Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

Journal of Statistical Planning and Inference 128 (2005) 43–59

journal of statistical planning and inference

www.elsevier.com/locate/jspi

On power and sample size calculations for Wald tests in generalized linear models

Gwowen Shieh

Department of Management Science, National Chiao Tung University, Hsinchu 30050, Taiwan, ROC

Received 27 February 2003; accepted 6 September 2003

Abstract

A Wald test-based approach for power and sample size calculations has been presented recently for logistic and Poisson regression models using the asymptotic normal distribution of the maximum likelihood estimator, which is applicable to tests of a single parameter. Unlike the previous procedures involving the use of score and likelihood ratio statistics, there is no simple and direct extension of this approach for tests of more than a single parameter. In this article, we present a method for computing sample size and statistical power employing the discrepancy between the noncentral and central chi-square approximations to the distribution of the Wald statistic with unrestricted and restricted parameter estimates, respectively. The distinguishing features of the proposed approach are the accommodation of tests about multiple parameters, the flexibility of covariate configurations and the generality of overall response levels within the framework of generalized linear models. The general procedure is illustrated with some special situations that have motivated this research. Monte Carlo simulation studies are conducted to assess and compare its accuracy with existing approaches under several model specifications and covariate distributions.

Keywords: Information matrix; Likelihood ratio test; Logistic regression; Noncentral chi-square; Poisson regression; Score test

1. Introduction

Generalized linear models were first introduced by Nelder and Wedderburn (1972) and are broadly applicable in almost all scientific fields. A thorough development can be found in McCullagh and Nelder (1989). The class of generalized linear models is specified by assuming that independent scalar response variables Y_i, $i=1, \ldots, N$, follow

E-mail address: gwshieh@mail.nctu.edu.tw (G. Shieh).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2003.09.017

a probability distribution belonging to the exponential family of probability distributions with probability function 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

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

where $\mathbf{X}=(X_1, \ldots, X_k)^{\mathrm{T}}$ is a $k \times 1$ vector of covariates, and $\beta=(\beta_1, \ldots, \beta_k)^{\mathrm{T}}$ represents the corresponding $k \times 1$ vector of unknown regression coefficients. The scale parameter ϕ is assumed to be known. Assume $(y_i, \mathbf{x}_i)$ is a random sample from the joint distribution of $(Y, \mathbf{X})$ with probability function $f(Y, \mathbf{X}) = f(Y|\mathbf{X})f(\mathbf{X})$, where $f(Y|\mathbf{X})$ has the form defined in (1) and $f(\mathbf{X})$ is the probability function for $\mathbf{X}$. The form of $f(\mathbf{X})$ is assumed to depend on none of the unknown parameters β. The likelihood function associated with the data is

$$L(\beta) = \prod_{i=1}^{N} f(y_i, \mathbf{x}_i) = \prod_{i=1}^{N} f(y_i|\mathbf{x}_i)f(\mathbf{x}_i).$$

Let $\beta_1 = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ and $\beta_2 = (\beta_{q+1}, \ldots, \beta_k)^{\mathrm{T}}$ represent the first q and the last p unknown regression coefficients of β, respectively ($k=q+p$, $q \geqslant 1$, $p \geqslant 1$). We wish to test the composite null hypothesis H_0: $\beta_2 = \mathbf{0}$ against the alternative hypothesis H_1: $\beta_2 \neq \mathbf{0}$, while treating β_1 as nuisance parameters. It follows from the standard asymptotic theory that the maximum likelihood estimator $\hat{\beta} = (\hat{\beta}_1^{\mathrm{T}}, \hat{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ is asymptotically normally distributed with mean $\beta = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ and with variance–covariance matrix given by the inverse of the $k \times k$ Fisher information matrix $\mathbf{I}(\beta_1, \beta_2)$, where the (i,j)th element of $\mathbf{I}$ is

$$I_{ij} = -E\left(\frac{\partial^2 \log L}{\partial \beta_i \partial \beta_j}\right), \quad i, j = 1, \ldots, k$$

and $E[\,\cdot\,]$ denotes the expectation taken with respect to the joint distribution of $(Y_1, \ldots, Y_N, \mathbf{X}_1, \ldots, \mathbf{X}_N)$. The Wald test statistic of the hypothesis is

$$W = \hat{\beta}_2^{\mathrm{T}} \hat{\mathbf{V}}^{-1} \hat{\beta}_2, \tag{2}$$

where $\hat{\mathbf{V}}$ is the lower-right $p \times p$ sub-matrix of $\mathbf{I}^{-1}(\hat{\beta}_1, \hat{\beta}_2)$. The actual test is performed by referring the statistic to its asymptotic distribution under the null hypothesis, which is a chi-square distribution with p degrees of freedom. In general, there is no simple closed-form expression for Fisher's information matrix except in some special cases.

For the purpose of power and sample size calculations, an approximate expression for Fisher's information matrix was provided in Whittemore (1981) for logistic regression model. The approximation employs the moment generating function of the covariates and is valid when the overall response probability is small. A formula for determining the sample size is developed from the resulting asymptotic variance of the maximum likelihood estimator of the parameters. Later, the technique was extended to the Poisson