Paper 260-27

Redesigning Experiments With Polychotomous Logistic Regression: A Power Computation Application

Charles Vaughan, Xoma US, LLC, Berkeley, CA
Serge Guzy, Xoma US, LLC, Berkeley, CA

ABSTRACT

Power and sample size calculations for experiments modeled with binary logistic regression are becoming more common, and are even available as freeware (e.g. Ralph O'Brien's *UnifyPow* application). Somewhat less common, if not altogether absent, is software that allows power analysis for multinomial logistic regression models. Using the algorithm introduced below, it is now possible to compute powers and sample sizes for arbitrary multinomial (ordinal) logistic regression models.

In particular, this application has been designed to allow investigators to explore subsets of data from previous studies that will, in turn, allow them to design or redesign experiments with optimal power.

This algorithm requires SAS/BASE®, SAS/STAT® and SAS/IML®. The algorithm currently runs on PC SAS, but may run on any platform. Finally, a heuristic will be presented for the general programmer. A knowledge of random number generation is helpful.

INTRODUCTION

We will assume the reader is familiar with the general form for ordinal logistic regression:

$$\Pr(response \le a_i \mid X_1,\ldots,X_m) = \frac{1}{1+\exp(-(\alpha_i + \beta^t X))},\ \text{for } 1 \le i \le k$$

where $a_1 < a_2 < \ldots < a_k$ are k ordinal response levels, $X_1,\ldots,X_m$ are m explanatory variables, $\beta^t = [\beta_1 \ldots \beta_m]$ is the vector of slope parameters, and $X^t = [X_1 \ldots X_m]$ is the vector of explanatory variables.

Geometrically, each of the k - 1 cumulative linear predictors $\alpha_i + \beta^t X$ forms an m-dimensional hyperplane. If we can quantify corresponding measures of dispersion, say, ε_i, about each of these hyperplanes from the observed data, then we can simulate response probabilities, and hence patient responses (see figure 1).

In figure 1, we have, for ease of explanation, illustrated a trichotomous model for a given treatment group (placebo or drug) with one explanatory variable X_1. In this case the hyperplanes are simply parallel lines. For a fixed cross-section of these parallel lines, the measures of dispersion ε_i about each line at x_1 is given by

$$\hat{\varepsilon}_i = \sqrt{\begin{bmatrix}1 & x_1\end{bmatrix} \cdot \hat{V}_i \cdot \begin{bmatrix}1 & x_1\end{bmatrix}^t},\ i = 1,2.$$

where

$$\hat{V}_i = \begin{bmatrix} \hat{\sigma}^2_{\alpha_i} & \hat{\sigma}_{\alpha_i\beta_1} \\ \hat{\sigma}_{\beta_1\alpha_i} & \hat{\sigma}^2_{\beta_1} \end{bmatrix}, i = 1,2$$

are submatrices of the larger 3 by 3 variance-covariance matrix whose entries are maximum likelihood estimates of the variances and covariances of all the ordinal logistic regression model parameters.

METHODS

Following a preliminary ordinal logistic regression modeling, there are four groups of algorithmic steps which follow each other sequentially: covariate simulation, response probability simulation, response simulation, and p-value computation.

Preliminary Modeling

Step 1: Input the raw data. If there are, say, m continuous covariates $X_1,\ldots,X_m$, prompt the investigator for an interval of interest on each covariate. These intervals may be thought of as inclusion/exclusion criteria for the study. In the case of figure 1, there is only one covariate, so the investigator's interval of interest may be expressed simply as a closed interval $[a,b]$.

Step 2: For each treatment group, construct a cumulative logistic regression model, and store the hyperplane parameter estimates in respective matrices.

1. Covariate Simulation

Step 3: For each treatment group (placebo and drug), simulate a vector from the approximate (multivariate normal) distribution of the covariates. In the case of figure 1, we would simulate from the approximate univariate normal distribution of X_1. Call these simulated vectors (scalars in figure 1) simvec0 and simvec1 respectively. These vectors may be thought of as corresponding to a single placebo patient and a single drug patient respectively.

Step 4: Repeat step 3 until simvec0 and simvec1 fall in $[a,b]$.

Step 5: Augment simvec0 and simvec1 with leading 1s for appropriate matrix algebra.

2. Response Probability Simulation

Steps 6 - 9: Through random number generation and further matrix manipulations, the end result of these steps is a pair of vectors of cumulative "linear predictors + respective measures of dispersion" for the simulated placebo and drug patients respectively. Call these vectors pred0 and pred1 respectively. In the case of figure 1, the vector of cumulative "linear predictors + measures of dispersion" is given by:

$$\begin{bmatrix}(\hat{\alpha}_1 + \hat{\beta}_1 x_1) + \hat{\varepsilon}_1 & (\hat{\alpha}_2 + \hat{\beta}_1 x_1) + \hat{\varepsilon}_2\end{bmatrix}$$

Step 10: Append pred0 and pred1 to respective lists Predlst0 and Predlst1, or set Predlst0 = pred0 and Predlst1 = pred1 on the first pass through the algorithm.

Step 11: Repeat steps 3 through 10 N - 1 times, where $N = N_{control} = N_{drug}$ is a prespecified sample size assigned by the investigator.

We now have two lists, Predlst0 and Predlst1, which may be represented as $N \times (k - 1)$ matrices. In the trichotomous case shown in figure 1, one such matrix would look like:

$$\begin{bmatrix} (\hat{\alpha}_1 + \hat{\beta}_1 x_1^{(1)}) + \hat{\varepsilon}_1^{(1)} & (\hat{\alpha}_2 + \hat{\beta}_1 x_1^{(1)}) + \hat{\varepsilon}_2^{(1)} \\ \vdots & \vdots \\ (\hat{\alpha}_1 + \hat{\beta}_1 x_1^{(N)}) + \hat{\varepsilon}_1^{(N)} & (\hat{\alpha}_2 + \hat{\beta}_1 x_1^{(N)}) + \hat{\varepsilon}_2^{(N)} \end{bmatrix}$$

where $x_1^{(n)}$ represents a simulated value of X_1 for the n^{th} patient, $n = 1,\ldots,N$.

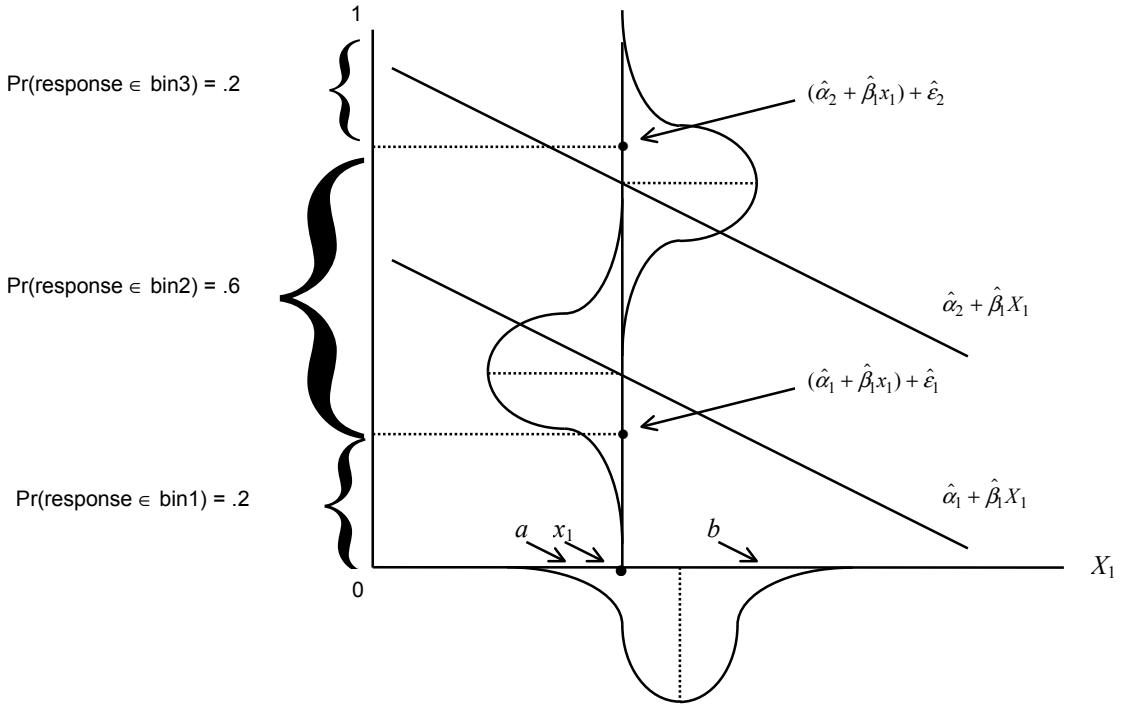


Figure 1. Simulating A Patient Response For A Treatment Group

Step 12: By applying the logit function to each entry in Predlst0 and Predlst1, we obtain two $N \times (k - 1)$ matrices SimProb0 and SimProb1 whose rows contain $k - 1$ cumulative probabilities for k possible ordinal responses. In the trichotomous case shown in figure 1, one such matrix would look like

$$\begin{bmatrix} \frac{1}{1+\exp(-(\hat{\alpha}_1 + \hat{\beta}_1 x_1^{(1)} + \hat{\varepsilon}_1^{(1)}))} & \frac{1}{1+\exp(-(\hat{\alpha}_2 + \hat{\beta}_1 x_1^{(1)} + \hat{\varepsilon}_2^{(1)}))} \\ \vdots & \vdots \\ \frac{1}{1+\exp(-(\hat{\alpha}_1 + \hat{\beta}_1 x_1^{(N)} + \hat{\varepsilon}_1^{(N)}))} & \frac{1}{1+\exp(-(\hat{\alpha}_2 + \hat{\beta}_1 x_1^{(N)} + \hat{\varepsilon}_2^{(N)}))} \end{bmatrix}$$

where $x_1^{(n)}$ represents a simulated value of X_1 for the n^{th} patient, $n = 1,\ldots,N$.

3. Response Simulation

Here it is helpful to visualize each row of SimProb0 and SimProb1 as a partitioning of the unit interval [0,1] into k subintervals, or bins. For simplicity of notation, we can rewrite the representative matrix of figure 1 from step 12 as

$$\begin{bmatrix} \partial bin1^{(1)} & \partial bin2^{(1)} \\ \vdots & \vdots \\ \partial bin1^{(N)} & \partial bin2^{(N)} \end{bmatrix}$$

where the notation ∂ may be read as "the upper bound of". Then, for each treatment group (SimProb0 and SimProb1 represent placebo and drug groups respectively), we have N different partitionings of [0,1]. Illustrating this with our representative matrix from figure 1, we have

$$0 < \partial bin1^{(n)} < \partial bin2^{(n)} < 1; n = 1,\ldots,N$$

Step 13: Generate two $N \times 1$ vectors $u_{placebo}$ and u_{drug} whose entries are sampled from the uniform(0,1) distribution, and augment SimProb0 and SimProb1 with $u_{placebo}$ and u_{drug} respectively. Call these augmented matrices SimAug0 and SimAug1 respectively. Letting $u = [u_1 \ldots u_N]^t$ represent one such vector, the matrix representative of the patients' cumulative probabilities in figure 1 becomes

$$\left[\begin{array}{cc|c} \partial bin1^{(1)} & \partial bin2^{(1)} & u_1 \\ \vdots & \vdots & \vdots \\ \partial bin1^{(N)} & \partial bin2^{(N)} & u_N \end{array}\right]$$

where $u_n \sim$ uniform(0,1) for $n = 1,\ldots,N$. Finally, in order to simulate response outcomes for each treatment group, we must determine the bins into which the simulated uniform(0,1) numbers fall. This can quickly be determined with a rank operator R. To understand how R works, suppose [.2 .8 | .66] is a row in the above augmented matrix. Then R([.2 .8 | .66]) = [1 3 | 2] indicates that .66 falls into the second bin, and hence represents a simulated ordinal response of 2. By letting R operate on each row of SimAug0 and SimAug1, the last column of each R-transformed matrix can be interpreted as an N-dimensional vector whose entries are simulated ordinal responses for N patients.

Having simulated responses for each treatment group, the total counts per response level can be summarized in a table as follows.

Table 1. Frequencies of Simulated Responses

Treatment	Responses a_1	a_2	…	a_k	Marginal Totals
Control	$s_{0,1}$	$s_{0,2}$	…	$s_{0,k}$	$\sum_{j=1}^{k} s_{0,j} = N_{control}$
Drug	$s_{1,1}$	$s_{1,2}$	…	$s_{1,k}$	$\sum_{j=1}^{k} s_{1,j} = N_{drug}$

Following the example of figure 1, $k = 3$, reducing table 1 to three columns.

4. Computing the *p*-value

At the beginning of the program, the user is prompted for the type of test to be performed on each such contingency table. These currently include chi-square, Fisher's exact test, and the option to see a comparison of the powers from both types of *p*-values. More detail regarding the relation between a simulated *p*-value and the final power will be given in step 16.

It should be noted that although Fisher's exact test is provided as an option, these authors do not recommend its use for testing the difference between a control group and a drug group. The reason is that Fisher's exact test assumes that *all* marginal totals in table 1 are fixed for a given experiment. While it is true that the row totals $N_{control}$ and N_{drug} can clearly be fixed by the experimenter, it is not true that the experimenter can know, *a priori*, the marginal totals for each response category. These marginal totals vary from experiment to experiment.

Thus, in these experiments the true sample space of contingency tables is a superset of the set of contingency tables that can be formed under the constraints of Fisher's exact test. Since *p*-values generated from Fisher's exact tests are computed from a subset of the true sample space of contingency tables, such *p*-values fail to account for probabilities of tables arising from unconstrained marginal response totals, hence Fisher's exact test is inappropriate.

Step 14: Append the simulated *p*-value to a list (or start a list on the first pass through the algorithm).

Step 15: Repeat steps 3 through 14 an arbitrarily large number of times (predetermined by the user). Call this value of repetitions *r*.

Step 16: We now have a list of *r* *p*-values from which we will be able to approximate the power. Recall that each *p*-value in the list was obtained by testing the following null hypothesis for each simulated contingency table:

$$H_0: \begin{array}{c} \pi_{0,1} = \pi_{1,1} \\ \pi_{0,2} = \pi_{1,2} \\ \vdots \\ \pi_{0,k} = \pi_{1,k} \end{array}$$

where

$j = 1,\ldots,k$ represents the levels of the response variable,

$\pi_{0,j} =$ true population proportion having response level j under control treatment

$\pi_{1,j} =$ true population proportion having response level j under drug treatment

and estimators can be formed as follows:

$$\hat{\pi}_{i,j} = \left\{ \begin{array}{l} \dfrac{s_{i,j}}{N_{control}}, \text{ if } i = 0 \\ \dfrac{s_{i,j}}{N_{drug}}, \text{ if } i = 1 \end{array} \right\}, j = 1,\ldots,k \quad \text{(see table 1)}$$

Before the algorithm is run, the user is prompted to choose one of the following alternative hypotheses to be used for each pass through the algorithm:

$H_{a,1}$: drug response "dominates" control response

$H_{a,2}$: control response "dominates" drug response

$H_{a,3}$: "any" difference between control and drug responses

Specifically, this notion of dominance can be expressed in terms of comparing the cumulative proportions for each treatment group:

$$H_{a,1}: \pi_{0,1} < \pi_{1,1}, \pi_{0,1} + \pi_{0,2} < \pi_{1,1} + \pi_{1,2}, \ldots, \sum_{j=1}^{k-1} \pi_{0,j} < \sum_{j=1}^{k-1} \pi_{1,j}$$

$$H_{a,2}: \pi_{0,1} > \pi_{1,1}, \pi_{0,1} + \pi_{0,2} > \pi_{1,1} + \pi_{1,2}, \ldots, \sum_{j=1}^{k-1} \pi_{0,j} > \sum_{j=1}^{k-1} \pi_{1,j}$$

$H_{a,3}: \pi_{0,j} \neq \pi_{1,j}$ for at least one value of $j = 1,\ldots,k$.

Let H_a represent the alternative hypothesis the user has selected. Given this choice, it is important to note that `Proc Freq`, by default, makes no distinction between $H_{a,1}$ and $H_{a,2}$. Without loss of generality, assume $H_a = H_{a,1}$. Then although the user may anticipate this type of dominance to occur often in the simulated contingency tables, there is no guarantee that the simulated tables will always follow this trend in dominance. In fact, tables may, if only infrequently, be simulated that satisfy $H_{a,2}$ to such an extent as to produce a *p*-value < α. If this is allowed to occur, then we will have *p*-values < α in our list of *r* *p*-values that satisfy both $H_{a,1}$ and $H_{a,2}$. The result, as we will see, would be an artificial inflation of our approximation of the power. Therefore, every *p*-value (<α) that came from a simulated contingency table satisfying $H_{a,2}$ is set equal to 1. Similarly, if $H_a = H_{a,2}$, then every *p*-value (<α) that came from a simulated contingency table satisfying $H_{a,1}$ is set equal to 1. Finally, if $H_a = H_{a,3}$, then the experimenter is testing for a difference in either direction, and nothing is done to the list of *r* *p*-values.

Having made the appropriate adjustments to the simulated *p*-values, we can now approximate the power. For the list of *r* *p*-values, define an indicator variable as follows.

$$reject_t = \left\{ \begin{array}{l} 1, p - value < \alpha \\ 0, p - value \geq \alpha \end{array} \right\}, t = 1\ldots,r$$

Then

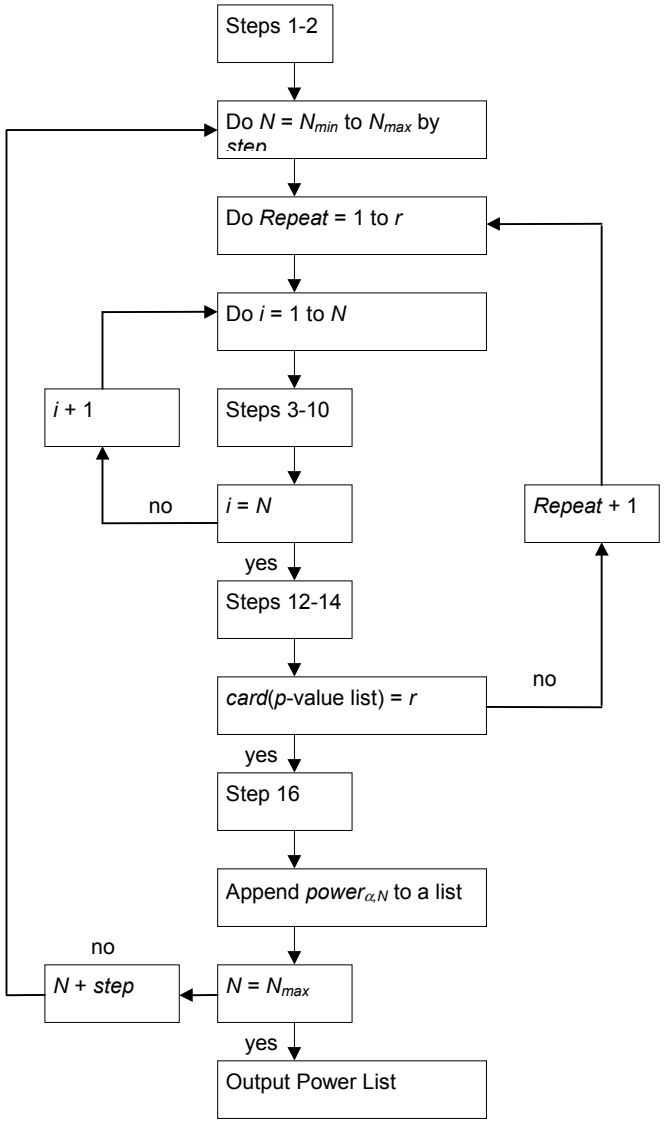
$$\lim_{r \to \infty} \frac{1}{r} \sum_{t=1}^{r} reject_t = \hat{power}_{\alpha, N_{control}}$$

That is, the larger the number of repetitions *r* the user chooses (see step 15), the more precise the approximation of the power (excluding any bias that may be inherent in the original sample data).

It should be noted that this algorithm has been generalized for the case of $N_{control} \neq N_{drug}$. This is done by nesting all the looping portions of this algorithm within another loop of sample sizes. Thus the investigator may construct power surfaces instead of power curves to explore the possibility of an unbalanced design yielding greater power.

The algorithm for $N_{control} = N_{drug}$ is summarized schematically in figure 2.

Figure 2. Algorithm Flowchart



RESULTS

To illustrate the algorithm in use, we have chosen data from a previous phase III study. In this study, drug X is used, which is derived from a certain protein found in the neutrophils (white blood cells) of all healthy human beings. The response variable to be presented, for our purposes of illustration, will be an amputation severity score. The independent variable to be presented is called Base Excess, which is a blood measure indicative of acidosis or alkalosis. Its empirical distribution can be described as follows: min(Base Excess) = -20.8, max(Base Excess) = 5.4, mean(Base Excess) = -8.13 and stddev(Base Excess) = 4.25. Finally, the response data may be summarized as follows:

Table 2. Summary of Raw Data

	Amputation Severity Score				
Treatment	0	1	2	3	Marginal Totals
Control	171	9	4	15	199
Drug X	167	9	4	6	186

Preliminary Modeling

The raw data allow us to construct an ordinal logistic regession model for each treatment group. It is from these two models, along with the approximate $N(-8.13, 4.25)$ distribution of Base Excess that we can simulate contingency tables, which in turn allow us to compute p-values, hence finally allowing us to build power curves. The two models' slopes, intercepts and variance-covariance maximum likelihood estimates are summarized in a SAS® data set as follows:

Table 3. Summary of Model Parameter Estimates

trt	_TYPE_	_NAME_	Intercept	Intercept2	Intercept3	BASEXS
0	PARMS	amputate	3.7654	4.2378	4.5075	0.1982
0	COV	Intercept	0.3476	0.3510	0.3529	0.0275
0	COV	Intercept2	0.3510	0.3788	0.3800	0.0282
0	COV	Intercept3	0.3529	0.3800	0.3996	0.0285
0	COV	BASEXS	0.0275	0.0282	0.0285	0.0025
1	PARMS	amputate	2.0362	2.7314	3.2648	-0.0182
1	COV	Intercept	0.2572	0.2537	0.2523	0.0266
1	COV	Intercept2	0.2537	0.3035	0.3010	0.0265
1	COV	Intercept3	0.2523	0.3010	0.3698	0.0265
1	COV	BASEXS	0.0266	0.0265	0.0265	0.0036

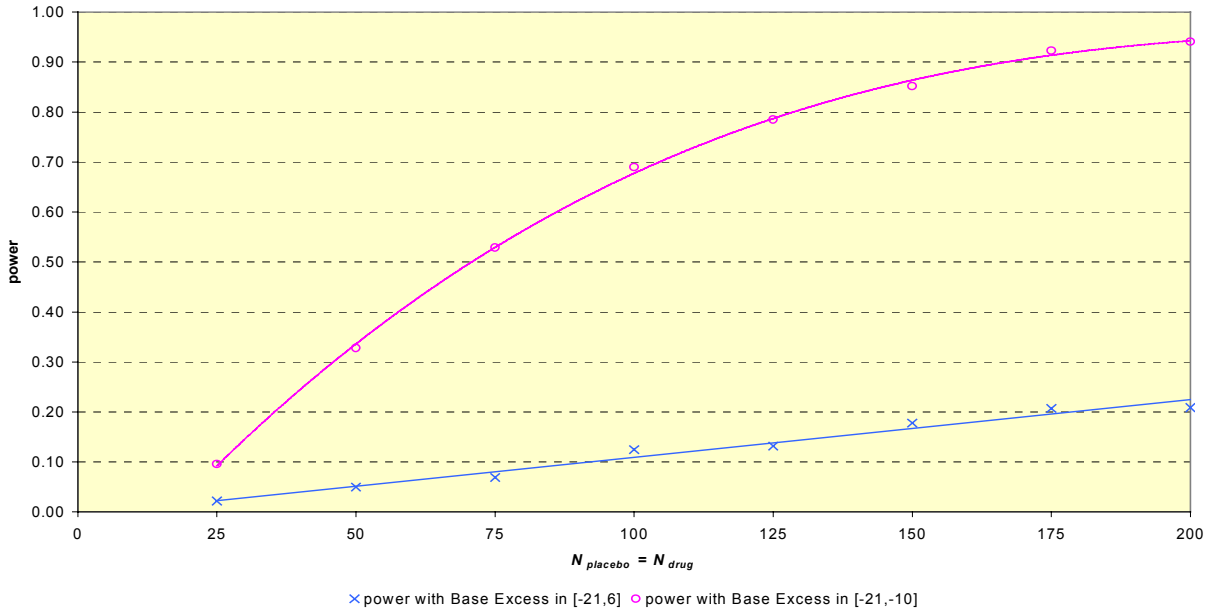
where trt = 0 indicates the parameter estimates for the model for the patients receiving placebos, and trt = 1 indicates the parameter estimates for the model for the patients receiving Drug X.

Powers Under Different Ranges of Base Excess

Although restrictions on Base Excess were not considered as part of the inclusion/exclusion criteria of the Phase III study from which the data came, an investigator may, for example, have reason to believe that certain ranges of Base Excess may be more predictive of differences in Amputation Severity Scores than others. That is, he can estimate the power of a new experiment under two conditions; one, in which patients' simulated values may range over the entire empirical distribution of Base Excess, i.e. –20.8 to 5.4. The other is a restricted range chosen by the investigator.

For example, if the investigator has evidence that leads him to suspect that patients who are more acidotic are more likely to benefit from Drug X, then he can restrict patients' simulated values of Base Excess to a more acidotic range, say –20.8 to –10. The investigator can then compare the power obtained under the restricted range to the power obtained under the unrestricted range to determine if the experiment under the restricted range is worth while, i.e. has a greater power than that obtained from the unrestricted range.

Figure 3. Power Curves Under Different Inclusion/Exclusion Criteria



Running the algorithm off of the same data, but under these two different inclusion/exclusion criteria, yields two different power curves (see figure 3).

By comparing these two power curves we see that experiments simulating patients who are more acidotic (lower Base Excess) are much more powerful in detecting a beneficial difference between placebo and Drug X treatments than experiments that simulate patients whose blood pHs span the entire spectrum.

The conclusion for the investigator is that if he wishes to design a new study based on the data gathered from the original phase III study, then restricting the enrolled patients to those who are more acidotic will be a more powerful study than one without such a restriction. If this type of screening is not practical or feasible, all is not lost. The next logical step for the investigator would be to find another measure 1) by which it is easy (easier) to screen patients, and 2) which is strongly correlated with Base Excess. This would then give the investigator an indirect way of screening out patients who are not sufficiently acidotic for the treatment to be worthwhile. In addition, it appears that enrollment of far fewer patients is required for any desired benchmark of power. This naturally makes such a study more affordable from a financial perspective. While such in depth concerns of physical feasibility and financial cost must be addressed, they are beyond the scope of this paper.

CONCLUSION

Using simulation techniques, we have demonstrated that it is now possible to perform power analysis for data which has been fitted with an ordinal (cumulative) logistic regression model. In particular, we have provided investigators with a new way of determining inclusion/exclusion criteria based on their impact on the power of a proposed study.

There are also two other techniques for approximating power for ordinal data. Whitehead (1993) has developed a sample size formula "based on a normal approximation, and [which] is accurate only when it generates moderate to large sample sizes." Hilton and Mehta (1993) have developed a different algorithmic approximation. They have developed a "network" estimator of power which randomly samples from the sample space of contingency tables and then averages the exact power computed for each table. In their appendix they demonstrate that both their "network" estimator and our "crude" Monte Carlo estimator are unbiased estimators of the power, but that the variance of their estimator is less than the variance of our estimator, i.e. relatively more efficient.

Although space has not permitted, we would like to mention that our algorithm also allows for dichotomous explanatory variables, and any combination of dichotomous and continuous explanatory variables. It allows for a continuous response variable as well. In each of these cases, the investigator has the capacity see how changing the inclusion/exclusion criteria will impact the power of his proposed study. It also allows for the case of a simple, multinomial response with no explanatory variables (see comments above regarding Mehta's network estimator).

Future plans for our algorithm include allowing for polychotomous explanatory variables. We also plan to allow the investigator an option for using nonparametric bootstrap calculations, as opposed to the parametric ones used now. Finally, we plan to allow the investigator to calculate (100 - α)% tolerance limits for power curves or surfaces generated.

Finally, as with all large applications, we are continually running validations and checks for internal consistency, and make no claim to have eradicated all bugs. Given this caveat, if you have any need or desire for such an application, please feel free to contact us.

REFERENCES

Agresti, A. (1999). Modelling Ordered Categorical Data: Recent Advances and Future Challenges, *Statistics in Medicine*, Vol. 18, 2191-2207.

O'Brien RG (1998). A Tour of UnifyPow: A SAS Module/Macro for Sample-Size Analysis, *Proceedings of the 23rd SAS Users Group International Conference*, Cary, NC, SAS Institute, 1346-1355.

Hilton J., Mehta, C. (1993). Power and Sample Size Calculations for Exact Conditional Tests with Ordered Categorical Data, *Biometrics* 49(2), 609-616.

Whitehead, J. (1993). Sample Size Calculations for Ordered Categorical Data, *Statistics in Medicine*, Vol. 12, 2257-2271.

Bull, S.B. (1993). Sample Size and Power Determination for a Binary Outcome and an Ordinal Exposure when Logistic Regression Analysis Is Planned, *American Journal of Epidemiology*, Vol. 137, No. 6.

SAS Institute Inc., *SAS/STAT® User's Guide*, *Version 6*, *Fourth Edition*, *Volume 2*, Cary, NC: SAS Institute Inc., 1989. 846 pp.

SAS Institute Inc., *SAS/IML® User's Guide*, *Version 8*, Cary, NC: SAS Institute Inc., 1999. 846 pp.

ACKNOWLEDGEMENTS

I must first and foremost thank Serge Guzy, Ph.D. for his conceiving of this algorithm and for his endearing enthusiasm and patience in explaining all the statistical intricacies of simulation to me. Again, along with Serge, I would also like to thank my wife Misha Vaughan Ph.D., her mother Lorraine Walker, R.N., Ph.D., and my father George Vaughan, M.D., for their moral support and encouragement in writing this, my first paper.

CONTACT INFORMATION

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

Charles M. Vaughan

Xoma US, LLC

2910 7th Street

Berkeley, CA 94710

(510) 644-1170 x2079

vaughan@xoma.com

cmvaughan@earthlink.net