Sample size determination for logistic regression: A simulation study

Stephen Bush
School of Mathematical Sciences, University of Technology Sydney,
PO Box 123 Broadway NSW 2007, Australia

Abstract

This paper considers the different methods for determining sample sizes for Wald, likelihood ratio, and score tests for logistic regression. We review some recent methods, report the results of a simulation study comparing each of the methods for each of the three types of test, and provide Mathematica code for calculating sample size. We consider a variety of covariate distributions, and find that a calculation method based on a first order expansion of the likelihood ratio test statistic performs consistently well in achieving a target level of power for each of the three types of test.
Keywords: Power, Score Test, Likelihood Ratio Test, Wald Test, Mathematica.

1 Introduction

Sample size determination is an important step in the planning of an experiment or observational study, with serious consequences if not done carefully. If the chosen sample size is not large enough then the probability that an effect of practical significance is found to be statistically significant is not large enough for the study to be useful. If the sample size is too large then the study becomes more expensive than necessary.

The choice of sample size depends on a number of things. The first is the model that is to be used to analyse the data and the type of test that is used to determine significance. In addition, researchers need to set a level of significance, the desired level of power, and the effect size that they wish to be able to detect.

In this paper, we consider sample size determination for logistic regression when the Wald, likelihood ratio, and score tests are used to determine significance. There are several approaches to determining sample size based on each of these tests. While each of these methods have been tested to see whether the target power is achieved when the test the method was developed for is used, the effectiveness of these methods when using one of the other tests has not. The novel contribution of this paper is to, where possible, test all of the sample size determination methods using all three tests, to see whether there is much difference between the power obtained form each of the tests and whether some

of these methods consistently outperform others. We also present an implementation of the methods discussed in this paper in the computer algebra system Mathematica 8.

In the next section we review the Wald, likelihood ratio and score tests for logistic regression. In Section 3, we provide a more detailed review of some of the sample size determination methods. We consider a formulation of each of the tests that permit the testing of multiple parameters at once, as Shieh, 2005 considered for the Wald test. The benefit of this approach is that we can test the significance of factors that may be explained by multiple parameters, whether that be in the form of a multinomial factor or in the form of a polynomial response surface. In Section 4, we discuss the results of a simulation study comparing the performance of the methods discussed in Section 3, for each of the tests.

2 Logistic Regression

In this section, we review the Wald, likelihood ratio, and score tests for logistic regression. We use the concepts and notation described in this section later to describe the different approaches to determining sample size.

The logistic regression model is a member of the class of generalised linear models (Nelder and Wedderburn, 1972). In generalised linear models, the expected value of the response variable y is related to a linear combination of p predictor variables $\boldsymbol{x} = (x_1, \ldots, x_p)$ and q nuisance variables $\boldsymbol{z} = (z_1, \ldots, z_q)$ through an inverse link function $b'(\boldsymbol{\psi}, \boldsymbol{\lambda})$. This linear combination is denoted by $\eta_i = \boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i$, where $\boldsymbol{\psi}$ is a $1 \times p$ vector containing the regression coefficients of the predictor variables and $\boldsymbol{\lambda}$ is a $1 \times q$ vector containing the regression coefficients of the nuisance variables. For logistic regression the inverse link function is $b'(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \exp(\eta_i)/(1 + \exp(\eta_i))$.

Each of the Wald, likelihood ratio, and score tests are based on properties of the likelihood function of the model being fitted. For logistic regression, the log likelihood function is given by

$$\ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \sum_{i=1}^{n} \left(y_i \left(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i \right) - n_i \log \left(1 + \exp(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i) \right) \right),$$

(Agresti, 2003).

Using this expression, we define the score vector and the Fisher information matrix. For logistic regression, the j^{th} entry in the score vector is given by

$$S(\boldsymbol{\psi}, \boldsymbol{\lambda})_j = E\left(\frac{\partial \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}, \phi; \boldsymbol{y})}{\partial \psi_j} \right) = \sum_{i=1}^{n} y_i x_{ij} - n_i x_{ij} \frac{\exp(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i)}{1 + \exp(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i)}.$$

A similar expression exists when we differentiate $\ell(\boldsymbol{\psi}, \boldsymbol{\lambda})$ by an entry in $\boldsymbol{\lambda}$. The $(j, k)^{\text{th}}$ entry of the Fisher information matrix for the logistic regression model

is given by

$$I(\boldsymbol{\psi},\boldsymbol{\lambda})_{j,k} = E\left(-\frac{\partial^2 \ell(\boldsymbol{\psi},\boldsymbol{\lambda})}{\partial\psi_j \partial\psi_k}\right) = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}\exp(\boldsymbol{\psi}^T\boldsymbol{x}_i + \boldsymbol{\lambda}^T\boldsymbol{z}_i)}{\left(1+\exp(\boldsymbol{\psi}^T\boldsymbol{x}_i + \boldsymbol{\lambda}^T\boldsymbol{z}_i)\right)^2}.$$

Similar expressions exist when we differentiate $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ by two entries in $\boldsymbol{\lambda}$, or by one entry in $\boldsymbol{\psi}$ and one entry in $\boldsymbol{\lambda}$. When nuisance variables exist, it is useful to partition the score vector into the derivatives of entries in $\boldsymbol{\psi}$ and the derivatives of entries in $\boldsymbol{\lambda}$. So

$$S(\boldsymbol{\psi},\boldsymbol{\lambda}) = \begin{bmatrix} S_\psi(\boldsymbol{\psi},\boldsymbol{\lambda}) \\ S_\lambda(\boldsymbol{\psi},\boldsymbol{\lambda}) \end{bmatrix},$$

where $S_\psi(\boldsymbol{\psi},\boldsymbol{\lambda})$ contains the derivatives of $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ with respect to entries in $\boldsymbol{\psi}$ and $S_\lambda(\boldsymbol{\psi},\boldsymbol{\lambda})$ contains the derivatives of $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ with respect to entries in $\boldsymbol{\lambda}$. Similarly, we partition the entries in the Fisher information matrix into entries where $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ has been differentiated with respect to two entries in $\boldsymbol{\psi}$, entries where $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ has been differentiated with respect to two entries of $\boldsymbol{\lambda}$, and entries where $\ell(\boldsymbol{\psi},\boldsymbol{\lambda})$ has been differentiated with respect to one entry in $\boldsymbol{\psi}$ and one entry in $\boldsymbol{\lambda}$. So the Fisher information matrix and the its inverse, the covariance matrix, can be expressed as

$$I(\boldsymbol{\psi},\boldsymbol{\lambda}) = \begin{bmatrix} I_{\psi\psi}(\boldsymbol{\psi},\boldsymbol{\lambda}) & I_{\psi\lambda}(\boldsymbol{\psi},\boldsymbol{\lambda}) \\ I_{\lambda\psi}(\boldsymbol{\psi},\boldsymbol{\lambda}) & I_{\lambda\lambda}(\boldsymbol{\psi},\boldsymbol{\lambda}) \end{bmatrix} \text{ and } \Sigma(\boldsymbol{\psi},\boldsymbol{\lambda}) = \begin{bmatrix} \Sigma_{\psi\psi}(\boldsymbol{\psi},\boldsymbol{\lambda}) & \Sigma_{\psi\lambda}(\boldsymbol{\psi},\boldsymbol{\lambda}) \\ \Sigma_{\lambda\psi}(\boldsymbol{\psi},\boldsymbol{\lambda}) & \Sigma_{\lambda\lambda}(\boldsymbol{\psi},\boldsymbol{\lambda}) \end{bmatrix},$$

respectively.

Following Cox and Hinkley, 1979, we define the Wald, likelihood ratio, and score test statistics for testing the hypothesis $H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0$ against the hypothesis $H_1 : \boldsymbol{\psi} \neq \boldsymbol{\psi}_0$ as

$$W_{\text{Wald}} = (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0)^T \Big(\Sigma_{\psi\psi}(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\lambda}})\Big)^{-1} (\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0) \tag{1}$$

$$W_{\text{LR}} = 2\Big(\ell(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\lambda}}) - \ell(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}_0})\Big) \tag{2}$$

$$W_{\text{Score}} = S_\psi(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}_0})^T \Big(\Sigma_{\psi\psi}(\widehat{\boldsymbol{\psi}},\widehat{\boldsymbol{\lambda}})\Big)^{-1} S_\psi(\boldsymbol{\psi}_0,\widehat{\boldsymbol{\lambda}_0}) \tag{3}$$

respectively, where the entries of $\widehat{\boldsymbol{\lambda}_0}$ are the maximum likelihood estimates of $\boldsymbol{\lambda}$ under the null hypothesis $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. We note that $\widehat{\boldsymbol{\lambda}_0}$ is not a consistent estimator for $\boldsymbol{\lambda}$, rather it converges to some value $\boldsymbol{\lambda}_0^*$. Under the null hypothesis, each of these test statistics have a central chi–squared distribution with p degrees of freedom. Under the alternative hypothesis, the test statistics have a non–central chi–squared distribution with p degrees of freedom and non–centrality parameter equal to the expected value of the test statistic under the alternative hypothesis. Then the power of the test can be expressed as

$$1 - \beta = P(\chi_p^2(\gamma) > \chi_{p,\alpha}^2), \tag{4}$$

where $\chi^2_p(\gamma)$ is a non–central chi–square distribution with p degrees of freedom and non–centrality parameter γ, and $\chi^2_{p,\alpha}$ is the upper α percentile of of a central chi square distribution with p degrees of freedom. The value for γ is W_{Wald}, W_{LR}, or W_{Score} when we use the Wald, likelihood ratio, or score tests, respectively.

3 Methods for Calculating Sample Size

In this section, we discuss some methods for determining sample size for logistic regression. Later we compare these methods in a simulation study to determine how well the sample sizes obtained from each method achieve the target power when the Wald, likelihood ratio, and score tests are used.

3.1 Wald Test

Several authors have considered sample size determination for the Wald test for logistic regression. Whittemore, 1981 considers sample size determination for logistic regression on a single parameter when the probability of response is small and the Wald test is used to test hypotheses. This method involves approximating the variance of the parameter estimates, and then correcting the sample size to account for this approximation. Shoenfeld and Borenstein, 2005 shows that the approach described by Whittemore works well for response probabilities as large as 0.27. Wilson and Gordon, 1986 extend Whittemore's approach to incorporate nuisance variables into the calculations.

Hsieh et. al., 1998 consider an approach based on treating the response probability as continuous and comparing two samples using a 2-sample t test, where the two samples are obtained from two different predictor values, 0 and 1 for a Bernoulli distributed predictor, and μ and $\mu+\sigma$ for a normally distributed predictor. This approach is modified in Novikov et. al., 2010 to incorporate the sample size formula presented in Schouten, 1999. Hsieh et. al., 1998 also extends the work of Whittemore to allow for nuisance variables using the variance inflation factor (VIF) to adjust the variance function, and then using this adjusted variance to obtain sample sizes.

Shieh, 2005 presents two approaches, a so–called direct approach and an approach that calculates an adjusted significance level that corrects for the different Fisher information matrices under the null and alternative hypotheses. In both cases, we begin by calculating the Fisher information matrix evaluated at $\boldsymbol{\psi}_0$ and $\boldsymbol{\lambda}_0^*$,

$$I(\boldsymbol{\psi},\boldsymbol{\lambda}) = E_{X,Z}\left(\frac{\exp(\boldsymbol{\psi}^T\boldsymbol{x}_i+\boldsymbol{\lambda}^T\boldsymbol{z}_i)}{\left(1+\exp(\boldsymbol{\psi}^T\boldsymbol{x}_i+\boldsymbol{\lambda}^T\boldsymbol{z}_i)\right)^2}\left(\boldsymbol{X}^T,\boldsymbol{Z}^T\right).\left(\boldsymbol{X}^T,\boldsymbol{Z}^T\right)^T\right).$$

In the direct method, we calculate the target non–centrality parameter γ, such that $\chi^2_{p,1-\beta}(\lambda) = \chi^2_{p,\alpha}$, where $1-\beta$ is the desired level of power. The sample size is then $N = \lambda/\delta$, where $\delta = \boldsymbol{\psi}^T\left(\Sigma_{\psi\psi}(\boldsymbol{\psi},\boldsymbol{\gamma})\right)^{-1}\boldsymbol{\psi}$.

In the adjusted method, we begin by calculating the adjusted level of significance,

$$\alpha^* = P\bigg(Z^T \Big(\Sigma_{\psi\psi}(\boldsymbol{\psi}, \boldsymbol{\lambda}) \Big)^{-1} Z > \chi^2_{p,\alpha} \bigg), \tag{5}$$

where Z is a vector of uncorrelated standard normal random variables. In this paper, the author presents a simplification of Equation 5 using a three parameter F distribution. Once we have obtained the adjusted level of significance, we calculate the target non–centrality parameter γ^*, such that $\chi^2_{p,1-\beta}(\gamma^*) = \chi^2_{p,\alpha^*}$. The sample size is then $N = \gamma^*/\delta$ for the adjusted method, where $\delta = \boldsymbol{\psi}^T \big(\Sigma_{\psi\psi}(\boldsymbol{\psi}, \boldsymbol{\lambda}) \big)^{-1} \boldsymbol{\psi}$. Shoenfeld and Borenstein, 2005 presents a theorem that reduces the numerical complexity of sample size determination for the Wald test to a single integral.

Demidenko, 2007 presents explicit derivations for the Fisher information matrix of the logistic regression model with a single Bernoulli distributed predictor, and the logistic regression model with one Bernoulli distributed predictor variable and one Bernoulli distributed nuisance variable. The author then uses these matrices to obtain explicit formulae for sample sizes. For logistic regression with one Bernoulli distributed predictor and no nuisance variables, the author obtains

$$n \geq (Z_{1-\alpha/2} + Z_{1-\beta})^2 \times \frac{p_x(1+A)^2 B + (1-p_x)(1+AB)^2}{p_x(1-p_x)AB(\psi - \psi_0)^2},$$

where $A = \exp(\beta_0)$, where β_0 is the intercept term, $B = \exp(\psi)$, and p_x is the proportion of observations with $X = 1$. **A sample size calculator for the methods introduced by this author can be found at**

`http://www.dartmouth.edu/~eugened/power-samplesize.php`.

Novikov et. al., 2010 present a sample size determination method for logistic regression with a single normally distributed predictor variable. The authors extend a method initially presented in Hsieh et. al., 1998 that approximates the sample size for logistic regression with that of a two–sample t–test. This extension is made by using Schouten's sample size formula (Schouten, 1999) to allow for unequal sample sizes in the $Y = 0$ and $Y = 1$ groups. The authors implement this method in the SAS software system.

Lyles et. al., 2006 use expanded data sets to determine power for a given sample size. In this approach, we choose a value for N, and construct a data set with N entries. The values of the predictor and nuisance variables in each of these entries are chosen such that the joint probability distribution of the variables is reflected in the data set. This data set is then copied so we obtain one copy of the data set with $Y = 0$ and one copy with $Y = 1$. Each entry in the data set is weighted by $P(Y = y | \boldsymbol{Z} = \boldsymbol{z}_i, \boldsymbol{X} = \boldsymbol{x}_i)$. We then use the weights to fit a model, calculate either W_{Wald} or W_{LR}, and use this as the non–centrality parameter in Equation 4 to calculate the power of the test. The sample size is

then adjusted until the desired power is achieved. The authors argue that the benefit of this approach is that we are able to model more complex relationships between the predictor and nuisance variables, so long as we are able to list the distinct combinations of these variables and can calculate joint probabilities.

3.2 Likelihood Ratio Test

Self et. al., 1992 considers sample size determination for testing multiple parameters using the likelihood ratio test. Their approach incorporates nuisance variables, but both the predictors and nuisance variables must either be discrete or be discretised; see Shieh, 2000a for instance. Shieh, 2000b presents a generalisation of the likelihood ratio test approach considered in Self et. al., 1992. Both papers consider the first order expansion of the Wald test statistic

$$W_{\text{LR}} = 2\Big(\ell(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\lambda}}) - \ell(\boldsymbol{\psi}, \boldsymbol{\lambda})\Big) - 2\Big(\ell(\boldsymbol{\psi}_0, \widehat{\boldsymbol{\lambda}_0}) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\Big) + 2\Big(\ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\Big).$$

Both papers argue that the first term has an expected value of $p + q$ and that, if $(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)$ is equal to the true parameter values, the first order expansion of the second term is equal to q. Self et. al., 1992 shows that the third term is equal to

$$\triangle_{\text{SMO}} = 2\sum_{i=1}^{k}\left(\frac{\big((\boldsymbol{\psi} - \boldsymbol{\psi}_0)^T\boldsymbol{x}_i + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0^*)^T\boldsymbol{z}_i\big)\exp(\boldsymbol{\psi}^T\boldsymbol{x}_i + \boldsymbol{\lambda}^T\boldsymbol{z}_i)}{1 + \exp(\boldsymbol{\psi}^T\boldsymbol{x}_i + \boldsymbol{\lambda}^T\boldsymbol{z}_i)} - \log\Big(\frac{1 + \exp(\boldsymbol{\psi}^T\boldsymbol{x}_i + \boldsymbol{\lambda}^T\boldsymbol{z}_i)}{1 + \exp(\boldsymbol{\psi}_0^T\boldsymbol{x}_i + (\boldsymbol{\lambda}_0^*)^T\boldsymbol{z}_i)}\Big)\right),$$

where there are $k < \infty$ covariate configurations. Shieh, 2000b generalises this expression to allow for an infinite number of covariate configurations, as is required if one or more of the variables are continuous. He obtains

$$\triangle_{\text{Shieh}} = 2E_{X,Z}\left(\frac{\big((\boldsymbol{\psi} - \boldsymbol{\psi}_0)^T\boldsymbol{X} + (\boldsymbol{\lambda} - \boldsymbol{\lambda}_0^*)^T\boldsymbol{Z}\big)\exp(\boldsymbol{\psi}^T\boldsymbol{x} + \boldsymbol{\lambda}^T z)}{1 + \exp(\boldsymbol{\psi}^T\boldsymbol{x} + \boldsymbol{\lambda}^T\boldsymbol{z})} - \log\Big(\frac{1 + \exp(\boldsymbol{\psi}^T\boldsymbol{x} + \boldsymbol{\lambda}^T\boldsymbol{z})}{1 + \exp(\boldsymbol{\psi}_0^T\boldsymbol{x} + (\boldsymbol{\lambda}_0^*)^T\boldsymbol{z})}\Big)\right).$$

To calculate the sample size using either method we calculate the target non–centrality parameter γ by solving $\chi^2_{p,1-\beta}(\gamma) = \chi^2_{p,\alpha}$, and then find $N = \gamma/\triangle$.

3.3 Score Test

Self and Mauritsen, 1988 considers sample size determination for testing multiple parameters using the score test. The authors approximate $S_\psi(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)$ using a first order Taylor series expansion,

$$S_\psi(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \approx S_\psi(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) - I_{\psi\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\Big(I_{\lambda\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\Big)^{-1} S_\lambda(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*), \qquad (6)$$

as discussed in Cox and Hinkley, 1979. The authors then state that W_{score} follows a non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\boldsymbol{\xi}_N^T \Sigma_N^{-1} \boldsymbol{\xi}_N$, where $\boldsymbol{\xi}_N$ and Σ_N are the mean vector and covariance matrix of Equation 6, respectively. For GLMs with discrete predictors and nuisance variables, the authors use

$$\boldsymbol{\xi}_N = N \sum_{i=1}^{m} \pi_i \left(\frac{\exp(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i)}{1 + \exp(\boldsymbol{\psi}^T \boldsymbol{x}_i + \boldsymbol{\lambda}^T \boldsymbol{z}_i)} - \frac{\exp(\boldsymbol{\psi}_0^T \boldsymbol{x}_i + (\boldsymbol{\lambda}_0^*)^T \boldsymbol{z}_i)}{1 + \exp(\boldsymbol{\psi}_0^T \boldsymbol{x}_i + (\boldsymbol{\lambda}_0^*)^T \boldsymbol{z}_i)} \right) \times Z_i^*,$$

$$\Sigma_N = N \sum_{i=1}^{m} \pi_i \frac{\exp(\boldsymbol{\psi} \boldsymbol{x}_i^T + \boldsymbol{\lambda} \boldsymbol{z}_i^T)}{\left(1 + \exp(\boldsymbol{\psi} \boldsymbol{x}_i^T + \boldsymbol{\lambda} \boldsymbol{z}_i^T)\right)^2} \times Z_i^* (Z_i^*)^T,$$

where $Z_i^* = Z_i - I_{\psi\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \big(I_{\lambda\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\big)^{-1} X_i$, and π_i is the proportion of the sample with the i^{th} combination of variable levels. We then calculate the non–centrality parameter $\gamma_N = \boldsymbol{\xi}_N^T \Sigma_N^{-1} \boldsymbol{\xi}_N$, and find the required sample size by solving $\chi^2_{p,1-\beta}(\gamma_N) = \chi^2_{p,\alpha}$ for N.

We also consider an approach to the score test where we use expectation to construct $\boldsymbol{\xi}_N$ and Σ_N, rather than weighted sums over covariate distribution. This is reminiscent of the generalisation of the method of Self et. al., 1992 by Shieh, 2000b to allow for continuous predictor variables. So

$$\boldsymbol{\xi}_N = N \times \mathrm{E}_{X,Z} \left(S_\psi(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) - I_{\psi\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \big(I_{\lambda\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\big)^{-1} S_\lambda(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \right)$$

$$\Sigma_N = N \times \mathrm{Var}_{X,Z} \left(S_\psi(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) - I_{\psi\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \big(I_{\lambda\lambda}(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*)\big)^{-1} S_\lambda(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0^*) \right)$$

Then to calculate the sample size we find N such that Equation 4 is satisfied, where the non–centrality parameter is $\gamma_N = \boldsymbol{\xi}_N^T \Sigma_N^{-1} \boldsymbol{\xi}_N$. We find the required sample size by solving $\chi^2_{p,1-\beta}(\gamma_N) = \chi^2_{p,\alpha}$ for N.

4 Simulation Study

In this section we present the results of a simulation study that investigates how well sample sizes obtained using each of the methods described in the previous section achieve the target level of power.

In these simulations, five different combinations of predictor and nuisance variables will be considered. These are described in Table 1. We code the variables using effects coding, since this is an example of a set of contrasts, which provide a flexible method of conducting different comparisons between levels of a factor. To determine the intercepts under the null and alternate hypotheses, we solve $E_{X,Z}\left(b'(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)\right) = k$, and $E_{X,Z}\left(b'(\boldsymbol{\psi}, \boldsymbol{\lambda})\right) = k$, respectively. In each scenario, we estimate the sample size required to estimate a given set of effects at the 5% significance level with 90% power, and with 95% power.

To perform the simulations, we use the methods in Section 3 to obtain the sample size estimates. For each sample size obtained, we construct 10000 simulated data sets in R by sampling N observations from the joint distributions

	Linear Predictor	**Distributions**	**Coding**	**k**		**H_0**	**H_1**
1	$\beta_0 + \psi_1 x_1$	$x_1 \sim \text{Bin}(r)$ $r = 0.1, 0.3, 0.5, 0.7, 0.9$	$x_1 = -\frac{1}{2}, \frac{1}{2}$	0.2	ψ_1	0	ln(2)
2	$\beta_0 + \psi_1 x_1$	$x_1 \sim N(0,1)$		0.2	ψ_1	0	ln(2)
3	$\beta_0 + \psi_1 x_1 + \psi_2 x_2$	$(x_1, x_2, 1 - x_1 - x_2) \sim \text{Multinomial}(p1, p2, p3)$ $(p1, p2, p3) = (0.3, 0.3, 0.4), (0.1, 0.2, 0.7)$, and $(0.7, 0.1, 0.2)$	$x_1 = -\frac{1}{2}, \frac{1}{2}$ $x_2 = -\frac{1}{2}, \frac{1}{2}$	0.2	ψ_1 ψ_2	0 0	ln(1.5) ln(2)
4	$\beta_0 + \psi_1 x_1 + \psi_2 x_2 + \lambda_1 z_1$	$P\Big((x_1, x_2) = \big((-\frac{1}{2}, -\frac{1}{2}), (-\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, -\frac{1}{2}), (\frac{1}{2}, \frac{1}{2})\big)\Big)$ $= (0.76, 0.19, 0.01, 0.04), (0.4, 0.1, 0.1, 0.4)$, and $(0.04, 0.01, 0.19, 0.76)$ $z_1 \sim N(0,1)$		0.1 λ_1	ψ_1 ψ_2 0.1	0 0 0.1	ln(1.5) ln(2)
5	$\beta_0 + \psi_1 x_1 + \psi_2 x_2 + \psi_3 x_3$	$P\Big((x_1, x_2) = \big((-\frac{1}{2}, -\frac{1}{2}), (-\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, -\frac{1}{2}), (\frac{1}{2}, \frac{1}{2})\big)\Big)$ $= (0.76, 0.19, 0.01, 0.04)$, $(0.4, 0.1, 0.1, 0.4)$, and $(0.04, 0.01, 0.19, 0.76)$ $x_3 \sim N(0,1)$		0.1 	ψ_1 ψ_2 ψ_3	0 0 0	ln(1.5) ln(2) 0.1

Table 1: Summary of Simulations

of X and Z, and then sample a response given the values of X and Z. Once the model is fitted to the simulated data, we calculate either the Wald, likelihood ratio or score test statistic. For each simulation we report the proportion of the 10000 tests that have a test statistic that exceeds the critical value. This procedure is repeated for each of the three types of test.

Across each of the scenarios, we can compare the simulated power for each of the calculated sample size estimates for the Wald, likelihood ratio, and score tests. We notice that the simulated power values are typically very similar, rarely deviating more than 2% between the three tests, despite the use of separate simulations to obtain each of the different power estimates for the three tests. The only exceptions to this observation are the cases where the simulated level of power was much lower than the target level of power.

In general, the likelihood ratio test approach of Shieh, 2000b gave sample sizes that best achieve the target level of power. The adjusted method of Shieh, 2005 performs well for the first three scenarios, but less well in the final two scenarios, where the direct method of the same paper performs better. Both methods based on the score test perform quite poorly in general, giving the most overpowered and the most underpowered tests in the study. The methods of Demidenko, 2007 and Lyles et. al., 2006 gave underpowered tests in scenario 1 and overpowered tests in scenario 2.

5 Conclusions

We conclude with a summary of our findings, some recommendations, and a discussion of the implementation of these algorithms in the Mathematica software package.

First, we find that the methods that most accurately and consistently achieve the desired level of power are those presented in Shieh, 2000b based on the likelihood ratio test, which is a generalisation of the method presented in Self et. al., 1992, as well as the method presented in Novikov et. al., 2010 based on the Wald test. This is the case regardless of whether the Wald, likelihood ratio or score test is used to perform the test. So while it seems sensible to determine sample size based on the test that will ultimately be used to test the hypotheses on parameters, as suggested by Demidenko, 2008, there appears to be little difference in the performance of a certain sample size determination method for the three tests in the simulations considered in this paper.

We note that the formulae of Demidenko, 2008 assume that dummy coding has been used to code the Bernoulli variable in Scenario 1. The author also assumes that $P(Y = 1|H_0) = P(Y = 1|X = 0, H_1)$. If we repeat the simulations under these assumptions with $\psi = \ln(2)$, we find that the simulated powers do not differ substantially from those in Table 2.

On the website `http://sites.google.com/site/stephenabush/`, we provide the Mathematica notebooks for the direct and adjusted methods of Shieh, 2005, the likelihood ratio test method of Shieh, 2000b, and the score test methods of Section 3.

Acknowledgements

The author thanks the referees for their constructive comments on this article, which have substantially improved the clarity of the presentation of this article.

References

Agresti A. (2003). *Categorical Data Analysis*. Hoboken NJ: Wiley-interscience.

Cox D, Hinkley D. (1979). *Theoretical Statistics*. London UK: Chapman & Hall/CRC.

Demidenko E. (2007). Sample size determination for logistic regression revisited. *Statistics in Medicine* **26**(18):3385–3397.

Demidenko E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine* **27**(1):36–46.

Hsieh F, Bloch D, Larsen M. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* **17**(14):1623–1634.

Lyles R, Lin H, Williamson J. (2006). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine* **26**(7):1632–1648.

Nelder J, Wedderburn R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society Series A-General* **135**(3):370–384.

Novikov I, Fund N, Freedman L. (2010). A modified approach to estimating sample size for simple logistic regression with one continuous covariate. *Statistics in Medicine* **29**(1):97–107.

Schoenfeld D, Borenstein M. (2005). Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation* **75**(10):771–785.

Schouten H. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine* **18**(1):87–91.

Self S, Mauritsen R. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44**(1):79–86.

Self S, Mauritsen R, Ohara J. (1992). Power calculations for likelihood ratio tests in generalized linear models. *Biometrics* **48**(1):31–39.

Shieh G. (2000). A comparison of two approaches for power and sample size calculations in logistic regression models. *Communications in Statistics-Simulation and Computation* **29**(3):763–791.

Shieh G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* **56**(4):1192–1196.

Shieh G. (2005). On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference* **128**(1):43–59.

Whittemore A. (1981). Sample size for logistic regression with small response probability. *Journal of the American Statistical Association* **76**(373):27–32.

Wilson S, Gordon I. (1986). Calculating sample sizes in the presence of confounding variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **35**(2):207–213.

	Shieh (2005) Direct		Shieh (2005) Adjusted		Shieh (2000)		Demidenko (2007)		Lyles et al (2007)		Self Mauritsen (1988)		Score Test using Expectation	
Power	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95
0.1														
Sample Size	1173	1451	1377	1677	1261	1559	1137	1405	1135	1405	1490	1842	1473	1822
Power - Wald Test	0.8749	0.9331	0.9187	0.9603	0.8978	0.9471	0.8658	0.9230	0.8583	0.9290	0.9377	0.9688	0.9322	0.9710
Power - LR Test	0.8622	0.9231	0.9122	0.9544	0.8865	0.9400	0.8570	0.9175	0.8513	0.9131	0.9290	0.9707	0.9275	0.9662
Power - Score Test	0.8776	0.9283	0.9222	0.9608	0.8997	0.9456	0.8653	0.9306	0.8670	0.9232	0.9375	0.9747	0.9359	0.9698
0.3														
Sample Size	587	726	626	789	602	744	524	648	524	648	658	814	638	788
Power - Wald Test	0.8944	0.9457	0.9119	0.9604	0.8953	0.9498	0.8551	0.9185	0.8597	0.9190	0.9218	0.9647	0.9115	0.9576
Power - LR Test	0.8872	0.9403	0.9083	0.9562	0.8974	0.9504	0.8453	0.9169	0.8519	0.9153	0.9133	0.9616	0.9139	0.9544
Power - Score Test	0.8931	0.9464	0.9128	0.9593	0.9012	0.9448	0.8571	0.9209	0.8602	0.9168	0.9232	0.9635	0.9148	0.9580
0.5														
Sample Size	583	720	561	696	565	698	471	582	471	582	560	692	536	662
Power - Wald Test	0.9074	0.9522	0.8993	0.9506	0.8938	0.9492	0.8377	0.9090	0.8340	0.9104	0.8921	0.9494	0.8836	0.9376
Power - LR Test	0.9151	0.9569	0.9030	0.9525	0.9040	0.9485	0.8415	0.9128	0.8455	0.9119	0.9041	0.9500	0.8881	0.9431
Power - Score Test	0.9158	0.9593	0.8949	0.9485	0.8933	0.9528	0.8358	0.9086	0.8472	0.9098	0.8969	0.9506	0.8846	0.9402
0.7														
Sample Size	822	1016	716	899	751	928	597	738	596	739	662	818	634	784
Power - Wald Test	0.9252	0.9673	0.8848	0.9401	0.9008	0.9488	0.8200	0.8938	0.8281	0.8999	0.8571	0.9299	0.8461	0.9138
Power - LR Test	0.9398	0.9713	0.9021	0.9490	0.9085	0.9563	0.8340	0.8979	0.8320	0.9048	0.8718	0.9272	0.8609	0.9205
Power - Score Test	0.9279	0.9680	0.8878	0.9509	0.9061	0.9541	0.8321	0.8979	0.8266	0.8998	0.8615	0.9253	0.8477	0.9177
0.9														
Sample Size	2267	2803	1797	2278	1953	2415	1477	1826	1476	1826	1497	1851	1467	1814
Power - Wald Test	0.9397	0.9775	0.8678	0.9488	0.9069	0.9596	0.8012	0.8793	0.8029	0.8809	0.8006	0.8884	0.7906	0.8809
Power - LR Test	0.9511	0.9806	0.8903	0.9496	0.9140	0.9614	0.8293	0.8956	0.8270	0.8978	0.8316	0.8954	0.8260	0.8952
Power - Score Test	0.9426	0.9760	0.8720	0.9457	0.9046	0.9550	0.8066	0.8922	0.8020	0.8866	0.8083	0.8924	0.7982	0.8859

Table 2: Power simulations for scenario 1

	Shieh (2005) Direct		Shieh (2005) Adjusted		Shieh (2000)		Demidenko (2007)		Lyles et al (2007)		Score Test using Expectation		Novikov et al (2010)	
Power	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95
-Log(2)														
Sample Size	172	213	151	189	153	190	169	208	163	201	127	158	156	191
Power - Wald Test	0.9312	0.9646	0.8938	0.9481	0.8952	0.9483	0.9236	0.9659	0.9139	0.9595	0.8331	0.9073	0.9064	0.9515
Power - LR Test	0.9310	0.9694	0.8995	0.9522	0.9076	0.9495	0.9242	0.9660	0.9228	0.9628	0.8454	0.9132	0.9130	0.9506
Power - Score Test	0.9311	0.9709	0.8947	0.9511	0.9022	0.9527	0.9193	0.9644	0.9147	0.9592	0.8389	0.9055	0.9084	0.9520
-Log(1.5)														
Sample Size	433	535	413	513	416	514	422	522	425	525	389	482	420	515
Power - Wald Test	0.9136	0.9592	0.8988	0.9490	0.9001	0.9510	0.9006	0.9518	0.9047	0.9532	0.8809	0.9374	0.8982	0.9494
Power - LR Test	0.9127	0.9572	0.8991	0.9496	0.8978	0.9495	0.9094	0.9552	0.9078	0.9531	0.8786	0.9400	0.9049	0.9501
Power - Score Test	0.9174	0.9579	0.8946	0.9516	0.9039	0.9535	0.9073	0.9550	0.9075	0.9515	0.8798	0.9360	0.9037	0.9523
Log(1.5)														
Sample Size	433	535	413	513	416	514	422	522	425	525	389	482	420	515
Power - Wald Test	0.9133	0.9603	0.8956	0.9481	0.8949	0.9483	0.9113	0.9517	0.9044	0.9533	0.8815	0.9364	0.8988	0.9482
Power - LR Test	0.9176	0.9591	0.8998	0.9559	0.9003	0.9484	0.9049	0.9526	0.9049	0.9518	0.8855	0.9385	0.9039	0.9509
Power - Score Test	0.9115	0.9597	0.8955	0.9483	0.9014	0.9533	0.8987	0.9538	0.9050	0.9551	0.8827	0.9387	0.8991	0.9554
Log(2)														
Sample Size	172	213	151	189	153	190	169	208	163	201	127	158	156	191
Power - Wald Test	0.9324	0.9663	0.8885	0.9498	0.8934	0.9457	0.9247	0.9625	0.9094	0.9591	0.8339	0.9087	0.8963	0.9523
Power - LR Test	0.9306	0.9711	0.8965	0.9569	0.9088	0.9520	0.9277	0.9676	0.9182	0.9639	0.8398	0.9087	0.9106	0.9538
Power - Score Test	0.9298	0.9684	0.8994	0.9514	0.8974	0.9497	0.9247	0.9645	0.9134	0.9616	0.8387	0.9094	0.9015	0.9521

Table 3: Power simulations for scenario 2

	Shieh (2005) Direct		Shieh (2005) Adjusted		Shieh (2000)		Self Mauritsen 1988		Score Test using Expectation	
Power	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95
(0.3,0.3,0.4)										
Sample Size	699	853	639	787	625	763	571	697	508	620
Power - Wald Test	0.9302	0.9723	0.9098	0.9588	0.9017	0.9544	0.8659	0.9265	0.8136	0.8986
Power - LR Test	0.9393	0.9751	0.9172	0.9626	0.9097	0.9577	0.8778	0.9391	0.8446	0.9048
Power - Score Test	0.9378	0.9707	0.9059	0.9566	0.8975	0.9516	0.8670	0.9374	0.8218	0.8996
(0.1,0.2,0.7)										
Sample Size	1824	2226	1467	1833	1519	1854	2003	2444	1066	1300
Power - Wald Test	0.9497	0.9797	0.8835	0.9480	0.8904	0.9502	0.9680	0.9892	0.7232	0.8289
Power - LR Test	0.9598	0.9858	0.9108	0.9587	0.9157	0.9594	0.9735	0.9917	0.7899	0.8725
Power - Score Test	0.9505	0.9813	0.8860	0.9528	0.9046	0.9538	0.9709	0.9900	0.7428	0.8415
(0.2,0.7,0.1)										
Sample Size	1156	1411	1034	1276	1013	1237	890	1086	788	962
Power - Wald Test	0.9339	0.9740	0.9034	0.9582	0.8992	0.9495	0.8509	0.9195	0.7959	0.8814
Power - LR Test	0.9484	0.9778	0.9213	0.9644	0.9135	0.9596	0.8734	0.9311	0.8334	0.9027
Power - Score Test	0.9422	0.9729	0.9097	0.9596	0.9013	0.9529	0.8637	0.9211	0.8118	0.8864
(0.7,0.1,0.2)										
Sample Size	463	565	527	635	474	578	547	667	503	614
Power - Wald Test	0.8969	0.9467	0.9285	0.9659	0.9071	0.9483	0.9401	0.9696	0.9154	0.9595
Power - LR Test	0.8855	0.9400	0.9204	0.9653	0.8976	0.9418	0.9359	0.9696	0.9073	0.9534
Power - Score Test	0.9001	0.9450	0.9298	0.9703	0.9032	0.9512	0.9413	0.9753	0.9256	0.9649

Table 4: Power simulations for scenario 3

	Shieh (2005) Direct		Shieh (2005) Adjusted		Shieh (2000)		Self Mauritsen (1988)		Score Test using Expectation	
Power	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95
(0.76,0.19,0.01,0.04)										
Sample Size	971	1185	1245	1484	1048	1279	1649	2013	1237	1510
Power - Wald Test	0.8769	0.9314	0.9425	0.9717	0.8998	0.9480	0.9847	0.9949	0.9415	0.9743
Power - LR Test	0.8542	0.9210	0.9369	0.9679	0.8871	0.9378	0.9816	0.9931	0.9332	0.9644
Power - Score Test	0.8852	0.9327	0.9463	0.9764	0.9044	0.9496	0.9819	0.9938	0.9476	0.9742
(0.4,0.1,0.1,0.4)										
Sample Size	676	824	620	763	622	759	634	773	542	622
Power - Wald Test	0.9210	0.9655	0.8943	0.9530	0.8989	0.9462	0.8999	0.9509	0.8426	0.8957
Power - LR Test	0.9244	0.9651	0.9081	0.9536	0.9069	0.9506	0.9069	0.9567	0.8556	0.9038
Power - Score Test	0.9275	0.9636	0.9029	0.9527	0.9027	0.9507	0.9114	0.9539	0.8525	0.9037
(0.04,0.01,0.19,0.76)										
Sample Size	2052	2505	1413	1798	1727	2108	1268	1548	1233	1505
Power - Wald Test	0.9356	0.9784	0.7678	0.8914	0.8712	0.9473	0.6985	0.8225	0.6770	0.8045
Power - LR Test	0.9585	0.9827	0.8494	0.9295	0.9155	0.9612	0.8059	0.8824	0.7992	0.8729
Power - Score Test	0.9438	0.9790	0.8097	0.9109	0.8873	0.9519	0.7466	0.8546	0.7297	0.8391

Table 5: Power simulations for scenario 4

	Shieh (2005) Direct		Shieh (2005) Adjusted		Shieh (2000)		Score Test using Expectation	
Power	0.9	0.95	0.9	0.95	0.9	0.95	0.9	0.95
(0.76,0.19,0.01,0.04)								
Sample Size	1022	1238	1196	1427	1094	1326	1269	1537
Power - Wald Test	0.8832	0.9329	0.9284	0.9611	0.9065	0.9503	0.9416	0.9732
Power - LR Test	0.8690	0.9247	0.9170	0.9568	0.8833	0.9392	0.9334	0.9684
Power - Score Test	0.8896	0.9429	0.9320	0.9679	0.9082	0.9514	0.9417	0.9741
(0.4,0.1,0.1,0.4)								
Sample Size	725	878	674	822	668	810	584	707
Power - Wald Test	0.9138	0.9630	0.8973	0.9487	0.8947	0.9501	0.8324	0.9073
Power - LR Test	0.9297	0.9656	0.9083	0.9583	0.9032	0.9509	0.8544	0.9178
Power - Score Test	0.9230	0.9684	0.9044	0.9565	0.9042	0.9492	0.8461	0.9196
(0.04,0.01,0.19,0.76)								
Sample Size	2013	2439	1471	1845	1726	2091	1268	1536
Power - Wald Test	0.9251	0.9733	0.7868	0.8953	0.8695	0.9386	0.6847	0.8087
Power - LR Test	0.9537	0.9810	0.8619	0.9364	0.9171	0.9645	0.7986	0.8817
Power - Score Test	0.9410	0.9757	0.8205	0.9141	0.8860	0.9476	0.7438	0.8396

Table 6: Power simulations for scenario 5