Biometrika (2001), **88**, 4, *pp*. 1193–1199

Printed in Great Britain

Sample size calculations for logistic and Poisson regression models

By GWOWEN SHIEH

Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30050, R.O.C.

gwshieh@cc.nctu.edu.tw

Summary

A method is proposed for improving sample size calculations for logistic and Poisson regression models by incorporating the limiting value of the maximum likelihood estimates of nuisance parameters under the composite null hypothesis. The method modifies existing approaches of Whittemore (1981) and Signorini (1991) and provides explicit formulae for determining the sample size needed to test hypotheses about a single parameter at a specified significance level and power. Simulation studies assess its accuracy for various model configurations and covariate distributions. The results show that the proposed method is more accurate than the previous approaches over the range of conditions considered here.

Some key words: Generalised linear model; Information matrix; Logistic regression; Maximum likelihood estimator; Poisson regression; Power; Sample size; Wald statistic.

1. Introduction

The class of generalised linear models, first introduced by Nelder & Wedderburn (1972) and later expanded by McCullagh & Nelder (1989, Ch. 2), is specified by assuming independent scalar response variables Y_i $(i = 1, \ldots, N)$ to have an exponential-family probability density of the form

$$\exp[\{Y\theta - b(\theta)\}/a(\phi) + c(Y, \phi)]. \tag{1}$$

The expected value $E(Y) = \mu$ is related to the canonical parameter θ by the function $\mu = b'(\theta)$, where b' denotes the first derivative of b. The link function g relates the linear predictors η to the mean response $\eta = g(\mu)$. The linear predictors can be written as

$$\eta = \beta_0 + X^{\mathrm{T}}\beta,$$

where X is a $K \times 1$ vector of covariates, and β_0 and $\beta = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ represent the corresponding $K + 1$ unknown regression coefficients. The scale parameter ϕ is assumed to be known. Assume (y_i, x_i), for $i = 1, \ldots, N$, is a random sample from the joint distribution of (Y, X) with probability density function $f(Y, X) = f(Y|X)f(X)$, where $f(Y|X)$ has the form defined in (1) and $f(X)$ is the probability density function for X. The form of $f(X)$ is assumed to depend on none of the unknown parameters β_0 and β. The likelihood function associated with the data is

$$L(\beta_0, \beta) = \prod_{i=1}^{N} f(y_i, x_i) = \prod_{i=1}^{N} f(y_i|x_i)f(x_i).$$

It follows from the standard asymptotic theory that the maximum likelihood estimator $(\hat{\beta}_0, \hat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ is asymptotically normally distributed with mean $(\beta_0, \beta^{\mathrm{T}})^{\mathrm{T}}$ and with variance–covariance matrix given by the inverse of the $(K + 1) \times (K + 1)$ Fisher information matrix $I(\beta_0, \beta)$, where the (i, j)th element of I is

$$I_{ij} = -E\left(\frac{\partial^2 \log L}{\partial\beta_i\,\partial\beta_j}\right) \quad (i, j = 0, \ldots, K).$$

We wish to test the composite null hypothesis $H_0: \beta_1 = 0$ against the alternative hypothesis $H_1: \beta_1 \neq 0$, while treating $(\beta_0, \beta_2, \ldots, \beta_K)$ as a nuisance parameter. The Wald-type test of this hypothesis is based on $\hat{\beta}_1/\{\mathrm{var}(\hat{\beta}_1)\}^{1/2}$, where $\hat{\beta}_1$ is the first entry of $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)^{\mathrm{T}}$ and $\mathrm{var}(\hat{\beta}_1)$ is the second diagonal term of $I^{-1}(\hat{\beta}_0, \hat{\beta})$. The actual test is performed by referring the statistic to its asymptotic distribution under the null hypothesis, which is the standard normal distribution. In general, there is no simple closed-form expression for Fisher's information matrix except in some special cases. In view of their practical importance and the existence of explicit formulae, we restrict attention to the cases of logistic and Poisson regression models. For the logistic regression model, an approximate expression for Fisher's information matrix was provided in Whittemore (1981). The approximation is based on the moment generating function of the distribution of the covariates and is valid when the overall response probability is small. A formula for determining the sample size is based on the resulting asymptotic variance of the maximum likelihood estimator of the parameters. Later, the technique was extended to the Poisson regression model in Signorini (1991). However, in this case, the expression of Fisher's information matrix is exact and there is no restriction of use in terms of the overall response level. This paper presents a direct modification of the approaches of Whittemore (1981) and Signorini (1991) to the question of sample size calculations. The major difference between our approach and theirs is that the value of the nuisance parameter under the null model is different from that under the alternative model. We use the limiting value of the maximum likelihood estimator of $(\beta_0, \beta_2, \ldots, \beta_K)$ under the constraint $\beta_1 = 0$ as specified in the null hypothesis. Self & Mauritsen (1988) and Self et al. (1992) took a similar approach to the determination of sample sizes for the score statistic and likelihood ratio test, respectively, within the framework of generalised linear models. Although their methods are applicable to logistic and Poisson regression models, they assumed all the covariates in the model to be categorical with a finite number of covariate configurations. However, here we allow the covariate to be continuous or discrete with an infinite number of configurations, as in the cases of normal and Poisson covariates.

In § 2, the methodology is described and the procedures are illustrated with examples. In § 3, simulation studies are performed and comparison made with the methods of Whittemore (1981), Signorini (1991) and Self et al. (1992).

2. The proposed method

It was shown in Whittemore (1981) that

$$I_{ij} \simeq N \exp(\beta_0) E\{X_i X_j \exp(X^{\mathrm{T}}\beta)\} \quad (i, j = 0, \ldots, K),$$

with $X_0 = 1$, for logistic regression with small response probabilities, while the equation is exact for the case of Poisson regression as described by Signorini (1991). Based on this question and the moment generating function of the distribution of covariates, $m(t) = E\{\exp(X^{\mathrm{T}}t)\}$, the asymptotic variance of the maximum likelihood estimator $\hat{\beta}_1$ of β_1 can be expressed as $V(\beta_0, \beta)/N$, where $V(\beta_0, \beta) = v(\beta)/\exp(\beta_0)$; $v(\beta)$ is the second diagonal element of $M^{-1}(\beta)$ with $M = (m_{ij})$, $m_{00} = m_0 = m$, $m_{i0} = m_{0i} = m_i$, $m_i = \partial m/\partial t_i$, and $m_{ij} = \partial^2 m/\partial t_i\,\partial t_j$, for $i, j = 0, \ldots, K$.

In order to approximate the required sample size, we need to examine the asymptotic mean and variance of $\hat{\beta}_1$ under the null model, as in Self & Mauritsen (1988). Let $(\beta_0^*, \beta_2^*, \ldots, \beta_K^*)$ denote the solution of the equation $\lim_{N\to\infty} N^{-1}E\{S_N(\beta_0, 0, \beta_2, \ldots, \beta_K)\} = 0$, where S_N represents derivatives of the loglikelihood function with respect to $(\beta_0, \beta_2, \ldots, \beta_K)$ and $E\{.\}$ denotes expectation taken with respect to the true value of $(\beta_0, \beta_1, \beta_2, \ldots, \beta_K)$. We synthesise the ideas of Whittemore (1981) and Self & Mauritsen (1988) by incorporating the value (β_0^*, β^*) with $\beta^* = (0, \beta_2^*, \ldots, \beta_K^*)$ as the asymptotic mean under the null model. Thus, the sample size needed to test the hypothesis

$H_0: \beta_1 = 0$ with specified significance level α and power $1-\gamma$ against the alternative $H_1: \beta_1 \neq 0$ is

$$N_{\text{PM}} \geqslant \left\{ \frac{V^{1/2}(\beta_0^*, \beta^*) Z_{\alpha/2} + V^{1/2}(\beta_0, \beta) Z_\gamma}{\beta_1} \right\}^2, \tag{2}$$

where Z_p is the $100(1-p)$th percentile of the standard normal distribution. In contrast, Whittemore (1981) and Signorini (1991) set the values of the nuisance parameter equal to $(\beta_0, \beta^{(0)})$ with $\beta^{(0)} = (0, \beta_2, \ldots, \beta_K)$, under both the null and alternative models. In our notation this gives

$$N_{\text{WS}} \geqslant \exp(-\beta_0) \left\{ \frac{v^{1/2}(\beta^{(0)}) Z_{\alpha/2} + v^{1/2}(\beta) Z_\gamma}{\beta_1} \right\}^2. \tag{3}$$

In order to improve the accuracy, some modification was provided in Whittemore (1981). For the univariate case, $K = 1$, the sample size is more accurately calculated as

$$N_{\text{W1}} = N_{\text{WS}} \{1 + 2\exp(\beta_0)\delta(\beta_1)\}, \tag{4}$$

where

$$\delta(\beta_1) = \frac{v^{1/2}(0) + v^{1/2}(\beta_1) R(\beta_1)}{v^{1/2}(0) + v^{1/2}(\beta_1)},$$

$$R(\beta_1) = v(\beta_1)\{m_{11}(2\beta_1) - 2m^{-1}(\beta_1) m_1(\beta_1) m_1(2\beta_1) + m^{-2}(\beta_1) m(2\beta_1) m_1^2(\beta_1)\}.$$

For the multivariate case, $K \geqslant 2$, the correction is too complicated, and a simple version was proposed for routine use:

$$N_{\text{W2}} = N_{\text{WS}} \{1 + 2\exp(\beta_0)\}. \tag{5}$$

In summary, the actual implementation of the proposed approach is as follows. For a chosen logistic or Poisson regression model with specified parameter value (β_0, β), distribution of covariates $f(X)$ and required significance α and power $1-\gamma$, the minimum sample size N needed is determined from equation (2). The test statistic is

$$\hat{\beta}_1 \{N \exp(\hat{\beta}_0)/v(\hat{\beta})\}^{1/2}, \tag{6}$$

which is referred to the standard normal distribution. The null hypothesis is rejected if the absolute value of the statistic exceeds $Z_{\alpha/2}$.

To illustrate the general formula we continue the sample size calculations in Whittemore (1981) and Signorini (1991) for logistic and Poisson regression models, respectively.

Whittemore (1981) presented the problem of testing whether or not the incidence of coronary heart disease among white males aged 39–59 is related to their serum cholesterol level. During an 18-month follow-up period, the probability of a coronary heart disease event for a subject with the population mean serum cholesterol level is judged to be 0·07, and the cholesterol levels in this population are well represented by a standard normal distribution. Hence, we consider a simple logistic regression with $\eta = \beta_0 + X\beta_1$, $\beta_0 = \log(0{\cdot}07)$, and $X \sim N(0, 1)$. To detect the odds ratio of $e^{0\cdot 1}$, corresponding to $\beta_1 = 0{\cdot}1$, and $e^{0\cdot 5}$, corresponding to $\beta_1 = 0{\cdot}5$, for a subject with a cholesterol level of one standard deviation above the mean at a significance level 0·05 and with power 0·95, equation (4) yields $N_{\text{W1}} = 21\,147$ and $N_{\text{W1}} = 839$, respectively. Here

$$\delta(\beta_1) = \frac{1 + (1 + \beta_1^2)\exp(5\beta_1^2/4)}{1 + \exp(-\beta_1^2/4)}, \quad v(\beta_1) = \exp(-\beta_1^2/2).$$

To detect the same effects with the proposed method (2), the required sample sizes are $N_{\text{PM}} = 18\,478$ and $N_{\text{PM}} = 662$, respectively. The respective values of β_0^* are $-2{\cdot}6549$ and $-2{\cdot}5541$.

Signorini (1991) discussed a study of water pollution in terms of the number of illnesses and infections contracted per swimming season for ocean swimmers versus non-ocean or infrequent swimmers. We continue to consider a simple Poisson regression for the number of infections, with

Table 1. *Calculated sample sizes and estimates of actual power at specified sample size for logistic regression*

	Proposed method		Whittemore (1981)		Self et al. (1992)	
	Power		Power		Power	
	0·90	0·95	0·90	0·95	0·90	0·95
			Bernoulli (0·5)			
Sample size[a] (N_{PM}, N_{W1}, N_S)	1739	2165	2405	2923	1938	2396
Nominal power[b] at N_{PM}	0·9000	0·9500	0·7737	0·8644	0·8667	0·9288
Estimated power	0·8802	0·9392	0·8802	0·9392	0·8658	0·9304
Error	−0·0198	−0·0108	0·1065	0·0748	−0·0009	0·0016
			Normalised Poisson (5)[c]			
Sample size[a] (N_{PM}, N_{W1}, N_S)	364	441	530	634	484	599
Nominal power[b] at N_{PM}	0·8997	0·9499	0·7302	0·8260	0·8020	0·8711
Estimated power	0·8792	0·9290	0·8792	0·9290	0·8404	0·9022
Error	−0·0205	−0·0209	0·1490	0·1030	0·0384	0·0311
			Standard normal[d]			
Sample size[a] (N_{PM}, N_{W1}, N_S)	411	507	547	666	545	675
Nominal power[b] at N_{PM}	0·9002	0·9500	0·7933	0·8755	0·8030	0·8777
Estimated power	0·8612	0·9222	0·8612	0·9222	0·8346	0·9014
Error	−0·0390	−0·0278	0·0679	0·0467	0·0316	0·0237
			Multinomial (0·76, 0·19, 0·01, 0·04)			
Sample size[a] (N_{PM}, N_{W2}, N_S)	5785	6929	6305	7545	5783	7152
Nominal power[b] at N_{PM}	0·9000	0·9500	0·8684	0·9290	0·9001	0·9439
Estimated power	0·9216	0·9542	0·9216	0·9542	0·8834	0·9262
Error	0·0216	0·0042	0·0532	0·0252	−0·0167	−0·0177
			Multinomial (0·40, 0·10, 0·10, 0·40)			
Sample size[a] (N_{PM}, N_{W2}, N_S)	3311	4075	3869	4711	3245	4014
Nominal power[b] at N_{PM}	0·9000	0·9500	0·8457	0·9152	0·9056	0·9528
Estimated power	0·9274	0·9602	0·9274	0·9602	0·9170	0·9546
Error	0·0274	0·0102	0·0817	0·0450	0·0114	0·0018
			Multinomial (0·25, 0·25, 0·25, 0·25)			
Sample size[a] (N_{PM}, N_{W2}, N_S)	1726	2150	2315	2813	1947	2408
Nominal power[b] at N_{PM}	0·8999	0·9500	0·7875	0·8756	0·8627	0·9260
Estimated power	0·8790	0·9376	0·8790	0·9376	0·8612	0·9256
Error	−0·0209	−0·0124	0·0915	0·0620	−0·0015	−0·0004

[a] Sample sizes needed to achieve power 0·9 and 0·95, respectively.
[b] Nominal powers at calculated sample sizes of the proposed method in [a].
[c] The categorical approximation of Po(5) is (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) with probabilities (0·0404, 0·0842, 0·1404, 0·1755, 0·1755, 0·1462, 0·1044, 0·0653, 0·0363, 0·0318).
[d] The categorical approximation is (−2·2, −1·7, −1·2, −0·7, −0·2, 0·2, 0·7, 1·2, 1·7, 2·2) with probabilities (0·0228, 0·0441, 0·0918, 0·1499, 0·1915, 0·1915, 0·1499, 0·0918, 0·0441, 0·0228).

a single Bernoulli covariate indicating ocean swimming if $X = 1$ and with $\pi = \text{pr}(X = 1) = 0{\cdot}5$. In this case, the estimated infection rate of non-ocean or infrequent swimmers is 0·85, corresponding to $\beta_0 = \log(0{\cdot}85)$, and the ratio of mean response for $X = 1$ to mean response for $X = 0$ is 1·3, corresponding to $\beta_1 = \log(1{\cdot}3)$. It follows from equation (3) with $v(\beta_1) = \{\pi \exp(\beta_1)\}^{-1} + (1 - \pi)^{-1}$ that the sample sizes N_{WS} needed for a significance level 0·05 with power 0·80, 0·90 and 0·95 are 518, 685 and 841, respectively. Our equation (2) gives 469, 629 and 779 at the three corresponding power levels and $\beta_0^* = -0{\cdot}0228$.

Table 2. *Calculated sample sizes and estimates of actual power at specified sample size for Poisson regression*

	Proposed method		Signorini (1991)		Self et al. (1992)	
	Power		Power		Power	
	0·90	0·95	0·90	0·95	0·90	0·95
			Bernoulli (0·5)			
Sample size[a] (N_{PM}, N_{WS}, N_S)	1835	2285	2354	2861	1856	2295
Nominal power[b] at N_{PM}	0·9001	0·9500	0·8069	0·8905	0·8968	0·9492
Estimated power	0·8880	0·9494	0·8880	0·9494	0·8918	0·9514
Error	−0·0121	−0·0006	0·0811	0·0589	−0·0050	0·0022
			Normalised Poisson (5)[c]			
Sample size[a] (N_{PM}, N_{WS}, N_S)	389	472	464	554	455	563
Nominal power[b] at N_{PM}	0·8999	0·9498	0·8299	0·9060	0·8505	0·9104
Estimated power	0·8966	0·9283	0·8966	0·9382	0·8802	0·9278
Error	−0·0033	−0·0116	0·0667	0·0322	0·0297	0·0174
			Standard normal[d]			
Sample size[a] (N_{PM}, N_{WS}, N_S)	437	541	508	619	512	633
Nominal power[b] at N_{PM}	0·8997	0·9500	0·8485	0·9185	0·8502	0·9154
Estimated power	0·8762	0·9404	0·8762	0·9404	0·8748	0·9390
Error	−0·0235	−0·0096	0·0277	0·0219	0·0282	0·0203
			Multinomial (0·76, 0·19, 0·01, 0·04)			
Sample size[a] (N_{PM}, N_{WS}, N_S)	6165	7383	6218	7442	4933	6101
Nominal power[b] at N_{PM}	0·9000	0·9500	0·8971	0·9483	0·9519	0·9776
Estimated power	0·9462	0·9706	0·9462	0·9706	0·9364	0·9656
Error	0·0462	0·0206	0·0491	0·0223	−0·0155	−0·0120
			Multinomial (0·40, 0·10, 0·10, 0·40)			
Sample size[a] (N_{PM}, N_{WS}, N_S)	3537	4353	3963	4825	3143	3886
Nominal power[b] at N_{PM}	0·9000	0·9500	0·8615	0·9265	0·9305	0·9682
Estimated power	0·9342	0·9692	0·9342	0·9692	0·9400	0·9710
Error	0·0342	0·0192	0·0727	0·0427	0·0095	0·0028
			Multinomial (0·25, 0·25, 0·25, 0·25)			
Sample size[a] (N_{PM}, N_{WS}, N_S)	1835	2285	2354	2861	1856	2295
Nominal power[b] at N_{PM}	0·9001	0·9500	0·8069	0·8905	0·8968	0·9492
Estimated power	0·8904	0·9484	0·8904	0·9484	0·8954	0·9506
Error	−0·0097	−0·0016	0·0835	0·0579	−0·0014	0·0014

[a] Sample sizes needed to achieve power 0·9 and 0·95, respectively.
[b] Nominal powers at calculated sample sizes of the proposed method in [a].
[c] The categorical approximation of Po(5) is (1, 2, 3, 4, 5, 6, 7, 8, 9, 10) with probabilities (0·0404, 0·0842, 0·1404, 0·1755, 0·1755, 0·1462, 0·1044, 0·0653, 0·0363, 0·0318).
[d] The categorical approximation is (−2·2, −1·7, −1·2, −0·7, −0·2, 0·2, 0·7, 1·2, 1·7, 2·2) with probabilities (0·0228, 0·0441, 0·0918, 0·1499, 0·1915, 0·1915, 0·1499, 0·0918, 0·0441, 0·0228).

3. Simulation studies

The finite-sample adequacy of our formula was assessed through simulations, in which we also compared the proposed method to those of Whittemore (1981), Signorini (1991) and Self et al. (1992). The results are presented in Tables 1 and 2 for logistic and Poisson regression models, respectively.

For both regression models, two linear predictors are examined, namely $\eta = \beta_0 + X_1\beta_1$ and $\eta = \beta_0 + X_1\beta_1 + X_2\beta_2$. In the case of the simple linear predictor $\eta = \beta_0 + X_1\beta_1$, we consider Ber(0·5), normalised Po(5) and standard normal distributions for the covariate X_1. For the second predictor

$\eta = \beta_0 + X_1\beta_1 + X_2\beta_2$, the joint distribution of (X_1, X_2) is assumed to be multinomial with probabilities π_1, π_2, π_3 and π_4, corresponding to (x_1, x_2) values of (0, 0), (0, 1), (1, 0) and (1, 1), respectively. Three sets of $(\pi_1, \pi_2, \pi_3, \pi_4)$ are studied to represent different distributional shapes, namely (0·76, 0·19, 0·01, 0·04), (0·4, 0·1, 0·1, 0·4) and (0·25, 0·25, 0·25, 0·25). Both the parameter of interest β_1 and the confounding parameter β_2 are taken to be log 2. The intercept parameter β_0 is chosen to satisfy the overall response $\bar{\mu} = 0{\cdot}05$, where $\bar{\mu} = E[\exp(\eta)/\{1 + \exp(\eta)\}]$ and $\bar{\mu} = E\{\exp(\eta)\}$ for the logistic and Poisson regression models, respectively.

First, we calculated from equations (2)–(5) the sample sizes $(N_{\text{PM}}, N_{\text{WS}}, N_{\text{W1}}, N_{\text{W2}})$ required to achieve the selected significance 0·05 and power (0·90, 0·95) within the model specifications. Let N_{S} denote the sample size for the approach of Self et al. (1992). Since they assumed all of the covariates to be categorical with a finite number of configurations, discretisation schemes are needed for the cases of Poisson and standard normal covariates; the chosen schemes are listed in the footnotes of Tables 1 and 2. These estimates of sample size allow comparison of relative efficiencies of the approaches. Since the magnitude of the sample size affects the accuracy of the asymptotic distribution and the resulting formulae, we unify the sample sizes in the simulations; the sample size N_{PM} is chosen as the benchmark and is used to recalculate the nominal powers for all competing approaches.

Estimates of the true power associated with given sample size and model configuration are then computed through Monte Carlo simulation based on 5000 independent datasets. For each replicate, N_{PM} covariate values are generated from the selected distribution. These covariate values determine the incidence rates for generating N_{PM} Bernoulli or Poisson outcomes. Then the test statistic is computed and the estimated power is the proportion of the 500 replicates whose test statistic values exceed the critical value. The adequacy of the sample size formula is determined by the difference between the estimated power and nominal power specified above. All calculations are performed using programs written with SAS/IML (SAS Institute, 1989).

The results in Tables 1 and 2 suggest that there is a close agreement between the estimated power and the nominal power for the proposed method regardless of the model configuration and covariate distribution. The only exceptions are with the extremely skewed multinomial covariate distribution (0·76, 0·19, 0·01, 0·04). The approach of Self et al. (1992) is also very good at achieving the nominal levels, but the approaches of Whittemore (1981) and Signorini (1991) incur much larger errors. We conclude that the proposed method maintains the accuracy within a reasonable range of nominal power and is much more accurate than the previous approaches proposed by Whittemore (1981) and Signorini (1991). Nevertheless, unbalanced allocations appear to degrade the accuracy of sample size calculations.

ACKNOWLEDGEMENT

I would like to thank the editor and referees for their helpful comments that improved the presentation of this paper. This work was partially supported by the National Science Council of Taiwan.

REFERENCES

MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hall.

NELDER, J. A. & WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–84.

SAS INSTITUTE (1989). *SAS/IML Software: Usage and Reference*, Version 6. Carey, NC: SAS Institute Inc.

SELF, S. G. & MAURITSEN, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44**, 79–86.

Self, S. G., Mauritsen, R. H. & Ohara, J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* **48**, 31–9.
Signorini, D. F. (1991). Sample size for Poisson regression. *Biometrika* **78**, 446–50.
Whittemore, A. S. (1981). Sample size for logistic regression with small response probability. *J. Am. Statist. Assoc.* **76**, 27–32.

[*Received March* 2000. *Revised April* 2001]