WILEY
InterScience®
DISCOVER SOMETHING GREAT

Sample size determination for logistic regression revisited

Eugene Demidenko*,†

Dartmouth Medical School, Hanover, NH 03755, U.S.A.

SUMMARY

There is no consensus on the approach to compute the power and sample size with logistic regression. Some authors use the likelihood ratio test; some use the test on proportions; some suggest various approximations to handle the multivariate case. We advocate the use of the Wald test since the Z-score is routinely used for statistical significance testing of regression coefficients. The null-variance formula became popular from early studies, which contradicts modern software, which utilizes the method of maximum likelihood estimation (MLE), when the variance of the MLE is estimated at the MLE, not at the null. We derive general Wald-based power and sample size formulas for logistic regression and then apply them to binary exposure and confounder to obtain a closed-form expression. These formulas are applied to minimize the total sample size in a case–control study to achieve a given power by optimizing the ratio of controls to cases. Approximately, the optimal number of controls to cases is equal to the square root of the alternative odds ratio. Our sample size and power calculations can be carried out online at www.dartmouth.edu/ ∼eugened. Copyright © 2006 John Wiley & Sons, Ltd.

1. INTRODUCTION

Power analysis and sample size determination are essential in the design of clinical trials and epidemiology. Several specialized statistical packages compute power and sample size for logistic regression under various scenarios: PASS 2000, nQuery, EGRET SIZ.

Three general statistical approaches to power analysis and sample size determination are available: the likelihood ratio test [1], the score test [2, 3] and the Wald test [4]. The classical theory of maximum likelihood estimation says that all these tests have asymptotically the same type I error (significance level or size). Moreover, they are *locally* equivalent, so that the power functions are close when the alternative approaches the null [5, 6]. However, *globally*, the three tests are different, so there should be no surprise if different tests produce different sample sizes. Consequently, the

---

*Correspondence to: Eugene Demidenko, Dartmouth Medical School, Hanover, NH 03755, U.S.A.
†E-mail: eugened@dartmouth.edu

test used to compute the sample size should match the test used in future regression significance testing. Having this in mind, we consistently apply the Wald test to the sample size determination because it is routinely used as the significance test for the logistic regression coefficients. One of the drawbacks of existing sample size formulas is that they are often based on tests different from Wald, and therefore, do not match the routine of statistical significance testing.

It is customary to see the sample size formula

$$n = \frac{(Z_{1-\alpha/2}\sqrt{V_0} + Z_P\sqrt{V})^2}{(\beta - \beta_0)^2} \tag{1}$$

in various statistical texts, such as Shieh [1], Whittemore [7], Smith and Day [8], Wilson and Gordon [9], Bull [10], Hwang *et al.* [11] and many others. In this formula, the null hypothesis is that the unknown parameter takes the value β_0, or symbolically,

$$H_0 : \beta = \beta_0 \tag{2}$$

with the alternative $\beta \neq \beta_0$. Formula (1) gives a required sample size for a two-sided test to have a significance level α and power P, where V_0 is the variance of the test statistic evaluated at the null and V is the variance evaluated at the alternative; $Z_{1-\alpha/2}$ and Z_P are the $(1 - \alpha/2)$ and P quantiles of the standard normal distribution. It is instructive to derive formula (1) following the line of standard argumentation for testing one-sample binomial proportion [12, p. 249].

Let $\widehat{p}$ be the proportion of cases in n independent and identically distributed (i.i.d.) observations. We want to test that the probability of the case is p_0, so the null hypothesis is $H_0 : p = p_0$. Under the null hypothesis, from the Central Limit Theorem

$$\widehat{p} \sim \mathcal{N}\left(p_0, \frac{p_0(1 - p_0)}{n}\right)$$

with the test statistic

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \tag{3}$$

Under the alternative, Z has an asymptotic normal distribution with mean $(p - p_0)/V_0\sqrt{n}$ and variance V/V_0, where $V = p(1 - p)$ and $V_0 = p_0(1 - p_0)$. This immediately leads to formula (1) assuming that the power of the test is P and the size is α.

We challenge formula (1), suggesting that $V_0 = V$, so that the formula becomes

$$n = \frac{(Z_{1-\alpha/2} + Z_P)^2}{(\beta - \beta_0)^2} V \tag{4}$$

Our point is that in practice, and particularly in logistic regression, we do not use test statistic (3) but

$$Z_* = \frac{\widehat{p} - p_0}{\sqrt{\widehat{p}(1 - \widehat{p})/n}} \tag{5}$$

Obviously, statistic Z_* leads to the sample size given by (4). As another argument, we draw a parallel to the coefficient significance testing, $H_0: \beta_i = 0$ in linear model using the *t*-test by computing $\widehat{\beta}_i/\sqrt{s^2(\mathbf{X}'\mathbf{X})_{ii}^{-1}}$. Following (1), to obtain s^2, we should compute the residual sum of squares at $\beta_i = 0$ but we never do this. A rigorous, more general derivation is found in the next section.

Both statistics, Z and Z_*, have the same asymptotic distribution under the null but different power. Hereafter, we refer to formula (1) as to the *null-variance* formula because the variance is evaluated at $\beta = \beta_0$. In the next section, we derive conditions under which the null-variance formula over- or underestimates the sample size, assuming that the test (3) is used as a test statistic (in a more general setting, it corresponds to test (9)).

The organization of the paper is as follows. In Section 2, we derive the power function and the sample size for the Wald test in multivariate logistic regression in general forms. In Section 3, we illustrate the computation with binary exposure and compare our formula with the null-variance formula. We apply our formula to case–control studies to optimally define the proportion of controls and cases. In Section 4, we extend the results to $m = 2$, when both variables are binary.

2. WALD TEST FOR LOGISTIC REGRESSION

The key step of the Wald test is computation of the Fisher information matrix as a function of parameters. First, we rigorously derive the power function and the sample size for logistic regression in general terms. Second, we apply this approach to binary exposure and covariate. Third, we compare our sample size formula with a popular null-variance formula analytically and *via* simulations.

The probability of a binary event (such as disease status), y_i, is modelled *via* a multivariate logistic regression model as a conditional probability

$$\Pr(y = 1 | \mathbf{x}) = \frac{e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}}{1 + e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}} \tag{6}$$

where α_0 is the intercept term (on the logit scale), and $\mathbf{x}$ is the $m \times 1$ vector of covariates, which may include exposure and confounder. We assume that the data $\{\mathbf{x}_i, i = 1, 2, \ldots, n\}$ are i.i.d. with the distribution, $\mathscr{F}$, independent of $\boldsymbol{\beta}$, $\mathbf{x}_i \sim \mathscr{F}$, McCullagh and Nelder [13]. The $(m + 1) \times (m + 1)$ Fisher information matrix for $\boldsymbol{\theta} = (\alpha_0, \boldsymbol{\beta}')$ takes the form

$$\mathbf{I}_{\boldsymbol{\theta}} = \begin{bmatrix} I_{\alpha_0} & \mathbf{I}'_{\alpha_0 \boldsymbol{\beta}} \\ \mathbf{I}_{\alpha_0 \boldsymbol{\beta}} & \mathbf{I}_{\boldsymbol{\beta}} \end{bmatrix} \tag{7}$$

where the entries for the respective blocks are the expectations over the $\mathbf{x}$-distribution, namely,

$$I_{\alpha_0} = E_{\mathbf{x} \sim \mathscr{F}} \frac{e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}}{(1 + e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}})^2}, \quad \mathbf{I}_{\alpha_0 \boldsymbol{\beta}} = E_{\mathbf{x} \sim \mathscr{F}} \frac{e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}}{(1 + e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}})^2} \mathbf{x}, \quad \mathbf{I}_{\boldsymbol{\beta}} = E_{\mathbf{x} \sim \mathscr{F}} \frac{e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}}{(1 + e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}})^2} \mathbf{x} \mathbf{x}' \quad (8)$$

respectively, scalar, $m \times 1$ vector and $m \times m$ matrix. Asymptotically, when the sample size, n, goes to infinity, the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is normally distributed, $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{ML}} - \boldsymbol{\theta}) \simeq \mathcal{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V} = \mathbf{I}_{\boldsymbol{\theta}}^{-1}$ is the asymptotic variance–covariance matrix of the $\sqrt{n}\widehat{\boldsymbol{\theta}}_{\mathrm{ML}}$. Based on this fact, we derive the power function and the required sample size, n, to achieve a predefined power probability, P. Since we are dealing only with the maximum likelihood estimate, the subscript ML will be omitted from the notation. Whittemore [7] assumed that the disease incidence is small enough (α_0 is a large negative number) to allow replacing the right-hand side of (6) with $e^{\alpha_0 + \boldsymbol{\beta}' \mathbf{x}}$. Note that while such an assumption facilitates obtaining a closed-form solution to expectations

(8) for normally distributed $\mathbf{x}_i$, it makes the calculation inadequate for studies with high incidence rates, such as in case–control studies.

Now, we specify the Wald-based sample size determination in general terms. Let the coefficient of interest be β, the jth component of vector $\boldsymbol{\theta}$. From the information matrix, we find that the asymptotic variance of $\sqrt{n}\widehat{\beta}$ is $V = V_{jj}$, the corresponding diagonal element of matrix $\mathbf{V}$. According to the Wald test, the test statistic for the null hypothesis (2) is

$$Z = \frac{\sqrt{n}(\widehat{\beta} - \beta_0)}{\sqrt{\widehat{V}}} \simeq \mathcal{N}(0, 1) \tag{9}$$

where $\widehat{V}$ is the variance evaluated at the MLE, $\widehat{\beta}$. In the current software, such as SAS or S-Plus, the variance, V, is indeed evaluated at the MLE, not at β_0. Let V be the variance of $\sqrt{n}\widehat{\beta}$ under the alternative $H_A: \beta \neq \beta_0$. By Slutsky's theorem, the distributions of Z and $\sqrt{n}(\widehat{\beta} - \beta_0)/\sqrt{V}$ are equivalent for large n, Bickel and Doksum [4, p. 512], and Demidenko [14, p. 644]. If α is the specified significance level and $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)$th quantile of the standard normal distribution, the power function of the Wald test is the probability of $|Z| > Z_{1-\alpha/2}$, computed at the alternative. Mathematically, the power function takes the form

$$\Pr(|Z| > Z_{1-\alpha/2}) = \Pr\left(\left|\sqrt{n}\frac{\widehat{\beta} - \beta}{\sqrt{\widehat{V}}} + \sqrt{n}\frac{\beta - \beta_0}{\sqrt{\widehat{V}}}\right| > Z_{1-\alpha/2}\right)$$

$$\simeq \Phi\left(-Z_{1-\alpha/2} + \sqrt{n}\frac{\beta - \beta_0}{\sqrt{V}}\right) + \Phi\left(-Z_{1-\alpha/2} - \sqrt{n}\frac{\beta - \beta_0}{\sqrt{V}}\right) \tag{10}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution. It is worthwhile to remember that the left and the right sides of (10) are close only in large samples and therefore, the sample size determination is valid when n is large. Fortunately, when the power is close to 1, say greater or equal 80 per cent, the contribution of the second cdf in (10) becomes negligible. In general form, we approximate the Wald power and the required sample size as

$$\text{Power} = \Phi\left(-Z_{1-\alpha/2} + \frac{(\beta - \beta_0)\sqrt{n}}{\sqrt{V}}\right) \tag{11}$$

$$n = \frac{(Z_{1-\alpha/2} + Z_P)^2 V}{(\beta - \beta_0)^2} \tag{12}$$

where Z_P is the Pth quantile (typically, $P = 0.8$ or 0.9). In an important special case, $\beta_0 = 0$, we come to the formula

$$n = \frac{(Z_{1-\alpha/2} + Z_P)^2 V}{\beta^2} \tag{13}$$

where V is computed at the alternative. For a popular choice of $\alpha = 0.05$ and $P = 0.8$, from (13) we assess the sample size as $n = 8V/\beta^2$, Lehr [15]. Note that formulas (11), (12) and (13) work for any distribution of $\mathbf{x}$, discrete or continuous.

3. BINARY EXPOSURE

Expectations (8) and sample size determination do not admit a closed-form expression except in some special cases. One of those cases is when the exposure and covariate are binary as well. In this section, we derive Wald power and sample size formula for $m = 1$ and in the next section for $m = 2$. Binary exposure and covariate are typical scenarios in power analysis for designing epidemiologic studies.

To begin, we apply formulas (11) and (12) to a logistic regression with binary exposure ($m = 1$)

$$\Pr(y = 1|x) = \frac{e^{\alpha_0 + \beta x}}{1 + e^{\alpha_0 + \beta x}}$$

and x_i takes value 1 with probability p_x and 0 with probability $1 - p_x$. For example, y may code the occurrence of lung cancer and x smoking status. In this setting, the power analysis for logistic regression with binary exposure can be reduced to a well-developed analysis of proportions, where exact tests are available. See Sahai and Khurshid [16] and Duchateau $et\ al.$ [17] for a recent review. However, since it is likely that future statistical analysis will employ logistic regression with Wald significance testing, the sample size should be determined based on the information matrix. To simplify the exposition, we shall assume that the null hypothesis is $\beta_0 = 0$. Thus, the null hypothesis is that there is no association between y and x, or the odds ratio (OR), $e^{\beta} = 1$. The sample size required to reject the null with the type I error α and the type II error P is given by

$$n = (Z_{1-\alpha/2} + Z_P)^2 \frac{p_x(1 + A)^2 B + (1 - p_x)(1 + AB)^2}{p_x(1 - p_x)AB \ln^2(B)} \tag{14}$$

where we denote $A = e^{\alpha_0}$ and $B = e^{\beta}$ as the alternative OR. Throughout this paper, we use capital letters to denote exponential functions of the corresponding Greek letters. The derivation of formula (14) is found in the Appendix. This sample size formula is referred to as *corrected*. One obtains a rough estimate of the sample size assuming that the alternative is close to 0 (the definition of the variance operator, $\mathcal{V}$, is found in the Appendix):

$$n = \frac{(Z_{1-\alpha/2} + Z_P)^2}{p_x(1 - p_x)\mathcal{V}_A \beta^2} \tag{15}$$

With a popular choice, $\alpha = 0.05$ and $P = 0.8$, we have $n = 8/(p_x(1 - p_x)\mathcal{V}_A \beta^2)$. Since for a binary random variable the variances, $p_x(1 - p_x)$ and $p_y(1 - p_y)$, are less than or equal to $\frac{1}{4}$, we obtain an easy-to-memorize lower bound for the sample size to achieve 80 per cent power with a 5 per cent significance level

$$n \geqslant \frac{125}{\beta^2} \tag{16}$$

This lower bound for the sample size is attained at the completely balanced design, $p_x = p_y = \frac{1}{2}$. In observational studies with a rare disease ($p_y \simeq 0$), the required sample size is the reciprocal of p_y, which encourages the use of case–control studies.

The appearance of the y-variance, $\mathcal{V}_A$, in the denominator of (15) seems paradoxical because in the linear model, $y = \alpha_0 + \beta x + \varepsilon$ with $\text{var}(y|x) = \sigma^2$, the formula for n is the same except that $\text{var}(y|x)$ is in the numerator. This means that, in the linear regression, a larger variance of the

dependent variable *increases* the sample size, but in logistic regression it *decreases* the sample size. The explanation is that in the linear model, the variance of the dependent variable is independent of regression coefficients, but in logistic regression, it is a function of them.

3.1. Comparison with the null-variance formula

Recall, the corrected formula (14) is derived under the assumption that the variance is evaluated at the MLE. This is typical for all logistic regression software, including popular statistical packages such as SAS, S-Plus, and STATA. In contrast, the null-variance formula (1) uses the variance at the null. To compare the sample sizes produced by those formulas, we compute V_0 from (A3) letting $B = 1$

$$V_0 = \frac{(1 + A)^2}{p_x(1 - p_x)A}$$

Under the same conditions (power, size, alternative OR), the null-variance formula produces a larger n if, and only if, $V_0 - V > 0$. After some algebra, we express the difference as

$$V_0 - V = \frac{(B - 1)(1 - BA^2)}{p_x A B}$$

Hence, the null-variance formula overestimates or underestimates the sample size depending on the sign of the numerator,

$$(B - 1)(1 - BA^2)$$

where $B = e^\beta$ is the alternative OR and $A = e^{\alpha_0}$, so that $A/(1 + A)$ is the prevalence rate among the non-exposed subjects. In the case of the exposure variable, we have $B > 1$, so the null-variance formula leads to oversampling if the squared prevalence rate is less than the reciprocal of the alternative OR, or equivalently, $2\alpha_0 + \beta < 0$ on the logit scale. Otherwise, the null-variance formula leads to undersampling.

   We illustrate the formula comparison in Figure 1, where two situations are shown for binary exposure with probability $p_x = \frac{1}{100}$. At the left panel, $A = e^{\alpha_0} = \frac{1}{4}$ with the prevalence rate $\frac{1}{5}$. In the range of the alternative OR from 2 to 3 we have $B \times (0.25)^2 < 1$ and the null-variance formula overestimates the required sample size by almost 30 per cent. Conversely, in the right panel, $B \times (0.9)^2 > 1$, and therefore, the null-variance formula underestimates the required sample size. To test formula (14), we show simulation results with n computed by the formula, and $B = 2.2$, 2.6 (3000 simulations). The empirical power is satisfactorily close to the nominal 80 per cent.

3.2. Optimizing a case–control study

The corrected sample size formula can be applied to case–control studies, as follows from work by Prentice and Pyke [18]. Then A is a controllable parameter, namely, the proportion of controls to cases in the study of n subjects. Thus, to achieve the fixed power P under specified size α, assuming that probability p_x is fixed, we should find a minimum of (14) over positive A. Elementary calculus gives the optimal proportion of cases to controls

$$A_{\text{opt}} = \frac{1}{\sqrt{B}} \sqrt{\frac{p_x B + (1 - p_x)}{p_x + B(1 - p_x)}} \tag{17}$$
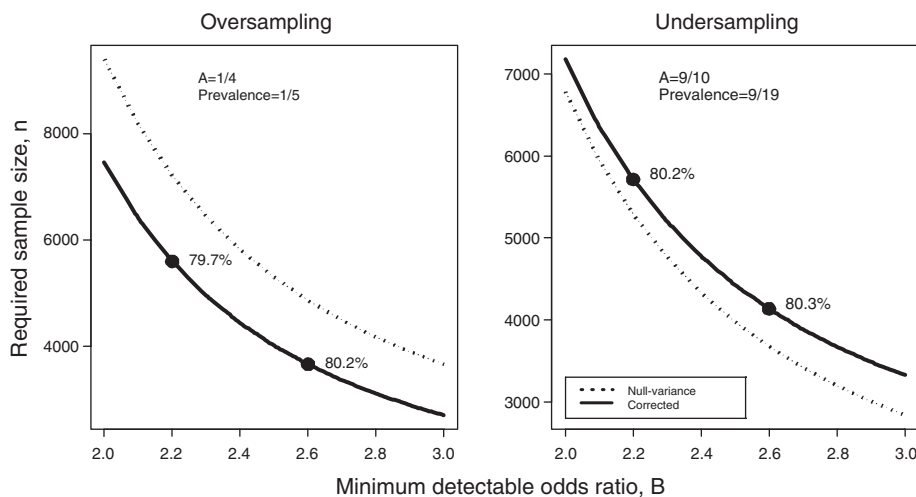
Figure 1. Required sample size for logistic regression with binary exposure with probability $\Pr(x=1)=0.01$. With $A=0.25$, the null-variance formula leads to oversampling and with $A=0.9$ it leads to undersampling. The numbers at $B=2.2$ and 2.6 represent the empirical power obtained from simulations (theoretical power $=80$ per cent), indicated by dots.
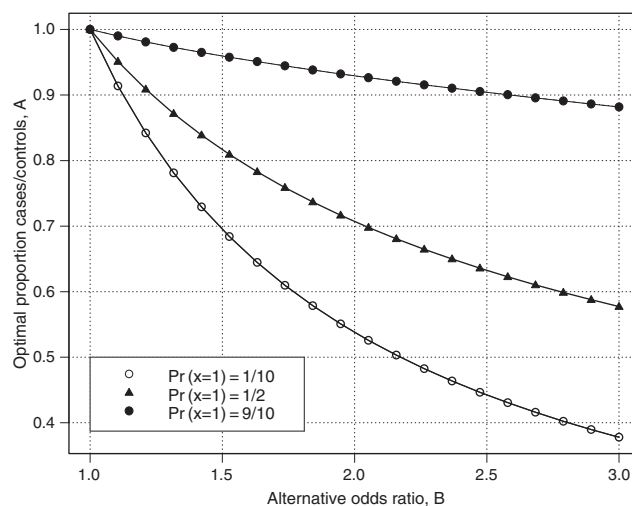

Figure 2. Optimal proportion cases/controls as a function of the alternative odds ratio, B, for different occurrence rates, $p_x = \Pr(x=1)$.

Interestingly, when B is close to 1, the optimal proportion is $\frac{50}{50}=1$ regardless of the prevalence of the exposure. But when the alternative OR departs from 1, the optimal proportion of controls may be less or greater than 1 depending on p_x. We illustrate the computation of the optimal ratio graphically in Figure 2. When exposure is a risk factor ($B>1$), the proportion of cases to controls

is a decreasing function of B. There should be a significantly smaller number of cases when the occurrence rate of exposure is small. As a rule of thumb: if exposure is a risk factor ($B > 1$), there should be more controls than cases. If exposure is a preventive factor ($B < 1$), there should be more cases than controls.

We can find an easy-to-memorize approximation to (17)

$$\frac{n_{\text{control}}}{n_{\text{case}}} = \sqrt{\text{alternative OR}} \tag{18}$$

This formula follows from (17) as reciprocal of A_{opt} after approximation of the numerator and denominator in the neighbourhood of $B = 1$. For example, if, in a case–control study, we want to detect OR $B = 2$, the number of controls to cases should be around $\sqrt{2} = 1.41$. Thus, the number of controls should be 41 per cent higher than the number of cases.

Example
A case–control study is designed to identify the effect of smoking on the occurrence of lung cancer. We want to derive the number of cancer patients and controls (cancer-free) to detect OR $B = 2$ with power 80 per cent and significance level 5 per cent. Assuming that one out of five people smoke, $p_x = \frac{1}{5}$, we determine from formula (17) that the number of cases to patients should be

$$\frac{n_{\text{control}}}{n_{\text{case}}} = \frac{1}{A} = \sqrt{2}\sqrt{\frac{\frac{1}{5} + \frac{8}{5}}{\frac{2}{5} + \frac{4}{5}}} = 1.73 \tag{19}$$

This means that there should be 73 per cent more non-cancer participants than patients with lung cancer (recall that from formula (18) it follows that the number of cancer-free participants should be 41 per cent higher). From formula (14), we compute the total number of participants

$$n = (1.96 + 0.84)^2 \frac{2(1 + A)^2/5 + 4 \times (1 + 2A)^2/5}{8A \ln^2(2)/25} = 416$$

As follows from (19), among 416 participants there should be 152 cases and 264 controls.

4. BINARY EXPOSURE AND CONFOUNDER

Now, the logistic regression takes the form

$$\Pr(y = 1 | x, z) = \frac{e^{\alpha_0 + \beta x + \gamma z}}{1 + e^{\alpha_0 + \beta x + \gamma z}}$$

where exposure x and confounder z are binary. We want to find the power and the sample size, n, to detect the alternative OR $B = e^\beta$ in terms of the marginal probabilities, $\Pr(x = 1) = p_x$ and $\Pr(z = 1) = p_z$. Now, besides previously defined parameters, we need to specify the OR of the confounder $G = e^\gamma$, and the relationship between the exposure and the confounder. This relationship is also defined *via* logistic regression as

$$\Pr(x = 1 | z) = \frac{e^{c + dz}}{1 + e^{c + dz}} \tag{20}$$

Parameter $C = \mathrm{e}^c$ is found from equation

$$1 - p_x = \frac{p_z}{1 + CD} + \frac{1 - p_z}{1 + C}$$

which is reduced to a quadratic equation ($D = \mathrm{e}^d$). As shown in the Appendix, the elements of the 3×3 information matrix are given as follows:

$$I_{11} = L + F + J + H, \quad I_{12} = F + H, \quad I_{13} = J + H$$
$$I_{22} = F + H, \quad I_{23} = H, \quad I_{33} = J + H \tag{21}$$

where

$$L = \frac{A(1 - p_z)}{(1 + A)^2 (1 + C)}, \quad H = \frac{ABGCDp_z}{(1 + ABG)^2 (1 + CD)}$$
$$F = \frac{ABC(1 - p_z)}{(1 + AB)^2 (1 + C)}, \quad J = \frac{AGp_z}{(1 + AG)^2 (1 + CD)} \tag{22}$$

We are interested in the coefficient at the exposure; therefore, taking the $(2, 2)$th element of the inverse information matrix, we arrive at the variance of $\sqrt{n}\widehat{\beta}$ as

$$V = \frac{(1/L + 1/F)(1/H + 1/J)}{1/L + 1/H + 1/F + 1/J} \tag{23}$$

Then, the power function and the sample size are computed by formulas (11) and (12).

It is interesting to note that, unlike linear regression, the addition of the confounder changes the variance of $\widehat{\beta}$, and consequently the power and the sample size, even if the exposure and confounder are independent ($D = 1$). This paradoxical phenomenon is illustrated in Figure 3,
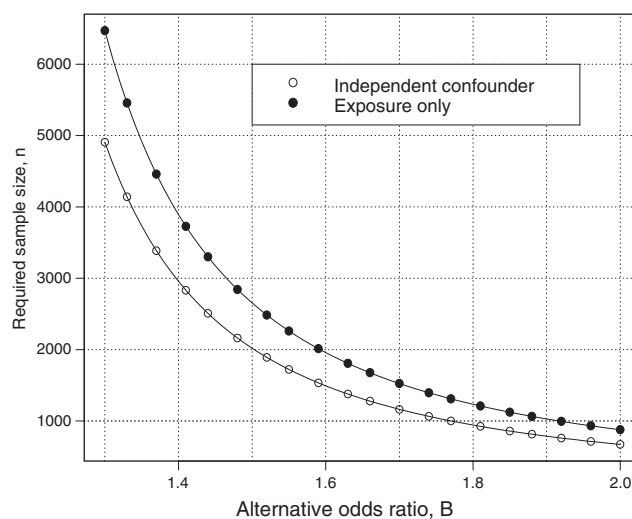

Figure 3. A paradoxical situation: the presence of independent confounder may improve the study.

where the OR between y and confounder is $G = 2$, the prevalence rate of y is $\frac{1}{10}$, $p_x = \frac{1}{4}$ and $p_z = \frac{1}{2}$, $\alpha = 5$ per cent, and $P = 80$ per cent. A similar paradox has been reported by Robinson and Jewell [19].

5. SUMMARY AND DISCUSSION

There is no consensus on what approach to use for the sample size formula in studies with logistic regression. Some authors apply various approximations, such as the variance-inflation technique [20], some advocate the likelihood ratio approach [2]. We argue that the power calculation should use the same test statistic, which is used in coefficient significance testing. Since the Wald test is typically used, the same test should be used as the base for power calculations and the required sample size. We have developed simple Wald-based sample size formulas for binary exposure and confounder. Those are the simplest scenarios when a clinical or epidemiologic study is planned where closed-form formulas still exist. Our sample size and power calculations can be carried out online at www.dartmouth.edu/ ∼eugened.

The null-variance formula became very popular in the early stages of sample size studies, as an extension of the test on proportions. We argue that the routine test statistic for the logistic regression coefficients is the Z-score, $\widehat{\beta}_i/\mathrm{SE}(\widehat{\beta}_i)$, where the standard error

$$\mathrm{SE}(\widehat{\beta}_i) = \sqrt{V_{ii}(\widehat{\boldsymbol{\beta}})}$$

is computed at the MLE value, not at the null hypothesis $\beta_i = 0$. This implies the corrected formulas (11) and (12).

We consistently apply the Wald approach to the sample size determination and illustrate it with logistic regression models for $m = 1$ and 2, when exposure and confounder are binary—perhaps the simplest to specify and the most important scenarios when planning a clinical or epidemiologic study.

Generally, we caution extending the assertions based on the linear regression model, such as 'an addition of a covariate increases the variance of the coefficient estimate at exposure', to logistic regression. Logistic regression is another animal. For example, Robinson and Jewell [19] uncovered surprising results about covariate adjustment in logistic regression (*surprising* from a linear regression standpoint).

Hauck and Donner [21] reported an aberrant behaviour of the Wald test in logistic regression, meaning that the power goes down when the alternative moves away from the null value. Væth [22] extended the analysis of the Wald test to exponential families in the framework of generalized linear model. Does this mean that the likelihood ratio and score tests are preferable? Unfortunately, there is no rigorous answer—additional theoretical and empirical work is needed to compare the tests globally, in the wide range of alternatives. In any event, the same test should be used for the sample size determination and coefficient significance testing.

As a final note, the power analysis and sample size determination based on asymptotic tests, such as Wald or likelihood ratio tests, are implicitly controversial. On one hand, these tests work on large samples but on the other hand, we want to find a finite n that produces the desired power. More work should be done to develop a small-sample version of the Wald test and appropriate adjustments to the sample size formulas discussed in this paper. The work by King and Ryan [23]

is a good example of studying finite sample properties of logistic regression estimates. On the other hand, high power dictates a large number of observations that revive application of the Central Limit Theorem.

APPENDIX A

A.1. Derivation of formula (14)

Denoting the probability (prevalence) of the exposure as $\Pr(x = 1) = p_x$, the information matrix (7) for $\boldsymbol{\theta} = (\alpha_0, \beta)'$ simplifies to a 2×2 matrix

$$\mathbf{I} = \begin{bmatrix} \dfrac{ABp_x}{(1 + AB)^2} + \dfrac{A(1 - p_x)}{(1 + A)^2} & \dfrac{ABp_x}{(1 + AB)^2} \\[2mm] \dfrac{ABp_x}{(1 + AB)^2} & \dfrac{ABp_x}{(1 + AB)^2} \end{bmatrix} \tag{A1}$$

Also, to simply the notation, we introduce the variance operator

$$\mathscr{V}_g = \frac{g}{(1 + g)^2} \tag{A2}$$

For example, the quantities $\mathscr{V}_A$ and $\mathscr{V}_{AB}$ represent the variance of y among the non-exposed ($x = 0$) and exposed ($x = 1$) subjects, respectively. Letting $p_y = A/(1 + A)$, the conditional prevalence rate (the probability of disease among unexposed subjects), we can write $\mathscr{V}_A = p_y(1 - p_y)$.

In the $\mathscr{V}$-notation, it is easy to check that the asymptotic variance of the MLE, as the (2, 2)th element of the inverse information matrix (A1), can be expressed as

$$V = \frac{1}{1 - p_x} \frac{1}{\mathscr{V}_A} + \frac{1}{p_x} \frac{1}{\mathscr{V}_{AB}} \tag{A3}$$

After some algebra, V can be represented *via* a well-known formula of the 2×2 contingency table $V = \sum_{k, j=0}^{1} 1/\Pr(y = k, x = j)$ using the relationship $\Pr(y = k, x = j) = \Pr(y = k | x = j) \Pr(x = j)$; see, Robinson and Jewell [19] and Rosner [12], among others. Combining formula for the variance (A3) with the general sample size formula (12) produces formula (14).

A.2. Information matrix with binary exposure and confounder

Using general formula (8), the 3×3 information matrix is

$$\frac{A}{(1 + A)^2} \Pr(x = 0, z = 0) \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{AB}{(1 + AB)^2} \Pr(x = 1, z = 0) \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$+ \frac{AG}{(1 + AG)^2} \Pr(x = 0, z = 1) \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} + \frac{ABG}{(1 + ABG)^2} \Pr(x = 1, z = 1) \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \frac{A}{(1+A)^2} \frac{1-p_z}{1+C} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{AB}{(1+AB)^2} \frac{C(1-p_z)}{1+C} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$+ \frac{AG}{(1+AG)^2} \frac{p_z}{1+CD} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} + \frac{ABG}{(1+ABG)^2} \frac{CDp_z}{1+CD} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Combining the terms and inverting the information matrix, we obtain formula (23).

ACKNOWLEDGEMENTS

REFERENCES

1. Shieh G. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* 2000; **56**(4):1192–1196.
2. Self SG, Mauritsen RH. Power/sample size calculations for generalized linear models. *Biometrics* 1988; **44**(1): 79–86.
3. Lubin JH, Gail MH. On power and sample size for studying features of the relative odds disease. *American Journal of Epidemiology* 1990; **131**(3):552–566.
4. Bickel PJ, Doksum KA. *Mathematical Statistics* (2nd edn). Prentice-Hall: Upper Saddle River, NJ, 2001.
5. Rao CR. *Linear Statistical Inference and its Applications* (2nd edn). Wiley: New York, 1973.
6. Serfling RJ. *Approximation Theorems of Mathematical Statistics*. Wiley: New York, 1980.
7. Whittemore AS. Sample size for logistic regression with small response probability. *Journal of the American Statistical Association* 1981; **76**(1):27–32.
8. Smith PG, Day NE. The design of case–control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology* 1984; **13**(3):356–365.
9. Wilson SR, Gordon I. Calculating sample size in the presence of confounding variables. *Applied Statistics* 1986; **35**(2):307–213.
10. Bull SB. Sample size and power determination for a binary outcome and an ordinal exposure when logistic regression analysis is planned. *American Journal of Epidemiology* 1993; **137**(6):676–684.
11. Hwang S-J, Beaty TH, Liang K-L, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene–environment interaction in case–control designs. *American Journal of Epidemiology* 1994; **140**(11): 1029–1037.
12. Rosner B. *Fundamentals of Biostatistics* (5th edn). Duxbury: Pacific Grove, CA, 2000.
13. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: Boca Raton, FL, 1989.
14. Demidenko E. *Mixed Models*: *Theory and Applications*. Wiley: New York, 2004.
15. Lehr R. Sixteen s-squared over d-squared: a relation for crude sample size estimates. *Statistics in Medicine* 1992; **11**(8):1099–1102.
16. Sahai H, Khurshid A. Formulae and tables for the determination of sample sizes and power in clinical trials for testing differences in proportions for the two-sample design: a review. *Statistics in Medicine* 1996; **15**(1):1–21.
17. Duchateau L, McDermott B, Rowlands GR. Power evaluation of small drug and vaccine experiments with binary outcomes. *Statistics in Medicine* 1998; **17**(1):111–120.
18. Prentice RL, Pyke R. Logistic disease incidence models and case–control studies. *Biometrika* 1979; **66**(3): 403–411.

19. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 1991; **58**(2):227–240.
20. Hsieh FY, Bloch DA, Larsen MD. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* 1998; **17**(14):1623–1634.
21. Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association* 1977; **72**(360):851–853.
22. Væth M. On the use of Wald's test in exponential families. *International Statistical Review* 1985; **53**(2):199–214.
23. King EN, Ryan TP. A preliminary investigation of maximum likelihood logistic regression versus exact logistic regression. *The American Statistician* 2002; **56**(3):163–170.