https://doi.org/10.56093/ijas.v90i3.101496

Indian Journal of Agricultural Sciences **90** (3): 601–5, March 2020/Article

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 n_{dg} 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 N_0 controls (D = 0) and N_1 cases (D = 1).

*Corresponding author e-mail: annmaria1986@gmail.com

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 $\hat{\beta}_{Aa}$ and $\hat{\beta}_{aa}$, 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 λ. 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:

$$\hat{a}^{Aa} = \begin{cases} \hat{a}_R^{Aa} & \text{if test rejects } H_0 : \hat{a}_P^{Aa} = \hat{a}_R^{Aa} \\ \hat{a}_P^{Aa} & \text{otherwise} \end{cases} \quad (1)$$

where $\hat{a}_R^{Aa}$, estimator based on retrospective approach; $\hat{a}_P^{Aa}$, estimator based on prospective approach.

The test statistic can be written as:

$$\ddot{e}^{Aa} = \frac{(\hat{\beta}_R^{Aa} - \hat{\beta}_P^{Aa})^2}{\hat{V}(\hat{\beta}_R^{Aa} - \hat{\beta}_P^{Aa})} \quad \square F_{(1,k)} \quad (2)$$

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

It can be seen that:

$$E(\hat{\beta}^{Aa}) = E(\hat{\beta}_P^{Aa} / \lambda^{Aa} < f_\alpha)\, P(\lambda^{Aa} < f_\alpha) + E(\hat{\beta}_R^{Aa} / \lambda^{Aa} > f_\alpha)\, P(\lambda^{Aa} > f_\alpha)$$

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

$$\hat{\beta}_P^{Aa} = \log \frac{n_{11} n_{00}}{n_{01} n_{10}} \quad \text{and} \quad \hat{\beta}_R^{Aa} = \log \frac{n_{11} \hat{n}_{00}^E}{\hat{n}_{01}^E n_{10}} \quad (3)$$

The bias of PTE $(\hat{\beta}^{Aa})$ can be obtained as:

$$bias(\hat{\beta}^{Aa}) = E(\hat{\beta}^{Aa}) - \beta^{Aa} = \sigma_u \sum_{j=0}^{k/2-1} \frac{1}{j!} \left(\frac{k}{2f_\alpha}\right)^j \left(\frac{f_\alpha}{f_\alpha + k}\right)^{j+1} \mu_{2j+1} \quad (4)$$

Mean square error of PTE $(\hat{\beta}^{Aa})$ can be

$$E(\hat{\beta}^{Aa^2}) = E\left[\left(\hat{\beta}_P^{Aa}\right)^2 / \lambda < f_\alpha\right] P(\lambda < f_\alpha) + E\left[\left(\hat{\beta}_R^{Aa}\right)^2 / \lambda > f_\alpha\right] P(\lambda > f_\alpha)$$

$$MSE(\hat{\beta}^{Aa}) = Var(\hat{\beta}^{Aa}) + Bias^2(\hat{\beta}^{Aa}) = E(\hat{\beta}^{Aa^2}) - [E(\hat{\beta}^{Aa})]^2 + Bias^2(\hat{\beta}^{Aa})$$

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

$$MSE(\hat{\beta}^{Aa}) = \sigma^2{}_v + \sigma^2{}_u \sum_{j=0}^{k/2-1} \frac{1}{j!} \left(\frac{k}{2f_\alpha}\right)^j \left(\frac{f_\alpha}{f_\alpha + k}\right)^{j+3/2} \mu_{2j+2}$$

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 (θ = 0), (ii) small deviation from HWE (θ = 0.5 log (1.2), (iii) moderate deviation from HWE (θ = 0.5 log (1.6) and (iv) large deviation from HWE (θ = 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.

$$\%\ gain\ in\ efficiency = \left| \frac{MSE(\beta_*^{Aa}{}_{(f,\theta)}) - MSE(\beta_R^{Aa}{}_{(f=0.3,\theta=0)})}{MSE(\beta_*^{Aa}{}_{(f,\theta)})} \right| \quad (7)$$

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

$$\%\ gain\ in\ efficiency = \left| \frac{MSE(\beta_*^{Aa}{}_{(f,\theta)}) - MSE(\beta_R^{Aa}{}_{(f=0.3,\theta=0)})}{MSE(\beta_*^{Aa}{}_{(f,\theta)})} \right| \quad (8)$$

where, $MSE(\beta_*^{Aa}{}_{(f,\theta)})$ is the Mean Square Error for the estimators, i.e prospective, retrospective and Preliminary Test Estimator ∀ f and θ and $MSE(\beta_R^{Aa}{}_{(f=0.3,\theta=0)})$ is Mean

Table 1 Simulation results with estimate, percent bias and gain in efficiency for $\hat{\beta}_{aa}$.

		Prospective			Retrospective			Preliminary test estimator		
		Estimate	Percent bias	Gain in efficiency	Estimate	Percent bias	Gain in efficiency	Estimate	Percent bias	Gain in efficiency
					$n_1=n_0=500$					
θ=0	f=0.1	-	-	-	-	-	-	-	-	-
	f=0.2	0.685	1.791	58.29	0.666	1.032	34.23	0.669	0.586	40.65
	f=0.3	0.681	1.197	17.98	0.672	0.14	0	0.674	0.157	2.67
θ=0.5log(1.2)	f=0.1	-			-					
	f=0.2	0.69	2.534	53.21	0.777	15.463	37.61	0.745	10.708	41.13
	f=0.3	0.68	1.048	10.98	0.743	10.41	3.95	0.719	6.844	5.19
θ=0.5log(1.6)	f=0.1	-			-					
	f=0.2	0.681	1.197	44.7	0.928	37.901	55.21	0.77	14.423	50.68
	f=0.3	0.678	0.751	1.35	0.842	25.122	25.51	0.717	6.547	10.98
θ=0.5log(2)	f=0.1	-			-					
	f=0.2	0.68	1.048	39.17	1.039	54.396	68.53	0.727	8.033	48.95
	f=0.3	0.675	0.305	-5.8	0.91	35.227	42.06	0.685	1.791	0
					$n_1=n_0=1000$					
θ=0	f=0.1	0.706	4.912	89.944	0.649	3.558	77.5	0.653	2.964	81.91
	f=0.2	0.679	0.9	57.647	0.669	0.586	34.545	0.671	0.289	40.984
	f=0.3	0.675	0.305	18.182	0.673	0.008	0	0.674	0.157	5.263
θ=0.5log(1.2)	f=0.1	0.697	3.575	88	0.801	19.029	77.215	0.755	12.194	80.645
	f=0.2	0.68	1.048	52.632	0.777	15.463	42.857	0.738	9.667	46.269
	f=0.3	0.675	0.305	10	0.742	10.262	10	0.71	5.506	10
θ=0.5log(1.6)	f=0.1	0.691	2.683	84.549	1.018	51.275	85.185	0.824	22.447	84.681
	f=0.2	0.679	0.9	45.455	0.931	38.347	68.966	0.721	7.141	54.43
	f=0.3	0.676	0.454	2.703	0.843	25.27	43.75	0.685	1.791	10
θ=0.5log(2)	f=0.1	0.686	1.94	81.633	1.181	75.497	90.323	0.784	16.503	85.246
	f=0.2	0.675	0.305	38.983	1.041	54.693	80.328	0.682	1.346	41.935
	f=0.3	0.677	0.603	-5.882	0.914	35.821	61.29	0.678	0.751	-5.882

The odds ratios are assumed to follow a "recessive" pattern with ψAa = 1, ψaa = (1.4)2 (βAa=0, βaa= 0.6729444)

Square Error for the retrospective estimator at f=0.3 and θ=0, which has the minimum MSE among all the other estimators.

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

		prospective			Retrospective			Preliminary test estimator		
		Estimate	percent bias	gain in efficiency	Estimate	percent bias	gain in efficiency	Estimate	percent bias	gain in efficiency
					$n_1=n_0=500$					
θ=0	f=0.1	0.339	0.751	45.098	0.336	0.14	39.13	0.338	0.454	39.13
	f=0.2	0.339	0.751	24.324	0.337	0.157	9.677	0.337	0.157	9.677
	f=0.3	0.337	0.157	24.324	0.336	0.14	0	0.337	0.157	6.667
θ=0.5log(1.2)	f=0.1	0.34	1.048	45.098	0.32	4.896	39.13	0.328	2.518	41.667
	f=0.2	0.336	0.14	24.324	0.298	11.434	12.5	0.313	6.976	15.152
	f=0.3	0.336	0.14	24.324	0.282	16.189	12.5	0.302	10.245	15.152
θ=0.5log(1.6)	f=0.1	0.34	1.048	47.17	0.285	15.298	44	0.314	6.679	45.098
	f=0.2	0.338	0.454	28.205	0.233	30.752	34.884	0.304	9.651	31.707
	f=0.3	0.338	0.454	26.316	0.188	44.126	45.098	0.306	9.056	34.884
θ=0.5log(2)	f=0.1	0.337	0.157	48.148	0.249	25.997	49.091	0.309	8.165	49.091
	f=0.2	0.337	0.157	30	0.176	47.693	51.724	0.318	5.49	34.884
	f=0.3	0.339	0.751	28.205	0.113	66.416	65	0.331	1.626	31.707
					$n_1=n_0=1000$					
θ=0	f=0.1	0.34	1.048	44	0.339	0.751	36.36	0.339	0.751	39.13
	f=0.2	0.338	0.454	22.22	0.337	0.157	75	0.338	0.454	6.67
	f=0.3	0.338	0.454	22.22	0.337	0.157	0	0.337	0.157	6.67
θ=0.5log(1.2)	f=0.1	0.337	0.157	44	0.318	5.49	39.13	0.326	3.112	41.67
	f=0.2	0.337	0.157	26.32	0.299	11.137	17.65	0.316	6.084	22.22
	f=0.3	0.336	0.14	22.22	0.281	16.486	17.65	0.307	8.759	22.22
θ=0.5log(1.6)	f=0.1	0.335	0.438	46.15	0.28	16.784	48.15	0.315	6.382	46.15
	f=0.2	0.337	0.157	26.32	0.232	31.049	48.15	0.321	4.598	33.33
	f=0.3	0.337	0.157	26.32	0.187	44.423	62.16	0.329	2.221	30

The odds ratios are assumed to follow a "multiplicative" pattern with ψAa = 1.4, ψaa = (1.4)2 (βAa= 0.336472236, βaa= 0.6729444) (Aa)

The disease-genotype odds ratios are assumed to follow a "multiplicative" pattern with $\psi_{Aa} = 1.4$, $\psi_{aa} = (1.4)^2$ (β_{Aa}= 0.336472236, β_{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

Andersen E B. 1970. Asymptotic Properties of Conditional Maximum-Likelihood Estimators. *Journal of Royal Statistical Society Series B* **32**: 283–01.

Chatterjee N, Chen Y, Luo S and Carroll R J. 2009. Analysis of Case-Control Association Studies: SNPs, Imputation and Haplotypes. *Statistical Science* **24**(4): 489–502.

Chatterjee N and Carroll R J. 2005. Semiparametric Maximum Likelihood Estimation in Case-Control Studies of Gene-Environment Interactions. *Biometrika* **92**: 399–418.

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

from case-control studies. *Human Heredity* **63**: 196–204.

Chen Y H, Chatterjee N and Carroll R J. 2009. Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *Journal of American Statistical Association* **104**: 220–33.

Das G and Bez K. 1995. Preliminary Test Estimators in Double Sampling with Two Auxiliary Variables. *Communication in Statistics: Theory and Methods* **24(**5): 1211–26.

Epstein M P and Satten G A. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. *American Journal of Human Genetics* **73**: 1316–29.

Han C P and Bancroft T A. 1968. On Pooling Means when Variance is Unknown. *Journal of American Statistical Association* **62**: 1333–42.

Luo S, Mukherjee B, Chen J and Chatterjee N. 2009. Shrinkage estimation for robust and efficient screening of single-SNP association from case-control genome-wide association studies. *Genetic Epidemiology* **33**(8): 740–50.

Mukherjee B and Chatterjee N. 2008. Exploiting gene–environment independence for analysis of case-control studies: An empirical Bayes approach to trade-off between bias and efficiency. *Biometrics* **64**: 685–94.

Prentice R L and Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* **66**: 403–12.

Rai A and Srivastava A K. 1998. Estimation of regression coefficient from survey data based on test of significance. *Communications in Statistics - Theory and Methods* **27**(3): 761–73.

Satten G A and Epstein M P. 2004. Comparison of Prospective and Retrospective Methods for Haplotype Inference in Case-Control Studies. *Genetic Epidemiology* **27**: 192–201.

Spinka C, Carroll R J and Chatterjee N. 2005. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genetic Epidemiology* **29**: 108–27.

Sukhatme B V and Tang V K T. 1975. Allocation in Stratified Sampling Subsequent to Preliminary Test of Significance. *Journal of the American Statistical Association* **70**: 176–79.