

# **An Algorithm and Macro for Estimating Power and Sample Size for Logistic Models with One or More Independent Variables of Interest in the Presence of Covariates**

D. Keith Williams and Zoran Bursac  
University of Arkansas for Medical Sciences

## **ABSTRACT**

Commonly when designing studies, researchers propose to measure several independent variables in a regression model, a subset of these are identified as the main variables of interest while the remainders are retained in a model as covariates or confounders. Power for linear regression in this setting can be calculated using SAS PROC POWER. There exists a void in estimating power for the logistic regression models in this same setting. Currently, an approach that calculates power for only one variable of interest in the presence of other covariates is in common use and works well for this special case. We propose an algorithm and a SAS macro that extends power estimation for one or more primary variables of interest in the presence of several confounders.

Keywords: Logistic regression, power, sample size, SAS *%PowerLog* macro

## **BACKGROUND**

The motivation for this work stems from methods that are in use to estimate power and sample size for standard linear regression models [4,6]. SAS PROC POWER [2,3] allows the investigator to determine the power to detect significance for a model with set of primary predictors of interest in the presence of covariates which are included in the model but not of primary interest. For example, suppose an investigator proposes a model with four total predictors X1, X2, X3, and X4 but is primarily interested in X1 and X2 while controlling for X3 and X4. To power this setting the full model would be:

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \beta_4 X4$$

While the reduced model would be:

$$Y = \beta_0 + \beta_3 X3 + \beta_4 X4$$

In the best case scenario to accurately estimate power, we would like to know the difference in the R-square of the full model and R-square for the reduced model. Then the short SAS code below would result in the immediately following output.

## SAS Code

```
proc power ;  
multreg  
model=fixed  
alpha= .05  
nfullpredictors= 4  
ntestpredictors= 2  
rsqfull=0.45  
rsqreduced=0.34  
ntotal= 60  
power=. ;  
run;
```

## Resulting Output

The POWER Procedure  
Type III F Test in Multiple Regression

Fixed Scenario Elements

Method	Exact
Model	Fixed X
Number of Predictors in Full Model	4
Number of Test Predictors	2
Alpha	0.05
R-square of Full Model	0.45
R-square of Reduced Model	0.34
Total Sample Size	0

Computed Power

Power

0.864

The setting above works nicely but uses information that is rarely known to an investigator in advance. A practical approach that provides a good guess for the difference in full and reduced model R-squares only requires the investigator to provide a correlation between each predictor X and Y plus the correlation of each of the X's among themselves. The straight forward matrix expression:

$$R^2 = \rho_{yx} R_{xx}^{-1} \rho_{yx}'$$

where  $\rho_{yx}$  is the vector of simple correlations between each predictor and the y response and  $R_{xx}^{-1}$  is the inverse of the correlation matrix among each of the predictors. Next one can correspondently calculate R-square for the reduced model by doing the identical calculation with the removal of the predictors of interest from the rows of  $\rho_{yx}$  and the rows and columns of  $R_{xx}^{-1}$ .

An example of these calculations is as follows:

$$R_{Full}^2 = [.6 \ .5 \ .6 \ .4] * \begin{bmatrix} 1 & .6 & .7 & .5 \\ .6 & 1 & .7 & .4 \\ .7 & .7 & 1 & .5 \\ .5 & .4 & .5 & 1 \end{bmatrix}^{-1} * \begin{bmatrix} .6 \\ .5 \\ .6 \\ .4 \end{bmatrix} = 0.45$$

where the leading vector is the set of simple correlations of Y with each of the four predictors, the middle matrix is the correlation of all four predictors among themselves, and the last vector is the transpose of the leading vector. The correlation values are from a particular data set and are intended for demonstration. If we are interested in the first two predictors controlling for the second two predictors we would use the calculation:

$$R_{Reduced}^2 = [.6 \ .4] * \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}^{-1} * \begin{bmatrix} .6 \\ .4 \end{bmatrix} = 0.34$$

in which it can be seen that the first two columns of the leading vector and the first two rows and columns of the middle matrix have been omitted. The difference in these two calculations results in:

$$R_{Full}^2 - R_{Reduced}^2 = 0.45 - 0.34 = 0.11$$

which represents another approach to providing the difference in R-squares, a quantity needed in order to calculate power for this regression model setting. A corresponding set of calculations can be done for any size set of predictors with a set of predictors of interest with the compliment of this set representing the predictors that are serving for controls. It is a reasonable approach in that researchers in many instances will have some idea of the simple correlations among the response and the predictors before their study, so this approach does have its merit.

## METHODS

We present the simple example above, because there does not exist similar corresponding approach in a logistic regression setting [1]. Currently, all the software the authors are aware of, estimates logistic model power of only one predictor of interest in the presence of some number of other covariates. A well written and documented SAS macro intended for this scenario is the `%PowerLog` macro [5]. This works nicely for this one setting but is not able to estimate power for a corresponding setting as was discussed above, that is, having more than one predictor of interest in a model controlling for other

predictors. Our objective was to provide a power estimation method that works in a somewhat corresponding manner to the matrix approach above. The proposed approach has the user providing the conjectured *odds ratios* associated with each predictor and the binary outcome in addition to the correlations among all the predictors. In the next section we outline our algorithm to estimate power for a given sample size in this manner.

### Algorithm Steps

1. Simulate a six variate uniform distribution  $(-6,6)$  for  $x_1, \dots, x_6$  with a pair wise correlation value  $\rho$ .
2. Calculate:  $\text{logit} = \ln(\text{avep}/(1-\text{avep})) + \ln(\text{OR}_1)X_1 + \dots + \ln(\text{OR})X_6$ .
3. If  $\text{logit}$  is less than or equal to a random uniform  $(0,1)$  draw then  $y=1$ , otherwise = 0.
4. Using SAS PROC LOGOSTIC, fit the full model  $y=X_1 X_2 X_3 X_4 X_5 X_6$  and save the  $-2 \log$  likelihood value.
5. Using SAS PROC LOGOSTIC, fit the reduced model which has the predictors of interest omitted from the full model and save the  $-2 \log$  likelihood value.
6. Save the difference in the full and reduced likelihood values and determine if this value is greater or equal to the appropriate critical value. If this is the case, record this single simulation run as a 'rejection'.
7. Repeat steps 1-5  $m$  times and tabulate the proportion of rejections. This proportion will be the estimate of the power of the conjectured scenario.

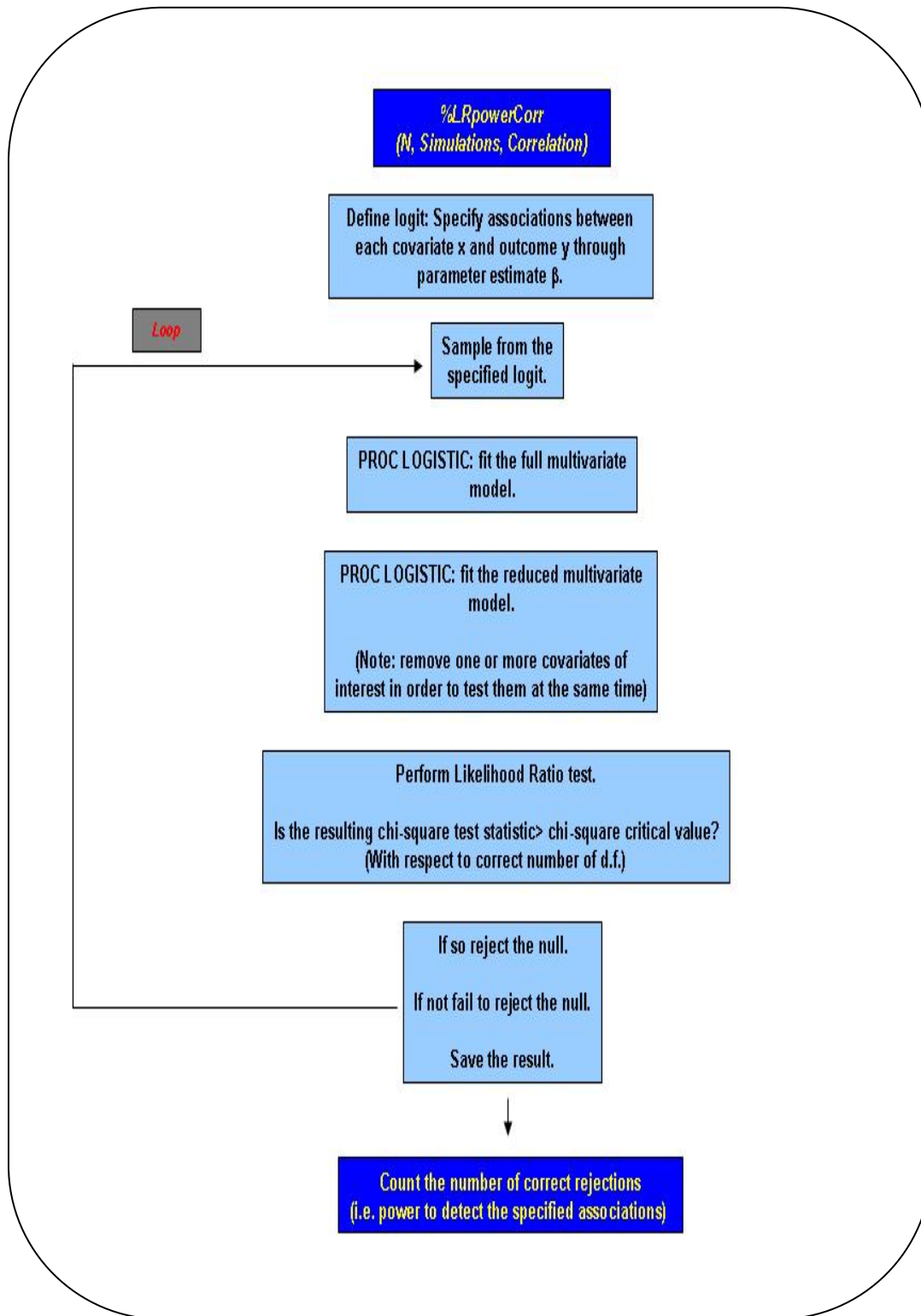


Figure 1. The %LRpowerCorr Macro Flowchart

## The %LRpowerCorr Macro

The user must define several variables as shown in Table 1. The macro variable SAMPLESIZE corresponds to the sample size that the macro is evaluating. NSIMS is the number of simulation runs required by the user, while P is the correlation among all of the predictors. AVEP is the average proportion of ‘yes’ responses ( $y = 1$ ) when all the predictor values are theoretically equal to zero. OR1 through OR6 are odds ratio values associated with the predictor variables CX1-CX6. The FULLMODEL macro variable has the user list the predictor variables in the full model. It should be noted that this is the literal script that is placed to the right of the equal sign in the model statement of the PROC LOGISTIC routine inside the macro, so care should be taken for accuracy. In a like manner, the REDUCEDMODEL variable is the list of predictors left in the model after the terms of interest are removed from the FULLMODEL list. ALPHA is the level of significance and DFTEST is the degrees of freedom for the likelihood ratio test. This will correspond to the number of predictor terms of interest, that is, the difference in the number of terms in the FULLMODEL and REDUCEDMODEL lists.

Table 1. Macro Variables

SAMPLESIZE	The sample size to be evaluated
NSIMS	The number of simulation runs
P	The correlation among the predictors
AVEP	The average number of “1” responses in the samples
OR1	The odds ratio associated with CX1
OR2	The odds ratio associated with CX2
OR3	The odds ratio associated with CX3
OR4	The odds ratio associated with CX4
OR5	The odds ratio associated with CX5
OR6	The odds ratio associated with CX6
FULLMODEL	The predictor terms in the full model among CX1 CX2 CX3 CX4 CX5 CX6
REDUCEDMODEL	The predictor terms in the reduced model among CX1 CX2 CX3 CX4 CX5 CX6
ALPHA	The significance level of the testing
DFTEST	The degrees freedom of the testing

## User Notes and Cautions

1. The sample size and the average proportion of  $Y = 1$  should have a product of at least 10 to minimize problems of complete separation, which generally causes numerical problems. Additionally, the sample size and one minus the average proportion should also have a product of at least 10.

2. Caution and thought should go into the value(s) of OR and average sample proportion being evaluated for multiple logistic model power. If one evaluates  $OR_1 = 2$  along with  $AVEP = 0.10$  in the setting in which the CX is from the default uniform(-6,6) distribution this roughly implies that at the  $P(Y = 1)$  approximately doubles for each one unit increase in CX1. Additionally, given that  $CX_1 = -6$ ,  $P(Y = 1)$  stretches from low probability of 0.002 up to 0.877 at  $CX_1 = 6$ . This is a wide range and is perhaps far from matching the reality of a research scenario. Conversely, if  $OR_1 = 1.1$  and  $AVEP = 0.10$  then the same range of  $P(Y = 1)$  goes from 0.059 to 0.164, perhaps matching the reality of what would happen in a real research study. It is common for instance to have a highly significant continuous covariate (like age) with OR of 1.04, suggesting 4% increase in odds per every year increase in age. In the Appendix 1 there is a short SAS program that calculates the particular values of  $P(Y = 1)$  for a range of CX values, this gives one a quick calculation for evaluation.
3. The `%LRpowerCorr` macro uses the likelihood ratio (LR) chi-square statistic to evaluate power. Other power approaches use the Wald chi-square for the power evaluation. Most statisticians would agree that the LR chi-square is generally a bit more sensitive and this implies that if one compared equivalent scenarios, it is likely that the LR chi-square approach would be slightly more powerful but still very close.
4. This algorithm and macro used at this stage of development and testing are a work in progress. More evaluation and increased sophistication of the macro itself is appropriate. Other distributions for the predictor variables as well as combinations of those would be another evolution of the approach and macro. Additionally, talented SAS programmers could likely speed up the evaluation of a given scenario.

## DISCUSSION

The `%LRpowerCorr` macro and the algorithm it is based on, shows promise to fill a void for estimating power for multivariable logistic models. It is able to match the approach that researchers can use for multiple regression when estimating the power of a model in which one or more predictors are of interest while controlling for a number of other predictors. We can specify the amount of correlation among all the predictors and attempt to match real data analysis setting that researcher commonly encounter. The current `%LRpowerCorr` macro [is given in Appendix 2](#).

## REFERENCES

1. Hosmer, D.W., and Lemeshow, S. 2000. *Applied Logistic Regression*. New York: Wiley.
2. SAS Institute Inc. 2004. *SAS/STAT User's Guide, Version 9.1*. Cary, NC: SAS Institute Inc.
3. Castelloe, J.M. (2000), "Sample Size Computations and Power Analysis with the SAS System," Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Paper 265-25, Cary, NC: SAS Institute Inc.
4. Maxwell, S. E. (2000). Sample size and multiple regression analysis. *Psychological Methods*, 5, 434–458.

5. Powerlog macro <http://www.math.yorku.ca/SCS/vcd/powerlog.html> Friendly, M. Visualizing Categorical Data. SAS Institute, Cary, NC, 2000.
6. Hsieh, F.Y., Block, D.A., and Larsen, M.D. (1998). A Simple Method of Sample Size Calculation for Linear and Logistic Regression. *Statistics in Medicine*, Volume 17, pages 1623-1634.

## CONTACT INFORMATION

The %LRpowerCorr macro will be provided as requested.

D. Keith Williams

Biostatistics

Fay W. Boozman College of Public Health

University of Arkansas for Medical Sciences

4301 W. Markham, Slot 781

Little Rock, AR 72205

Work Phone: (501) 526-6723

Fax: (501) 526-6721

E-mail: [williamsdavidk@uams.edu](mailto:williamsdavidk@uams.edu)

Web: [www.uams.edu/biostat/williams/](http://www.uams.edu/biostat/williams/)

## Appendix 1

The SAS *phats* program that provides insight in the range of estimated probabilities suggested by ones choice of OR and P0.

```
data phats;
input x@@;
or=1.1;
p0=.05;
B0=log(p0/(1-p0));
/* Next 2 statements convert to binary for example.
/*Comment out for uniform(-6,6) */
*xunit=(x+6)/12;
*xunit=round(xunit);
xunit=x;
logit=B0+(log(or))*xunit;ps=exp(logit)/(1+exp(logit));
cards;
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
;
proc print;run; proc gplot data=phats; plot ps*xunit;run;
```

## Appendix 2

### The Current %LRpowerCorr SAS Macro

```
%macro LRpowerCorr (samplesize,nsims,p,avep,or1,or2,or3,or4,or5,or6,
                    fullmodel,
                    reducedmodel,
                    alpha,
                    dftest);

proc datasets library=work; delete base; run;quit;
options nosource nonotes nosource2 noprintmsglist nosymbolgen nomprint;

%let size=&samplesize; /* input sample size */

/* Generate sample values with data step */

%do sim=1 %to &nsims;

data multiple2(drop=samp);
  retain Seed_1 0 Seed_2 0
         Seed_3 0 seed_4 0 seed_5 0 seed_6 0 seed_7 0;

  do samp=1 to &size;

/* generate random 0,1 uniform */
x1=12.01*ranuni(Seed_1)-6.01;
  x2=12.01*ranuni(Seed_2)-6.01;
  x3=12.01*ranuni(Seed_3)-6.01;
  x4=12.01*ranuni(Seed_4)-6.01;
  x5=12.01*ranuni(Seed_5)-6.01;
  x6=12.01*ranuni(Seed_6)-6.01;
xy=ranuni(Seed_7);
  cx1=x1/((12)**.5);
  cx2=x2/((12)**.5);
  cx3=x3/((12)**.5);
  cx4=x4/((12)**.5);
  cx5=x5/((12)**.5);
  cx6=x6/((12)**.5);

  output;
end;

run;
/* standardize the logit converted values */
ods listing close;
/* Performs matrix algebra to obtain a correlation structure */
/* creates a binary value y */
/* Performs matrix algebra to obtain a correlation structure */
proc iml;
  use multiple2;
  read all var{cx1 cx2 cx3 cx4 cx5 cx6} into center;
  read all var{xy} into xy;
```

```

corr = {1 &p &p &p &p &p &p, &p 1 &p &p &p &p &p, &p &p 1 &p &p &p &p, &p &p &p 1
&p &p,
      &p &p &p &p 1 &p, &p &p &p &p &p 1};
cy = center * half(corr);
cx1=cy[,1]; cx2=cy[,2]; cx3=cy[,3]; cx4=cy[,4]; cx5=cy[,5]; cx6=cy[,6];
cx1=cx1*((12)**.5); cx2=cx2*((12)**.5); cx3=cx3*((12)**.5);
cx4=cx4*((12)**.5); cx5=cx5*((12)**.5); cx6=cx6*((12)**.5);
xy=xy;
create multiple3 var{cx1 cx2 cx3 cx4 cx5 cx6 xy};
append ;
run; quit;
/* specify your logit */
data multiple3 ; set multiple3;
/* f is a small add on to help with numeric problems */
f=0.00001;
b0=log(&avep/(1-&avep));
*logit=-
0.0+(1/3.5)*0.693*cx1+0.00001*cx2+.00001*cx3+0.00001*cx4+0.00001*cx5+0.
00001*cx6;

/* Binary cx1 cx2 if desired comment out if uniform(-6,6) is desired.
*/
cx1=round((cx1+6)/12);
cx2=round((cx2+6)/12);
/*      */
logit=b0+
(log(&or1)+f)*(cx1)+
(log(&or2)+f)*(cx2)+
(log(&or3)+f)*cx3+
(log(&or4)+f)*cx4+
(log(&or5)+f)*cx5+
(log(&or6)+f)*cx6;
prob=(exp(logit))/(1+exp(logit));
if xy<=prob then y=1;
if xy>prob then y=0;
output;
run;
/* fit the full model */
ods output fitstatistics=fits1;

proc logistic data=multiple3 descending ;

model y=&fullmodel;
run;

data fits1;set fits1;
if criterion='AIC' or criterion='SC' then delete;
drop interceptonly;
run;

/* fit the reduced model */
ods output fitstatistics=fits2;

proc logistic data=multiple3 descending ;
model y= &reducedmodel ;
run;

```

```

data fits2;set fits2
(rename=(interceptandcovariates=interceptandcovariates2));
if criterion='AIC' or criterion='SC' then delete;
drop interceptonly;
run;

/* specifies chi-square critical value */
/* 3.84 5.99 7.81 9.49 11.07 */
data both;merge fits1 fits2;
likelihoodratio=interceptandcovariates2-interceptandcovariates;
critval=cinv(1-&alpha,&dfest);
if likelihoodratio>critval then reject=1;
else reject=0;
run;

proc append base=base data=both;
run;

/*dm 'log;clear;output;clear;'; */
/*dm 'log;clear;'; */
%end;
ods listing;
proc freq data=base;
tables reject;
run;

%mend LRpowerCorr;

```

### Appendix 3

An example of `%LRpowerCorr` use.

Here we demonstrate a scenario with six total predictors in which X1 and X2 are binary and of primary interest with **OR1 = 2** and **OR2 = 2**, while X3, X4, X5, and X6 are covariates with OR values all equal to 1. The sample size of interest is **200**, correlation among the predictors is **0.2**, and the average proportion of outcome Y=1 is **0.25**. One wishes to perform **1000** simulations comparing the full model with all 6 predictors versus 4 predictors at the 0.05 level of significance, resulting in a chi-square critical value with  $df = 2$ .

The following macro call invokes above described scenario,

```
%LRpowerCorr (200,1000,.2,.25,2,2,1,1,1,1,  
              cx1 cx2 cx3 cx4 cx5 cx6,  
              cx3 cx4 cx5 cx6,  
              0.05,  
              2);run;
```

and gives the output below, suggesting the approximate power is **82.7%**.

The FREQ Procedure				
reject	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	173	17.30	173	17.30
1	827	82.70	1000	100.00