SAS® GLOBAL FORUM 2021

Paper 1158-2021

Do I Have Enough? - A Macro for Sample Size Determination in Simple Binary Logistic Regression and Multiple Binary Logistic Regression Models

Carl P. Wilson, Spectrum Health Office of Research and Education and Grand Valley State University

ABSTRACT

Binary logistic regression finds plenteous usage throughout many scientific disciplines. Despite its multitudinous applications, there lacks a universal method of determining the sample size for a binary logistic regression model. Some suggest using 'magic number' methods to estimate the sample size, but such methods do not include statistical power, effect size, or significance level into their approximations. To remedy this dilemma, this paper presents a SAS macro that automates the sample size determinations for both simple and multiple binary logistic regression models with continuous predictors. This macro—based on Hsieh's (1998) formula—lacks the limitations from which the magic number method tolerates. Researchers will benefit from using this macro to determine the true sample sizes for their studies and to overcome the imbroglios of under and overestimating a sample size.

INTRODUCTION

The application of binary logistic regression is prevalent throughout various scientific disciplines, such as biomedicine, social science, business, and genetics (Agresti, 2014). The ubiquity of binary logistic modeling compels the desideratum to obtain a proper sample size calculation. Generally speaking, with an insufficient sample size, a study may fail to detect even large treatment effects (Wang & Ji, 2020). Conversely, a study may allocate expensive resources (e.g. research time, personnel labor, overall cost) towards obtaining an excessively inflated sample size.

To further complicate things, there is an egregious amount of disagreement regarding the best method of sample size determination for logistic regression. One proposed solution, the EPV method (Events Per Variable), suggests having a specified minimum number of events/subjects per explanatory variable (Harrell, Lee, Matchar & Reichert, 1985). Both cox regression and logistic regression used this methodology with the generally advocated EPV value being ten (Peduzzi, et. al, 1996), (Concato, et. al, 1995). Unfortunately, there is even discordance for the ideal EPV value, with some disputing it should be at least 20 (Austin & Steyerberg, 2017), and others suggesting the minimum should be 50 (Bujang, et. al, 2018). There is also speculation regarding whether the EPV method is even valid (Smeden et al., 2016). Further inspecting validity, these "Magic Number" methods do not take into consideration an effect size, significance level, or power threshold for their sample size calculations, and are therefore less apt than methods that do (Courvoisier, et. al, 2011). In order to rectify these issues, this macro uses an approach by Hsieh (1998).

BACKGROUND

Hsieh's method of sample size determination is a formulaic approach predicated on a binary logistic regression model being generalized by a two-sample framework. In a simple logistic regression relating a predictor X_1 to a binary response Y, the model is as follows:

$$log(\frac{P}{(1-P)}) = \beta_0 + \beta_1 X_1,$$

In such a scenario, we are interested in whether or not the predictor X_1 is related to the binary response variable, and we verify this by testing the null hypothesis H_0: $\beta_1 = 0$, against the alternative hypothesis H_1: $\beta_1 \neq 0$. When X_1 is a continuous variable with a normal distribution, the log odds value $\beta_1 = 0$, if and only if, the group means (assuming homogeneity of variances) is equal between the two response categories; therefore, we can derive a sample size formula from a two-sample test (see Hsieh (1998) for full derivation and theory). As suggested by Whittemore (1980), the sample size approximation can be further improved for small response probabilities via multiplying by a correction factor of $1 + 2P\delta$, yielding a sample size formula of:

$$n_{simple} = (Z_\alpha + Z_\beta\, e^{(\frac{-\theta^{*2}}{4})})^2\, (\frac{1+2P\delta}{P\theta^{*2}}),$$

$$where\ \delta = [1 + (1 + \theta^{*2})\, e^{(\frac{5\theta^{*2}}{4})}](1 + e^{(\frac{-\theta^{*2}}{4})})$$

Z_α and Z_β are the upper quantiles for the significance and power thresholds respectively, $P = P(Y \mid \bar{x}_{X1} + s_{X1})$, the probability of the response at one standard deviation above the mean of the predictor X_1, and $\theta^* = log(\frac{P(Y \mid \bar{x}_{X1} + s_{X1})}{P(Y \mid \bar{x}_{X1})})$, the overall effect size derived as the log odds ratio of the probability of the response at one standard deviation above the mean of X_1, relative to the probability of the response at the mean of X_1. As the probability of the response at the mean of X_1 approaches 0.50, and as $|\theta^*|$ increases, the smaller the necessary sample size to achieve the desired power will be (see Agresti (2014) for more information).

Whittemore (1981) and Hsieh (1998) demonstrated that this formula can also be used to approximate the sample size for a multiple logistic regression model with n continuous predictors, through inflating the aforementioned simple logistic regression sample size via multiplying by $\frac{1}{1-R^2}$, where R^2 is the proportion of the variance of X_1 (the original predictor from the univariable model) explained by the regression with the new quantitative predictors in the multiple variable model, X_2, …, X_n. This yields a sample size formula of:

$$n_{multiple} = \frac{n_{simple}}{1-R^2}$$

EXPLANATION OF THE MACRO FUNCTIONALITY AND CODE

This macro has the capacity to automatically calculate the effect size for the user, a feature that most sample size determination software's lack. However, this is at the expense of requiring a sample dataset to derive the calculations from. This macro will be useful for researchers conducting small-scale pilot studies, as they can determine an effect size and required sample size to achieve their desired power and significance levels for their full-scale research projects.

As aforementioned, the macro is derived from Hsieh and Whittemore's formulas and is flexible in that it can automate the determination of the sample size both for simple and multiple binary logistic regressions. The macro has 6 parameters, and are outlined as follows:

- dataset – The name of the dataset.

- response – The binary response variable.
- x – The singular continuous predictor for the simple logistic regression case, or the main predictor of interest for the multiple logistic regression case.
- other_x – The other continuous predictors for the multiple logistic regression case; to utilize only a simple logistic regression, set this parameter equal to 0.
- alpha – The specified significance level.
- beta – The specified beta level to calculate power (power = 1 - β).

EXAMPLE OF MACRO CODE USE

SASHELP library contains the dataset HEART, which we will use to illustrate the functionality of this macro. The HEART dataset contains data from the Framingham heart study (see framinghamheartstudy.org/fhs-about/ for more information), and with it, we can build simple and multiple logistic regressions to predict the binary variable *status* (Dead or Alive) based on continuous predictors. To parallel how this macro may be used in a research setting, a random sample of size 100 will be taken via PROC SURVEYSELECT. The random sample is comparable to how a researcher would use this macro with a small-scale pilot dataset. From this pilot dataset, they use the macro to determine the observed effect size, and with this value, the macro calculates the necessary sample size to reach the power and significance levels desired for their full-scale study. The source code used to create the random sample is given below:

```
proc surveyselect data=sashelp.heart out=sample
sampsize=100 seed=12345;
run;
```

Simple Logistic Regression

If we are only interested in predicting status based one main predictor of interest (i.e. systolic blood pressure), and we want a significance level of α = 0.05, and a power of 0.8, the macro code is as follows:

```
%mlog_n(dataset = sample,
        response = Status,
        x = systolic,
        other_x = 0,
        alpha = .05,
        beta =  .2);
run;
```

For this example, the *dataset* parameter is sample (the random sample of 100 observations from the HEART dataset), the *response* is status, the predictor *x* is systolic, the parameter for the other predictors *other_x* will be set to zero as we are only interested in one predictor, systolic. Our specified *alpha* threshold is 0.05, and our *beta* level is 0.2 (i.e. 1 - .2 = .8 power).

The macro outputs summary values of the input parameters, such as the significance level, the power threshold, the respective upper quantiles of the two aforementioned values, the probability of the response at the mean level of the predictor variable *X*, the probability of the response at 1 standard deviation above the mean of the predictor *X*, the overall log-odds (effect size), and the value of Whittemore's (1980) correction factor. Subsequent to the summary values, the macro outputs the estimated sample size based on the input parameters via Hsieh's (1998) formula (Output 1).

Parameter Values

Alpha	Beta	Z alpha/2	Z beta/2	Probability at mean level of x	Probability at 1 stddev above mean	Effect Size (Overall Log Odds)	Correction Factor (Delta)
0.05	0.2	1.95996	1.28155	0.52719	0.22228	-0.37506	1.20073

Page Break

Estimated Sample Size

Sample Size
501.419

Output 1. Output for Simple Logistic Regression

Multiple Logistic Regression

We can also create a multiple binary logistic regression model by including more predictors (e.g., systolic, diastolic, weight, height, and cholesterol) to predict status. Using the same significance and power as the previous example, the macro code is as follows:

```
%mlog_n(dataset = sample,
        response = Status,
        x = systolic,
        other_x = diastolic height weight cholesterol,
        alpha = .05,
        beta =  .2);
run;
```

For this example, the *dataset* parameter is still the random sample, the *response* remains as status; the *x* parameter remains as systolic for being our main predictor of interest; the *other_x* parameters are now diastolic, height, weight; and cholesterol instead of 0, and lastly our *alpha* and *beta* thresholds are still 0.05 and 0.2, respectively.

The macro outputs the same summary values of the input parameters and sample size that the simple logistic regression version does, as well as the sample size estimation for the multiple logistic regression (Output 2).

Parameter Values

Alpha	Beta	Z alpha/2	Z beta/2	Probability at mean level of x	Probability at 1 stddev above mean	Effect Size (Overall Log Odds)	Correction Factor (Delta)
0.05	0.2	1.95996	1.28155	0.52719	0.22228	-0.37506	1.20073

Page Break

Estimated Sample Size

Sample Size
501.419

Page Break

Estimated Sample Size for Multiple Predictors

n_multi
1205.17

Output 2. Output for Simple Logistic Regression

CONCLUSION

With the importance of a proper sample size for conducting research, this macro is beneficial in that it streamlines the process of sample size determination for simple and

multiple binary logistic regression models. It provides novel usefulness in that the macro can automate the calculation of the effect size θ^* and the sample size n from a smaller representative sample, allowing the researcher to know how much data they will need to collect or pull from databases/registries in order to achieve their desired power and significance levels.

In both of these scenarios, the macro circumvents the issue of failing to detect effects attributed to having too small of a sample size—a complication that can arise from using "magic number" methods like the Events Per Variable method, which do not consider statistical power, significance, or an effect size. This macro also saves resources such as cost, labor, and research time that would be drained from excessive data collection to achieve a larger than necessary sample size.

REFERENCES

Agresti, A. (2014). *Categorical Data Analysis*. Hoboken: Wiley.

Austin, P. C., & Steyerberg, E. W. 2017. "Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models." *Statistical methods in medical research*, *26*(2), 796–808. https://doi.org/10.1177/0962280214558972

Bujang, M. A., Sa'at, N., Sidik, T., & Joo, L. C. 2018. "Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data." *The Malaysian journal of medical sciences : MJMS*, *25*(4), 122–130. https://doi.org/10.21315/mjms2018.25.4.12

Concato, J., Peduzzi, P., Holford, T. R., & Feinstein, A. R. 1995. "Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy." *Journal of clinical epidemiology*, *48*(12), 1495–1501. https://doi.org/10.1016/0895-4356(95)00510-2

Courvoisier DS, Combescure C, Agoritsas T, Gayet-Ageron A, & Perneger TV. 2011. "Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure." *J Clin Epidemiol*.

Harrell FE Jr, Lee KL, Matchar DB, Reichert TA. 1985. "Regression models for prognostic prediction: advantages, problems, and suggested solutions." *Cancer Treat Rep*.

Hsieh, F. Y., Bloch, D. A., & Larsen, M. D. 1998. "A simple method of sample size calculation for linear and logistic regression." *Statistics in Medicine, 17*(14), 1623-1634. http://dx.doi.org/10.1002/(SICI)1097-0258(19980730)17:14<1623::AID-SIM871>3.0.CO;2-S

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. 1996. "A simulation study of the number of events per variable in logistic regression analysis." *J Clin Epidemiol.*

Smeden, M. V., Groot, J. A., Moons, K. G., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. 2016. "No rationale for 1 variable per 10 events criterion for binary logistic regression analysis." *BMC Medical Research Methodology, 16*(1). https://doi.org/10.1186/s12874-016-0267-3

Wang, X., & Ji, X. 2020. "Sample Size Estimation in Clinical Research From Randomized Controlled Trials to Observational Studies." *An Overview Of Study Design And Statistical Considerations, 158*(1). https://doi.org/10.1007/s11707-018-0727-7

Whittemore, A. 1981. "Sample size for logistic regression with small response probability." *Journal of the American Statistical Association*.

ACKNOWLEDGMENTS

I would like to acknowledge the entire Scholarly Activity and Scientific Support team at Spectrum Health for their feedback on this paper. A special thanks to Nicholas Andersen, PhD, for his guidance in the development of this paper, and also to Jessica Parker for her continued mentoring throughout the entirety of my SAS Global Forum experience.

RECOMMENDED READING

- *(PDF) A Simple Method of Sample Size Calculation for Linear and Logistic Regression (researchgate.net)*

CONTACT INFORMATION

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

Carl P. Wilson
Spectrum Health Office of Research and Education
Grand Valley State University
wilsocar@mail.gvsu.edu

APPENDIX: MACRO CODE

```
/*************************************************************************/
/*      Fully Automated Binary Logistic Regression Sample Size MACRO     */
/*************************************************************************/
/*                                                                       */
/* Author: Carl Wilson.                                                  */
/*                                                                       */
/*                                                                       */
/* Required parameters:                                                  */
/*                                                                       */
/*  dataset       = The dataset name                                     */
/*  response      = The binary response variable y                       */
/*  x             = quantitative predictor of interest                   */
/*  other_x       = other quantitative predictors (other_x = 0 for simple) */
/*  alpha         = desired significance level                           */
/*  beta          = 1 minus the desired power level                      */
/*                                                                       */
/*************************************************************************/

%macro mlog_n  (dataset = ,     /*dataset name*/
                response = ,    /*binary response variable y*/
                x = ,           /*Quantitative predictor of interest*/
                other_x = ,     /*other quantitative predictors */
                alpha = ,       /*Significance level*/
                beta =  );      /*1 minus the desired power level*/

/*finding the mean and standard deviation of the predictor of interest, x*/
      proc means data=&dataset mean std noprint;
            var &x;
                  output out=mean_std mean=mean std=std;
      run;

      data _null_;
            set mean_std;
            call symput("mean",mean);
            call symput("std",std);
      run;

      data data1;
            set &dataset;
            &x = &mean;
      run;

/*Finding the probability of the response at the mean value of x*/
      proc logistic data = &dataset outest=log_out noprint;
            model &response = &x;
      run;

      proc score data=data1 score=log_out out=pred1 type=parms;
            var &x;
      run;

      ods select Variables;
      proc contents data=pred1;
            ods output Variables=var1;
      run;
      proc sort data=var1;
```

```
      by Num;
run;

data _null_;
      set var1;
      by num;
      if last.num then call symputx('lastvar', Variable);
run;

%put The last variable is &lastvar;

data want;
      set pred1;
      keep &lastvar;
run;

data p1;
  set want;
  odds = exp(&lastvar);
  p_1 = odds /(1+odds);
run;

data _null_;
      set p1;
      call symput("p1",p_1);
run;

/*Finding the probability of the response 1 standard deviation above the mean
value of x*/
data data2;
      set &dataset;
      &x = &mean + &std;
run;

proc logistic data = &dataset outest=log_out2 noprint;
      model &response = &x;
run;

proc score data=data2 score=log_out2 out=pred2 type=parms;
      var &x;
run;

ods select Variables;
proc contents data=pred2;
      ods output Variables=var2;
run;

proc sort data=var2;
      by Num;
run;

data _null_;
      set var2;
      by num;
      if last.num then call symputx('lastvar_p2', Variable);
run;

%put The last variable is &lastvar_p2;
```

```
	data want2;
		set pred2;
		keep &lastvar_p2;
	run;

	data p2;
	  set want2;
	  p2_odds = exp(&lastvar_p2);
	  p2 = p2_odds /(1+p2_odds);
	run;

	data _null_;
		set p2;
		call symput("p2",p2);
	run;

/*finding R-squared*/
	%if &other_x ne 0 %then %do;

		proc reg data=&dataset noprint outest=outreg;
			model &x = &other_x / rsquare;
		run;

		data _null_;
			set outreg;
			call symput("Rsq", _RSQ_);
		run;
	%end;

/*Calculations for estimating the sample size*/
	data sample_size;
		alpha = &alpha;
		beta = &beta;
		za = quantile("Normal", 1-&alpha/2);
		zb = quantile("Normal", 1-&beta/2);
		p1 = &p1;
		p2 = &p2;
		t = (log10(p2/p1));
		l = (1+(1+t**2)*(exp(5*t**2/4)))/(1+exp((-t**2/4)));
		n = ((za + zb*exp(-t**2/4))**2)*((1+2*p2*l)/(p2*t**2));

		label alpha = "Alpha"
			beta = "Beta"
			za = "Z alpha/2"
			zb = "Z beta/2"
			p1 = "Probability at mean level of x"
			p2 = "Probability at 1 stddev above mean"
			t = "Effect Size (Overall Log Odds)"
			l = "Correction Factor (Delta)"
			n = "Sample Size";
	run;

/*outputting the parameter values that go into the calculation*/
	title "Parameter Values";
	proc print data=sample_size label noobs;
		var alpha beta za zb p1 p2 t l;
	run;
```

```
      title;

/*outputting the estimated sample size*/
      title "Estimated Sample Size";
      proc print data=sample_size label noobs;
            var n;
      run;
      title;

/*calculating sample size for multi log reg*/
      %if &other_x ne 0 %then %do;

      data multi;
            set sample_size;
            n_multi = n / (1 - &Rsq);
      run;

/*outputting the estimated multi sample size*/
      title "Estimated Sample Size for Multiple Predictors";
      proc print data=multi label noobs;
            var n_multi;
      run;
      title;

      %end;

%mend mlog_n;
```