Abstract

In longitudinal studies, the change in the response variable over time is often of interest. Linear growth curve analysis allows you to estimate the mean slope as a function of time. Analysis becomes more difficult when there is individual variation about the mean slope and repeated observations are not independent.

The MIXED procedure in the SAS/STAT software can model random effects which allows for individual variation. This procedure allows various models for the random effects component as well as for the error component. The method of generalized estimating equations (GEE) provides a robust covariance matrix even when the correlation structure of repeated observations is misspecified. A SAS macro is available that applies GEE to longitudinal data analysis. The macro does not explicitly model random effects, but provides valid results when data actually contain random effects.

This paper discusses how you can use either the MIXED procedure or the GEE macro to perform linear growth curve analysis. Various models were fit to simulated data to validate the correct model and examine effects of model misspecification.

Introduction

In longitudinal studies, repeated observations of a certain outcome are taken over the study period on the same unit. A special case of longitudinal data analysis is growth curve analysis where you examine the change in the outcome as a function of time. In the special case of linear growth curve analysis, you can estimate slopes that represent the assumed linear change in time. A test of non-zero slope or comparison of slopes among treatment groups is commonly done.

In linear growth curve analysis it is important to consider whether to model the intercept and slope parameters as fixed or random effects. Random effects allow for individual variation from group means. Fixed effects are used to model a population and assumes negligible variance within the population. In either case, observations within a subject may still be correlated after accounting for subject means. In a randomized clinical trial where treatment group i (i=1,...,I) has j subjects (j=1,...,ni) with observations at time tk (k=1,...,nij), treatment group and time are fixed effects, while subjects may be random.

In the simple linear model with only fixed effects,

\[ Y = X \beta + \epsilon \]

where Y is a univariate vector of observations, X a model matrix, \( \beta \) a vector of intercept \( \beta_{0i} \) and slope \( \beta_{1i} \) parameters for each of the I groups and \( \epsilon \) an error vector distributed \( \mathcal{N}(0,R) \). Assuming the number of observations per subject is a constant K, the observation vector for subject j in treatment group i from the simple linear model is:

\[ Y_{ij} = \beta_{0i} + \beta_{1i} * t + \epsilon_{ij} \]

where \( Y_{ij} \) is the Kx1 observation vector \([Y_{ij1}, Y_{ij2},...,Y_{ijK}]\), \( t \) is the Kx1 time vector \([t_1, t_2, ..., t_K] \) and \( \epsilon_{ij} \) is the Kx1 error vector \([\epsilon_{ij1}, \epsilon_{ij2}, ..., \epsilon_{ijk}]\).

Since the intercept and slope parameters are fixed,

\[ V_{ij} = \text{Var}(Y_{ij}) = \text{Var}(\epsilon_{ij}) = R_{ij} \]

is the KxK variance matrix for subject j in the simple linear model. The variance structure may vary by group but subjects are assumed to be independent and have the same variance structure within a group. In most analyses, repeated observations are assumed to be independent such that \( R_{ij} = \sigma^2 I \) where I is the identity matrix. The assumption of a fixed effects model and independent repeated observations could be costly and lead to invalid results in terms of parameter estimation and hypothesis tests when these assumptions do not hold true.

Random Effects Model

Laird and Ware (1982) discuss two-stage random effects models in longitudinal data with growth models as a special case. In the first stage,
population parameters, subject effects and within-subject variation are modeled. The mixed general linear models approach, or simply mixed models, allows for specification of both fixed and random effects. The mixed model has the following form:

\[ Y = X\beta + Zv + \varepsilon \]

where \( Y \), \( X \), \( \beta \) and \( \varepsilon \) as before, \( Z \) is a model matrix for random effects and \( v \) is a vector of random effects. We assume both the intercept and slope are random effects. In this case, \( Z \) is the same as \( X \) and \( v \) contains \( v_{0i} (= b_{0i} \beta_{0}) \) and \( v_{1i} (= b_{1i} \beta_{1}) \) to model subject \( j \)'s individual variation from the mean intercept and slope for group \( i \), respectively.

In the second stage, the distribution describing the between-subject variation is given. If \( r \) is the number of random effects parameters in the model, then \( v \) is distributed \( N(0, G) \) where \( G \) is an \( rxr \) variance matrix. The corresponding variance matrix for subject \( j \) in the mixed model is

\[ V_{ij} = \text{Var}(Y_{ij}) = \text{Var}(Z_{ij}v_{ij}) + \text{Var}(\varepsilon_{ij}) 
= Z_{ij}\text{Var}(v_{ij})Z_{ij}^{\prime} = Z_{ij}GZ_{ij}^{\prime} + R_{ij} \]

where \( G \) estimates the variability due to a subject's departure from the group mean and \( R_{ij} \) estimates the within-subject variability. The random effects model, regardless of the structure of \( R_{ij} \), imposes a covariance structure with correlated observations within subjects. In a fixed effects model, \( V_{ij} \) reduces to \( R_{ij} \) since \( Z = 0 \) and \( G = 0 \).

**Fitting a Mixed Model in SAS**

Let the data set `simdat` contain the variables `outcome` for the outcome variable, `trt` for group ('C' for control or 'T' for treatment), `indiv` for subject number and `time` for time period. The following MIXED procedure code using SAS version 6.08 estimates the intercept and slope for each group allowing random effects for both parameters:

```sas
proc mixed data=simdat;
  class trt indiv;
  model outcome=trt trt*time / noint s;
  random int time / type=un(1) subject=indiv 
    group=trt g;
  run;
```

The 'noint' and 's' options in the MODEL statement give explicit parameter results. The `int` variable in the RANDOM statement is interpreted as containing the value 1 for all observations. The 'subject=' and 'group=' options in the RANDOM statement specifies that variation occurs at the `indiv` variable level about the mean intercept and slope for each level of the `trt` variable. The 'g' option in the RANDOM statement allows you to examine the resulting \( G \) matrix. In the above code, the structure of \( G \) is banded main diagonal and \( R \) is simple as explained below. The estimated mean squared error, \( \sigma^{2} \), is shown in the Estimate column for the Residual parameter under the Covariance Parameter Estimates section of the output.

**Specification of the R and G matrices**

The structure of both the \( G \) and \( R \) matrices can be specified using the 'type=' option in the RANDOM and REPEATED statements of the MIXED procedure, respectively. In the second stage of the random effects model, assume all random effects parameters are independent. This means we have the following \( G \) matrix structure:

\[
G = \begin{bmatrix}
\sigma_1 & 0 & 0 & 0 \\
0 & \sigma_2 & 0 & 0 \\
0 & 0 & \sigma_3 & 0 \\
0 & 0 & 0 & \sigma_4
\end{bmatrix}
\]

where \( \sigma_1 \) and \( \sigma_3 \) denote the variances of the random intercept effect and \( \sigma_2 \) and \( \sigma_4 \) denote the variances of the random slope effect for the control and treatment groups, respectively.

There are two matrix structures which can be specified to Impose independent random effects. The simple structure, which is the default, is assumed when the 'type=' option is missing or explicitly specified as 'type=sim' in the RANDOM statement. The simple structure restricts \( \sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 \) so that the variances of the random intercepts and slopes are all equal. This would rarely be appropriate.

You can specify the less restrictive banded main diagonal structure with the 'type=un(1)' option as shown in the example code. This allows \( \sigma_1 \neq \sigma_2 \neq \sigma_3 \neq \sigma_4 \) so that there are no constraints on the variances of the random effects. Unfortunately there is no 'type=' option that constrains \( \sigma_1^2 = \sigma_3^2 \) and \( \sigma_2^2 = \sigma_4^2 \) so that the random intercepts in the two groups would have equal variance and the random slopes would have equal variance. The 'group=' option adds some flexibility.
and partitions the G matrix into submatrices with the same structure but different variance parameters.

The common matrix structures for $R_{ij}$ are simple and compound symmetry. The structure is as follows:

$$
R_{ij} = \sigma^2 \Sigma_{ij} = \sigma^2 \begin{bmatrix}
1 & \rho & \ldots & \rho \\
\rho & 1 & \ldots & \rho \\
\ldots & \ldots & \ldots & \ldots \\
\rho & \ldots & \rho & 1 \\
\rho & \ldots & \rho & 1
\end{bmatrix}
$$

where $\rho = 0$ if simple and $\rho \neq 0$ if compound symmetry. The simple structure is the default and assumed when the REPEATED statement is missing or the REPEATED statement is included with a ‘type = ’ option missing or explicitly specified as ‘type = sim’. You can specify a compound symmetry structure for $R_{ij}$ with the following REPEATED statement:

```
repeated / type = cs subject = indiv r;
```

The ‘subject = ’ option specifies that the repeated observations are at the indiv variable level. The ‘r’ option allows you to examine the resulting $R_{ij}$ matrix. The resulting R matrix is block-diagonal since subjects are assumed to be independent. The ‘group = ’ option is also available with the REPEATED statement.

Depending on the specification of the G and R matrix structures, the model may not converge due to overparameterization of the resulting V matrix. This is the case with a banded main diagonal G matrix and a compound symmetry R matrix when both the intercepts and slopes are random.

**Generalized Estimating Equations**

Correlated observations have become easier to deal with since the advent of generalized estimating equations (GEE) by Liang and Zeger (1988) and Zeger and Liang (1988). They have shown that GEE provides robust mean estimates and consistent variance estimates even when the correlation between observations is not correctly specified, but the mean structure is correctly specified. GEE is an extension of the quasi-likelihood approach in the generalized linear models framework to longitudinal data with multiple observations per subject.

The generalized linear model introduced by Nelder and Wedderburn (1972) and McCullagh and Nelder (1983) extended the simple linear model in that some transformation of the marginal expectation of $Y$ is a linear function of the covariates. It is assumed that the observations in $Y$ come from a distribution in the exponential family which have the following form:

$$
f_Y(y; \theta; \phi) = \exp((y \beta - b(\beta))a(\phi) + c(y, \phi))
$$

for some specific functions $a(\phi)$, $b(\theta)$ and $c(y, \phi)$. We have the observations in $Y$ having marginal expectation $E(Y) = \mu$ and constant variance $\sigma^2$. A linear predictor, denoted by $\eta$, is given by $\eta = X\beta$. The generalization is in $\eta = g(\mu)$ where $g(\mu)$ is referred to as the “link” function.

The form of the distribution of observations is needed to define a likelihood, but Wedderburn (1974) showed that in quasi-likelihood theory only the relationship between the mean and variance of the observations is needed to specify a likelihood function for estimation. In quasi-likelihood, $\text{Var}(Y) = V = h(\mu)/\delta$ where $\delta$ is a scale parameter and treated as a nuisance parameter. Ignoring the group subscript and letting $Y_j$ denote the observation vector for subject $j = 1, 2, \ldots, J$, $\mu_j = E(Y_j)$ and $V_j = \text{Cov}(Y_j)$, the quasi-likelihood estimator of the parameters in $\beta$ is the solution to the following score-like equations:

$$
U(\beta) = \sum_{j=1}^{J} \frac{\partial \mu_j}{\partial \beta} V_j^{-1} (Y_j - \mu_j)
$$

where

$$
V_j (\alpha) = A_j^{1/2} R_j(\alpha) A_j^{1/2}/\phi.
$$

The matrix $A_j$ is a diagonal matrix with $g(\mu_j)$ on the diagonal and $\phi$ is a nuisance parameter. $R_j(\alpha)$ is called the “working” correlation matrix since the correlation between observations is not expected to be correctly specified and is fully specified by a $sx1$ vector of parameters, $\alpha$.

**GEE in SAS**

To use the GEE macro, written by Karim and Zeger (1988) using the SAS/IML® software, you must specify a “link” function (LINK option), mean-variance relation (VARI option) and “working” correlation matrix (CORR, M and R options). As outlined by Zeger, Liang and Albert (1988), GEE can be applied to the mixed model with subjects
specified as random effects, which they refer to as the subject-specific model. Unfortunately, the GEE macro does not currently allow for specification of random effects and assumes all effects are fixed.

The GEE macro also does not contain a CLASS statement as in the MIXED procedure so you must explicitly create variables for each group. Let \( \text{trt1} \) be a variable equal to 1 if observations are from the control group; 0 if not. Similarly, let \( \text{trt2} \) equal 1 if observations are from the treatment group; 0 if not. The \( \text{time1} \) and \( \text{time2} \) variables are equal to the element by element multiplication of \( \text{trt1} \) by \( \text{time} \) and \( \text{trt2} \) by \( \text{time} \), respectively. You can use the following code to estimate the intercept and slope parameter for each group based on GEE:

```sas
%INCLUDE 'c:\directory\geemacro.sas';
%GEE (DATA=simdat,
  YVAR=outcome,
  XVAR=trt1 trt2 time1 time2,
  ID=indiv, /* Clustered by individual */
  LINK= 1, /* Identity link */
  VARI = 1, /* Gaussian error */
  CORR = 1, /* Simple */
);
```

The 'ID=' statement, which is similar to the 'subject=' option in the REPEATED statement of the MIXED procedure, specifies the variable level at which observations are repeated. Assuming a normally distributed continuous outcome, we use the identity link where \( g(\mu) = \mu \). The mean-variance relation is Gaussian (\( h(\mu) = 1 \)) and the "working" correlation matrix is simple. To specify compound symmetry (exchangeable in GEE) as the "working" correlation matrix, replace CORR=1 with CORR=4. The GEE macro is not as flexible as the MIXED procedure in handling unbalanced data.

### Model Validation

A simulation procedure was performed to verify the correct model specification using the MIXED procedure and GEE macro in linear growth curve analysis. The simulation data sets consist of two treatment groups with an equal number of subjects. Each subject has a baseline measurement plus six subsequent measures that are equally spaced having a linear relationship with time.

Data sets were simulated to have either (1) fixed effects \( Z=0 \) or (2) random effects \( Z=X \) for the intercept and slope parameters. The random effects data sets have \( b_{11} \sim \mathcal{N}(0,1.0) \) and \( b_{12} \sim \mathcal{N}(0,0.1) \). We set \( \beta_{01} = \beta_{02} = \beta_{11} = 0 \) and vary the slope for the second treatment group, \( \beta_{12} \), from 0 to 0.06 in increments of 0.02. The error variance matrix, or \( R \), was simulated as having simple (SIM, \( \rho = 0 \)) or compound symmetry (CS, \( \rho = 0.5 \)) structure with \( \sigma^2 = 1 \). In each analysis, the dependent outcome variable is one of the following:

<table>
<thead>
<tr>
<th>Outcome Variable Name</th>
<th>Type of Simulated Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>fixsim</td>
<td>Fixed effects and SIM error</td>
</tr>
<tr>
<td>fixcs</td>
<td>Fixed effects and CS error</td>
</tr>
<tr>
<td>randsim</td>
<td>Random effects and SIM error</td>
</tr>
<tr>
<td>randcs</td>
<td>Random effects and CS error</td>
</tr>
</tbody>
</table>

where the intercept and slope random effects are assumed to be independent of each other.

To each simulated data set, we apply the following five models:

<table>
<thead>
<tr>
<th>Model</th>
<th>Procedure</th>
<th>Random Effects</th>
<th>( R ) matrix structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMF(SIM)</td>
<td>MIXED</td>
<td>No SIM</td>
<td>SIM</td>
</tr>
<tr>
<td>MMF(CS)</td>
<td>MIXED</td>
<td>No CS</td>
<td>CS</td>
</tr>
<tr>
<td>MMR(SIM)</td>
<td>MIXED</td>
<td>Yes SIM</td>
<td>SIM</td>
</tr>
<tr>
<td>GEE(SIM)</td>
<td>GEE</td>
<td>No</td>
<td>SIM</td>
</tr>
<tr>
<td>GEE(CS)</td>
<td>GEE</td>
<td>No</td>
<td>CS</td>
</tr>
</tbody>
</table>

and examine slope estimates, power for test size and confidence interval coverage. To examine power for test size and confidence interval coverage, we include the following ESTIMATE statement:

```sas
estimate 'Slope' time -1 /alpha=.1;
```

in the MIXED procedure code. This conducts a hypothesis test at the \( \alpha = 0.05 \) significance level with the null being no difference in slopes. The 'cl' and 'alpha=' options create a 90% confidence interval around the estimated slope difference and the simulation procedure checks to see if the true difference is contained in the interval. In the MMR(SIM) model, the \( G \) matrix structure was specified as banded main diagonal.

There is no equivalent ESTIMATE statement in the GEE macro, so a different parameterization is needed to get an estimate of the difference in slopes between the two treatment groups. Let the variable \( \text{intercept} \) be equal to 1 for all observations and specify \( \text{intercept} \), \text{trt2}, \text{time} and \text{time2} as
variables in the XVAR statement of the GEE macro. The \texttt{time2} variable gives slope difference results. Hypothesis tests and confidence interval coverage were based on the robust standard errors.

Based on the mean results from 500 simulations, the MIXED procedure gave accurate parameter estimates when specification of random effects, G matrix structure and R matrix structure were correct. The power to reject the null hypothesis was approximately 0.05 when the true slope difference was zero and confidence interval coverage of the true parameter was approximately correct. In the interest of time, the maximum likelihood (ML) estimation method was used. Both GEE models performed well on all simulated data sets, even those with random effects. Valid results were seen for both small (N=25 subjects per group) and large (N=100 subjects per group) sample sizes.

**Model Misspecification**

The simulation procedure also gave results of model misspecification. The misspecification of random effects as fixed in the MIXED procedure gives anti-conservative results in that the null hypothesis of no difference in slopes is rejected too often. This is worrisome since fixed effects models are commonly used. To determine whether or not random effects are present, you can fit a random effects model and examine the G matrix. Variances that are close to zero imply a fixed effects model may be appropriate while variances significantly far from zero indicate a random effects model may be better.

Assuming a fixed effects model is correct, the MIXED procedure gives valid results if you misspecify the error structure as compound symmetry when it is truly simple. The MMF(CS) model just reduces to the MMF(SIM) model with $p = 0$. On the other hand, you obtain conservative results when a compound symmetry error structure is misspecified as simple in that the null hypothesis of no difference in slopes is not rejected often enough.

Assuming a random effects model is correct, the MIXED procedure gives reasonable conclusions if you specify the error structure as simple when the true structure is compound symmetry. The effect of specifying a compound symmetry correlation structure when the true structure is simple could not be examined because the model is overparameterized when intercepts and slopes are random.

The simulation results also show that the GEE model with simple or compound symmetry as the “working” correlation matrix performed well on all data sets. This is true even for random effects data, even though the GEE macro is not modeling the random effects. Thus, the MMR(SIM), GEE(SIM) and GEE(CS) models still provide valid results in terms of parameter estimation, test size and confidence interval coverage even if the model is misspecified. This is demonstrated with the example data set described below.

**Example**

The data set used to illustrate the different modeling approaches is an example from a published paper by Laird and Wang (1990). The data set comes from a study by Follinsbee, Donnell and Horstman (1988) and compares the effects of ozone exposure to clean air on forced expiratory volume in one second (FEV\textsubscript{1}). The FEV\textsubscript{1} measurements were taken at baseline and at 50 minute intervals thereafter for a total of seven measurements per subject. There are ten subjects per group with the same ten subjects in both the control and treatment group. As was done by Laird and Wang, we ignore the crossover design for simplicity and assume the study was conducted as a randomized trial. The forced expiratory volume data set is completely balanced with no missing data.

The variances of the slope random effects for the MMR(SIM) model taken from the G matrix are 734.4 for the control group and 9022.1 for the treatment group. The non-zero slope variances support a random effects model. The widely different standard variances of the slope random effects by group strongly indicate that the variances should not be constrained to be equal in the G matrix structure. This is not surprising since one expects to see more fluctuation in the treatment group compared to a control group. The MMF(SIM), MMF(CS), GEE(SIM), GEE(CS) and MMR(SIM) models give the same estimate of the slope difference. The GEE(SIM), GEE(CS) and MMR(SIM) models give approximately the same standard error estimate for the slope difference.

The variances of the slope random effects for the MMR(SIM) model taken from the G matrix are 734.4 for the control group and 9022.1 for the treatment group. The non-zero slope variances support a random effects model. The widely different standard variances of the slope random effects by group strongly indicate that the variances should not be constrained to be equal in the G matrix structure. This is not surprising since one expects to see more fluctuation in the treatment group compared to a control group. The MMF(SIM), MMF(CS), GEE(SIM), GEE(CS) and MMR(SIM) models give the same estimate of the slope difference. The GEE(SIM), GEE(CS) and MMR(SIM) models give approximately the same standard error estimate for the slope difference.

Due to the properties of GEE, the robust estimates and standard errors based on the GEE(SIM) and GEE(CS) models are the same. GEE also gives a naive standard error which is identical to that for the MMF model when the error structure is simple or compound symmetry.

In addition to the above models, we also tried...
auto-regressive (AR) and unstructured (UN) covariance structures. The more complex the covariance structure, the longer the model fitting procedure takes. Generally, GEE will converge faster than the MIXED procedure as the size of the data set increases and the covariance structure becomes more complex. The estimate and standard error for each group slope and the difference in slopes for the different model specifications are summarized in Table 1 below.

Table 1. Slope estimates (est.) and standard errors (s.e.) from ozone example data set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Control</th>
<th>Treatment</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est. S.E</td>
<td>Est. S.E</td>
<td>Est. S.E.</td>
</tr>
<tr>
<td>MMF(SIM)</td>
<td>15.6 24.9</td>
<td>-88.8 24.9</td>
<td>-104.4 35.3</td>
</tr>
<tr>
<td>MMF(CS)</td>
<td>15.6 11.7</td>
<td>-88.8 11.7</td>
<td>-104.4 16.6</td>
</tr>
<tr>
<td>MMF(AR)</td>
<td>12.4 20.5</td>
<td>-90.1 20.5</td>
<td>-102.5 29.1</td>
</tr>
<tr>
<td>MMF(UN)</td>
<td>-1.0 18.8</td>
<td>-91.3 18.8</td>
<td>-90.3 26.6</td>
</tr>
<tr>
<td>GEE(SIM)</td>
<td>15.6 11.8</td>
<td>-88.8 30.2</td>
<td>-104.4 32.4</td>
</tr>
<tr>
<td>GEE(CS)</td>
<td>15.6 11.8</td>
<td>-88.8 30.2</td>
<td>-104.4 32.4</td>
</tr>
<tr>
<td>GEE(AR)</td>
<td>12.5 12.9</td>
<td>-90.0 31.8</td>
<td>-102.6 34.3</td>
</tr>
<tr>
<td>GEE(UN)</td>
<td>9.5 18.4</td>
<td>-73.8 35.0</td>
<td>-83.3 39.5</td>
</tr>
<tr>
<td>MMR(SIM)</td>
<td>15.6 11.5</td>
<td>-88.8 31.0</td>
<td>-104.4 33.1</td>
</tr>
</tbody>
</table>

Conclusions

Linear growth curve analysis can be performed in SAS using the MIXED procedure or GEE macro. The MIXED procedure has nice modeling capabilities but depends on correct assumptions. Valid results rely on correct specification of between-subject variability (Z and G matrices) and within-subject variability (R matrix).

The unique property of GEE is robustness to misspecification of the correlation between repeated observations. Although the GEE macro does not currently model random effects, a simulation procedure found GEE to perform well when random effects are present in the data. GEE is not as flexible with unbalanced data but may converge faster compared to the MIXED procedure.

SAS, SAS/STAT and SAS/IML are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

References


Author Contact
Laura Ichikawa
Center for Health Studies
Group Health Cooperative
1730 Minor Ave., Suite 1600
Seattle, WA 98101