BIOS6643 **HW4 Solutions, updated** **2014**

(1) Consider a study in which subjects' blood pressures are observed over time (4 time points, equally spaced, no missing data). The model will have fixed effects for time, age (at start of experiment) and gender; a random intercept for subjects will also be included.

a. Write out the mixed model, specifying all parts, if time is modeled as a continuous variable (linear term only) and the AR(1) structure is used to model the errors.

Solution: $Y_{ijk} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2i} + \alpha_k + b_i + \varepsilon_{ijk}$

where x_{1ij} is age for subject *i* at the start of the experiment, x_{2i} is time, α_k is gender effect for *k*=1,2 (females and males), b_i is a random intercept, ε_{ijk} is error that follows an AR(1) process.

b. How many fixed effect (β) parameters are in the model (considering full-rank parameterization)? How many covariance (α) parameters?

Solution: 4 fixed effects (intercept, slope for age, slope for time, and 1 for gender); 3 covariance parameters (variance for random intercept, residual variance and AR(1) correlation).

c. Write the form of Var($\mathbf{Y}_i$) in terms of the covariance parameters.

Solution: Recall that $Var(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^t + \mathbf{R}_i$. Here, $\mathbf{Z}_i$ is just a column of 1's since there is only a random intercept, so $\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i^t = \sigma_b^2\mathbf{J}$. Also, $\mathbf{R}_i = \sigma_\varepsilon^2\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix}$ based on the AR(1), so we have

$$Var(\mathbf{Y}_i) = \sigma_b^2\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + \sigma_\varepsilon^2\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{pmatrix} = \begin{pmatrix} \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi^3 \\ \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2\phi^2 \\ \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 & \sigma_b^2+\sigma_\varepsilon^2\phi \\ \sigma_b^2+\sigma_\varepsilon^2\phi^3 & \sigma_b^2+\sigma_\varepsilon^2\phi^2 & \sigma_b^2+\sigma_\varepsilon^2\phi & \sigma_b^2+\sigma_\varepsilon^2 \end{pmatrix}$$

d. Repeat parts a-c if time is modeled as a class variable, the UN structure is used for the covariance matrix of the errors and there is no random intercept term.

Solution: $Y_{ijk} = \beta_0 + \beta_1 x_i + \tau_j + \alpha_k + \varepsilon_{ij}$

where x_{1ij} is age for subject *i* at the start of the experiment, τ_j is time effect for j=1,2,3,4, α_k is gender effect for *k*=1,2 (females and males), b_i is a random intercept, ε_{ij} is error that has an unstructured covariance matrix. This is a less-than-full-rank model. There are 2 linear dependencies, one for time and one for gender.

There are 6 fixed effect parameters in the full rank model (after dropping 2 that are linearly dependent); there are 10 covariance parameters introduced with the UN structure on $\mathbf{R}_i$.

$$\mathbf{V}_i = \mathbf{R}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 \end{pmatrix}$$

e. **Not to turn in**: In the model, age at start of experiment (i.e., baseline) was used. How would estimates change if you used continuous age in the model instead? In order to answer the question, write out the statistical models for both approaches.

Can discuss in class! Also see lecture notes.

(2) Consider the linear mixed model (in subject form): $\mathbf{Y}_i$=$\mathbf{X}_i\boldsymbol{\beta}$+$\mathbf{Z}_i\mathbf{u}_i$+$\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_\mathrm{i}$~$N(\mathbf{0},\mathbf{R}_i)$ independent of $\mathbf{u}_i$~N($\mathbf{0}$,$\mathbf{G}_i$), and subjects i=1,…,n are independent. [Here, $\mathbf{G}_i$ and $\mathbf{R}_i$ are the same for all subjects; the index is just used to indicate dimensions.] A longitudinal experiment is conducted where subjects are observed at 3 equally spaced time points – for simplicity you can let the time points be 0,1, and 2. The data will be modeled using $\mathbf{b}_i$=(b_{i0},b_{i1}) defined for subjects (b_{i0} = random intercept; b_{i1} = random slope for time) and $\mathbf{R}$=$\sigma_\varepsilon^2\mathbf{I}$. There will be fixed-effect terms for time and possibly other variables, but their specification isn't relevant to the questions below. Justify all responses / show work.

a. Determine $\mathbf{V}_\mathrm{i}$=Var($\mathbf{Y}_\mathrm{i}$) (the model covariance matrix for subject i) if the unstructured covariance structure is used for **G**.

Solution: $\mathbf{Z}_i = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$, so

$$\mathbf{V}_i = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix} + \sigma_\varepsilon^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_0^2 + \sigma_\varepsilon^2 & \sigma_0^2 + \sigma_{01} & \sigma_0^2 + 2\sigma_{01} \\ \sigma_0^2 + \sigma_{01} & \sigma_0^2 + 2\sigma_{01} + \sigma_1^2 + \sigma_\varepsilon^2 & \sigma_0^2 + 3\sigma_{01} + 2\sigma_1^2 \\ \sigma_0^2 + 2\sigma_{01} & \sigma_0^2 + 3\sigma_{01} + 2\sigma_1^2 & \sigma_0^2 + 4\sigma_{01} + 4\sigma_1^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

b. Write the SAS, PROC MIXED code for the analysis of the data, if we allow for a simple linear time trend for fixed effects. (You can use generic names, e.g., *dat* for the data set, *time* for time.)

```
proc mixed data=dat;
class id;
model y=time;
random intercept time / type=un subject=id;
run;
```

c. Often a realistic covariance structure for $\mathbf{V}_i$ for longitudinal data is one where covariance between responses is positive but decays as time between responses increases. With the covariance structure you came up with in part a, is it possible for this structure to have covariance that decays as time between responses increases? Justify your response.

Solution: Yes, for example when the covariance σ_{01} is negative, you will see a decay. Also, the have the covariance for 2nd and 3rd time points greater than that of 1st and 3rd time points, we need $\sigma_{01} > -2\sigma_1^2$. Note that since variances are changing over time, we could also consider finding constraints on the correlations, which might be even more meaningful. However, this is an algebraic nightmare and would be aided by some algebra software (e.g., Mathematica).

(3) Consider the peak flow data fit with a simple random effect model, with SAS output presented in the LMM V slides and in the *5 LMM inference* notes, Section 4.

a. Interpret the tests for random effects, e.g., interpret the effect for subject 101S.

Solution: this is the deviation for a subject from the population mean; e.g., H0: bi=0 for 'I' related to subject 101S is the test of whether subject 101S's average differs from the population mean.

b. Based on the estimated covariance parameters, are you surprised that the majority of kids have p-values below 0.05? Explain.

Solution: No, because the between-subject variability is very high in relation to the within-subject error.

c. **Not to turn in**: For subject 101S, determine the following, replacing unknown parameters with their estimates. The data is posted on the course web site; you only need the January values for subject 101S. The quantities can be found in the slides and notes as mentioned above.

1. The shrinkage factor, $\hat{\lambda}$.

Solution: Note that $\lambda = \dfrac{r_i\sigma_u^2}{\sigma_\varepsilon^2 + r_i\sigma_u^2}$. Replacing unknown parameters with their estimates yields $\hat{\lambda}$=0.9917.

2. The random effect estimate, $\hat{b}_{0i}$.

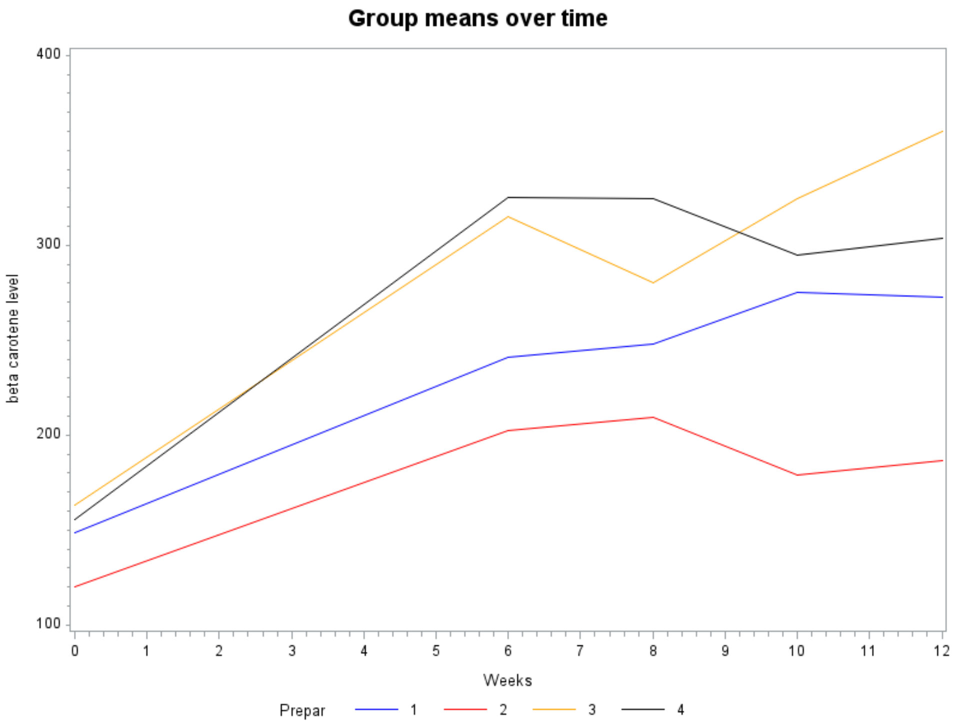
Solution: $\hat{b}_i\left(\boldsymbol{\theta}\right) = \lambda[\bar{Y}_i - \beta_0]$. Note that this is the Bayes estimator; the empirical Bayes estimator, $\hat{b}_i\left(\hat{\boldsymbol{\theta}}\right)$, is obtained by replacing unknown parameters on the right side with their estimators. We obtain $\hat{b}_i\left(\hat{\boldsymbol{\theta}}\right)$= 52.67, same as reported in the SAS output.

3. The predicted PEF, $\hat{Y}_i$. Is more weight given to the subject's data or the population average (estimate)? Does the result surprise you? Explain.

Solution: $\hat{\lambda}\overline{Y}_i + (1-\lambda)\hat{\beta}_0$ **= 262.88. See the 'practice questions HW1-5' document for more detail.**

(4) (From Rosner, 2006.) A clinical trial was planned comparing the incidence of cancer in a group taking beta-carotene in capsule form compared with a group taking beta-carotene placebo capsules. One issue in planning such a study is which preparation to use for the beta-carotene capsules. Four preparations were considered: (1) Solatene (30mg capsules), (2) Roche (60mg capsules), (3) BASF (30mg capsules), (4) BASF (60mg capsules). To test efficacy of the four agents in raising plasma-carotene levels, a small bioavailability study was conducted. After two consecutive-day fasting blood samples, 23 volunteers were randomized to one of the four preparations mentioned above, taking 1 pill every other day for 12 weeks. The primary endpoint was level of plasma carotene attained after moderately prolonged steady ingestion. For this purpose, blood samples were drawn at 6, 8, 10 and 12 weeks. In order to model the data, consider group and time as class variables. Other model specifications may depend on the question. **Just use the second baseline measure as the 'time 0' measure for parts a through f.**

a. Create a graph for the data. You have some flexibility on what type of graph to include.



b. Consider the model with a random intercept for subjects (in addition to other terms mentioned above); let **R** have the independent structure (**R**=σ_ε^2**I**). Mention 2 approaches to fitting the model and conducting tests.

Solution: (1) use repeated measures ANOVA; can be carried out with PROC GLM, and then adjusting tests as necessary using the correct denominator MS (as indicated by the expected mean squares); (2) use LMM methods for the random intercept model. (Specifying the R matrix to have the CS structure and not including a random intercept in the model will yield the same fit although it does not allow us to estimate subject random effects).

Synopsis of results using LMM methods:

```
proc mixed data=long.beta_carotene_univar;
class id prepar time;
model y= prepar time prepar*time;
random intercept / subject=id(prepar);
*Note:  could also just use 'id' in the subject option above since IDs are
unique study-wide; run;
```

AIC (smaller is better) 1093.1

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	**Id(Prepar)**	9025.37
Residual		2064.32

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Prepar	3	19	1.52	0.2405
time	4	76	34.66	<.0001
Prepar*time	12	76	1.99	0.0366

c. For the model above, determine the ICC and interpret the quantity.

Solution: The estimated ICC is 9025.37/(9025.37+2064.32)=81%. This means that 81% of the variability in the data (after accouting for fixed effects) is due to differences between subjects.

d. Now consider the same model but include a random slope for subjects in addition to the random intercept. Try the UN structure for **G** and compare results using the VC structure. Which model is better? Interpret the covariance term between the intercept and slope.

Solution: note that right now we have time as a class variable. If you want to keep time as a class variable for fixed effects and add a random slope for continuous time, you need to define another time variable:

```
data test; set
long.beta_carotene_univar; t=time; run;
proc mixed data=test;
  class id prepar time;
  model y= prepar time prepar*time;
  random intercept t /
subject=id(prepar); run;
```

```
proc mixed data=test;
  class id prepar time;
  model y= prepar time prepar*time;
  random intercept t /
      subject=id(prepar) type=un;
run;
```

```
        Covariance Parameter Estimates

        Cov Parm   Subject      Estimate

        Intercept  Id(Prepar)    4232.07

        t          Id(Prepar)    41.5106

        Residual                 1399.42


                Fit Statistics

         -2 Res Log Likelihood     1071.8

         AIC (smaller is better)   1077.8

         AICC (smaller is better)  1078.1

         BIC (smaller is better)   1081.2


          Type 3 Tests of Fixed Effects

Effect        Num DF  Den DF  F Value  Pr > F

Prepar             3      19     2.16  0.1264

time               4      57    21.42  <.0001

Prepar*time       12      57     1.87  0.0581
```

```
        Covariance Parameter Estimates

        Cov Parm   Subject      Estimate

        UN(1,1)    Id(Prepar)    2827.57

        UN(2,1)    Id(Prepar)     344.51

        UN(2,2)    Id(Prepar)    25.9845

        Residual                 1513.45


                Fit Statistics

         -2 Res Log Likelihood     1060.7

         AIC (smaller is better)   1068.7

         AICC (smaller is better)  1069.2

         BIC (smaller is better)   1073.3


          Type 3 Tests of Fixed Effects

Effect        Num DF  Den DF  F Value  Pr > F

Prepar             3      19     1.52  0.2405

time               4      57    24.52  <.0001

Prepar*time       12      57     1.90  0.0539
```

There is an improvement in the model fit by adding the covariance between intercept and slope (i.e., the UN structure). Interesting, the covariance is positive, which is less common in general…

e. Now consider a model with no random effects, but specifying **R** to have the UN structure. Fit this model and compare with the previous ones.

```
proc mixed data=test;
  class id prepar time;
  model y= prepar time prepar*time;
     repeated / subject=id(prepar) type=un r rcorr; run;
```

```
  Estimated R Matrix for Id(Prepar) 71 1

Row     Col1     Col2     Col3     Col4     Col5

  1  2824.78  4114.06  5333.61  3939.28  4304.46

  2  4114.06    14162    13215    11600    12287

  3  5333.61    13215    14514    11697    12450

  4  3939.28    11600    11697    11396    11313

  5  4304.46    12287    12450    11313    12552
```

```
 Estimated R Correlation Matrix for Id(Prepar) 71
                        1

    Col1      Col2      Col3      Col4      Col5

  1.0000    0.6505    0.8330    0.6943    0.7229

  0.6505    1.0000    0.9218    0.9131    0.9216

  0.8330    0.9218    1.0000    0.9095    0.9224

  0.6943    0.9131    0.9095    1.0000    0.9459

  0.7229    0.9216    0.9224    0.9459    1.0000
```

Fit Statistics

-2 Res Log Likelihood	1024.0
AIC (smaller is better)	1054.0
AICC (smaller is better)	1060.1
BIC (smaller is better)	1071.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
14	165.33	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Prepar	3	19	1.52	0.2405
time	4	19	16.00	<.0001
Prepar*time	12	19	2.23	0.0573

f. Write a paragraph about your findings. Include a couple of specific tests to support your story (with custom needs, if you'd like). You can also include a table with goodness-of-fit statistics for models fit in the previous parts. Out of all the models, which one would you use for your 'final' model? Why?

Solution: I am going to go with the last approach (in 10e, with UN structure for **R**) as the final model. The model with random intercept and slope and using the UN structure wasn't bad either. For the 10e approach, we get the best AIC, plus there is a little more flexibility in modeling the correlation of repeated measures. To show this, consider the correlation matrix for V [obtained by including VCORR as an option in the RANDOM statement for the 'Approach 1' (10d; random intercept and slope; UN structure for G) and by including RCORR as an option in the REPEATED statement for 'Approach 2' (10e; our 'final model'):

Estimated V Correlation Matrix for Id(Prepar) 71 1

Row	Col1	Col2	Col3	Col4	Col5
1	1.0000	0.7658	0.7897	0.8096	0.8263
2	0.7658	1.0000	0.8547	0.8677	0.8787
3	0.7897	0.8547	1.0000	0.8801	0.8900
4	0.8096	0.8677	0.8801	1.0000	0.8994
5	0.8263	0.8787	0.8900	0.8994	1.0000

Estimated R Correlation Matrix for Id(Prepar) 71 1

Col1	Col2	Col3	Col4	Col5
1.0000	0.6505	0.8330	0.6943	0.7229
0.6505	1.0000	0.9218	0.9131	0.9216
0.8330	0.9218	1.0000	0.9095	0.9224
0.6943	0.9131	0.9095	1.0000	0.9459
0.7229	0.9216	0.9224	0.9459	1.0000

We see a pattern using Approach 1 (left) that has some of the same characteristics of the unstructured, but we do lose a little flexibility. In particular, correlations between baseline and post-BL time points tend to be higher and correlations within post-BL time points tend to be lower with Approach 1 compared with Approach 2, indicating that Approach 1 does not have quite the flexibility in modeling these two types of correlations. Notice that Approach 1 does capture the trend of increasing correlation between equally spaced time points as time increases…the AR(1) and Toeplitz structure would probably not work well with these data; the AR(1) will force decaying correlations between responses as times get further apart, and the Toeplitz assumes that correlations between measures that have the same time spacing are the same.

Here are some findings for fixed-effect part of the model: The significance of the time variable indicates that overall, the beta carotene seems to be storing up in the body ($p<0.0001$). The prepar*time interaction is marginally significant ($p=0.06$ for last model), indicating that differences in drugs changes over time (with marginal significance). I was interested particularly in the BASF 30 versus 60 mg. I ran one test for equal means, which again yielded a marginally significant p value ($p=0.02$). However, when comparing at individual time points, no significant differences were found.

This demonstrates that taking the extra 30mg did not yield significant gains in beta carotene storage. In fact, the sample mean beta carotene level at the last time point was greater for the 30mg treatment, as shown in the previous graph. Below are the tests for BASF. I used the model with the UN structure for repeated measures since it yielded the lowest AIC.

```
proc mixed data=test;
  class id prepar time;
  model y= prepar time prepar*time;
  repeated / subject=id(prepar) type=un r rcorr;
  estimate 'basf 30 vs 60 t1' prepar 0 0 1 -1
    prepar*time 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0 0 0;
  estimate 'basf 30 vs 60 t2' prepar 0 0 1 -1
    prepar*time 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0 0;
  estimate 'basf 30 vs 60 t3' prepar 0 0 1 -1
    prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0;
  estimate 'basf 30 vs 60 t4' prepar 0 0 1 -1
    prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0;
  estimate 'basf 30 vs 60 t5' prepar 0 0 1 -1
    prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1;run;
  contrast 'basf 30 vs 60'
      prepar 0 0 1 -1 prepar*time 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0 0 0,
      prepar 0 0 1 -1 prepar*time 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0 0,
      prepar 0 0 1 -1 prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0 0,
      prepar 0 0 1 -1 prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1 0,
      prepar 0 0 1 -1 prepar*time 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 -1;
```

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
basf 30 vs 60 t1	7.5333	32.1831	19	0.23	0.8174
basf 30 vs 60 t2	-10.5333	72.0605	19	-0.15	0.8853
basf 30 vs 60 t3	-44.3333	72.9504	19	-0.61	0.5506
basf 30 vs 60 t4	29.7333	64.6404	19	0.46	0.6508
basf 30 vs 60 t5	56.3333	67.8418	19	0.83	0.4166

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
basf 30 vs 60	5	19	3.67	0.0172

g. **Not to turn in**: There are two baseline variables. There are different ways these variables could be used in an analysis. For example, you could simply use one or the other, or perhaps the average of the two. Suggest what you would use for a baseline measure. [Note: you receive the data not knowing for certain whether all the values are correct or not.]

Solution: One of the variables has an outlier. It is unclear whether it is erroneous or not, but looks very suspicious. It is for that reason I did not use that (entire) first baseline variable. Otherwise, you might consider the average of the two. However, another consideration would variability. By averaging 2 variables, the baseline variable would naturally have less variability that the others. We could account for that in the model, but we'd lose consistency between measurement types of the repeated measures (one would involve the average of two, the others would involve only one).