STATISTICS IN MEDICINE
Statist. Med. **17**, 1623–1634 (1998)

A SIMPLE METHOD OF SAMPLE SIZE CALCULATION FOR LINEAR AND LOGISTIC REGRESSION

F. Y. HSIEH[1]*, DANIEL A. BLOCH[2] AND MICHAEL D. LARSEN[3]

[1] *CSPCC, Department of Veterans Affairs, Palo Alto Health Care System (151-K), Palo Alto, California 94304, U.S.A.*
[2] *Division of Biostatistics, Department of Health Research and Policy, Stanford University, Stanford, California 94305, U.S.A.*
[3] *Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.*

SUMMARY

A sample size calculation for logistic regression involves complicated formulae. This paper suggests use of sample size formulae for comparing means or for comparing proportions in order to calculate the required sample size for a simple logistic regression model. One can then adjust the required sample size for a multiple logistic regression model by a variance inflation factor. This method requires no assumption of low response probability in the logistic model as in a previous publication. One can similarly calculate the sample size for linear regression models. This paper also compares the accuracy of some existing sample-size software for logistic regression with computer power simulations. An example illustrates the methods. © 1998 John Wiley & Sons, Ltd.

INTRODUCTION

In a multiple logistic regression analysis, one frequently wishes to test the effect of a specific covariate, possibly in the presence of other covariates, on the binary response variable. Owing to the nature of non-linearity, the sample size calculation for logistic regression is complicated. Whittemore[1] proposed a formula, derived from the information matrix, for small response probabilities. Hsieh[2] simplified and extended the formula for general situations by using the upper bound of the formula. Appendix I presents a simple closed form, based on an information matrix, to approximate the sample size for both continuous and binary covariates in a simple logistic regression. In a different approach, Self and Mauritsen[3] used generalized linear models and the score tests to estimate the sample size through an iterative procedure. These published methods are complicated and may not be more accurate than the conventional sample size formulae for comparing two means or a test of equality of proportions. In the next section, we present a simple formula for the approximate sizes of the sample required for simple logistic regression by using formulae for calculating sample size for comparing two means or for

* Correspondence to: F. Y. Hsieh, CSPCC, Department of Veterans Affairs, Palo Alto Health Care System (151-K), Palo Alto, California 94304, U.S.A.

Contract/grant sponsor: Department of Veterans Affairs Cooperative Studies Program
Contract/grant sponsor: NIH
Contract/grant number: AR20610
Contract/grant sponsor: National Institute on Drug Abuse
Contract/grant number: Y01-DA-40032-0

CCC 0277–6715/98/141623–12$17.50

Received February 1997
Revised October 1997

comparing two proportions. We can then adjust the sample size requirement for a multiple logistic regression by a variance inflation factor. This approach applies to multiple linear regression as well.

SIMPLE LOGISTIC REGRESSION

In a simple logistic regression model, we relate a covariate X_1 to the binary response variable Y in a model $\log(P/(1-P)) = \beta_0 + \beta_1 X_1$ where $P = \text{prob}(Y = 1)$. We are interested in testing the null hypothesis $H_0\colon \beta_1 = 0$ against the alternative $H_1\colon \beta_1 = \beta^*$, where $\beta^* \neq 0$, that the covariate is related to the binary response variable. The slope coefficient β_1 is the change in log odds for an increase of one unit[4] in X_1. When the covariate is a continuous variable with a normal distribution, the log odds value β_1 is zero if and only if the group means, assuming equal variances, between the two response categories are the same. Therefore we may use a sample size formula for a two-sample t-test to calculate the required sample size. For simplicity, we use a normal approximation instead, as the sample size formula (see formula (7) in Appendix I) may be easily changed to include t-tests if required:

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})^2/[P1(1-P1)\beta^{*2}] \tag{1}$$

where n is the required total sample size, β^* is the effect size to be tested, $P1$ is the event rate at the mean of X, and Z_u is the upper uth percentile of the standard normal distribution.

When the covariate is a binary variable, say $X = 0$ or 1, the log odds value $\beta_1 = 0$ if and only if the two event rates are equal. The sample size formula for the total sample size required for comparing two independent event rates has the following form (see formula (10)):

$$n = \{Z_{1-\alpha/2}[P(1-P)/B]^{1/2} + Z_{1-\beta}[P1(1-P1) + P2(1-P2)(1-B)/B]^{1/2}\}^2 \\ /[(P1-P2)^2(1-B)] \tag{2}$$

where: $P(=(1-B)P1 + BP2)$ is the overall event rate; B is the proportion of the sample with $X = 1$; $P1$ and $P2$ are the event rates at $X = 0$ and $X = 1$, respectively. For $B = 0{\cdot}5$, the required sample size is bounded by the following simple form (see formula (11)):

$$n < 4P(1-P)(Z_{1-\alpha/2} + Z_{1-\beta})^2/(P1-P2)^2. \tag{3}$$

Appendix I presents two simpler forms, formulae (12) and (13), than formula (2). A later section presents the comparisons of these formulae with computer power simulations.

MULTIPLE LOGISTIC REGRESSION

When there is more than one covariate in the model, a hypothesis of interest is the effect of a specific covariate in the presence of other covariates. In terms of log odds parameters, the null hypothesis for multiple logistic regression is H_0: $[\beta_1, \beta_2, \ldots, \beta_p] = [0, \beta_2, \ldots, \beta_p]$ against the alternative $[\beta^*, \beta_2, \ldots, \beta_p]$. Let b_1 be the maximum likelihood estimate of β_1. Whittemore[1] has shown that, for continuous, normal covariates X, the variance of b_1 in the multivariate setting with p covariates, $\text{var}_p(b_1)$, can be approximated by inflating the variance of b_1 obtained from the one parameter model, $\text{var}_1(b_1)$, by multiplying by $1/(1-\rho^2_{1.23\ldots p})$ where $\rho_{1.23\ldots p}$ is the multiple correlation coefficient relating X_1 with $X_2, \ldots, X_p$. That is, approximately

$$\text{var}_p(b_1) = \text{var}_1(b_1)/(1-\rho^2_{1.23\ldots p})$$

The squared multiple correlation coefficient $\rho^2_{1.23\ldots p}$, also known as R^2, is equal to the proportion of the variance of X_1 explained by the regression relationship with $X_2, \ldots, X_p$. The term $1/(1-\rho^2_{1.23\ldots p})$ will be referred to as a variance inflation factor (VIF). The required sample size for the multivariate case can also be approximated from the univariate case by inflating it with the same factor $1/(1-\rho^2_{1.23\ldots p})$. Following the relationship of the variances, we have $n_p = n_1/(1-\rho^2_{1.23\ldots p})$ where n_p and n_1 are the sample sizes required for a logistic regression model with p and 1 covariates, respectively. The same VIF seems to work well for binary covariates (see Appendix III).

MULTIPLE LINEAR REGRESSION

For multiple linear regression models, we can easily derive the same VIF for p covariates (see Appendix II). Therefore, we can adjust similarly the sample size for a regression model with p covariates. It is known that in a simple linear regression model, the correlation coefficient ρ and the regression parameter β_1 have the relationship $\rho = \beta_1\sigma_X/\sigma_Y$. Hence $\rho = 0$ if and only if $\beta_1 = 0$. When both X and Y are standardized, testing the hypotheses that $\rho = 0$ and that $\beta_1 = 0$ are equivalent and the required sample sizes are the same.

Let r be the estimate of the correlation coefficient between X and Y. The sample size formula (see Sokal and Rohlf[5]) for testing H_0: $\rho = 0$ against the alternative H_1: $\rho = r$ is

$$n_1 = (Z_{1-\alpha/2} + Z_{1-\beta})^2/C(r)^2 + 3$$

where the Fisher's transformation $C(r) = \frac{1}{2}\log((1+r)/(1-r))$. If we add $p-1$ covariates to the regression model, the required sample size for testing H_0: $[\beta_1, \beta_2, \ldots, \beta_p] = [0, \beta_2, \ldots, \beta_p]$ against the alternative $[\beta^*, \beta_2, \ldots, \beta_p]$ is $n_p = n_1/(1-\rho^2_{1.23\ldots p})$, approximately. If we already have q covariates in the model and would like to expand the model to $p(>q)$ covariates, then, from Appendix II, $n_p = n_q\,((\mathrm{var}_p(b_1)/\mathrm{var}_q(b_1)) = n_q/(1-\rho^2_{(1\,q+1\ldots p)\cdot(23\ldots q)})$ where the partial correlation coefficient $\rho_{(1\,q+1\ldots p)\cdot(23\ldots q)}$ measures the linear association between covariates X_1 and $X_{q+1}, \ldots, X_p$ when the values of covariates $X_2, \ldots, X_q$ are held fixed.

COMPARISON OF SAMPLE-SIZE SOFTWARE

There are at least two computer programs available that use formula (4) (see Appendix I): nQuery from Dr. Janet Elashoff,[6] and SSIZE[7] from the first author. One program, EGRET SIZ from SERC,[8] uses the approach of Self and Mauritsen.[3] For logistic regression, the computer programs nQuery and SSIZE provide sample sizes only for continuous covariates while EGRET SIZ only provides estimates for discrete covariates. Both nQuery and EGRET SIZ are commercial software. Note that the sample size calculation for logistic regression is only one of the many features provided by the above three computer programs.

Table I presents sample size examples for a binary covariate using formula (4) and software EGRET SIZ as well as the corresponding sample size for comparing two proportions (without continuity correction from formulae (2), (3), (12) and (13)), and the results of power simulations. In the table, $P1$ and $P2$ are event rates at $X = 0$ and $X = 1$, respectively; B is the proportion of the sample with $X = 1$; OR is the odds ratio of $X = 1$ versus $X = 0$ such that $\mathrm{OR} = P2(1-P1)/(P1(1-P2))$; $P = (1-B)P1 + BP2$ is the overall event rate or case fraction. Table I is designed to show the relationship of sample sizes for different study designs. It is known that a balanced design ($B = 0{\cdot}5$) requires less sample size than an unbalanced design

Table I. Results of sample size calculations for a binary covariate from six different methods, power = 95 per cent, two-sided significance level 5 per cent

Design		Sample size	Power simulation
Balanced design with high event rates			
(4):	$P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	1367	96·0 ± 0·63%
(2):	$P = 0{\cdot}45$, $P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	1282	95·4 ± 0·66%
(3):	$P = 0{\cdot}45$, $P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	1287	94·7 ± 0·71%
(12)	$P = 0{\cdot}45$, $P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	1287	94·7 ± 0·71%
(13)	$P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	1274	94·6 ± 0·71%
SIZ:	OR = 1·5, case fraction $P = 0{\cdot}45$, sampling fraction 50/50	1285	95·9 ± 0·63%
Balanced design with low odds ratio			
(4):	$P1 = 0{\cdot}5$, $P2 = 0{\cdot}2$, $B = 0{\cdot}5$	141	96·3 ± 0·60%
(2):	$P = 0{\cdot}35$, $P1 = 0{\cdot}5$, $P2 = 0{\cdot}2$, $B = 0{\cdot}5$	126	95·0 ± 0·69%
(3):	$P = 0{\cdot}35$, $P1 = 0{\cdot}5$, $P2 = 0{\cdot}2$, $B = 0{\cdot}5$	131	96·6 ± 0·57%
(12):	$P = 0{\cdot}35$, $P1 = 0{\cdot}5$, $P2 = 0{\cdot}2$, $B = 0{\cdot}5$	131	96·6 ± 0·57%
(13):	$P1 = 0{\cdot}5$, $P2 = 0{\cdot}2$, $B = 0{\cdot}5$	119	94·9 ± 0·70%
SIZ:	OR = 0·25, case fraction $P = 0{\cdot}35$, sampling fraction 50/50	129	96·1 ± 0·61%
Balanced design with high odds ratio			
(4):	$P1 = 0{\cdot}2$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	166	99·0 ± 0·31%
(2):	$P = 0{\cdot}35$, $P1 = 0{\cdot}2$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	126	95·0 ± 0·69%
(3):	$P = 0{\cdot}35$, $P1 = 0{\cdot}2$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	131	96·6 ± 0·57%
(12):	$P = 0{\cdot}35$, $P1 = 0{\cdot}2$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	131	96·6 ± 0·57%
(13):	$P1 = 0{\cdot}2$, $P2 = 0{\cdot}5$, $B = 0{\cdot}5$	119	92·9 ± 0·81%
SIZ:	OR = 4·0, case fraction $P = 0{\cdot}35$, sampling fraction 50/50	129	95·4 ± 0·66%
Balanced design with high odds ratio			
(4):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}5$	1818	98·2 ± 0·42%
(2):	$P = 0{\cdot}075$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}5$	1437	94·4 ± 0·73%
(3):	$P = 0{\cdot}075$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}5$	1443	95·8 ± 0·63%
(12):	$P = 0{\cdot}075$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}5$	1443	95·8 ± 0·63%
(13):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}5$	1430	94·4 ± 0·73%
SIZ:	OR = 2·111, case fraction $P = 0{\cdot}075$, sampling fraction 50/50	1417	94·5 ± 0·72%
Low prevalence rate			
(4):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}2$	2612	97·4 ± 0·50%
(2):	$P = 0{\cdot}06$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}2$	2186	94·9 ± 0·70%
(12):	$P = 0{\cdot}06$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}2$	1833	91·2 ± 0·90%
(13):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}2$	2648	97·4 ± 0·50%
SIZ:	OR = 2·111, case fraction $P = 0{\cdot}06$, sampling fraction 80/20	2070	94·6 ± 0·71%
High prevalence rate			
(4):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}8$	3060	98·3 ± 0·41%
(2):	$P = 0{\cdot}09$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}8$	2257	95·0 ± 0·69%
(12):	$P = 0{\cdot}09$, $P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}8$	2661	97·8 ± 0·46%
(13):	$P1 = 0{\cdot}05$, $P2 = 0{\cdot}1$, $B = 0{\cdot}8$	1820	89·5 ± 0·97%
SIZ:	OR = 2·111, case fraction $P = 0{\cdot}09$, sampling fraction 20/80	2347	97·2 ± 0·52%

($B = 0{\cdot}2$ or $0{\cdot}8$); a low prevalence rate ($B = 0{\cdot}2$) requires less sample size than a high prevalence rate ($B = 0{\cdot}8$); sample size remains the same if the odds ratio is reversed. In addition to the significance level and the power of the test, the values of the following parameters, listed after the sample size methods, are specified in the table:

Formula (4): $P1$, $P2$ and B.
Formula (2): tests of proportions: P, $P1$, $P2$ and B.
Formula (3): simple form for a balanced design: P, $P1$ and $P2$.
Formula (12): simple form for an unbalanced design: P, $P1$, $P2$ and B.
Formula (13): simple form for an unbalanced design: $P1$, $P2$ and B.
SIZ: OR, sampling fractions $1 - B$ and B, and overall case fraction P.

The power simulations, obtained from SIZ with 1000 replications, use the likelihood ratio test for the logistic regression model. The simulations show that the sample sizes obtained from testing two proportions (formulae (2) and (3)) have statistical power within one standard deviation of the expected power of 95 per cent. Also, formulae (2) and (3) are more stable than the other four methods. Note that formula (4) calculates the required total number of events based on the event rate corresponding to $X = 0$, then inflates the number of events to obtain the total sample size. Therefore, formula (4) produces a larger sample size if the lower event rate is assigned to $P1$ instead of $P2$. Formula (4) tends to overestimate the required sample sizes especially when the event rates are low (see Table I). Formula (3) is a special case of formula (12) for a balanced design. As shown in Table I, formula (3) gives the same sample sizes as formula (12) when $B = 0{\cdot}5$, but slightly larger sample size than formula (2). Since formula (3) is designed for $B = 0{\cdot}5$, no sample sizes for formula (3) are given for low or high prevalence rate. Formulae (12) and (13) are simpler than formula (2), but lack accuracy when the sample size ratio is not close to 1 (say > 2 or $< 0{\cdot}5$), and should not be used when the accuracy of sample size calculation is important. It is known that a design with low prevalence rate requires less sample size than high prevalence rate. In Table I, formula (13) does not show this relationship which indicates that the formula overestimates the sample size for low prevalence rate and underestimates high prevalence rate.

Table II presents the results for a continuous covariate from sample size programs nQuery and SSIZE. The corresponding sample sizes from a two-sample t-test (formula (6) with Z-values replaced by t-values) and from formula (1) are also listed for comparison. The table specifies the following parameters indicated after the sample size methods:

Formula (1): $P1$, effect size $= \log(\text{OR}) = \beta^*$.
Two-sample t-test: effect size $= \log(\text{OR})$,
sample size ratio $= \text{prob}(Y = 1)/\text{prob}(Y = 0) = (1 - P1)/P1$.
nQuery: $P1$(event rate at the mean of X),
$P2$(event rate at one standard deviation above the mean of X).
SSIZE: $P1$(event rate at the mean of X),
OR (odds ratio at one standard deviation above the mean of X)
$= P2(1 - P1)/(P1(1 - P2))$.

Table II also provides power simulations obtained from 1000 replications generated by assuming a normally distributed variable X. We used the Wald test in the simulation of the logistic regression model. The results show that the sample sizes estimated by using the two-sample t-test formula and formula (1) seem to be more conservative, but still large enough to achieve the

Table II. Results of sample size calculations for a continuous covariate from four different methods, power = 95 per cent, two-sided significance level 5 per cent

Design		Sample size	Power simulation
Balanced design			
(1):	$P1 = 0{\cdot}5$, effect size $\beta^* = 0{\cdot}405$	317	$95{\cdot}0 \pm 0{\cdot}69\%$
t-test:	effect size = 0·405, sample size ratio = 1	320	$95{\cdot}5 \pm 0{\cdot}66\%$
nQuery:	$P1 = 0{\cdot}5$, $P2 = 0{\cdot}6$	342	$96{\cdot}1 \pm 0{\cdot}61\%$
SSIZE:	$P1 = 0{\cdot}5$, OR = 1·5	341	$95{\cdot}3 \pm 0{\cdot}67\%$
Unbalanced design, high event rates			
(1):	$P1 = 0{\cdot}4$, effect size $\beta^* = 0{\cdot}405$	330	$94{\cdot}4 \pm 0{\cdot}73\%$
t-test:	effect size = 0·405, sample size ratio = 1·5	333	$94{\cdot}8 \pm 0{\cdot}70\%$
nQuery:	$P1 = 0{\cdot}4$, $P2 = 0{\cdot}5$	380	$96{\cdot}7 \pm 0{\cdot}56\%$
SSIZE:	$P1 = 0{\cdot}4$, OR = 1·5	379	$96{\cdot}7 \pm 0{\cdot}56\%$
Unbalanced design, low event rates			
(1):	$P1 = 0{\cdot}1$, effect size $\beta^* = 0{\cdot}405$	880	$95{\cdot}5 \pm 0{\cdot}66\%$
t-test:	effect size = 0·405, sample size ratio = 9	890	$96{\cdot}1 \pm 0{\cdot}61\%$
nQuery:	$P1 = 0{\cdot}1$, $P2 = 0{\cdot}143$	951	$96{\cdot}6 \pm 0{\cdot}57\%$
SSIZE:	$P1 = 0{\cdot}1$, OR = 1·5	950	$96{\cdot}6 \pm 0{\cdot}57\%$

desired power. In other words, Table II seems to indicate that the t-test is a good estimate of sample size which preserves power. Since we used to upper bound of the required sample size in the formulae in both nQuery and SSIZE, both programs provide sample sizes slightly higher than those required. When the odds ratio is fixed, a balanced design (that is, response rate $P1 = 0{\cdot}5$) requires less sample size than an unbalanced design (for example, $P1 = 0{\cdot}4$ or 0·1). Note that due to the exponential nature of the correction term (see Appendix I), we do not recommended use of either software for logistic regression when the odds ratio is large (say $\geqslant 3$).

EXAMPLE

We use a Department of Veterans Affairs Cooperative Study entitled 'A Psychophysiological Study of Chronic Post-Traumatic Stress Disorder'[9] to illustrate the preceding sample size calculation for logistic regression with continuous covariates. The study developed and validated a logistic regression model to explore the use of certain psychophysiological measurements for the prognosis of combat-related post-traumatic stress disorder (PTSD). In the study, patients' four psychophysiological measurements – heart rate, blood pressures, EMG and skin conductance – were recorded while patients were exposed to video tapes containing combat and neutral scenes. Among the psychophysiological variables, the difference of the heart rates obtained while viewing the combat and the neutral tapes (DCNHR) is considered a good predictor of the diagnosis of PTSD. The prevalence rate of PTSD among the Vietnam veterans was assumed to be 20 per cent. Therefore, we assumed a four to one sample size ratio for the non-PTSD versus PTSD groups. The effect size of DCNHR is approximately 0·3 which is the difference of the group means divided by the standard deviation. With a two-sided significance level of 0·05 and a power of 95 per cent, the required sample size based on a two-sample t-test is 905. The squared multiple correlation coefficient of DCNHR versus the other three psychophysiological variables was estimated to be 0·1 and thus the VIF is 1·11. After adjusting for the VIF, a sample size of 1005 was needed for fitting a multiple logistic regression model.

CONCLUSION

The proposed simple methods to calculate sample size for linear and logistic regression models have several advantages. The formulae for the simple methods are well known and do not require specialized software. This paper also provides simple forms of the formulae for easy hand calculation. Compared to more accurate, but more complicated formulae, formulae (1) and (3) have high degrees of accuracy. Computer simulations suggest that the proposed sample size methods for comparing means and for comparing proportions are more accurate than SSIZE, nQuery and EGRET SIZ. This paper suggests not to use SSIZE or nQuery when the odds ratio is large (say $\geqslant 3$) and Liu and Liang's formula (13) when the sample size ratio is not close to 1 (say > 2 or $< 0{\cdot}5$). This paper derives the variance inflation factor (VIF) for the linear regression model and also shows, through computer simulations, that the same VIF applies to the logistic regression model with binary covariates. The usage of the VIF to expand the sample size calculation from one covariate to more than one covariate appears very useful and can be extended to other multivariate models. In conclusion, this paper presents more accurate and simple formulae for sample size calculation with extensions to multivariate models of various types.

APPENDIX I

In a simple logistic regression model $\log(P/(1-P)) = \beta_0 + \beta_1 X_1$, where $P = \text{prob}(Y=1)$, the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 = \beta^*$ is of interest. A power of $1-\beta$ and a two-sided significance level α are usually prespecified to calculate the sample size for the hypothesis test. The following sample size formula, used in both SSIZE and nQuery, is a combination of Whittemore[1] formulae (6) and (16):

$$n = (V(0)^{1/2} Z_{1-\alpha/2} + V(\beta^*)^{1/2} Z_{1-\beta})^2 (1 + 2P1\delta)/(P1\beta^{*2}) \tag{4}$$

where the log odds value $\beta^* = \log(P2(1-P1)/(P1(1-P2)))$, and $Z_{1-\beta}$ and $Z_{1-\alpha/2}$ are standard normal variables with a tail probability of β and $\alpha/2$, respectively.

For a continuous covariate, $V(0) = 1$, $V(\beta^*) = \exp(-\beta^{*2}/2)$, $P1$ and $P2$ are the event rates at the mean of X and one SD above the mean, respectively. The value of δ for continuous covariates is from Hsieh[2] formula (3): $\delta = (1 + (1 + \beta^{*2})\exp(5\beta^{*2}/4))(1 + \exp(-\beta^{*2}/4))^{-1}$.

For a binary covariate, the overall event rate $P = (1-B)P1 + BP2$, where $P1$ and $P2$ are the event rates at $X = 0$ and $X = 1$, respectively; B is the proportion of the sample with $X = 1$, $V(0) = 1/(1-B) + 1/B$, and $V(\beta^*) = 1/(1-B) + 1/(B\exp(\beta^*))$. The value of δ for binary covariates is from Whittemore[1] formula (14): $\delta = (V(0)^{1/2} + V(\beta^*)^{1/2}R)/(V(0)^{1/2} + V(\beta^*)^{1/2})$ where R is from Whittemore[1] formula (15): $R = V(\beta^*)B(1-B)\exp(2\beta^*)/(B\exp(\beta^*) + (1-B))^2$. Note that $R = \delta = 1$ when $\beta^* = 0$.

The proposed method is to use a two-sample test instead of a one-sample test for sample size calculation. The popular sample size formula for testing the equality of two independent sample means with equal sample sizes from two normally distributed groups has the familiar form (see Rosner[10]):

$$n = 2(\sigma_1^2 + \sigma_2^2)(Z_{1-\alpha/2} + Z_{1-\beta})^2/\Delta^2 \tag{5}$$

where n is the total sample size and Δ is the difference of the two group means to be detected; σ_1^2 and σ_2^2 are the variances of the two groups. For an unequal-sample-size design with a sample size ratio of k, the required total sample size should be inflated by a factor of $(k+1)^2/(4k)$. Assuming equal variances, the test statistic employs the common variance of the two groups and formula (5) reduces to

$$n = \sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2[(k+1)^2/k]/\Delta^2 \tag{6}$$

In a simple logistic regression model with a continuous covariate, the sample size ratio is $k = (1 - P1)/P1$ where $P1$ is the event rate of the response at $X = 0$. Therefore, $P1$ is also the overall event rate when X is standardized to have mean 0 and variance 1. By replacing the effect size Δ/σ by β^*, formula (6) becomes

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})^2/[P1(1-P1)\beta^{*2}]. \tag{7}$$

As derived by Whittemore,[1] $1 = V(0) \geqslant V(\beta^*)$, and therefore formula (4) can be bounded by

$$n \leqslant (Z_{1-\alpha/2} + Z_{1-\beta})^2(1 + 2P1\delta)/(P1\beta^{*2}). \tag{8}$$

Formula (7) is more general than the formula derived by Whittemore,[1] who assumed that $P1$ is small and therefore $1/(1 - P1)$ is negligible. Note that Hsieh[2] formula (3) implies that one should not use formula (4) when the odds ratio is large (say $\geqslant 3$).

When the covariate is a binary variable, say $X = 0$ or 1, the log odds values $\beta_1 = 0$ if and only if the two event rates are equal. We can calculate the total sample size from the formula for comparing the two independent event rates (see Rosner[10]):

$$n = (1+k)\{Z_{1-\alpha/2}[P(1-P)(k+1)/k]^{1/2} + Z_{1-\beta}[P1(1-P1) + P2(1-P2)/k]^{1/2}\}^2/(P1-P2)^2 \tag{9}$$

where: $k = B/(1-B)$ is the sample size ratio; B is the proportion of the sample with $X = 1$; $P = (1-B)P1 + BP2$ is the overall event rate; $P1$ and $P2$ are the event rates at $X = 0$ and $X = 1$, under the alternative hypothesis, respectively. By replacing k by $B/(1-B)$, formula (9) becomes

$$n = \{Z_{1-\alpha/2}[P(1-P)/B]^{1/2} + Z_{1-\beta}[P1(1-P1) + P2(1-P2)(1-B)/B]^{1/2}\}^2/[(P1-P2)^2(1-B)]. \tag{10}$$

For a balanced design, $k = 1$ or $B = 0{\cdot}5$, formula (10) is bounded by

$$n < 4P(1-P)(Z_{1-\alpha/2} + Z_{1-\beta})^2/(P1-P2)^2. \tag{11}$$

For an unbalanced design, similar to (6), we inflate formula (11) by a factor of $1/[4B(1-B)]$ to obtain a simple approximation:

$$n = P(1-P)(Z_{1-\alpha/2} + Z_{1-\beta})^2/[B(1-B)(P1-P2)^2]. \tag{12}$$

In a recent publication, Liu and Liang[11] extended Self and Mauritsen's method for correlated observations. As a special case, they provided a closed form for a logistic regression model with one binary covariate. Their closed form, without the adjustment of the design effect for correlated observations, is very similar to (12):

$$n = (Z_{1-\alpha/2} + Z_{1-\beta})^2[BP1(1-P1) + (1-B)P2(1-P2)]/[B(1-B)(P1-P2)^2]. \qquad (13)$$

Examples and comparisons of these formulae are provided in Table I.

APPENDIX II

Let $\mathrm{var}_p(b_1)$ and $\mathrm{var}_1(b_1)$ equal the variances of the parameter estimate obtained from multiple linear regression models with p and 1 covariates, respectively. We show that, most often, the ratio $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$ is bounded by $1/(1-\rho^2_{1.23\ldots p})$. In addition, $\mathrm{var}_p(b_1)/\mathrm{var}_q(b_1)$ is bounded by $1/(1-\rho^2_{(1\,q+1\ldots p)\cdot(23\ldots q)})$ where the partial correlation coefficient $\rho_{(1\,q+1\ldots p)\cdot(23\ldots q)}$ measures the linear association between covariates X_1 and $X_{q+1},\ldots,X_p$ when the values of covariates $X_2,\ldots,X_q$ are held fixed.

We begin with one covariate in a linear regression model $Y = \beta_0 + \beta_1 X_1 + e$ where the error term e is distributed as Normal $(0,\sigma_1^2)$ and, for simplicity, the sample mean of X_1 is 0. The variance of the least squares estimate b_1 is known to equal

$$\mathrm{var}_1(b_1) = \sigma_1^2/\Sigma X_1^2.$$

When there are two covariates X_1 and X_2 with sample means 0, the variance-covariance matrix of the estimates of the parameters is

$$\mathrm{var}_2(b_1,b_2) = \sigma_2^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma_2^2\begin{bmatrix}\Sigma X_1^2 & \Sigma X_1X_2\\ \Sigma X_1X_2 & \Sigma X_2^2\end{bmatrix}^{-1}$$

where $\mathbf{X}$ is the matrix of covariates. Through the inverse of the 2×2 $\mathbf{X}'\mathbf{X}$ matrix, we can obtain the variance of b_1 as

$$\begin{aligned}\mathrm{var}_2(b_1) &= \sigma_2^2\Sigma X_2^2/(\Sigma X_1^2\Sigma X_2^2 - (\Sigma X_1X_2)^2)\\ &= (\sigma_2^2/\sigma_1^2)\,\mathrm{var}_1(b_1)/(1-\rho_{12}^2).\end{aligned}$$

The value of σ_2^2/σ_1^2, in most cases, is less than 1 and close to 1. Since the additional covariate in the model also takes away a degree of freedom from the error term, the estimate of the variance ratio σ_2^2/σ_1^2 may sometimes slightly exceed 1. The squared multiple correlation coefficient, in this case the same as the simple correlation coefficient, is $\rho_{12}^2 = (\Sigma X_1X_2)^2/(\Sigma X_1^2\Sigma X_2^2)$.

When there are three covariates, the multiple correlation coefficient $\rho_{1.23}$ can be obtained from the matrix operation

$$\begin{aligned}\rho^2_{1.23} &= [\Sigma X_1X_2, \Sigma X_1X_3]\begin{bmatrix}\Sigma X_2^2 & \Sigma X_2X_3\\ \Sigma X_2X_3 & \Sigma X_3^2\end{bmatrix}^{-1}\begin{bmatrix}\Sigma X_1X_2\\ \Sigma X_1X_3\end{bmatrix}\Bigg/\Sigma X_1^2.\\ &= (2\Sigma X_1X_2\Sigma X_2X_3\Sigma X_1X_3 - \Sigma X_2^2(\Sigma X_1X_3)^2 - \Sigma X_3^2(\Sigma X_1X_2)^2]/\{\Sigma X_1^2[\Sigma X_2^2\Sigma X_3^2\\ &\quad - (\Sigma X_2X_3)^2]\}.\end{aligned}$$

With three covariates in the regression model, the variance-covariance matrix of the estimates of the parameters can be obtained from the inverse of the 3×3 $\mathbf{X}'\mathbf{X}$ matrix through the formula

$\mathrm{var}_3(b_1, b_2, b_3) = \sigma_3^2(\mathbf{X}'\mathbf{X})^{-1}$. Therefore

$$\begin{aligned}\mathrm{var}_3(b_1) &= \sigma_3^2[\Sigma X_2^2\,\Sigma X_3^2 - (\Sigma X_2X_3)^2]/[\Sigma X_1^2\Sigma X_2^2\Sigma X_3^2 + 2\Sigma X_1X_2\Sigma X_2X_3\Sigma X_1X_3\\ &\quad - \Sigma X_1^2(\Sigma X_2X_3)^2 - \Sigma X_2^2(\Sigma X_1X_3)^2 - \Sigma X_3^2(\Sigma X_1X_2)^2]\\ &= (\sigma_3^2/\sigma_1^2)\mathrm{var}_1(b_1)/(1-\rho_{1.23}^2).\end{aligned}$$

Usually, $\sigma_3^2/\sigma_1^2 \leqslant 1$ and $\mathrm{var}_3(b_1) \leqslant \mathrm{var}_1(b_1)/(1-\rho_{1.23}^2)$. In a linear regression model with p parameters, $\mathrm{var}_p(b_1, b_2, \dots, b_p) = \sigma_p^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma_p^2\Sigma$. By applying a result of Anderson[12] (equation 20) that $\Sigma_{11}^{-1} = \Sigma X_1^2(1-\rho_{1.23\dots p}^2)$, we obtain

$$\begin{aligned}\mathrm{var}_p(b_1) = \sigma_p^2\Sigma_{11} &= \sigma_p^2/\Sigma X_1^2(1-\rho_{1.23\dots p}^2)\\ &= (\sigma_p^2/\sigma_1^2)\mathrm{var}_1(b_1)/(1-\rho_{1.23\dots p}^2).\end{aligned}$$

Again, in most situations, $\sigma_p^2/\sigma_1^2 \leqslant 1$ and $\mathrm{var}_p(b_1) \leqslant \mathrm{var}_1(b_1)/(1-\rho_{1.23\dots p}^2)$. Then, the VIF $= 1/(1-\rho_{1.23\dots p}^2)$ is the approximate upper bound of the ratio $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$. The upper bound does not hold in the rare situation when $\sigma_p^2/\sigma_1^2 > 1$ but the approximation is still good enough. When p is not too large, the bound is tight; when p is large and $\rho_{1.23\dots p}$ is near 1, the bound is inaccurate. A similar result holds for nested models. We would like to expand the model from the situation of q covariates to p covariates where $p > q$. Then, reasoning as above

$$\begin{aligned}\mathrm{var}_p(b_1)/\mathrm{var}_q(b_1) &= (\sigma_p^2/\sigma_q^2)(1-\rho_{1.23\dots p}^2)/(1-\rho_{1.23\dots p}^2) \leqslant (1-\rho_{1.23\dots q}^2)/(1-\rho_{1.23\dots p}^2)\\ &= 1/(1-\rho_{(1\,q+1\dots p)\cdot(23\dots q)}^2).\end{aligned}$$

where the partial corelation coefficient $\rho_{(1\,q+1\dots p)\cdot(23\dots q)}$ measures the linear association between covariates X_1 and $X_{q+1}, \dots, X_p$ when the values of covariates $X_2, \dots, X_q$ are held fixed. The value of the ratio σ_p^2/σ_q^2 should be closer to 1 than σ_p^2/σ_1^2.

APPENDIX III

We use simulations to investigate, in a multiple logistic regression model with p independent binary covariates, how well the ratio of the maximum likelihood estimates of the variances $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$ is approximated by $1/(1-\rho_{1.23\dots p}^2)$, where the multiple correlation coefficient relating binary covariates X_1 with $X_2, \dots, X_p$ has the same formula as continuous covariates with a normal distribution:

$$\rho_{1.23\dots p}^2 = [\Sigma X_1X_2, \Sigma X_1X_3 \dots, \Sigma X_1X_p]\begin{bmatrix}\Sigma X_2^2 & \Sigma X_2X_3 & \cdots & \Sigma X_2X_p\\ \Sigma X_2X_3 & \Sigma X_3^2 & \cdots & \Sigma X_3X_p\\ \dots & \dots & \dots & \dots\\ \Sigma X_2X_p & \Sigma X_3X_p & \cdots & \Sigma X_p^2\end{bmatrix}^{-1}\begin{bmatrix}\Sigma X_1X_2\\ \Sigma X_1X_3\\ \dots\\ \Sigma X_1X_p\end{bmatrix}\Bigg/\Sigma X_1^2 .$$

The 80 computer simulations each use a sample size of 1000 with eight binary covariates. When all eight covariates are generated independently, the estimate values of $\rho_{1.23\dots 7}^2$ are near zero. In order that the response variable Y and the covariates X's be somewhat correlated, and the estimates of $\rho_{1.23\dots 7}^2$ have a broad range of values, say from 0 to 0·7, the generation of the eight covariates requires some special care.

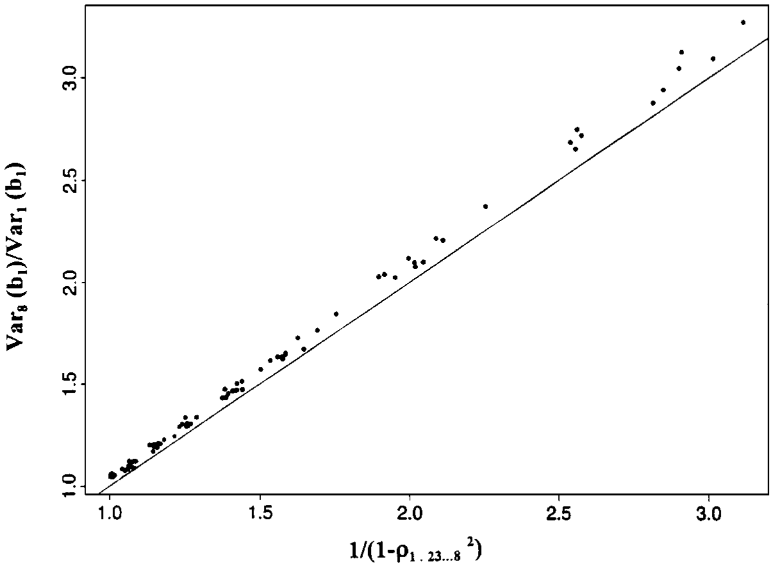

Figure 1. Results of 80 simulations: estimates of $\mathrm{var}_8(b_1)/\mathrm{var}_1(b_1)$ versus $1/(1-\rho^2_{1.23\ldots 8})$

Let U, V_1, V_2, ..., and V_8 be uniform random variates obtained from a generator in SAS.[13] The response variable Y is Bernoulli with a parameter value 0·5. The eight covariates X_1, X_2, ..., and X_8 are also Bernoulli with parameters B_1, B_2, ..., and B_8 which have values 0·5, 0·6, 0·65, 0·7, 0·75, 0·8, 0·85 and 0·9, respectively. The response variable Y is generated such that $Y=1$ when $U>0{\cdot}5$ and $Y=0$, otherwise. In the first simulation, the covariates are generated such that $X_i=1$ when $0{\cdot}1\,U+0{\cdot}9\,V_i>B_i$ and $X_i=0$, otherwise, for $i=1,2,\ldots,8$. The same process was repeated for the second simulation except for the generation of X_2 where the same random value for X_1 was used: $X_2=1$ when $0{\cdot}1U+0{\cdot}9V_1>B_2$ and $X_2=0$, otherwise. In the third simulation, the same random value for X_1 was used for X_3. The similar process continued until the completion of the eighth simulation. After finishing the first eight simulations, the whole process was then repeated ten times to obtain a total of 80 simulations.

In practice, the estimated values of $\mathrm{var}_p(b_1)$ and $\rho^2_{1.23\ldots p}$ (same as R^2) can be obtained from SAS PROC LOGISTIC and PROC REG,[13] respectively. The estimates of $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$ versus $1/(1-\rho^2_{1.23\ldots p})$ from the simulations are plotted in Figure 1. The simulation results show that, for binary covariates, the estimates of $1/(1-\rho^2_{1.23\ldots p})$ closely approximate the value of the estimates of the ratio $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$. Figure 1 shows that the estimates of $1/(1-\rho^2_{1.23\ldots p})$ very slightly underestimate the variance ratio $\mathrm{var}_p(b_1)/\mathrm{var}_1(b_1)$.

ACKNOWLEDGEMENTS

The authors thank Drs. Philip Lavori, Kelvin Lee, a referee and the editor for valuable comments and editorial suggestions which strengthened the content. This work was supported in part by the DVA Cooperative Studies Program of the Veteran Health Administration, NIH grant AR20610 (Multipurpose Arthritis Center) and Y01-DA-40032-0 (National Institute on Drug Abuse), the latter to the VA Cooperative Studies Program.

REFERENCES

1. Whittemore, A. 'Sample size for logistic regression with small response probability', *Journal of the American Statistical Association*, **76**, 27–32 (1981).
2. Hsieh, F. Y. 'Sample size tables for logistic regression', *Statistics in Medicine*, **8**, 795–802 (1989).
3. Self, S. G. and Mauritsen, R. H. 'Power/sample size calculations for generalized linear models', *Biometrics*, **44**, 1, 79–86 (1988)
4. Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*, Wiley, New York, 1989, p. 56.
5. Sokal, R. R. and Rohlf, F. J. *Biometry*, W. H. Freeman and Company, New York, 1995, p. 578.
6. Elashoff, J. *nQuery Advisor Sample Size and Power Determination*, Statistical Solutions Ltd., Boston, MA, 1996.
7. Hsieh, F. 'SSIZE: A sample size program for clinical and epidemiologic studies', *American Statistician*, **45**, 338 (1991).
8. SERC. *EGRET SIZ sample size and power for nonlinear regression models*, Statistics and Epidemiology Research Corp. Seattle, WA, 1992.
9. Keane, T. M., Kolb, L. C. and Thomas, R. G. 'A Psychophysiological Study of Chronic Post-Traumatic Stress Disorder', Cooperative Study No. 334, Cooperative Studies Program Coordinating Center, VA Medical Center, Palo Alto, California, U.S.A., 1988.
10. Rosner, B. *Fundamentals of Biostatistics*, 4th edn, PWS-KENT Publishing Company, 1995, p. 283 and 384.
11. Liu, G. and Liang, K. Y. 'Sample size calculation for studies with correlated observations', *Biometrics*, **53**, 537–547 (1997).
12. Anderson, T. W. *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958, p. 32.
13. SAS Institute Inc. *SAS/STAT User's Guide*, *Version 6* (*Vol. 1 and 2*), Cary, NC, 1990.