行政院國家科學委員會專題研究計畫 成果報告

基因體資料中的特徵選取(第 2 年)
研究成果報告(完整版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 96-2118-M-004-004-MY2
執 行 期 間 ：97 年 08 月 01 日至 98 年 07 月 31 日
執 行 單 位 ：國立政治大學統計學系

計 畫 主 持 人 ：薛慧敏

處 理 方 式 ：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中 華 民 國 98 年 10 月 30 日

Optimal sampling in retrospective logistic regression via two-stage method

Chih-Yi Chien [1], **Yuan-chin Ivan Chang** [1,2], and **Huey-Miin Hsueh**[*1]

[1] Department of Statistics, National Cheng-Chi University, Taipei, Taiwan (R.O.C.)
[2] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan (R.O.C.)

Received 8 October 2009

Summary

Case-control sampling is popular in epidemiological research because of its cost and time saving. With limited knowledge on the covariance matrix of the point estimator a priori, there exists no fixed sample size analysis and sequential methods are applied. We consider a two-stage sequential analysis of a retrospective logistic regression model. An interim analaysis is conducted midway through the experiment for estimation of the optimal sample fraction and the required sample size to achieve a predetermined volume of a joint confidence set. Additional required observations are collected in the second stage according to the estimated optimal sample fraction. At the end of the experiment, data from these two stages are combined and analyzed for statistical inference. A similar technique as Chen(2000) to deal with the unidentifiable intercept is employed in our procedure. Simulation studies are conducted to justify the proposed two-stage procedure and an example is presented for illustration.

Key words: Joint confidence region, Optimal sample fraction of case to control, Retrospective logistic regression model, Sample size determination,Two-stage sequential procedure.

1 Introduction

The case-control sampling is a sampling scheme which draws samples from the disease population (the cases) and the non-disease population (the controls) independently and then investigates whether the selected subjects have been exposed to the covariates of interest. This kind of retrospective sampling scheme usually requires a smaller sample size comparing with that of a prospective study (Farewell (1979)), and therefore is popular in epidemiological research because of its cost and time saving (Anderson (1972)). Here we focus on the statistical inference of a retrospective logistic regression model; in particular, the required sample size to achieve the predetermined precision of a confidence region. On the other hand, it is shown that the sample fraction of control to case affects the efficiency of the point estimator. An optimal sample fraction is determined as well as the sample size.

In logistic regression, determination of sample size naturally involves the covariance matrix of the point estimators of the unknown regression parameters, which is also a function of unknown parameters to be estimated. Several sample size formula of logistic regression analysis have been proposed in literature. Demidenko (2007) gave a thorough review. For small response probability, Whittemore (1981) proposed a simple approximated sample size formulae, in which the information matrix is approximated by the derivatives of the moment generating function of the covariates. Subsequently, Hsieh (1989) presented the sample size tables from Whittemore's formula. Shieh (2001) and Demidenko (2007) developed two variants of Whittemore's sample size formulae. However, all these methods rely on a specification of the covariance matrix, which involves the distribution of the covariates and the true parameter values. With limited knowledge a priori, the specification is often challenging. Investigators are usually suggested to

[*] Corresponding author: e-mail: hsueh@nccu.edu.tw, Phone: +886 2 2939 3091 Ext. 81138, Fax: +886 2 2939 8024

take the information from prior similar trials, which are difficult to find in practice. Friede and Kieser (2004) showed from some real examples where the results can considerably differ between studies.

Lacking a reliable guess of the covariance matrix, there exists no fixed sample size and sequential methods can be applied. Chang and Martinsek (1992) proposed a sequential logistic regression analysis, and the correspondent random sample size has been shown to stop and converge to the fixed sample size almost surely under some regularity conditions. Grambsch (1989) showed that the sequential maximum likelihood estimator has asymptotic normality and gave applications to retrospective logistic regression. Chen (2000) proposed an optimal case-control sequential design, in which an optimal sampling fraction of case to control is estimated and a fully sequential sampling scheme is proposed. Ghosh and Mukherjee (2006) proposed a Bayes stopping rule for an optimal sequential case-control design. However, in practice, the fully sequential statistical approach is not feasible or desirable (Betensky and Tierney (1997)). Multi-stage sequential analysis is an useful alternative. In this study, we consider a two-stage sequential analysis, in which an interim analaysis is conducted midway through the experiment. The optimal sample fraction and the required sample size are estimated based on the initial sample data. The additional required observations are collected in the second stage according to the estimated optimal fraction. At the end of the experiment, data from the two stages are combined and analyzed for statistical inference.

The two-stage sequential procedure is similar to the so-called internal pilot study introduced by Wittes and Brittain (1990). Both approaches estimate the required sample size based on an initial data set. However, the internal pilot study use the initial sample to re-calculate the sample size, but not for any interim statistical analysis. This method has become more and more popular recently due to its flexibility (Proschan (2005), Proschan (2009), Friede and Kieser (2004), Kieser and Friede (2000), Coffey and Muller (1999), Betensky and Tierney (1997), Wittes *et al.* (1999), Gould and Shih (1992)). The two-stage method can be dated back to Stein (1945), where the dependency between the pilot data set and the recalculated sample size is successfully taken into account. Compared with the internal pilot study, the two-stage sequential procedure, which allows an early stopping of the study and thus usually requires smaller total sample size, is superior in terms of resource saving.

It's known that, under both prospective and retrospective sampling scheme, the regression parameters of a logistic regression remains the same except for the intercept term, which is confounded with the population odds of the disease and hence is not estimable from a case-control dataset. Luckily, as the research goal is to identify the association between covariates and disease, the parameters of main interest are the slopes. A brief review on a retrospective logistic model and the maximal likelihood inference is given in Section 2. It can be seen that not only the sample size but also the sample fraction of control to case affect the statistical inference. Section 3 introduces a two-stage sequential procedure of a retrospective logistic regression analysis for estimation of the optimal sample fraction and the required sample size. The determination of the optimal sample fraction uses a similar technique as that of Chen (2000) to deal with the unidentifiable intercept. Simulation studies are conducted and their empirical results are are reported in Section 4 to justify the proposed two-stage procedure. In Section 5, we apply our method to real data to illustrate the proposed procedure. Concluding remarks are given in Section 6.

2 Retrospective Logistic Regression Models

Suppose $Y = 1(0)$ denotes the diseased (non-diseased) case and let Z be a corresponding p-dimensional vector of covariates. Therefore a prospective logistic regression model for describing the relation of disease status and covariates of interest can be specified as

$$\text{logit } Pr(Y = 1|z) = \theta_0 + \beta_0^T z, \tag{1}$$

where the intercept θ_0 and the slopes $\beta_0 = (\beta_{01}, \cdots, \beta_{0p})^T \in R^p$ denote the unknown true regression coefficients to be estimated. In a case-control study, for a given sample of size n, $\{(z_i, y_i), i = 1, \ldots, n\}$,

the likelihood function of the covariates given y can be written as

$$L_n = \prod_{i=1}^{n} Pr(z_i|y_i) = \left\{\prod_{i=1}^{n} Pr(y_i|z_i)\right\}\left\{\prod_{i=1}^{n} \frac{Pr(z_i)}{Pr(y_i)}\right\} \equiv L_{1n}L_{2n},$$

where for $i = 1, \cdots, n$, the three probabilities satisfy

$$\int Pr(y_i|z_i)Pr(z_i)dz = Pr(y_i). \tag{2}$$

The first component L_{1n} directly involves the parameters $\beta = (\beta_1, \cdots, \beta_p)^T$ of main interest, and the second component L_{2n}, which consists of the marginal distributions of Z and Y, is closely related to the regression coefficients by the constraints (2).

Prentice and Pyke (1979) proposed to estimate L_{2n} by the empirical density function of Z and the sampling proportions of control or case, denoted by $\hat{Pr}(z_i)$ and $\hat{Pr}(y_i)$. Subsequently, the odds between the diseased and non-diseased group given z becomes

$$\log \frac{\hat{Pr}(Y=1|z)}{\hat{Pr}(Y=0|z)} = \log \frac{Pr(z|Y=1)\hat{Pr}(Y=1)}{Pr(z|Y=0)\hat{Pr}(Y=0)} = \theta_0(\hat{\rho}) + \beta_0^T z.$$

Hence, the intercept becomes

$$\theta_0(\hat{\rho}) = \theta_0 + \log(\rho_0) - \log(\hat{\rho}) \tag{3}$$

where $\rho_0 = Pr(Y=0)/Pr(Y=1)$ is the population odds of non-disease and $\hat{\rho}$ is its sample version counterpart. The likelihood function becomes

$$\hat{L}_n \propto \prod_{i=1}^{n} \hat{Pr}(y_i|z_i) \equiv \hat{L}_{1n},$$

and it should satisfy that

$$\int \hat{Pr}(y_i|z_i)\hat{Pr}(z_i)dz = \hat{Pr}(y_i), \quad i = 1, \cdots, n. \tag{4}$$

Note that the original prospective logistic model is distorted by these estimates and the intercept of the current prospective model is now a function of θ_0, ρ_0 and $\hat{\rho}$. Unless $\hat{\rho}$ is equal to ρ_0 or ρ_0 is known, which is not possible in practice, one cannot draw statistical conclusions on θ_0 from the inference of $\theta_0(\hat{\rho})$.

By taking the first derivative of the logarithm of $\hat{L}_{1n}$, we have the score function of $\eta(\hat{\rho}) = (\theta(\hat{\rho}), \beta^T)^T$ as

$$S_n(\eta) = \frac{\partial \log \hat{L}_{1n}}{\partial \eta(\hat{\rho})} = \sum_{i=1}^{n} x_i \left\{y_i - \frac{\exp(\eta(\hat{\rho})^T x_i)}{1 + \exp(\eta(\hat{\rho})^T x_i)}\right\}, \tag{5}$$

where $x_i = (1, z_i^T)^T$. Setting the score to zero and solving this score equation by the Newton-Raphson method, we obtain the MLEs of the regression parameters, say $\hat{\eta}_n(\hat{\rho}) = (\hat{\theta}_n(\hat{\rho}), \hat{\beta}_n^T)^T$, where $\hat{\beta}_n = (\hat{\beta}_{1n}, \cdots, \hat{\beta}_{pn})^T$. It is shown that the constraints (4) are satisfied by plugging in $\hat{\eta}_n(\hat{\rho})$ into the first element of the score function (5). The score function coincides with the score function from a prospective study. Prentice and Pyke (1979) further showed the strong consistency and the asymptotical normality of $\hat{\beta}_n$. As $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \to_{\mathcal{L}} N(0, e(\hat{\rho})), \tag{6}$$

where $e(\hat{\rho})$ is the corresponding matrix partition in the inverse of the information matrix, $e(\hat{\rho}) = (\Sigma(\hat{\rho})^{-1})_{22}$, and

$$\Sigma(\hat{\rho}) = E\left\{ \left(XX^T\right) \frac{\exp(\eta_0(\hat{\rho})^T X)}{\{1+\exp(\eta_0(\hat{\rho})^T X)\}^2} \right\}. \tag{7}$$

Here $\eta_0(\hat{\rho}) = (\theta_0(\hat{\rho}), \beta_0^T)^T$ is the true value of $\eta(\hat{\rho})$ and $X = (1, Z^T)^T$. Note that the expectation of $\Sigma(\hat{\rho})$ is taken under $\hat{\rho}$. Wang and Wang (1995) proposed a consistent estimate of the information matrix $\Sigma(\hat{\rho})$

$$\hat{\Sigma}(\hat{\rho}) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \left(x_i x_i^T\right) \frac{\exp(\hat{\eta}_n(\hat{\rho})^T x_i)}{\{1+\exp(\hat{\eta}_n(\hat{\rho})^T x_i)\}^2} \right\}, \tag{8}$$

and a consistent estimate of the covariance matrix is $\hat{e}(\hat{\rho}) = (\hat{\Sigma}(\hat{\rho})^{-1})_{22}$. The estimates are of the same forms as those derived in a prospective study.

In summary, for slope parameters, the estimation and subsequent inference can be conducted in the same way under both prospective and case-control sampling schemes. One can apply a prospective logistic regression data analysis to case-control data sets. For the intercept, because the coefficient in (1) is distorted and confounded with the population prevalence rate of the disease, it is no longer identifiable and estimable. In the following, we propose a two-stage design for sample size determination for case-control logistic regression models and to construct a fixed width confidence set of β_0.

3 Two-stage Procedure

In planning a case-control study with a logistic model, the sample size determination involves the covariance matrix of the MLE of regression coefficients. Additionally, it can be seen that the sample fraction of control to case $\hat{\rho}$ affects the covariance matrix and hence the efficiency of the estimators. Without reliable information of the covariance matrix from prior studies, the sample size estimation is difficult and tends to become arbitrary. Thus a two-stage procedure with an internal pilot study is applied to solve the problem. At the first stage, a pilot dataset is collected to determine both the optimal sample allocation and the necessary sample size. In the second stage of study, additional subjects are drawn by using the suggested sample allocation until the desired sample size is achieved. Finally, a confidence region for β_0 is constructed based on all observations by the end of study.

From (6), as $n \to \infty$,

$$n(\hat{\beta}_n - \beta_0)^T e(\hat{\rho})^{-1}(\hat{\beta}_n - \beta_0) \to_{\mathcal{L}} \chi^2(p),$$

where $\chi^2(p)$ is a chi-square distribution with p degrees of freedom. Let $\chi^2_{p,\alpha}$ be the $100(1-\alpha)\%^{th}$ percentile of $\chi^2(p)$, i.e. $P(\chi^2(p) \le \chi^2_{p,\alpha}) = 1-\alpha$. It follows that the region

$$R_{E,n} = \{\beta \in R^p : n(\hat{\beta}_n - \beta)^T\, e(\hat{\rho})^{-1}\, (\hat{\beta}_n - \beta) \le \chi^2_{p,\alpha}\}.$$

is a $100(1-\alpha)\%$ confidence ellipsoid for β_0 approximately. On the other hand, the $100(1-\alpha)\%$ rectangular simultaneous confidence region is given by

$$R_{R,n} = \{\hat{\beta}_{in} - \sqrt{\chi^2_{p,\alpha}\, e(\hat{\rho})_{ii}/n} \le \beta_i \le \hat{\beta}_{in} + \sqrt{\chi^2_{p,\alpha}\, e(\hat{\rho})_{ii}/n},\ \ i = 1, \cdots, p\},$$

where $e(\hat{\rho})_{ii}$ is the i-th diagonal entry of $e(\hat{\rho})$ (Johnson and Wichern(1992, P193)).

When considering a confidence ellipsoid, the width is defined as the length of the maximal axis of the ellipsoid; while the width of a rectangular confidence region is the maximal length among each individual

confidence interval. Let the desired width equal $2d$ for some pre-specified $d > 0$. The ellipsoid defined by $R_{E,n}$ is centered at $\hat{\beta}_n$ and the length of its maximal axis is equal to

$$2\sqrt{\chi^2_{p,\alpha}/\left\{n\lambda_{\min}(e(\hat{\rho})^{-1})\right\}} = 2\sqrt{\chi^2_{p,\alpha}\lambda_{\max}(e(\hat{\rho}))/n},$$

where $\lambda_{\min}(e(\hat{\rho})^{-1}), \lambda_{\max}(e(\hat{\rho}))$ represent the minimal and maximal eigenvalue of $e(\hat{\rho})^{-1}$ and $e(\hat{\rho})$, respectively. On the other hand, the largest diagonal entry of e, denoted by $(\max_{i=1,\cdots,p} e(\hat{\rho})_{ii})$, is proportional to the maximal length of $R_{R,n}$. Subsequently, it can be easily derived that the necessary sample size for a $100(1-\alpha)\%$ joint confidence ellipsoid or simultaneous confidence hyperrectangle for β_0 with desired length $2d$ is, respectively,

$$n_{E,opt} = \left[\frac{\lambda_{max}(e(\hat{\rho}))\ \chi^2_{p,\alpha}}{d^2}\right],\ \ n_{R,opt} = \left[\frac{(\max_{i=1,\cdots,p} e(\hat{\rho})_{ii})\ \chi^2_{p,\alpha}}{d^2}\right], \tag{9}$$

provided $e(\hat{\rho})$ is known. Notation $[x]$ denotes the smallest integer that is not less than x. In real applications, $e(\hat{\rho})$ is often unknown. Therefore there exists no fixed sample size formulae. Besides, it is clear that the sample size depends on the sampling fraction $\hat{\rho}$. Hence, for a given desired width as defined above, the optimal sampling fraction of case (or control) that minimizes the required sample size, and the resultant sample size can be determined.

Consider an internal pilot study of size $n^{(0)}$ in the first stage of the study with a given initial sampling fraction $\rho^{(0)}$, the numbers of case and control are given as $n_1^{(0)} = n^{(0)}\rho^{(0)}/\{1+\rho^{(0)}\}, n_0^{(0)} = n^{(0)}/\{1+\rho^{(0)}\}$. The MLEs $\hat{\eta}_{n^{(0)}}(\rho^{(0)}) = (\hat{\theta}_{n^{(0)}}(\rho^{(0)}), \hat{\beta}^T_{n^{(0)}})^T$ from the first stage data are consistent estimators for $\eta_0(\rho^{(0)}) = (\theta_0(\rho^{(0)}), \beta_0^T)^T$, where $\theta_0(\rho^{(0)}) = \theta_0 + \log(\rho_0) - \log(\rho^{(0)})$ by (3). Hence, $\hat{\theta}_{n^{(0)}}(\rho^{(0)}) + \log(\rho^{(0)})$ is a consistent estimator of $\theta_0 + \log(\rho_0)$. Moreover, for any $\rho > 0$, let $\hat{\theta}_{n^{(0)}}(\rho) = \hat{\theta}_{n^{(0)}}(\rho^{(0)}) + \log(\rho^{(0)}/\rho)$. Then $\hat{\theta}_{n^{(0)}}(\rho)$ is a consistent estimator of the intercept $\theta_0(\rho) = \theta_0 + \log(\rho_0/\rho)$. Subsequently, based on the first stage data, the information matrix $\Sigma(\rho)$ can be estimated by

$$\hat{\Sigma}(\rho) = \frac{1}{n^{(0)}}\sum_{i=1}^{n^{(0)}} \left(x_i x_i^T\right) \frac{\exp(\hat{\eta}_{n^{(0)}}(\rho)^T x_i)}{\{1+\exp(\hat{\eta}_{n^{(0)}}(\rho)^T x_i)\}^2},$$

where $\hat{\eta}_{n^{(0)}}(\rho) = (\hat{\theta}_{n^{(0)}}(\rho), \hat{\beta}^T_{n^{(0)}})^T$. Note that by using the relationship between the two intercept terms at different sampling fractions, we successfully avoid the estimation of the original intercept and the true population odds in the sample size determination. Taking the correspondent partition of the inverse of $\hat{\Sigma}(\rho)$, we obtain an estimate of $e(\rho)$; that is, $\hat{e}(\rho) = \{\hat{\Sigma}(\rho)^{-1}\}_{22}$.

It follows that an optimal sampling fraction can be found by minimizing the required sample size at fixed width of the confidence region. From (9), the sample size is proportional to the maximal eigenvalue of $\hat{e}(\rho)$ for a confidence ellipsoid and to the maximal diagonal entry of $\hat{e}(\rho)$ for a confidence hyperrectangle. Consequently, the two estimated optimal sample fractions can be obtained as

$$\hat{\rho}_{E,opt} = \arg\min_{\rho>0}\lambda_{max}(\hat{e}(\rho)),\ \ \hat{\rho}_{R,opt} = \arg\min_{\rho>0}\left(\max_{i=1,\cdots,p}\hat{e}(\rho)_{ii}\right). \tag{10}$$

A numerical method for solving the optimal sampling fraction is used. Once the optimal sample fraction is determined, based on (9), we can calculate the required sample size, which is random and usually called a stopping time in sequential analysis:

$$\tau_{E,d} = \left[\frac{\lambda_{max}(\hat{e}(\hat{\rho}_{E,opt}))\ \chi^2_{p,\alpha}}{d^2}\right],\ \ \tau_{R,d} = \left[\frac{(\max_{i=1,\cdots,p}\hat{e}(\hat{\rho}_{R,opt}))\ \chi^2_{p,\alpha}}{d^2}\right]. \tag{11}$$

By the end of the first stage, we already have $n^{(0)}$ subjects with $n_1^{(0)}$ cases and $n_0^{(0)}$ controls and $\rho^{(0)} = n_0^{(0)}/n_1^{(0)}$. Based on this information, an allocation rule on collecting new cases and controls in the second stage is proposed below. Following (11), one needs to take $\tau_{l,d} - n^{(0)}$, more subjects with $n_1^{(1)}$ cases and $n_0^{(1)}$ controls, where $l = E$ or R, and $n_1^{(1)} = \tau_{l,d}/(1+\hat{\rho}_{l,opt}) - n_1^{(0)}$, $n_0^{(1)} = \tau_{l,d}\hat{\rho}_{l,opt}/(1+\hat{\rho}_{l,opt}) - n_0^{(0)}$. If $n_1^{(1)} \leq 0$ then we take $\tau_{l,d} - n^{(0)}$ controls without any cases, and vice versa. After drawing new samples, the MLEs are re-calculated and the confidence set can be established for final statistical inference. This procedure is a unrestricted design because there is no pre-planned sample size, hence the estimated sample size does not have a lower bound.

The properties of the MLE of logistic regression under sequential sampling have been studied by many authors; in particular, Grambsch (1989) and Chen (2000) considered this problem under case-control design. In addition, Chang (2001) proved the asymptotic properties of MLE of logistic regression under adaptive design. Thus, it follows that the asymptotic property of the estimate, such as its asymptotic normality, under two-stage case-control sampling remain, and the proposed confidence region is then based on such a property and the total sample is used. In the next section, the validity of the proposed procedure is justified through simulation studies.

4 Empirical Studies

The simulation presented below is to evaluate the proposed two-stage sample size estimating procedure for a fixed width confidence region. The estimated optimal sample fractions, the random sample sizes and the resultant confidence sets are investigated. The exposures of cases and controls are generated separately from their own group. Consider one single exposure variable, Z. In the control group, Z_i are independently generated from a standard normal distribution; while in the case group, Z_i are independently generated from a normal distribution with mean μ and variance σ^2. Then the logarithm of the density of Z between a case and a control is

$$\begin{aligned} \log \frac{Pr(z|Y=1)}{Pr(z|Y=0)} &= -\frac{\mu^2}{2\sigma^2} - \log\sigma + \frac{\mu}{\sigma^2}z + (\frac{-1}{2\sigma^2} + \frac{1}{2})z^2 \\ &= \theta_0^* + \beta_{10}z + \beta_{20}z^2, \end{aligned}$$

which corresponds to a prospective logistic regression model of two predictors, Z and Z^2. The true values of the regression coefficients are $\theta_0^* = -\mu^2/2\sigma^2 - \log\sigma$, $\beta_{10} = \mu/\sigma^2$ and $\beta_{20} = -1/2\sigma^2 + 1/2$. It's seen that a valid β_{20} should be greater than 0.5 in this model. Here we consider $\beta_{10} = 0, 0.5, 1, 1.5, 2$ and β_{20} from -2, -1.5, -1, -0.5, 0. It's known that the distorted intercept is the sum of the original intercept and the population fraction of control to the case. Since the two parameters are not of our primary research interest, there is no need to distinguish the two components. The sample size is determined to construct a 95% confidence region of (β_{10}, β_{20}) with width $2d = 1$. 500 simulations are conducted for the empirical results.

Consider a balance study for the initial sample, $n_0^{(0)} = n_1^{(0)} = 100$, and $\rho^{(0)} = 1$. Denote the sample proportion of cases as $\pi = Pr(Y = 1)$ and $\pi = 1/(1+\rho)$. In the following, the results are expressed by π instead of ρ. By the end of the first stage, the optimal proportion of cases and the necessary sample size are both estimated. A 95% confidence ellipsoid or hyperrectangle of β_0 will be obtained after sufficient subjects are additionally drawn. The empirical mean and standard error of the estimates from 500 repetitions are reported in Table 1 and Table 2. Meanwhile, the empirical coverage probability of the confidence region, and the mean and standard deviation of the half width of the region are also calculated and presented in Table 3.

The true optimal sampling fraction and the optimal sample size, denoted by $\pi_{l,opt}, n_{l,opt}, l = E, R$, are found based on a moment estimate of e by using 1,000,000 replicates. Figure 1 gives an example for $\beta_{10} = 0.5, \beta_{20} = -1.0$(i.e.$\mu = 0.17, \sigma = 0.58$). The upper panel presents the plots of $\lambda_{\max}(e)$ and the

required sample size, while the lower panel gives the plots of $(\max_i e_{ii})$ and the required sample size. It's seen that in this case the maximal eigenvalue and the maximal diagonal entry of e, which are proportional to the width of the confidence region, becomes large as π approaches to 0 or 1. It indicates that an extremely unbalance study leads to an inefficient statistical inference in terms of requiring a large sample size or obtaining a wide confidence region. The minimum occurs at $\pi_{E,opt} = 0.63$ and $\pi_{R,opt} = 0.63$, and the correspondent optimal sample size is $n_{E,opt} = 380$ and $n_{R,opt} = 303$ for an elliptical and a rectangular confidence region, respectively. The optimal study is close to the 1:1 match design. In fact, when one considers a 1:1 match design, the necessary optimal sample size is 402 for an elliptical confidence region and 322 for a rectangular confidence region. Also present are the required sample sizes for 1:1, 1:3, 1:5 and 1:10 match studies in the figures. We can see that a 1:10 match study needs nearly four to five times the sample size when compared to the study under the optimal sampling.

Table 1 and Table 2 give the numerical results of the estimated optimal sampling fraction and the estimated sample size, obtained from the initial sample, of a confidence ellipsoid and of a confidence hyperrectangle. From the two tables, we find that the true optimal sample fraction is insensitive to the true slopes and the values all are around 60%, which is close to the 1:1 match design. In this case, when one searches for an effortless and quick solution, the 1:1 match is a good choice and its performance is near that of the optimal design. The true optimal sample size increases as the absolute value(s) of either or both slopes becomes large. Because under such circumstances, the success probability of the logistic regression model moves toward 0 or 1, more subjects are necessary to maintain the efficiency of the statistical inference. The estimated optimal sample fraction is close to the true value on the average, with moderate variation. Overall the estimated sample size overestimates the fixed optimal sample size on the average and tends to draw a conservative conclusion. The standard error of the estimated sample size increases as the magnitude of the sample size. In general, the required sample size of an ellipsoid is more than that of a hyperrectangle due to the use of a stricter criterion.

Table 3 reports the empirical coverage probability, and the empirical mean and standard deviation of the width of the two 95% confidence regions. The confidence ellipsoid is likely to have a coverage deficiency. On the other hand, because the confidence hyperrectangle is developed in a conservative way, the coverage probabilities are all above the nominal level 95%. Further, except at $\beta_{10} = 2, \beta_{20} = 0$, all the mean half widths of the confidence region are well controlled and below $d = 0.5$.

Another simulation is conducted to study the effect of the initial sample size $n^{(0)}$. Consider $\beta_{10} = 0.5, \beta_{20} = -1.0$. By Table 1, Table 2 and Figure 1, the optimal sampling fractions are $\pi_{E,opt} = 0.63, \pi_{R,opt} = 0.63$, and the fixed sample sizes are $n_{E,opt} = 380, n_{R,opt} = 303$, for a confidence ellipsoid and a confidence hyperrectangle, respectively. From Table 4, as $n^{(0)}$ increases, on the average the estimated optimal sample fraction approaches the true value, the random sample size decreases and approaches the fixed sample size. Further, while the coverage probability remains at the nominal level, the half width of the confidence region gets close to the pre-determined $d = 0.5$. The variation of all estimators decreases as expected. When the pilot study is small, it leads to oversampling and hence the loss of cost. However, the coverage probability of the confidence region can still be achieved.

5 Application: The Data

We apply the proposed two-stage procedure to a real example from the Framingham data (Carroll *et al.* (2006)) for illustration. The data set includes 1615 men, among them 128 had a coronary heart disease (CHD) and 1487 were normal. Four variables, age (AGE), smoke (SMOKE), log(serum cholesterol at exam 3) (CHOL) and log(average of two systolic blood pressure at exam 3 -50) (LSBP), are included in a logistic regression model.

From (10) and (11), the optimal fraction of control to case and the necessary sample size are determined for a rectangular 95% confidence region of the four slopes $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)^T$ with the maximal axis not exceeding $2d$. Here $d = 1, 1.5, 2, 2.5$. The first 50 subjects of each group were selected

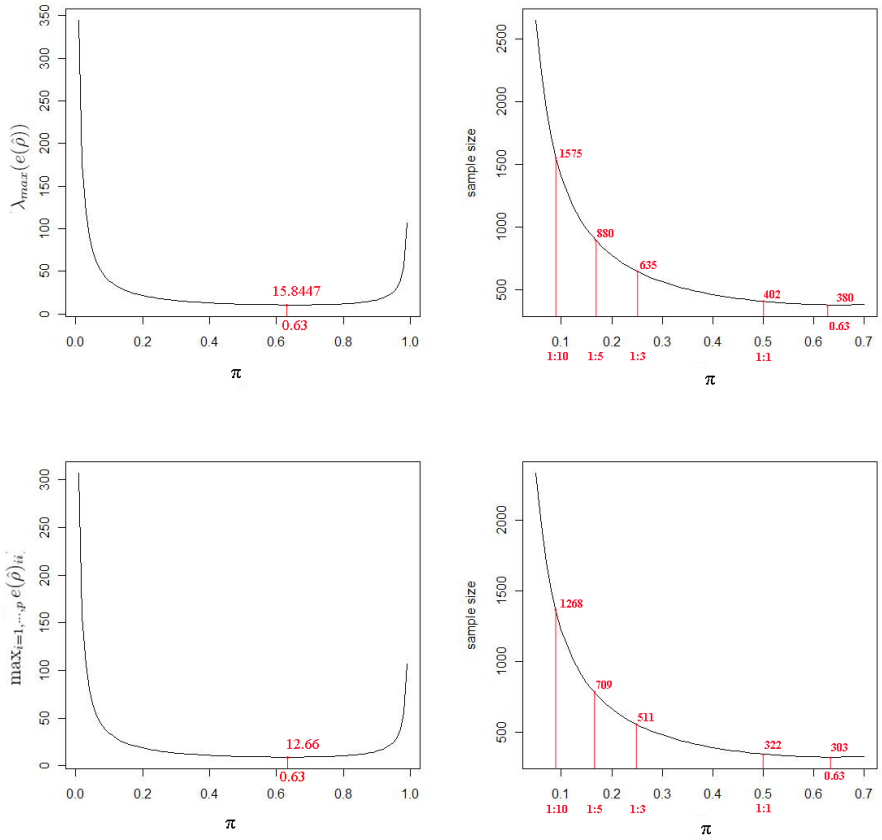

Fig. 1 The upper panel shows the plots of $\lambda_{\max}(e(\rho))$(left), and the required sample size(right) versus π for a 95% confidence ellipsoid. The lower panel shows the plots of $\max_{i=1,\cdots,p} e(\rho)$(left), and the required sample size(right) versus π for a 95% confidence hyperrectangle. The true slopes are $\beta_{01} = 0.5$, $\beta_{01} = -1.0$ and the pre-specified width $2d = 1$.

for the internal pilot study. From the data of the first stage, the estimated optimal sampling fraction is $\hat{\pi}_{R,opt} = 0.37$, $\hat{\rho}_{R,opt} = 1.7$. Table 5 presents the estimated required sample size by using different sample allocations. Specifically, when $d = 2.5$, the required total sample size is 277, which includes 103 cases and 174 controls. Consequently, additional 53 cases and 124 controls are further drawn. The data analysis, which include the MLE, the standard error and the individual confidence interval, are given in Table 6. The maximal half length of the individual interval occurs at the slope correspondent to the variable CHOL and is 2.6015, slightly greater than $d = 2.5$.

6 Concluding Remarks

We propose a two-stage procedure for developing a fixed width confidence region. In the first stage, a small-scale internal pilot study is conducted for determination of required sample size and the optimal sampling fraction. Additional subjects are drawn in the second stage. A confidence region is constructed by utilizing the data from both stages. Two types of joint confidence regions are considered: confidence ellipsoid and confidence hyperrectangle. Although the confidence ellipsoid is guaranteed to more precisely

Table 1 The empirical mean, standard deviation of the estimated optimal sample fraction and the estimated sample size for a 95% confidence ellipsoid with maximal axis not exceeding 1.

			$\hat{\pi}_{E,opt}$			$\tau_{E,d}$	
β_{10}	β_{20}	$\pi_{E,opt}$	mean	s.e.	$n_{E,opt}$	mean	s.e.
0.0	0.0	0.50	0.501	0.028	95	106	13.0
	-0.5	0.56	0.577	0.031	160	190	43.4
	-1.0	0.64	0.618	0.027	302	363	110.0
	-1.5	0.64	0.628	0.024	539	648	248.3
	-2.0	0.64	0.633	0.027	856	1000	362.2
0.5	0.0	0.58	0.562	0.043	130	146	37.8
	-0.5	0.61	0.607	0.028	229	251	72.7
	-1.0	0.63	0.625	0.026	380	423	136.3
	-1.5	0.64	0.632	0.026	608	670	223.2
	-2.0	0.64	0.634	0.025	914	1046	359.3
1.0	0.0	0.65	0.635	0.035	252	282	100.8
	-0.5	0.63	0.629	0.026	354	399	127.2
	-1.0	0.64	0.632	0.024	523	585	193.1
	-1.5	0.64	0.634	0.024	759	852	295.4
	-2.0	0.64	0.633	0.024	1066	1230	480.3
1.5	0.0	0.69	0.675	0.044	591	664	309.4
	-0.5	0.64	0.637	0.028	560	628	230.4
	-1.0	0.64	0.634	0.025	730	825	276.5
	-1.5	0.64	0.635	0.026	980	1099	405.0
	-2.0	0.64	0.632	0.025	1293	1475	506.5
2.0	0.0	0.70	0.680	0.069	1517	1675	975.0
	-0.5	0.63	0.633	0.034	905	1040	443.9
	-1.0	0.64	0.632	0.025	1010	1148	397.7
	-1.5	0.64	0.634	0.024	1257	1419	478.7
	-2.0	0.64	0.636	0.026	1591	1781	681.5

achieve the confidence level, its expression is complicated and the application is difficult when $p \geq 4$. On the other hand, the rectangular confidence region, which provides confidence intervals of individual parameters, is popular.

In this study, the sampling cost and the budget limit are not considered. We assume that the recruiting expense of a case and that of a control are the same. When the sampling costs are not constant, it should be adequately included in the criterion. Here the matching is taken with respect to the disease status solely.

Table 2 The empirical mean, standard deviation of the estimated optimal sample fraction and the estimated sample size for a 95% confidence hyperrectangle with maximal marginal width not exceeding 1.

			$\hat{\pi}_{R,opt}$			$\tau_{R,d}$	
β_{10}	β_{20}	$\pi_{R,opt}$	mean	s.e.	$n_{R,opt}$	mean	s.e.
0.0	0.0	0.50	0.501	0.025	95	104	12.2
	-0.5	0.56	0.572	0.027	160	177	41.3
	-1.0	0.64	0.622	0.029	299	338	110.3
	-1.5	0.64	0.635	0.025	541	611	217.2
	-2.0	0.64	0.635	0.026	851	960	338.7
0.5	0.0	0.55	0.542	0.036	122	133	27.4
	-0.5	0.58	0.581	0.022	190	208	48.2
	-1.0	0.63	0.615	0.027	303	359	110.7
	-1.5	0.64	0.634	0.026	541	620	201.1
	-2.0	0.64	0.633	0.025	857	972	345.6
1.0	0.0	0.64	0.620	0.036	218	241	81.4
	-0.5	0.62	0.614	0.024	279	305	92.2
	-1.0	0.61	0.608	0.022	361	411	126.9
	-1.5	0.64	0.625	0.027	549	631	208.6
	-2.0	0.64	0.634	0.024	865	985	325.5
1.5	0.0	0.70	0.674	0.040	494	569	312.9
	-0.5	0.64	0.636	0.024	447	504	191.1
	-1.0	0.63	0.627	0.024	521	593	185.3
	-1.5	0.62	0.619	0.023	612	708	215.4
	-2.0	0.64	0.625	0.028	874	991	343.4
2.0	0.0	0.72	0.696	0.061	1292	1388	816.8
	-0.5	0.64	0.647	0.035	744	824	301.1
	-1.0	0.64	0.636	0.024	764	863	301.0
	-1.5	0.63	0.630	0.025	845	924	319.4
	-2.0	0.62	0.622	0.026	946	1122	401.2

Practically, the matching may be taken with respect to other important confounding covariates. However, the problem becomes more complicated and is beyond the scope of this study.

References

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

Table 3 The empirical coverage probability, and the mean and standard deviation of the half width of the maximal axis of the constructed confidence ellipsoid and confidence hyperrectangle with maximal marginal width of β.

		Ellipsoid			Hyperrectangle		
			Half width			Half width	
β_{10}	β_{20}	Coverage rate	mean	s.e.	Coverage rate	mean	s.e.
0.0	0.0	0.986	0.50	0.01	1.000	0.50	0.01
	-0.5	0.972	0.49	0.04	0.992	0.49	0.04
	-1.0	0.964	0.48	0.06	0.992	0.49	0.06
	-1.5	0.954	0.48	0.07	0.988	0.49	0.07
	-2.0	0.950	0.48	0.07	0.988	0.50	0.08
0.5	0.0	0.974	0.49	0.03	0.996	0.50	0.02
	-0.5	0.940	0.49	0.05	0.990	0.49	0.04
	-1.0	0.954	0.49	0.06	0.996	0.49	0.06
	-1.5	0.934	0.49	0.07	0.992	0.50	0.07
	-2.0	0.956	0.49	0.07	0.986	0.49	0.08
1.0	0.0	0.954	0.48	0.06	0.992	0.49	0.06
	-0.5	0.938	0.48	0.06	0.992	0.49	0.05
	-1.0	0.952	0.49	0.07	0.992	0.49	0.06
	-1.5	0.952	0.49	0.07	0.986	0.48	0.07
	-2.0	0.950	0.49	0.08	0.990	0.49	0.08
1.5	0.0	0.930	0.49	0.09	0.984	0.49	0.10
	-0.5	0.952	0.49	0.08	0.980	0.49	0.07
	-1.0	0.948	0.49	0.07	0.978	0.49	0.07
	-1.5	0.938	0.50	0.08	0.982	0.49	0.07
	-2.0	0.960	0.49	0.08	0.982	0.49	0.08
2.0	0.0	0.932	0.53	0.14	0.968	0.70	0.13
	-0.5	0.952	0.49	0.08	0.980	0.50	0.08
	-1.0	0.960	0.49	0.08	0.978	0.49	0.08
	-1.5	0.960	0.49	0.07	0.984	0.49	0.07
	-2.0	0.956	0.49	0.08	0.988	0.49	0.08

Betensky, R.A. and Tierney, C. (1997). An examination of methods for sample size recalculation during an experiment. *Statistics in Medicine* **16**, 2587–2598.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models* (2nd ed.). Chapman & Hall/CRC, London.

Table 4 The empirical mean, standard deviation of the estimated optimal sample fraction and the random sample size; and the empirical coverage probability, and the mean and standard deviation of the half width of the maximal axis of the constructed confidence ellipsoid of β with various initial sample sizes. The parameters are $\beta_{10} = 0.5, \beta_{20} = -1.0$ and $\pi_{E,opt} = 0.63, n_{E,opt} = 380, \pi_{R,opt} = 0.63, n_{R,opt} = 303$.

		$\hat{\pi}_{opt}$		τ_d		Coverage	Half width	
	$n^{(0)}$	mean	s.e.	mean	s.e.	rate	mean	s.e.
Ellipsoid	50	0.61	0.050	662.0	419.0	0.978	0.43	0.120
	100	0.61	0.038	505.7	273.8	0.958	0.47	0.097
	150	0.62	0.033	464.3	181.3	0.970	0.48	0.072
	200	0.62	0.026	426.4	139.6	0.970	0.50	0.062
	250	0.63	0.023	422.1	123.3	0.974	0.50	0.057
	300	0.63	0.021	405.3	100.0	0.982	0.52	0.045
Hyperrectangle	50	0.59	0.044	514.8	437.3	0.990	0.44	0.127
	100	0.61	0.037	388.3	178.9	0.998	0.47	0.090
	150	0.61	0.031	370.3	142.2	0.996	0.49	0.073
	200	0.62	0.028	343.8	102.2	0.996	0.50	0.059
	250	0.62	0.027	342.0	103.0	0.998	0.51	0.050
	300	0.62	0.026	337.1	92.0	1.000	0.52	0.043

Table 5 Sample size estimation $\tau_{R,d}$ of Framingham data

d	1:1.7[a]	1:1	1:2	1:5	1:10
1.0	1729	1787	1812	2068	2954
1.5	769	794	805	919	1313
2.0	432	447	453	517	739
2.5	277	286	290	331	473

[a]:Optimal sampling fraction, $\hat{\pi}_{R,opt} = 0.37$.

Chang, Y-c.I. (2001). Sequential confidence regions of generalized linear models with adaptive designs. *Journal of Statistical Planning and Inference* **93**, 277–293.

Chang, Y-c.I. and Martinsek, A.T. (1992). Fixed size confidence regions for parameters of a logistic regression model. *The Annals of Statistics* **20**, 1953–1969.

Chen, K. (2000). Optimal sequential designs of case-control studies. *The Annals of Statistics* **28**, 1452–1471.

Coffey, C.S. and Muller, K.E. (1999). Exact test size and power of a Gaussian error linear model for an internal pilot study. *Statistics in Medicine* **18**, 1199–1214.

Table 6 Data analysis of Framingham data at $\pi = 0.37, d = 2.5$

Variable	Estimate	s.e.	$(\hat{\beta}_{in} - \sqrt{\chi 2_{p,\alpha}\, e_{ii}/n},\ \hat{\beta}_{in} + \sqrt{\chi 2_{p,\alpha}\, e_{ii}/n})$
AGE	0.050	0.019	(-0.007 , 0.108)
SMOKE	1.006	0.339	(-0.039 , 2.050)
LSBP	2.463	0.618	(0.559 , 4.367)
CHOL	2.339	0.844	(-0.262 , 4.941)

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

Farewell, V.T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika* **66**, 27–32.

Friede, T. and Kieser, M. (2004). Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics* **3**, 269–279.

Ghosh, M. and Mukherjee, B. (2006). Data-adaptive sequential design for case-control studies. *Statistica Sinica* **16**, 697–719.

Gould, A.L. and Shih, W.J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics: Theory and Methods* **21**, 2833–2853.

Grambsch, P. (1989). Sequential maximum likelihood estimation with applications to logistic regression in case-control studies. *Journal of Statistical Planning and Inference* **22**, 355-369.

Hsieh, F.Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine* **8**, 795–802.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate statistical Analysis* (3rd ed.). Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901–911.

Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411

Proschan, M.A.(2005) Two-stage sample size re-rstimation based on a nuisance parameter: a review. *Journal of Biopharmaceutical Statistics* **15**, 559–574.

Proschan, M.A.(2009) Sample size re-estimation in clinical trials. *Biometrical Journal* **51**, 348–357.

Shieh, G. (2001). Sample size calculations for logistic and Poisson regression models. *Biometrika* **88**, 1193–1199.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* **16**, 243–258.

Wang, C.Y. and Wang, S. (1995). On information matrices in case-control studies. *Statistics and Probability Letters* **22**, 269–274.

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

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E. and Proschan, M. (1999). Internal pilot studies I: Type I error rate of the naive t-test. *Statistics in Medicine* **18**, 3481–3491.