JP Journal of Biostatistics
Volume 3, Number 3, 2009, Pages 215-224

SCREENING VARIABLE TO DETERMINE BINARY RESPONSES TO MAXIMIZE THE ASYMPTOTIC RELATIVE EFFICIENCY

LAKSHMI DAMARAJU

Centocor R & D, Inc.
200 Great Valley Parkway
Malvern, PA 19355-1307, U. S. A.

Abstract

Classifying an individual unit into normal (N) or abnormal $(\overline{N})$ categories will often be expensive and time consuming. Instead, if a screening variable, which can be easily measured is available, one may use the screening variable measurement to classify that individual. Assuming that the individual units in the abnormal category takes smaller values of the screening variable, we need to find a threshold u such that units with screening variable measurement at most u are classified $\overline{N}$ and others are classified N. In this paper, we determine u so that the asymptotic relative efficiency in testing the prevalence rate of $\overline{N}$ with and without screening variable measurement is maximum.

1. Introduction

Classification of an experimental unit into either normal (N) or abnormal $(\overline{N})$ categories can be very expensive and time consuming. Instead of directly classifying the experimental unit, observations on a screening variable can be used to classify the experimental unit. However, this comes with a price of introducing misclassification errors. By using a screening variable, an experimental unit might be

---

2000 Mathematics Subject Classification: Primary 62P10; Secondary 62P99.

Keywords and phrases: false positive, false negative, normal distribution.

Received July 25, 2009

classified as N when it is really $\overline{N}$, and vice versa. Let ℓ_1 be the probability of committing a false positive error and ℓ_2 be the probability of committing a false negative error. In this paper, we will restrict ourselves to continuous screening variables. We will also assume that each of ℓ_1 and ℓ_2 is less than 0.5, so that the correlation between true and classified binary responses is positive (see [3]).

We will now illustrate the use of screening variables through some examples:

Example 1.1. Catheterization is an invasive procedure, which cardiologists perform only if they are pretty certain that the patient has arteriosclerosis. Thus a screening variable X, where X is the number of minutes a patient can walk on the treadmill, is used to determine if a patient has arteriosclerosis. For example, if $X > 15$ minutes, then the cardiologist may determine that the patient is okay and there is no need to do the catheterization. On the other hand, if $X \leq 15$ minutes, then the cardiologist may conclude that the patient has some blocked artery and will need a catheterization.

Example 1.2. SAT score may be used by the admissions office in a university to determine if a student can obtain his/her B.S. degree. For example, if $X > 900$, then the admissions office will assume that the student can complete the degree without any problem, but if $X \leq 900$, then the student may have a problem obtaining a degree.

In both of these examples, it is cheaper and faster to use screening variables to increase the probability of classifying the $\overline{N}$ cases. However, the chances of wrong decisions also increase.

Let π be the true prevalence rate of abnormal in a population. We will now consider the problem of testing $\pi = \pi_0$. Let X be a screening variable which is assumed to take smaller values for units classified as $\overline{N}$. We find a threshold value u, such that if X is in the interval $(-\infty, u]$, we will classify the corresponding unit as $\overline{N}$ or else we will classify the corresponding unit as N. The threshold value for a screening variable u can be set using different criteria like:

(1) minimizing the misclassification error rate;

(2) controlling the cost for treatment;

(3) maximizing the asymptotic relative efficiency to test the prevalence rate with and without the screening variable.

This u may depend on the distributions of the screening variable in the N and $\overline{N}$ cases and also the prevalence rate of the abnormal.

In this paper, we will discuss how to determine this threshold value of u by using asymptotic relative efficiency in testing the hypothesis, that is, sample size with misclassification errors compared to sample size without misclassification errors for a given power. The threshold determined in this way, when the screening variable has a normal distribution in both the groups with equal variances, is same as the threshold determined by minimizing the total misclassification error. This is not the case when the variances are unequal.

The screening variables for getting binary responses associated with misclassification errors from a different perspective were discussed by Dunsmore and Boys [2], and Boys and Glazebrook [1].

In Section 2, we will give some general methodology, such as the test statistic, sample size calculation and an expression for the asymptotic relative efficiency. In Sections 3 and 4, we will illustrate our methodology using the normal distribution.

2. General Methodology

Let us now consider the random variables (X, T), where X is a screening variable and T is a binary variable showing the true status, such that if $T = 1$, then the unit belongs to the $\overline{N}$ category and if $T = 0$, then the unit belongs to the N category. Let

$$\Pr(T = 1) = \pi, \quad \Pr(T = 0) = \overline{\pi}.$$

Let us assume that $X \mid (T = 1)$ has a continuous density function $f(x)$ and a cumulative distribution function $F(x)$, and $X \mid (T = 0)$ has a continuous density function $g(x)$ and a cumulative distribution function $G(x)$. We will assume that $F(x) \geq G(x)$, for all real x, that is, F is stochastically smaller than G.

One would like to determine a threshold value of u for the continuous screening variable X, such that, a unit with $X \leq u$ is classified as $\overline{N}$ and a unit with $X > u$ is classified as N. Now

$$\Pr(\text{a unit is classified as } \overline{N} \text{ based on } X) = \Pr(X \leq u) = \pi F(u) + \overline{\pi} G(u) = \theta, \quad \text{say.}$$

Let us consider the problem of testing the hypotheses

$$H_0 : \pi = \pi_0, \quad H_A : \pi = \pi_0 + \varepsilon,$$

where $\varepsilon > 0$ and π_0 is a given value. Let $\theta_\varepsilon = (\pi_0 + \varepsilon)F(u) + (\bar{\pi}_0 - \varepsilon)G(u)$. Then the hypotheses are equivalent to

$$H_0 : \theta = \theta_0, \quad H_A : \theta = \theta_\varepsilon. \tag{2.1}$$

Let a random sample of size *n* be taken from the population and be classified into *N* and $\bar{N}$ categories, based on the screening variable. Let n_1 observations belong to the $\bar{N}$ category. A test statistic for testing the hypotheses given in equation (2.1) is

$$Z = \frac{\arcsin\left(\sqrt{\frac{n_1}{n}}\right) - \arcsin(\sqrt{\theta_0})}{\sqrt{\frac{1}{4n}}},$$

with critical region $Z > z_\alpha$, where z_α is the 100α upper percentile point of a standard normal variable. Let us determine the sample size *n* such that the power of the test is at least $1 - \beta$. Then

$$n = \frac{(z_\alpha + z_\beta)^2}{4[\arcsin\sqrt{\theta_\varepsilon} - \arcsin\sqrt{\theta_0}]^2}.$$

By using Taylor's expansion for $\arcsin\sqrt{\theta_\varepsilon}$, we obtain

$$\arcsin(\sqrt{\theta_\varepsilon}) = \arcsin\sqrt{\theta_0} + \frac{\varepsilon(F(u) - G(u))}{2\sqrt{\theta_0(1 - \theta_0)}}.$$

We can now rewrite the sample size *n* as

$$n = \frac{(z_\alpha + z_\beta)^2\theta_0(1 - \theta_0)}{\varepsilon^2(F(u) - G(u))^2}. \tag{2.2}$$

Our false positive and false negative error rates correspond to $\ell_1 = G(u)$ and $\ell_2 = 1 - F(u)$.

Let *R* be the asymptotic relative efficiency for the test with misclassification errors compared to without misclassification errors for a given power. Then $R = n_0/n$, where n_0 is obtained by putting $F(u) = 1$ and $G(u) = 0$ in equation

(2.2), and is

$$R(u) = (F(u) - G(u))^2 \frac{\pi_0(1 - \pi_0)}{\theta_0(1 - \theta_0)}. \qquad (2.3)$$

Since π_0 is unknown, let us consider the worst case scenario, of $\pi_0 = 0.5$, when n_0 is maximum for a given ε. Then equation (2.3) becomes

$$R(u) = \frac{(F(u) - G(u))^2}{1 - (1 - F(u) - G(u))^2} = \frac{(1 - l_1 - l_2)^2}{1 - (l_1 - l_2)^2}. \qquad (2.4)$$

Let us now consider *R* to be a function of *u* and try to maximize *R* of equation (2.4). Note that the first derivative of *R* with respect to *u*, $R(u)$ is

$$R'(u) = \frac{2(F(u) - G(u))^2(1 - F(u) - G(u))(f(u) + g(u))}{(F(u) + G(u))^2(2 - F(u) - G(u))^2} H(u),$$

where

$$H(u) = \frac{f(u) - g(u)}{f(u) + g(u)} \; \frac{F(u) + G(u)}{F(u) - G(u)} \; \frac{2 - F(u) - G(u)}{1 - F(u) - G(u)} - 1.$$

We cannot get the critical value for *u* satisfying $H(u) = 0$ in a mathematically closed form in the general case. One needs to determine *u* in a reasonable interval (a, b), so that the false positive and false negative error rates are each less than half and plot $R(u)$ against *u* to find a *u* that maximizes $R(u)$ or one can use NLIN in SAS to obtain the threshold *u*, where $R(u)$ is maximized. This threshold determined by using NLIN need not correspond to a point *u*, where each of the error rates is less than half.

3. Screening Variable Normally Distributed with Equal Variances

Let $X \mid (T = 1) \sim N(\mu_1, \sigma^2)$ and $X \mid (T = 0) \sim N(\mu_2, \sigma^2)$, where $\mu_1 < \mu_2$. Then the false positive error rate

$$\ell_1 = \Phi((u - \mu_2)/\sigma)$$

and the false negative error rate

$$\ell_2 = 1 - \Phi((u - \mu_1)/\sigma).$$

In order that $\ell_1 < 0.5$ and $\ell_2 < 0.5$, we restrict *u* to the open interval (μ_1, μ_2). We will now maximize R given in equation (2.4). The threshold value u that maximizes R in equation (2.4), when $X \mid (T = 1) \sim N(\mu_1, \sigma^2)$ and $X \mid (T = 0) \sim N(\mu_2, \sigma^2)$, where $\mu_1 < \mu_2$ is

$$u = \frac{\mu_1 + \mu_2}{2}.$$

This can be shown by assuming ℓ_1 and ℓ_2 as functions of u. Let ℓ'_1 and ℓ'_2 be the first order derivatives, and let ℓ''_1 and ℓ''_2 be the second order derivatives, with respect to u. Then

$$l'_1 = \left(\frac{1}{\sigma}\right)\phi\left(\frac{u - \mu_2}{\sigma}\right),$$

$$l'_2 = -\left(\frac{1}{\sigma}\right)\phi\left(\frac{u - \mu_1}{\sigma}\right),$$

$$l''_1 = -\left(\frac{u - \mu_2}{\sigma^3}\right)\phi\left(\frac{u - \mu_2}{\sigma}\right),$$

$$l''_2 = \left(\frac{u - \mu_1}{\sigma^3}\right)\phi\left(\frac{u - \mu_1}{\sigma}\right).$$

The first order derivative of R with respect to u satisfies

$$\frac{(1 - (l_1 - l_2)^2)^2}{-2(1 - l_1 - l_2)} R'(u) = (1 - (l_1 - l_2)^2)(l'_1 + l'_2) - (1 - l_1 - l_2)(l_1 - l_2)(l'_1 - l'_2). \quad (3.1)$$

We will now prove that $R'(u) = 0$, when $u = (\mu_1 + \mu_2)/2$. Let us denote by ℓ_1^*, ℓ_2^*, $\ell_1^{*'}$, $\ell_2^{*'}$, $\ell_1^{*''}$, $\ell_2^{*''}$ the values of ℓ_1, ℓ_2, ℓ'_1, ℓ'_2, ℓ''_1, ℓ''_2, when $u = (\mu_1 + \mu_2)/2$. Clearly,

$$l_1^* = l_2^* = \Phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right),$$

$$l_1^{*'} = -l_2^{*'} = \frac{1}{\sigma}\phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right),$$

$$l_1^{*''} = l_2^{*''} = \frac{\mu_2 - \mu_1}{2\sigma^3}\phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right).$$

Substituting $\ell_1^* - \ell_2^* = 0$ and $\ell_1^{*'} + \ell_2^{*'} = 0$ in equation (3.2), we get $R'(u) = 0$. Thus $u = (\mu_1 + \mu_2)/2$ is a critical value of $R(u)$.

We will now show that the second derivative of $R(u)$ with respect to u, $R''(u)$, is negative when $u = (\mu_1 + \mu_2)/2$. We have

$$-\frac{1}{2}R''\left(\frac{\mu_1 + \mu_2}{2}\right) = \frac{(1 - 2l_1^*)2l_1^{*''} - (1 - 2l_1^*)^2(2l_1^{*''})^2}{(1 - (l_1^* - l_2^*)^2)^4}. \tag{3.2}$$

When $x > 0$, from the Mean Value Theorem of Differential Calculus, we obtain

$$\frac{\Phi(x) - \Phi(-x)}{2x} = \phi(a), \quad \text{where} \quad -x < a < x.$$

Consequently,

$$\Phi(x) - \Phi(-x) \leq 2x\frac{1}{\sqrt{2\pi}}, \tag{3.3}$$

since $e^{-\left(\frac{x^2}{2}\right)} \leq 1$. Thus, using equation (3.3),

$$1 - 2l_1^* = \Phi\left(\frac{\mu_2 - \mu_1}{2\sigma}\right) - \Phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right) \leq \left(\frac{\mu_2 - \mu_1}{\sigma}\right)\frac{1}{\sqrt{2\pi}}. \tag{3.4}$$

Replacing equation (3.4) in equation (3.2), we obtain

$$-\frac{1}{2}(1 - (l_1^* - l_2^*)^2)^4 R''\left(\frac{\mu + \mu_2}{2}\right)$$

$$= (1 - 2l_1^*)\frac{2e^{-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}}}{\sqrt{2\pi}\,\sigma}\left[\frac{(\mu_2 - \mu_1)}{2\sigma^2} - \frac{2(1 - 2l_1^*)e^{-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}}}{\sqrt{2\pi}\,\sigma}\right]$$

$$\geq (1 - 2l_1^*)\sqrt{\frac{2}{\pi}}\,\frac{1}{\sigma^3}\,e^{-\frac{(\mu_1-\mu_2)^2}{8\sigma^2}}(\mu_2 - \mu_1)\left(\frac{1}{2} - \frac{1}{\pi}\right)$$

$$> 0.$$

Thus $R''((\mu_1 + \mu_2)/2)$ is negative. Note that the threshold value is the same threshold value that is used in discriminant analysis to classify an individual into one of two normal populations with equal variance so that the probability of misclassification is minimum.

4. Screening Variable Normally Distributed with Unequal Variances

When the screening variable is normally distributed with equal variances we obtain closed expressions as given in the last section. However, it is not the case when we have unequal variances. Here we will try to find the threshold *u* using NLIN in SAS. The proofs of these results in this section are given in the thesis of Lakshmi [4].

Let $X \mid (T = 1) \sim N(\mu_1, \sigma_1^2)$ and $X \mid (T = 0) \sim N(\mu_2, \sigma_2^2),$ where $\mu_1 < \mu_2$ and $\sigma_1 \neq \sigma_2.$ The false positive error rate $\ell_1 = \Phi((u - \mu_2)/\sigma_2)$ and the false negative error rate $\ell_2 = 1 - \Phi((u - \mu_1)/\sigma_1).$ In order that $\ell_1 < 0.5$ and $\ell_2 < 0.5,$ we restrict *u* to the open interval $(\mu_1, \mu_2).$ We will now maximize

$$R = \frac{(1 - l_1 - l_2)^2}{1 - (l_1 - l_2)^2}$$

as in Section 3.

4.1. $\sigma_1 < \sigma_2$

In this case, define

$$u^* = \frac{\frac{1}{\sigma_1}\mu_1 + \frac{1}{\sigma_2}\mu_2}{\frac{1}{\sigma_1} + \frac{1}{\sigma_2}}$$

$$u^{**} = \frac{\frac{1}{\sigma_1}\mu_1 - \frac{1}{\sigma_2}\mu_2}{\frac{1}{\sigma_1} - \frac{1}{\sigma_2}},$$

$$u_1 = u^{**} + \frac{(\mu_1 - \mu_2)\delta}{\sigma_1\sigma_2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)},$$

$$u_2 = u^* - \frac{(\mu_1 - \mu_2)\delta}{\sigma_1\sigma_2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)},$$

$$\delta = \sqrt{1 - \frac{\sigma_1^2 - \sigma_2^2}{(\mu_1 - \mu_2)^2}\ln\frac{\sigma_1^2}{\sigma_2^2}} - 1 \quad (> 0).$$

If $(\mu_2 - \mu_1)^2/\sigma_1^2 > \ln(4\sigma_2^2/\sigma_1^2)$, then $u_1 < u^{**} < \mu_1 < u^* < u_2 < \mu_2$, and $R(u)$ has a maximum at u in the open interval (u_2, μ_2).

Example 4.1.1. If $\sigma_1 = 5$, $\sigma_2 = 10$, $\mu_1 = 10$, $\mu_2 = 30$, $u_1 = -11.63$, $u_2 = 18.29$, $u^* = 16.67$ and $u^{**} = -10$, then, by plotting u against $R(u)$ (see Figure 1), or by using PROC NLIN, we find the maximum at $u = 18.61$ which is between u_2 and μ_2.

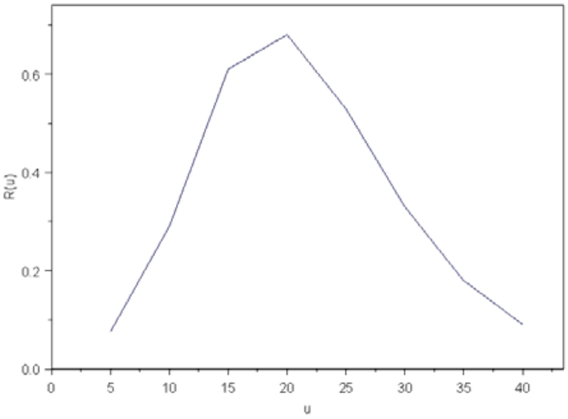

Figure 1. Graph of $R(u)$ vs. u.

Example 4.1.2. An example of a violation of the condition is when $\mu_1 = 10$, $\mu_2 = 15$, $\sigma_1 = 5$, and $\sigma_2 = 10$. Here $u_1 = 0.76$, $u^{**} = 5$, $u^* = 11.67$, $u_2 = 15.9$ and $R(u)$ has a maximum at $u = 18.0539$.

4.2. $\sigma_1 > \sigma_2$

As in Subsection 4.1, $(\mu_1 - \mu_2)^2/\sigma_1^2 > \ln(4\sigma_2^2/\sigma_1^2)$, then $\mu_1 < u_2 < u^* < \mu_2 < u^{**} < u_1$, and $R(u)$ has a maximum at u in the open interval (μ_1, u_2).

The condition $(\mu_2 - \mu_1)^2 / \sigma_2^2 > \ln(4\sigma_1^2 / \sigma_2^2)$ is needed so that R has a maximum between the narrow interval of (μ_1, μ_2).

Example 4.2.1. If $\sigma_1 = 10$, $\sigma_2 = 5$, $\mu_1 = 10$, $\mu_2 = 30$, $u_1 = 51.63$, $u_2 = 21.70$, $u^* = 23.33$ and $u^{**} = 50$, then, by plotting u against $R(u)$, we find the maximum at $u = 21.38$ which is between μ_1 and u_2.

Example 4.2.2. An example of a violation of the condition is when $\mu_1 = 10$, $\mu_2 = 15$, $\sigma_1 = 10$ and $\sigma_2 = 5$. Here $u_1 = 24.23$, $u^{**} = 20$, $u^* = 13.33$, $u_2 = 9.10$ and $R(u)$ has a maximum at $u = 6.9421$, which is not in the open interval $(10, 15)$.

In the case, when $\sigma_1 \neq \sigma_2$, the threshold value u determined by our method will be different from the threshold determined by minimizing the misclassification errors. In the case of Example 4.1.1, the threshold u that minimizes the total misclassification error is $u = 18.3$ whereas our method gave $u = 18.61$.

Acknowledgement

I am very thankful to Professor Woollcott K. Smith of Temple University for his helpful and constructive comments in preparing this manuscript.

References

[1] R. J. Boys and K. D. Glazebrook, A robust design of a screen for a binary response, Biometrics 79 (1992), 643-650.

[2] I. R. Dunsmore and R. J. Boys, Predictive screening methods in binary response models, Probability and Bayseian Statistics, R. Viertl, ed., Plenum, New York, 1987, pp. 151-158.

[3] S. W. Lagakos, Effects of mismodelling and mismeasureing explanatory variables on tests of their association with a response variable, Statist. Medi. 7 (1988), 257-274.

[4] D. Lakshmi, Inferences in the presence of misclassification errors, Ph.D. Dissertation, Temple University, 1995.