th} row and k^{th} column of $\mathbf{\Omega}_i$. When we eventually discuss the full correlation and covariance matrix of all the errors collectively, we will drop the subscript i.

1.1 No correlation

In the absence of correlation between errors of repeated measures of subject i, we have

$$\text{corr}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Omega}_i = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I}_m$$

where $\mathbf{I}_m$ stands for the *identity matrix* with m rows and m columns, where m is the number of repeated measurements for the subject. Without correlation, the correlation matrix for the errors within a subject is the identity matrix since all random variables have a correlation of 1 with themselves, but zero correlation among different errors.

1.2 A matrix with correlation

With correlation among errors within a subject, we might have something like

$$\text{corr}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Omega}_i = \begin{bmatrix} 1 & 0.2 & -0.1 \\ 0.2 & 1 & 0.5 \\ -0.1 & 0.5 & 1 \end{bmatrix}$$

for a case when $m = 3$. Again, we have 1's along the diagonal since a random variable is always perfectly correlated with itself, but some non-zero correlations in the *off-diagonals*. This example also makes it easier to see the symmetry of a correlation matrix. There is a positive correlation between ϵ_{i1} and ϵ_{i2} equal to 0.2, which we can see in the elements $\boldsymbol{\Omega}_{i,12}$ and $\boldsymbol{\Omega}_{i,21}$.

1.3 Symbolic representation

It is common to use ρ_{jk} to denote the correlation between variables j and k (the errors in this case). Thus, a symbolic representation you will see is

$$\text{corr}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Omega}_i = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1m} \\ \rho_{12} & 1 & \dots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1m} & \rho_{2m} & \dots & 1 \end{bmatrix}. \tag{1}$$

Notice that we don't switch the order of the subscripts since the correlation in row j and column k has to be equal to the correlation in row k and column j. This makes the symmetry clearer when we switch to a symbolic representation of the matrix. Similarly, if all errors within a subject have the same correlation, we would make this clear by using the notation

$$\text{corr}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Omega}_i = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}. \tag{2}$$

Here, we dropped the subscripts to make it clear that all the off-diagonal elements are the same.

2 The covariance matrix

The covariance is the "scaled" version of the correlation matrix. It contains variances and covariances, and is directly related to the correlation matrix. Specifically,

$$\mathbf{\Sigma} = \mathbf{D}\Omega\mathbf{D}^\top = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \dots & \sigma_1\sigma_m\rho_{1m} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & \dots & \sigma_2\sigma_m\rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_m\rho_{1m} & \sigma_2\sigma_m\rho_{2m} & \dots & \sigma_m^2 \end{bmatrix}$$

where

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m \end{bmatrix}$$

is a diagonal matrix of the standard deviations of each of the random variables. Most often, we assume equal variances among repeated measurements (and therefore equal standard deviations), so $\mathbf{D} = \sigma\mathbf{I}_m$, where σ is the common standard deviation. The standard deviation(s) therefore take correlations, which are on a standard scale between -1 and 1, and scale them to be on the scale of variation in the errors.

3 Correlation structures

There are an infinite number of correlation structures we could dream up, but below are a handful of some commonly used structures for repeated measures.

3.1 Variance components

As discussed in previous labs and class, there is a specific form of covariance / correlation that is implied or *induced* when we use random effects in a mixed model. However, once we account for the fixed and random effects, we assume the model errors are independent. Another way to say "accounting for" is *conditioning on*. The errors are independent *conditional* on the variables included in the model. We can therefore refer to these errors as conditional errors. The conditional errors from a linear mixed model are assumed to be independent.

As we have discussed, we believe that observations taken from the same physical unit are correlated with each other. This is one of the reasons we have been fitting mixed models. Such correlations are sometimes called *marginal* correlations; these are the correlations among errors without conditioning on the random effects. The marginal errors are the errors after taking only the fixed effects into account.

The marginal errors from a linear mixed model are correlated. There is an overall within-subject correlation of errors for all pairs of the repeated measurements. This makes sense, since we believe there is some systematic effect of the variable we are using as a random effect on the response variable.

These within-subject marginal correlations from a linear mixed model must be strictly positive. We cannot model negative correlation among errors using a mixed model. This is usually fine since we expect errors to be more alike within groups than between groups (i.e., positive correlation).

The correlation matrix that represents the marginal correlation induced by the mixed model is shown in 2 where ρ is constrained to be positive.

3.2 Compound symmetry

All the remaining correlation structures we will talk about are correlations of the conditional errors, the leftover correlation after we take into account both the fixed and random effects from a mixed model. Compound symmetry is the simplest kind of correlation structure. When using a compound symmetry structure we calculate a single correlation, which is shared among all within-subject repeated measurements.

This correlation from compound symmetry is rarely different than the correlation induced by a mixed model. The main instance where compound symmetry adds something more than what we got from having a random effect in the model is if the within-subject correlation is *negative*. Otherwise, we don't get much extra since the compound symmetry correlation matrix is identical to that of the variance components correlation structure (equation 2), but ρ can be negative or positive.

3.3 AR(1)

Autoregressive of order 1 correlation, more commonly known as AR(1), is frequently used for time series but can be used for some repeated measures in space. In an AR(1) structure, correlations systematically diminish as observations get farther apart in space or time. It is important to recognize that, when using an AR(1) correlation structure, we estimate only a single correlation. This correlation is used to describe how correlations diminish with increasing space or time between measurements.

The AR(1) correlation matrix is

$$\mathbf{\Omega}_i = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^m \\ \rho & 1 & \rho & \dots & \rho^{m-1} \\ \rho^2 & \rho & 1 & \dots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^m & \rho^{m-1} & \rho^{m-2} & \dots & 1 \end{bmatrix}.$$

Note that there is just one parameter, ρ, but that it is raised to higher and higher powers as we move away from the diagonal. Since $\rho \in (-1, 1)$, raising it to higher and higher powers results in smaller and smaller values.

3.4 Unstructured or general correlation

The general correlation structure, often called *unstructured* correlation, is the most complex correlation structure we have for package `nlme`.

The general correlation structure estimates a separate correlation for every pair of levels from the repeated measurement factor. This means we end up estimating a lot of correlations if we have a lot of repeated measurements. This structure can lead to overly complex models when we have relatively few data and many repeated measurements. Because of this, the general correlation structure is rarely advised for situations with many repeated measurements. In many cases we simply can't justify estimating many separate correlations given our sample size.

The general correlation structure has no pattern and is represented by Equation 1. Every pair of positions has a unique correlation, unrelated to the other correlations.

3.5 Correlations available in nlme

There are more correlation structures available in the `nlme` package than we will see in this class. Below is the complete list. See the documentation for more details, `?corStruct`.

- `corCompSymm()`: compound symmetry
- `corSymm()`: general
- `corAR1()`: autoregressive order 1
- `corARMA()`: autoregressive-moving average
- `corCAR1()`: continuous AR1
- `corExp()`: exponential spatial
- `corGaus()`: Gaussian spatial
- `corLin()`: linear spatial
- `corRatio()`: rational quadratic spatial
- `corSpher()`: spherical spatial