

Calculating the power or sample size for the logistic and proportional hazards models

DAVID A. SCHOENFELD*†‡ and MICHAEL BORENSTEIN†‡

†Biostatistics Center, Massachusetts General Hospital, 50 Staniford Street, Boston, MA 02114, USA

‡Biostatistical Programming Associates Inc., 14 North Dean Street, Englewood, NJ 07631, USA

(Revised 11 July 2002; in final form 23 May 2004)

An algorithm is presented for calculating the power for the logistic and proportional hazards models in which some of the covariates are discrete and the remainders are multivariate normal. The mean and covariance matrix of the multivariate normal covariates may depend on the discrete covariates.

The algorithm, which finds the power of the Wald test, uses the result that the information matrix can be calculated using univariate numerical integration even when there are several continuous covariates. The algorithm is checked using simulation and in certain situations gives more accurate results than current methods which are based on simple formulae. The algorithm is used to explore properties of these models, in particular, the power gain from a prognostic covariate in the analysis of a clinical trial or observational study. The methods can be extended to determine power for other generalized linear models.

Keywords: Sample size; Power; Logistic model; Proportional hazards model; Generalized linear models; Multivariate normal integrals; Wald test

1. Introduction

Non-linear data models such as the proportional hazards and logistic models are commonly used in the analysis of medical data. Whittemore [1], Hsieh [2] and Hsieh *et al.* [3], give formulas for the power of the logistic model for continuous normal covariates. Schoenfeld [4] gives a sample size formula for the proportional hazards model for a randomized clinical trial where the distribution of the continuous covariates are independent of the distribution of a single binary covariate. Schmoor *et al.* [5] give an approximate formula for this case which does not depend on this restriction, and Bernardo *et al.* [6] give a more general approximate formula.

While these methods provide simple formula, none actually uses the true asymptotic distribution of the test statistic other than in limited circumstances. Thus, there are situations where these methods cannot be applied or where the approximations that these methods use will fail even for large samples.

*Corresponding author. Email: dschoenfeld@partners.org

This paper shows that the calculation of the asymptotic non-centrality parameter of the Wald statistic for an arbitrary number of normally distributed covariates only involves one-dimensional numerical integration. This allows the development of an algorithm for calculating power for these models. The methods described here also apply to the generalized linear models such as Poisson regression and polychotomous logistic regression.

Section 2 specifies the covariate distribution for which power will be calculated for both the models. Section 3 presents a theorem which is used to reduce the multivariate integrals involved in the calculation of the non-centrality parameter into univariate integrals. The proof is given in Appendix A. Section 4 describes the elements of the power calculation. In section 5, power calculations based on this algorithm are compared to those in the literature. Sections 7 and 8 use the algorithm to investigate properties of the models, section 9 discusses its computational efficiency, and section 10 discusses how the algorithm can be used.

2. Description of the covariate model

Most covariate models are partially discrete and partially continuous. To simplify the description of the model assume that the first few covariates are discrete and the remaining are continuous. The vector of discrete covariates can be thought of as dividing the patient population into K distinct groups. For instance, suppose a study has two treatments and, in addition, an ordinal covariate with three values indicating disease severity. Then for the logistic model, the discrete covariate vector would have length 3, two covariates that code treatment and one that codes the disease severity. In this case, there would be six values of the discrete covariate vector. Since the proportional hazards model does not need a constant term, the discrete covariate vector would have length 3 but there would still be six distinct groups.

We assume that for each of the K distinct groups the continuous covariates can have a different multivariate distribution. We assume that this distribution is multivariate normal, although in certain cases this requirement could be relaxed.

To fix notation let \mathbf{X} be the covariate and first part, composed of a vector of discrete variables, has a finite number of values $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K$, where each occur with relative frequency w_1, \dots, w_K . For each value \mathbf{v}_s , $s = 1, \dots, K$, the second part of the covariate vector has a multivariate normal distribution with mean μ_s and variance covariance matrix Σ_s .

This is a general way of describing a joint multinomial, multivariate normal distribution. In the hypothetical example, suppose that there were two continuous covariates, say age and blood pressure. Then $K = 6$ and $\mathbf{v}_1 = (1, 0, 1)'$, $\mathbf{v}_2 = (1, 1, 1)'$, $\mathbf{v}_3 = (1, 0, 2)'$, \dots , if the patient groups were not randomized then the mean age and the blood pressure of group 1, $\mu_1 = (\mu_{1,1}, \mu_{1,2})'$ might be different than that of group 2, $\mu_2 = (\mu_{2,1}, \mu_{2,2})'$. The variance covariance matrices Σ_s for $s = 1, \dots, 6$ could also be different.

3. Theoretical basis for the algorithm

As we show later, the expression for the power and sample size involves finding the information matrix for the model. The calculation of this matrix will depend on the calculation of $E\{\mathbf{X}\mathbf{X}'f(\beta'\mathbf{X})\}$ where f is a specified function which depends on the model. The algorithm depends on a theorem that shows that this multivariate integral can be reduced to a univariate integral, which can be evaluated numerically. The proof, which is given in Appendix A, is based on the fact that the conditional expectation of $\mathbf{X}\mathbf{X}'f(\beta'\mathbf{X})$ given $\beta'\mathbf{X}$ has a closed form solution in terms of β' .

THEOREM 1 Let \mathbf{u}_s be $(\mathbf{v}'_s, \mu'_s)'$ and \mathbf{D}_s be the square matrix with row dimension the same as \mathbf{u}_s whose lower right hand corner is Σ_s and is zero elsewhere. (Note that \mathbf{u}_s and \mathbf{D}_s are the mean and variance-covariance matrix of the covariate vector for patients in group s .)

Define $\sigma_s = \sqrt{\beta' \mathbf{D}_s \beta}$, $\tau_s = \beta' \mathbf{u}_s$ and $\gamma = \mathbf{D}_s \beta / \sigma_s$. Furthermore, let $e_i(\tau, \sigma) = \int z^i f(\sigma z + \tau) \phi(z)$, where $\phi(z)$ is the normal probability density function.

Then

$$\begin{aligned} E\{\mathbf{X}\mathbf{X}' f(\beta' \mathbf{X})\} &= \sum_{s=1}^K w_s \{(\mathbf{u}_s \mathbf{u}'_s + \mathbf{D}_s - \gamma_s \gamma'_s) e_0(\tau_s, \sigma_s) \\ &\quad + (\mathbf{u}_s \gamma'_s + \gamma_s \mathbf{u}'_s) e_1(\tau_s, \sigma_s) + \gamma_s \gamma'_s e_2(\tau_s, \sigma_s)\} \\ E\{\mathbf{X} f(\beta' \mathbf{X})\} &= \sum_{s=1}^K w_s \{\mathbf{u}_s e_0(\tau_s, \sigma_s) + \gamma_s e_1(\tau_s, \sigma_s)\} \\ E\{f(\beta' \mathbf{X})\} &= \sum_{s=1}^K w_s e_0(\tau_s, \sigma_s). \end{aligned} \quad (1)$$

4. Algorithm to calculate power

The logistic model specifies that the probability of success for the i th experimental unit is given by $\text{logit}(\beta' \mathbf{X}) = p$, whereas the proportional hazards model specifies that the hazard function is $\lambda(t) \exp(\beta' \mathbf{X})$.

Assume that the null hypothesis to be tested is of the form $H_0: \mathbf{A}(\beta - \beta_0) = 0$. This allows compound null hypotheses as well as hypotheses about single parameters. The Wald test statistic is

$$T = (\hat{\beta} - \beta_0)' \mathbf{A}' (\mathbf{A} \hat{\mathbf{I}}^+ \mathbf{A}')^{-1} \mathbf{A} (\hat{\beta} - \beta_0),$$

where $\hat{\mathbf{I}}$ is the empirical information matrix evaluated at $\hat{\beta}$ the maximum likelihood estimate of β . Under the null hypothesis T has a χ^2 distribution with $\text{rank}(\mathbf{A})$ degrees of freedom.

Let β_a be the value of β under the alternative hypothesis and \mathbf{I} is the information matrix for n observations evaluated at β_a . Then, under the alternative hypothesis, T has a non-central χ^2 distribution with non-centrality parameter

$$\delta = (\beta_a - \beta_0)' \mathbf{A}' (\mathbf{A} \mathbf{I}^+ \mathbf{A}')^{-1} \mathbf{A} (\beta_a - \beta_0), \quad (2)$$

Thus, the power is calculated from δ using a program for calculating the non-central χ^2 distribution function. The only part of expression (2) that needs to be calculated is \mathbf{I} .

Under the logistic model,

$$\mathbf{I} = n E \left[\mathbf{X}\mathbf{X}' \frac{\exp(\beta' \mathbf{X})}{\{1 + \exp(\beta' \mathbf{X})\}^2} \right].$$

Let

$$f(u) = \left[\frac{\exp(u)}{\{1 + \exp(u)\}^2} \right]. \quad (3)$$

Then

$$\mathbf{I} = n E(\mathbf{X}\mathbf{X}' f(\beta' \mathbf{X})), \quad (4)$$

which is calculated using Theorem 1.

The derivation above can be easily extended to the generalized linear models, the likelihood is of the form $l = \sum \log(h(y_i, \mu_i, \phi))$ where ϕ is a dispersion parameter which does not depend on the covariates and may or may not be present. In addition, $g(\mu_i) = \beta' X_i$ is a link function that relates the mean parameter to the covariates. As in the logistic model, the value of I for such a model can be calculated using Theorem 1.

The proportional hazards model is somewhat more complicated. Let $r_j(t)$ be the indicator function of whether the j th patient is observed (alive and not censored) at time t . The empirical information matrix is given by

$$I = \sum \frac{\left\{ \sum \mathbf{X}\mathbf{X}' \exp(\beta' \mathbf{X}) r_j(t) \right\} \left\{ \sum \exp(\beta' \mathbf{X}) r_j(t) \right\} - \left\{ \sum \mathbf{X} \exp(\beta' \mathbf{X}) r_j(t) \right\} \left\{ \sum \mathbf{X}' \exp(\beta' \mathbf{X}) r_j(t) \right\}}{\left\{ \sum \exp(\beta' \mathbf{X}) r_j(t) \right\}^2}, \quad (5)$$

where the first sum is over death times and the other sums are over all patients.

Divide the period of observation into k intervals with endpoints t_1, t_2, \dots, t_{k+1} and assume that the hazard function, $\lambda(t)$, is constant within each interval. Let $C(t)$ be the censoring distribution and $\Lambda(t)$ be the cumulative hazard function.

First consider the first term in the numerator of 5, $\sum \mathbf{X}\mathbf{X}' \exp(\beta' \mathbf{X}) r_j(t)$. Its value during the h th interval is approximated by its expectation at time $t = (t_h + t_{h+1})/2$. Conditioning first on \mathbf{X} , replace $r_j(t)$ by the probability that a patient with covariate vector \mathbf{X} is observed at time t , $E(r_j(t)|\mathbf{X}) = \{1 - C(t)\} \exp\{-\Lambda(t) \exp(\beta' \mathbf{X})\}$. Thus, the sum is approximated by $nE[\mathbf{X}\mathbf{X}' \exp(\beta' \mathbf{X}) \{1 - C(t)\} \exp\{-\Lambda(t) \exp(\beta' \mathbf{X})\}]$. This can be found using Theorem 1. The other terms are found similarly. Denote the summands of 5 for the intervals 1, 2, \dots , k as S_1, S_2, \dots, S_k .

Let P_h be the probability that a patient dies in the h th interval. Then

$$P_h = E \left(\int_{t_h}^{t_{h+1}} \exp[-\Lambda(t_h) + \lambda_h t] \exp(\beta' \mathbf{X}) \lambda \exp(\beta' \mathbf{X}) \{1 - C(t)\} dt \right).$$

The inner integral has a closed form expression when patients enter the trial uniformly and are followed for a fixed length of time and, in addition, there is an exponential drop out rate. The expression for this term can be found using a symbolic integration program such as Mathematica or Maple. Then P_h can be found using Theorem 1.

The expected value of \mathbf{I} is approximated by $\sum_{s=1}^k n P_h S_h$ which can be made arbitrarily accurate by increasing the number of intervals.

5. Performance for logistic regression

There is no formula described in the literature for obtaining sample size when there are both discrete and continuous covariates. Thus the situation, common in the analysis of clinical trials and observational studies, when logistic regression is used to compare patient groups 'correcting' for some possibly confounding covariate, is not described. However, it is instructive, to see how our algorithm compares with other algorithms in the literature for the situations where these algorithms can be used. Whittemore [1], Hsieh [2] and Hsieh *et al.* [3], give formulas for the power of the logistic model for continuous normal covariates. First, we compare their formula to ours for a single normal covariate, aside from a constant term, and then for the situation where there are two covariates.

5.1 Comparison for a single covariate

We consider the case with a single continuous covariate with a standard normal distribution and a constant term. Define $\mathbf{X} = (x_1, x_2)'$, where $x_1 = 1$ and x_2 has a standard normal distribution. The null hypothesis is that $\beta_2 = 0$ with the alternative that $\beta_2 \neq 0$, we present power for a two sided test with a significance level of 0.025 and a power of 0.90.

Then, using the notation of Theorem 1, there is only one value of s and $\mathbf{u}_s = (1, 0)'$,

$$\Sigma^a = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

and $\sigma = \beta_2$, $\tau = \beta_1$ and $\gamma = (0, 1)'$. Thus, writing $e_i(\tau, \sigma)$ as e_i ,

$$\mathbf{I} = n \begin{Bmatrix} e_0 & e_1 \\ e_1 & e_2 \end{Bmatrix}.$$

Substituting these values into equation (1), the standard deviation of $\hat{\beta}_2$ can therefore be shown to be equal to

$$v = \left\{ \frac{e_0}{e_2 e_0 - e_1^2} \right\}^{1/2},$$

and the sample size is given by

$$N = \frac{(z_{0.975} + z_{0.9}v)^2}{\beta_2^2}. \quad (6)$$

Whittemore proposes the following formula, for situations where the response probability is small,

$$N = \frac{\{z_{0.975} + z_{0.9} \exp(-\beta_2^2/4)\}^2}{\exp(\beta_1)\beta_2^2} \{1 + 2 \exp(\beta_1)\delta\},$$

where

$$\delta = \frac{1 + (1 + \beta_2^2) \exp(5\beta_2^2/4)}{1 + \exp(-\beta_2^2/4)}.$$

Whittemore's formula is derived in a similar manner as ours except that

$$E \left[\mathbf{X}\mathbf{X}' \frac{\exp(\beta'\mathbf{X})}{\{1 + \exp(\beta'\mathbf{X})\}^2} \right]$$

is approximated up to terms of order $O(\exp(2\beta))$ rather than evaluated numerically. Thus, Whittemore's formula should be accurate as long as $\exp(\beta_1)$ is small.

The formula given in Hsieh [2], used in nQuery advisor, which he ascribes to Whittemore, replaces $\exp(\beta_1)$ wherever it appears by

$$P = \frac{\exp(\beta_1)}{1 + \exp(\beta_1)}.$$

Hsieh *et al.* [3] also propose

$$n = \frac{(z_{0.975} + z_{0.9})^2}{P(1 - P)\beta_2^2}, \tag{7}$$

which is justified because the score test for a univariate logistic regression is asymptotically equivalent to a *t*-test comparing responders and non-responders.

Each of the formula above can be rearranged to give a power formula for a fixed sample size. The first section of table 1 shows the power predicted for each of these methods, and by simulation for $N = 500$, $\beta_1 = -3, -2, -1, 0, 1$ and $\beta_2 = 0.68, 0.45, 0.33, 0.29, 0.33$. The values for β_1 were chosen to give a wide range of response probabilities. Since Whittemore's formula is designed for small response probabilities it is important to know where it breaks down and whether the proposed method is superior for larger response probabilities. The value of N was chosen to be large in order to test the asymptotic properties of the sample size formulas and not their small sample properties, which is outside the scope of this paper. Finally, the values of β_2 were chosen to give 90% power with a two-sided $\alpha = 0.05$ significance level using formula (7) since the correct estimation of sample size for a power of 80–90% is the most common us of sample size formulae. The accuracy of a simulation of power based on 1000 samples is ≈ 0.01 .

In the univariate case, in table 1, the algorithm gives reasonable results over the full range of values. The method used in nQuery advisor was inaccurate for intermediate values of P while Whittemore's method is inaccurate for high values of P . The empty spaces in the nQuery rows occur when the parameters are outside the range supported by nQuery. It is notable that Whittemore's method works well for values of P as high as 0.27. The method based on the

Table 1. Difference between calculated and simulated power.

β_1	-3	-2	-1	0	1
Odds ratio ($\exp(\beta_2)$)	1.97	1.57	1.39	1.39	1.38
P	0.047	0.12	0.27	0.50	0.73
Univariate case					
Simulation	0.92	0.92	0.91	0.90	0.91
Algorithm	+0.00	-0.02	-0.02	-0.02	-0.02
Whittemore	-0.02	-0.02	0.00	+0.05	+0.09
nQuery	-0.03	-0.05	-0.06	-0.01	
Hsieh (<i>t</i> -test)	-0.02	-0.02	-0.02	-0.00	-0.01
Multivariate $\rho = 0, \beta_3 = 0.9$ case					
Simulation	0.96	0.94	0.89	0.98	0.87
Algorithm	-0.01	-0.03	-0.03	-0.01	-0.01
Whittemore	-0.08	-0.03	+0.09	+0.02	+0.13
nQuery	-0.07	-0.07	-0.03	+0.01	
Hsieh (<i>t</i> -test)	-0.06	-0.06	+0.02	+0.01	+0.03
Multivariate $\rho = 0.5, \beta_3 = 0.9$ case					
Simulation	0.92	0.83	0.76	0.71	0.75
Algorithm	-0.01	0.00	-0.01	-0.01	0.00
Whittemore	-0.06	+0.12	+0.20	+0.27	+0.25
nQuery	-0.15	-0.08	-0.01	+0.03	
Hsieh (<i>t</i> -test)	-0.13	-0.02	+0.05	+0.05	+0.06
Multivariate $\rho = 0.5, \beta_3 = 0$ case					
Simulation	0.86	0.83	0.80	0.95	0.80
Algorithm	-0.02	-0.02	-0.01	0.00	-0.01
Whittemore	-0.07	-0.03	+0.02	+0.04	+0.18
nQuery	-0.08	-0.07	-0.05	-0.00	
Hsieh (<i>t</i> -test)	-0.06	-0.02	+0.01	+0.02	+0.01

t -test worked very well as would be expected, since it does not use any approximation other than the asymptotic equivalence of the t -test and univariate logistic regression.

5.2 Comparison for multiple covariates

Suppose interest focuses on β_2 the coefficient of the first continuous covariate but there is an additional covariate in the model that is correlated to x_2 . Assume that this covariate, x_3 has mean 0 and variance 1 and the correlation of x_2 and x_3 is ρ . Although proposed algorithm has been incorporated into a general purpose program which is available from the authors, we include an explicit formula for this case in order to provide a comparison to the other methods. We found this formula from Theorem 1 using a symbolic manipulation program. First $\sigma = \sqrt{\beta_2^2 + \beta_3^2 + 2\beta_2\beta_3\rho}$ and $\tau = \beta_1$, then defining e_0, e_2, e_3 as in Theorem 1 the standard deviation of β_2 is given by

$$v = \left\{ \frac{\beta_3^2(e_1^2 - e_0e_2) + 2\beta_2\beta_3(e_1^2 - e_0e_2)\rho + \beta_2^2(e_1^2\rho^2 - e_0e_2\rho^2 + e_0^2(-1 + \rho^2))}{e_0(-e_1^2 + e_0e_2)(\beta_2^2 + \beta_3^2 + 2\beta_2\beta_3\rho)(-1 + \rho^2)} \right\}^{1/2}.$$

The sample size then can be found using formula (6).

Whittemore's formulation for the sample size in this case is,

$$N = \frac{(1 + 2 \exp \beta_1)(z_\alpha v_1^{1/2} + z_\beta v_2^{1/2})^2}{\exp(\beta_1)\beta_2^2}$$

where in this case

$$v_1 = \left[\exp \left\{ \frac{\beta_3^2}{2} \right\} (1 - \rho^2) \right]^{-1}$$

and

$$v_2 = \left[\exp \left\{ \frac{\beta_2^2 + \beta_3^2 + 2\beta_2\beta_3\rho}{2} \right\} (1 - \rho^2) \right]^{-1}.$$

This formula is not equivalent to the previous formula when $\beta_3 = 0$ because she does not compute the 'correction factor' in the same way. Hsieh corrects the univariate sample size by dividing by a 'variance inflation factor,' which in this case would equal $(1 - \rho^2)$. He justifies this by analogy with the linear regression and by simulation. It appears that his simulations are conducted with $\beta_3 = 0$. The method used in nQuery advisor is to adjust their univariate sample size by the same 'variance inflation factor.'

It is instructive to see how these methods perform when $\beta_3 \neq 0$. Consider the simplest case when $\rho = 0$, shown in the second section of table 1. Then the nQuery and the Hsieh method give the same power calculation as before. In both cases these methods underestimate the power, while Whittemore's method and the algorithm, which incorporate β_3 , give answers much closer to the simulation. Whittemore's method breaks down, as expected, when $\exp(\beta_1)$ is high.

Finally, we consider the case when $\rho = 0.5$, shown in the last two sections of the table, the Hsieh method gives adequate approximations as long as $\beta_3 = 0$. When $\beta_3 \neq 0$ only the algorithm gives adequate approximations.

5.3 Discrete analysis of covariance

The most common use of the logistic model is for comparisons of response rates between treatment groups in clinical trials or disease rates between exposure groups in observational

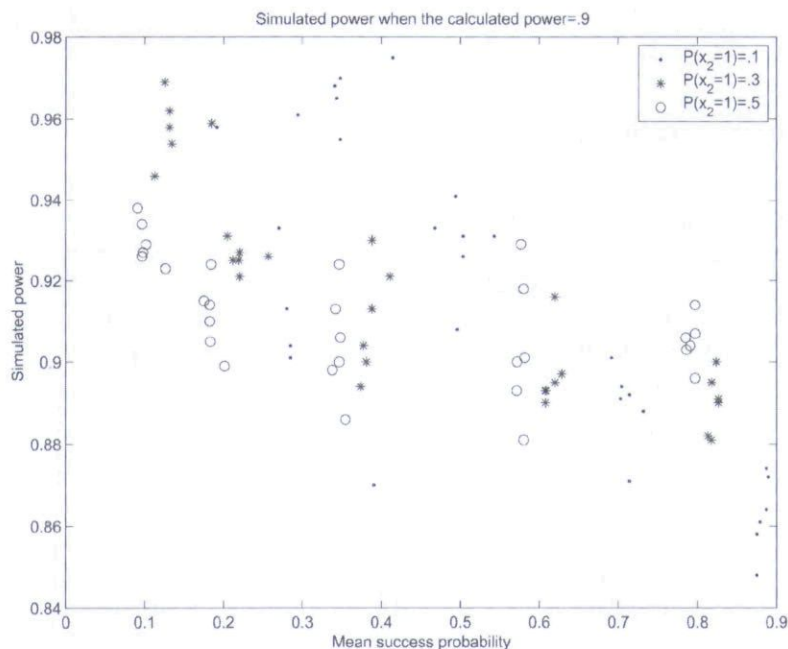


Figure 1. Simulated power when the calculated power = 0.9.

studies ‘controlling’ for possibly confounding covariates. We ran a simulation to assess the accuracy of the algorithm. We consider the case where there is one continuous covariate and one discrete covariate. We assume that $\mathbf{X} = (x_1, x_2, x_3)$, where $x_1 = 1$, $x_2 = 1$ for the treatment or exposure group and zero otherwise, and x_3 is the (possibly) confounding covariate with a normal distribution. We assume the mean of x_3 is $-\mu/2$ when $x_2 = 0$ and $\mu/2$ when $x_2 = 1$. Thus if $\mu = 0$, x_3 does not confound the analysis because its distribution is independent of x_2 . We assume the variance of x_3 is 1.

Let the proportion of patients in the first group, when $x_1 = 1$ be p . To simplify the situation, we consider values of β_2 where the power would be 90% when total sample size is 500. Thus the parameters that affect the power are β_1 , β_3 , μ and p . We used the values of $\beta_1 = -3, -2, -1, 0, 1$ described previously, we set p to be 0.1, 0.3 and 0.5 and we set β_3 to be either zero or the values used previously, to simulate the case where the covariate is influential, μ was set to $-1, 0$ or $+1$. Figure 1 shows the result of this simulation. The accuracy is good with most simulated powers being within 0.02 of the nominal power of 0.9. The samples where the success probabilities were high or low and where $p = 0.1$ were the situations where the simulated power was most different from the nominal power. A similar relationship was found using β_1 on the abscissa.

6. Performance for the proportional hazards model

Schoenfeld [4] gives a sample size formula for a survival ‘Analysis of Covariance’ where there is a treatment, and a set of continuous explanatory variables. This formula is valid for the case of a randomized study where the distribution of the covariates is independent of the treatment. Schmoor *et al.* [5] and Bernardo *et al.* [6] show, when there is no censoring, that this formula can be corrected for the situation where independence does not hold by the addition

of a 'variance inflation factor' similar to that of Hsieh [2]. With censoring they argue that it still provides a reasonably good approximation.

To evaluate the algorithm, consider the case where x_1 is a binary variable that denotes treatment and x_2 is a normal random variable with mean $\mu/2$ or $-\mu/2$ for x_1 equal 0 or 1, respectively. Suppose further that the censoring distribution is uniform on the interval (1, 4), and that the probability of death for a patient with $x_1 = 0$, $x_2 = 0$ is d and the underlying hazard is constant. We consider trials with 500 patients where $p = P(x_1 = 1) = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $d = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and β_1 adjusted to give 90% power for an increase and decrease in mortality when $x_1 = 1$, $\beta_2 = \{0.0 \pm 0.33 \pm 1.0\}$ and $\mu = -1, 0, 1$.

Note that Schoenfeld's formula is only applicable when $\mu = 0$. In that case, under the conditions above the formula would be

$$\frac{(z_\beta + z_\alpha)^2}{Dp(1-p)\beta_1^2}.$$

The value of D in this formula and in the formulas which use a 'variance inflation factor' is the proportion of patients dying on the trial. It is not equal to d above which is the proportion when $x_1 = 0$, $x_2 = 0$. To get from one to the other involves first solving the equation

$$d = 1 - \frac{1 - e^{-3\theta}}{3e^\theta},$$

which is the expression for d as a function of the exponential parameter θ when the censoring distribution is uniform on the interval (1, 4). Then

$$D = (1-p) \int \frac{1 - e^{-3\theta e^{\beta_2 x}}}{3e^{\theta e^{\beta_2 x}} \theta e^{\beta_2 x}} \phi(x) dx + p \int \frac{1 - e^{-3\theta e^{\beta_1 + \beta_2 x}}}{3e^{\theta e^{\beta_1 + \beta_2 x}} \theta e^{\beta_1 + \beta_2 x}} \phi(x) dx.$$

Figure 2 shows the performance of the algorithm. The data has been divided into 'unbalanced data' shown with a '+' and 'other data' shown with a '*'.

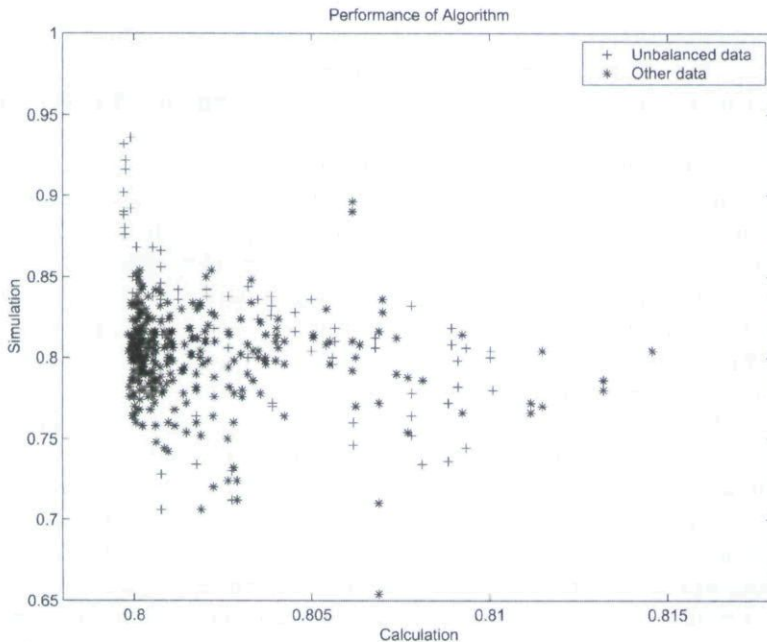


Figure 2. Performance of algorithm.

$d = 0.1$. Most of the cases where the algorithm failed to perform properly were in the unbalanced case. The problem is that despite a sample size of 500 small sample properties predominate.

7. Use of the algorithm to study the properties of logistic regression

One advantage of having an algorithm that quickly calculates the power and sample size under a wide range of circumstances is that it enables research into the properties of these models. We provide some examples here.

7.1 Power for large values of β

One problem with Wald's test is that for very large parameter values the power decreases. The test is therefore inconsistent although the power may reach one well before it begins to decrease. It is easy to see what happens in the case where the covariate has two values $(-1, +1)$ each with probability 0.5. Assume that there is no constant term. Then the non-centrality parameter is

$$\beta^2 \left[\frac{\exp(\beta)}{\{1 + \exp(\beta)\}^2} + \frac{\exp(-\beta)}{\{1 + \exp(-\beta)\}^2} \right]$$

which approaches zero as β approaches infinity. Thus, the power will decrease for very large values of β .

7.2 Effect of covariates on power

Covariates are added to the logistic regression models in randomized clinical trials to increase the power of a treatment comparison. In observational studies covariates are also included to correct for confounding. The algorithm can be used to determine the effect of both these uses on power.

Suppose that the logarithm of the odds ratio is $\mu + \beta_1 x_1 + \beta_2 x_2$ where x_1 is a binary treatment indicator and x_2 is a normally distributed covariate. Suppose we wish to test the hypothesis that $\beta_1 = 0$ and x_2 is included in the model either to increase the power of the test of $\beta_1 = 0$ or to remove the confounding caused by an association between x_1 and x_2 .

Assume that the mean of x_2 is $\Delta/2$ or $-\Delta/2$ depending on whether $x_1 = 0$ or $x_1 = 1$. Assume that the standard deviation of x_2 is 1. We wish to see how the power of a test $H_0: \beta_1 = 0$ is affected by β_2 and Δ .

First consider the case where $\Delta = 0$. This occurs in a randomized study where the mean of the normally distributed variable is the same for each treatment. The calculation of the effect of including a covariate in the analysis of a randomized study is complicated by an 'Amalgamation Paradox' [7]. The value of β_1 is dependent on whether x_2 is included in the model. If x_2 is in the model then β_1 estimates the log of the odds ratio for fixed values of x_2 . When x_2 is not included in the model β_1 measures the odds ratio for the entire population which is less than the odds ratio for fixed values of x_2 whenever $\beta_2 \neq 0$.

In order to find the effect of adding a covariate x_2 as a function of its coefficient β_2 , we fix β^* which we define as the attenuated treatment effect without the covariate. For each value of β_2 , we solve for β_1 which is a function of β_2 and β^* and calculate the power using the

algorithm. Define

$$p(x_1) = \int \frac{\exp(\mu + \beta_1 x_1 + \beta_2 z)}{1 + \exp(\mu + \beta_1 x_1 + \beta_2 z)} \phi(z) dz.$$

Then

$$\beta^* = \log \left[\frac{p(1)\{1 - p(2)\}}{p(2)\{1 - p(1)\}} \right]$$

which is the attenuated, uncorrected log-odds ratio.

In an example with $\mu = 0$, if we set the sample size to get a power of 0.70 when $\beta_3 = 0$, if $\beta_3 = 1$, an odds ratio of 2.7, the power increases to 0.84 and if $\beta_3 = 2$, an odds ratio of 7.4, the power increases to 0.9. The increase is nearly linear so a figure is not included. If β_1 , rather than β^* , were held constant the power decreases with increased values of β_2 . Notice that the power gain is small for even relatively large values of β_2 , for instance suppose that a one standard deviation increase in x_2 increases the odds ratio by 3. Then $\beta_2 = 1.1$ and the power increases from 0.79 to 0.85, the same power increase one would see from an 18% increase in sample size. These calculations are based on total sample size of 100 with $\beta_1 = 1.15$. These calculations would imply that the inclusion of covariates have a minor effect on power unless they are extremely prognostic.

Next we consider the effect of confounding on power, and to do this we plot power as function of increasing values of Δ . Consider the case where $\beta_2 = 0$ and where $\beta_2 = 1$. Fix β_1 so that $\beta^* = 1.15$ when $\Delta = 0$.

The result of this calculation is as expected: a drop in power as the extent of confounding increases (see figure 3).

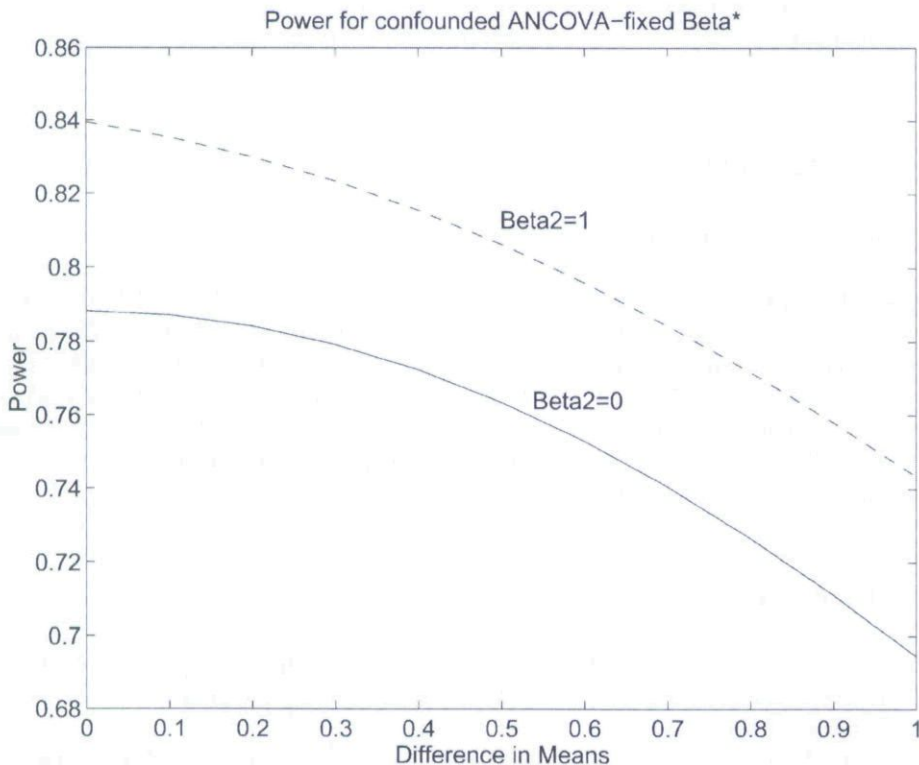


Figure 3. Power for confounded ANCOVA-fixed β^* .

8. Use of the algorithm to study the properties of the proportional hazards model

Unlike the logistic model, the proportional hazards model is no longer valid if a covariate is omitted. If the log hazard ratio is $\beta_1 x_1 + \beta_2 x_2$ and $\beta_2 > 0$ then the hazard ratio for the two values of x_1 will not be constant over time [8]. Despite the fact that the model does not fit the data, it still provides a valid test of $H_0: \beta_1 = 0$ when x_2 is independent of x_1 because under the null hypothesis the proportional hazards model holds. In order to determine the affect of covariate correction on power let β^* be the limit of $\hat{\beta}_1$ as the sample size gets large. Then it is necessary to calculate the power as β_2 increases with a fixed value of β^* .

Gail *et al.* [9] give an expression for β^* as a function of β_1 and β_2 involving several numerical integrals which can be simplified using Theorem 1. In order to provide an example of the effect of covariate adjustment, we study the most common situation which is a clinical trial. We assume that x_1 has a value of 0 or 1 with probability of 0.5, as would occur in a two arm clinical trial. We assume that the clinical trial has 24 months of accrual and 12 months of follow up and the underlying hazard (with $x_1 = 0$) is 0.03. This would give a 36 months survival of 34%. We assume that $\beta^* = -0.38$. This was chosen to be a typical value for a large trial, without a covariate the power would be 80% if for a trial with a total of 500 patients, this value corresponds to a 32% decrease in hazard.

Now suppose that there was an important covariate, say $\beta_2 = 0.5$, a hazard ratio of 1.65. Then β_1 would have to be 0.41 to achieve a value of β^* of 0.38 and the power would be 86%. Suppose the covariate had a hazard ratio of two ($\beta_2 = 0.6931$). Then, following the same logic, the power would be 0.89. The covariate had a hazard ratio of 3 then the power would be 0.97. Note that in the latter case if $\beta_1 = 0.38$ the power is only 83%.

Correcting for a common important covariate can substantially increase power in a clinical trial but the power increase is not apparent unless you take into account the fact that the hazard ratio before the introduction of the covariate will be less than the hazard ratio after its introduction.

9. Computational efficiency

An alternative method of computing power is simulation. Programming a general simulation for these problems takes approximately the same amount of code as the programs using the proposed algorithm but simulation programs are easier to conceptualize and debug. A sample problem was run using the algorithm and a simulation with 500 replications. The programs were written in Matlab and run on a 800 MHz computer. A logistic model took 0.63 s using the algorithm and 2.2 s using a simulation. A proportional hazards model took 1.3 s with 5 intervals and 2.6 s with 10 intervals and a simulation took 8.4 s. If the sample size was increased to 400, the logistic model simulation time increased to 3.1 s and the proportional hazards model simulation time increased to 27 s.

10. Discussion

The algorithm finds the power for the Wald test that a specified linear combination of the parameters equal zero. There are two other tests which are commonly used for the same purpose: the likelihood ratio test and the Rao efficient score test.

The likelihood ratio test is invariant to reparameterization and has better small sample properties. The test statistic can be calculated using popular software but the estimation program must be run twice. Confidence intervals for a single parameter require calculating the profile

likelihood for the parameter, which is computationally demanding, although this calculation is implemented in SAS.

The Rao efficient score test [10] is based on the derivatives of the log-likelihood under the null hypothesis. It has the advantage that it only requires parameter estimation under the null hypothesis and thus is useful for stepwise variable selection algorithms.

Rao shows that all three test statistics converge to each other in probability under the null hypothesis and, therefore, will have the same distribution under contiguous alternative distributions [11]. Thus the algorithm presented here should give reasonable approximations to the power for any of these tests. A direct method of calculating the non-centrality parameter of any likelihood ratio test is given in Strawderman [12] and for those used for the generalized linear models in Shieh [13]. Both methods would involve multivariate integration if there were more than one continuous covariate, although the integration could be made tractable using Theorem 1 if the continuous covariates had a normal distribution.

The current paper focuses on the Wald test because it is the most commonly used, largely because it is easy to calculate, and the statistic for testing the significance of a single covariate is its estimate divided by its standard error. Confidence intervals based on the test are very simple to compute. However, the Wald test has been criticized because it is not invariant to a reparameterization and its distribution may not be normal in some circumstances [14] and should be used with care. One problem with the statistic is discussed in section 7.

Theorem 1 reduces the multivariate integral needed to calculate the information matrix to a univariate integral in $\beta'X$. The same method could be applied for non-normal covariates but the result would only be a univariate integral if the distribution of $\beta'X$ could be expressed in closed form. Thus, the results in this paper could have been presented in a much more general but less useful context.

A computer program based on this algorithm has several useful applications. One application is to allow the use of a single program to calculate sample size in many situations. Many discrete data analysis problems can be solved using a logistic regression, thus the same computer application could be used to calculate the sample size to compare two proportions, to compare more than two proportions, to test for the effect of a continuous covariate and to determine the effect of a covariate when proportions are compared.

Software using the algorithm was written using Matlab. A compiled version of the software for Windows 95/NT is available from the authors for individual use (www.power-analysis.com/power.htm). The algorithm has also been incorporated into a commercial program 'Power and Precision' [15].

The competitor to this software is simulation which can always be used to accurately calculate power. The problem with simulation is that, even with modern computers, it is too slow for many study design purposes. Simulation calculates the power from the sample size. In order to reverse the process four or five simulations must be performed for each calculation of the sample size from the power, while the algorithm can find sample size directly because the information matrix is proportional to the sample size. For larger samples, it would take 3 or 4 minutes, to calculate a sample size for the proportional hazards model. This is feasible for single calculations but it would take too much time to generate the extensive graphs and user aids provided in 'Power and Precision.' It is clear, however, that in a few more generations of computers, there will be no need to develop efficient algorithms such as the one described here.

Acknowledgements

This work was funded by NIH under grants CA 74302, SBIR-MH 52969 and SBIR-MH 60033. The algorithm was developed for inclusion in the commercial software application 'Power and Precision' developed by Biostatistical Programming Associates Inc. David Schoenfeld is a paid

consultant to Biostatistical Programming Associates Inc. and Michael Borenstein is the owner and president of the company.

References

- [1] Whittemore, A.S., 1981, Sample size for logistic regression with small response probability. *Journal of the American Statistical Association*, **76**, 27–32.
- [2] Hsieh, F.Y., 1989, Sample size tables for logistic regression. *Statistics in Medicine*, **8**, 795–802.
- [3] Hsieh, F.Y., Bloch, D.A. and Larsen, M.D., 1998, A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, **17**, 1623–1634.
- [4] Schoenfeld, D.A., 1983, Sample-size formula for the proportional-hazards regression model. *Biometrics*, **39**, 499–503.
- [5] Schmoor, C., Sauerbrei, W. and Schumacher, M., 2000, Sample size considerations for the evaluation of prognostic factors in survival analysis. *Statistics in Medicine*, **19**(4), 441–452.
- [6] Bernardo, P.M.V., Lipsitz, S.R., Harrington, D.P. and Catalano, P.J., 2000, Sample size calculations for failure time random variables in non-randomized studies. *The Statistician*, **49**(1), 31–40.
- [7] Samuels, M.L., 1993, Simpson's paradox and related phenomena. *Journal of the American Statistical Association*, **88**, 81–88.
- [8] Lagakos, S.W. and Schoenfeld, D.A., 1984, Properties of proportional-hazards score tests under misspecified regression models. *Biometrics*, **40**, 1037–1048.
- [9] Gail, M.H., Wieand, S. and Piantadosi, S., 1984, Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71**, 431–444.
- [10] Rao, R.C., 1973, *Linear Statistical Inference and its Applications* (John Wiley & Sons).
- [11] Simons, G., 1982, Contiguity. In: *Encyclopedia of Statistical Sciences (9 volumes plus Supplement)*, Vol. 2, pp. 159–161.
- [12] Strawderman, W.E., 1983, Likelihood ratio tests. In: *Encyclopedia of Statistical Sciences (9 volumes plus Supplement)*, Vol. 4, pp. 647–650.
- [13] Shieh, G., 2000, On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, **56**(4), 1192–1196.
- [14] Pawitan, Y., 2000, A reminder of the fallibility of the Wald statistic: likelihood explanation. *The American Statistician*, **54**(1), 54–56.
- [15] Borenstein, M., 2001, Power and precision. Available online at: www.power-analysis.com.
- [16] Mathworks, Matlab. www.mathworks.com, 3 Apple Hill Drive, Natick, MA 01760, USA.

Appendix A

A.1 Proof of Theorem

Recall the first part of the covariate vector which is discrete and has K distinct values each with probabilities w_1, w_2, \dots, w_K and the distribution of $\mathbf{X}\mathbf{X}'$ is different for each of these values. Let E_1, E_2, \dots be the expectations with respect to these distributions.

Therefore,

$$E\{\mathbf{X}\mathbf{X}'f(\beta'\mathbf{x})\} = \sum_{s=1}^m w_s E_s\{\mathbf{X}\mathbf{X}'f(\beta'\mathbf{X})\} \quad (\text{A1})$$

The following method is used to compute $E_s\{\mathbf{X}'\mathbf{X}f(\beta'\mathbf{X})\}$. The subscript s is suppressed in what follows.

First note that

$$\begin{aligned} E\{\mathbf{X}\mathbf{X}'f(\beta'\mathbf{X})\} &= E\{f(\beta'\mathbf{X})E(\mathbf{X}\mathbf{X}'|\beta'\mathbf{X})\} \\ &= E[f(\beta'\mathbf{X})\{E(\mathbf{X}|\beta'\mathbf{X})'E(\mathbf{X}|\beta'\mathbf{X}) + V(\mathbf{X}|\beta'\mathbf{X})\}] \end{aligned} \quad (\text{A2})$$

The matrix Σ^a is the variance covariance matrix of \mathbf{X} where the first d components, being constant, have zero variance. Then using standard formulas for conditional normal

distributions,

$$E(\mathbf{X}'|\beta'\mathbf{X}) = \mathbf{u} + \Sigma^a \beta (\beta' \Sigma^a \beta)^{-1} (\beta' \mathbf{X} - \beta' \mathbf{u}) \quad (\text{A3})$$

and

$$V(\mathbf{X}'|\beta'\mathbf{X}) = \Sigma^a - (\Sigma^a \beta) (\beta' \Sigma^a \beta)^{-1} (\Sigma^a \beta)'. \quad (\text{A4})$$

Let $\sigma = \sqrt{\beta' \Sigma^a \beta}$, $\tau = \beta' \mathbf{u}$, $z = (\beta' \mathbf{X} - \tau)/\sigma$ and $\gamma = \Sigma^a \beta/\sigma$. Note that z has a standard normal distribution.

Then, from equations (A3) and (A4),

$$E(\mathbf{X}'|\beta'\mathbf{X}) = \mathbf{u} + \gamma z \quad (\text{A5})$$

and

$$V(\mathbf{X}'|\beta'\mathbf{X}) = \Sigma^a - \gamma \gamma'. \quad (\text{A6})$$

Define $e_i(\tau, \sigma) = E(z^i f(\sigma z + \tau))$, for $i = 0, 1, 2$, where z has a standard normal distribution. This value will be found by numerical integration.

Substituting equations (A5) and (A6) into equation (A2) we get that

$$E\{\mathbf{X}\mathbf{X}' f(\beta'\mathbf{X})\} = (\mathbf{u}\mathbf{u}' + \Sigma^a - \gamma \gamma') e_0(\tau, \sigma) + (\mathbf{u}\gamma' + \gamma \mathbf{u}') e_1(\tau, \sigma) + \gamma \gamma' e_2(\tau, \sigma) \quad (\text{A7})$$

The derivations for $E\{\mathbf{X} f(\beta'\mathbf{X})\}$ and $E\{f(\beta'\mathbf{X})\}$ are similar.

Copyright of Journal of Statistical Computation & Simulation is the property of Taylor & Francis Ltd. The copyright in an individual article may be maintained by the author in certain cases. Content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.