# Appendix E: SAS PROC MIXED Syntax

We summarize the basic syntax of SAS `proc mixed`. The usual linear mixed effects model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,$$

$$E(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{D}, \quad E(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = E(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{0}, \quad \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i, \boldsymbol{b}_i) = \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \boldsymbol{R}_i$$

(`proc mixed` does not directly accommodate specifications of $\boldsymbol{R}_i$ that depend on $\boldsymbol{b}_i$). The model is usually written in software documentation in a streamlined form by "stacking" the contributions from each individual. Define

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix}, \quad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_m \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{R}_m \end{pmatrix},$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \\ \vdots \\ \boldsymbol{b}_m \end{pmatrix}, \quad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{Z}_m \end{pmatrix}, \quad \widetilde{\boldsymbol{D}} = \begin{pmatrix} \boldsymbol{D} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{D} \end{pmatrix}.$$

Here, $\widetilde{\boldsymbol{D}}$ has been displayed in the case where $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i) = \boldsymbol{D}$ for all individuals (so is independent of $\boldsymbol{x}_i$), but can be modified if this is relaxed, as in the dental study with the girls and boys having different matrices $\boldsymbol{D}_G$ and $\boldsymbol{D}_B$ (so depending on $\boldsymbol{a}_i$). The model can be written consisely with these definitions as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e}, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{D}}\boldsymbol{Z}^T + \boldsymbol{R}. \tag{B.1}$$

The SAS documentation refers to $\widetilde{\boldsymbol{D}}$ as $\boldsymbol{G}$.

The syntax for `proc mixed` is geared to the ***subject-specific*** linear mixed effects model; however, the procedure can also be used to fit ***population-averaged*** linear models of the form

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \epsilon_i$$

or in "stacked" form

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad E(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{X}\beta, \quad \text{var}(\epsilon|\tilde{\boldsymbol{x}}) = \text{var}(\boldsymbol{Y}|\tilde{\boldsymbol{x}}) = \boldsymbol{V},$$

where $\boldsymbol{V}$ does not have a specific structure; the structure of $\boldsymbol{V}$ is specified fully by the analyst and is not induced by the model.

For either type of model, the form of the ***population mean*** is either induced (SS model) or is specified explicitly (PA model). The `model` statement is of course the mechanism by which the analyst specifies the form the population mean, which is $\boldsymbol{X}_i\boldsymbol{\beta}$ for $i$ or $\boldsymbol{X}\boldsymbol{\beta}$ for all individuals, stacked, in the usual SAS way.

In the context of a PA model, we have used the `repeated` statement to specify the overall covariance matrix $\boldsymbol{V}$. For the SS linear mixed effects model, the `repeated` statement is used to specify the form of the ***within-individual*** covariance model $\boldsymbol{R}_i$ or, equivalently, $\boldsymbol{R}$ above. For this model the `random` statement is used to specify the assumption on var($\boldsymbol{b}_i|\boldsymbol{x}_i$) ($\widetilde{\boldsymbol{D}}$).

Here is a summary of the basic form of a call to `proc mixed`.

```
proc mixed data=dataset method= (ML,REML);
class  classification variables;
model response =  columns of  X / solution;
random columns of  Z / type=  subject=  group=   ;
repeated  classification variable for time / type=  subject=  group=   ;
run;
```

`proc mixed` statement

- `method=REML` is the default; no `method=` required in this case

`model` statement

- ***columns of X*** are variables (`class` or continuous) corresponding to variables associated with fixed effects $\boldsymbol{\beta}$
- Intercept is assumed unless `noint` option after slash
- `solution` is an option and must be invoked to get the estimates of $\boldsymbol{\beta}$

`random` statement

- Describes the matrix $\widetilde{\boldsymbol{D}} = \text{var}(\boldsymbol{b}|\tilde{\boldsymbol{x}})$ (i.e. the matrices $\text{var}(\boldsymbol{b}_i|\boldsymbol{x}_i)$ making up the blocks of $\widetilde{\boldsymbol{D}}$
- ***columns of Z*** are variables (`class` or continuous), i.e. variables associated with random effects $\boldsymbol{b}$
- `subject=` tells `mixed` what `class` variable denotes the grouping determining the ***individuals***
- `type=` allows choice of matrix (e.g. `un`, unstructured)
- `group=` allows $\boldsymbol{D}$ to be different according to this `class` variable (e.g. dental study, boys, girls)

`repeated` statement

- Describes the matrix $\boldsymbol{R} = \text{var}(\boldsymbol{e}|\tilde{\boldsymbol{x}})$ (i.e. the matrices $\boldsymbol{R}_i = \text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i)$ )
- If $\text{var}(\boldsymbol{e}_i \boldsymbol{x}_i) = \sigma^2 \boldsymbol{I}_{n_i}$ is the same for all $i$ `repeated` statement is ***NOT*** needed
- `subject=` tells `mixed` what `class` variable denotes the grouping determining the individuals
- `type=` allows choice other than diagonal (e.g. `ar(1), cs, etc.`)
- The optional classification variable before the slash is for situations with unbalanced data and nondiagonal `type` so that observations can be correctly attributed to the times at which they were taken
- `group=` allows $\boldsymbol{R}_i$ to be different depending on group membership (e.g. dental study, $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \sigma_G^2 \boldsymbol{I}_{n_i}$ girls, $\text{var}(\boldsymbol{e}_i|\boldsymbol{x}_i) = \sigma_B^2 \boldsymbol{I}_{n_i}$ boys)

The foregoing syntax makes clear that, to implement a linear PA model using `proc mixed` with the `repeated` statement, we simply make a correspondence between this model and the model (B.1) with ***no*** random effects $\boldsymbol{b}$, From purely ***operational*** point of view (but ***not*** an ***interpretation*** point of view), the models have the same structure – a mean plus a deviation with components of length $n_i$, each of which has a covariance matrix. Thus, purely to specify these covariance matrices for the PA model, the `repeated` statement can be used.

See the SAS documentation for `proc mixed` for much more detail on the use of these statements and available options.