

Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies

BY NILANJAN CHATTERJEE

*Division of Cancer Epidemiology and Genetics, National Cancer Institute,
National Institutes of Health, Department of Health and Human Services, Rockville,
Maryland 20852, U.S.A.*

chattern@mail.nih.gov

AND RAYMOND J. CARROLL

Texas A&M University, College Station, Texas 77843-3143, U.S.A.

carroll@stat.tamu.edu

SUMMARY

We consider the problem of maximum-likelihood estimation in case-control studies of gene-environment associations with disease when genetic and environmental exposures can be assumed to be independent in the underlying population. Traditional logistic regression analysis may not be efficient in this setting. We study the semiparametric maximum likelihood estimates of logistic regression parameters that exploit the gene-environment independence assumption and leave the distribution of the environmental exposures to be nonparametric. We use a profile-likelihood technique to derive a simple algorithm for obtaining the estimator and we study the asymptotic theory. The results are extended to situations where genetic and environmental factors are independent conditional on some other factors. Simulation studies investigate small-sample properties. The method is illustrated using data from a case-control study designed to investigate the interplay of BRCA1/2 mutations and oral contraceptive use in the aetiology of ovarian cancer.

Some key words: Case-control study; Gene-environment interaction; Genetic epidemiology; Logistic regression; Population stratification; Profile likelihood; Retrospective study; Semiparametric method.

1. INTRODUCTION

The case-control study design gives an efficient way of collecting covariate information for epidemiological studies of rare diseases. Cornfield (1956) showed that the prospective odds ratio of a disease given a covariate is equivalent to the retrospective odds ratio of the covariate given the disease and thus prospective odds ratios are estimable from case-control designs. For discrete covariates, Andersen (1970) and then more generally Prentice & Pyke (1979) showed that fitting a standard prospective logistic regression that ignores the retrospective sampling nature of the design yields the maximum likelihood estimates of the regression parameters under a ‘semiparametric’ model that allows the covariate distribution to be nonparametric. More recently, Rabinowitz (1997) and Breslow et al.

(2000) used modern semiparametric theory to show that the prospective logistic regression analysis of case-control data is efficient in the sense that it achieves the variance lower bound of the underlying semiparametric model.

It is now believed that the risks of many complex diseases are determined by the combined effects of genetic susceptibility G and environmental or non-genetic exposures E , and, since studies of interactions, especially for rare exposures, typically require a large sample size, efficient designs and analytical methods for gene-environment interaction are vital.

A special feature of the gene-environment interaction problem is that it may often be reasonable to assume that a subject's genetic susceptibility, a factor which is determined from birth, is independent of his/her subsequent environmental exposure. Standard logistic regression analysis, being the semiparametric maximum likelihood solution for the problem that allows an arbitrary covariate distribution, clearly remains a valid option for analysing case-control data. However, the method may not be efficient because it fails to exploit the gene-environment independence assumption. In general, under the case-control design, the variance lower bound for estimators of the regression parameters under particular constraints or models for the covariate distribution will be lower than that of the more general model that allows a completely nonparametric covariate distribution.

In the past, several researchers have presented analytical methods that exploit the gene-environment independence assumption. Piegorsch et al. (1994) noted that, under gene-environment independence and the rare disease assumption, the multiplicative interaction parameter in the logistic regression model can be estimated as the odds ratio between G and E among cases alone. Moreover, they observed that the corresponding case-only estimator of interaction is more precise than the estimator of the interaction parameter from traditional logistic regression analysis involving both cases and controls. When data on both cases and controls are available, assuming rare disease and categorical exposures, Umbach & Weinberg (1997) showed that maximum-likelihood estimators of all the parameters of a logistic regression model can be obtained in a fairly general setting by fitting a suitably constrained log-linear model to the data. They showed that, for simple scenarios that involve dichotomous G , dichotomous E and no confounder, the log-linear model and case-only analysis approach yields the same estimator of the multiplicative interaction parameter in the logistic regression model. Modan et al. (2001), in a specific application, noted that, under gene-environment independence and the rare disease assumption, $\text{pr}(E|G, D=0) = \text{pr}(E|D=0)$, where $D=0$ corresponds to disease-free, i.e. control, subjects. Based on this, they argued that the disease odds ratio associated with E among subjects with genotype $G=g$ can be estimated by a logistic regression analysis that compares the distribution of E among all controls, $\text{pr}(E|D=0)$, with the exposure distribution among cases with $G=g$, $\text{pr}(E|D=1, G=g)$.

The methods have some limitations. First, they all require the risk of the disease to be small for all levels of both genetic and environmental exposures. This assumption can lead to substantial bias in the estimation of the odds ratio parameters even for diseases like cancer, for which the marginal probability of the disease may be small in the population but the disease risk may be high for certain combinations of genetic and environmental exposures (Schmidt & Schaid, 1999). Secondly, the methods of Piegorsch et al. (1994) and Modan et al. (2001) allow estimation of some, but not all, of the parameters of interest in the general logistic regression model. Thirdly, some of the above methods have been described in very simple settings involving only two factors G and E , and it is often not clear how to exploit the gene-environment independence assumption in the most general

setting that will, for example, allow for potential confounders or account for factors that could induce association between G and E . The log-linear model framework described by Umbach & Weinberg (1997) for categorical co-factors gives the most general method to date for exploiting the gene-environment independence assumption and can handle some of these issues. For a rich model with many covariates, however, the log-linear modelling approach can easily become cumbersome and intricate. Moreover, in a rich model, the log-linear specification would typically involve a large number of ‘nuisance parameters’ that characterise the covariate distribution among the controls. When continuous covariates are involved, the number of such nuisance parameters would even increase with the sample size. The asymptotic theory for the lower-dimensional regression parameters of interest in the presence of the high-dimensional nuisance parameters is nonstandard and has not been studied rigorously under the underlying semiparametric setting.

In this paper, we develop a general framework for maximum-likelihood estimation under the gene-environment independence assumption. The proposed method has several unique aspects. First, it is exact in not requiring any rare-disease assumption. Secondly, we develop the methodology in a very general setting so that it retains all the flexibility of traditional logistic regression analysis, such as adjustment for confounders, incorporation of continuous exposures and/or confounders and complex modelling of the regression effects of the risk factors. Thirdly, we show how to incorporate external information about the marginal probability of the disease in the population and hence improve efficiency of parameter estimation. Fourthly, we show how to adjust for bias that may arise when G and E may be related because of their dependence on other common measured factors. Finally, we develop the methodology in a semiparametric framework that allows the distribution of the environmental factors $F(e)$ to be completely nonparametric. Given that in a typical application E might include many factors, both discrete and continuous variables, nonparametric treatment of $F(e)$ is attractive both for avoiding complex modelling and for robustness.

2. ESTIMATION THEORY AND METHODOLOGY

2.1. Model and identification

Let D be the binary indicator of presence, $D = 1$, or absence, $D = 0$, of a disease. Suppose the prospective risk model for the disease given a subject’s genetic factors, G , and environmental risk factors, E , is given by the logistic regression model $\text{pr}(D = 1|G, E) = H\{\beta_0 + m(G, E; \beta_1)\}$, where $H(x) = \{1 + \exp(-x)\}^{-1}$ is the logistic distribution function and $m(\cdot)$ is a known but arbitrary function. Typically, in the standard logistic regression model, one has $m(G, E, \beta_1) = (G, E, G * E)\beta_1$ with the exponents of the parameters in β_1 having the standard exposure odds ratio interpretation. However, more general forms of $m(\cdot)$ could be of interest, especially for interaction studies where different forms of $m(\cdot)$ can be chosen to assess interaction at different scales; see Khouri et al. (1993, § 5.5.3). We assume that the joint distribution of G and E is given by the product form $\mathcal{H}(e, g) = Q(g)F(e)$, where Q and F are the marginal distribution functions of G and E , respectively. Suppose that N_0 controls and N_1 cases are sampled from the conditional distributions $\text{pr}(G, E|D = 1)$ and $\text{pr}(G, E|D = 0)$, respectively, and let $(G_i, E_i)_{i=1}^{N_0+N_1}$ denote the corresponding covariate data of the $N_0 + N_1$ study subjects.

Before we describe estimation, it is useful to study the identifiability of the parameters. In a nonparametric setting where no assumption is made about the form of the covariate

distribution \mathcal{H} , it is well known that neither \mathcal{H} nor the intercept parameter β_0 is identifiable from case-control data (Prentice & Pyke, 1979). Under the assumption of gene-environment independence, however, these results may not necessarily be true. Let \mathcal{B} denote the parameter space for β_1 and let $\mathcal{B}^0 \subset \mathcal{B}$ denote the values of β_1 so that $m(G, E, \beta_1)$ depends only on G or only on E , but not both. For example, suppose that $m(G, E, \beta_1)$ corresponds to a standard logistic regression model with $\beta_1 = (\beta_G, \beta_E, \beta_{GE})$, where β_G , β_E and β_{GE} denote the main effect of G , the main effect of E and the interaction between G and E , respectively. In this case, the set \mathcal{B}^0 would consist of parameter values of the form $\beta_1 = (\beta_G, 0, 0)$ or $\beta_1 = (0, \beta_E, 0)$, which correspond to either only the main effect of G or only the main effect of E , respectively. Since β_1 is well known to be identifiable from case-control data under general nonparametric \mathcal{H} , it follows trivially that β_1 remains identifiable when H is assumed to be of the form $\mathcal{H} = Q \times F$. The identifiability result for the remaining parameters can be stated as follows.

LEMMA 1. For all $\beta_1 \notin \mathcal{B}^0$,

$$\text{pr}(E = e, G = g | D = d, \beta_0, \beta_1, Q, F) = \text{pr}(E = e, G = g | D = d, \beta_0^*, \beta_1, Q^*, F^*)$$

if and only if $\beta_0 = \beta_0^*$, $Q = Q^*$ and $F = F^*$.

The proof of Lemma 1 is given in the Appendix. Thus, we note that a somewhat surprising consequence of the gene-environment independence assumption is that, except for some boundary situations, the intercept parameter of the logistic regression model β_0 is theoretically identifiable from the retrospective likelihood of case-control data. Although this may seem counter-intuitive, it is easy to see from the proof of Lemma 1 that in general the identifiability of β_0 is intrinsically related to the class of \mathcal{H} that is under consideration.

2.2. Profile likelihood estimation

We begin with the following parameterisation of the exposure distributions Q and F . We assume that the genetic factor G for a subject can take values in a fixed set $\{g_0, \dots, g_J\}$. Thus the distribution Q can be parameterised by the corresponding probability masses $\{q_0, \dots, q_J\}$. Moreover, using population genetics theory, in many situations the probabilities q_j ($j = 1, \dots, J$) can be further modelled as $q_j = q_j(\theta)$, for some known function q_j and some parameter vector θ . For example, if G represents one of the three possible genotypes a subject can have corresponding to a bi-allelic locus, the population frequencies of the three genotypes could be specified in terms of the allele frequency of one of the alleles under the Hardy–Weinberg equilibrium assumption for the underlying population. If no population genetics model assumption is made to specify the q_j 's, we will assume in the above notation that θ represents the vector of q_j 's themselves.

For parameterisation of the environmental covariate E , we first assume that the non-parametric maximum likelihood estimator of F can allow positive masses only within the set $\mathcal{E} = \{e_1, \dots, e_K\}$ that represents the unique values of E that are observed in the case-control sample of $N_0 + N_1$ study subjects. Thus, for obtaining the maximum likelihood estimator it is sufficient to consider the class of discrete F that have support points within the set \mathcal{E} . Any F in this class can be parameterised with respect to the probability masses $\{\delta_1, \dots, \delta_K\}$ that it assigns to the points $\{e_1, \dots, e_K\}$. Let n_{ijk} denote the number of subjects with $D = i$, $G = g_j$ and $E = e_k$. Let the corresponding marginal frequencies for the i th category of disease be $n_{i++} = N_i$, for the j th category of G be n_{+j+} and for the k th

category of E be n_{++k} . The loglikelihood for the case-control data is then

$$\begin{aligned} L &= \log \{ \ell_{CC}(\beta_0, \beta_1, \theta, \delta) \} = \sum_{u=1}^{N_0+N_1} \log \{ \text{pr}(D_u | G_u, E_u) \text{pr}(G_u) \text{pr}(E_u) / \text{pr}(D_u) \} \\ &= \sum_{ijk} n_{ijk} \log \{ P_{ij}(e_k, \beta_0, \beta_1) \} + \sum_j n_{+j+} \log \{ q_j(\theta) \} + \sum_k n_{++k} \log(\delta_k) \\ &\quad - \sum_i n_{i++} \log \left\{ \sum_{lm} P_{il}(e_m, \beta_0, \beta_1) q_l(\theta) \delta_m \right\}, \end{aligned} \quad (1)$$

where $P_{ij}(e_k, \beta_0, \beta_1) = \text{pr}(D = i | G = j, E = e_k)$.

When the dimension of δ is large, as could be expected when E consists of multiple covariates and/or some of its components are continuous, direct maximisation of the loglikelihood with respect to $(\beta_0, \beta_1, \theta, \delta)$ may be numerically challenging or even infeasible. An alternative approach is first to derive the profile likelihood of the data that is obtained by maximising the likelihood with respect to δ for fixed values of $\gamma = (\beta_0, \beta_1, \theta)$ and then to maximise the profile likelihood with respect to γ . If $\hat{\delta}(\gamma)$ denotes the value of δ that maximises the likelihood for fixed γ , the profile loglikelihood is then $L(\gamma, \hat{\delta}(\gamma))$. In Lemma 2, we state an equivalent representation of $L\{\gamma, \hat{\delta}(\gamma)\}$ that is computationally useful.

LEMMA 2. Define the parameters $\mu_i = n_{i++} / \{N \text{pr}(D = i)\}$ for $i = 0, 1$ and let

$$P_{ij}^* \{ e_k; \gamma, \mu = (\mu_0, \mu_1) \} = \frac{P_{ij}(e_k; \beta_0, \beta_1) \mu_i q_j(\theta)}{\sum_i \sum_j P_{ij}(e_k; \beta_0, \beta_1) \mu_i q_j(\theta)}. \quad (2)$$

The profile loglikelihood $L\{\gamma, \hat{\delta}(\gamma)\}$ can be computed as $L^*\{\gamma, \hat{\mu}(\gamma)\}$, where

$$L^*(\gamma, \mu) = \sum_{ijk} n_{ijk} \log P_{ij}^* \{ e_k; \gamma, \mu \}, \quad (3)$$

and $\hat{\mu}(\gamma) = \{ \hat{\mu}_0(\gamma), \hat{\mu}_1(\gamma) \}$ is defined by the solution of the equations

$$n_{i++} = \sum_k \sum_j n_{++k} P_{ij}^* \{ e_k; \gamma, \mu \} \quad (i = 0, 1). \quad (4)$$

The proof of the lemma is given in the Appendix and is developed following techniques in Scott & Wild (1997). The main consequence of Lemma 2 is that $L\{\gamma, \hat{\delta}(\gamma)\}$ can be computed without having to maximise the likelihood $L(\gamma, \delta)$ numerically with respect to the potentially high-dimensional nuisance parameter δ . Instead, $L\{\gamma, \hat{\delta}(\gamma)\}$ can be obtained in closed form up to only two additional parameters $\mu = (\mu_0, \mu_1)$, which in turn are defined as the solution of two equations given in (4). The result of this lemma can also be compared to the classical result of Prentice & Pyke (1979) that, when the exposure distribution is unspecified, maximisation of the retrospective likelihood can be achieved by simply fitting the prospective logistic model to the data ignoring the retrospective design. Lemma 2 gives the corresponding simplification for maximum-likelihood estimation under the gene-environment independence assumption and unspecified distribution of E . Prentice & Pyke (1979) also essentially showed that the maximum likelihood estimator of the logistic regression parameters can be obtained by maximising the prospective likelihood of the form $\text{pr}(D|X, \delta = 1)$, where δ is the indicator of whether or not a subject has been selected in the case-control sample and $\text{pr}(\delta = 1|D)$, the probability of selection of a subject given his/her disease status, is fixed at its asymptotic value, which in turn is proportional to μ_D . Similarly, Lemma 2 shows that the maximum likelihood estimator of the regression parameter under the gene-environment independence assumption can

be obtained by solving score equations corresponding to the prospective likelihood $P_{DG}^*(E) = \text{pr}(D, G|E, \delta = 1)$. For derivation of the asymptotic distribution theory, however, we will show later that $P_{DG}^*(E)$ cannot be treated as an ordinary likelihood.

For computational convenience we consider further reparameterisation of the problem. Let $G = g_0$ define a reference category for the genetic exposure G . We can now write

$$P_{ij}^*(e_k, \gamma, \mu) = \frac{\exp\{\theta_{ij}(e_k; \gamma, \mu)\}}{1 + \sum_{i':(i'j) \neq (0,0)} \exp\{\theta_{i'j}(e_k; \gamma, \mu)\}}, \quad (5)$$

where

$$\begin{aligned} \theta_{ij}(e_k; \gamma, \mu) &= \log \left\{ \frac{P_{ij}^*(e_k; \gamma, \mu)}{P_{00}^*(e_k; \gamma, \mu)} \right\} \\ &= i\{\beta_0 + \log(\mu_1/\mu_0)\} + im(g_j, e_k; \beta_1) + \log\{q_j(\theta)/q_0(\theta)\} \\ &\quad + \log \left[\frac{1 + \exp\{\beta_0 + m(g_0, e_k; \beta_1)\}}{1 + \exp\{\beta_0 + m(g_j, e_k; \beta_1)\}} \right]. \end{aligned} \quad (6)$$

Thus, $L^*(\gamma, \mu) = \sum_{ijk} n_{ijk} \log P_{ij}^*(e_k; \gamma, \mu)$ depends on μ_0 and μ_1 only through the parameter $\kappa = \beta_0 + \log(\mu_1/\mu_0)$. Moreover, since

$$\frac{\partial}{\partial \kappa} L^*(\gamma, \kappa) = n_{1++} - \sum_k \sum_j n_{++k} P_{1j}^*(e_k, \gamma, \mu),$$

it follows that $\hat{\kappa}(\gamma) = \beta_0 + \log\{\hat{\mu}_1(\gamma)/\hat{\mu}_0(\gamma)\}$, where $\hat{\mu}_1(\gamma)$ and $\hat{\mu}_0(\gamma)$ are defined in equation (4), will satisfy the equation $(\partial/\partial \kappa)L^*(\gamma, \kappa) = 0$. Thus, the semiparametric maximum likelihood estimate of γ can be obtained by solving the equation $\partial L^*(\gamma, \kappa)/\partial(\gamma, \kappa) = 0$ jointly with respect to γ and κ .

Estimation of β_0 in the above approach requires special attention. From the expression for $\theta_{ij}(e_k; \gamma, \mu)$ given in (6), it can be seen that the intercept parameter β_0 is involved in $L^*(\gamma, \kappa)$ not only through κ but also through the term

$$\tau(e_k, g_j, \beta_0, \beta_1) = \log \left[\frac{1 + \exp\{\beta_0 + m(g_0, e_k; \beta_1)\}}{1 + \exp\{\beta_0 + m(g_j, e_k; \beta_1)\}} \right].$$

Thus, in principle, β_0 is identifiable from $L^*(\gamma, \kappa)$ independently of the parameter κ . However, for diseases that are rare for all combinations of g_j and e_k , that is $\tau(e_k, g_j, \beta_0, \beta_1) \simeq 0$ for all j and k , there would be little information about β_0 from $L^*(\gamma, \kappa)$ that is not absorbed in κ . Since the corresponding information matrix is nearly singular, direct optimisation of $L^*(\gamma, \kappa)$ with respect to β_0 using standard methods such as Newton–Raphson can be numerically unstable. To overcome this problem, one strategy that we have found useful is to consider the profile likelihood of β_0 obtained as $L^*\{\beta_0, \hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$, where $\{\hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$ denotes the solution of the equation $\partial L^*(\beta_0, \beta_1, \theta, \kappa)/\partial(\beta_1, \theta, \kappa) = 0$ for fixed β_0 . One can then perform a one-dimensional grid search for the optimal value of β_0 that maximises $L^*\{\beta_0, \hat{\beta}_1(\beta_0), \hat{\theta}(\beta_0), \hat{\kappa}(\beta_0)\}$, possibly on a fixed interval of values.

In the above approach, for rare diseases, the estimate of the parameter β_0 itself can be expected to be imprecise because of intrinsic noninformativeness of the retrospective likelihood. Much more precise estimation of β_0 is possible when the marginal probability of the disease, $\text{pr}(D = 1)$, in the underlying population is known. We can then fix the parameters μ_i for $i = 0, 1$ in $L^*(\gamma, \mu_0, \mu_1)$ at their true values $n_{i++}/\{N \text{pr}(D = i)\}$ for

$i = 0, 1$, respectively. In the corresponding expression for $\theta_{ij}^*(e_k, \gamma, \mu)$ given in (6), $\log(\mu_1/\mu_0)$ will be fixed and β_0 will be identifiable from the first term of (6) itself. In this case, the parameterisation $\eta = \{\beta_0, \beta_1, \theta, \kappa = \beta_0 + \log(\mu_1/\mu_0)\}$ is unnecessary and instead the original parameterisation $\eta = (\beta_0, \beta_1, \theta)$ should be used. Hereafter, we will use the generic notation η so that our results are valid for both the cases of $\text{pr}(D = 1)$ being known and $\text{pr}(D = 1)$ being unknown.

2.3. Asymptotic theory

In this section, we study the asymptotic properties of the semiparametric maximum likelihood estimator of η . Since we have shown that the estimator can be obtained by solving the equation $\partial L^*(\eta)/\partial \eta = 0$, the asymptotic properties can be studied by estimating-equation theory. Since $L^*(\eta) = \sum_{ijk} n_{ijk} \log P_{ij}^*(e_k, \eta)$, where $P_{ij}^*(e_k, \eta)$ is defined in (5), the estimating function $\partial L^*(\eta)/\partial \eta$ can be expressed as

$$\begin{aligned} \frac{\partial L^*}{\partial \eta} &= \sum_{ijk} n_{ijk} \left[\frac{\partial \theta_{ij}(e_k; \eta)}{\partial \eta} - \sum_{i'j'} \frac{\exp\{\theta_{i'j'}(e_k; \eta)\}}{\sum_{i''j''} \exp\{\theta_{i''j''}(e_k; \eta)\}} \frac{\partial \theta_{i'j'}(e_k; \eta)}{\partial \eta} \right] \\ &= \sum_{u=1}^N \left[\frac{\partial \theta_{D_u G_u}(E_u; \eta)}{\partial \eta} - E_{DG}^* \left\{ \frac{\partial \theta_{DG}(E; \eta)}{\partial \eta} \middle| E = E_u \right\} \right], \end{aligned} \tag{7}$$

where $E_{DG}^*(\cdot|E)$ denotes expectation with respect to the joint probability distribution for D and G given E that was defined by P^* in (2). Define $\Psi(D_u, G_u, E_u; \eta)$ to be the summand in the second expression of formula (7). We will develop the asymptotic theory in a scenario in which the total sample size $N = N_0 + N_1$ goes to infinity, but the sampling proportions for the cases and controls, namely N_0/N and N_1/N , remain fixed at π_1 and $\pi_0 = 1 - \pi_1$, respectively. We first state a lemma, proved in the Appendix, that will be used repeatedly in the development of the asymptotic theory, because in various places we will need to compute expectations and limits of functions in the case-control sampling scheme.

LEMMA 3. Under the case-control sampling design described above and for any measurable function $Q(D, G, E)$ of data (D, G, E) ,

$$N^{-1} \sum_{u=1}^N Q(D_u, G_u, E_u) \rightarrow \int E_{DG}^* \{Q(D, G, E)|E = e\} h(e) dF(e),$$

where the convergence is in probability and $h(e) = \sum_{ij} P_{ij}(e; \beta_0, \beta_1) \mu_i q_j(\theta)$, if we assume that the integral in the above equation exists.

At this point, we note an important subtlety of studying asymptotic theory under the case-control sampling design when the assumption of gene-environment independence is made. If no assumption is made about the joint distribution of (G, E) , that is the form of $\mathcal{H}(g, e)$ is left completely unspecified, then from standard case-control sampling theory it follows that $N^{-1} \sum_{u=1}^N Q(D_u, G_u, E_u) \rightarrow \tilde{E}_{D,G,E} \{Q(D, G, E)\}$, where the convergence is in probability and where $\tilde{E}_{D,G,E}$ corresponds to expectation with respect to a joint distribution function $\tilde{\text{pr}}(D, G, E)$, so that $\tilde{\text{pr}}(G, E|D) = \text{pr}(G, E|D)$ and $\tilde{\text{pr}}(D) = N_D/N$. This follows because, when the form of \mathcal{H} is left unspecified, one can vary the parameters β_0 and \mathcal{H} without changing the value of the retrospective likelihood $\text{pr}(G, E|D)$ (Roeder et al., 1996, Lemma 1). In particular, one can choose $\tilde{\beta}_0$ and $\tilde{\mathcal{H}}$ so that $\text{pr}_{\beta_0, \beta_1, \mathcal{H}}(G, E|D) = \text{pr}_{\tilde{\beta}_0, \beta_1, \tilde{\mathcal{H}}}(G, E|D)$ and $\text{pr}_{\beta_0, \beta_1, \mathcal{H}}(D) = N_D/N$. However, these results do

not hold when one assumes \mathcal{H} to be of the form $Q \times F$ as in this case we have shown that α and \mathcal{H} are uniquely identifiable from the retrospective likelihood, except for some boundary parameter values. Similarly, other standard theories for case-control sampling may not be applicable under the gene-environment independence model.

In Lemma 4, proved in the Appendix, we state the limiting form of the second derivatives of $L^*(\eta)$.

LEMMA 4. *We have that*

$$\frac{1}{N} \frac{\partial^2 L^*}{\partial \eta \partial \eta^T} \rightarrow \int V_{DG}^* \left\{ \frac{\partial \theta_{DG}(E; \eta)}{\partial \eta} \middle| E = e \right\} h(e) dF(e) \equiv \mathcal{I},$$

in probability, where $V_{DG}^*(\cdot|E)$ denotes variance with respect to the joint probability distribution for D and G given E that is defined by P^* .

Finally, we state the main asymptotic limiting results, proved in the Appendix.

PROPOSITION 1. *Under suitable regularity conditions, the following results hold:*

- (i) *the estimating equations $\partial L^*/\partial \eta \equiv \sum_{i=1}^N \Psi(D_i, G_i, E_i; \eta) = 0$ have a unique, consistent sequence of solutions, $\{\hat{\eta}^N\}_{N \geq 1}$;*
- (ii) *if $\Omega = \sum_{d=0}^1 \mu_d [E\{\Psi(D, G, E)|D = d\}]^{\otimes 2}$, then $N^{\frac{1}{2}}(\hat{\eta}^N - \eta_0) \rightarrow N(0, \Sigma)$ in distribution, with*

$$\Sigma = \mathcal{I}^{-1} - \mathcal{I}^{-1} \Omega \mathcal{I}^{-1}. \quad (8)$$

3. EXTENSIONS

3.1. Population stratification

Although genetic susceptibility and environmental exposures are unlikely to be causally related at an individual level, these factors may be correlated at a population level because of their dependence on other factors, such as ethnicity. In this section, we briefly describe how to generalise our methods to handle ‘population stratification’. Most of the details and proofs of the theoretical results follow from straightforward generalisation of the results derived in § 2.

We will assume that G and E are independent conditional on a set of variables S so that the joint distribution of G , E and S is given by the product form $H(g, e, s) = Q_s(g) \times F(e, s)$, where $Q_s(g)$ corresponds to the distribution of G given $S = s$ and $F(e, s)$ denotes the joint distribution of E and S . The distribution function $F(e, s)$ will be treated nonparametrically. Let $\text{pr}(G = g_j | S = s)$ be denoted by $q_j(s; \theta)$ with θ being a fixed set of parameters characterising the conditional distribution. If S involves only discrete variables that define a relatively small number of strata, then no modelling of $\text{pr}(G = g_j | S = s)$ is necessary and θ may denote the vector of conditional probabilities themselves. If S involves a relatively large number of variables, possibly including continuous ones, parametric modelling of the distribution $\text{pr}(G|S)$ will be necessary. When G is a binary variable indicating the presence or absence of a certain genetic variation, for example, $\text{pr}(G|S)$ can be parametrically specified through a logistic regression model. We further assume that the disease-risk model is given by $\text{pr}(D = 1 | G, E, S) = H\{\beta_0 + m(G, E, S; \beta_1)\}$. Thus, we allow the stratum variables S to be covariates of interest in the disease model. Let (e_k, s_k) , for $k = 1, \dots, K$, be the unique observed values for (E, S) and let n_{ijk} denote the number of subjects in the data with $D = i$, $G = j$ and $(E, S) = (e_k, s_k)$. As before, let $\mu_i = n_{i++} / \{N \text{pr}(D = i)\}$.

With this notation, the results of Lemma 4 can be generalised to show that the semi-parametric maximum likelihood estimator of $\gamma = (\beta_0, \beta_1, \theta)$ can be obtained by solving the equation $\partial L^*(\gamma, \mu)/\partial(\gamma, \mu) = 0$ jointly with respect to (γ, μ) , where

$$L^*(\gamma, \mu) = \sum_{ijk} n_{ijk} \log P_{ij}^*(e_k, s_k; \gamma, \mu)$$

and $P_{ij}^*(e_k, s_k; \gamma, \mu)$ is defined by formula (2) with $P_{ij}(e_k, \beta_0, \beta_1)$ and $q_j(\theta)$ replaced by $P_{ij}(e_k, s_k, \beta_0, \beta_1)$ and $q_j(s; \theta)$, respectively. Moreover, $P_{ij}^*(e_k, s_k; \gamma, \mu)$ can be written in the form of expression (5) with $\theta_{ij}(e_k; \gamma, \mu)$ replaced by $\theta_{ij}(e_k, s_k; \gamma, \mu)$, which in turn is defined by equation (6) with $q_j(\theta)/q_0(\theta)$ and $m(g, e_k; \beta_1)$ replaced by $q_j(s; \theta)/q_0(s; \theta)$ and $m(g, e_k, s_k; \beta_1)$, respectively. All the theory that we developed in § 2.3 can now be generalised by replacing E with $E' = (E, S)$ and $q_j(\theta)$ by $q_j(s; \theta)$ everywhere.

3.2. Frequency-matched case-control studies

In this section, we comment briefly on the modifications needed for the proposed methodology while dealing with frequency-matched case-control studies in which controls are selected in numbers proportional to the number of cases within strata defined by some matching variables W . The problem of individually-matched case-control studies is addressed in a separate article (Chatterjee et al., 2005). Let $W = w_m$ ($m = 1, \dots, M$) denote M strata used for matching. To allow for factors, such as race, which may be candidates for both matching and population stratification, we write $W = (W^S, W^{\bar{S}})$, so that W^S represents the elements of W that are included in S , the factors for population stratification. Similarly, we write $S = (S^W, S^{\bar{W}})$, so that S^W denotes elements of S that are included in W . We will assume that G is independent of (E, W^S) conditional on S . We further assume that the regression model is given by

$$\text{pr}(D = 1 | G, E, S^W, W) = H\{\beta_{0W} + m(G, E, S^W, W; \beta_1)\},$$

so that it corresponds to the standard practice of allowing for an independent intercept term for each level of the matching variable $W = w$. Let $\beta_0 = (\beta_{01}, \dots, \beta_{0M})$ be the vector of intercept parameters corresponding to the M different values of W .

With the above notation and definitions, the retrospective likelihood for the matched case-control design can be written as

$$\ell_{\text{MCC}} = \prod_{i=1}^{N_0 + N_1} \text{pr}(G_i, E_i, S_i^{\bar{W}} | D_i, W_i),$$

where the conditioning on (D, W) represents the fact that in a matched case-control design subjects are selected into the study based on both the disease status D and the matching variable W . The semiparametric maximum likelihood estimator of $\gamma = (\beta_0, \beta_1, \theta)$ that leaves the joint distribution of (E, S, W) completely unspecified can be derived by following techniques of §§ 2.2, 2.3 and 3.1, with μ_i replaced throughout by μ_{wi} , where $\mu_{wi} = n_{wi++}/\{N \text{pr}(D = i | W)\}$, in which n_{wi++} is the number of subjects with $D = i$ and $W = w$ in the sample. In particular, we can show that the semiparametric maximum likelihood estimate of γ can be obtained by jointly solving a set of equations of the form $\partial L^*(\gamma, \kappa)/\partial(\gamma, \kappa) = 0$, where $\kappa = (\kappa_1, \dots, \kappa_M)$ with $\kappa_m = \beta_{0m} + \log(\mu_{1m}/\mu_{0m})$ and $L^*(\gamma, \kappa) = \sum_{ijk} n_{wijk} \log P_{wij}^*(e_k, s_k; \gamma, \mu)$, in which $P_{wij}^*(e_k, s_k; \gamma, \mu)$ is defined by formula (5)

with $\theta_{ij}(e_k; \gamma, \mu)$ replaced by

$$\begin{aligned} \theta_{wij}(e_k, s_k; \gamma, \mu) &= \log \left\{ \frac{P_{wij}^*(e_k, s_k; \gamma, \mu)}{P_{w00}^*(e_k, s_k; \gamma, \mu)} \right\} \\ &= ik_w + im(g_j, e_k, s_k^{\bar{w}}, w; \beta_1) + \log \{q_j(s_k; \theta)/q_0(s_k; \theta)\} \\ &\quad + \log \left[\frac{1 + \exp \{\beta_{0w} + m(g_0, e_k, s_k^{\bar{w}}, w; \beta_1)\}}{1 + \exp \{\beta_{0w} + m(g_j, e_k, s_k^{\bar{w}}, w; \beta_1)\}} \right]. \end{aligned} \quad (9)$$

Using the structure of $\theta_{wij}(\cdot)$ we observe that β_{0w} is involved in $L^*(\gamma, \kappa)$ not only through κ_w but also through the last term of expression (9), which we will denote by $\tau(g_j, e_k, s_k^{\bar{w}}, w; \beta_{0w}, \beta_1)$. For rare diseases, however, for which $\tau(g_j, e_k, s_k^{\bar{w}}, w; \beta_{0w}, \beta_1) \simeq 0$ for all values of j and k there would be little information about β_{0w} from $L^*(\gamma, \kappa)$ that is not absorbed in κ_w . In most case-control studies the matching factor W consists of basic demographic factors such as race, sex and age-groups, for which $\text{pr}(D = 1|W)$ is available externally, for example from a population registry. In this case, μ_{wi} can be treated as a fixed parameter in the definition of κ_w and hence β_{0w} can be identified through κ_w itself. Use of external information about $\text{pr}(D = 1|W)$ is recommended as it would not only resolve any numerical problems that may arise with estimation of the barely identifiable parameters, but would also improve efficiency of estimation of the other regression parameters of interest. An alternative solution for diseases that are extremely rare, such as the example of ovarian cancer we consider in § 5, is to ignore the term $\tau(g_j, e_k, s_k^{\bar{w}}, w; \beta_{0w}, \beta_1)$ in the calculations. Under the rare-disease assumption, we note that the functional form of L^* becomes exactly the same under frequency-matched and traditional unmatched case-control sampling designs, and thus the estimates under the matched design can be obtained by the method described in § 2 for the traditional case-control design with a disease risk model that allows for an independent intercept term for each level of the matching variable W . The estimator for the main effects for W would yield an unbiased estimator not of β_{0w} but of κ_w .

4. SIMULATION STUDY

4.1. *The factors G and E are independent*

In the first experiment, we study the relative performance of the standard logistic regression analysis and the proposed semiparametric maximum-likelihood estimator under the gene-environment independence model. We assumed that the genetic covariate G is a binary variable, where for example $G = 1$ or $G = 0$ corresponds to presence or absence of a genetic mutation, respectively. We considered two scenarios: (a) $\text{pr}(G = 1) = 0.065$ and (b) $\text{pr}(G = 1) = 0.26$, corresponding to a rare and a common genetic mutation, respectively. We generated the environmental covariate as $E = \min(10, X)$ where X follows the log-normal distribution for which the mean and variance of the underlying normal distribution are 0 and 1. Given the values of (G, E) , we generated a binary disease outcome D from the logistic regression model $\text{logit}\{\text{pr}(D|G, E)\} = \beta_0 + \beta_G G + \beta_E E + \beta_{GE} G \times E$, with $(\beta_G, \beta_E, \beta_{GE}) = (0.26, 0.10, 0.3)$. We choose the intercept parameter β_0 to be, respectively, -3.2 and -3.45 for scenarios (a) and (b) so that in both cases the marginal probability of the disease in the population is 0.05. The parameter values were chosen to reflect modest main effects for both G and E , but strong interaction between G and E . For example, the

odds ratio associated with the lower versus the upper quartile of the distribution of E was 1.3 for $G = 0$ and 3.1 for $G = 1$. The marginal odds ratios for G and E were 2.6 and 2.5, respectively. In each replication of our simulation experiment, we generated data for 500 cases and 500 controls from the above model by sampling the cases and controls from a larger random sample of subjects. We analyse each such case-control dataset using three procedures: standard logistic regression; SPMLE_1 , which denotes the proposed semiparametric maximum likelihood method under the gene-environment independence model when $\text{pr}(D = 1)$ is known; and SPMLE_2 , which denotes the same procedure but with $\text{pr}(D = 1)$ unknown.

Table 1 summarises the simulation results for scenarios (a) and (b). Based on these simulation results we make the following key observations. First, as expected from theory, both the logistic regression and the semiparametric maximum likelihood estimators under the correct conditional independence assumption provide essentially unbiased estimators of all regression parameters. Secondly, the variance ratios of the semiparametric maximum likelihood and logistic regression estimator show that when the gene-environment independence assumption is exploited there is a major efficiency gain for the estimation of β_G and β_{GE} ; the gain is quite dramatic for estimation of the interaction parameter β_{GE} and is larger for the study of the rare mutation than for the common mutation. Thirdly, under the gene-environment independence model, incorporating the known $\text{pr}(D = 1)$ in the estimation leads to major efficiency gains in the estimation of the regression parameters, the gain being particularly striking for β_{GE} . This observation is particularly interesting given that it is well known that in the standard logistic regression setting, when no assumption is made about the exposure distribution, use of the known marginal probability of the disease in the population only identifies the intercept parameter of the logistic regression model, but does not have any effect on the efficiency of the estimators of the other regression parameters of interest. Fourthly, comparison of the empirical standard errors and the means of the estimated standard errors of the semiparametric maximum likelihood estimator shows that the proposed sandwich variance estimator performs well for realistic parameter values and modest sample sizes.

Table 1. *Simulation study for studying bias and efficiency of semiparametric maximum-likelihood estimators when G and E are independent: SPMLE_1 , the proposed method when the marginal probability $\text{pr}(D = 1)$ is known; SPMLE_2 , the proposed method when $\text{pr}(D = 1)$ is unknown*

	Logistic regres.	Bias		Var ratio		Empirical SE		Estimated SE	
		SPMLE_1	SPMLE_2	$\frac{\text{SPMLE}_1}{\text{Logistic}}$	$\frac{\text{SPMLE}_2}{\text{Logistic}}$	SPMLE_1	SPMLE_2	SPMLE_1	SPMLE_2
Scenario (a): $\text{pr}(G = 1) = 0.05$									
β_G	0.033	0.021	0.034	0.629	0.818	0.282	0.322	0.275	0.327
β_E	-0.002	0.000	0.002	0.900	0.991	0.035	0.037	0.034	0.035
β_{GE}	-0.032	-0.009	-0.023	0.264	0.535	0.090	0.128	0.087	0.126
θ	.	0.020	0.021	.	.	0.164	0.187	0.167	0.194
Scenario (b): $\text{pr}(G = 1) = 0.2$									
β_G	0.016	0.004	0.015	0.709	0.905	0.175	0.198	0.171	0.195
β_E	0.001	0.002	0.001	0.769	0.987	0.038	0.043	0.037	0.041
β_{GE}	-0.013	-0.006	-0.011	0.360	0.717	0.053	0.075	0.052	0.074
θ	.	0.003	-0.001	.	.	0.092	0.105	0.095	0.108

regres., regression; Var, variance; SE, standard error.

The above simulation set-up also allows us to study bias in parameter estimation in existing approximate methods that rely on the rare-disease assumption. Schmidt & Schaid (1999) noted that, even for rare diseases like breast cancer, the ‘case-only’ analysis approach to interaction that is based on the rare disease assumption can seriously underestimate the logistic regression interaction parameters for studying major susceptibility genes such as the BRCA1 and BRCA2 genes, which are known to confer a very high risk of breast and ovarian cancer. Our simulation gives an alternative relevant scenario involving a continuous environmental exposure variable where the gene or the environmental exposures themselves do not pose a very high risk of the disease, but among the mutation carriers there is a strong dose-response relationship between the risk of the disease and the continuous exposure. We examined the bias in estimation of the interaction parameter β_{GE} in two approximate methods, the case-only analysis (Piegorisch et al., 1994) and the combined control group approach of Modan et al. (2001). We did not implement the log-linear model approach for categorical covariates (Umbach & Weinberg, 1997) as in our simulation the environmental covariate E was continuous. We found that on average the case-only estimates of β_{GX} were 0.189 and 0.212 in scenarios (a) and (b), respectively. The corresponding average estimates obtained from the approach of Modan et al. are 0.194 and 0.229, respectively. Given that the true value of the interaction parameter was 0.30, in each of the scenarios we considered, the approximate methods seriously underestimate the odds-ratio interaction parameter.

4.2. Factors G and E are independent conditional on S

We considered a second simulation experiment in which the independence assumption between G and E holds only within subpopulations defined by a stratum variable S . As before, we considered two scenarios, one for a rare mutation and one for a common mutation, but in each situation we now assume that the gene frequency differs across strata defined by S : we took $\theta_1 = \text{pr}(G = 1|S = 1) = 0.05$ and $\theta_2 = \text{pr}(G = 1|S = 2) = 0.1$ in scenario (a) and $\theta_1 = \text{pr}(G = 1|S = 1) = 0.2$ and $\theta_2 = \text{pr}(G = 1|S = 2) = 0.4$ in scenario (b). We assumed that $\text{pr}(S = 2) = 0.3$. Also, as before, we generated the environmental covariate as $E = \min(10, X)$, where X follows a log-normal distribution, but we allowed the mean parameter for the underlying normal distribution to be different across strata defined by S . In particular, we used the values of $\mu_1 = 0$, $\mu_2 = 0.67$ and $\sigma_1 = \sigma_2 = 1$ so that the 75th percentile of the distribution of $X|S = 1$ corresponds to only the 50th percentile of the distribution of $X|S = 2$. We also assumed that the stratification variable S is a risk factor for the disease and hence is part of the risk model. We allowed both a main effect, β_S , and an interaction of S with G , β_{GS} , in the disease risk model with the true parameter values being $\log(2)$ and $\log(3)$, respectively. As before, we assumed $(\beta_G, \beta_X, \beta_{GX}) = (0.26, 0.1, 0.3)$. We generated 500 simulated datasets, each dataset consisting of observations on (G, X, S) for 500 cases and 500 controls. We analysed each such case-control dataset using three procedures: standard logistic regression; SPMLE(CS), which denotes the proposed method under the correctly specified independence model that assumes G is independent of E given S ; and SPMLE(MS), based on a misspecified independence model that assumes G is independent of (E, S) . For both of the latter two procedures, we assumed $\text{pr}(D)$ was known.

The results in Table 2 stimulate the following key observations. First, when the correct model is that G and E are independent given S , but we assume the misspecified model in which G is independent of both E and S , estimators of β_G , β_S and β_{GS} can be seriously

Table 2. Simulation study for studying bias and efficiency of semiparametric maximum-likelihood estimators when G and E are independent conditional on a stratification variable S : SPMLE(CS), our method when the probability model for G and E given S is correctly specified; SPMLE(MS), our method when this model is misspecified

	Logistic regres.	Bias		MSE ratio		Empirical SE		Estimated SE	
		SPMLE (CS)	SPMLE (MS)	<u>SPMLE(CS)</u> Logistic	<u>SPMLE(MS)</u> Logistic	SPMLE (CS)	SPMLE (MS)	SPMLE (CS)	SPMLE (MS)
pr($G = 1 S = 1$) = 0.05 and pr($G = 1 S = 2$) = 0.1									
β_G	0.036	0.016	0.518	0.680	1.605	0.397	0.323	0.393	0.338
β_E	-0.002	-0.001	0.008	0.827	0.847	0.030	0.029	0.029	0.029
β_S	-0.008	-0.009	0.078	0.981	1.195	0.153	0.150	0.154	0.150
β_{GE}	-0.044	-0.021	-0.026	0.273	0.242	0.088	0.081	0.088	0.085
β_{GS}	-0.005	0.018	-0.940	0.879	3.386	0.508	0.338	0.481	0.343
pr($G = 1 S = 1$) = 0.2 and pr($G = 1 S = 2$) = 0.4									
β_G	0.014	0.017	0.473	0.874	3.292	0.278	0.263	0.269	0.252
β_E	-0.002	0.001	0.016	0.736	0.865	0.037	0.036	0.039	0.038
β_S	0.003	0.003	0.342	0.976	3.235	0.221	0.213	0.222	0.211
β_{GE}	-0.007	-0.007	-0.025	0.472	0.569	0.054	0.054	0.054	0.056
β_{GS}	-0.025	-0.025	-1.101	0.953	11.468	0.326	0.273	0.337	0.269

regres., regression; MSE, mean squared error; SE, standard error.

biased, with the bias of the interaction parameter being the most striking. Secondly, the ratio of the mean squared error for SPMLE(CS) and for the logistic regression analysis shows that when the correct conditional independence model was exploited there was a major efficiency gain in estimating β_G , β_E and β_{GE} , the gain being most dramatic for estimation of β_{GE} . The corresponding ratio of the mean squared errors for SPMLE(MS) shows that for those parameters, where SPMLE(MS) produces large bias, the mean squared error for SPMLE(MS) tends to be much larger than that for the logistic regression analysis. For the parameters β_E and β_{GE} , however, where there is very little bias in SPMLE(MS), both SPMLE(MS) and SPMLE(CS) have similar mean squared errors. Thus, if we adjust properly for the stratification variable S in the independence model, we can correct for bias in estimating β_G , β_S and β_{GS} and yet can retain the efficiency advantage resulting from the gene-environment independence assumption. Thirdly, comparison of the empirical standard errors and the means of the estimated standard errors of the semiparametric maximum likelihood estimators shows that the proposed variance estimator performs well under the population-stratification model.

5. ISRAELI OVARIAN CANCER STUDY

In this section, we apply the proposed methodology to data from a population-based case-control study based on all ovarian cancer patients identified in Israel between 1 March 1994 and 30 June 1999 (Modan et al., 2001). For each case, two controls were selected from the central population registry matched by age within two years, area of birth and place and length of residence. Blood samples were then collected from the cases and the controls in order to test for the presence of mutation in the two major breast and ovarian cancer susceptibility genes BRCA1 and BRCA2. In addition, the subjects were interviewed to collect data on reproductive/gynaecological history such as parity, number of years of

oral contraceptive use and gynaecological surgery. The main goal of the study was to examine the interplay of the BRCA1/2 genes and known reproductive/gynaecological risk factors of ovarian cancer.

Modan et al. (2001) studied the interaction between BRCA1/2 mutations and two known reproductive risk factors for ovarian cancer, namely oral contraceptive use and parity. They pointed out that, since BRCA1/2 mutations were very rare among ovarian cancer controls, traditional logistic regression analysis would yield very imprecise estimates of the various regression parameters of interest. Thus they considered alternative efficient methods of analysis that exploit the likely scenario that the status of BRCA1/2 mutations is independent of the reproductive risk factors. In particular, they estimated the odds ratio of ovarian cancer associated with the reproductive risk factors separately for carriers and non-carriers by using the combined common control group approach that we described in § 1. In addition, to test if the effects of the reproductive risk factors are different for BRCA1/2 mutation carriers and non-carriers, the authors performed the ‘case-only’ analysis of interaction (Piegorisch et al., 1994).

We reanalysed the data using the proposed maximum likelihood method under the gene-environment independence assumption. Our analysis included 832 cases and 747 controls who did not have bilateral oophorectomy, were interviewed for risk factor information and successfully tested for BRCA1/2 mutations. There were 240 carriers, but only 12 among the controls. Similarly to Modan et al., we coded reported parity values greater than 10 to be 10. In addition, we deleted three women with extreme oral contraceptive use, of at least 250 months, as they became highly ‘influential’ for the estimation of regression parameters. We considered the following logistic regression model for risk of ovarian cancer:

$$\begin{aligned} \text{logit}\{\text{pr}(D = 1)\} = & \beta_0 + \beta_{\text{BRCA1/2}}I(\text{BRCA1/2}) + \beta_{\text{OC}}\text{OC} + \beta_{\text{Par}}\text{Parity} \\ & + \beta_{\text{BRCA1/2*OC}}I(\text{BRCA1/2})*\text{OC} \\ & + \beta_{\text{BRCA1/2*Par}}I(\text{BRCA1/2})*\text{Parity} + \gamma^T Z, \end{aligned}$$

where $I(\text{BRCA1/2})$ denotes the 0–1 indicator of carrying at least one BRCA1/2 mutation, OC denotes years of oral contraceptive use, Parity denotes the number of children and Z denotes the set of all co-factors that Modan et al. used to adjust their regression analysis; Z included the main effects of age, as a categorical variable defined by decades, ethnic background, being Ashkenazi or non-Ashkenazi, the presence of personal history of breast cancer, PHB, history of gynaecological surgery, and family history of breast or ovarian cancer, FHBO, where 0 corresponds to no history in the family, 1 to one breast cancer case in the family and 2 to ovarian cancer or two or more breast cancer cases in the family.

Next we considered an appropriate model for gene-environment independence. Clearly, a personal history of breast cancer and family history of breast/ovarian cancer cannot be assumed to be independent of BRCA1/2 status as mutations in these genes are known to increase dramatically the risk of these familial cancers. Moreover, BRCA1/2 mutation frequency has been reported in the past to vary by age and ethnicity. Given that some of these factors can also be related to oral contraceptive use and parity, we make the assumption of independence between mutation and reproductive risk factors only conditional on $S = (\text{Age}, \text{Ethnicity}, \text{PHB}, \text{FHBO})$. Given that the total number of strata defined by S is large, estimation of the genotype frequencies individually for each stratum would be imprecise. Thus, we considered the following parametric model for specification of the

carrier frequencies:

$$\begin{aligned} \text{logit}\{\text{pr}(G = 1|S)\} = & \theta_0 + \theta_{\text{Age}}I(\text{Age} \geq 50) + \theta_{\text{Eth}}I(\text{Non-Ashkenazi}) \\ & + \theta_{\text{PH}}I(\text{PHB} = 1) + \theta_{\text{1FH}}I(\text{FHBO} = 1) + \theta_{\text{2FH}}I(\text{FHBO} = 2). \quad (10) \end{aligned}$$

Modan et al. had reported a total of 1326 cases of peritoneal or epithelial ovarian cancer during the five-year study period, in a baseline population of approximately 1.5 million. Thus, the marginal probability for the disease for the underlying population is small, at about $\text{pr}(D = 1) = 8.7 \times 10^{-4}$. Therefore, based on the discussion in § 3.2, we note that we can analyse data from this age-matched case-control study using methods developed for ordinary case-control studies as long as we allow for an independent intercept term for each of the age-strata that were used for matching. Although the cases and controls were matched to within two years, to avoid problems with sparse cells we allowed an independent intercept term only for every 10-year interval. This approximation, the validity of which requires assumptions similar to those required for unconditional logistic regression analysis of matched data, is reasonable for this study.

Table 3 shows the estimates and 95% confidence intervals corresponding to the regression parameters associated with the main covariates of interest: BRCA1/2, oral contraceptive use and parity. Two sets of estimates and confidence intervals are shown, one corresponding to an ordinary logistic regression analysis of the case-control data and the other corresponding to our method estimated under the conditional gene-environment independence model. Based on the ordinary logistic regression estimates of the main effect parameters, we first observe that, among childless women, for whom Parity = 0, and who never used oral contraceptives, BRCA1/2 mutation is associated with a dramatic increase in risk of ovarian cancer, with odds ratio $\exp(3.58) = 35.87$. Among BRCA1/2 non-carriers, both higher parity and longer use of oral contraceptives are associated with decreased risk of ovarian cancer, with the associated odds ratio parameters estimated to be respectively 0.95 and 0.94 for Parity; both of these results are borderline statistically significant at the 5% level. The estimates of the interaction parameters from the logistic regression analysis suggest that, among BRCA1/2 carriers, the risk of ovarian cancer decreases even more strongly with increasing parity, with odds ratio $\exp(-0.058) \times \exp(-0.199) = 0.77$, but increases slightly with longer oral contraceptive use, with odds ratio $\exp(-0.047) \times \exp(0.056) = 1.01$. However, the confidence intervals for the interaction parameters are very wide, suggesting that the point estimates are imprecise and hence hard to interpret.

Table 3. *Parameter estimates and confidence intervals for the risk model in the Israeli ovarian cancer study*

	Ordinary logistic regression		MLE with G-E independence given <i>S</i>	
	Estimate	95% CI	Estimate	95% CI
BRCA1/2	3.58	(2.27, 4.89)	3.15	(2.51, 3.79)
oc use	-0.047	(-0.098, 0.003)	-0.051	(-0.102, -0.001)
Parity	-0.058	(-0.121, 0.004)	-0.061	(-0.125, 0.002)
oc*BRCA1/2	0.056	(-0.149, 0.260)	0.089	(0.021, 0.150)
Parity*BRCA1/2	-0.199	(-0.626, 0.229)	-0.036	(-0.141, 0.068)

MLE, maximum likelihood estimate; G-E, gene-environment; CI, confidence interval; *S* = (Age, Ethnicity, Personal history of breast cancer, Family history of breast/ovarian cancer); oc, oral contraceptive.

Inspection of the parameter estimates from the semiparametric maximum likelihood method with the gene-environment independence model suggests similar types of association to those from the logistic-regression analysis. However, the precisions of the estimates are greater for all the terms involving BRCA1/2, the gain being particularly striking for the interaction terms. In particular, under the gene-environment independence model, the interaction between BRCA1/2 mutation and oral contraceptive use is statistically significant, suggesting that, unlike for non-carriers, the risk of breast cancer for carriers did not decrease with increasing oral contraceptive use. For carriers, the association between oral contraceptive use and risk of ovarian cancer, if any, is positive, with odds ratio $\exp(-0.051 + 0.089) = 1.034$, and 95% confidence interval (0.977, 1.095). The interaction estimate between Parity and BRCA1/2 suggests that the decrease in risk of ovarian cancer associated with increased parity is modestly larger for carriers than for non-carriers, but this difference is not statistically significant.

Table 4 shows the maximum likelihood estimates corresponding to the model for carrier frequency $\text{pr}(G = 1|S)$. Although these parameters do not have any causal interpretation and are not generally of biological interest, they can be useful for descriptive purposes. For example, as expected, prevalence of a BRCA1/2 mutation is significantly higher among women with either a personal history of breast cancer or family history of breast/ovarian cancer. Moreover, we observe that BRCA1/2 mutation frequency is significantly lower among non-Ashkenazi Jewish women compared to Ashkenazi women. There is also some evidence, with p -value = 0.05, that carrier frequency was smaller among women older than 50 than among younger women.

Table 4. *Parameter estimates and confidence intervals for the logistic regression model for $\text{pr}(G = 1|S)$ in the Israeli ovarian cancer study, with risk factors ethnicity, age, personal history of breast cancer, family history of breast cancer and family history of breast/ovarian cancer*

	θ_0	θ_{Eth}	θ_{Age}	θ_{PH}	$\theta_{1\text{FH}}$	$\theta_{2\text{FH}}$
Estimate	-3.78	-1.31	-0.28	1.59	0.71	1.32
95% CI	(-4.40, -3.16)	(-1.74, -0.890)	(-0.638, 0.071)	(1.01, 2.18)	(0.20, 1.21)	(0.74, 1.90)

CI, confidence interval.

A version of the dataset is available together with the software from the website <http://dceg.cancer.gov/people/ChatterjeeNilanjan.html> under the software link. This dataset consists of the real data on disease status, D , and non-genetic co-factors, X . For reasons of privacy, however, the real genetic data are not publicly available. Instead, the data consist of simulated genetic data, G , generated using the conditional distribution of $[G|D, X]$ as specified by the parameter estimates obtained from the real data.

6. DISCUSSION

Case-control studies with modest sample sizes often have very little power for studying interaction and other hypotheses of interest using the standard logistic regression analysis. In such situations, epidemiological researchers currently have been prone to exploit the efficiency advantage from the gene-environment independence assumption through the case-only approach that yields estimate of the multiplicative interaction parameter in

the logistic regression model under the rare disease assumption (Piegorsch et al., 1994). This analysis, however, is limited. It discards all the information from controls and hence loses the ability to estimate the main effect parameters of the logistic regression model which are required for deriving the various alternative scientific parameters of interest. In this paper, we have considered estimation of regression parameters under the gene-environment independence assumption in a very general logistic regression model that uses data from both cases and controls and hence can estimate all of the parameters of interest.

However, we recommend cautious use of the gene-environment independence assumption. Simulation studies reported in § 4.2, as well as those in Albert et al. (2001), show that methods that use the gene-environment independence assumption when the assumption is not true may produce severe bias in parameter estimation. We have proposed a possible remedy for minimising such bias by explicitly accounting for observable factors, denoted by S , that can potentially be related to both G and E .

Methods for exploiting the gene-environment independence assumption could be practically useful without concerns about bias in many important situations. For 'randomised exposure' such as the treatment assigned in a randomised trial, the gene-environment independence assumption would be satisfied by the definition of randomisation. The assumption of gene-environment independence is also very likely to be satisfied for external environmental agents, e.g. carcinogens from a nearby chemical factory, exposure to which is not directly controlled by an individual's own behaviour. When an exposure depends on subject's individual behaviour, on the other hand, the independence assumption should be used more cautiously. There could be spurious association between G and E for established risk factors such as smoking because family history of lung cancer, which is associated with G , may also influence a subject to change his/her smoking behaviour. There could also be direct association. For example, genetic polymorphisms in the smoking metabolism pathway may not only modify a subject's risk from smoking, but may also influence a subject's degree of addiction to smoking.

When violation of the gene-environment independence seems plausible, because of direct or indirect association, effort should be made to validate the assumption empirically. However, tests for independence within a given study may have very little power, and empirical evidence from external data sources should be investigated. When substantial uncertainty remains about the validity of the assumption because of lack of empirical data or for other reasons, positive findings based on proposed methodology should be considered as preliminary screen which should be pursued with high 'priority' in future epidemiological studies.

In practice, genetic and/or environmental exposure data can be also missing on certain study subjects, by design or by change. Umbach & Weinberg (1997) described a number of alternative designs in which genetic and/or environmental exposure data are collected only on a subset of controls. They showed how different parameters of interest can be estimated under different designs using the approximate log-linear model approach for categorical variables. Further research is warranted to extend the proposed maximum-likelihood methodology to handle missing data in genetic as well as environmental exposures. Such extensions will also be useful to haplotype-based associated studies where genetic effects are modelled in terms of 'haplotypes', the combination of alleles at multiple loci in a single chromosome, but the exact haplotype configuration in two chromosomes of some subjects cannot be derived with certainty from available locus-specific genotype data.

ACKNOWLEDGEMENT

Carroll’s research was supported by a grant from the National Cancer Institute, and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences. We would like to thank Dr Sholom Wacholder for sharing his valuable insights about gene-environment studies.

APPENDIX

Proofs

Proof of Lemma 1. By Lemma 1 of Roeder et al. (1996), it follows that the probability equality of our Lemma 1 holds only if

$$d\mathcal{H}^*(G, E) = \frac{[1 + \exp\{\beta_0^* + m(G, E; \beta_1)\}]/[1 + \exp\{\beta_0 + m(G, E; \beta_1)\}]d\mathcal{H}(G, E)}{\sum_g \int_e [1 + \exp\{\beta_0^* + m(g, e; \beta_1)\}]/[1 + \exp\{\beta_0 + m(g, e; \beta_1)\}]d\mathcal{H}(g, e)}.$$

If \mathcal{H} is of the product form $Q \times F$, $\mathcal{H}^* \neq \mathcal{H}$ could be of the product form only if $\beta_1 \in \mathcal{B}^0$. Thus, if $\beta_1 \notin \mathcal{B}^0$, then $F = F^*$ and $Q = Q^*$. Moreover, since $\mathcal{H}^* = \mathcal{H}$, it also follows that $\beta_0 = \beta^*$. \square

Proof of Lemma 2. By equating the partial derivatives of the loglikelihood given in equation (1) with respect to $\delta_1, \dots, \delta_K$, we can easily show that $\hat{\delta}_k(\gamma)$ will satisfy the equation

$$\delta_k = \frac{n_{++k}}{\sum_{ij} P_{ij}(e_k, \beta) \hat{\mu}_i q_j(\theta)}, \tag{A1}$$

where

$$\hat{\mu}_i = \frac{n_{i++}}{\text{pr}(D = i)} = \frac{n_{i++}}{\sum_{j'k'} P_{ij'}(e_{k'}, \beta) q_{j'}(\theta) \delta_{k'}}. \tag{A2}$$

If we now substitute the left-hand side of (A1) for δ_k into the loglikelihood of the data defined in (1), we obtain

$$\begin{aligned} L\{\gamma, \hat{\delta}(\gamma)\} &= \sum_{ijk} n_{ijk} \log P_{ij}(e_k, \beta_0, \beta_1) + \sum_j n_{+j+} \log q_j(\theta) \\ &+ \sum_k n_{++k} \log \frac{n_{++k}}{\sum_{ij'} P_{ij'}(e_k; \beta_0, \beta_1) \hat{\mu}_i q_{j'}(\theta)} - \sum_i n_{i++} \log \frac{n_{i++}}{\hat{\mu}_i(\gamma)}, \end{aligned}$$

which is equivalent to $L^*\{\gamma, \hat{\mu}(\gamma)\}$ up to constant terms. Moreover, if we substitute (A1) into (A2) it can be seen that $\hat{\mu}_i(\gamma)$, for $i = 0, 1$, are given by solutions of the equations

$$n_{i++} = \sum_{k'} n_{++k'} \frac{\sum_{j'} P_{ij'}(e_{k'}; \beta) q_{j'}(\theta) \mu_i}{\sum_{ij} P_{ij}(e_{k'}; \beta) q_j(\theta) \mu_i} \quad (i = 0, 1),$$

which are in turn equivalent to the equations given in (4). Thus Lemma 2 is proved. \square

Proof of Lemma 3. First, we note that, by the law of large numbers,

$$\begin{aligned} \frac{1}{N} \sum_{u=1}^N Q(D_u, G_u, E_u) &\equiv \frac{N_0}{N} \frac{1}{N_0} \sum_{u=1}^{N_0} Q(D_u, G_u, E_u) + \frac{N_1}{N} \frac{1}{N_1} \sum_{u=1}^{N_1} Q(D_u, G_u, E_u) \\ &= \mu_0 \text{pr}(D = 0) E\{Q(D, G, E) | D = 0\} \\ &+ \mu_1 \text{pr}(D = 1) E\{Q(D, G, E) | D = 1\} + o_p(1). \end{aligned} \tag{A3}$$

Using Bayes’ rule we can write

$$\text{pr}(D = d) E\{Q(D, G, E) | D = d\} = \int \left[\sum_j \{Q(d, g_j, e) q_j(\theta)\} \right] dF(e).$$

Thus, we can write the limiting expression in (A3) as

$$\int_e \left\{ \sum_{ij} \frac{Q(i, g_j, e) P_{ij}(e; \beta_0, \beta_1) \mu_i q_j(\theta)}{h(e)} \right\} h(e) dF(e).$$

The proof of Lemma 3 follows if we note that $P_{ij}^*(e; \gamma, \mu) = \mu_i P_{ij}(e; \beta_0, \beta_1) q_j(\theta)/h(e)$. \square

Proof of Lemma 4. By applying the chain rule of derivatives to formula (7) we have

$$\begin{aligned} \frac{1}{N} \frac{\partial^2 L^*}{\partial \eta \partial \eta'} &= \sum_{u=1}^N \left[\frac{\partial^2 \theta_{D_u G_u}(E_u, \eta)}{\partial \eta \partial \eta'} - E_{DG}^* \left\{ \frac{\partial^2 \theta_{DG}(E, \eta)}{\partial \eta \partial \eta'} \middle| E = E_u \right\} \right] \\ &\quad - \sum_{u=1}^N \sum_{ij} \frac{\partial \theta_{ij}(E_u, \eta)}{\partial \eta'} \frac{\partial}{\partial \eta} P_{ij}^*(E_u; \eta). \end{aligned}$$

Using Lemma 3, we can now show that the first term in the above expression goes to zero in probability. Furthermore, with some algebra it can be shown that

$$\sum_{ij} \frac{\partial \theta_{ij}(E_u, \eta)}{\partial \eta} \frac{\partial}{\partial \eta} P_{ij}^*(E_u; \eta) = V^* \left\{ \frac{\partial \theta_{DG}(E_u, \eta)}{\partial \eta} \middle| E = E_u \right\}.$$

The proof of Lemma 4 now easily follows from the result of Lemma 3. \square

Proof of Proposition 1. (i) The main condition for consistency, that is the asymptotic unbiasedness of the score equation $\sum_{i=1}^N \Psi(D_i, G_i, E_i; \eta) = 0$, follows from direct application of Lemma 3. In Lemma 4, we have further shown that $-\partial/\partial \eta \{N^{-1} \sum_{i=1}^N \Psi(D_i, G_i, E_i; \eta)\} \rightarrow \mathcal{I}$ in probability, where \mathcal{I} is a positive definite matrix. Moreover, from (6) it is easy to see that the first and second derivatives of $\theta_{ij}(E; \eta)$ with respect to η can be uniformly bounded in an open neighbourhood of η_0 . This can be used to show that the convergence in Lemma 4 holds uniformly in an open neighbourhood of η_0 . The proof now follows using results of Foutz (1977).

(ii) The asymptotic normality of the estimator follows from standard application of the central limit theorem. To derive the form of the asymptotic variance, we need to prove that

$$\Gamma \equiv \text{cov } N^{-1/2} \sum_{u=1}^N \Psi(D_u, G_u, E_u; \eta) = \mathcal{I} - \Omega. \quad (\text{A4})$$

Let $\Phi(D; \eta) = E\{\Psi(D, G, E; \eta) | D\}$ and $\tilde{\Psi}(D, G, E; \eta) = \Psi(D, G, E; \eta) - \Phi(D; \eta)$. We can now write

$$\begin{aligned} \Gamma &= \sum_d \frac{N_d}{N} \text{cov} \{ \Psi(D, G, E; \eta) | D = d \} = \sum_d \frac{N_d}{N} E \{ \tilde{\Psi}^{\otimes 2}(D, G, E; \eta) | D = d \} \\ &= \sum_d \mu_d \sum_j \int \tilde{\Psi}^{\otimes 2}(D, G, E; \eta) \text{pr}(D = d | G = g_j, E = e) q_j dF(e). \end{aligned}$$

By reordering the sums and the integral in the last expression we can easily show that

$$\Gamma = \int E^* \{ \tilde{\Psi}^{\otimes 2}(D, G, E; \eta) | E = e \} h(e) dF(e).$$

Since

$$\tilde{\Psi}^{\otimes 2}(D, G, E; \eta) = \Psi^{\otimes 2}(D, G, E; \eta) + \Phi^{\otimes 2}(D; \eta) - 2\Psi(D, G, E; \eta)\Phi(D; \eta)^T$$

and $E^* \{ \Psi(D, G, E; \eta)^{\otimes 2} | E = e \} = V^* \{ \Psi(D, G, E; \eta) | E = e \}$, the proof of formula (A4) will follow if we can show that

$$\sum_d \mu_d \Phi(d; \eta)^{\otimes 2} = \int E^* \{ \Psi(D, G, E; \eta)\Phi(D; \eta)^T | E = e \} h(e) dF(e), \quad (\text{A5})$$

$$\sum_d \mu_d \Phi(d; \eta)^{\otimes 2} = \int E^* \{ \Phi(D; \eta)^{\otimes 2} | E = e \} h(e) dF(e). \quad (\text{A6})$$

To prove (A5), we first define

$$W(D, E; \eta) = E\{\Psi(D, G, E; \eta) | D, E\} = E^*\{\Psi(D, G, E; \eta) | D, E\}$$

and note that $E\{W(D, E; \eta) | D\} = \Phi(D; \eta)$. It is easily seen that the right-hand side of (A5) can be written as

$$\int E_D^* \{\Phi(D; \eta) W(D, e; \eta) | E = e\} h(e) dF(e). \quad (\text{A7})$$

Now we observe that $\text{pr}^*(D|E) = \text{pr}(D|E)\mu_D/h(E)$ and write (A7) as

$$\begin{aligned} \int \sum_D \mu_D \text{pr}(D|E = e) \Phi(D; \eta) W(D, e; \eta) dF(e) &= \sum_D \mu_D \text{pr}(D) \Phi(D; \eta) \int \frac{W(D, e; \eta) \text{pr}(D|E = e) dF(e)}{\text{pr}(D)} \\ &= \sum_D \mu_D \Phi^{\otimes 2}(D; \eta). \end{aligned}$$

This proves (A5). The proof of (A6) follows from similar steps. \square

REFERENCES

- ALBERT, P. S., RATNASINGHE, D., TANGREA, J. & WACHOLDER, S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *Am. J. Epidemiol.* **154**, 687–93.
- ANDERSEN, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Statist. Soc. B* **32**, 283–301.
- BRESLOW, N. E., ROBINS, J. M. & WELLNER, J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**, 447–55.
- CHATTERJEE, N., KALAYLIOGLU, Z. & CARROLL, R. J. (2005). Exploiting gene-environment independence in family-based case-control studies: Increased power for detecting associations, interactions and joint effects. *Genet. Epidemiol.* **28**, 138–56.
- CORNFIELD, J. (1956). A statistical problem arising from retrospective studies. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob.*, **4**, Ed. J. Neyman, pp. 135–48. Berkeley, CA: University of California Press.
- FOUTZ, R. V. (1977). On the unique consistent solution to the likelihood equations. *J. Am. Statist. Assoc.* **72**, 147–9.
- KHOURI, M. J., BEATY, T. H. & COHEN, B. H. (1993). *Fundamentals of Genetic Epidemiology*. Oxford University Press.
- MODAN, M. D., HARTGE, P. et al. (2001). Parity, oral contraceptives and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New Engl. J. Med.* **345**, 235–40.
- PIEGORSCH, W. W., WEINBERG, C. R. & TAYLOR, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies. *Statist. Med.* **13**, 153–62.
- PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.
- RABINOWITZ, D. (1997). A note on efficient estimation from case-control data. *Biometrika* **84**, 486–8.
- ROEDER, K., CARROLL, R. J. & LINDSAY, B. G. (1996). A nonparametric mixture approach to case-control studies with errors in covariables. *J. Am. Statist. Assoc.* **91**, 722–32.
- SCHMIDT, S. & SCHAID, D. J. (1999). Potential misinterpretation of the case-only study to assess gene-environment interaction. *Am. J. Epidemiol.* **150**, 878–85.
- SCOTT, A. J. & WILD, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57–71.
- UMBACH, D. M. & WEINBERG, C. M. (1997). Designing and analyzing case-control studies to exploit independence of genotype and exposure. *Statist. Med.* **16**, 1731–43.

[Received September 2003. Revised October 2004]