How to run a longitudinal GEE model with very large datasets in a reasonable amount of CPU time

Ellen Hertzmark and Donna Spiegelman

August 23, 2012

Abstract

This document contains instructions on how to run mixed (repeated measures) models for longitudinal data using a whole Channing cohort.
Keywords: SAS, mixed models, repeated measures

1 Motivation

Until now, people interested in longitudinal analysis with continuous outcomes have been sampling their data to make them runnable in PROC MIXED. This is NOT NECESSARY. In fact, with a bit of tweaking, PROC MIXED runs quite efficiently, even with a very large data set and a large number of repeated measures.

2 Setting up the data

Suppose you have the typical Channing dataset, with a record for each person-questionnaire cycle. It is already sorted by subject and time period.

If you have another type of dataset, you must sort it by subject and the index for the repeated measures.

If the data are sorted by the subject id, there is no need to make the subject a class variable. This saves a lot of memory when you

run the PROC MIXED.

Furthermore, in the example below we specify `DDFM=BW` (degrees of freedom method = Between-Within), which also makes things run faster.

Note that in the example below, where `BMI` is modeled in relation to protein intake, 49357 observations are deleted because of missing data (17109 with missing BMI or diet data in period 1, 17624 missing both in period 2, and 14624 missing both in period 3). Even though data are missing, it is important to keep all observations from subjects who have any data in the dataset, if the covariance structure is different at or between different times, such as when the working variance-covariance structure is unstructured (`TYPE=UN`), because they are needed to keep the place of each observation within the subject, so that the (1, 2) place in the covariance matrix will always be the relation between periods 1 and 2, and never between 1 and 3 (if 2 is missing) or between 2 and 3 (if 1 is missing).

We recommend using the `empirical` option to avoid requiring multivariate normality for valid inference.

3 Examples

Here is an example of models for the continuous outcome `BMI` with 3 time periods.

3.1 Random intercept and random slope for one predictor

First we specify a covariance with a random intercept and a random slope for `timepd`. The partial .saslog is shown.

```
1          title1 '/udd/stleh/helpme/yli/mixedmod.sas';
2
3          filename indat '/proj/nhdbxs/nhdbx0v/NHS2/formixed.data';
4          options fullstimer nocenter ps=78 ls=130 formdlim='[' ;
           /* the option FULLSTIMER allows us to know particulars about
```

```
              how much memory a step used */
5
9          data rs;
10         infile indat;
11         input id 1-6 bmi 8-18 aprot91a 20-25 timepd 27;
12         run;

13         proc sort data=rs;  by id;  run;
26
27         title2 'random intercept, and slope for timepd';
28         proc mixed data=rs;
29         model bmi=aprot91a timepd/s ddfm=bw;
30         random intercept timepd/type=un subject=id;
31         run;

NOTE: 49357 observations are not included because of missing values.
NOTE: Convergence criteria met.
NOTE: The PROCEDURE MIXED printed pages 1-2.
NOTE: PROCEDURE MIXED used (Total process time):
      real time           20.18 seconds
      user cpu time       14.98 seconds
      system cpu time     3.84 seconds
      Memory                              8233k
      Page Faults                         210
      Page Reclaims                       0
      Page Swaps                          0
      Voluntary Context Switches          1116
      Involuntary Context Switches        799
      Block Input Operations              230
      Block Output Operations             21403
```

So the first model, with one random slope, used about 19 seconds of (user+system) cpu and about 8 megabytes of memory. Here is the output.

```
[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[[
```

random intercept, and slope for timepd

The Mixed Procedure

Model Information

Data Set	WORK.RS
Dependent Variable	bmi
Covariance Structure	Unstructured
Subject Effect	id
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Empirical
Degrees of Freedom Method	Between-Within

Dimensions

Covariance Parameters	4
Columns in X	3
Columns in Z Per Subject	2
Subjects	84073
Max Obs Per Subject	3

Number of Observations

Number of Observations Read	252219
Number of Observations Used	202862
Number of Observations Not Used	49357

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	1303435.3381153	

1	3	1049443.8409174	0.00071731
2	1	1049179.5513401	0.00001630
3	1	1049173.9505007	0.00000001
4	1	1049173.9469250	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	31.3619
UN(2,1)	id	0.05804
UN(2,2)	id	0.6954
Residual		1.9588

Fit Statistics

-2 Log Likelihood	1049174
AIC (smaller is better)	1049188
AICC (smaller is better)	1049188
BIC (smaller is better)	1049253

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
3	254261.39	<.0001

[[

/udd/stleh/helpme/yli/mixedmod.sas 12:14 Wednesday, August 16, 2006 2
random intercept, and slope for timepd

```
The Mixed Procedure

                   Solution for Fixed Effects

                         Standard
Effect       Estimate       Error       DF    t Value    Pr > |t|

Intercept     21.5501     0.08191     82E3     263.10      <.0001
aprot91a      0.05662    0.001273     82E3      44.47      <.0001
timepd         0.5905    0.005225     12E4     113.02      <.0001


        Type 3 Tests of Fixed Effects

              Num     Den
Effect         DF      DF    F Value    Pr > F

aprot91a        1    82E3    1977.40    <.0001
timepd          1    12E4    12773.9    <.0001
```

[[

3.2 Random intercepts and slopes for two predictors

Now we specify a covariance with a random intercept and random slopes for `timepd` and protein intake (`aprot91a`).

```

          title2 'random intercept,  random slopes for timepd and aprot91a';
          proc mixed data=rs method=ml empirical;
          model bmi=aprot91a timepd/s ddfm=bw;
          random intercept aprot91a timepd/type=un subject=id;
          run;

NOTE: 49357 observations are not included because of missing values.
NOTE: Convergence criteria met.
NOTE: The PROCEDURE MIXED printed pages 3-4.
```

```
NOTE: PROCEDURE MIXED used (Total process time):
      real time           27.25 seconds
      user cpu time       22.51 seconds
      system cpu time     3.69 seconds
      Memory                           10283k
      Page Faults                      0
      Page Reclaims                    0
      Page Swaps                       0
      Voluntary Context Switches       1074
      Involuntary Context Switches     1081
      Block Input Operations           0
      Block Output Operations          19391
```

This model, with 2 random slopes, used about 26 seconds of (user+system) cpu and about 10 megabytes of memory. The output is

```
/udd/stleh/helpme/yli/mixedmod.sas        12:14 Wednesday, August 16, 2006   3
random intercept,  random slopes for timepd and aprot91a

The Mixed Procedure

                  Model Information

Data Set                     WORK.RS
Dependent Variable           bmi
Covariance Structure         Unstructured
Subject Effect               id
Estimation Method            ML
Residual Variance Method     Profile
Fixed Effects SE Method      Empirical
Degrees of Freedom Method    Between-Within


            Dimensions

Covariance Parameters             7
Columns in X                      3
Columns in Z Per Subject          3
Subjects                      84073
```

Max Obs Per Subject 3

Number of Observations

Number of Observations Read	252219
Number of Observations Used	202862
Number of Observations Not Used	49357

Iteration History

Iteration	Evaluations	-2 Log Like	Criterion
0	1	1303435.3381153	
1	3	1048888.8981202	0.00074013
2	1	1048616.0334194	0.00001751
3	1	1048610.0144801	0.00000001
4	1	1048610.0102814	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	id	17.5010
UN(2,1)	id	0.08272
UN(2,2)	id	0.000735
UN(3,1)	id	0.6342
UN(3,2)	id	-0.00861
UN(3,3)	id	0.6946
Residual		1.9591

Fit Statistics

-2 Log Likelihood	1048610
AIC (smaller is better)	1048630
AICC (smaller is better)	1048630
BIC (smaller is better)	1048723

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
6	254825.33	<.0001

[[

/udd/stleh/helpme/yli/mixedmod.sas 12:14 Wednesday, August 16, 2006 4
random intercept, random slopes for timepd and aprot91a

The Mixed Procedure

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	21.6098	0.08011	82E3	269.75	<.0001
aprot91a	0.05572	0.001244	82E3	44.80	<.0001
timepd	0.5905	0.005222	12E4	113.07	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
aprot91a	1	82E3	2007.12	<.0001
timepd	1	12E4	12785.5	<.0001

4 Credits

Written by Donna Spiegelman and Ellen Hertzmark for the Channing Lab.
Questions and comments may be directed to Ellen Hertzmark, `stleh@channing.harvard.edu`, (617) 432-4597.