University of Pennsylvania

UPenn Biostatistics Working Papers

Year 2006 *Paper* 14

Conditional Likelihood Methods for Haplotype-based Association Analysis Using Matched Case-Control Data

Jinbo Chen[*] Carmen Rodriguez[†]

[*]

[†]

Conditional Likelihood Methods for Haplotype-based Association Analysis Using Matched Case-Control Data

Jinbo Chen and Carmen Rodriguez

Abstract

Genetic epidemiologists routinely assess disease susceptibility in relation to haplotypes, i.e., combinations of alleles on a single chromosome. We study statistical methods for inferring haplotype-related disease risk using SNP genotype data from matched case-control studies, where controls are individually matched to cases on some selected factors. Assuming a logistic regression model for haplotype-disease association, we propose two conditional likelihood approaches that address the issue that haplotypes cannot be inferred with certainty from SNP genotype data (phase ambiquity). One approach is based on the likelihood of disease status conditioned on the total number of cases, genotypes, and other covariates within each matching stratum, and the other is based on the joint likelihood of disease status and genotypes conditioned only on the total number of cases and other covariates. The joint-liklihood approach is generally more efficient, particularly for assessing haplotype-environment interactions. Simulation studies demonstrated that the first approach was more robust to model assumptions on the the diplotype distribution conditioned on environmental risk variables and matching factors in the control population. We applied the two methods to analyze a matched case-control study of prostate cancer.

Conditional Likelihood Methods for Haplotype-based Association Analysis Using Matched Case-Control Data

Jinbo Chen[1,*] and Carmen Rodriguez[2]

[1]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd, Rockville, MD 20852

[2]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Georgia 30329-4251

*Email: chenjin@mail.nih.gov

SUMMARY. Genetic epidemiologists routinely assess disease susceptibility in relation to haplotypes, i.e., combinations of alleles on a single chromosome. We study statistical methods for inferring haplotype-related disease risk using SNP genotype data from matched case-control studies, where controls are individually matched to cases on some selected factors. Assuming a logistic regression model for haplotype-disease association, we propose two conditional likelihood approaches that address the issue that haplotypes cannot be inferred with certainty from SNP genotype data (phase ambiguity). One approach is based on the likelihood of disease status conditioned on the total number of cases, genotypes, and other covariates within each matching stratum, and the other is based on the joint likelihood of disease status and genotypes conditioned only on the total number of cases and other covariates. The joint-likelihood approach is generally more efficient, particularly for assessing haplotype-environment interactions. Simulation studies demonstrated that the first approach was more robust to model assumptions on the diplotype distribution conditioned on environmental risk variables and matching factors

keywords: conditional likelihood; logistic regression; matched case-control study; phase ambiguity; unphased genotype data.

in the control population. We applied the two methods to analyze a matched case-control study of prostate cancer.

1. Introduction

To assess the association between a disease outcome and a candidate gene using genotype data for single nucleotide polymorphisms (SNPs), it has been generally agreed that the disease risk in relation to the set of alleles on each chromosome (haplotype), should be examined (Risch and Merikangas, 1996; Botstein and Risch, 2003; Schaid, 2004). Haplotype-based association analysis has been proven to be useful when alleles on the same chromosome interact with each other (cis interaction) or for fine mapping of disease causing loci (Schaid, 2004). When studies involve unrelated individuals, however, the haplotypes may not be inferred with certainty from the multi-locus SNP genotype data. This phenomenon is usually referred to as phase ambiguity. To address this issue, various statistical methods have been developed for estimating haplotype-related odds ratio (OR) parameters from retrospective case-control studies (Schaid et al., 2002; Epstein and Satten, 2003; Stram et al., 2003; Lin and Zeng, 2005; Spinka, Carroll, and Chatterjee, 2005), and for estimating haplotype-related relative-hazard parameters from cohort studies (Lin, 2004; Chen and Chatterjee, 2006) or from individually matched case-control studies nested within a cohort (Chen et al., 2004; Chen and Chatterjee, 2006). In this article, we consider a similar problem for the analysis of matched case-control studies: we study methods for estimating haplotype-related OR parameters accounting for phase ambiguity and the case-control matching.

The matched case-control design is often adopted for epidemiologic studies to efficiently control for variables that confound environmental factors. Because genetic epidemiology studies investigate joint gene-environment effects or gene-environment interactions as the norm rather than the exception, matched case-control design is also very useful for many genetic epidemiologic investigation. To implement this design, the population under study is classified into strata by disease status and values of some confounding factors chosen *a priori*. Then n cases and m controls are randomly sampled from each stratum, and n and m are allowed to vary across different strata. The analysis model of choice for this design is usually the logistic regression model, and the OR parameters are estimated by maximizing the standard conditional likelihood. For estimating haplotype-related ORs using genotype data, novel statistical methods are needed to address phase ambiguity. When n and m are large, methods for unmatched case-control studies could yield approximately valid results when stratum-specific intercept parameters are included in the logistic regression model (Breslow and Day, 1982). When n and m are small or the number of matching strata is large, the conditional analysis would be preferred. In fact, whenever possible, it would always be preferred to perform conditional analysis (Breslow and Day, 1982). Kraft et al. (2005) compared several *ad hoc* approaches, including a standard conditional likelihood approach but replacing haplotype-specific covariates with their expected values conditioned on the genotype data. Because the expected values are usually not the same as the true values, this approach usually yields biased OR estimates.

In this article, we propose two novel conditional likelihood methods for consistent estimation of haplotype and haplotype-environment interaction effects using unphased genotype data from matched case-control studies. We introduce the notation and model in Section 2. In Section 3, we present the two approaches and show how they are used for assessing SNP-disease association. In Section 4, we apply the proposed methods to the analysis of data from a collaborative study between the National Cancer Institute (NCI) and American Cancer Society (ACS) on the genetic susceptibility of prostate cancer. In a scenario similar to this real study, we perform simulation studies to assess the consistency and efficiency of the two methods, to examine the consistency of their asymptotic variance estimators, and to evaluate the robustness of the two methods to critical assumptions. Details of these simulation studies are given in Section 5. We make some final remarks in Section 6.

2. Notation and Model Specification

We use $\mathcal{H} = \{\ldots, h_a, \ldots, h_b, \ldots\}$ to denote the set of all possible haplotypes within a suitable chromosome segment (e.g., haplotype block) chosen *a priori*. Let $H = (h_a, h_b)$ denote the haplotype pair that an individual carries in his/her pair of homologous chromosomes (diplotype). The diplotype data H is usually not directly observable. Instead, the multi-locus genotype data G, which record the pair of alleles a subject carries on the pair of homologous chromosomes at each locus, are observed. To discern H from G, one needs to know the arrangement of alleles, i.e., the phase information, along each individual chromosome, but such information is not contained in G. Consequently, the same genotype data G could be compatible with multiple diplotypes. We denote $\mathcal{H}_G$ to be the set of all possible diplotypes consistent with the genotype data G. Let D denote the binary case-control status, S indicate the sampling stratum defined by matching factors, and Z be a vector of covariates. We assume that the disease risk (penetrance) given H, Z, and S is described by the logistic regression model

$$\mathrm{p}(D = 1|H, Z, S) = \left[1 + \exp\{-\alpha_s - \beta_h X^H - \beta_z Z - \gamma k(Z, H)\}\right]^{-1}, \quad (1)$$

where $\{e^{\alpha_s}, s = 2, \ldots, S\}$ are stratum-specific OR parameters in reference to a baseline stratum, e^{β_h} is a vector of haplotype-specific OR parameters in reference to a chosen baseline haplotype, and $k(Z, H)$ is an interaction term between Z and H. Let δ_{ij} be a function taking value one if $i = j$ and zero otherwise. Depending on the numerical coding of X^H, one can fit for a haplotype h^* a multiplicative model ($X^H = \delta_{h_a h^*} + \delta_{h_b h^*}$), a dominant model ($X^H = \delta_{h_a h^*} + \delta_{h_b h^*} - \delta_{h_a h^*}\delta_{h_b h^*}$), or a recessive model ($X^H = \delta_{h_a h^*}\delta_{h_b h^*}$) (e.g., Wallenstein, Hodge, and Weston, 1998). If, for example, four haplotypes (h_1, h_2, h_3, h_4) are present in the study population, then to fit a dominant model, one could construct $X^H = X_{h_2}$ where X_{h_2} indicates whether a subject has a copy of haplotype h_2 or not (0 or 1).

Let $\mathbf{D}_s = (D_{s1}, \ldots, D_{sJ_s})$ be the vector of disease indicators for subjects $j = 1, \ldots, J_s$ in the s^{th} matched set, and $\mathbf{D}_s^*$ be a permutation of $\mathbf{D}_s$. We similarly define $\mathbf{Z}_s$, $\mathbf{G}_s$, and $\mathbf{H}_s$. Let n_{1s} be the number of cases within the s^{th} stratum. If the diplotype H were observed for all subjects, then $\beta \equiv (\beta_h, \beta_z, \gamma)$ could be estimated by maximizing the standard conditional likelihood $L^f = \prod_s p(\mathbf{D}_s|n_{1s}, \mathbf{H}_s, \mathbf{Z}_s)$:

$$L^f = \prod_{s=1}^{S} \frac{\prod_{j:D_{sj}=1} e^{\beta_h X^{H_{sj}} + \beta_z Z_{sj} + \gamma k(Z_{sj}, H_{sj})}}{\sum_{\mathbf{D}_s^*} \prod_{l:D_{sl}^*=1} e^{\beta_h X^{H_{sl}} + \beta_z Z_{sl} + \gamma k(Z_{sl}, H_{sl})}}.$$

An attractive feature of this likelihood is that nuisance parameters α_s fall out because the likelihood is conditioned on their sufficient statistic $\{n_{1s}\}$. Consequently, parameters $\{\alpha_s, s = 2, \ldots, S\}$ are not involved in the inference of β.

3. Semiparametric Approaches for the Estimation of OR Parameters

In this section, we present two novel conditional likelihood approaches for the estimation of β. In the absence of phase information, we consider a novel conditional likelihood based on the observed genotype data G, $\prod_s p(\mathbf{D}_s, \mathbf{G}_s|n_{1s}, \mathbf{Z}_s, S)$, which is a function of β and the conditional diplotype distribution $p(H|D, Z, S)$. We factorize this likelihood as

$$\left\{\prod_s p\left(\mathbf{G}_s|\mathbf{D}_s, \mathbf{Z}_s, S\right)\right\} \left\{\prod_s p\left(\mathbf{D}_s|n_{1s}, \mathbf{Z}_s, S\right)\right\}. \tag{2}$$

The first method we propose is based only on the first part $\prod_s p\left(\mathbf{G}_s|\mathbf{D}_s, \mathbf{Z}_s, S\right)$, and the second approach utilize both parts. Both methods involve the unknown nuisance distribution $p(H|D, Z, S)$, which is non-identifiable from SNP genotype data without any model assumptions (Epstein and Satten, 2003). We thus impose two restrictions to the distribution of H in the control population. The first one is that the haplotype distribution varies only with a summary variable of Z and S, A_{ZS}, that takes a small fixed number of values. This assumption can be formally expressed as $p(h_a|D = 0, Z, S) = p(h_a|D = 0, A_{ZS}) \equiv f_a^0(A_{ZS})$ and $p(H|D = 0, Z, S) = p(H|D = 0, A_{ZS}) \equiv p^0(H|A_{ZS})$. We use subscript "$ZS$" to indicate that A could be a function of both Z and S. For example, if cases and controls are matched on ethnicity and family history is included in Z as a risk factor, then A_{ZS} could indicate strata defined jointly by ethnicity and family history. We assume that these control population substrata are defined before applying the proposed methods, possibly by examining differences in haplotype frequency estimates among different candidate substrata. The second restriction we impose is that the distribution of diplotype H follows Hardy-Weinberg equilibrium in each control sub-population defined by A_{ZS}. That is, within each control sub-group defined by A_{ZS}, we have $p^0\{H = (h_a, h_b)|A_{ZS}\} = f_a^0(A_{ZS})f_b^0(A_{ZS})$ if $a = b$ and $p^0\{H = (h_a, h_b)|A_{ZS}\} = 2f_a^0(A_{ZS})f_b^0(A_{ZS})$ if $a \neq b$. Let $\mathbf{f}^0(A_{ZS})$ be the vector of haplotype frequencies within stratum A_{ZS} and $\mathbf{f}^0$ be the pool of vectors $\mathbf{f}^0(A_{ZS})$.

Under the two assumptions above, the full conditional likelihood (2) is a function of unknown parameters β and $\mathbf{f}^0$. A useful result owing to the first assumption is that haplotype frequencies for cases are free of α_s and depend on S only through A_{ZS}, following a similar result in Epstein and Satten (2003):

$$p(H|D = 1, Z, S) = \frac{e^{\beta_h X^H + \gamma k(Z,H)} p^0(H|A_{ZS})}{\sum_G \sum_{H \in \mathcal{H}_G} e^{\beta_h X^H + \gamma k(Z,H)} p^0(H|A_{ZS})} \equiv p(H|D = 1, Z, A_{ZS}).$$

Here we also make a natural assumption that the support of genotype G is the collection of all unique genotypes observed in cases and controls.

3.1 *Method I: an Estimated Conditional Likelihood Approach*

The first part of the factorization in (2), $\prod_s p(\mathbf{D}_s|\mathbf{G}_s, n_{1s}, \mathbf{Z}_s, S)$, which we denote as $L^I(\beta, \mathbf{f}^0)$, is an appropriate conditional likelihood. We propose to estimate β parameters based on L^I. The disease odds associated with (G, Z, S), similar as a result in Epstein and Satten (2003), is

$$\widetilde{\theta}(G, Z, S) = \frac{p(Y = 1|G, Z, S)}{p(Y = 0|G, Z, S)} = e^{\alpha_s + \beta_z Z} \sum_{H \in \mathcal{H}_G} e^{\beta_h X^H + \gamma k(Z,H)} p(H|D = 0, G, A_{ZS}).$$

The corresponding conditional likelihood L^I is then

$$L^I\{\beta, \mathbf{f}^0\} = \prod_{s=1}^{S} \frac{\prod_{j:D_{sj}=1} e^{\beta_z Z_{sj}} \sum_{H\in\mathcal{H}_{G_{sj}}} e^{\beta_h X^H+\gamma k(Z_{sj},H)} p(H|D=0, G_{sj}, A_{Z_{sj}s})}{\sum_{\mathbf{D_s^*}} \prod_{l:D^*_{sl}=1} e^{\beta_z Z_{sl}} \sum_{H\in\mathcal{H}_{G_{sl}}} e^{\beta_h X^H+\gamma k(Z_{sl},H)} p(H|D=0, G_{sl}, A_{Z_{sl}s})}, \quad (3)$$

which has a similar form as standard conditional likelihood L^f but with $e^{\beta X^H+\gamma k(Z,H)}$ replaced by its expectation with respect to the distribution $p^0(H|G, A_{ZS})$. We note that the stratum-specific intercept α_s falls out of L^I, as in the standard conditional likelihood analysis. The conditional likelihood L^I involves haplotype frequencies in controls, $\mathbf{f}^0$, but it does not contain information for making inference on $\mathbf{f}^0$. This is because the genotype data G, which contain the information for inferring haplotype frequencies, are conditioned out. This is analogous to the fact that the conditional likelihood function L^f cannot be used for inferring parameters of the covariate distribution. We propose to estimate $\mathbf{f}^0$ by applying the EM algorithm (Excoffier and Slatkin, 1995) to the genotype data for controls within each substrata A_{ZS}. We then estimate β by maximizing L^I with respect to β with $p^0(H|A_{ZS})$ fixed at its estimate. That is, the resultant $\hat{\beta}$ maximizes the "estimated" likelihood $L^I\{\beta, \hat{\mathbf{f}}^0\}$. When there is only one case in each matched set indexed by $s1$ and the haplotype distribution in controls is independent of Z and S, the conditional likelihood (3) reduces to

$$L^{I1} = \prod_{s=1}^{S} \frac{e^{\beta_z Z_{s1}} \sum_{H\in\mathcal{H}_{G_{s1}}} e^{\beta_h X^H+\gamma K(Z_{s1},H)} p(H|D=0, G_{s1})}{\sum_{l=0,1} e^{\beta_z Z_{sl}} \sum_{H\in\mathcal{H}_{G_{sl}}} e^{\beta_h X^H+\gamma K(Z_{sl},H)} p(H|D=0, G_{sl})}.$$

This reduced likelihood L^{I1} takes the same form as the partial likelihood function proposed in Chen and Chatterjee (2006), where they studied methods for haplotype analysis of nested case-control studies assuming a Cox proportional hazards model for disease penetrance. The asymptotic distribution of $\hat{\beta}$ is derived in Appendix (A3). The asymptotic variance-covariance matrix, not surprisingly, takes a similar form as that in Chen and Chatterjee (2006). Specifically, under the null hypothesis that the disease risk is not related to any diplotype $H \in \mathcal{H}$, one can perform the score test using the approach of Chen et al. (2004).

3.2 *Method II: a Full Conditional Likelihood Approach*

We denote the full conditional likelihood (2) as L^{II}, which we show is free of the stratum-specific odds ratio parameter α_s and only involves β and $\mathbf{f}^0$. Because subjects within each stratum defined by (D, S) are independent, $\prod_s p\left(\mathbf{G}_s|\mathbf{D}_s, \mathbf{Z}_s, S\right)$ can be factorized as $\prod_{l:D_{sl}=0} p(G_{sl}|D_{sl} = 0, A_{Z_{sl}s}) \prod_{r:D_{sr}=1} p(G_{sr}|D_{sr} = 1, Z_{sr}, A_{Z_{sr}s})$. Thus,

$$\prod_s p\left(\mathbf{G}_s|\mathbf{D}_s, \mathbf{Z}_s, S\right) = \left\{\prod_{l:D_{sl}=0} \sum_{H\in\mathcal{H}_{G_{sl}}} p^0(H|A_{Z_{sl}s})\right\}\left\{\prod_{r:D_{sr}=1} \sum_{H\in\mathcal{H}_{G_{sr}}} p(H|D_{sr} = 1, Z_{sr}, A_{Z_{sr}s})\right\},$$

which is free of α_s and depends only on β and $\mathbf{f}^0$. Furthermore, as is shown in Appendix (A1),

$$p\left(\mathbf{D}_s|n_{1s}, \mathbf{Z}_s, S\right) = \frac{\prod_{j:D_{sj}=1} \theta(Z_{sj}, A_{Z_{sj}s})}{\sum_{\mathbf{D}_s^*} \prod_{l:D_{sl}^*=1} \theta(Z_{sl}, A_{Z_{sl}s})}$$

where $\theta(Z, A_{ZS}) = e^{\beta_z Z} \sum_G \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H + \gamma k(Z,H)} p^0(H|A_{ZS})$. Thus, $L^{II}(\beta, \mathbf{f}^0)$ is free of α_s and is a function only of OR parameters β and haplotype frequencies $\mathbf{f}^0$ in controls. Estimates of $(\beta, \mathbf{f}^0)$ can be obtained by maximizing $\log L^{II}(\beta, \mathbf{f}^0)$.

When there is no interaction effect between Z and H ($\gamma = 0$) and the haplotype distribution in controls is independent of Z and S ($p\{H|D = 0, Z, S\} = p\{H|D = 0\}$), careful examination of $L^{II}(\beta, \mathbf{f}^0)$ reveals that the information for β_z is fully contained in $p(\mathbf{D}_s|n_{1s}, \mathbf{Z}_s, S)$, which reduces to an ordinary conditional likelihood for the effect of Z as if H were not associated with the disease risk:

$$p(\mathbf{D}_s|n_{1s}, Z_s, S) = \prod_{s=1}^{S} \frac{\prod_{j:D_{sj}=1} e^{\beta_z Z_{sj}}}{\sum_{\mathbf{D}_s^*} \prod_{l:D_{sl}^*=1} e^{\beta_z Z_{sl}}}.$$

Information for β_h and $\mathbf{f}^0$ is then fully contained in the part of the likelihood $p(\mathbf{G}_s|\mathbf{D}_s, \mathbf{Z}_s, S)$, which takes the same form as the retrospective likelihood (Epstein and Satten, 2003) for an "unmatched" case-control study. The implication for this observation is that one could perform "unconditional case-control analysis" if he/she is only interested in assessing main haplotype effects and willing to entertain the assumption that haplotype frequencies do not vary with environmental risk factors and matching variables. This result can be intuitively explained as follows. If no Z were involved, this assumption simply states that the stratum factor S is not related to haplotype frequencies in cases or controls. Consequently, the most efficient analysis would be the unconditional analysis breaking the matching. When variables Z and H are independent in controls and do not interact with each other for determining the disease risk, any difference in haplotype frequencies between cases and controls is not informative of the difference in the distributions of Z between cases and controls. Consequently, the odds ratio for Z could be estimated as if no genetic effects were present. The "conditional" analysis using, for example, one of our proposed methods to explicitly account for matching is necessary when it is of interest to assess haplotype-environment interaction or when the haplotype distribution in controls depends on Z or S.

To compute the estimates of $\{\beta, \mathbf{f}^0\}$ and obtain the asymptotic variance-covariance matrix, one can adopt the Newton-Raphson algorithm. Because we observed that the Newton-Raphson algorithm may fail to converge in the presence of rare haplotypes, we propose to use a one-step approximation to the Newton-Raphson solution (one-step estimator). This one-step estimation method leads to estimates with the same asymptotic property as the Newton-Raphson estimates (Bickel et al., 1993). Following Epstein and Satten (2003), we re-parametrize $f_a(A_{ZS})$ using parameters $\alpha_a(A_{ZS})$ with $f_a = e^{\alpha_a}/\{1 + \sum_{j=1}^{J-1} e^{\alpha_j}\}$, where J is the total number of haplotypes in the A_{ZS} stratum. Let α be the collection of all $\alpha_a(A_{ZS})$. We denote the score functions for β and α, $\{\partial \log L^{II}(\beta, \alpha)/\partial\beta, \partial \log L^{II}(\beta, \alpha)/\partial\alpha\}$, as $\{\dot{l}_\beta^{II}(\beta, \alpha), \dot{l}_\alpha^{II}(\beta, \alpha)\}$. The corresponding information matrix $I(\beta, \alpha)$ is composed of sub-matrices $I_{\beta\beta} = -\partial \dot{l}_\beta^{II}/\partial\beta$, $I_{\beta\alpha} = -\partial \dot{l}_\beta^{II}/\partial\alpha$, and $I_{\alpha\alpha} = -\partial \dot{l}_\alpha^{II}/\partial\alpha$. Denote estimates from method I as $(\hat{\beta}^I, \hat{\alpha}^I)$. The one-step estimates, denoted as $(\hat{\beta}^{II}, \hat{\alpha}^{II})$, are obtained as

$$\begin{pmatrix} \hat{\beta}^{II} \\ \hat{\alpha}^{II} \end{pmatrix} \approx \begin{pmatrix} \hat{\beta}^{I} \\ \hat{\alpha}^{I} \end{pmatrix} + I^{-1}\left(\hat{\beta}^I, \hat{\alpha}^I\right) \left\{ \begin{array}{c} \dot{l}_\beta^{II}\left(\hat{\beta}^I, \hat{\alpha}^I\right) \\ \dot{l}_\alpha^{II}\left(\hat{\beta}^I, \hat{\alpha}^I\right) \end{array} \right\}.$$

We could obtain $\hat{\beta}^{II}$ and $\hat{\alpha}^{II}$ separately to avoid the difficulty in inverting the full information matrix $I(\hat{\beta}^I, \hat{\alpha}^I)$. Define $\hat{I}_{\beta\beta} = I_{\beta\beta}(\hat{\beta}^I, \hat{\alpha}^I)$ and define $\hat{I}_{\beta\alpha}$ and $\hat{I}_{\alpha\alpha}$ similarly. Then $\hat{\beta}^{II}$ and $\hat{\alpha}^{II}$

can be obtained as

$$\hat{\beta}^{II} = \hat{\beta}^{I} + \left(\hat{I}_{\beta\beta} - \hat{I}_{\beta\alpha}\hat{I}^{-1}_{\alpha\alpha}\hat{I}_{\alpha\beta}\right)^{-1}\left(\hat{\dot{l}}^{II}_{\beta} - \hat{I}_{\beta\alpha}\hat{I}^{-1}_{\alpha\alpha}\hat{\dot{l}}^{II}_{\alpha}\right)$$

and

$$\hat{\alpha}^{II} = \hat{\alpha}^{I} + \left(\hat{I}_{\alpha\alpha} - \hat{I}_{\alpha\beta}\hat{I}^{-1}_{\beta\beta}\hat{I}_{\beta\alpha}\right)^{-1}\left(\hat{\dot{l}}^{II}_{\alpha} - \hat{I}_{\alpha\beta}\hat{I}^{-1}_{\beta\beta}\hat{\dot{l}}^{II}_{\beta}\right),$$

respectively. The key to the implementation of the one-step estimator is the availability of the consistent initial estimates $(\hat{\beta}^I, \hat{\alpha}^I)$, which are plugged into $(\dot{l}^{II}_{\beta}, \dot{l}^{II}_{\alpha})$ and $I(\beta, \alpha)$ to obtain their consistent estimates. In the Appendix (A4), for the one-one matched case, we give analytical formulas for score functions and the information matrix, which turn out to be remarkably simple.

3.3 *Application of Method II to Single SNP Analysis*

The two conditional likelihood approaches presented above can be applied to the analysis involving only one-locus genotype, where each allele is a "single-locus haplotype." Of course no phase ambiguity is involved in this scenario. Let a and A indicate the two alleles, and ρ be the frequency of minor allele a. Under the HWE assumption and gene-environment independence $\rho = p(a|Z, S)$, the frequencies of genotypes AA, Aa, and aa in controls are $(1-\rho)^2$, $\rho(1-\rho)$, and ρ^2, respectively. Let G indicate genotype AA, Aa, or aa, and let X^G indicate the presence or absence of minor allele a in G. That is, X^G takes value one if $G = Aa$ or $G = aa$ and zero otherwise. Certainly X^G could use other numerical codings as well. It is easy to see that method I reduces to the ordinary conditional likelihood analysis. For method II, the log likelihood $\log L^{II}$ for the one-one matched design can be simplified as

$$\begin{aligned}\log L^{II} \;=\; & \sum_{s=1}^{S}\sum_{i=0,1} D_{si}\{\beta_z Z_{si} + \beta_h X^{G_{si}} + \gamma k(Z_{si}, G_{si})\} \\ & + \{2(n_{00}+n_{10}) + (n_{01}+n_{11})\}\log(1-\rho) + \{2(n_{02}+n_{12}) + (n_{01}+n_{11})\}log\rho \\ & - \sum_{s=1}^{S}\log\left[\sum_{i=0,1} e^{\beta_z Z_{si}}\left\{(1-\rho)^2 + e^{\beta_h + \gamma Z_{si}}(2\rho - \rho^2)\right\}\right],\end{aligned}$$

where (n_{00}, n_{01}, n_{02}) and (n_{10}, n_{11}, n_{12}) are counts of genotype AA, Aa, and aa in controls and cases, respectively. Parameter estimates can then be obtained by maximizing $\log L^{II}$ jointly over $(\beta_z, \beta_h, \gamma, \rho)$ or by applying the one-step approximation.

4. Analysis of Data from a Prostate Cancer Study

Researchers at the National Cancer Institute (NCI) and the American Cancer Society (ACS) are collaborating to investigate the role of insulin resistance and chronic inflammation in the development of prostate cancer using the ACS Cancer Prevention Study-II nutrition cohort (Calle, 2002). A nested case-control study of 1,209 prostate cancer cases and an equal number of controls has been undertaken to evaluate the relationship between prostate cancer risk and a number of key genes involved in the insulin signaling and chronic inflammation pathways. One control was matched to a case on ethnicity, age within 6 months, and date of blood collection. Because 97.0% of the study subjects were Caucasians, and because we did not have convenient access to the ethnicity status due to data security concerns, we included the small number of non-Caucasians in the current analysis.

We assessed whether the gene coding for tumor necrosis factor α (TNF-α) was associated with prostate cancer risk. The current analysis focused on demonstrating the proposed methods, and scientific results on genetic epidemiology of prostate cancer in relation to this gene will be presented in future manuscripts. TNF-α is a protein produced by macrophages in the presence of an endotoxin, and it has been shown to contribute to the progression of several cancers and thus may play a role in prostate cancer progression. In this collaborative study, 5 SNPs were genotyped in the TNF-α gene. First, we assessed the association between each SNP and prostate cancer risk by examining the OR associated with the presence of a variant allele, using both the standard conditional likelihood method (method I) and method II. The two methods yielded similar ORs and confidence intervals. In particular, the presence of a minor allele at one locus, which we named as TNF1, appeared to be significantly associated with lower risk

(OR[95% CI]: 0.776[0.616, 0.980] and 0.761[0.609, 0.950] by methods I and II, respectively). In addition, the estimates of minor allele frequencies from method II were also very similar to those obtained by simply calculating the proportion of minor alleles in controls.

We then investigated the joint effect of SNP TNF1 and haplotypes formed by the other four SNPs (Table 1). In particular, let X^h indicate the presence ($X^h = 1$) or absence ($X^h = 0$) of haplotype h, and let X^1 indicate the presence ($X^1 = 1$) or absence ($X^1 = 0$) of a minor allele at locus TNF1. We applied methods I and II to fit the following model:

$$\mathrm{p}(D = 1|X^h, X^1, S) = \left[1 + \exp\left\{-\alpha_s - \beta_h X^h - \beta_z X^1 - \gamma X^h \times X^1)\right\}\right]^{-1}.$$

We note that this analysis is appropriate when TNF1 is outside of the chromosome region spanned by the other four SNPs. Because TNF1 was in linkage disequilibrium with the other four SNPs (correlation coefficients in controls were -0.21, -0.09, -0.14, and -0.18, respectively), we estimated haplotype frequencies in controls for those with $X^1 = 0$ and $X^1 = 1$ separately. Following the notations in Section 3, $Z = X^1$ and $A_{ZS} = X^1$. In this analysis, we chose to ignore haplotypes with estimated frequencies below 0.025%, so that the expected number of each haplotype in the study sample was at least one in the absence of association. Nine haplotypes for controls with $X^h = 0$ and 6 for controls with $X^h = 1$ were included in the analysis. We chose to present results for haplotype 1122. Neither the main effect of haplotype 1122 nor its interaction with TNF1 was significant by either method, suggesting a lack of evidence that the genomic region spanned by the four TNF SNPs other than TNF1 may contain loci related to prostate cancer risk. As in the single SNP analysis, the two methods yielded similar OR estimates, although the confidence intervals by method II were slightly shorter. Because the number of subjects with both $X^h = 1$ and haplotype 1122 was small (less than 4.0% in controls), the confidence intervals (CIs) of $\hat{\gamma}$ by both methods were large. The haplotype frequency estimates by the two methods were also very close, and the CIs by method II were slightly shorter. In unreported analysis, we also examined the multiplicative effect of haplotype 1122 ($X^h = 0$, 1, or 2 according to the number of haplotype 1122 that a subject had). The results were essentially the same with those in Table 1.

For the purpose of comparison, we also performed an analysis similar to that in Table 1 but assumed that the haplotype distribution was the same for controls with $X^h = 0$ or $X^h = 1$. The respective OR estimates and 95% CIs for β_h, β_z and γ were $-0.03(-0.26, 0.21)$, $-0.17(-0.43, 0.06)$, and $-0.52(-1.18, 0.15)$ by method I and $0.05(-0.16, 0.26)$, $-0.03(-0.28, 0.22)$, and $-1.09(-1.53, -0.66)$ by method II. We observed that method I yielded essentially the same results as those in Table 1, but method II yielded dramatically different results. This indicated that method II is much more sensitive to the assumption $p(H|D = 0, Z, S) = p(H|D = 0)$.

In unreported analysis, we also evaluated the effect of haplotypes formed by all five TNF SNPs. Haplotype 21111 appeared to be significantly associated with prostate cancer risk (OR[95% CI] is 0.773[0.610, 0.978] by method I and 0.761[0.608, 0.953] by method II). However, it appeared that this association was mainly due to the presence of a minor allele at the first locus (TNF1), which in itself was shown to be significant with a similar OR.

5. Simulation Studies

We performed simulation studies to assess the proposed methods in the following aspects: (i) the consistency of methods I and II; (ii) the relative efficiency of the two approaches; (iii) the consistency of the asymptotic variance estimators: the asymptotic variance formula for method I (Appendix $A3$), and the inverse of the information matrix as a variance-covariance estimator for method II; and (iv) the robustness of the methods with respect to the two critical assumptions: $p(H|D, Z, S) = p(H|D, A_{ZS})$ and HWE in controls with the same value for A_{ZS}.

5.1 *Basic study design*

We generated data for 500 one-one matched case-control pairs sampled from 500 distinct strata. We first generated $p(Z = 1|D = 0, S)$ by sampling S numbers from a normal distribution with mean minus one and variance one and then transforming them by the logistic function $\exp(\cdot)/\{1 + \exp(\cdot)\}$. These 500 values were kept fixed in all simulations. For each simulation study, we then generated Z for controls in each of the S strata from a Bernoulli distribution with success probability $p(Z = 1|D = 0, S)$. The prevalence of exposure Z was approximately 0.3 in the sampled controls. We considered the estimation of OR parameters for the main effect of one haplotype and the interaction between this haplotype covariate and variable Z. We considered three models for the haplotype effect: the multiplicative model, the dominant model, and the recessive model.

To generate SNP genotype data for controls, we used 8 common haplotypes estimated from the genotype data for all 5 SNPs in the NCI-ACS study: 11111, 21111, 11122, 11211, 12112, 11112, 11121, and 21121. Their estimated frequencies were 0.285, 0.144, 0.138, 0.111, 0.095, 0.085, 0.076, and 0.066, respectively. For each control, we generated a haplotype pair under the assumption of HWE and then deleted phase information to obtain the SNP genotype data. In this simulation study, we focused on assessing the main effect of haplotype "21111" and its interaction with Z.

We generated Z for cases from the distribution

$$p(Z = 1|D = 1, S) = \frac{p(Z = 1|D = 0, S)e^{\beta_z Z} \sum_G \sum_{H \in \mathcal{H}_G} e^{\beta_h X^H + \gamma k(Z,H)} p^0(H)}{\sum_{i=0}^{1} p(Z = i|D = 0, S)e^{\beta_z Z_i} \sum_G \sum_{H \in \mathcal{H}_G} e^{\beta_h X^H + \gamma k(Z_i,H)} p^0(H)}.$$

We noted that this distribution was compatible with the penetrance function (1). This could be shown in a way similar to that of deriving $p(H|D = 1, Z, A_{ZS})$. The genotype data for cases were generated using diplotype frequencies $p(H|D = 1, Z, A_{ZS})$. The data were analyzed using three methods: methods I and II using the unphased genotype data and standard conditional logistic regression using the simulated diplotype data without phase deletion. The simulation studies were conducted in software R and repeated 200 times.

5.2 *Results*

Table 2 contains results on the consistency of the two methods and their asymptotic variance estimators. We used 1000 matched pairs for the recessive model to compensate for the fact that less than 1% of controls had two copies of haplotype "21111" and $Z = 1$. We used 500 matched pairs for multiplicative and dominant models to save computing time. Averaged estimates for β were sufficiently close to the true values, suggesting the consistency of the two methods. For both methods, the averaged asymptotic standard deviations (SDs) (column $\overline{SD}_A$) are all close to the empirical SDs for 200 β estimates (column SD_E), and all the 95% confidence intervals nearly achieve the nominal coverage probability (column "Cover"). Interestingly, comparison of SDs for standard conditional logistic regression (column "CLREG") and the two proposed methods showed that method II led to the smallest SD, and that method I gave the largest SD. That is, in this simulation study, method II, which did not use phase information, was more efficient than the standard conditional logistic regression analysis that exploited the diplotype data. We note, however, that these two methods were not comparable as they adopted different likelihoods and used different data. The efficiency gain for method II over method I was most apparent for the estimation of interaction effects and for the recessive model. Such gain, quantified by the square of the ratio between estimated variances for methods I and II minus quantity 1, was 103% under the null and 245% with the interaction effect OR being 1.5 (i.e., $\gamma = 0.405$) when the effect of haplotype "21111" was recessive. For the estimation of the main effect β_h, under multiplicative, dominant, and recessive models, the respective efficiency gain of method II was around 16%, 27%, and 86% under the null and 21%, 35%, and 96% under the alternative. Similar efficiency advantage of method II was also observed for the estimation of $\mathbf{f}^0$ (data not shown).

We assessed the sensitivity of the two methods to the HWE assumption by allowing a particular form of departure from HWE when simulating diplotype data for controls. In particular, we assumed that $p\{H = (h_a, h_b)\} = (1-\phi)f_a f_b + \phi f_a$, if $a = b$ and $p\{H = (h_a, h_b)\} = 2(1-\phi)f_a f_b$ if $a \neq b$, where ϕ is the fixation index, with larger absolute value of ϕ indicating more serious departure from HWE. Positive ϕ indicates excess homozygous diplotypes, and negative ϕ indicates excess heterozygous diplotypes. It appears that both methods were insensitive to small deviations from HWE (ϕ=0.05). The test sizes for β were all close to the nominal level. The estimated coverage probabilities of the 95% confidence intervals for all parameters were all close to 95%, although under the alternative, the coverage probabilities for β_h and frequency for haplotype "21111" were more noticeably lower than 95% (90.5% and 86.5%, respectively). Nevertheless, the averaged asymptotic SDs were still close to the empirical SDs. Table 3 displays results corresponding to $\phi = 0.14$, which indicates serious violation of the HWE assumption. Method I performed satisfactorily in all aspects, although the coverage probability for the frequency of haplotype "21111" was slightly lower under the alternative (88.0%). However, method II became problematic. Under the dominant and recessive models, the bias in the main effect β_h was intolerable. For example, the average of 200 estimates by method II was 1.419 compared with the true value 0.405. Consequently, the confidence intervals for β_h had very poor coverage probabilities. Under the dominant model, the 95% confidence intervals for β_h and frequency of haplotype "21111" had respective coverage 62.0% and 77.5% under the null and 52.5% and 51.0% under the alternative. Surprisingly, the averaged estimates of the interaction effect were close to to the truth (0.405) in all scenarios.

When the assumption $p(H|D = 0, Z, S) = p(H|D = 0, A_{ZS})$ does not hold, we note that the degree of bias would be similar to that when the HWE assumption is violated. This is because $p(H|D = 0, Z, S)$ would be the mixture of haplotype distributions in all control strata. This mixture mostly does not follow HWE if each control stratum is at HWE but haplotype frequencies differ among strata. Unreported simulation results confirmed this observation and showed that both the main effect β_h and the interaction effect γ could be seriously biased.

6. Discussion

The two conditional likelihood approaches have different merits. The full conditional likelihood approach (method II) could yield a much more precise estimate of the haplotype-environment interaction effect than the estimated conditional likelihood method (method II). Method I, on the other hand, appeared to be very robust to parametric model assumptions on the distribution of diplotypes conditioned on environmental covariates and matching factors (namely, $p(H|D = 0, Z, S) = p(H|D = 0, A_{ZS})$ and HWE). Certainly these assumptions could be relaxed to a certain extent, as suggested in the previous literature (eg., Satten and Epstein, 2004; Spinka, Carroll, and Chatterjee, 2005). For example, the HWE assumption could be relaxed by incorporating an unknown fixation index to allow excess homozygous or heterozygous diplotypes (Satten and Epstein, 2004). The assumption $p(H|D = 0, Z, S) = p(H|D = 0, A_{ZS}$ could be partially examined by assessing whether the genotype frequency of each single SNP varies with Z. In the situation that these assumptions seem doubtful, method I is probably more assuring to the mind.

The robustness of method I to the two assumptions carries over to the hypothesis testing using the Wald statistic: the size of the tests is generally close to the nominal level. In particular, Chen et al. (2004) reported that the score test based on method I maintains the correct size even when the HWE assumption is violated. They showed that the expectation of the score functions for OR parameters under the null are zero even if wrong haplotype frequencies are used. It is interesting to explore similar robust score tests in the setting of method II.

We proposed to implement method II by the one-step estimation method (Bickel et al., 1993), which essentially yielded identical results as the Newton-Raphson algorithm. But it does not require iteration, it only involves the inversion of sub-matrices of the full Hessian matrix, and it saves the computing time. We implemented these methods in software R, and the program is available upon request to the first author. For fitting a dataset with 500 matched pairs, this program takes approximately 1.5 minutes for method I and 30 additional seconds for method II. The one step method could also be applied to the maximum likelihood method of haplotype analysis for other study designs, upon the availability of an initial consistent estimate.

The two novel methods we studied are closely related to the previous literature. The condition likelihood in method I extends the retrospective likelihood approach in Epstein and Satten (2003) to the matched case-control setting. In the absence of phase ambiguity, method I reduces to the ordinary conditional likelihood analysis. In the absence of haplotype-environment interaction and haplotype-environment dependence in controls, for estimating haplotype-related disease risk parameters, method II reduces to the unconditional maximum likelihood method proposed by Epstein and Satten (2003). We observed in the simulation study that method II is much more efficient than method I for the estimation of interaction effects, particularly under the non-multiplicative models. This observation is consistent with results in the previous literature in the setting of unmatched case-control studies (Satten and Epstein, 2004; Spinka, Carroll, and Chatterjee, 2005).

Methods proposed in this paper are also closely related to the work of Rathouz (2003) and Satten and Carroll (2000). For matched case-control studies, Rathouz (2003) proposed semiparametric efficient estimation methods for parameters in generalized linear models with stratum-specific intercepts and missing covariates. For logistic regression models with covariates modeled parametrically, their efficient method reduces to that in Satten and Carroll (2000). Our method II has a similar flavor. It would be interesting to investigate theoretically whether this full conditional likelihood approach is fully efficient under the HWE and haplotype-environment independence assumptions and to investigate efficient approaches that make less restrictive assumptions. Method I takes a similar form as a “suboptimal” conditional likelihood that Rathouz (2003) studied.

ACKNOWLEDGMENTS

The authors thank Drs. Mitchell H. Gail and Nilanjan Chatterjee at the National Cancer Institute for their helpful comments.

REFERENCES

Bickel, P., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). Efficient and Adaptive Estimation of Semiparametric Models. *The John Hopkins University Press.*

Botstein, D. and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genetics* **33**, 228-237.

Breslow, N. E. and Day, N. E. (1984). Statistical methods in cancer research Volume I: The analysis of case-control studies. *Oxford University Press.*

Calle, E. (2002). The American Cancer Society Nutrition Survey - Rationale, Study Design, and Baseline Characteristics. *Cancer***94**, 2490-501.

Chen, J., Peters, U., Foster, C., Chatterjee, N. (2004). A haplotype-based test of association using data from cohort and nested case-control epidemiological studies. *Human Heredity* **58**, 18-29.

Chen, J. and Chatterjee N. (2006). Haplotype-based association analysis in cohort and nested case-control studies. *Biometrics*, 62:28-35.

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**, 1316-1329.

Excoffier, L. and Slatkin, M. (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Molecular Biology and Evolution* **12**, 921-927.

Kraft, P., Cox, D. G., Paynter, R. A., Hunter, D., and De Vivo, I. (2005). Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques. *Genetic Epidemiology* **28**, 261-272.

Lin, D. Y. (2004). Haplotype-based association analysis in cohort studies of unrelated individuals. *Genetic Epidemiology* **26**, 255-264.

Lin, D. Y. and Zeng D. L.(2005). Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*, in press.

Rathouz, P. J. (2003). Likelihood methods for missing covariate data in highly stratified studies. *Journal of the Royal Statistical Society (B)* **65**, 711-723.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517.

Satten, G. A. and Carroll, R. J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384-388.

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.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425-434.

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**, 108-127.

Stram, D., Pearce, C. L., Bretsky, P., Freedman, M., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., and Thomas, D. C. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Human Heredity* **55**, 179-190.

Wallenstein, S., Hodge, S., and Weston, A. (1998). A logistic regression model for analyzing extended haplotype data. *Genetic Epidemiology* **15**, 173-181.

Zhao, L. P., Li, S., and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *American Journal of Human Genetics* **72**, 1231-1250.

APPENDIX

To simplify the notation, we present all derivations assuming $p(H|D=0,Z,S) = P(H|D=0)$, but all results apply to the situation when the distribution of H in controls depends on A_{ZS}, that is, $p(H|D=0,Z,S) = P(H|D=0,A_{ZS})$.

A1 The derivation of $p(\mathbf{D}_s|n_{1s},\mathbf{Z}_s,s)$

We take the support of genotype $\mathbf{G}$ to be the collection of all unique genotypes in the case-control sample. Applying a result in Epstein and Satten (2003), we derived

$$p(D=1|G,Z,S) = p(D=0|G,Z,S)e^{\alpha_s+\beta_z Z}\sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)}p(H|D=0,G).$$

Thus, we have

$$
\begin{aligned}
& p(D=1|Z,S) = \sum_G p(D=1|G,Z,S)p(G|Z,S) \\
= \;& e^{\alpha_s+\beta_z Z}\sum_G \left\{ p(D=0|G,Z,S)p(G|Z,S) \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0,G,Z,S) \right\} \\
= \;& p(D=0|Z,S)e^{\alpha_s+\beta_z Z}\sum_G \left\{ p(G|D=0,Z,S) \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0,G,Z,S) \right\} \\
= \;& p(D=0|Z,S)e^{\alpha_s+\beta_z Z}\sum_G \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0).
\end{aligned}
$$

This leads to the result

$$\theta(Z,S) = \frac{p(D=1|Z,S)}{p(D=0|Z,S)} = e^{\alpha_s+\beta_z Z}\sum_G \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0).$$

Let $\theta(Z) = e^{\beta_z Z}\sum_G \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0)$. Then $p(\mathbf{D}_s|n_{1s},\mathbf{Z}_s,S)$ follows.

A2 The derivation of $p(\mathbf{G}|\mathbf{D}_s,\mathbf{Z}_s,S)$

By the assumption that $\mathrm{p}(H|D=0,Z,S) = \mathrm{p}(H|D=0)$, we have

$$
\begin{aligned}
& \mathrm{p}(H|D=1,Z,S) = \mathrm{p}(D=1|H,Z,S)p(H|Z,S)/\mathrm{p}(D=1|Z,S) \\
= \;& e^{\alpha_s+\beta_z Z+\beta_h X^H+\gamma k(Z,H)} p(H|D=0,Z,S)p(D=0|Z,S)/p(D=1,Z,S) \\
= \;& \frac{e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0)}{\sum_G \sum_{H\in\mathcal{H}_G} e^{\beta_h X^H+\gamma k(Z,H)} p(H|D=0)}.
\end{aligned}
$$

In particular, when $\gamma = 0$, this formula reduces to

$$\mathrm{p}(H|D=1,Z,S) = \frac{e^{\beta_h X^H} p(H|D=0)}{\sum_{\mathbf{G}} \sum_{H\in\mathcal{H}_{\mathbf{G}}} e^{\beta_h X^H} p(H|D=0)} \equiv \mathrm{p}(H|D=1,Z).$$

Thus, haplotype frequencies in cases are independent of S when those of controls are independent of S. In the absence of haplotype-environment interaction, such frequencies are independent of Z as well.

A3 The asymptotic property of method I

The proof largely follows that in Chen and Chatterjee (2006). Let $U_{\beta\alpha}(\beta,\hat{\alpha}) = \partial L^I(\beta,\hat{\alpha})/\partial\beta$. A standard Taylor's series expansion of $U(\hat{\beta},\hat{\alpha})$ around the true parameter values (β,α) leads to

$$\hat{\beta} - \beta \approx I_{\beta\beta}^{-1} U(\beta, \alpha) - I_{\beta\beta}^{-1} I_{\beta\alpha}(\hat{\alpha} - \alpha),$$

where $I_{\beta\beta}$ and $I_{\beta\alpha}$ are the large-sample limits of $-\partial U(\beta,\alpha)/\partial\beta$ and $-\partial U(\beta,\alpha)/\partial\alpha$, respectively. Since $L^I(\beta,\alpha)$ is a proper conditional likelihood function, following the standard asymptotic theory for matched case-control studies, we have $\text{cov}\{U(\beta,\alpha)\} = I_{\beta\beta}$, and $\sqrt{n_c}U(\beta,\alpha) \sim \text{Normal}(0, I_{\beta\beta})$. Moreover, from standard parametric maximum likelihood inference theory, we have $\sqrt{n_c}(\hat{\alpha} - \alpha) \sim \text{Normal}\{0, (I_{\alpha\alpha}^c)^{-1}\}$, where n_c is the total number of matched controls and $I_{\alpha\alpha}^c$ is the asymptotic information matrix for α (Excoffier and Slatkin, 1995).

Furthermore, $U(\beta,\alpha)$ and $(\hat{\alpha} - \alpha)$ are asymptotically uncorrelated. The results above show that $\hat{\beta}$ follows an asymptotically normal distribution with mean β and variance $\Sigma = I_{\beta\beta}^{-1} + I_{\beta\beta}^{-1} I_{\beta\alpha}(I_{\alpha\alpha}^c)^{-1} I_{\beta\alpha}^T I_{\beta\beta}^{-1}$. Σ can be estimated as follows. Let $\hat{I}_{\beta\beta} = -\partial U(\beta,\alpha)/\partial\beta|_{\hat{\beta},\hat{\alpha}}$, $\hat{I}_{\beta\alpha} = -\partial U(\beta,\alpha)/\partial\alpha|_{\hat{\beta},\hat{\alpha}}$, and $\hat{I}_{\alpha\alpha}^c$ be the estimated information matrix for α using controls only. Then Σ can be consistently estimated as

$$\hat{\Sigma} = \hat{I}_{\beta\beta}^{-1} + \hat{I}_{\beta\beta}^{-1} \hat{I}_{\beta\alpha}(\hat{I}_{\alpha\alpha}^c)^{-1} \hat{I}_{\beta\alpha}^T \hat{I}_{\beta\beta}^{-1}.$$

Above, $I_{\beta\beta}$ can also be consistently estimated as $\sum_{s=1}^S U^s(\hat{\beta},\hat{\alpha}) \left[U^s(\hat{\beta},\hat{\alpha})\right]^T$, where U^s is $U_{\beta\alpha}(\hat{\beta},\hat{\alpha})$ for the s^{th} matched set.

A4 The score functions and Hessian matrix for method II

We give score functions for $(\beta, \mathbf{f}^0)$ only for the one-one matching case. Let $\mathcal{X} = \{X^H, Z, k(Z,H)\}$, let $\text{E}_{si}^1(\cdot) = \text{E}(\cdot|D=1, Z_{si})$, and let

$$W_{\beta_h,\gamma}(Z_{si}, H) = \sum_G \sum_{H \in H_G} e^{\beta_h X^H + \gamma k(Z,H)} p(H|D=0).$$

Then the score functions for β are

$$U_\beta \;=\; \sum_{s=1}^S \left[\text{E}(\mathcal{X}|D=1, G_{s1}, Z_{s1}) - \frac{\sum_{i=0,1} \text{E}(\mathcal{X}|D=1, Z_{si}) e^{\beta_z Z_{si}} W_{\beta_h,\gamma}(Z_{si}, H)}{\sum_{i=0,1} e^{\beta Z_{si}} W_{\beta_h,\gamma}(Z_{si}, H)}\right].$$

Let Δ_a be the number of counts $(0, 1, 2)$ for haplotype h_a. The score functions for f_a^0 without applying the restriction $\sum_a f_a^0 = 1$ are

$$U_{f_a^0} = \frac{1}{f_a^0}\sum_{s=1}^{S}\left[\sum_{i=0,1} \mathrm{E}(\Delta_a|D=1,G_{si},Z_{si}) - \frac{\sum_{i=0,1}\mathrm{E}(\Delta_a|D=1,Z_{si})e^{\beta_z Z_{si}}W_{\beta_h,\gamma}\{Z_{si},H\}}{\sum_{i=0,1}e^{\beta Z_{si}}W_{\beta_h,\gamma}\{Z_{si},H\}}\right].$$

It can be shown that $\hat{f}_a^0$ can be obtained as $f_a^0 U_{f_a^0}/2n_c$, where n_c is the total number of controls in the matched case-control sample. The score function for α_a can be easily obtained from $U_{f_a^0}$. The Hessian matrix also has a very simple form.

$$\begin{aligned} I_{\beta\beta} = & -\sum_{s=1}^{S}\left[\frac{\sum_{i=0,1}\mathrm{E}_{si}^1(\mathcal{X}\mathcal{X}^T)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)}{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)} + \mathrm{cov}(\mathcal{X}\mathcal{X}^T|D=1,G_{s1},Z_{s1})\right. \\ & \left. -\frac{\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\mathcal{X})e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\mathcal{X})e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^T}{\left\{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^2}\right], \end{aligned}$$

$$\begin{aligned} I_{\beta\mathbf{f}^0} = & -\frac{1}{f_a}\sum_{s=1}^{S}\left[\frac{\sum_{i=0,1}\mathrm{E}_{si}^1(\mathcal{X}\Delta^T)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)}{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)} + \mathrm{cov}(\mathcal{X}\Delta^T|D=1,G_{s1},Z_{s1})\right. \\ & \left. -\frac{\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\mathcal{X})e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\Delta)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^T}{\left\{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^2}\right], \end{aligned}$$

and

$$\begin{aligned} I_{f_a f_b} = & -\frac{1}{f_a f_b}\sum_{s=1}^{S}\left[\frac{\sum_{i=0,1}\mathrm{E}_{si}^1(\Delta_a\Delta_b)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)}{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)} + \sum_{i=0,1}\mathrm{cov}(\Delta_a\Delta_b|D=i,G_{si},Z_{si})\right. \\ & \left. -\frac{\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\Delta_a)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}\left\{\sum_{i=0,1}\mathrm{E}_{si}^1(\Delta_b)e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^T}{\left\{\sum_{i=0,1}e^{\beta_z Z_{si}}W_{\beta_h,\gamma}(Z_{si},H)\right\}^2}\right] \\ & -I(a=b)(f^a)^{-2}U_{f_a}. \end{aligned}$$

$I_{\beta,\mathbf{f}^0}$ and $I_{\mathbf{f}^0,\mathbf{f}^0}$ can then be easily transformed to get $I_{\beta\alpha}$ and $I_{\alpha\alpha}$, the Hessian matrix corresponding to $\alpha = \{\alpha_a,\ a = 1, \ldots, J-1\}$.

Table 1

The ASC Study: Prostate Cancer Risk Associated with the Presence of Haplotype 1122 in Reference to its absence

	Method I		Method II	
$\hat{\beta}_h$(95% CI)	-0.03(-0.26, 0.21)		-0.10(-0.31, 0.12)	
$\hat{\beta}_z$(95% CI)	-0.19(-0.45, 0.06)		-0.22(-0.47, 0.04)	
$\hat{\gamma}$(95% CI)	-0.50(-1.16, 0.16)		-0.34(-1.00, 0.33)	
	$X^h = 0$	$X^h = 1$	$X^h = 0$	$X^h = 1$
$100 \times \hat{f}^{0a}_{1111}$(95% CI)	56.15^b(53.54, 58.76)	80.66(76.80, 84.52)	56.36^c(54.34, 58.38)	80.29(76.94, 83.64)
$100 \times \hat{f}^{0}_{2111}$(95% CI)	17.99(15.97, 20.01)	8.39(5.68, 11.10)	17.39(15.99, 18.79)	8.38(5.67, 11.09)
$100 \times \hat{f}^{0}_{1122}$(95% CI)	16.27(14.35, 18.19)	5.69(3.43, 7.95)	16.57(14.69, 18.44)	6.30(4.45, 8.16)
$100 \times \hat{f}^{0}_{1212}$(95% CI)	5.13(3.96, 6.30)	1.98(0.62, 3.34)	5.63(4.74, 6.53)	2.52(1.24, 3.80)
$100 \times \hat{f}^{0}_{1112}$(95% CI)	2.16(1.40, 2.92)	1.27(0.16, 2.38)	1.79(1.34, 2.23)	0.47(0.16, 0.78)
$100 \times \hat{f}^{0}_{1121}$(95% CI)	1.51(0.78, 2.24)	2.01(0.63, 3.39)	1.62(1.08, 2.16)	2.04(1.00, 3.09)
$100 \times \hat{f}^{0}_{2121}$(95% CI)	0.65(0.12, 1.18)	-^{d}	0.58(0.27, 0.89)	-
$100 \times \hat{f}^{0}_{2221}$(95% CI)	0.07(0.01, 0.21)	-	0.03(0.01, 0.06)	-
$100 \times \hat{f}^{0}_{1222}$(95% CI)	0.05(0.01, 0.18)	-	0.07(0.01, 0.33)	-

a: Subscripts refer to haplotypes, with "1" representing the wild-type allele and "2" the variant allele.

b: Haplotype frequency estimates by applying EM algorithm to controls only.

c: Haplotype frequency estimates using method II.

d: Haplotype frequency estimate was less than 0.0001.

Table 2

Consistency and Efficiency of Proposed Methods, and Consistency of Asymptotic Variance Estimators: Simulation Studies Using 500 Matched Case-Control Pairs and 200 Repeated Runs.

		CLREG	Method I		Method II	
Model		$\overline{\hat{\beta}}(SD_E)^a$	$\overline{\hat{\beta}}(\overline{SD}_A/SD_E)^b$	Cover^d	$\overline{\hat{\beta}}(\overline{SD}_A/SD_E)^c$	Cover^d
		Under the Null Hypothesis: $\beta_{h_2} = \beta_z = \gamma = 0$				
Multiplicative	β_{h_2}	-0.001(0.154)	0.003(0.163/0.160)	0.950	-0.005(0.150/0.146)	0.960
	β_z	-0.001(0.162)	-0.004(0.172/0.167)	0.965	-0.004(0.161/0.157)	0.955
	γ	0.001(0.281)	0.004(0.293/0.297)	0.945	0.001(0.209/0.200)	0.960
Dominant	β_{h_2}	-0.019(0.183)	-0.022(0.185/0.188)	0.930	-0.010(0.164/0.163)	0.960
	β_z	-0.011(0.170)	-0.014(0.175/0.175)	0.965	-0.007(0.162/0.155)	0.965
	γ	0.014(0.310)	0.024(0.332/0.329)	0.955	-0.004(0.234/0.225)	0.970
Recessive^e	β_{h_2}	-0.005(0.380)	-0.005(0.387/0.380)	0.950	-0.019(0.284/0.281)	0.955
	β_z	-0.002(0.110)	-0.002(0.108/0.110)	0.940	-0.002(0.107/0.110)	0.930
	γ	0.058(0.718)	-0.058(0.742/0.718)	0.980	-0.043(0.520/0.509)	0.975
		Under the Alternative Hypothesis: $\beta_{h_2} = 0.405$, $\beta_z = 0.916$, $\gamma = 0.405$				
Multiplicative	β_{h_2}	0.411(0.161)	0.409(0.170/0.169)	0.965	0.382(0.154/0.157)	0.940
	β_z	0.928(0.189)	0.930(0.187/0.190)	0.935	0.938(0.174/0.180)	0.940
	γ	0.418(0.251)	0.411(0.269/0.254)	0.965	0.388(0.158/0.166)	0.935
Dominant	β_{h_2}	0.413(0.194)	0.405(0.202/0.202)	0.950	0.402(0.174/0.178)	0.955
	β_z	0.930(0.188)	0.927(0.189/0.190)	0.940	0.927(0.174/0.177)	0.950
	γ	0.413(0.300)	0.418(0.317/0.310)	0.965	0.404(0.195/0.196)	0.950
Recessive^e	β_{h_2}	0.425(0.419)	0.425(0.386/0.419)	0.940	0.439(0.276/0.277)	0.915
	β_z	0.919(0.116)	0.919(0.109/0.116)	0.930	0.918(0.109/0.113)	0.935
	γ	0.464(0.658)	0.464(0.634/0.658)	0.925	0.441(0.341/0.365)	0.935

a: The mean (standard deviation) of $\hat{\beta}$ by standard conditional logistic regression using known phase.

b: The mean (averaged estimated asymptotic/empirical standard deviation) using method I.

c: The mean (averaged estimated asymptotic/empirical standard deviation) using method II.

d: 95% coverage probabilities.

e: 1000 matched pairs were used for the recessive model.

Table 3

Robustness to the Violation of HWE due to Excess Homozygosity (Fixation Index is 0.14): Simulation Studies Using 500 Matched Case-Control Pairs and 200 Repeated Runs.

		CLREG	Method I		Method II	
Model		$\bar{\hat{\beta}}(SD_E)^a$	$\bar{\hat{\beta}}(\overline{SD}_A/SD_E)^b$	Cover^c	$\bar{\hat{\beta}}(\overline{SD}_A/SD_E)^d$	Cover^c
		Under the Null Hypothesis: $\beta_{h_2} = \beta_z = \gamma = 0$				
	β_{h_2}	0.001(0.105)	0.001(0.128/0.125)	0.955	-0.001(0.092/0.117)	0.860
Multiplicative	β_z	0.010(0.122)	0.011(0.120/0.123)	0.940	0.011(0.114/0.117)	0.940
	γ	-0.009(0.190)	-0.014(0.188/0.189)	0.955	-0.016(0.136/0.152)	0.930
	β_{h_2}	0.005(0.124)	0.000(0.132/0.135)	0.955	-0.199(0.115/0.116)	0.620
Dominant	β_z	0.009(0.123)	0.008(0.122/0.125)	0.940	0.002(0.114/0.120)	0.925
	γ	-0.016(0.223)	-0.014(0.238/0.241)	0.955	0.008(0.169/0.175)	0.940
	β_{h_2}	-0.006(0.273)	-0.006(0.270/0.273)	0.615	1.048(0.284/0.252)	0.035
Recessive	β_z	0.001(0.099)	0.001(0.108/0.099)	0.925	0.000(0.101/0.097)	0.985
	γ	-0.033(0.535)	-0.033(0.443/0.535)	0.895	0.023(0.502/0.497)	0.945
		Under the Alternative Hypothesis: $\beta_{h_2} = 0.405$, $\beta_z = 0.916$, $\gamma = 0.405$				
	β_{h_2}	0.416(0.106)	0.417(0.112/0.110)	0.935	0.440(0.088/0.106)	0.945
Multiplicative	β_z	0.906(0.128)	0.907(0.130/0.130)	0.950	0.877(0.120/0.122)	0.945
	γ	0.412(0.182)	0.410(0.171/0.184)	0.945	0.480(0.097/0.120)	0.945
	β_{h_2}	0.408(0.141)	0.401(0.142/0.148)	0.945	0.165(0.121/0.124)	0.525
Dominant	β_z	0.928(0.135)	0.929(0.131/0.139)	0.945	0.926(0.122/0.126)	0.945
	γ	0.397(0.227)	0.393(0.224/0.241)	0.945	0.404(0.138/0.147)	0.925
	β_{h_2}	0.390(0.251)	0.390(0.248/0.251)	0.825	1.419(0.272/0.260)	0.040
Recessive	β_z	0.910(0.110)	0.910(0.111/0.110)	0.950	0.909(0.110/0.109)	0.960
	γ	0.461(0.484)	0.461(0.433/0.484)	0.885	0.454(0.338/0.385)	0.925

a: The mean (standard deviation) of $\hat{\beta}$ by standard conditional logistic regression.

b: The mean (averaged estimated asymptotic/empirical standard deviation) using Method I.

c: 95% coverage probabilities.

d: The mean (averaged estimated asymptotic/empirical standard deviation) using Method II.