Paper SAS030-2014

Power and Sample Size for MANOVA and Repeated Measures with the GLMPOWER Procedure

John Castelloe, SAS Institute Inc.

ABSTRACT

Power analysis helps you plan a study that has a controlled probability of detecting a meaningful effect, giving you conclusive results with maximum efficiency. SAS/STAT® provides two procedures for performing sample size and power computations: the POWER procedure provides analyses for a wide variety of different statistical tests, and the GLMPOWER procedure focuses on power analysis for general linear models. In SAS/STAT 13.1, the GLMPOWER procedure has been updated to enable power analysis for multivariate linear models and repeated measures studies. Much of the syntax is similar to the syntax of the GLM procedure, including both the new MANOVA and REPEATED statements and the existing MODEL and CONTRAST statements. In addition, PROC GLMPOWER offers flexible yet parsimonious options for specifying the covariance. One such option is the two-parameter linear exponent autoregressive (LEAR) correlation structure, which includes other common structures such as AR(1), compound symmetry, and first-order moving average as special cases. This paper reviews the new repeated measures features of PROC GLMPOWER, demonstrates their use in several examples, and discusses the pros and cons of the MANOVA and repeated measures approaches.

PROLOGUE

You are a consulting statistician at a manufacturer of herbal medicines, charged with calculating the required sample size for an upcoming repeated measures study of a new product called SASGlobalFlora (SGF), comparing it to a placebo. Your boss communicates the study plans and assumptions to you as follows:

- The outcome to be measured is a “wellness score,” which ranges from 0 to 60 and is assumed to be approximately normally distributed.
- Wellness is to be assessed at one, three, and six months.
- Subjects are to be allocated to the placebo and SGF at a ratio of 2 to 1, respectively.
- SGF is expected to increase the wellness score almost twice as much as the placebo over the six-month study period, according to the conjectured wellness score means over time shown in Table 1.
- The wellness standard deviation is expected to be approximately constant across time at a value of about 3.2.
- The planned data analysis is a chi-square test of the treatment-by-time interaction.
- The goal is to determine the number of subjects that are needed to achieve a power of 0.9 at a 0.05 significance level.

Table 1 Conjectured Wellness Score Means by Treatment and Time

		Time	
Treatment	**1 Month**	**3 Months**	**6 Months**
Placebo	32	36	39
SGF	35	40	48

There’s no information here about the expected within-subject correlations across time. But rather than bother the boss about that, you decide to proceed as best you can without it. You figure it’s probably OK in power analysis to assume a univariate model with a fixed subject effect instead of the repeated measures model, because that will hopefully yield only a slightly conservative sample size. You aren't sure how to properly account for the correlations anyway.

You proceed assuming the univariate model and use PROC GLMPOWER to compute a required sample size of 30 subjects. (See the section "APPENDIX: A TALE OF TWO POWER ANALYSES" on page 18 for the computational details of this particular power analysis.)

To corroborate your answer, you assign your bright new intern first to bother the boss for conjectured correlations and then to do a simulation to check the required sample size with the *multivariate* model, the one that will actually be used in the data analysis. Based on what you've heard about the power analysis relationship between univariate and repeated measures models, you expect the intern's estimated number of subjects to be lower than the 30 that you estimated, but only a little—maybe 24 or 27, because the 2:1 sampling plan forces subjects to come in groups of three.

But your boss stops you in the hallway a few days later with a concerned look on her face. "Your intern seems to think we'll need only *half* as many subjects as you figured," she says. "He seems to know what he's doing. I assume you'll double-check. This SGF doesn't just grow on . . . doesn't grow on *inexpensive* trees, you know. And it will be so much quicker to recruit only 15 subjects."

You hustle back to your office wondering if maybe you were wrong about assuming the univariate model. Is there anything you can do about it? You could always fall back on the simulation approach, but you know that it's awkward for producing power curves and for conducting sensitivity analyses about the conjectured means and variability. Also, the suits at your company will be more comfortable with a closed-form solution.

Furiously, you comb the SAS/STAT 13.1 documentation for a better way to do a repeated measures power analysis. Stay tuned . . . Your story continues in the section "EXAMPLE: TIME BY TREATMENT" on page 7.

INTRODUCTION

Statistical power analysis determines the ability of a study to detect a meaningful effect size—for example, the difference between two population means. It also finds the sample size that is required to provide a desired power for an effect of scientific interest. Proper planning reduces the risk of conducting a study that will not produce useful results and determines the most sensitive design for the resources available. Power analysis is now integral to the health and behavioral sciences, and its use is steadily increasing wherever empirical studies are performed.

Before SAS/STAT 13.1, the GLMPOWER procedure enabled you to conduct power analyses for tests and contrasts of fixed effects in univariate linear models. In SAS/STAT 13.1, it has been updated to handle multivariate linear models (MANOVA) and repeated measures studies. You can use these new features to help design studies for a wide variety of applications, such as industrial split-plot designs, agricultural variety studies, and advertising campaigns. The examples in this paper focus on the designs and analyses most commonly encountered in clinical trials.

The syntax of PROC GLMPOWER is most closely associated with that of the GLM procedure, and as with PROC GLM you can use it for several common special cases of mixed models that can also be analyzed using PROC MIXED. This property of being able to analyze a mixed model by using an equivalent MANOVA is very important in the approach discussed in this paper, and for that reason such models are given a name: "reversible," a term coined by Muller to cover the methods discussed in Muller and Stewart (2006). The examples in this paper illustrate two scenarios that involve reversible models:

- testing a treatment-by-time interaction in a repeated measures analysis ("EXAMPLE: TIME BY TREATMENT" on page 7 and "EXAMPLE: MULTILEVEL CORRELATION STRUCTURE" on page 12)
- testing for a treatment effect in a clustered data analysis ("EXAMPLE: CLUSTERED DATA" on page 10)

The primary syntax elements for the new PROC GLMPOWER features for MANOVA and repeated measures are summarized in Table 2.

Table 2 New Statements and Options in the GLMPOWER Procedure

Statement	Option	Description
REPEATED		Defines within-subject linear tests of model parameters in terms of common repeated measures transformations of the dependent variables (contrast, identity, polynomial, profile, Helmert, and mean)
MANOVA	M=	Defines within-subject linear tests of model parameters in terms of the matrix coefficients of the dependent variable transformation
POWER	MTEST=	Specifies the test statistic
POWER	CORRMAT=	Specifies the correlation matrix of the dependent variables
POWER	SQRTVAR=	Specifies the vector of error standard deviations of the dependent variables

REVIEW OF POWER AND SAMPLE SIZE

To help get you up to speed for the rest of the story about SASGlobalFlora that began in the Prologue, this section reviews the concepts and terminology that you encounter in power analysis, including a clarification of prospective versus retrospective analyses and a breakdown of the components of a power analysis for a multivariate linear model.

Concepts

Most of the time, you undertake a study to confirm an effect that you hypothesize will be there. This approach can go wrong in two ways:

1. The noise in your measurements might be too large or the study too small to declare statistical significance.
2. The study might be so large that the effect is hugely significant—encouraging, but wasteful.

How can you make sure that your study is not too small and not too large, but just right?

Power analysis is just such a way to get a “Goldilocks Solution” for resource usage and study design, improving your chances of obtaining conclusive results with maximum efficiency. Power analysis is most effective when performed at the study planning stage, and therefore it encourages early collaboration between researcher and statistician. It also focuses attention on effect sizes and variability in the underlying scientific process, concepts that both researcher and statistician should consider carefully at this stage. Muller and Benignus (1992) and O’Brien and Muller (1993) cover these and related concepts. These references also provide a good general introduction to power analysis.

A power analysis involves many factors, such as the research objective, design, data analysis method, power, sample size, Type I error, variability, and effect size. By performing a power analysis, you can learn about the relationships among these factors, optimizing those that are under your control and exploring the implications of those that are not.

Terminology

In statistical hypothesis testing, you usually express the belief that some effect exists in a population by specifying an alternative hypothesis, H_1. You state a null hypothesis, H_0, as the assertion that the effect does *not* exist and attempt to gather evidence to reject H_0 in favor of H_1. You gather evidence in the form of sample data, and you perform a statistical test to assess H_0. If H_0 is rejected but there really is *no* effect, this is called a *Type I error*. The probability of a Type I error is usually designated as “alpha” or α, and statistical tests are designed to ensure that α is suitably small (for example, less than 0.05).

If there really is an effect in the population but H_0 is *not* rejected in the statistical test, then that’s a *Type II error*. The probability of a Type II error is usually designated as “beta” or β. The probability $1 - \beta$ of avoiding a Type II error—that is, correctly rejecting H_0 and achieving statistical significance—is called the *power*.

(Note, however, that another, more technical definition of power is the probability of rejecting H_0 for any given set of circumstances, even those corresponding to H_0 being true.)

An important goal in study planning is to ensure an acceptably high level of power. Sample size plays a prominent role in power computations, because the focus is often on determining a sufficient sample size to achieve a certain power, or conversely, on assessing the power for a range of different sample sizes. For this reason, terms like *power analysis*, *sample size analysis*, and *power computations* are often used interchangeably to refer to the investigation of relationships among power, sample size, and other factors involved in study planning.

Prospective versus Retrospective

It is crucial to distinguish between *prospective* and *retrospective* power analyses. A prospective power analysis looks ahead to a future study, whereas a retrospective power analysis attempts to characterize a completed study. Sometimes the distinction is a bit fuzzy: for example, a retrospective analysis of a recently completed study can become a prospective analysis if it leads to the planning of a new study to address the same research objectives but with improved resource allocation.

Although a retrospective analysis is the most convenient type of power analysis to perform, it is often uninformative or misleading, especially when power is computed for the observed effect size. (For more information, see Lenth 2001.)

Power analysis is most effective when performed as part of study planning, and this paper considers only prospective power analysis.

Components

Power and sample size computations for multivariate linear models present a somewhat greater level of complexity than that required for simple hypothesis tests. You need to perform a number of steps to gather the required information to perform these computations. After settling on a clear research question, you must (1) define the *study design*; (2) make specific conjectures about the *means*, *variances*, and particularly the *correlations*; (3) specify the *statistical tests* that will best address the research question; and (4) characterize the *goal* as either a power or sample size computation. In hypothesis testing, you usually want to compute the powers for a range of sample sizes or vice versa. All this work has strong parallels to ordinary data analysis.

Even when the research questions and study design seem straightforward, the ensuing sample size analysis can seem technically daunting. It is often helpful to break the process down into four components:

- **Study Design**

  What is the structure of the planned design? This must be clearly and completely specified. What groups or treatments will you assess, and what will be the relative sample sizes across their levels? Will there be repeated measurements, clusters, or multiple outcomes?

- **Means, Variances, and Correlations**

  What are your beliefs about patterns in the data? What levels of “signals and noises” do you suspect in these patterns (or alternatively for the signals, what levels are you interested in detecting)? Imagine that you had unlimited time and resources to execute the study design, so that you could gather an “infinite data set.” Construct an “exemplary” data set that characterizes the cell means of this infinite data set, representing each design profile as a single observation, with the dependent variables containing the means. Also characterize the variance for each level of repeated measurement, cluster, and outcome, and the correlation structure among those levels.

  Positing correlations is such a big new wrinkle in power analysis for multivariate models, compared to univariate models, that an entire subsequent section (“SPECIFYING CORRELATIONS” on page 6) is devoted to it.

  Usually you will conjecture the means, variances, and correlations based on educated guesses of their true values. If instead you are interested in minimal clinical significance, then you can specify values that produce an effect size representing this. However, this minimal effect size is often so small that it requires excessive resources to detect. Or you can consider a variety of realistic possibilities for the effect size by performing a sensitivity analysis: you can construct multiple exemplary data sets that capture competing views of the cell mean patterns, and you can specify a range of values for variances and correlations. Your choice of strategy is ultimately determined by the goal of your power analysis.

- **Statistical Tests**

  How will you cast your model in statistical terms and conduct the eventual data analysis? Define the statistical model that will be used to embody the study design and test the effects central to the research question. What between-subject contrasts and within-subject transformations do you plan to test? Which significance level will you use, and what multivariate or univariate test statistics? Consider the covariance structure in your choice of test statistic: Muller et al. (2007) mention that univariate tests tend to be more powerful for situations close to compound symmetry, and multivariate tests are usually better under more complicated covariance structures.

- **Goal**

  Finally, what do you need to determine in the power analysis? Most often you want to examine the statistical powers across various scenarios for the means, variances, correlations, statistical tests, and feasible total sample sizes. Or you might want to find sample size values that provide given levels of power, say, 80%, 90%, or 95%. Are you interested in the presumed actual effect size or in minimal clinical significance?

NEW PROC GLMPOWER FEATURES FOR MANOVA AND REPEATED MEASURES

The statistical analyses that are newly covered in SAS/STAT 13.1 are Type III F tests and contrasts of fixed effects in multivariate linear models. You can choose among Wilks' likelihood ratio, the Hotelling-Lawley trace, and Pillai's trace as the basis of F tests for multivariate analysis of variance (MANOVA) and among uncorrected, Greenhouse-Geisser, Huynh-Feldt, and Box conservative F tests for the univariate approach to repeated measures. Tests and contrasts that involve random effects are not supported.

You can use either the MANOVA statement or the REPEATED statement in the GLMPOWER procedure to specify a multivariate linear model and transformations of the dependent variables. These statements are similar to their respective analogues in the GLM procedure.

Methodology

Power computations for multivariate linear models in PROC GLMPOWER are primarily based on noncentral F calculations, exact where possible and approximate elsewhere. Sample size is computed by inverting the power equation. Power computation methods for multivariate tests are based on Muller and Peterson (1984); Muller and Benignus (1992); and O'Brien and Shieh (1992). Methods for univariate tests for multivariate models are based on Muller and Barton (1989) and Muller et al. (2007). For more information about the methodology, see the section "Contrasts in Fixed-Effect Multivariate Models" in the GLMPOWER procedure chapter of the *SAS/STAT User's Guide*.

PROC GLM MANOVA Analyses

The new MANOVA statement in the GLMPOWER procedure enables you to define custom within-subject linear transformations of the responses by specifying an $\mathbf{M}$ vector or matrix for testing the hypothesis $\mathbf{L}\boldsymbol{\beta}\mathbf{M} = 0$.

To handle repeated measures of the same experimental unit, you would usually use the REPEATED statement instead of the MANOVA statement. But you can use the MANOVA statement in repeated measures situations, in addition to situations where you have clusters or multiple outcome variables.

PROC GLM Repeated Measures Analyses

The new REPEATED statement in the GLMPOWER procedure enables you to specify commonly used within-subject transformations such as pairwise comparisons, overall tests, polynomial trends, factor levels versus the mean of other levels, and Helmert contrasts. You use keywords rather than specifying coefficients of the $\mathbf{M}$ matrix, but you are limited to only these special cases of the $\mathbf{M}$ matrix.

Usually the REPEATED statement is used for handling repeated measurements on the same experimental unit, but you can also use the REPEATED statement for other situations, such as clusters or multiple outcome variables.

In SAS/STAT 13.1, you can specify only a single repeated factor in the REPEATED statement.

PROC MIXED Analyses

Muller et al. (2013) and Maldonado-Molina, Barón, and Kreidler (2013) discuss two common cases of "reversible" analyses, which are analyses that have equivalent PROC GLM and PROC MIXED formulations for Type III F tests of fixed effects in a "nice design":

- A PROC MIXED repeated measures analysis with unstructured covariance (REPEATED / TYPE=UN) that uses the Kenward-Roger degrees-of-freedom method (MODEL / DDFM=KR) is equivalent to a PROC GLM analysis that uses the F test based on the Hotelling-Lawley trace.
- A PROC MIXED repeated measures analysis with compound symmetry (REPEATED / TYPE=CS) that uses the default degrees-of-freedom method (MODEL / DDFM=BW) is equivalent to a PROC GLM analysis that uses the uncorrected univariate F test. (This model is also equivalent to a random-intercept-only model in PROC MIXED.)

A "nice design" in this context is one that has the following properties:

- no missing or mistimed data
- balanced within the independent sampling unit (ISU)
- treatment assignment unchanging over time
- no repeated covariates
- saturated in time and time-by-treatment effects

Thus you can use the GLMPOWER procedure for mixed models that have the preceding properties. These two special cases of reversible analyses are established in Edwards et al. (2008) and Gurka, Edwards, and Muller (2011). For more information about reversible analyses, see Muller and Stewart (2006).

In Defense of Classic MANOVA and Repeated Analyses

When PROC MIXED hit the scene in 1992, its generality and versatility helped it quickly become the de facto standard, eclipsing the classic MANOVA and repeated measures analyses in PROC GLM. But power computations for these classic analyses are solid, with exact formulas for many cases and good closed-form approximations for the others.

For many situations, the MANOVA approach also provides the best data analysis. Muller et al. (2007) caution that for small to moderate sample sizes, the standard mixed model analyses (except for the reversible cases) do not guarantee control of test size, whereas PROC GLM analyses *do* control the test size. So, for situations that don't require the generality of a mixed model, you can avoid inflated test sizes by using PROC GLM instead of PROC MIXED to perform the data analysis.

SPECIFYING CORRELATIONS

One of the most important and most difficult aspects of power analysis for multivariate linear models is conjecturing the covariance matrix of within-subject measurements. The GLMPOWER procedure supports several flavors of correlation structures and enables you to specify the correlations and variances either separately or together as a covariance matrix.

The following subsections discuss three common types of correlation patterns that you can encounter in repeated measures.

The LEAR Model

For repeated measurements of a subject over time, the correlation often decays at a rate somewhere between compound symmetry (no decay) and first-order autoregressive ("AR(1)," fast decay). A useful family of correlation structures, called "linear exponent AR(1)" (or "LEAR") by Simpson et al. (2010), covers the entire spectrum between these two cases by using a simple two-parameter specification. The LEAR model also extends to even faster exponential decay, all the way up to the first-order moving average model ("MA(1)"). The LEAR correlation model is related to the spatial covariance structures in PROC MIXED (Simpson et al. 2010, Appendix A) and has a parameterization particularly suitable for repeated measures designs.

All three of the examples that are discussed starting in the section "EXAMPLE: TIME BY TREATMENT" on page 7 use the LEAR model.

Unstructured Correlation

For multiple-outcome situations (such as concentrations of several different chemicals in a blood test), the correlation pattern often has no particular structure. This is referred to as an "unstructured" covariance, and it requires all $p(p-1)/2$ variances and covariances between the p responses to be specified.

Multilevel Models

When you have multiple sources of correlation, the overall correlation matrix is the Kronecker product of the individual correlation matrices. This type of correlation structure is often called a "multilevel model." For example, you might measure several different outcomes over time at different sites. A multilevel model is discussed in the section "EXAMPLE: MULTILEVEL CORRELATION STRUCTURE" on page 12.

EXAMPLE: TIME BY TREATMENT

In the Prologue you discovered the danger of basing a power analysis on a simplified version of the planned data analysis. Now, with SAS/STAT 13.1 in hand, you are ready to do the *right* power analysis for the SASGlobalFlora study.

The exemplary data set is easy to create, consisting simply of the same numbers as in Table 1 along with the sample size allocation weights:

```
data WellnessMult;
   input Treatment $7-13 WellScore1 WellScore3 WellScore6 Alloc;
   datalines;
      Placebo 32 36 39 2
      SGF     35 40 48 1
;
```

Your planned PROC MIXED analysis is as follows:

```
proc mixed;
   class Treatment Time Subject;
   model WellScore = Treatment|Time / ddfm=kr;
   repeated Time / subject=Subject type=un;
run;
```

This is equivalent to the following PROC GLM analysis:

```
proc glm;
   class Treatment;
   model WellScore1 WellScore3 WellScore6 = Treatment;
   repeated Time;
run;
```

This analysis uses the Hotelling-Lawley trace statistic (Edwards et al. 2008), as discussed in the section "PROC MIXED Analyses" on page 6.

Your boss tells you she's quite familiar with the correlation patterns from previous similar studies, and they are well represented by a LEAR model that has a base correlation of 0.85 and a decay rate of 1 over one-month intervals. The correlations according to this LEAR model, shown in Table 3, decay more slowly than in an AR(1) pattern. An AR(1) pattern here has a decay a rate of 3, the difference between maximum and minimum distances between time points.

Table 3 Conjectured Correlation Matrix

	1	3	6
1	1	0.722	0.614
3	0.722	1	0.684
6	0.614	0.684	1

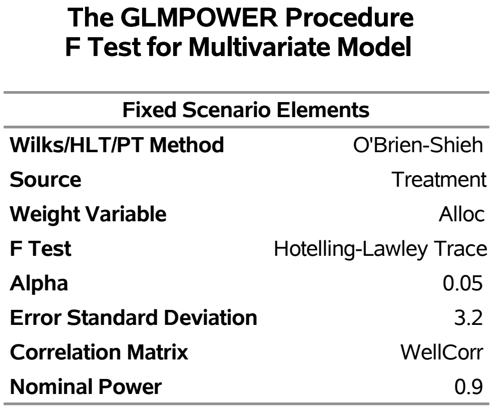
You are now ready to perform the power analysis for the model *actually planned for the data analysis*. Use the following statements to determine the number of subjects that are required to achieve a power of 0.9:

```
proc glmpower data=WellnessMult;
   class Treatment;
   weight Alloc;
   model WellScore1 WellScore3 WellScore6 = Treatment;
   repeated Time;
   power
      effects=(Treatment)
      mtest = hlt
      alpha = 0.05
      power = .9
      ntotal = .
      stddev = 3.2
      matrix ("WellCorr") = lear(0.85, 1, 3, 1 3 6)
      corrmat = "WellCorr";
run;
```

Note that the first four statements after the PROC GLMPOWER statement exactly match the PROC GLM formulation of the analysis. In the POWER statement, the EFFECTS= option chooses which between-subject effects to include in the power analysis (in this case, only the **Treatment** effect, excluding the intercept). The MTEST=HLT option specifies the Hotelling-Lawley trace statistic. The MATRIX= option defines the LEAR correlation structure, and the CORRMAT= option identifies it for use in the power analysis. The parameters in the LEAR specification are the base correlation (0.85), correlation decay rate (1), number of time points (3), and time values (one, three, and six months). For more information about the syntax, see the section "Syntax: GLMPOWER Procedure" in the GLMPOWER procedure chapter of the *SAS/STAT User's Guide*.

The results in Figure 1 are consistent with your intern's simulation, showing that only 15 subjects are needed. "Source" in the output is the between-subjects effect, "Transformation" is the dependent variable transformation, and "Effect" is the combination of Source and Transformation. The reported sample size "N Total" is the fractional solution rounded up to the nearest multiple of the allocation weight sum of 3 to ensure integer group sizes.

Figure 1 Sample Size Determination Assuming Multivariate Model



```
              The GLMPOWER Procedure
           F Test for Multivariate Model

         Fixed Scenario Elements
Wilks/HLT/PT Method               O'Brien-Shieh
Source                                Treatment
Weight Variable                           Alloc
F Test                    Hotelling-Lawley Trace
Alpha                                      0.05
Error Standard Deviation                    3.2
Correlation Matrix                     WellCorr
Nominal Power                               0.9
```

Figure 1 *continued*

Computed N Total

Index	Transformation	Effect	Num DF	Den DF	Actual Power	N Total
1	Time	Time*Treatment	2	12	0.918	15
2	Mean(Dep)	Treatment	1	16	0.942	18

Now you're curious what the power is for 30 subjects, the sample size you very nearly suggested to your boss. You submit the following statements to find out, and also to see the power curve:

```
ods graphics on;

proc glmpower data=WellnessMult;
   class Treatment;
   weight Alloc;
   model WellScore1 WellScore3 WellScore6 = Treatment;
   repeated Time;
   power
      effects=(Treatment)
      mtest = hlt
      alpha = 0.05
      ntotal = 30
      power = .
      stddev = 3.2
      matrix ("WellCorr") = lear(0.85, 1, 3, 1 3 6)
      cormat = "WellCorr";
   plot x=n min=6 step=3
        vary(symbol by stddev, panel by dependent source)
        xopts=(ref=15 30 crossref=yes);
run;

ods graphics off;
```

This is all the same as before, with these exceptions:

- The NTOTAL=30 value is specified, and the POWER= option value is left missing and thus to be computed.
- You've added a PLOT statement to see how the power depends on the sample size.

The output in Figure 2 and Figure 3 reveals that the power with 30 subjects is practically 100%!

Figure 2 Sample Size Determination Assuming Multivariate Model

The GLMPOWER Procedure
F Test for Multivariate Model

Fixed Scenario Elements

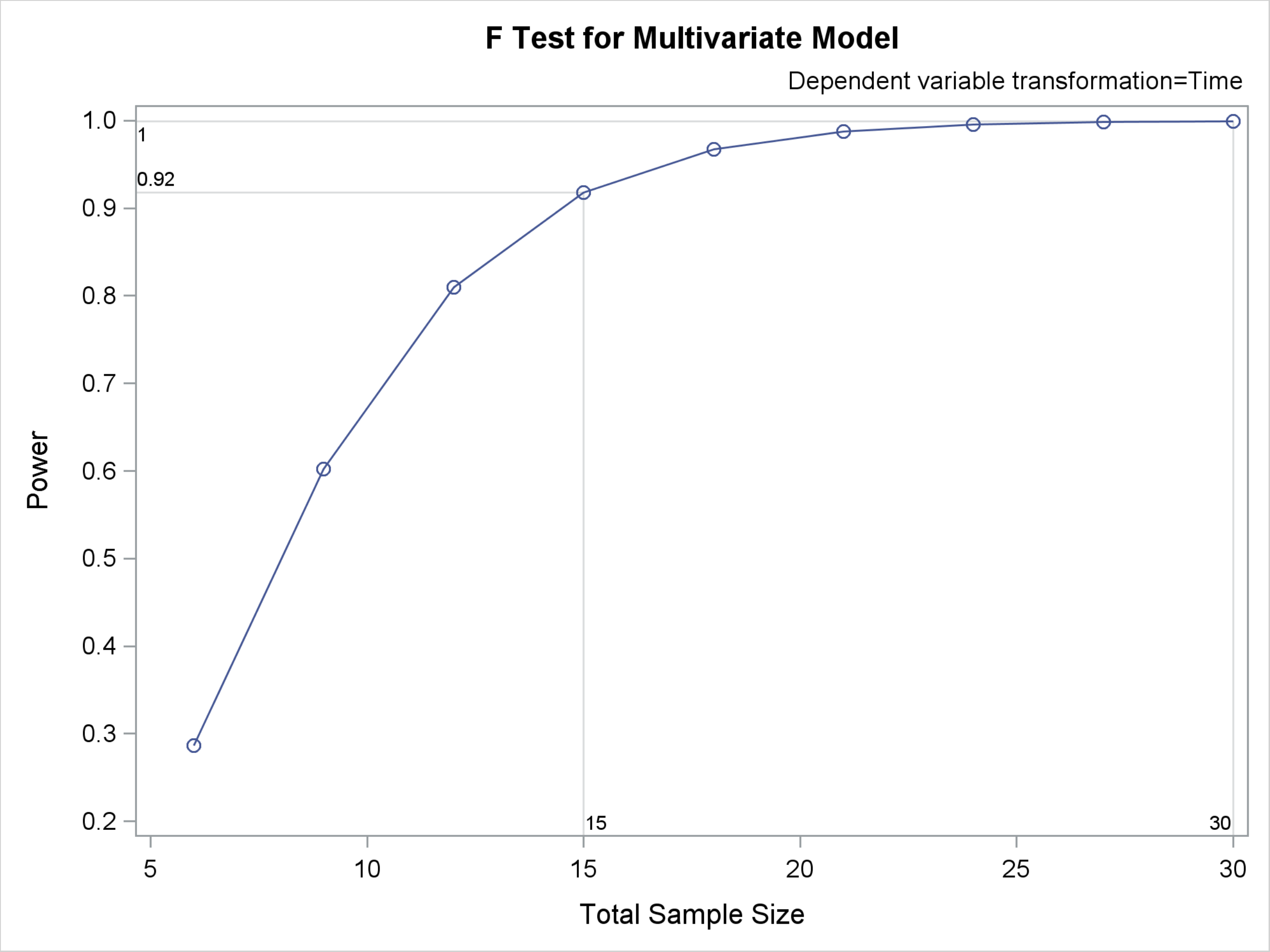
Wilks/HLT/PT Method	O'Brien-Shieh
Source	Treatment
Weight Variable	Alloc
F Test	Hotelling-Lawley Trace
Alpha	0.05
Error Standard Deviation	3.2
Correlation Matrix	WellCorr
Total Sample Size	30

Figure 2 *continued*

Computed Power

Index	Transformation	Effect	Num DF	Den DF	Power
1	Time	Time*Treatment	2	27	>.999
2	Mean(Dep)	Treatment	1	28	0.997

Figure 3 Plot of Power versus Sample Size for Repeated Measures Analysis



EXAMPLE: CLUSTERED DATA

Your boss is excited to hear that you can now do a power analysis that is properly aligned with the data analysis for the classic time-by-treatment repeated measures situation. “But I’m not sure if we’ll be able to run the study as planned,” she says. “We’re being pressured by upper management to finish it within three months instead of six and to assess the wellness scores only once, at the end. How many subjects will we need if we go with this ‘Plan B’ for the study? We know that in losing the benefit of the multiple time points, we’ll need more subjects to achieve the same power. So we’ll split them up among multiple doctors, assigning nine subjects to each (six on placebo and three on SGF). How many doctors will we need to get a power of 0.9?”

She also tells you that the planned data analysis for “Plan B” is a Type III F test for treatment in a compound symmetry model in PROC MIXED. The corresponding SAS statements are as follows:

```
proc mixed;
   class Treatment Doctor;
   model WellScore = Treatment;
   repeated / subject=Doctor type=cs;
run;
```

You need one more piece of information to do the power analysis. You ask your boss for an educated guess of the correlation between subjects under the same doctor (intraclass correlation), and she posits a value of 0.25.

First, you construct the exemplary data set, with the multivariate structure that you would use with PROC GLM. You do this by extracting the three-month means from the WellnessMult exemplary data set from the section “EXAMPLE: TIME BY TREATMENT” on page 7 and creating nine copies of them (one for each doctor), as in the following SAS statements:

```
data WellnessMult;
   input Treatment $7-13 WellScore1 WellScore3 WellScore6 Alloc;
   datalines;
      Placebo 32 36 39 2
      SGF     35 40 48 1
;
%let nSubjPerDoc = 9;
data WellnessClus;
   set WellnessMult;
   array WS{&nSubjPerDoc} WS1-WS&nSubjPerDoc;
   do iSubj = 1 to &nSubjPerDoc;
      WS{iSubj} = WellScore3;
   end;
   keep Treatment WS1-WS&nSubjPerDoc Alloc;
run;
```

Note that the preceding mixed model is equivalent to the following PROC GLM analysis of data that string the wellness scores out for each subject as multiple responses, **WS1**, . . . , **WS9**:

```
proc glm;
   class Treatment;
   model WS1-WS9 = Treatment;
   repeated Subject;
run;
```

This PROC GLM analysis uses the uncorrected univariate F test (Gurka, Edwards, and Muller 2011), as discussed in the section “PROC MIXED Analyses” on page 6. Note that in PROC MIXED it is the variable **Doctor** that plays the role of the “subject” in syntax, because that is the level at which the repeated measures occur (multiple subjects per **Doctor**). The REPEATED statement in PROC GLM, on the other hand, names the factor whose levels represent the repeated measurements themselves (**Subject** within doctor).

This power analysis calls for the NFRACTIONAL option in the POWER statement of PROC GLMPOWER, because the calculated sample size is the number of *doctors*. You don’t want the default behavior without the NFRACTIONAL option, which is the rounding of the number of doctors to the nearest multiple of the allocation weight sum of $2 + 1 = 3$.

Because compound symmetry is a special case of the LEAR model with a zero decay parameter, you use the LEAR option to specify the correlation structure.

The following statements compute the required sample size for the “Plan B” clustered data analysis:

```
proc glmpower data=WellnessClus;
   class Treatment;
   weight Alloc;
   model WS1-WS9 = Treatment;
   repeated Subject;
   power
```

```
         nfractional
         effects=(Treatment)
         mtest = uncorr
         alpha = 0.05
         power = .9
         ntotal = .
         stddev = 3.2
         matrix ("WellCorr") = lear(0.25, 0)
         corrmat = "WellCorr";
   run;
```

The results in Figure 4 reveal that you need 13 doctors (and thus a total of $13 \times 9 = 117$ subjects) to achieve the target power of 0.9.

Figure 4 Sample Size Determination for Clustered Data Analysis

The GLMPOWER Procedure
F Test for Multivariate Model

Fixed Scenario Elements	
Source	Treatment
Weight Variable	Alloc
F Test	Uncorrected Univariate Repeated
Nominal Alpha	0.05
Error Standard Deviation	3.2
Correlation Matrix	WellCorr
Nominal Power	0.9

Computed Ceiling N Total

								Ceiling	
			Num	Den	Actual	Fractional	Actual	N	
Index	**Transformation**	**Effect**	**DF**	**DF**	**Alpha**	**N Total**	**Power**	**Total**	**Error**
1	Subject	Subject*Treatment	.	.	.	.	.	.	Invalid input
2	Mean(Dep)	Treatment	1	11	0.05	12.365981	0.917	13	

You can safely ignore the first row of the “Computed Ceiling N Total” table in Figure 4 because you are not interested in testing the subject-by-treatment interaction. The “Invalid input” result stems from the lack of any interaction effect in the exemplary data set (thus making it impossible to achieve the target power for the interaction test by using any sample size).

You report the need for 13 doctors and 117 subjects in “Plan B” to upper management, and you persuade them to go with the original six-month treatment-over-time study plan instead. They didn’t expect the sample size requirement to jump so dramatically, but you explain that the expected treatment difference is greater at six months than at three, and the repeated measurements also increase the power by reducing variance estimates. “Plan B” loses both these benefits of the original plan.

EXAMPLE: MULTILEVEL CORRELATION STRUCTURE

Several months later your boss announces that it’s time to plan another study of the SASGlobalFlora herbal treatment. Your company has established an improved wellness scoring system, separating the wellness score into two dimensions, physical and mental. Your boss wants to know how this new scoring system affects the required sample size for a study like the one for which you conducted the power analysis in the section “EXAMPLE: TIME BY TREATMENT” on page 7, except that in this case you will incorporate both physical and mental wellness scores.

She expects better standard deviations than in the earlier study: 2.4 for the physical score and 2.8 for the mental score, with a correlation of about 0.4. And she expects a greater benefit for mental wellness than for physical wellness and posits the mean scores for SGF and placebo as shown in Table 4.

Table 4 Conjectured Physical and Mental Wellness Score Means by Treatment and Time

	Time					
	1 Month		3 Months		6 Months	
	Wellness Dimension					
Treatment	**Physical**	**Mental**	**Physical**	**Mental**	**Physical**	**Mental**
Placebo	33	31	36	36	37	41
SGF	36	34	40	40	46	50

Assuming that her conjectures for the means and variances are reasonable, has this effort to revamp the wellness scoring system produced a more efficient study, requiring fewer subjects to achieve the same level of power? The lower variability is likely to increase the power, but the extra test degrees of freedom and the correlation between the two wellness dimensions are likely to lower it. Which change wins?

The planned data analysis is similar to that in the section "EXAMPLE: TIME BY TREATMENT" on page 7 except for the addition of the classification variable **Dim** to indicate the dimension of the wellness score (physical or mental). The PROC MIXED statements for the doubly multivariate analysis are as follows:

```
proc mixed;
   class Treatment Time Dim Subject;
   model WellScore = Treatment|Time|Dim / ddfm=kr;
   repeated Time Dim / subject=Subject type=un@un;
run;
```

The exemplary data set for use with PROC GLMPOWER is constructed as follows:

```
data Wellness2Mult;
   input Treatment $7-13 WS1P WS1M WS3P WS3M WS6P WS6M Alloc;
   datalines;
      Placebo 33 31 36 36 37 41 2
      SGF     36 34 40 40 46 50 1
;
```

Again, the preceding mixed model is equivalent to the following PROC GLM analysis of data that string the wellness scores out for each subject as multiple responses, **WS1P**, . . . , **WS6M**:

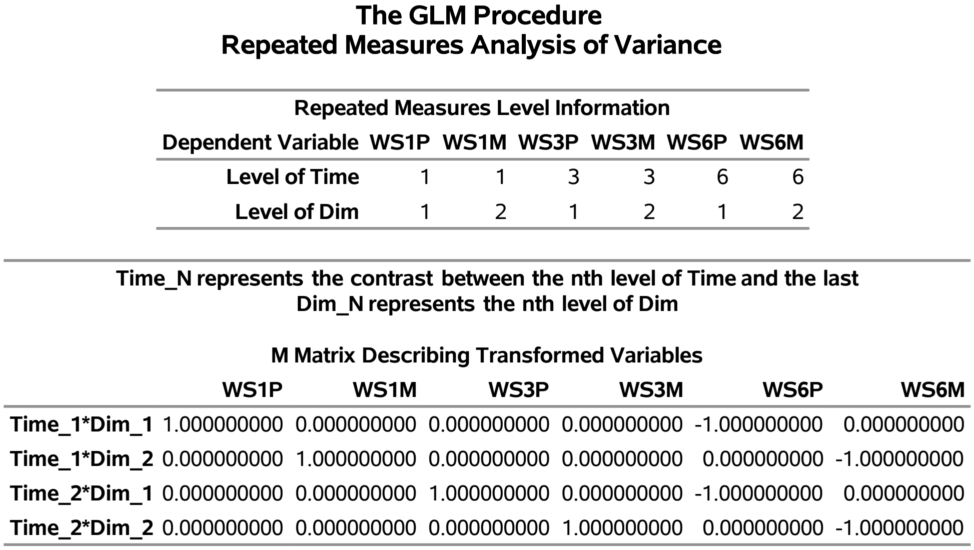
```
proc glm;
   class Treatment;
   model WS1P WS1M WS3P WS3M WS6P WS6M = Treatment;
   repeated Time 3 (1 3 6), Dim 2 identity;
run;
```

In SAS/STAT 13.1, you can specify only one repeated factor in the REPEATED statement in PROC GLMPOWER. But you can use the MANOVA statement instead to accomplish the same thing. Start by getting the GLM procedure itself to compute the appropriate $\mathbf{M}$ matrix by using the following statements:

```
proc glm data=Wellness2Mult;
   class Treatment;
   weight Alloc;
   model WS1P WS1M WS3P WS3M WS6P WS6M = Treatment / nouni;
   repeated Time 3 (1 3 6), Dim 2 identity / printm;
run;
```

The dependent variable transformation and the $\mathbf{M}$ matrix coefficients are shown in Figure 5. Note that you could use ODS to put the coefficients into a data set and unpack it again by using a macro, so that printing the matrix and parroting it by hand back as syntax isn't actually necessary.

Figure 5 M Matrix from PROC GLM



```
                    The GLM Procedure
          Repeated Measures Analysis of Variance

              Repeated Measures Level Information
   Dependent Variable  WS1P  WS1M  WS3P  WS3M  WS6P  WS6M
        Level of Time     1     1     3     3     6     6
         Level of Dim     1     2     1     2     1     2

  Time_N represents the contrast between the nth level of Time and the last
                   Dim_N represents the nth level of Dim

                 M Matrix Describing Transformed Variables
                      WS1P         WS1M         WS3P         WS3M          WS6P          WS6M
Time_1*Dim_1   1.000000000  0.000000000  0.000000000  0.000000000  -1.000000000   0.000000000
Time_1*Dim_2   0.000000000  1.000000000  0.000000000  0.000000000   0.000000000  -1.000000000
Time_2*Dim_1   0.000000000  0.000000000  1.000000000  0.000000000  -1.000000000   0.000000000
Time_2*Dim_2   0.000000000  0.000000000  0.000000000  1.000000000   0.000000000  -1.000000000
```

You can specify the correlation structures for **Time** and **Dim** separately and use the Kronecker product operator (@) in the MATRIX= option in the POWER statement of PROC GLMPOWER to build the overall within-subject correlation matrix, as follows:

```
matrix ("TimeCorr") = lear(0.85, 1, 3, 1 3 6)
matrix ("DimCorr") = lear(0.4, 0, 2)
matrix ("WellCorr") = "TimeCorr" @ "DimCorr"
corrmat = "WellCorr"
```

The resulting overall correlation matrix is shown in Table 5.

Table 5 Conjectured Correlation Matrix across Time and Wellness Score Dimensions

		1		3		6	
		Physical	**Mental**	**Physical**	**Mental**	**Physical**	**Mental**
1	**Physical**	1	0.4	0.722	0.289	0.614	0.246
	Mental	0.4	1	0.289	0.722	0.246	0.614
3	**Physical**	0.722	0.289	1	0.4	0.684	0.274
	Mental	0.289	0.722	0.4	1	0.274	0.684
6	**Physical**	0.614	0.246	0.684	0.274	1	0.4
	Mental	0.246	0.614	0.274	0.684	0.4	1

You can use the SQRTVAR= option in the POWER statement to specify the standard deviations, as follows:

```
matrix ("StdDevs") = (2.4 2.8 2.4 2.8 2.4 2.8)
sqrtvar = "StdDevs"
```

The resulting covariance matrix $\mathbf{\Sigma}$, shown in Table 6, is computed by PROC GLMPOWER as

$$\mathbf{\Sigma} = \left(\boldsymbol{\sigma}\boldsymbol{\sigma}'\right) \circ (\mathbf{R}_t \otimes \mathbf{R}_d)$$

where $\boldsymbol{\sigma}$ is the vector of standard deviations, $\mathbf{R}_t$ is the correlation matrix across time, $\mathbf{R}_d$ is the correlation matrix across dimension, "$\circ$" denotes elementwise (Hadamard) product, and "$\otimes$" denotes Kronecker product.

Table 6 Conjectured Covariance Matrix across Time and Wellness Score Dimensions

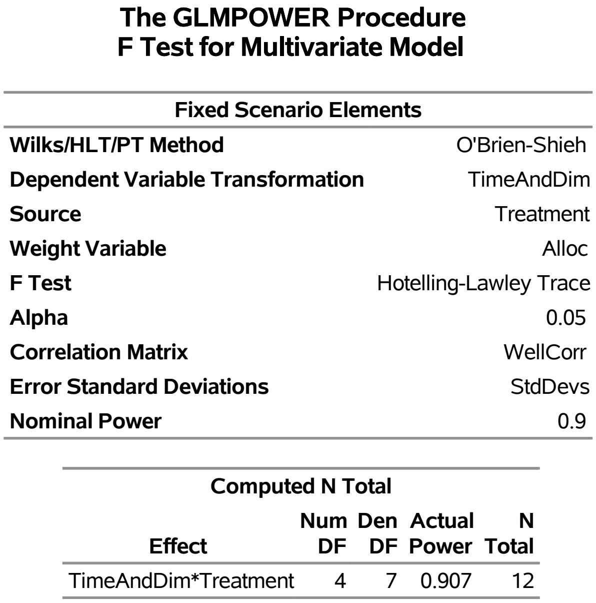
		1		3		6	
		Physical	**Mental**	**Physical**	**Mental**	**Physical**	**Mental**
1	**Physical**	5.76	2.69	4.16	1.94	3.54	1.65
	Mental	2.69	7.84	1.94	5.66	1.65	4.81
3	**Physical**	4.16	1.94	5.76	2.69	3.94	1.84
	Mental	1.94	5.66	2.69	7.84	1.84	5.37
6	**Physical**	3.54	1.65	3.94	1.84	5.76	2.69
	Mental	1.65	4.81	1.84	5.37	2.69	7.84

You now have all the ingredients to calculate the required sample size in PROC POWER:

```
proc glmpower data=Wellness2Mult;
   class Treatment;
   weight Alloc;
   model WS1P WS1M WS3P WS3M WS6P WS6M = Treatment;
   manova "TimeAndDim" M=(1 0 0 0 -1  0,
                          0 1 0 0  0 -1,
                          0 0 1 0 -1  0,
                          0 0 0 1  0 -1);
   power
      effects=(Treatment)
      mtest = hlt
      alpha = 0.05
      power = .9
      ntotal = .
      matrix ("TimeCorr") = lear(0.85, 1, 3, 1 3 6)
      matrix ("DimCorr") = lear(0.4, 0, 2)
      matrix ("WellCorr") = "TimeCorr" @ "DimCorr"
      matrix ("StdDevs") = (2.4 2.8 2.4 2.8 2.4 2.8)
      corrmat = "WellCorr"
      sqrtvar = "StdDevs";
run;
```

The results in Figure 6 show that you need only 12 subjects to achieve the desired power of 0.9.

Figure 6 Sample Size Determination for Multilevel Model



```
                   The GLMPOWER Procedure
                F Test for Multivariate Model

                  Fixed Scenario Elements
Wilks/HLT/PT Method                          O'Brien-Shieh
Dependent Variable Transformation              TimeAndDim
Source                                          Treatment
Weight Variable                                     Alloc
F Test                              Hotelling-Lawley Trace
Alpha                                                0.05
Correlation Matrix                               WellCorr
Error Standard Deviations                         StdDevs
Nominal Power                                         0.9

                     Computed N Total
                          Num  Den  Actual      N
          Effect           DF   DF   Power  Total
   TimeAndDim*Treatment     4    7   0.907     12
```

So the new study does appear to be slightly more efficient than the study that you analyzed in the section

“EXAMPLE: TIME BY TREATMENT” on page 7, which requires 15 subjects to attain the same level of power.

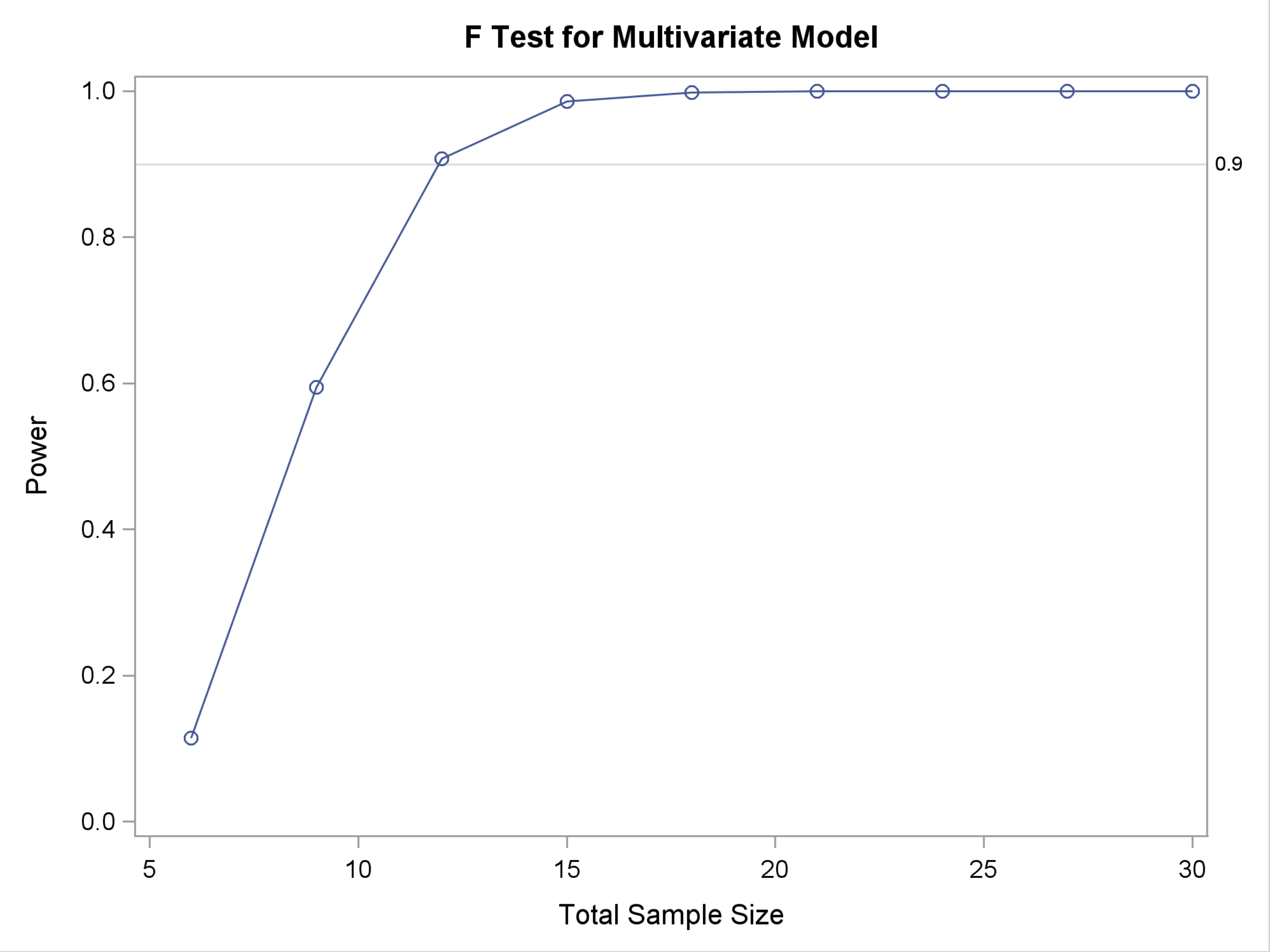
Finally, you use the following statements to produce a power curve for comparison with Figure 2:

```
ods graphics on;

proc glmpower data=Wellness2Mult plotonly;
   class Treatment;
   weight Alloc;
   model WS1P WS1M WS3P WS3M WS6P WS6M = Treatment;
   manova "TimeAndDim" M=(1 0 0 0 -1  0,
                          0 1 0 0  0 -1,
                          0 0 1 0 -1  0,
                          0 0 0 1  0 -1);
   power
      effects=(Treatment)
      mtest = hlt
      alpha = 0.05
      power = .
      ntotal = 12
      matrix ("StdDevs") = (2.4 2.8 2.4 2.8 2.4 2.8)
      matrix ("TimeCorr") = lear(0.85, 1, 3, 1 3 6)
      matrix ("DimCorr") = lear(0.4, 0, 2)
      matrix ("WellCorr") = "TimeCorr" @ "DimCorr"
      corrmat = "WellCorr"
      sqrtvar = "StdDevs";
   plot x=n min=6 max=30 step=3 yopts=(ref=.9);
run;

ods graphics off;
```

The PLOTONLY option in the PROC GLMPOWER statement suppresses tabular output, and X=N says to solve for power instead of sample size in order to produce a smoother curve; otherwise, the plotted points would be rounded sample size solutions, and the haphazardly varying degrees of rounding would be distracting. Other than those two revisions and the addition of the plot statement, the SAS code is identical to the previous PROC GLMPOWER code.

The resulting power curve is shown in Figure 7.

Figure 7 Plot of Power versus Sample Size for Multilevel Model



CONCLUSION AND FUTURE WORK

The methods that are discussed in this paper show how the new features for multivariate and repeated measures power analysis in PROC GLMPOWER in SAS/STAT 13.1 can help you design important classes of studies involving correlated measurements. In particular, for classic time-by-treatment repeated measures analysis, for situations involving clustered data, and for multiple outcomes, you can use the new MANOVA and REPEATED statements in PROC GLMPOWER to get just-right "Goldilocks" solutions to study resource problems. In pursuing such solutions, as the SASGlobalFlora herbal treatment examples demonstrate, you can use the LEAR structure to specify correlation behavior in a practical, parsimonious, and pliant way.

As mentioned before, in SAS/STAT 13.1 the REPEATED statement in PROC GLMPOWER accepts only a single repeated factor. This is not a huge practical limitation, because you can always use the MANOVA statement to roll your own analysis, but it does make the implementation in PROC GLMPOWER different from the familiar one in PROC GLM. Thus, you can expect this restriction to be lifted in future releases. Looking further down the road, there are related areas of ongoing research for which practical power analysis is desirable, including the following:

- accounting for data that are missing at random in a power analysis for multivariate linear models
- so-called reducible mixed model analyses, which are PROC MIXED analyses that aren't reversible—that is, they can't be equivalently fit using MANOVA—but can be projected onto similar PROC GLM analyses for reasonable power approximations

APPENDIX: A TALE OF TWO POWER ANALYSES

The details of the the misaligned power analysis that you performed in the Prologue are presented here in an appendix because they are unusually cumbersome. But they deserve at least appendix status because they demonstrate how to handle the situation where the "sample size" you're trying to compute is not actually the number of independent sampling units, but rather the number of levels of a factor.

Your intern incorporated the data analysis that was planned for the SASGlobalFlora study into his simulation, exemplified by the following statements:

```
proc mixed;
   class Treatment Time Subject;
   model WellScore = Treatment|Time / ddfm=kr;
   repeated Time / subject=Subject type=un;
run;
```

You based your first power analysis, however, on the standard F test in a univariate model without any within-subject correlation, as represented by the following statements:

```
proc mixed;
   class Treatment Time Subject;
   model WellScore = Treatment|Time Subject / ddfm=satterthwaite;
run;
```

This is equivalent to the following analysis in PROC GLM:

```
proc glm;
   class Treatment Time Subject;
   model WellScore = Treatment|Time Subject;
run;
```

So why exactly does this make your power analysis cumbersome? In the univariate model, **Subject** is no longer the independent sampling unit, but rather a model factor that affects the model degrees of freedom. As such, its number of levels affects the structure of the exemplary data set.

Although the GLMPOWER procedure isn't equipped to *directly* solve for the number of levels of a classification variable required to attain a certain power, you can solve for it *indirectly* by computing power iteratively over different exemplary data sets (constructed with different inherent numbers of subjects).

To build your exemplary data sets for the univariate power analysis, you used a macro variable called **nRep** for the number of repetitions of two placebo subjects plus one SGF subject. You started out with the same WellnessMult exemplary data set that you used for the multivariate power analysis in the section "EXAMPLE: TIME BY TREATMENT" on page 7, transposed it to univariate structure, and replicated it nRep times.

Then you repeated these steps for different values of nRep until you found a value that minimally attains the target power. You determined that nRep=10 (or 30 subjects) accomplishes this goal, and you ran the following statements in your final iteration:

```
data WellnessMult;
   input Treatment $7-13 WellScore1 WellScore3 WellScore6 Alloc;
   datalines;
      Placebo 32 36 39 2
      SGF     35 40 48 1
;
%let nRep=10;
data WellnessUni;
   set WellnessMult;
   retain Subject 1;
   do iRep = 1 to &nRep;
      do iSubj = 1 to Alloc;
         Time = 1; WellScore = WellScore1; output;
         Time = 3; WellScore = WellScore3; output;
         Time = 6; WellScore = WellScore6; output;
         Subject = Subject + 1;
      end;
   end;
   keep Treatment Time Subject WellScore;
run;
```

To compute the power, you set the NTOTAL option in the POWER statement of PROC GLMPOWER to your nRep value times 3 (to obtain the number of subjects) times 3 (the number of time points). You used the following statements:

```
   proc glmpower data=WellnessUni;
      class Treatment Time Subject;
      model WellScore = Treatment|Time Subject;
      power
         effects=(Treatment*Time)
         alpha = 0.05
         ntotal = %sysevalf(&nRep * (2+1) * 3)
         power = .
         stddev = 3.2;
   run;
```

The results (Figure 8) show that nRep=10 (that is, 30 subjects and NTOTAL=90) just attains the target power of 0.9.

Figure 8 Power Analysis Assuming Univariate Model with 30 Subjects

The GLMPOWER Procedure

Fixed Scenario Elements	
Dependent Variable	WellScore
Source	Treatment*Time
Alpha	0.05
Error Standard Deviation	3.2
Total Sample Size	90
Test Degrees of Freedom	2
Error Degrees of Freedom	56

Computed Power
Power
0.902

By comparison, your intern's simulation and your multivariate power analysis in the section "EXAMPLE: TIME BY TREATMENT" on page 7 concluded that only 15 subjects are needed.

REFERENCES

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., and Schabenberger, O. (2008), "An R-Square Statistic for Fixed Effects in the Linear Mixed Model," *Statistics in Medicine*, 27, 6137–6157.

Gurka, M. J., Edwards, L. J., and Muller, K. E. (2011), "Avoiding Bias in Mixed Model Inference for Fixed Effects," *Statistics in Medicine*, 30, 2696–2707.

Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," *American Statistician*, 55, 187–193.

Maldonado-Molina, M. M., Barón, A. E., and Kreidler, S. M. (2013), "Finding Power and Sample Size for Mixed Models in Study Designs with Repeated Measures and Clustering," Annual Meeting of the Society of Behavioral Medicine.

Muller, K. E., Barón, A. E., Kreidler, S. M., Chi, Y. Y., and Glueck, D. H. (2013), "Easy Power and Sample Size for Most of the Mixed Models You Will Ever See," Association of Clinical and Translational Statisticians (ACTS).

Muller, K. E. and Barton, C. N. (1989), “Approximate Power for Repeated-Measures ANOVA Lacking Sphericity,” *Journal of the American Statistical Association*, 84, 549–555, also see “Correction to *Approximate Power for Repeated-Measures ANOVA Lacking Sphericity*,” *Journal of the American Statistical Association* (1991), 86:255–256.

Muller, K. E. and Benignus, V. A. (1992), “Increasing Scientific Power with Statistical Power,” *Neurotoxicology and Teratology*, 14, 211–219.

Muller, K. E., Edwards, L. J., Simpson, S. L., and Taylor, D. J. (2007), “Statistical Tests with Accurate Size and Power for Balanced Linear Mixed Models,” *Statistics in Medicine*, 26, 3639–3660.

Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992), “Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications,” *Journal of the American Statistical Association*, 87, 1209–1226.

Muller, K. E. and Peterson, B. L. (1984), “Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis,” *Computational Statistics and Data Analysis*, 2, 143–158.

Muller, K. E. and Stewart, P. W. (2006), *Linear Model Theory: Univariate, Multivariate, and Mixed Models*, New York: Wiley-Interscience.

O’Brien, R. G. and Muller, K. E. (1993), “Unified Power Analysis for *t*-Tests through Multivariate Hypotheses,” in L. K. Edwards, ed., *Applied Analysis of Variance in Behavioral Science*, 297–344, New York: Marcel Dekker.

O’Brien, R. G. and Shieh, G. (1992), “Pragmatic, Unifying Algorithm Gives Power Probabilities for Common *F* Tests of the Multivariate General Linear Hypothesis,” Poster presented at the American Statistical Association Meetings, Boston, Statistical Computing Section.

Simpson, S. L., Edwards, L. J., Muller, K. E., Sen, P. K., and Styner, M. A. (2010), “A Linear Exponent AR(1) Family of Correlation Structures,” *Statistics in Medicine*, 29, 1825–1838.

ACKNOWLEDGMENTS

The author is grateful to Randy Tobias, Ed Huddleston, and Tim Arnold of the Advanced Analytics Division at SAS Institute Inc., and to Jill Tao and Kathleen Kiernan of the Technical Support Division at SAS, for their valuable assistance in the preparation of this manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

John Castelloe
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-5728
919-677-4444
john.castelloe@sas.com