Autoregression in Mixed Models
Ed Stanek

Introduction

We develop an autoregression structure for the variance in the context of a simple example. The development illustrates the basic ideas that underlie first order autoregressive processes.

The Example

We take as an example data from the Seasons study of cholesterol. Weather data were collected over a three year period as part of the study. An example of part of these data is contained in the SAS data set (weat1.sd2). This data set has the daily temperature recorded for the first four days of each month for the study period. We sketch a model for these data.

Study Population and Parameters

We assume the study population consists of a large number of clusters (ie. Months) that are numbered from \( s = 1, ..., N \), where each cluster contains the days \( t = 1, ..., M \) (varying from 28 to 31). The index for days enumerates consecutive days in a month. The values of the index have meaning, since pairs of values, (i.e. \( t=2, t=4 \)) indicate that there is two days separating the values. The response variable that we consider is average temperature on the day. We represent the potentially observable temperature on day \( t \) in month \( s \) as a random variable

\[
Y_{stk} = \mu_s + E_{stkh}
\]

where

\[
E_{stkh} = \rho E_{s(t-1)k} + V_{stkh}.
\]

We assume that \( E(V_{stkh}) = 0 \) and \( \text{var}(V_{stkh}) = \sigma^2 \) for all \( s \) and \( t \), and that for any \( s \neq s^* \), \( t \neq t^* \), or \( k \neq k^* \), \( \text{cov}(V_{stkh}, V_{st^*k^*}) = 0 \). The index \( k \) represents a measure of average temperature on day \( t \) in month \( s \). This measure is one of a potentially infinite number of measures that could be taken. Note that the expression for \( E_{stkh} \) is recursive. Thus, we can express

\[
E_{stkh} = \rho E_{s(t-1)k} + V_{stkh} = \rho \left( \rho E_{s(t-2)k} + V_{s(t-1)k} \right) + V_{stkh} = \rho^2 E_{s(t-2)k} + \rho V_{s(t-1)k} + V_{stkh}.
\]

Also, expanding \( E_{s(t-2)k} \),
\[ E_{stk} = \rho^2 E_{s(t-2)k} + \rho V_{s(t-1)k} + V_{stk} \]
\[ = \rho^2 \left( \rho V_{s(t-3)k} + V_{s(t-2)k} \right) + \rho V_{s(t-1)k} + V_{stk} \]
\[ = \rho^3 E_{s(t-3)k} + \rho^2 V_{s(t-2)k} + \rho V_{s(t-1)k} + V_{stk} \]
or equivalently, \[ E_{stk} = \sum_{u=0}^{\infty} \rho^u V_{s(t-u)k} \]. Since \[ E(V_{stk}) = 0 \],
\[ E(E_{stk}) = \sum_{u=0}^{\infty} \rho^u E(V_{s(t-u)k}) = 0. \]
Also, since \[ \text{var}(V_{stk}) = \sigma^2 \] and the random variables \[ V_{stk} \] are independent and identically distributed,
\[ \text{var}(E_{stk}) = \sum_{u=0}^{\infty} \rho^{2u} \text{var}(V_{s(t-u)k}) \]
\[ = \sigma^2 \left( \sum_{u=0}^{\infty} (\rho^2)^u \right) \].
Now when \(|x| < 1\), based on a Taylor series expansion, \[ \sum_{u=0}^{\infty} x^u = \frac{1}{1-x} \]. As a result,
\[ \text{var}(E_{stk}) = \frac{\sigma^2}{1 - \rho^2} = \sigma^2. \]
This is a first order autoregressive structure with equal spacing. The equal spacing results from the equal time period between days.

**Developing the Covariance Between Days**

The first order autoregressive model is given by
\[ Y_{stk} = \mu_{stk} + E_{stk} \]
where \[ E_{stk} = \rho E_{s(t-1)k} + V_{stk} \] and \[ \text{var}(E_{stk}) = \frac{\sigma^2}{1 - \rho^2} = \sigma^2 \]. We develop an expression for the covariance between days, first considering the covariance between adjacent days,
\[ \text{cov}(Y_{stk}, Y_{s(t+1)k}) = \text{cov}(E_{stk}, E_{s(t+1)k}) \]. Since \[ E_{s(t+1)k} = \rho E_{stk} + V_{s(t+1)k} \],
\[ \text{cov}(Y_{stk}, Y_{s(t+1)k}) = \text{cov}(E_{stk}, \rho E_{stk} + V_{s(t+1)k}) \]
\[ = \rho \text{var}(E_{stk}) + \text{cov}(E_{stk}, V_{s(t+1)k}) \]
\[ = \rho \sigma^2 \].
Similarly, we can show that
\[ \text{cov}(Y_{stk}, Y_{(t+2)k}) = \text{cov}(E_{stk}, \rho^2 E_{stk} + \rho V_{s(t+1)k} + V_{s(t+2)k}) \]
\[ = \rho^2 \text{var}(E_{stk}) + \text{cov}(E_{stk}, \rho V_{s(t+1)k} + V_{s(t+2)k}), \]
\[ = \rho^2 \sigma^2 \]

or more generally that \( \text{cov}(Y_{stk}, Y_{(t+s)k}) = \rho^s \sigma^2 \).

We use these results to summarize the first order autoregressive model for \( m \) consecutive days in a month. Let \( \mathbf{Y}_s = (Y_{s1k}, Y_{s2k}, \ldots, Y_{smk})' \) and \( \mathbf{E}_s = (E_{s1k}, E_{s2k}, \ldots, E_{smk})' \). Then the model is given by
\[ \mathbf{Y}_s = \mathbf{I}_m \mu_s + \mathbf{E}_s \]
where \( \mathbf{E}(\mathbf{Y}_s) = \mathbf{I}_m \mu_s \), and
\[ \text{var}(\mathbf{Y}_s) = \sigma^2 \begin{pmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{m-1} \\
\rho & 1 & \rho & \ldots & \rho^{m-2} \\
\rho^2 & \rho & 1 & \ldots & \rho^{m-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \ldots & 1
\end{pmatrix}. \]

In SAS Proc Mixed, the first order autoregressive structure is specified by TYPE=AR(1).

**Extensions of the Autoregressive model.**

One extension of an autoregressive model allows for different variances for different days. Such a model can be fit using PROC Mixed in SAS using the TYPE=ARH(1) structure. We consider this variance structure in more detail. We might conjecture that the basic underlying model for this structure is given by
\[ Y_{stk} = \mu_s + E_{stk} \]
where
\[ E_{stk} = \rho E_{s(t-1)k} + V_{stk}. \]

We assume that \( E(V_{stk}) = 0 \) and \( \text{var}(V_{stk}) = \sigma_i^2 \) for all \( s \), and that for any \( s \neq s^* \), \( t \neq t^* \), or \( k \neq k^* \), \( \text{cov}(V_{stk}, V_{s^*t^*k^*}) = 0 \). Assuming this structure, we develop expressions for the variance. By analogy, we note that at any day, \( t \), \( \text{var}(E_{stk}) = \frac{\sigma_i^2}{1-\rho^2} = \sigma_i^2 \). We develop an expression for
the covariance between days, first considering the covariance between adjacent days,
\[
\text{cov}(Y_{stk}, Y_{s(t+1)k}) = \text{cov}(E_{stk}, E_{s(t+1)k}).
\]
Since \( E_{s(t+1)k} = \rho E_{stk} + V_{s(t+1)k} \),
\[
\text{cov}(Y_{stk}, Y_{s(t+1)k}) = \text{cov}(E_{stk}, \rho E_{stk} + V_{s(t+1)k})
\]
\[
= \rho \text{var}(E_{stk}) + \text{cov}(E_{stk}, V_{s(t+1)k}).
\]
\[
= \rho \sigma_t^2
\]
Similarly, \( \text{cov}(Y_{stk}, Y_{s(t+2)k}) = \rho^2 \sigma_t^2 \). However, note that
\[
\text{cov}(Y_{s(t+1)k}, Y_{s(t+2)k}) = \text{cov}(E_{s(t+1)k}, \rho E_{s(t+1)k} + V_{s(t+2)k})
\]
\[
= \rho \text{var}(E_{s(t+1)k}) + \text{cov}(E_{s(t+1)k}, V_{s(t+2)k}).
\]
\[
= \rho \sigma_{t+1}^2
\]
As a result, this heterogeneity of variance assumption will result in the following variance matrix for equally spaced days:
\[
\text{var}(Y_i) = \sigma_1^2 \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1^2 & \rho^2 \sigma_1^2 & \cdots & \rho^{m-1} \sigma_1^2 \\
\rho \sigma_1^2 & \sigma_2^2 & \rho \sigma_2^2 & \cdots & \rho^{m-2} \sigma_2^2 \\
\rho^2 \sigma_1^2 & \rho \sigma_2^2 & \sigma_3^2 & \cdots & \rho^{m-3} \sigma_3^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{m-1} \sigma_1^2 & \rho^{m-2} \sigma_2^2 & \rho^{m-3} \sigma_3^2 & \cdots & \sigma_m^2
\end{pmatrix}.
\]
Note that this structure does not match the variance structure resulting from using PROC Mixed in SAS using the TYPE=ARH(1) structure. \{At this point, I’m not sure what model would result in this variance structure.\}

The autoregressive model can be extended in several other ways. The first order autoregressive model can be extended to unequal spacing of days. For example, suppose that measures were made on days 1, 2, 5, and 6. Then
\[
\text{var}(Y_{s(1:k)}) = \sigma^2 \begin{pmatrix}
1 & \rho & \rho^4 & \rho^5 \\
\rho & 1 & \rho^3 & \rho^4 \\
\rho^4 & \rho^3 & 1 & \rho \\
\rho^5 & \rho^4 & \rho & 1
\end{pmatrix}.
\]
In a REPEATED statement in PROC MIXED, the “repeated effect”, when sorted within subjects, can represent unequally spaced measures.

Another extension is to consider second (or higher) order auto-regressive models. Such models have error structures of the form,
\[
E_{stk} = \rho E_{s(t-1)k} + V_{stk}
\]
for a second order model, or for a \( \rho^p \) order auto-regressive model, the error has the form,
\[
E_{stk} = \rho^p E_{s(t-p)k} + V_{stk}.
\]

Other models combine an autoregressive and moving average structure. We do not discuss these models, but note that SAS Proc Mixed has an option to fit such models.
Adding Random Effects to an Autoregressive model

We add random effects to an autoregressive model in a simple setting where we have a population of clusters (i.e., Months) that are numbered from \( s = 1, \ldots, N \), where each cluster contains the days \( t = 1, \ldots, M \) (varying from 28 to 31). From each month, we assume that we have measures of average temperature on the first four consecutive days. We summarize the autoregressive model on month \( s \) as follows:

\[
Y_s = \mathbf{1}_4 \mu_s + \mathbf{E}_s
\]

where \( E(Y_s) = \mathbf{1}_4 \mu_s \), and \( \text{var}(Y_s) = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \).

Notice that we have not defined \( \mu_s \) apart from the definition of \( E(Y_{ss}) = \mu_s \). This implies that for each day \( t \) in month \( s \), the true mean temperature is the same. This assumption corresponds to the ‘stationary’ assumption in autoregressive models.

We describe parameters for the population of \( s = 1, \ldots, N \) clusters next. We define the mean and variance as

\[
\mu = \frac{1}{N} \sum_{s=1}^{N} \mu_s \quad \text{and} \quad \left( \frac{N-1}{N} \right) \sigma^2 = \left( \frac{N-1}{N} \right) \left( \frac{\sum_{s=1}^{N} (\mu_s - \mu)^2}{N-1} \right).
\]

We use these definitions of parameters to define a

\[
\mu_s = \mu + (\mu_s - \mu) = \mu + \beta_s
\]

where \( \beta_s = (\mu_s - \mu) \) represent the deviation of a cluster mean from the overall mean.
Selection of Subjects in the Study

The study is conducted by randomly selecting \( i = 1, \ldots, n \) clusters (subjects), and within each selected cluster, observing the temperature on days \( t = 1, \ldots, m \). The days are not assumed to be a random sample of days, but rather are consecutive days. The index \( i \) represents the position of a cluster in the sample. Since the cluster assigned to position \( i \) in the sample is considered to be random, we represent response as the random variable

\[
Y_{itk} = \text{response for the } k^{th} \text{ measure at time } t \text{ in the } i^{th} \text{ selected cluster.}
\]

The response can also be represented via the model:

\[
Y_{itk} = \mu + B_i + E_{itk}
\]

where \( B_i \) is a random variable that represents the difference between the \( i^{th} \) selected cluster from the population mean, and \( E_{itk} \) is a random variable that represents the difference between the \( k^{th} \) measure at time \( t \) in the \( i^{th} \) selected cluster and the \( i^{th} \) selected cluster’s mean.

Vector and Matrix Representation of the Population and Model

The study consists of selecting a simple random sample of clusters and observing response at times corresponding to a consecutive set of times. Although only a subset of the population will be selected in the study, the entire population can be represented as a vector of random variables. We use this representation, and subsequently describe the subset that will correspond to the realized sample.

Simple random sampling without replacement can be represented using a vector of random variables. The random variables are defined such that they can take on the values corresponding to any permutations of the values in the population. Each potential realization of the random variables is equally likely. With this representation, a sample of size \( n \) can be represented as the first \( n \) random variables in vector representing the permutation of population values.

We extend the fixed effect autocorrelation model to a mixed model such that

\[
Y_i = X_i \mu + Z_i B + E_i
\]

or for the sample of size \( n \), as

\[
Y = X \mu + Z B + E.
\]

With four measures on each cluster, \( Y_i = (Y_{i1k} \ Y_{i2k} \ Y_{i3k} \ Y_{i4k})' \), \( X_i = 1_n \), \( Z = 1_n \otimes 1_4 \), and \( E_i = (E_{i1k} \ E_{i2k} \ E_{i3k} \ E_{i4k})' \). Taking the expected value over all possible selections of
clusters, $E(Y) = X\mu$, and $\text{var}(Y) = Z\text{var}(B)Z' + \text{var}(E)$. This variance matrix is block diagonal (assuming that $n$ is small relative to the population size, $N$), and is given by

$$\text{var}(Y_i) = \sigma^2 J + \sigma^2_\rho \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$  

This variance matrix has both compound symmetry and auto-correlation.