

## SAS® Macro for calculating sample size and power on GLM models with correlated binary outcomes

Patricia R Houck, MSH, UPMC, University of Pittsburgh, Pittsburgh, PA

Qianyu Dang, PhD, Biostatistics and Medicine, University of Pittsburgh, Pittsburgh, PA

Sati Mazumdar, PhD, Biostatistics, University of Pittsburgh, Pittsburgh, PA

### ABSTRACT

SAS® PROC GLIMMIX has become a popular procedure to run generalized linear mixed models. One of the most useful applications is for binary outcomes in longitudinal studies. We have derived formulas for power and sample size calculations for longitudinal designs with attrition over time (Dang et al, in press). These estimates depend on the within-subject correlations and the size of random effects. SAS® macros have been created to calculate sample size (glimmixsamplesize) and power (glimmixpower). An application on a late-life psychiatric clinical trial is demonstrated.

### INTRODUCTION

Correlated binary outcome data are common in longitudinal clinical trials. The use of GLIMMIX has become a popular procedure in SAS to analyze such data. The generalized linear mixed model has more lenient assumptions than generalizing estimating equations (GEE) with regard to missing data. GEE assumes that the missingness mechanism is missing completely at random (MCAR). GLIMMIX assumes that the missing data is missing at random (MAR) which means that the missing data can depend on the observed values but not on the unobserved values.

Power and sample size calculations for GEE models have been developed (Pan, 2001) based upon the z-test and a robust variance estimator. Rochon (1998) developed a program (GEESIZE) that accounts for unequal allocation and loss to followup. More recently Dang and his colleagues proposed power and sample size calculations for mixed effect models (Dang, in press) using SAS GLIMMIX. They demonstrated good approximation with simulations. We have modified the method into a macro. We will briefly explain the method and apply the macro to a late-life psychiatric clinical trial.

### METHOD

The macro is based upon a method proposed by Dang et al, for calculating power and sample size for GLIMMIX models. The method is designed for a two-group comparison with either compound symmetry (CS) or autoregressive (AR(1)) within-subject correlation structure.

The general form of the GLIMMIX model with a link function  $g$  can be written as

$$g(E(Y_{ij} | \gamma_i)) = X_i \beta + Z_i \gamma_i \quad (1)$$

where  $Y_{ij}$  denotes the outcome for subject  $i$  at time  $j$ , where  $X_i$  denotes the design matrix for fixed effects, and where  $Z_i$  denotes the design matrix for random effects.

In the example we present here, we assume the treatment group to be the fixed effect and each subject within the group to be the random effect. The purpose is to test if the binary responses of the treatment group are better than those of the control group. The design matrices for fixed and random effects for subject  $i$  are:

Control group:  $X_i = [1_{ni} \ 0_{ni}]$ ,  $Z = 1_{ni}$  ;  
 Treatment group:  $X_i = [1_{ni} \ 1_{ni}]$ ,  $Z = 1_{ni}$  .

The null hypothesis is  $H_0: \beta = 0$ , and the alternative hypothesis is  $H_a: \beta = b > 0$ . The test for a hypothesis such as this is a Wald-type test based on asymptotically normal distributions as in the marginal model (Pan, 2001). Thus, for given sample size  $N$ , the power  $1 - \Delta$  can be obtained by

$$1 - \Delta \approx 1 - \Phi \left( Z_{\alpha/2} - b \sqrt{\frac{N}{\text{VAR}(\hat{\beta})}} \right), \quad (2)$$

and the required sample size  $N$  for a given power  $1 - \Delta$  is

$$N = \frac{\text{VAR}(\hat{\beta})(Z_{\alpha/2} - Z_{1-\Delta})^2}{b^2}. \quad (3)$$

We denoted  $\pi$  as the sample allocation ratio and denoted  $p_0$  and  $p_1$  as the marginal average rates for control and treatment groups. The variance estimator is a function of  $\pi$ ,  $p_0$ ,  $p_1$ ,  $R^*$  and  $\sigma_i^2$ :

$$\text{VAR}(\hat{\beta}) = \left( \sum_{i=1}^N \frac{1'_{n_i} R^{*-1} 1_{n_i} M}{\sigma_i^2} \right)^{-1} = \left( \sum_{i=1}^N \frac{1'_{n_i} R^{*-1} 1_{n_i}}{\sigma_i^2} \right)^{-1} M^{-1}, \quad (4)$$

where

$$M^{-1} = \frac{1}{\pi p_0(1-p_0)} \begin{bmatrix} 1 & \\ -1 & 1 + \frac{\pi p_0(1-p_0)}{(1-\pi)p_1(1-p_1)} \end{bmatrix}. \quad (5)$$

When  $\text{VAR}(\hat{\beta})$  becomes available, we can plug it into equation (2) or (3) to obtain the estimated power or sample size.

In the special case in which the  $R$  matrix has a CS covariance structure with  $\rho_0^*$  and  $\rho_1^*$  as the components in  $R^*$  and with  $\sigma_0$  and  $\sigma_1$  as the combined variance parameter mentioned above for the control and treatment groups, the variance estimation formula is

$$\text{VAR}(\hat{\beta}) = \left( \sum_{N_0} \frac{n_i}{1 + (n_i - 1)\rho_0^*\sigma_0^2} + \sum_{N_1} \frac{n_i}{1 + (n_i - 1)\rho_1^*\sigma_1^2} \right)^{-1} \left[ \frac{1}{\pi p_0(1-p_0)} + \frac{1}{(1-\pi)p_1(1-p_1)} \right]. \quad (6)$$

But if the  $R$  matrix has an AR(1) covariance structure, the formula becomes

$$\text{VAR}(\hat{\beta}) = \left( \sum_{N_0} \frac{n_i - (n_i - 2)\rho_0^*}{(1 + \rho_0^*)\sigma_0^2} + \sum_{N_1} \frac{n_i - (n_i - 2)\rho_1^*}{(1 + \rho_1^*)\sigma_1^2} \right)^{-1} \left[ \frac{1}{\pi p_0(1-p_0)} + \frac{1}{(1-\pi)p_1(1-p_1)} \right]. \quad (7)$$

In both cases,  $N_0$  and  $N_1$  is the sample size in control and treatment group.

All of the formulas described above assume that the sample size is constant across time. In practice, however, this assumption rarely holds true in longitudinal studies. To account for attrition, we can begin by denoting  $N_{0k}$  and  $N_{1k}$  as the number of subjects who have  $k$  observations ( $k = 1, \dots, n$ ) in the control group and the treatment group, respectively. If the assumed covariance structure for  $R^*$  is CS, then the estimated variance will be

$$\text{VAR}(\hat{\beta}) = \left( \sum_k \sum_{N_{0k}} \frac{k}{1 + (j-1)\rho_0^*\sigma_0^2} + \sum_k \sum_{N_{1k}} \frac{k}{1 + (j-1)\rho_1^*\sigma_1^2} \right)^{-1} \left[ \frac{1}{\pi p_0(1-p_0)} + \frac{1}{(1-\pi)p_1(1-p_1)} \right] \quad (8)$$

If the covariance structure is AR(1), then the estimated variance will be

$$\text{VAR}(\hat{\beta}) = \left( \sum_k \sum_{N_{0k}} \frac{k - (k-2)\rho_0^*}{(1 + \rho_0^*)\sigma_0^2} + \sum_k \sum_{N_{1k}} \frac{k - (k-2)\rho_1^*}{(1 + \rho_1^*)\sigma_1^2} \right)^{-1} \left[ \frac{1}{\pi p_0(1-p_0)} + \frac{1}{(1-\pi)p_1(1-p_1)} \right]. \quad (9)$$

## EXAMPLE

The macro is applied to a psychiatric clinical trial in late-life mood disorders. The study is a randomized clinical trial for maintenance therapies with a primary outcome of clinical dementia. The Mattis Dementia Rating Scale (MDRS) is a clinical instrument to measure cognitive impairment and is assessed over 4 timepoints for 2 years in elderly recovered depressed subjects randomized to donepezil or placebo. Maximum score on the MDRS is 144 and the cut-off score for dementia is 128. Most subjects are expected to deteriorate over time, however some subjects may improve or oscillate around the cut-off score. The within-subject structure was assumed to be compound symmetry with correlations between 0.4 and 0.6. The attrition over time was varied between 20% and 40%. Power was set at 80% and two-tailed alpha of 0.05. Marginal rate of outcome was set at 0.1 for subjects randomized to donepezil and 0.2 for subjects randomized to placebo.

```
title 'Sample Size';
%glimmixsamplesize
(RQPOWER=.8, BN=100, N=4, DROP=.2, PI=.5, P0=.2, P1=.1, G=1, RHO=.4, CORR=1);
bN = starting total sample size
```

```

n = number of time points for each subject
drop = drop out rate, assuming evenly distributed over time
pi = sample allocation ratio
p0 = marginal rate for group 0
p1 = marginal rate for group 1
G = G side random effects
rho = within subject correlation
corr= within subject correlation structure: 1 = CS, 2= AR(1)

title 'Power Calculation';
%glimmixpower(BN=100,N=4,DROP=.2,PI=.5,P0=.2,P1=.1,G=1,RHO=.4);
bN = starting total sample size
n = number of time points of the study
drop =drop out rate, assuming evenly distributed over time
pi = sample allocation ratio
p0 = marginal rate for group 0
p1 = marginal rate for group 1
G = G side random effects
rho = within subject correlation

```

Table 1a: Number needed to be randomized for varying level of attrition and within-subject correlation (T=4)

Attrition	Within-subject correlation		
	r=0.4	r=0.5	r=0.6
20%	194	218	242
30%	204	228	250
40%	214	238	262

Table 1b: Number needed to be randomized for varying level of attrition and within-subject correlation (T=6)

Attrition	Within-subject correlation		
	r=0.4	r=0.5	r=0.6
20%	186	214	242
30%	202	230	260
40%	218	250	280

## CONCLUSION

SAS macros have been created to calculate either sample size or power for correlated binary outcome data based upon mixed effect models. The macros have been applied to a late-life psychiatric longitudinal clinical trial.

## MACROS

```

%macro glimmixsamplesize
(RQPOWER=&RQPOWER,BN=&BN,N=&N,DROP=&DROP,PI=&PI,P0=&P0,P1=&P1,G=&G,RHO=&RHO,CORR=&CORR);
/* parameters
rqpower = required power
bN = starting total sample size
n = number of time points for each subject
drop = drop out rate, assuming evenly distributed over time
pi = sample allocation ratio
p0 = marginal rate for group 0
p1 = marginal rate for group 1
G = G side random effects
rho = within subject correlation
corr= within subject correlation structure: 1 = CS, 2= AR(1)
*/

proc iml;
/*treatment effect as beta*/
beta = abs((log((p0/(1-p0))/(p1/(1-
p1))))*sqrt(((16*sqrt(3))/(15*CONSTANT('PI')))**2*(G**2)+1)) ;
print "treatment effect coefficient is" beta;

/* CS within subject correlation*/
if corr = 1 then do;
powercs = 0.0;
DO until (powercs >= rqpower);

```

```

bNn = bN*(1-drop);
bN1 = bN*drop/n;
rhop0 = ((p0*(1-p0))**2*G + rho)/(1+ (p0*(1-p0))**2*G);
rhop1 = ((p1*(1-p1))**2*G + rho)/(1+ (p1*(1-p1))**2*G);
sigma20 = 1 + (p0*(1-p0))**2*G;
sigma21 = 1 + (p1*(1-p1))**2*G;

c = 1/(pi*bNn*n/((1+(n-1)*rhop0)*sigma20)+(1-pi)*bNn*n/((1+(n-1)*rhop1)*sigma21)
+ pi*bN1*3/((1+(3-1)*rhop0)*sigma20)+(1-pi)*bN1*3/((1+(3-1)*rhop1)*sigma21)
+ pi*bN1*2/((1+(2-1)*rhop0)*sigma20)+(1-pi)*bN1*2/((1+(2-1)*rhop1)*sigma21)
+ pi*bN1*1/((1+(1-1)*rhop0)*sigma20)+(1-pi)*bN1*1/((1+(1-1)*rhop1)*sigma21));

var1 = c*(1/(pi*p0*(1-p0))+ 1/((1-pi)*p1*(1-p1)));
powercs = 1 - probnorm(1.96 - beta*sqrt(1/var1));
bN = bN + 2;
end;
print "Minimum sample size for &RQPOWER power with CS structure is " bN powercs ;
end;

/* AR(1) within subject correlation*/
if corr = 2 then do;
powerar1 = 0.0;
DO until (powerar1 >= rqpower);
bNn = bN*(1-drop);
bN1 = bN*drop/n;
rhop0 = ((p0*(1-p0))**2*G + rho)/(1+ (p0*(1-p0))**2*G);
rhop1 = ((p1*(1-p1))**2*G + rho)/(1+ (p1*(1-p1))**2*G);
sigma20 = 1 + (p0*(1-p0))**2*G;
sigma21 = 1 + (p1*(1-p1))**2*G;

c2 = 1/(pi*bNn*(n-(n-2)*rhop0)/((1+rhop0)*sigma20)+(1-pi)*bNn*(n-(n-2)*rhop1)/
((1+rhop1)*sigma21)+pi*bN1*(3-(3-2)*rhop0)/((1+rhop0)*sigma20)+(1-pi)*bN1*(3-(3-2)*rhop1)/
((1+rhop1)*sigma21) + pi*bN1*(2-(2-2)*rhop0)/(1+rhop0)*sigma20+(1-pi)*bN1*(2-(2-2)*rhop1)/
((1+rhop1)*sigma21) + pi*bN1*(1-(1-2)*rhop0)/(1+rhop0)*sigma20+(1-pi)*bN1*(1-(1-2)*rhop1)/
((1+rhop1)*sigma21));

var2 = c2*(1/(pi*p0*(1-p0))+ 1/((1-pi)*p1*(1-p1)));
powerar1 = 1 - probnorm(1.96 - beta*sqrt(1/var2));
bN = bN + 2;
end;
print "Minimum sample size for &RQPOWER power with AR(1) structure is "
bN powerar1 ;
end;
quit;
run;
%mend glimmixsamplesize;
run;

%macro glimmixpower(BN=&BN,N=&N,DROP=&DROP,PI=&PI,P0=&P0,P1=&P1,G=&G,RHO=&RHO);
/*parameters
bN = starting total sample size
n =number of time points of the study
drop =drop out rate, assuming evenly distributed over time
pi = sample allocation ratio
p0 = marginal rate for group 0
p1 = marginal rate for group 1
G = G side random effects
rho = within subject correlation
*/

proc iml;
/*treatment effect as beta*/
beta = abs((log((p0/(1-p0)))/(p1/(1-
p1))))*sqrt(((16*sqrt(3))/(15*CONSTANT('PI')))**2*(G**2)+1)) ;
print "treatment effect coefficient is" beta;

bNn = bN*(1-drop);
bN1 = bN*drop/n;

```

```

rho0 = ((p0*(1-p0))**2*G + rho)/(1+ (p0*(1-p0))**2*G);
rho1 = ((p1*(1-p1))**2*G + rho)/(1+ (p1*(1-p1))**2*G);
sigma20 = 1 + (p0*(1-p0))**2*G;
sigma21 = 1 + (p1*(1-p1))**2*G;

/* CS within subject correlation*/
c = 1/(pi*bNn*n/((1+(n-1)*rho0)*sigma20)+(1-pi)*bNn*n/((1+(n-1)*rho1)*sigma21)
+ pi*bN1*3/((1+(3-1)*rho0)*sigma20)+(1-pi)*bN1*3/((1+(3-1)*rho1)*sigma21)
+ pi*bN1*2/((1+(2-1)*rho0)*sigma20)+(1-pi)*bN1*2/((1+(2-1)*rho1)*sigma21)
+ pi*bN1*1/((1+(1-1)*rho0)*sigma20)+(1-pi)*bN1*1/((1+(1-1)*rho1)*sigma21));

var1 = c*(1/(pi*p0*(1-p0))+ 1/((1-pi)*p1*(1-p1)));
powerCS = 1 - probnorm(1.96 - beta*sqrt(1/var1));

/*AR(1) within subject correlation*/
c2 = 1/(pi*bNn*(n-(n-2)*rho0)/((1+rho0)*sigma20)+(1-pi)*bNn*(n-(n-
2)*rho1)/((1+rho1)*sigma21) +pi*bN1*(3-(3-2)*rho0)/((1+rho0)*sigma20)+(1-pi)*bN1*(3-(3-
2)*rho1)/((1+rho1)*sigma21) +pi*bN1*(2-(2-2)*rho0)/(1+rho0)*sigma20+(1-pi)*bN1*(2-(2-
2)*rho1)/((1+rho1)*sigma21) +pi*bN1*(1-(1-2)*rho0)/(1+rho0)*sigma20+(1-pi)*bN1*(1-(1-
2)*rho1)/((1+rho1)*sigma21 ));

var2 = c2*(1/(pi*p0*(1-p0))+ 1/((1-pi)*p1*(1-p1)));
powerAR1 = 1 - probnorm(1.96 - beta*sqrt(1/var2));

print "Power estimations are "
powerCS powerAR1;
quit;
run;
%mend glimmixpower;

```

## REFERENCES

Dang Q, Mazumdar S, Houck PR. Sample size and power calculations based on generalized linear mixed models with correlated binary outcomes. *Computer Methods and Programs in Biomedicine*, in press.

Pan W. 2001, Sample size and power calculations with correlated binary data. *Controlled Clinical Trials*, 22:211-227.

Rochon J. 1998, Application of GEE procedures for sample size calculations in repeated measures, *Statistics in Medicine*, 17:1643-1658.

## ACKNOWLEDGMENTS

SAS® is a Registered Trademark of the SAS Institute, Inc. of Cary, North Carolina.

This work supported in part by the Advanced Center for Interventions and Services Research for Late-Life Mood Disorders (P30-MH071944).

## CONTACT INFORMATION

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

Patricia R Houck  
 UPMC  
 E-1128 TDH, WPIC  
 3811 O'Hara Street  
 Pittsburgh, PA 15213  
 Work Phone: (412)-246-6402  
 Email: houckpr@upmc.edu

\*\*\*\*\*