

# A Monte Carlo Simulation Approach for Power Calculations in Repeated Measures Arising from Observational Studies: A SAS<sup>®</sup>/IML Application

Victor M. Gastañaga, Christine E. McLaren and Ralph J. Delfino

Epidemiology Division, College of Medicine, University of California, Irvine, CA 92697

## ABSTRACT

Repeated measurements arising from longitudinal studies occur frequently in epidemiologic studies. Methods to calculate power in the context of repeated measures are available for experimental settings where the covariate of interest is a discrete treatment indicator. However, no closed form expression exists to calculate power for generalized linear models with nonzero within-cluster correlation that are common in observational studies in which the covariate of interest varies over time and is often measured on a continuous scale, and where the researchers control for several potential confounders. We describe a SAS<sup>®</sup> implementation of a Monte Carlo simulation approach conducted to calculate power, and illustrate its application in a model frequently encountered in practice, the normal linear mixed model with repeated measurements and nonzero within-cluster correlation. This approach can be used to calculate the effect on power of changing various simulation conditions controlled by the researcher, such as sample size, within-cluster correlation structure, smallest meaningful difference to detect, and distributional assumptions.

## INTRODUCTION

Power and sample size calculations are carried out in the design phase of a research study to avoid choosing a sample size that is too large and costly, or too small and possibly resulting in a study with inadequate sensitivity. This paper presents a SAS<sup>®</sup> implementation of a Monte Carlo simulation approach developed to compute power for complex study designs for which no exact or approximate mathematical expressions exist. An observational (i.e., nonexperimental) study design with repeated measures and several confounders is a frequently observed complex study design for which no exact methods are available. We consider the context of repeated measures in observational studies for which the covariates of interest vary over time and are often measured on a continuous scale, and where all known factors that may affect disease risk must be controlled for in the analysis.

We completely specify a statistical model and data generation process that generates an outcome variable. First, predictor variables are generated so that their distribution mimics the sampling distribution of actual data, while model parameters are chosen so that the outcome variable also displays a distribution similar to that observed in actual data as reported in the peer-reviewed literature. Then the outcome variable is generated using this fully specified model. A linear mixed model is estimated with these data and relevant tests of hypotheses are conducted. The process is repeated a large number of times (as determined below) to ensure accuracy, and power is then computed based on frequencies associated with the test results.

## MODEL

Data are generated according to a linear mixed model with repeated measurements with the typical observation given by:

$$y_{ij} = \beta_0 + \sum_{k=1}^p \beta_k x_{kij} + \sum_{m=1}^g z_{mij} v_{mi} + \epsilon_{ij} \quad (1)$$

$i = 1, \dots, s; j = 1, \dots, n_i.$

In which there are  $p$  predictors,  $s$  subjects,  $n_i$  measurements for the  $i$ -th subject, and  $g$  random effects ( $v_{mi}$ ) associated with the  $i$ -th subject. Further,  $\epsilon_{ij}$  is an error term and the  $z_{mij}$  are known constants.

In matrix notation the model above is specified by:

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{v} + \boldsymbol{\epsilon} \quad (2)$$

in which  $\mathbf{y}$  is a column vector of dimension  $\sum n_i$ ,  $X$  is a  $(\sum n_i \times [p+1])$  design matrix,  $\boldsymbol{\beta}$  is a  $(p+1)$  vector of fixed effects, and  $\mathbf{v}$  is a column vector of dimension  $sg$  containing the random effects  $v_{mi}$ . Further,  $Z$  is a block diagonal  $(\sum n_i \times sg)$  between-subject design matrix, and  $\boldsymbol{\epsilon}$  is an  $\sum n_i$ -dimension column vector that contains the  $\epsilon_{ij}$  error terms.

The vector  $\boldsymbol{\epsilon}$  in (2) follows a multivariate normal distribution with means equal to zero, and variance covariance matrix  $R(\sum n_i \times \sum n_i)$ . In general, the  $R$  matrix is block diagonal with each of the  $s$  blocks a  $(n_i \times n_i)$  matrix corresponding to the  $i$ -th subject.

For this implementation we assume that each of the  $s$  blocks of  $R$  (the variance of  $\boldsymbol{\epsilon}$  in (2)) has an AR(1), covariance structure. The typical element of the  $i$ -th block of  $R$  is:

$$\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma_\epsilon^2 \rho^{|k-j|} \quad (3)$$

$i=1, \dots, s; j, k=1, \dots, n_i$

where  $\sigma_\epsilon^2$  is the variance associated with the error term  $\epsilon_{ij}$ , and  $\rho$  is the autoregressive coefficient, with  $|\rho| < 1$  for stationarity. A stationary process has constant mean and variance over time and the covariance between any two measurements will depend on their time difference only. Specifically, variance and covariance for an AR(1) process are undefined when  $|\rho| \geq 1$ .

This covariance structure for  $R$  allows for within-subject correlation that decreases proportionally with the time elapsed between repeated measurements. For this application  $\rho$  was drawn from a beta (2,3) distribution, which was chosen to ensure a value in the [0,1] range. A value for  $\sigma_\epsilon^2$  is determined below.

For this application we assume  $g=1$  (one random effect per cluster) and  $z_{mij}=1$  in equation (1), thus specifying a random individual intercept model. Furthermore, the  $s \times s$  matrix  $G$  (the variance of  $\mathbf{v}$  in (2)) is assumed to be:

$$G = \sigma_v^2 I_s \quad (4)$$

where  $\sigma_v^2$  is the variance associated with each random effect, and  $I_s$  an identity matrix of dimension  $s$ .

For the data generation process, values of  $X$ ,  $\boldsymbol{\beta}$ ,  $Z$ ,  $\mathbf{v}$ , and  $\boldsymbol{\epsilon}$  are created and then used to generate a vector  $\mathbf{y}$  as in equation (2). We now describe how these data are generated.

## EXPLANATORY VARIABLES

We consider one covariate of interest,  $x_1$ , and two potential confounders ( $x_2$  and  $x_3$ ), i.e., we set  $p=3$  in (1). We first generate individual ( $i$ -th subject) means for the  $k$ -th predictor,  $\mu_{ki}$ . We then use these means to generate within-cluster repeated measures,  $x_{kij}$ .

We generate subject-specific means in such a way as to replicate the first two moments of the distribution of observed between-subject data. The  $\mu_{ki}$  are generated as linear transformations of random variables with known mean and variance. For continuous variables we could use variates drawn from a standard beta distribution, which has desirable properties for simulation. Depending on the application, other suitable distributions may be used. In particular,  $\mu_{i1}$ , the cluster-specific means for  $x_1$  are computed as  $0.3738 + 85.4046 * b_{3,4}$ , where the  $b_{3,4}$  were random numbers drawn from a beta distribution with shape parameters 3 and 4. The shape parameters 3 and 4 are chosen because the mean-to-variance ratio matches that of the observed between-subject  $x_1$  data. Finally, the linear transformation used was chosen to yield a distribution that mimics the first two moments of the observed  $x_1$  distribution.

Similarly,  $\mu_{2i}$ , the subject-specific means for  $x_2$  are computed as  $-0.8006 + 90.5974 * b_{29,1}$ . Finally,  $\mu_{3i}$ , the between-subject means for  $x_3$ , a binary variable, are computed as  $0.0594 + 0.3451 * chsq(0.09)$ , where the  $chsq(0.09)$  are random numbers drawn from a chi-squared distribution with mean equal to 0.09. A chi-squared distribution is chosen because its range lies within  $[0, \infty)$ , and the degrees of freedom and linear transformation were chosen to mimic the first two moments of the observed distribution of  $x_3$ .

For binary explanatory variables, intracluster repeated measures are generated as realizations from a Bernoulli distribution with parameter  $\mu_{ki}$ .

For continuous variables we assume that intracluster repeated measures follow an autoregressive process. We focus on longitudinal studies, and therefore, within-cluster repeated measures are generated assuming an AR(1) process:

$$x_{kij} = \mu_{ki}(1 - \lambda_k) + \lambda_k x_{k,i,j-1} + w_{kij} \quad (5)$$

where  $\lambda_k$  is the autoregressive coefficient ( $|\lambda_k| < 1$ ) and  $w_{kij}$  is an independent and identically distributed error term drawn from a normal distribution with mean equal to zero and variance equal to  $\tau_{ki}^2$ .

It can be shown that (5) results in  $E(x_{kij}) = \mu_{ki}$  and

$$Var(x_{kij}) = \left( \frac{1}{1 - \lambda_k^2} \right) \tau_{ki}^2 \quad (6)$$

where  $Var(x_{kij})$  denotes within-subject variance.

The initial value of  $x_{kij}$ , say  $x_{ki0}$ , is set equal to zero and values of  $x_{ki1}$ ,  $x_{ki2}$ , etc. are generated according to equation (5). The first  $T$  values (say,  $T=20$ ) of  $x_{kij}$  are discarded to reduce any dependency of  $x_{kij}$  on the initial value, and the last  $n_i$  values of  $x_{kij}$  are retained for each subject.

Within-cluster repeated measures of  $x_3$  for the  $i$ -th subject are drawn from a Bernoulli distribution with mean  $\mu_{3i}$ . Within-subject repeated measures of  $x_1$  and  $x_2$  are generated according to an AR(1).

Choice of parameters. The autoregressive coefficients  $\lambda_k$  in (5) are drawn from a beta distribution. This ensured that the drawn parameters were bounded between 0 and 1, thereby prescribing positive autocorrelation, which we would expect for the variables under study, and ensuring stability of the autoregressive processes corresponding to the within-subject repeated measures  $x_{kij}$ . We set  $\tau_{ki}^2 = \tau_k^2$ , and determined  $\tau_k^2$  using equation (6) and replacing the

$Var(x_{kij})$  with sample within-subject variance computed from observed data.

## RESPONSE VARIABLE

Parameters other than those drawn from specified distributions are chosen to ensure that the generated outcome variable  $y_{ij}$  has mean and variance compatible to that of observed data, as described in this section.

We note the following relations. From equation (1) we have:

$$E(y_{ij}) = \beta_0 + \beta_k \sum_{k=1}^p E(x_{kij}) \quad (7)$$

The expected values of  $y_{ij}$  and  $x_{kij}$  can be replaced by their known sample means (table 1). Therefore,  $\beta_0$  can be determined in (7) for given values of the  $\beta_k$  ( $k=1,2,3$ ). Values of  $\beta_2$  and  $\beta_3$  can be arbitrarily assigned, as interest centers on  $\beta_1$ , the coefficient of  $x_1$ . To select these two values we consider the following expression, also derived from (1) assuming  $g=1$  and  $z_{mij}=1$ , where  $Var(\cdot)$  denotes between-cluster variance:

$$Var(y_{ij}) = \sum_{k=1}^p \beta_k^2 Var(x_{kij}) + \sigma_v^2 + \sigma_e^2 \quad (8)$$

The  $x_{kij}$  are generated for each of  $M$  replications, and therefore their variance is nonzero. Since the  $x_{kij}$  are generated so that their variance matches that of observed data, the  $Var(x_{kij})$  are known but values of  $\beta_k$ ,  $\sigma_v^2$ , and  $\sigma_e^2$  have to be selected. As can be seen in (8), the precision of the regression estimates can be made arbitrarily large by generating data such that the  $\beta_k^2 Var(x_{kij})$  are large relative to  $\sigma_v^2 + \sigma_e^2$ . We conservatively choose for  $\beta_2$  and  $\beta_3$  the smallest values that are meaningful for subject matter specialists. Simulations are run for different values of  $\beta_1$  to assess how small an effect of  $x_1$  on  $y$  could be detected while still retaining sufficient power. We compute  $\sigma_e^2$  as 60 percent of the residual variance  $D$  defined as  $D = Var(y_{ij}) - \sum_{k=1}^3 \beta_k^2 Var(x_{kij})$ .

Allocating more than half (such as 60 percent) of  $D$  to  $\sigma_e^2$  rather than  $\sigma_v^2$  is also conservative in the sense that the contribution of the fixed and random effects are not being made excessively large (relative to the error term variance), thus not arbitrarily increasing the precision of the regression estimates.

## DATA GENERATION AND POWER CALCULATION

For a given set of parameters  $\mu_{ki}$ ,  $\tau_{ki}^2$ , and  $\lambda_k$  the matrix  $X$  is generated  $M$  times. Vectors  $\mathbf{v}$  and  $\mathbf{e}$  are drawn from multivariate normal distributions with means equal to zero and variance covariance matrices  $G$  and  $R$ , respectively, as described by (4) and (3). For a given set of values of  $\sigma_e^2$ ,  $\rho$ , and  $\sigma_v^2$ , vectors  $\mathbf{e}$  and  $\mathbf{v}$  are generated  $M$  times. For each of the  $M$  sets of vectors  $\mathbf{e}$  and  $\mathbf{v}$  and matrices  $X$  and  $Z$ , we generate a vector  $\mathbf{y}$  according to (2).

A linear mixed model of  $y$ ,  $x_1$ ,  $x_2$ , and  $x_3$  is estimated by restricted maximum likelihood (REML) using PROC MIXED in SAS<sup>®</sup>/STAT, explicitly allowing for a subject-specific random effect and a covariance structure of type AR(1). Denominator degrees of freedom are calculated by the Kenward-Roger method. Note that it is possible to assess the sensitivity of power to incorrect specification of the covariance structure.

For each of  $M$  replications of data sets and estimation results, and assuming that  $x_{ij}$  is the covariate of interest, we test the false null hypothesis  $\beta_i=0$ , and note the  $p$ -value associated with a two-tailed test. Power, denoted  $P$ , is estimated as the frequency of rejection of the null hypothesis for a given significance level  $\alpha$ .

To determine  $M$ , consider that each Monte Carlo replication will generate a test statistic that either rejects or does not reject the false null hypothesis of no effect. Therefore, the replications can be considered as independent Bernoulli trials. If  $M$  replications are completed and  $Q$  rejections are observed, then  $P^* = Q/M$  is an estimate of  $P$ , with variance  $M^{-1}P(1-P)$ . Then we can use the normal approximation to the binomial to determine the number of replications needed to attain a 95% confidence interval of length  $c$ . For example, assuming  $P=0.9$  and  $c=0.04$ , we would need approximately 865 replications.

$M$  is set to 1000 and the values chosen for  $\beta_i$  are as shown in the source code below. Data are generated for 64 subjects with 10 repeated measurements. In each replication we test the false null hypothesis  $\beta_i=0$ , and note the  $p$ -value associated with a two-tailed test. Power is then computed as the frequency of rejection of the null hypothesis for significance level  $\alpha=0.05$ .

## SAS® CODE

This section presents the source code used to simulate data and compute power.

```
libname out "C:\Monte Carlo\ ";

** Define global values: No. of replications,
subjects, repeated measurements;
%let sim=200; %let sub=64; %let rep_m=10;
** Set parameters based on subject matter
knowledge;
%let beta1=-0.0007; *(True) slope of interest;
%let beta2=-0.005; *Partial slope for continuous
covariate;
%let beta3=-0.100; *Partial slope for binary
covariate;

%let seed_v=0; *Seed to draw random effects;
%let seed_e=0; *Seed to draw random error term;

** Response variable based on previous studies;
*between-ID mean, variance;
%let mean_y=0.699757638; %let var_y
=0.01121143872;
** Error Term;
%let res_var=0.60; *% of residual variance
assigned to error term vs random effect;
*Beta distr parameters m and n to draw AR(1)
coeff for error term;
%let rho_m=2; %let rho_n=3;

*** Parameters used to generate X1, the
predictor of interest;
%let mean_x1=36.9758; %let var_x1 =223.2838;
*Between-ID variance;
%let wn_var1=1.0; *Within-ID variance a fraction
of Between-ID variance;

* Subject-specific means for X1 will be a
```

```
linear transformation (=a+bx) ;
* of a beta(m,n) variate;
%let beta_m1=3; %let beta_n1=4;
*Coefficients for linear transformation;
%let a_beta1=0.373849; %let b_beta1=85.40455;

*Within-subject X1 data follow an AR(1)
process;
*AR coeff is rho1, drawn from a beta(m,n);
%let arbeta_m1=2; %let arbeta_n1=3;
%let burnin=-19; *No. of AR initial values to be
discarded is -burnin + 1.

*** Parameters used to generate X2, a continuous
covariate;
%let mean_x2=86.7769; %let var_x2 =8.5315;
%let wn_var2=3.5; *Within-ID variance a fraction
(or multiple) of Between-ID variance;

* Subject-specific means for X2 will be a
linear transformation (=a+bx) ;
* of a beta(m,n) variate;
%let beta_m2=29; %let beta_n2=1;
%let a_beta2=-0.8006; *Coefficients a and b for
linear transformation;
%let b_beta2=90.59741;

*Within-subject X2 data follow an AR(1)
process;
*AR coeff is rho2, drawn from a beta(m,n);
%let arbeta_m2=2;
%let arbeta_n2=3;

*** Parameters used to generate X3, a binary
covariate;
%let mean_x3=0.090415; %let var_x3 =0.021438;
* Subject-specific proportions (means) will be
a linear ;
* transformation (=a+bx) of a chisq(k)
variate;
%let chisq3=0.09; *Chi square parameter;
*Coefficients a and b for linear transformation;
%let a_chi3=0.059356; %let b_chi3=0.345109;

*** Macro to generate ID-specific means for X1,
X2;
%macro eta(k=1,a=0,b=1,m=2,n=3);
%do id=1 %to &sub;
eta&k&id=&a&b*rand('beta',%sysfunc(inputn(&m,3.
1)),%sysfunc(inputn(&n,3.1)) );
%end;
%mend eta;

*** Macro to generate ID-specific means for
binary predictor X3;
%macro eta3;
%do id=1 %to &sub;
eta3&id=&a_chi3+
&b_chi3*rand('chisq',&chisq3);
```

```

        if eta3&id gt 1.0 then eta3&id=1.0;
    %end;
%mend eta3;

*** Create scalar parameters common to all IDs;
data out.one;
    * Compute "true" intercept for the LINEAR link;
    beta0=&mean_y - &beta1*&mean_x1 -
&beta2*&mean_x2 - &beta3*&mean_x3;
    rho=rand('beta',&rho_m,&rho_n);
    lamb1=rand('beta',&beta_m1,&beta_n1);
*AR(1) coeff for predictor x1;
    lamb2=rand('beta',&beta_m2,&beta_n2);
*AR(1) coeff for cont. covariate x2;

    t1=sqrt( (&wn_var1*&var_x1)*(1-lamb1**2)); *Std
Dev for epsilon in AR(1) for predictor;
    t2=sqrt( (&wn_var2*&var_x2)*(1-lamb2**2)); *Std
Dev for epsilon in AR(1), cont. covariate;

    *Variance for the linear mixed model error
term;
    var_e=&res_var*(&var_y-(&beta1**2)*&var_x1-
(&beta2**2)*&var_x2-(&beta3**2)*&var_x3);
    if var_e le 0 then do; put "&errmsg";
error="&errmsg"; end;
    *Variance for the random effects;
    var_u=(1-&res_var)*(&var_y -
(&beta1**2)*&var_x1-(&beta2**2)*&var_x2-
(&beta3**2)*&var_x3);

    ** Next generate id-specific means for all
three predictors;

%eta(k=1,a=&a_beta1,b=&b_beta1,m=&beta_m1,n=&bet
a_n1)

%eta(k=2,a=&a_beta2,b=&b_beta2,m=&beta_m2,n=&bet
a_n2)
    %eta3
run;

*Macro to create data for all IDs/Subjects;
%macro mcdata;
    %do id=1 %to &sub;
        data mc&id(keep=id count x0 x1 x2 x3);
            set OUT.one;
            id=&id;
            do count=&burnin to &rep_m;
                retain x0 x1 x2 x3 0; * =0 for t=0;
                x0=1;
                x1=eta1&id*(1-lamb1) + lamb1*x1 +
(t1*normal(0));
                x2=eta2&id*(1-lamb2) + lamb2*x2 +
(t2*normal(0));
                x3=rand("bernoulli",eta3&id);
                if count>0 then output; *Discard
first -burnin+1 values;
            end;
        run;
    %end;

```

```

        proc append base=allmc data=mc&id force; run;
    %end;
%mend mcdata;

*Create Var-Cov matrix R for error term;
proc iml;
    use out.one;
    read all var{var_e} into sigesq;
    read all var{rho} into rho_e;
    B=j(&rep_m,&rep_m,1); *initialize AR(1) var-
cov for each ID;
    do i=1 to &rep_m;
        do j=i to &rep_m;
            B[i,j]=rho_e**abs(j-i);
            B[j,i]=B[i,j];
        end;
    end;
    R=sigesq#(I(&sub)@B);
    *Kronecker product above creates block diag
matrix;
    free B;
    create out.Rmatrix from R; *write out err var-
cov to SAS® dataset;
    append from R;
quit;

*This concludes the generation of global
parameters;
*Next we use these parameters to generate data M
times;

*** Monte Carlo Macro ***;
%macro m_carlo;
    %do rep=1 %to &sim;
        %mcdata
        *Create the dependent variable;
        proc iml;
            use out.one;
            read all var{beta0} into beta0;
            read all var{var_u} into sigusq;
            beta=j(4,1,0);
            beta[1,1]=beta0;

            beta[2,1]=&beta1;beta[3,1]=&beta2;beta[4,1]=&bet
a3;

            use allmc;
            read all var{x0 x1 x2 x3} into xvvars;
            read all var{id x1 x2 x3} into xdata;

            use out.Rmatrix;
            read all into R;

            y=xvvars*beta +
(I(&sub)@j(&rep_m,1,1))*
t(root(sigusq#I(&sub)))*normal(j(&sub,1,&seed_v)
))+
            t(root(R))*normal(j( (&sub*&rep_m)
,1,&seed_e));
            free R xvvars;
        quit;
    %end;

```

```

mix=y||xdata;
cname={"y_var" "ID" "x1" "x2" "x3"};
  create mix&rep from mix [ colname=cname ];
  append from mix;
quit;

** Estimated model with simulated data;
ods output solutionf=estout;
proc mixed data=mix&rep maxiter=1000;
  class ID;
  model y_var=x1 x2 x3 / s ddfm=kr;
  random int / subject=ID;
  repeated / type=ar(1) subject=id;
run; quit;

** Run t-test for the null hypothesis of no
effect and compute p-value;
data mixed;
  set estout(firstobs=2 obs=2);*keep first
slope estimate only;
  t=estimate/stderr; pvalue=2*(1-
cdf("t",abs(t),df));
  rep=&rep;
run;
proc append base=out.powerout data=mixed
force; run;

** Clear temp tables and output, log windows;
proc datasets library=work kill nolist; run;
quit;
dm "clear output"; dm "clear log";
%end;
%mend m_carlo;
%m_carlo

** Compute power as % of p-values under 0.05;
ods exclude ExtremeObs Quantiles;
proc univariate data=out.powerout mu0=.05
loccount;
  title "&head";
  var pvalue;
run;

```

## CONCLUSION

The SAS® implementation presented here allowed us to compute power for a relatively complex study for which no exact methods are available. Furthermore, this method inherently lends itself to a close alignment of power analysis and data analysis. Although the results of Monte Carlo experiments cannot be generalized to study designs other than the one being simulated, the approach is flexible enough to accommodate virtually any link function (in a GLIM framework), any number or type (continuous or discrete) of covariates, and random effects. By varying certain simulation conditions, the researcher can assess the effect on power of different sample sizes, within-cluster correlation structure, smallest meaningful difference to detect, significance level, number of predictors, the distribution assumed for the response and predictor variables, and whether the correlation structure is correctly specified or not.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the corresponding author at:

Victor M. Gastañaña, PhD  
Epidemiology Division, College of Medicine  
University of California, Irvine  
224 Irvine Hall  
Irvine CA 92697-7555  
Tel: 949-824-9861  
Fax: 949-824-1343  
Email: vgastana@uci.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.