Robust method of estimating of odd ratios in case-control association studies

MANJU MARY PAUL* and ANIL RAI

ICAR-Indian Agricultural Statistics Research Institute, New Delhi 110 012, India

Received: 24 July 2019; Accepted: 31 July 2019

ABSTRACT

In case-control studies, association of a disease with a genetic trait can be studied in terms of Single Nucleotide Polymorphisms (SNP). Prospective and retrospective likelihoods are the two common approaches used to study this association. In this paper, a method based on Preliminary test, has been proposed, which is efficient than prospective approach by exploiting model assumptions of Hardy Weinberg Equilibrium (HWE) and robust against failure of model assumptions as compared to retrospective approach (2015–16). In this proposed approach, a Preliminary Test Estimator (PTE) has been suggested based on both prospective and retrospective approach for estimating of association of a disease with genetic marker. The proposed PTE has been empirically evaluated through a simulation study. The Preliminary Test Estimator developed was found to be robust against the deviation from HWE.

Key words: Case-Control Study, Preliminary Test Estimator, Prospective Approach, Retrospective Approach, Single Nucleotide Polymorphisms

Plant disease is one of the major causes of reduced agricultural production. Association of a disease with a genomic region can be studied in terms of underlying molecular markers such as Single Nucleotide Polymorphisms (SNPs), SSR (Single Sequence Repeats) and haplotypes etc. Case- control study designs are often used to study association between genetic factors and a disease, which involves identification of individuals with (‘cases’, i.e. having disease) and without (‘controls’, i.e. not having disease) a particular disease or condition. The standard method for analysis of case-control data is the prospective logistic regression, ignoring the retrospective nature of the underlying design. Andersen (1970) and Prentice and Pyke (1979) showed that such a prospective approach is actually equivalent to the retrospective maximum likelihood analysis, provided that the covariates have nonparametric distribution. Chatterjee and Carroll (2005) developed a retrospective maximum-likelihood approach for analysis of case-control studies, exploiting gene-environment independence and HWE assumption. The “prospective” method does not depend on the assumptions of Hardy-Weinberg Equilibrium (HWE) and is more robust than retrospective method. “Retrospective” method is more efficient than the prospective methods under HWE conditions and thus produces more precise estimates of odd ratios. But when the underlying assumptions of HWE are violated, the retrospective estimator becomes biased (Chatterjee et al. 2009).

Bancroft introduced Preliminary Test Estimator (PTE) for a situation in which one has two alternative solution/estimator, where, each providing an estimate of an unknown parameter. Later on, several researchers like Sukhatme and Tang (1975), Das and Bez (1995) applied PTE in the area of sample surveys. Rai and Srivastava (1998) proposed a PTE based on a test statistic for estimation of regression coefficient from survey data. It has been observed from the above studies that PTE is likely to be either more efficient or robust than their respective approaches as both these approaches are being used optimally.

Hence, a PTE based method, which can gain efficiency by exploiting model assumptions of HWE and also resistant to bias when these model assumptions are violated, has been proposed in this article. The outcome of the preliminary test decides about the estimator to be used out of the two alternatives i.e. prospective or retrospective, in a given situation.

MATERIALS AND METHODS

Let D denote the disease status of an individual, with D=1 (cases) and D=0 (controls). Let G be the number of minor alleles carried by an individual (G = 0, 1, 2) and ndg denote number of subjects with genotype G = g and disease status D = d as observed in the case-control sample.

Prospective approach: Let data on some genetic (G) and environmental (E) exposures be collected in a case-control study involving N0 controls (D = 0) and N1 cases (D = 1).
The probability of developing a disease given a genetic and environmental condition, i.e. the prospective-likelihood is given by Chatterjee et al. (2009). The maximum likelihood estimates of the Odds Ratio parameters for genotypes Aa and aa, say denoted by $\hat{\beta}_{as}$ and $\beta_{as}$, where both defined in reference to the baseline genotype, i.e. AA, as well as the variance-covariance matrix of the estimated log odds ratio can be obtained as given by Chen et al. (2007). It can be seen that the estimation of odd ratios for association of disease with a genotype in prospective approach is not based on HWE assumptions.

**Retrospective approach:** The retrospective likelihood for the genotype data of a single-SNP is given by the product of two sets of multinomial probabilities as described by Chatterjee et al. (2009). Expression for log odds ratio as well as the variance-covariance matrix that estimates this log odd ratio is given by Chatterjee et al. 2009. Retrospective method produces much more precise estimates of odd ratio as compared to prospective approach under HWE assumptions (Satten and Epstein, 2004). But when the underlying assumptions are violated, the estimator becomes biased, whereas, the “prospective" methods does not depend on the assumptions of HWE and is considered more robust (Luo et al. 2009). Thus, it is desirable to develop a Preliminary Test statistic to decide, whether, to use prospective or retrospective method to estimate odds ratio based on its outcome. The test statistic is given by $\lambda$. The details of the proposed method are given in the section below.

**Proposed approach of Preliminary Test Estimator (PTE):** Preliminary Test Estimator (PTE) can be implemented using both prospective and retrospective methods. It has already demonstrated that prospective and retrospective odd ratios are equivalent in the absence of HWE assumption. Hence, acceptance of null hypothesis, i.e. the equivalence of both the estimates implies absence of HWE. In that case, prospective estimator will be preferred. But, when the population is in HWE, the null hypothesis will be rejected, retrospective estimator is the best and should be used for estimation of odd ratio. Therefore, PTE for estimating the odds ratio in case of heterozygotes can be written as:

$$\hat{\beta}_{as} = \begin{cases} \hat{\beta}_{as}^R & \text{if test rejects } H_0 : \hat{\beta}_{as}^R = \hat{\beta}_{as}^A \\ \hat{\beta}_{as}^A & \text{otherwise} \end{cases}$$

(1)

where $\hat{\beta}_{as}^R$, estimator based on retrospective approach; $\hat{\beta}_{as}^A$, estimator based on prospective approach.

The test statistic can be written as:

$$\bar{u}_{as} = \left( \hat{\beta}_{as}^R - \hat{\beta}_{as}^A \right)^2 \big/ \hat{\sigma}_{as}^2$$

(2)

Similarly, odds ratio in case of recessive homozygotes can also be estimated.

It can be seen that:

$$E(\hat{\beta}_{as}^A) = E(\hat{\beta}_{as}^R / \lambda < f_a) P(\lambda < f_a) + E(\hat{\beta}_{as}^R / \lambda > f_a) P(\lambda > f_a)$$

(3)

where $f_a$, denotes the tabulated value of F-distribution at $\alpha$-level of significance with 1 and n-2 degrees of freedom.

$$\hat{\beta}_{as}^A = \log \frac{n_{11}n_{00}}{n_{10}n_{01}}$$ and $\hat{\beta}_{as}^R = \log \frac{n_{11}n_{00}^2}{n_{10}n_{01}^2}$

(4)

The bias of PTE ($\hat{\beta}_{as}$) can be estimated as:

$$\text{bias}(\hat{\beta}_{as}) = E(\hat{\beta}_{as}^A) - \beta_{as}$$

$$\sigma^2 \sum_{j=0}^{n-2} \frac{1}{f_{a_j} - k} \left( f_{a_j} - f_{a_j + k} \right)$$

(4)

Mean square error of PTE ($\hat{\beta}_{as}$) can be:

$$E(\hat{\beta}_{as}^A) = E\left( E(\hat{\beta}_{as}^A / \lambda > f_a) P(\lambda > f_a) + \left( \hat{\beta}_{as}^R / \lambda < f_a \right) \right)$$

$$E(\hat{\beta}_{as}^R) = E\left( \left( \hat{\beta}_{as}^R / \lambda > f_a \right) \right)$$

$$\text{MSE}(\hat{\beta}_{as}) = \text{Var}(\hat{\beta}_{as}^A) + \text{Bias}^2(\hat{\beta}_{as}^A) = \text{E}(\hat{\beta}_{as}^A) - \text{E}(\hat{\beta}_{as}^A) + \text{Bias}^2(\hat{\beta}_{as}^A)$$

The MSE ($\hat{\beta}_{as}$) can be calculated as:

$$\text{MSE}(\hat{\beta}_{as}) = \sigma^2 + \sigma^2 \sum_{j=0}^{n-2} \frac{1}{f_{a_j} - k} \left( f_{a_j} - f_{a_j + k} \right)$$

(4)

**Empirical comparison of the proposed Preliminary Test Estimator (PTE):** The statistical performance of the proposed PTE is compared empirically through a simulation study following the approach of Luo et al. (2009). The genotype information corresponding to a single SNP was simulated for different combination of simulation parameters. This simulation has been done for two sample sizes, i.e. $n_0=n_1=1000$ and $n_0=n_1=500$ by taking into account different minor allele (a) frequencies (f), i.e. f = 0.1, 0.2, and 0.3. The data has been simulated for four situations in which population follows (i) HWE ($\theta = 0$), (ii) small deviation from HWE ($\theta = 0.5 \log (1.2)$), (iii) moderate deviation from HWE ($\theta = 0.5 \log (1.6)$) and (iv) large deviation from HWE ($\theta = 0.5 \log (2)$). Genotype probabilities for controls and cases are calculated as explained in Luo et al. (2009). After generating the probabilities for cases and controls, the genotype information for these groups were generated with help of multinomial distribution. The percentage bias and gain in efficiency for the estimators has been calculated using the equations given below.

$$\% \text{gain in efficiency} = \frac{\text{MSE}(\hat{\beta}_{as}^R (f, \theta)) - \text{MSE}(\hat{\beta}_{as}^A (f, \theta, 0, 0, 0, 0))}{\text{MSE}(\hat{\beta}_{as}^R (f, \theta))}$$

(7)

where $\hat{\beta}_{as}$ is the actual value of the parameter taken for simulation and $\hat{\beta}_{as}$ is the estimated value of the odds ratio

$$\% \text{gain in efficiency} = \frac{\text{MSE}(\hat{\beta}_{as}^R (f, \theta)) - \text{MSE}(\hat{\beta}_{as}^A (f, \theta, 0, 0, 0, 0))}{\text{MSE}(\hat{\beta}_{as}^A (f, \theta, 0, 0, 0, 0))}$$

(8)
The odds ratios are assumed to follow a “recessive” pattern with \( \psi_{Aa} = 1, \psi_{aa} = (1.4)^2 \) \( (\beta_{Aa}=0, \beta_{aa}= 0.6729444) \).

### RESULTS AND DISCUSSION

Results of the simulation study are presented in Tables 1 and 2. Table 1 gives the results for the estimation of \( \hat{\beta}_{aa} \) under the condition where the disease-genotype odds ratios are assumed to follow a “recessive” pattern with \( \psi_{Aa} = 1, \psi_{aa} = (1.4)^2 \) \( (\beta_{Aa}=0, \beta_{aa}= 0.6729444) \) sample size; \( n_1=n_0=500 \) and sample size \( n_1=n_0=1000 \). % bias in this table has been obtained with the help of equation (7) as compared to the actual value of \( \beta_{aa} \) of this simulation study.

It can be seen that the % bias decreases with increase in minor allele frequency. % bias for retrospective approach for estimation of odds ratio is less than prospective in case of Hardy Weinberg Equilibrium, whereas, % bias of PTE is the least i.e proposed PTE is almost unbiased. Also, it can be seen that generally % bias decreases as deviation from HWE increases in case of prospective approach. However, in case of PTE, % bias is considerably less as compared to the retrospective approach but higher than prospective approach due to the reasons described above. % gain in efficiency has been calculated using (8). The % gain in efficiency of the proposed PTE is in between % gain in efficiency of prospective and retrospective approach. However, under HWE condition, the % gain in efficiency in case of prospective approach is much higher as compared to the retrospective approach. However, as the deviation from HWE increase, the %gain in efficiency of retrospective approach is more as compared to prospective. It can also be noted that % gain in efficiency in case of all the estimators decrease with increase in minor allele frequency. % gain in efficiency of the proposed PTE is between % gain in efficiency of prospective and retrospective approach and under HWE condition.

Further, Simulation results with estimate, percent bias and gain in efficiency for \( \hat{\beta}_{aa} \) is given in Table 2.

<table>
<thead>
<tr>
<th>( \theta ) = 0</th>
<th>( f = 0.1 )</th>
<th>( f = 0.2 )</th>
<th>( f = 0.3 )</th>
<th>( f = 0.1 )</th>
<th>( f = 0.2 )</th>
<th>( f = 0.3 )</th>
<th>( f = 0.1 )</th>
<th>( f = 0.2 )</th>
<th>( f = 0.3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta = 0 )</td>
<td>( n_1=n_0=500 )</td>
<td>( n_1=n_0=1000 )</td>
<td>( n_1=n_0=500 )</td>
<td>( n_1=n_0=1000 )</td>
<td>( n_1=n_0=500 )</td>
<td>( n_1=n_0=1000 )</td>
<td>( n_1=n_0=500 )</td>
<td>( n_1=n_0=1000 )</td>
<td>( n_1=n_0=500 )</td>
</tr>
</tbody>
</table>
The disease-genotype odds ratios are assumed to follow a “multiplicative” pattern with $\psi_{Aa} = 1.4$, $\psi_{aa} = (1.4)^2$ ($\beta_{Aa} = 0.336472236$, $\beta_{aa} = 0.6729444$) and sample size is $n_1=n_0=500$ and $n_1=n_0=1000$. Results for % bias follows a similar pattern as described above. In case % gain in efficiency of the proposed PTE is in between % gain in efficiency of prospective and retrospective approach.

The Preliminary Test Estimator (PTE) developed can be seen as robust against the deviation from Hardy Weinberg Equilibrium. The percent bias is seen to be least at situation where HWE assumption is followed, as deviation from HWE increases, PTE is found to be less biased as compared to the retrospective methods. The % gain in efficiency also found in between the two existing methods which makes it a reliable and robust estimator at all situations. The PTE developed, due to its computational simplicity, advantages in bias; efficiency etc. will be very much helpful in genetic association studies. Further, this technique of utilizing the assumption-based and assumption-free methods in semi-parametric problems can be used in studies beyond genetic association studies which can lead to further research in this general area.

REFERENCES
Chen J and Chatterjee N. 2007. Exploiting Hardy–Weinberg equilibrium for efficient screening of single SNP associations

### Table 2  Simulation results with estimate, percent bias and gain in efficiency for $\hat{\beta}_{aa}$

<table>
<thead>
<tr>
<th></th>
<th>prospective</th>
<th></th>
<th></th>
<th>Retrospective</th>
<th></th>
<th></th>
<th>Preliminary test estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}_{aa}$</td>
<td>Estimate</td>
<td>percent bias</td>
<td>gain in efficiency</td>
<td>Estimate</td>
<td>percent bias</td>
<td>gain in efficiency</td>
</tr>
<tr>
<td>$\theta=0$</td>
<td>$f=0.1$</td>
<td>0.339</td>
<td>0.751</td>
<td>45.098</td>
<td>0.336</td>
<td>0.14</td>
<td>39.13</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.339</td>
<td>0.751</td>
<td>24.324</td>
<td>0.337</td>
<td>0.157</td>
<td>9.677</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.337</td>
<td>0.157</td>
<td>24.324</td>
<td>0.336</td>
<td>0.14</td>
<td>0</td>
</tr>
<tr>
<td>$\theta=0.5\log(1.2)$</td>
<td>$f=0.1$</td>
<td>0.34</td>
<td>1.048</td>
<td>45.098</td>
<td>0.32</td>
<td>4.896</td>
<td>39.13</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.336</td>
<td>0.14</td>
<td>24.324</td>
<td>0.298</td>
<td>11.434</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.336</td>
<td>0.14</td>
<td>24.324</td>
<td>0.282</td>
<td>18.189</td>
<td>12.5</td>
</tr>
<tr>
<td>$\theta=0.5\log(1.6)$</td>
<td>$f=0.1$</td>
<td>0.34</td>
<td>1.048</td>
<td>47.17</td>
<td>0.285</td>
<td>15.298</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.336</td>
<td>0.14</td>
<td>28.205</td>
<td>0.233</td>
<td>30.752</td>
<td>34.884</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.336</td>
<td>0.14</td>
<td>26.316</td>
<td>0.188</td>
<td>44.126</td>
<td>45.098</td>
</tr>
<tr>
<td>$\theta=0.5\log(2)$</td>
<td>$f=0.1$</td>
<td>0.34</td>
<td>1.048</td>
<td>48.148</td>
<td>0.318</td>
<td>5.49</td>
<td>39.13</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.337</td>
<td>0.157</td>
<td>26.32</td>
<td>0.299</td>
<td>11.137</td>
<td>17.65</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.336</td>
<td>0.14</td>
<td>26.32</td>
<td>0.281</td>
<td>16.486</td>
<td>17.65</td>
</tr>
<tr>
<td>$\theta=0$</td>
<td>$f=0.1$</td>
<td>0.34</td>
<td>1.048</td>
<td>44</td>
<td>0.339</td>
<td>0.751</td>
<td>36.36</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.338</td>
<td>0.454</td>
<td>22.22</td>
<td>0.337</td>
<td>0.157</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.338</td>
<td>0.454</td>
<td>22.22</td>
<td>0.337</td>
<td>0.157</td>
<td>0</td>
</tr>
<tr>
<td>$\theta=0.5\log(1.2)$</td>
<td>$f=0.1$</td>
<td>0.337</td>
<td>0.157</td>
<td>44</td>
<td>0.318</td>
<td>5.49</td>
<td>39.13</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.337</td>
<td>0.157</td>
<td>26.32</td>
<td>0.299</td>
<td>11.137</td>
<td>17.65</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.336</td>
<td>0.14</td>
<td>26.32</td>
<td>0.281</td>
<td>16.486</td>
<td>17.65</td>
</tr>
<tr>
<td>$\theta=0$</td>
<td>$f=0.1$</td>
<td>0.34</td>
<td>1.048</td>
<td>48.148</td>
<td>0.285</td>
<td>15.298</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>$f=0.2$</td>
<td>0.338</td>
<td>0.454</td>
<td>26.32</td>
<td>0.281</td>
<td>16.486</td>
<td>48.15</td>
</tr>
<tr>
<td></td>
<td>$f=0.3$</td>
<td>0.338</td>
<td>0.454</td>
<td>26.32</td>
<td>0.281</td>
<td>16.486</td>
<td>62.16</td>
</tr>
</tbody>
</table>