Journal of the Korean Statistical Society 43 (2014) 463–474

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Robust suptest for the genetic association study under genetic model uncertainty

Minjung Kwak*

Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk, 712-749, Republic of Korea

ARTICLE INFO

Article history:
Received 28 November 2013
Accepted 12 February 2014
Available online 27 February 2014

AMS 2000 subject classifications:
primary 62P05
secondary 62E99

Keywords:
Case–control study
Genome-wide association study
Robust test
Supremum test
Asymptotic property

ABSTRACT

In case–control studies the Cochran–Armitage trend test is powerful for detection of an association between a risk genetic marker and a disease of interest. To apply this test, a score should be assigned to the genotypes based on the genetic model. When the underlying genetic model is unknown, the trend test statistic is quite sensitive to the choice of the score. In this paper, we study the asymptotic property of the robust suptest statistic defined as a supremum of Cochran–Armitage trend test across all scores between 0 and 1. Through numerical studies we show that small to moderate sample size performances of the suptest appear reasonable in terms of type I error control and we compared empirical powers of the suptest to those of three individual Cochran–Armitage trend tests and the maximum of the three Cochran–Armitage trend tests. The use of the suptest is applied to rheumatoid arthritis data from a genome-wide association study.

1. Introduction

The case–control study is a powerful design to detect an association between a candidate marker and a disease. It is a commonly used design in recent genome-wide association studies where case and control samples are randomly drawn from the study populations. Genome-wide association studies involve rapidly scanning genetic markers across the genome to find genetic variations associated with a disease. Such studies are particularly useful to detect genetic variations that contribute to many common and complex diseases (Klein et al., 2005; Sladek et al., 2007; The Wellcome Trust Case Control Consortium, 2007). For genome-wide association studies, the issues include that thousands of individuals are genotyped on 100,000–500,000 SNPs, while usually only a few SNPs are associated with the disease. Another difficulty is that for many complex diseases, the underlying genetic model is unknown.

Typical tests applied to a genome-wide association study using case–control data include Pearson's Chi-square test and the Cochran–Armitage trend test discussed and illustrated in Agresti (1990), and Freidlin, Zheng, Li, and Gastwirth (2002). Three asymptotically optimal trend tests are available for the recessive, additive, and dominant genetic models respectively. However, the optimality of such a test depends on the information of true genetic model which is not known in real world. In this case, the allele based test or the trend test optimal for the additive model is usually used in practice, but they are not robust across the above three possible genetic models. Several robust methods using maximal type tests have been studied. Among many others Freidlin et al. (2002) and Zheng, Joo, and Yang (2009) considered as their test statistic the maximum of three trend tests under dominant, additive, and recessive genetic models. On the other hand the constrained likelihood ratio

* Tel.: +82 53 810 2321; fax: +82 53 810 4615.
E-mail address: mjkwak@yu.ac.kr.

http://dx.doi.org/10.1016/j.jkss.2014.02.002
1226-3192/

Table 1
Genotype distribution for a single SNP.

Genotype	*AA*	*AB*	*BB*	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

test where one maximizes the likelihood function in the constrained parameter space was studied by Wang and Sheffield (2005). A robust test recently proposed by WTCCC uses the minimum of the p-values of Pearson's test and the trend test for the additive model.

In this paper, we focus on the Cochran–Armitage trend test which has been used for testing case–control association. One advantage of the Cochran–Armitage trend test is that, when the genetic model is known, an optimal Cochran–Armitage trend test is available. If the underlying mode of inheritance is additive or multiplicative, then the Cochran–Armitage trend test is asymptotically equivalent to the allele based test (Sasieni, 1997). The second advantage of the Cochran–Armitage trend test is its robustness to departure from the Hardy–Weinberg equilibrium (Sasieni, 1997). Our study is new in that we propose as a robust test the supremum of the Cochran–Armitage trend test over all possible genetic models and provide the asymptotic property of the proposed test using the empirical process technique. We also show that our test can be implemented with critical values obtained easily by generating random numbers from a bivariate normal distribution. We compare our method to existing tests by simulation studies under various genetic models. We also apply the proposed test to a genome-wide association study to illustrate our method.

2. Robust tests in genetic association study

Suppose r cases and s controls are sampled and their genotypes are obtained at a biallelic candidate marker with alleles A and B, e.g. a single-nucleotide polymorphism (SNP). The case–control data can be summarized by a 2×3 table (Table 1), where (r_0, r_1, r_2) and (s_0, s_1, s_2) are genotype counts for $(G_0, G_1, G_2) = (AA, AB, BB)$ in case subjects and control subjects, respectively. We have that $r = r_0 + r_1 + r_2$ and $s = s_0 + s_1 + s_2$ and we let $n_i = r_i + s_i$, for $i = 0, 1, 2$. Let $n = r + s$ be the total sample size. Penetrances are denoted by $f_i = \Pr(\text{disease}|G_i),\ i = 0, 1, 2$. Genotype relative risks are given by $\lambda_1 = f_1/f_0$ and $\lambda_2 = f_2/f_0$. The null hypothesis H_0 is expressed as $H_0 : \lambda_1 = \lambda_2 = 1$. If allele B is the risk allele under the alternative hypothesis H_1, we expect under the dose–response model that $\lambda_2 \geq \lambda_1 \geq 1$ and $\lambda_2 \geq 1$. The dominant, additive, and recessive genetic models correspond to $\lambda_1 = \lambda_2,\ \lambda_1 = (1 + \lambda_2)/2$, and $\lambda_1 = 1$, respectively.

From Table 1, we have $\hat{p}' = (r_1 + 2r_2)/(2r)$ and $\hat{p} = (s_2 + 2s_2)/(2s)$, the estimated risk allele frequencies in cases and controls, respectively. The allele based test, comparing the frequency of disease-associated allele in cases and controls, can be written as

$$U_{AB} = \frac{\hat{p}' - \hat{p}}{\sqrt{\{\hat{p}'(1-\hat{p}') + \hat{p}(1-\hat{p})\}/(2n)}}.$$

Under H_0, U_{AB}^2 asymptotically follows a Chi-squared distribution with 1 degree of freedom, denoted by χ_1^2. The allelic test aims to identify significant differences in allelic proportions between case and control subjects. To this end, one supposes that alleles are binomially and independently sampled from cases and controls with probabilities corresponding to the proportions of the susceptibility allele in the two groups. However unless the observed genotypic proportions are strictly in the Hardy–Weinberg equilibrium the allelic test is biased (Guedj, Nuel, & Prum, 2008).

As an alternative approach, the genotype-based Cochran–Armitage (CA) trend test can be obtained as the score test from the logistic regression model (Sasieni, 1997). A set of increasing scores $(x_0, x_1, x_2) = (0, \theta, 1), \theta \in [0, 1]$ is used to code the three genotypes (G_0, G_1, G_2) in logistic regression analysis. And the CA trend test depending on the scores (x_0, x_1, x_2) can be written as

$$U_{n,\mathrm{CA}}(\theta) = \frac{n^{1/2} \sum_{i=0}^{2} x_i (s r_i - r s_i)}{\left[rs \left\{ n \sum_{i=0}^{2} x_i^2 n_i - \left(\sum_{i=0}^{2} x_i n_i \right)^2 \right\} \right]^{1/2}}.$$

For a given $\theta \in [0, 1]$, $U_{n,\mathrm{CA}}^2(\theta)$ asymptotically follows χ_1^2 under the null hypothesis. To apply the Cochran–Armitage trend test, the choices of θ for the recessive, additive, and dominant models are 0, 1/2 and 1, respectively. Previous studies show that $U_{n,\mathrm{CA}}(k)$, $k = 0, 1/2, 1$ are asymptotically optimal under each of the three specified models (Freidlin et al., 2002; Zheng, Freidlin, Li, & Gastwirth, 2003; Sasieni, 1997).

For many complex diseases, the underlying genetic model is unknown. Because there is no uniformly most powerful test across all possible alternative genetic models, it is not likely to have a single best test for all situations. However, it is possible to compare the tests under their worst situations across all scientifically plausible genetic models. A test with the