Summary of Input Parameters to GEE Sample Size Software

The software is written in **PROC IML** of **SAS**, and all calculations are performed in **PROC IML**. There are basically five steps involved. See the examples provided to see how these steps are invoked.

1. The user invokes **PROC IML**.

2. The software is located in the file **GEESIZE.SAS**. This software is read into **PROC IML** using the **%INCLUDE** command. Make sure to point to the appropriate directory where the software has been stored on your system. Use the **NOSOURCE2** option to avoid having the lines of code listed in the **LOG** window.

3. A call is made to the **BEGIN** subroutine by typing **RUN BEGIN;** This initializes the software and sets the various matrices to their default values. **BEGIN** would also be invoked at the beginning of a new problem or to re-initialize the various matrices for any reason.

4. The input parameters characterizing the sample size problem are specified to the software. This is performed by setting the corresponding matrices to their values in **PROC IML**. Note that many input parameters have already been set to default values and may not need to be specified explicitly.

5. The calculations are invoked by using the **RUN GEESIZE;** command.

Steps 4 and 5 can then be repeated to investigate different scenarios, e.g., different values for ϕ and θ. Note that all input matrices retain their values from one call of the **GEESIZE** subroutine to the next. It is only necessary to specify those matrices that have changed value since the previous call.

The following are user-defined input parameters to the software:

_MU_.....[*Required*] An ($S \times T$) matrix of expected values $\{\mu_{st}\}$ under the model, where there are *S* subpopulations of interest and *T* time points in the study. For count data, the expected values must be positive. For binary data, they must lie between 0 and 1 exclusive.

_X_.....An ($ST \times r$) design matrix for the study. This is the design matrix that we anticipate for the GEE analysis at the end of the study. A ($T \times r$) design matrix $\mathbf{X}_s$ is defined for each of the *S* subpopulations, whereupon they are stacked on top of each other. It defaults to a one-way ANOVA model across the *S* subpopulations.

SNAME.....An ($1 \times S$) character vector providing names for the subpopulations to be used in the printout. It defaults to non-descript, politically-correct labels!

LINK.....A (1×1) character scalar indicating the link $h(\mu)$ to be used in the analysis. Valid values include, **'Identity'**, **'Log'**, **'Recip'**, **'Logit'** and **'CLoglog'**. It defaults to **'Identity'**.

VARI.....A (1×1) character scalar indicating the variance function $g(\mu)$ to be used in the analysis. Valid values include **'Gaussian'**, **'Poisson'**, **'Gamma'** and **'Bernoulli'**. It defaults to **'Gaussian'**.

STD.....Standard deviation for the outcome variable. It is relevant only in the Gaussian case and is ignored for the other distributions. It can be a scalar, in which case it applies to all subpopulations at all time points. Otherwise, it must be an ($S \times T$) matrix providing for different standard deviations in each subpopulation $\times$ time combination.

TIMES.....A ($1 \times T$) vector indicating the follow-up times. It defaults to the equally-spaced time points, $\{1,2,\ldots,T\}$.

PHI.....A (1×1) scalar providing the autocorrelation parameter ϕ in the damped exponential correlation model of Muñoz *et al*. (1992). It defaults to 0.

THETA.....A (1×1) scalar providing the "damping" parameter θ in the damped exponential correlation model of Muñoz *et al*. (1992). It defaults to 1, namely, the AR(1) model.

PSI.....A (1×1) scalar providing the "scale" parameter ψ for the covariance structure. It is an inflation factor to the variance at the various time points. It defaults to 1.

Note that for Gaussian outcomes, **_PSI_** can be interpreted as the variance of the outcome for each subpopulation at each time point. Thus, in the Gaussian case (only), the parameters **_STD_** and **_PSI_** are redundant. One should specify one or the other but not both.

Moreover, when changing from a non-Gaussian distribution to the Gaussian distribution, one must reset one of these parameters to 1 and the other to the value of interest. Otherwise, they confuse each other!

_H_.....An ($h \times r$) hypothesis matrix **H** among the elements of linear model parameters β. This is the specific hypothesis for which control is sought and is written in the form $\mathbf{H}\beta = \mathbf{h}_0$. It defaults to a full-rank comparison of the mean value over time across the *S* subpopulations.

_h0_ .....An ($h \times 1$) vector of constant terms $\mathbf{h}_0$ in the linear hypothesis above. It defaults to a vector of zeros.

_TYPE_I.....A (1×1) scalar signifying the type-I error rate α. It is assumed to be two-sided and defaults to .05.

_TYPE_II....A (1×1) scalar signifying the type-II error rate γ. It defaults to .20, corresponding to power at 80% to detect a significant difference.

REL_SIZE....An ($1 \times S$) vector providing the relative sizes $\{\omega_s\}$ across the different subpopulations in the asymmetric allocation problem.

_PI_.....An ($S \times T$) matrix of probabilities $\{\pi_{st}\}$ for staggered entry and loss to follow-up. In each subpopulation (row), the t-th column is the marginal probability of providing the first t repeated measures (only) to the study before being censored or lost to follow-up. Since these are marginal probabilities they must sum to 1 in each row.

PRINT.....A (1×1) scalar indicating the amount of detail provided in the printout. Larger values imply greater detail. Legitimate values are 1, 2 and 3. It defaults to 1.