Using shared genetic controls in studies of gene-environment interactions

BY YI-HAU CHEN
Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan
yhchen@stat.sinica.edu.tw

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
Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, U.S.A.
carroll@stat.tamu.edu

SUMMARY

With the advent of modern genomic methods to adjust for population stratification, the use of external or publicly available controls has become an attractive option for reducing the cost of large-scale case-control genetic association studies. In this article, we study the estimation of joint effects of genetic and environmental exposures from a case-control study where data on genome-wide markers are available on the cases and a set of external controls while data on environmental exposures are available on the cases and a set of internal controls. We show that under such a design, one can exploit an assumption of gene-environment independence in the underlying population to estimate the gene-environment joint effects, after adjustment for population stratification. We develop a semiparametric profile likelihood method and related pseudolikelihood and working likelihood methods that are easy to implement in practice. We propose variance estimators for the methods based on asymptotic theory. Simulation is used to study the performance of the methods, and data from a multi-centre genome-wide association study of bladder cancer is further used to illustrate their application.

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

1. INTRODUCTION

In case-control epidemiological studies, it is generally desirable to sample the cases and controls from the same underlying study base (Wacholder et al., 1992), so that any difference in the exposure distribution between these groups is not confounded with differences in their population backgrounds. In genetic epidemiology, however, recent genome-wide association studies have tended to relax the study-base principle due to considerations of cost, convenience and efficiency of sampling. These studies have shown that bias arising from population stratification...
can often be corrected by employing methods, such as principal components analysis, that use a large panel of markers to adjust genomically for differences in population genetic background of the cases and controls. There is thus growing interest in reducing the cost of new case-control studies by using external controls, possibly drawn from different study bases, who may already have been genotyped.

One challenge to using external controls in genetic association studies is that it may limit one’s ability to study gene-environment interactions. Data on specific environmental exposures of interest for the disease under study may not be available for external controls, who may originally have been sampled to study an entirely different outcome. Even if data on relevant environmental exposures were available on external controls, such data may not be comparable to those for the cases, due to underlying population differences. Any resulting bias in estimation of the effect of environmental exposures cannot be corrected by genomic control methods, which are only suitable for adjusting for difference in genetic background between populations. To obtain estimates of interaction parameters, one can exploit an assumption of gene-environment independence in the underlying population and adopt a case-only approach (Piegorsch et al., 1994). The case-only analysis, however, is limited because it does not allow estimation of main effect parameters for genetic and/or environmental factors, and it does not use all the available data. Many procedures useful for analysing genome-wide association study data, such as tests for genetic association in the presence of interactions (Kraft et al., 2007), tests for additive interactions (Rothman & Greenland, 1998, Ch. 18), and tests for interaction in the presence of directional constraints (Song & Nicolae, 2009), require estimation of main effects along with interaction parameters.

In this article, we consider a natural setting for conducting genome-wide association studies that can take advantage of external controls and yet allow estimation of all the parameters of the gene-environment joint effects. We assume that the cases for a genome-wide association study are selected from an existing population-based case-control study for which data on environmental exposures have already been gathered for both the cases and the controls. In the absence of any genetic information, the cases and internal controls from such studies can be used in the usual manner to estimate the effects of the environmental exposures. We assume that, for the purpose of conducting a new genome-wide association scan, the cases from such a study can be genotyped and compared with existing genotype data from a set of external controls. We show that in such a study setting, it is possible to use both external and internal controls to estimate the joint effect of genetic and environmental factors on the risk of the disease by exploiting the assumption of gene-environment independence. We show how estimates of the genetic effects in such analyses can be appropriately adjusted for possible population stratification bias.

2. Methods

2.1. Notation and study design

Let $D$ be the case-control indicator, and let $G$ and $X$ be the genetic and exposure statuses of a subject, respectively. Let $S$ be a set of covariates that capture underlying population stratification information. In the analysis of genome-wide association studies, it is now common practice to adjust for population stratification based on the major principal components of large panels of single-nucleotide polymorphism, SNP, markers (Price et al., 2006). Thus, in our applications, $S$ could be a set of such principal components themselves or dummy variables associated with categorical strata derived from the principal components. We assume that
Shared genetic controls

in the underlying population from which the original case-control sample was drawn, the risk model is

\[ \Pr(D = 1 \mid G, X, S) = H \left\{ \beta_0 + r(S, \delta) + m(G, X, \beta) \right\}, \]  

(1)

where \( H(u) = \{1 + \exp(-u)\}^{-1} \) is the logistic distribution function, \( m(G, X, \beta) \) is a parametric function of \( G \) and \( X \), and \( r(S, \delta) \) is a parametric function of \( S \). The function \( m(G, X, \beta) \) specifies the form of the joint effects of \( G \) and \( X \) on the risk of \( D \). In standard logistic regression analysis, if \( G \) and \( X \) are scalar, then

\[ m(G, X, \beta) = \beta_G G + \beta_X X + \beta_{GX} GX, \]

where \( \beta_G \) corresponds to the main effect of \( G \), \( \beta_X \) corresponds to the main effect of \( X \), and \( \beta_{GX} \) corresponds to the interaction effect of \( G \) and \( X \) on the log odds ratio scale. The general form of \( m(G, X, \beta) \), however, allows the specification of joint effects in alternative forms, such as on the additive scale (Rothman & Greenland, 1998, Ch. 18), that may be nonlinear in the parameters. In model (1), inclusion of the term \( r(S, \delta) \), typically specified as \( r(S, \delta) = S^T \delta \), allows examination of the association of \( D \) with \( G \) and \( X \) after adjustment for population stratification \( S \).

We consider a study design where data on \( X \) have already been collected for \( n_{1P} \) cases and \( n_{0P} \) controls in the primary study, drawn from an underlying population denoted by \( \mathcal{P} \). We assume that data on \( G \) are available only for the \( n_{1P} \) cases and a set of \( n_{0E} \) controls, drawn from an external population denoted by \( \mathcal{P}^E \), that are unrelated to the subjects selected in the original primary case-control study. Under this design, information on \( S \) will also be available only on the cases and external controls. We assume that data on \( X \) for the external controls are either unavailable or discarded even if available, as they may not have a distribution comparable to that of the population from which the original case-control sample was drawn.

2.2. Assumptions

To estimate the parameters of the logistic model (1) from the design in § 2.1, we need the following assumptions.

**Assumption 1.** The disease under study is rare, so that the disease-free subjects, \( D = 0 \), approximately represent the underlying population.

**Assumption 2.** There is gene-environment independence conditional on population substructure \( S \); specifically, in the population \( \mathcal{P} \), \( \Pr(G \mid X, S, D = 0) = \Pr(G \mid S, D = 0) \).

**Assumption 3.** The distribution of the environmental variable \( X \) does not depend on \( S \), the genetic substructure in the underlying population; that is, \( \Pr(X \mid S, D = 0) = \Pr(X \mid D = 0) \).

**Assumption 4.** Conditional on the underlying genetic structure \( S \), the genotype distribution \( \Pr(G \mid S, D = 0) \) for the external population \( \mathcal{P}^E \) is the same as that for the population \( \mathcal{P} \).

Here we make some remarks on these assumptions. The rare-disease assumption, Assumption 1, should be reasonable since the case-control study is the most popular design for studying rare diseases. For rare diseases, assumptions that are natural for the underlying population, such as gene-environment independence, can be invoked for disease-free subjects.
Assumption 2, the independence of inherited genetic susceptibility and subsequent environmental exposure, is natural in many settings and has previously been used in many methods for improving efficiency in the analysis of case-control studies of gene-environment interaction (Piegorsch et al., 1994; Umbach & Weinberg, 1997; Chatterjee & Carroll, 2005). This assumption, when made conditional on $S$ as in Assumption 2, allows the distributions of both $G$ and $X$ to be influenced by population substructure (Chatterjee & Carroll, 2005; Chatterjee et al., 2005; Bhattacharjee et al., 2010). Assumption 3 implies that the distribution of $X$ does not depend on $S$, the genetic substructure in the underlying population. This assumption is essential for the estimation of main effect parameters associated with $X$, since one cannot adjust for $S$ in a comparison of the cases and the internal controls, due to the lack of data on $S$ for the latter group. The assumption is also implicit in standard analyses of case-control studies of environmental exposures, where one typically does not adjust for fine genetic substructure information as a potential confounder. Assumption 4 implies that after proper adjustment for $S$, the genotype distributions for the controls in the two populations are the same. For example, if $S$ is race, then the racial distributions might be different in the two underlying populations, but the distributions of genes are likely to be the same between the populations within each race. Assumption 4 is the minimal assumption needed to use the external controls for any association analysis method. Figure 1(a) shows a causal diagram of the relationships between $S, G, X$ and $D$ based on Assumptions 1–4.

If there is an auxiliary covariate, say $W$, that might be related to both $D$ and $X$, it can easily be incorporated by simply defining an augmented covariate set $\tilde{X} = (W, X)$ and assuming that $(W, X)$ are jointly independent of $(G, S)$; see Fig. 1(a). This formulation can be used to incorporate covariates such as age or any environmental variable other than the main exposure of interest that the disease model needs to account for. Some auxiliary covariates such as study centres, however, may be related to $S$; see Fig. 1(b). In § 6, we discuss an extension of the proposed method to account for such auxiliary covariates under a relaxed set of conditions, but assuming that such covariates are discrete. In either of the above situations, $W$ could be a factor by which the original case-control study has been matched. The proposed method can account for frequency-matched designs by simply allowing the disease model to incorporate saturated effects for the original matching strata.
3. Profile likelihood and related estimators

3-1. Basic likelihood calculations

We use a parametric model to describe \( \Pr(G \mid S, D = 0) \) and write this as \( \Pr(G \mid S, D = 0) = Q(G \mid S, \theta) \). For analysis of each SNP marker, where \( G \) typically takes values 0, 1 or 2 depending on the number of a specific allele that a subject carries on homologous chromosomes, we assume a trichotomous logistic model (Bhattacharjee et al., 2010)

\[
\Pr(G = g \mid S, D = 0, \theta) = Q(g \mid S, \theta) = \frac{\exp[\theta_{0g} + g \sum_k \theta_{1k} f_k(S)]}{1 + \sum_{g' = 1}^2 \exp[\theta_{0g'} + g' \sum_k \theta_{1k} f_k(S)]}
\]  

for \( g = 1, 2 \), where \( f_k(S) \) could, for example, denote principal components of markers for identifying population strata. Let \( \xi(x) = \Pr(X = x \mid D = 0) \) in the primary study population, and let \( \alpha(s) \) and \( \alpha^E(s) \) denote \( \Pr(S = s \mid D = 0) \) in the primary and external study populations, respectively; these distributions remain fully unspecified. For \( d = 0 \) or 1, let \( \pi_d = \Pr(D = d) \) in the underlying population \( \mathcal{P} \) and recall the identity

\[
\frac{\pi_1}{\pi_0} = \sum_g \int_x \int_S \exp[\beta_0 + r(s, \delta) + m(g, x, \beta)] Q(g \mid s, \theta) \xi(x) \alpha(s).
\]

The likelihood for the data is then

\[
L = \prod_{i=1}^{n_{1P}} \Pr(G_i, X_i, S_i \mid D_i = 1) \prod_{j=1}^{n_{0P}} \Pr(X_j \mid D_j = 0) \prod_{k=1}^{n_{0E}} \Pr(G_k, S_k \mid D_k = 0).
\]

By using the equality

\[
\Pr(G_i, X_i, S_i \mid D_i = 1) = \frac{\pi_0}{\pi_1} \exp[\beta_0 + r(S_i, \delta) + m(G_i, X_i, \beta)] \Pr(G_i, X_i, S_i \mid D_i = 0),
\]

which follows from (1) and Bayes’ rule, the likelihood can be rewritten as

\[
L = (\pi_1/\pi_0)^{-n_{1P}} \prod_{i=1}^{n_{1P}} \left[ \exp[\beta_0 + r(S_i, \delta) + m(G_i, X_i, \beta)] Q(G_i \mid S_i, \theta) \xi(X_i) \alpha(S_i) \right] \times \prod_{j=1}^{n_{0P}} \xi(X_j) \prod_{k=1}^{n_{0P}} \{Q(G_k \mid S_k, \theta) \alpha^E(S_k)\}.
\]

3-2. Profile likelihood

Here we derive the profile likelihood maximizing over the unknown distribution functions \( \xi(x) \) and \( \alpha(s) \). We consider the maximum likelihood estimator of \( \xi(\cdot) \) that allows masses within \( (x_1, \ldots, x_L) \), the support of the observed \( X \) in the primary case-control sample. We write \( \hat{\xi}_L = \xi(x_L) \), which is the probability mass associated with \( x_L \). Similarly, we consider the maximum likelihood estimator of \( \alpha(\cdot) \) that allows masses within \( (s_1, \ldots, s_M) \), the support of the observed \( S \) for the cases in the primary study. Let \( \alpha_m = \alpha(S_m) \). Recall the following identity parallel to (3):

\[
\frac{\Pr(D = 1 \mid S)}{\Pr(D = 0 \mid S)} = \sum_g \int_x \exp[\beta_0 + r(s, \delta) + m(g, x, \beta)] Q(g \mid s, \theta) \xi(x).
\]
Lemma 1. Let $\kappa = \beta_0 - \log(\pi_1/\pi_0)$, and let $h(S) = -\log\{p_r^*(D = 1 \mid S)/p_r^*(D = 0 \mid S)\}$, where $p_r^*(D = 1 \mid S)/p_r^*(D = 0 \mid S)$ is the right-hand side of (5) with $\beta_0$ replaced by $\kappa$. Set $\psi(\cdot) = h(\cdot) + r(\cdot, \delta)$. The profile likelihood is

$$L_{\text{prof}}(\kappa, \beta, \theta, \psi) \propto \prod_{i=1}^{n_{1P}} \left[ \exp\{\kappa + \psi(S) + m(G_i, X_i, \beta)\} Q(G_i \mid S_i, \theta) \right]$$

$$\times \prod_{i=1}^{n_{1P}} D(X_i, \kappa, \beta, \theta, \psi) \times \prod_{j=1}^{n_0} D(X_j, \kappa, \beta, \theta, \psi)$$

$$\times \prod_{k=1}^{n_{0E}} Q(G_k \mid S_k, \theta),$$

where

$$D(x_\ell, \kappa, \beta, \theta, \psi) = \left[ n_{0P} + \sum_g \sum_{m=1}^{n_{1P}} \exp\{\kappa + \psi(S_m) + m(g, x_\ell, \beta)\} Q(g \mid S_m, \theta) \right]^{-1}.$$

A sketch of the proof of Lemma 1 is given in the Appendix. A crucial observation is that in the presence of population stratification, the profile likelihood (6) depends on $\psi(S) = h(S) + r(S, \delta)$ and not separately on $h(S)$ and $r(S, \delta)$. Based on the above representation of the profile likelihood, we next propose a series of estimators.

3.3. Estimators

We consider a semiparametric maximum likelihood estimator, where we estimate the function $\psi(S)$ nonparametrically from the profile likelihood itself. If $S$ is modelled as a categorical variable, such as principal components strata, then the dimension of the parameters in the profile likelihood remains finite. If $S$ is continuous, then we allow $\psi(S)$ to have a mass point at each of the observed values of $S$ in the set of $n_{1P}$ cases.

Specifically, we obtain the semiparametric maximum likelihood estimator by treating each $\psi(S_m)$ value ($m = 1, \ldots, n_{1P}$) as a nuisance parameter and then maximizing the profile likelihood (6) over $\Omega = (\kappa, \beta, \theta)$ as well as $\{\psi(S_m) : m = 2, \ldots, n_{1P}\}$, with $\psi(S_1)$ absorbed into $\kappa$ to avoid redundancy. The denominator of $Q(g \mid S)$ displayed in (2) is a function of $S$ only. This and the fact that the specification of $\psi(S)$ is saturated imply that in the first three terms of (6), the denominator of $Q(g \mid S)$ can be absorbed into $\psi(S)$, so that the first term of the profile likelihood (6) can be expressed simply as $\prod_{i=1}^{n_{1P}} \exp[\eta(G_i, S_i, X_i, \Omega, \psi(S_i))]$ and the function $D$ appearing in the second and third terms of (6) can be defined as

$$D(X, \Omega, \psi) = \left( n_{0P} + \sum_g \sum_{m=1}^{n_{1P}} \exp[\eta(g, S_m, X, \Omega, \psi(S_m))] \right)^{-1},$$

where

$$\eta(G, S, X, \Omega, \psi(S)) = \kappa + \psi(S) + m(G, X, \beta) + \theta_0 G + G\theta_1^T f(S).$$
We now explain why treating each $\psi(S_i)$ value as a separate parameter leads to the semiparametric maximum likelihood estimate. Rewrite the profile likelihood (6) as

$$
\mathcal{L}_{\text{prof}}(\kappa, \beta, \theta, \psi) = \prod_{p=1}^N \text{pr}(D_p, G_p, S_p | X_p, \Omega, \psi)^{O_p} \times Q(G_p | S_p, \theta)^{1-O_p},
$$

(7)

where

$$
\text{pr}(d, g, s | X, \Omega, \psi) = \frac{\exp [d \eta(g, s, X, \Omega, \psi(s))]}{\sum_{u=1}^N \sum_{g'} O_u \exp [D_u \eta(g', S_u, X, \Omega, \psi(S_u))]} 
$$

and $\psi(s) = h(s) + r(s, \delta)$. In addition, in (7) the subscript $p$ indexes all $N = n_{1P} + n_{0P} + n_{0E}$ subjects in the primary and external samples, while $O_p = 1$ if subject $p$ belongs to the primary sample and $O_p = 0$ otherwise. By direct calculation, the score equation for $\psi_m = \psi(s_m)$ derived from (7), or equivalently from (6), is

$$
\sum_{p=1}^N O_p \left[ D_p I(S_p = s_m) - \frac{\sum_{u=1}^N \sum_{g} D_u I(S_u = s_m) O_u \exp\{D_u \eta(g, S_u, X_p, \psi_u, \Omega)\}}{\sum_{u=1}^N \sum_{g} O_u \exp\{D_u \eta(g, S_u, X_p, \psi_u, \Omega)\}} \right] = 0,
$$

which reduces to

$$
1 = \sum_{i=1}^{n_{1P}+n_{0P}} \sum_{g} \frac{\exp\{\eta(g, s_m, X_i, \psi_m, \Omega)\}}{n_{0P} + \sum_{g} \sum_{h=1}^{n_{1P}} \exp\{\eta(g, S_h, X_i, \psi_h, \Omega)\}}, \quad s_m \in \{S_i : i = 1, \ldots, n_{1P}\}. 
$$

(8)

On the other hand, the derivations in Appendix A1 demonstrate that for each $s_m \in \{S_i : i = 1, \ldots, n_{1P}\}$, the semiparametric maximum likelihood must satisfy the constraint posed by (A4), which, upon substituting the expression (A5) for $\xi(x_i)$, coincides exactly with the score equation (8) for $\psi(s_m)$. Hence, maximizing the profile likelihood (6) or (7) with each $\psi(s_m), s_m \in \{S_i : i = 1, \ldots, n_{1P}\}$, treated as a parameter leads to the semiparametric maximum likelihood.

The semiparametric maximum likelihood method described above is computationally demanding and may not be feasible for large-scale analyses when $S$ is not categorical. In what follows, we use the form of the profile likelihood described in Lemma 1 to propose two alternative, computationally simpler, estimators. We consider a working likelihood method, where we model the function $\psi(S)$ directly using a parametric model such as $\psi(S) = S^T \delta$, and estimate $\delta$ together with $(\kappa, \beta, \theta)$ by maximizing the profile likelihood (6). The method may result in biased estimators, because the form of $\psi(S) = h(S) + r(S, \delta)$ that is induced by our modelling framework may, in general, not coincide with the proposed parametric form. However, if we directly specify a model for $\psi(S)$, ignoring such constraints, then the resulting estimator, though possibly subject to bias, enjoys computational advantages which are important in genome-wide association studies. In §4, we show by simulation that the bias due to such approximations can be minimal, unless there exist strong population stratification effects.

Let $\Omega = (\kappa, \beta, \theta)$ and $\Omega^* = (\Omega, \psi)$ with $\psi = \{\psi(s_m) : m = 2, \ldots, n_{1P}\}$, and let $\hat{\Omega}^* = (\hat{\Omega}, \hat{\psi})$ be the semiparametric maximum likelihood estimator maximizing (6). In Appendix A2, we provide explicit formulae for the score function, the estimated information, and the probability limit of the estimated information. Call the latter $I^*$, and let its submatrices be $I^*_{s\Omega}, I^*_{\Omega\psi}$ and $I^*_{\psi\psi}$. In the Appendix we show the following result.
Proposition 1. As $N \to \infty$, $N^{1/2}(\hat{\Omega} - \Omega)$ converges in distribution to a normally distributed random variable with mean zero and covariance matrix $I_{\hat{\Omega}\Omega}^{-1}$, where $I_{\hat{\Omega}\Omega} = I_{\hat{\Omega}\Omega}^* - I_{\hat{\Omega}\psi}I_{\psi\psi}^{-1}I_{\psi\Omega\hat{\Omega}}$.

A sketch of the proof is given in Appendix A2. Evaluating $I_{\hat{\Omega}\Omega}$ at the maximum likelihood estimate $\hat{\Omega}^*$ provides a covariance matrix estimate for $\hat{\Omega}$.

The score function and information matrix for the working likelihood estimator can be obtained by the same formulae as those for the semiparametric maximum likelihood, with the parameter set $\Omega^*$ redefined as $\Omega^* = (\kappa, \beta, \theta, \delta)$, where $\delta$ contains parameters in the parametric model assumed for $\psi(S)$. The asymptotic distribution in Proposition 1 applies approximately to the working likelihood estimator.

3.4. A pseudolikelihood method

To avoid maximization with respect to high-dimensional nuisance parameters, we consider an alternative pseudolikelihood method that substitutes a one-step estimator for $\psi(S)$ in the profile likelihood (6). Based on (5) and the definition of $h(S)$ given in Lemma 1,

$$
\psi(S) = h(S) + r(S, \delta) = -\log[pr^*(D = 1 \mid S)/pr^*(D = 0 \mid S)] + r(S, \delta)
$$

with $\zeta(g) = \int_x \exp(m(g, x, \beta))d\xi(x)$. To obtain a proper maximum likelihood estimator, one could iteratively update $\psi(S)$ through $\zeta(g)$ based on the estimate $\hat{\xi}_\ell$ of $\xi(x_\ell)$; but, to avoid iterations, we adopt an alternative strategy and simply estimate $\zeta(g)$ using a one-step unbiased estimator based on the empirical distribution of $X$ in the primary control sample, given by $\hat{\xi}_\ell(g) = n^{-1}_{\text{op}} \sum_{i=1}^{n_{\text{op}}} \exp(m(g, X_i, \beta))$. When implementing this method, the denominator of $Q(g \mid S, \theta)$ is absorbed into $\psi(S)$ as in § 3.3, so that for given $\Omega = (\kappa, \beta, \theta)$, $\psi(S)$ is estimated as

$$
\tilde{\psi}(S, \Omega) = -\log \left[ \sum_g \exp[\kappa + \theta_0g + g\theta_1'S(S)] \xi(g) \right].
$$

Inserting $\tilde{\psi}(S_i, \Omega)$ into the profile likelihood (6), we obtain the pseudolikelihood $L_{\text{pse}}\{\Omega, \tilde{\psi}(\Omega)\}$ where $\tilde{\psi}(\Omega) = \tilde{\psi}(S_i, \Omega) : i = 1, \ldots, n_{\text{op}}$. The pseudolikelihood estimator $\tilde{\Omega}$ of $\Omega$ is then obtained by maximizing $L_{\text{pse}}$ with respect to $\Omega$.

We can write $(\partial/\partial \Omega) \log(L_{\text{pse}}) = \sum_{i=1}^{n_{\text{op}}} \tilde{U}_{\Omega,i} + o_p(N^{1/2})$, with the $(\tilde{U}_{\Omega,i})$ being independent zero-mean random variables, and $-(\partial^2/\partial \Omega \partial^{\top} \Omega) \log(L_{\text{pse}}) = \tilde{I}_{\Omega\Omega}$; explicit expressions for $\tilde{U}_{\Omega,i}$ and $\tilde{I}_{\Omega\Omega}$ are provided in Appendix A3. By conventional estimating-equation theory, we have the following proposition about the asymptotic distribution of the pseudolikelihood estimator $\tilde{\Omega}$.

Proposition 2. Suppose that $N^{-1}\tilde{I}_{\Omega\Omega} = \tilde{I}_{\Omega\Omega} + o_p(1)$ and $N^{-1}\sum_{i=1}^{n_{\text{op}}} \tilde{U}_{\Omega,i}\tilde{U}_{\Omega,i}^\top \equiv \tilde{V} + o_p(1)$, with both limiting matrices being finite and positive definite. Then, as $N \to \infty$, $N^{1/2}(\tilde{\Omega} - \Omega)$ converges in distribution to a normally distributed random variable with mean zero and covariance matrix $\tilde{I}_{\Omega\Omega}^{-1}\tilde{V}I_{\Omega\Omega}^{-1}$.

The variance estimator for $\tilde{\Omega}$ can be obtained by evaluating the above asymptotic covariance matrix at $\Omega$ and $\tilde{\psi}(\Omega)$. 
3-5. A conditional likelihood

In this subsection, we describe a simple conditional likelihood approach that allows estimation of genetic main effects and gene-environment interaction parameters using the genotype information from shared controls, while discarding exposure information from internal controls. The method is analogous to the pooled control analysis method that was first considered by Modan et al. (2001) and then more formally developed by Tchetgen Tchetgen (2011), but it is more general because it allows for proper adjustment for population stratification.

Suppose that we have a standard logistic regression of the form \( m(G, X, \beta) = \beta_G G + \beta_X X + \beta_{GX} X G \). Further, assume that we have specified \( \Pr(G \mid S, D = 0, \theta) \) using the trichotomous logistic model (2). Then we can estimate \( \beta_G \) and \( \beta_{GX} \) using the conditional likelihood

\[
L^C = \prod_{i=1}^{n_{1P}} \Pr(G_i \mid X_i, S_i, D_i = 1) \prod_{j=1}^{n_{0E}} \Pr(G_j \mid S_j, D_j = 0),
\]

where \( \Pr(G_i \mid X_i, S_i, D_i = 1) \propto \exp(\beta_G G_i + \beta_{GX} G_i X_i) \exp(\theta_0 G_i + G_i \sum_k \theta_{1k} f_k(S)) \), under the assumption that \( \Pr(G \mid X, S, D = 0) = \Pr(G \mid S, D = 0) \), i.e., that \( G \) and \( X \) are independent conditional on \( S \). More precisely,

\[
\Pr(G_i \mid X_i, S_i, D_i = 1) = \frac{\exp(\theta_0 G_i + \beta_G G_i + G_i \sum_k \theta_{1k} f_k(S_i) + \beta_{GX} G_i X_i)}{1 + \sum_{g'} \exp(\theta_0 g' + \beta_G g' + g' \sum_k \theta_{1k} f_k(S_i) + \beta_{GX} g' X_i)}. \tag{10}
\]

An advantage of this likelihood is that it is simple and requires only the assumption that \( G \) and \( X \) are independent conditional on \( S \); \( X \) need not be independent of \( S \). The drawback is that, unlike methods based on the profile likelihood, the conditional method cannot estimate \( \beta_X \), nor can it utilize exposure data collected for internal controls.

Redefine \( \Omega = (\beta, \theta) \). Let \( \hat{\Omega} \) be the conditional likelihood estimator maximizing (9), and let \( \hat{I}_{\hat{\Omega}} = -(\partial^2 / \partial \Omega \partial \Omega^T) \log L^C(\Omega) \). An explicit expression for \( \hat{I}_{\hat{\Omega}} \) is given in the Appendix; see equation (A13). By applying standard likelihood theory, we obtain the following proposition.

**Proposition 3.** Assume that \( N^{-1} \hat{I}_{\hat{\Omega}} = \tilde{I}_{\hat{\Omega}} + o_p(1) \), which is finite and positive definite. Then, as \( N \to \infty \), \( N^{1/2}(\hat{\Omega} - \Omega) \) converges in distribution to a normally distributed random variable with mean zero and covariance matrix \( \hat{I}_{\hat{\Omega}}^{-1} \).

4. Simulations

We conducted simulations to examine how the proposed alternative methods and their associated variance estimators perform in finite samples. We assume that the population structure information in the study can be captured using two normally distributed independent continuous random variables, \( S_1 \) and \( S_2 \). To allow for differences in structure of the underlying populations for the original case-control study and the external study, we assume that each \( S_i \) is distributed as a standard normal variate, \( N(0, 1) \), in \( P \) and as a 50:50 mixture of \( N(0, 1) \) and \( N(-0.5, 1) \) in \( P^E \). There exists an environmental factor \( X \) in \( P \), which is a standard normal variate independent of \( S \). For both \( P \) and \( P^E \), we assume that the genotype \( G \) for a putative susceptibility SNP follows the trichotomous logistic model (2) with \( \theta_{0g} = 0.5, 0 \) for \( g = 1, 2 \), respectively, \( f_k(S) = (S_1, S_2) \) and \( \theta_1 = (0.6, 0.4) \). The setting corresponds to a marginal allele frequency of the SNP of 0.497 and marginal genotype frequencies \( \Pr(G = 2) = 0.296, \Pr(G = 1) = 0.401 \) and \( \Pr(G = 0) = 0.303 \) in \( P \). In \( P^E \), the marginal allele frequency of the SNP is 0.443 and the marginal genotype frequencies are \( \Pr(G = 2) = 0.246, \Pr(G = 1) = 0.394 \) and \( \Pr(G = 0) = 0.360 \).
Table 1. Finite-sample performances (×10^{-2}) over 1000 simulations under the model with gene-environment interaction

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>( \beta_G = 0.2 )</th>
<th>( \beta_X = 0.3 )</th>
<th>( \beta_{GX} = 0.7 )</th>
<th>( \theta_{11} = 0.6 )</th>
<th>( \theta_{12} = 0.4 )</th>
<th>( \theta_{01} = 0.5 )</th>
<th>( \theta_{02} = 0.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPMLE</td>
<td>Mean</td>
<td>21</td>
<td>33</td>
<td>67</td>
<td>60</td>
<td>40</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>95</td>
<td>94</td>
<td>91</td>
<td>94</td>
<td>96</td>
<td>96</td>
<td>96</td>
</tr>
<tr>
<td>PL</td>
<td>Mean</td>
<td>22</td>
<td>33</td>
<td>67</td>
<td>60</td>
<td>40</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>95</td>
<td>94</td>
<td>92</td>
<td>94</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>WL</td>
<td>Mean</td>
<td>26</td>
<td>33</td>
<td>67</td>
<td>55</td>
<td>36</td>
<td>47</td>
<td>−3</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>12</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>12</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>93</td>
<td>94</td>
<td>91</td>
<td>82</td>
<td>89</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>CL</td>
<td>Mean</td>
<td>20</td>
<td>−</td>
<td>68</td>
<td>60</td>
<td>40</td>
<td>51</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>14</td>
<td>−</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>14</td>
<td>−</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>CP</td>
<td>95</td>
<td>94</td>
<td>93</td>
<td>94</td>
<td>96</td>
<td>96</td>
<td>95</td>
</tr>
</tbody>
</table>

SPMLE, semiparametric maximum likelihood estimate; PL, pseudolikelihood estimate; WL, working likelihood estimate; CL, conditional likelihood estimate; SE, standard error; \( \tilde{SE} \), mean of the standard error estimate; CP, coverage probability of a nominal 95% confidence interval.

We assume that the risk of the disease of interest given \( G, S \) and \( X \) in \( \mathcal{P} \) follows the logistic model (1) with

\[
m(G, X, \beta) = \beta_G G + \beta_X X + \beta_{GX} GX \quad \text{and} \quad r(S, \delta) = \delta_1 S_1 + \delta_2 S_2.
\]

We set the parameter values to be \( \beta_0 = -7.0, \beta_G = 0.2, \beta_X = 0.3, \delta_1 = \delta_2 = 0.2 \) and \( \beta_{GX} = 0.7 \) or 0.0, corresponding to scenarios with or without gene-environment interaction. Following this model, we simulate data for \( (G, X, S) \) for 500 cases and 500 controls from population \( \mathcal{P} \). In addition, we simulate data for \( (G, S) \) for 500 controls from population \( \mathcal{P}^E \). For the analysis of each simulated dataset, we assume that data on \( G \) and \( S \) are not available for controls drawn from \( \mathcal{P} \).

The results from 1000 replications are presented in Table 1. The semiparametric likelihood, the pseudolikelihood and the conditional likelihood produce almost unbiased estimates for all of the parameters in the disease-risk and genotype frequency models. The average standard error estimates for these estimators are also close to the simulation standard deviations. Moreover, the variance of the standard error estimates is on the order of \( 10^{-6} \) to \( 10^{-5} \), and the coefficient of variation of the standard error estimates is within the range of 2% to 7%, indicating that our standard errors are stable. The good agreement between the confidence interval coverage probabilities and their nominal value, 95%, indicates that the proposed estimators follow reasonably well the asymptotic normal distributions given in the propositions. The finite-sample performance of the semiparametric maximum likelihood method is particularly noteworthy, given that it involves nuisance parameters associated with the function \( \psi(S) \), which has dimensionality the same as the sample size. Although the working likelihood method incorrectly assumes the linear model \( \psi(S) = S^T \delta \), it still produces nearly unbiased estimates and good coverage probabilities for all the parameters in the disease-risk model; however, it produces noticeable bias for parameters in the genotype frequency model. In the Supplementary Material we show that under an extreme degree of population stratification, the working likelihood method may produce large bias for parameters of the disease-risk model, especially for genetic main effects.
Comparisons of the alternative estimators reveal that the semiparametric maximum likelihood, pseudolikelihood and working likelihood estimators have nearly the same efficiency for estimation of parameters in the disease-risk model. In contrast to these three methods, the conditional likelihood method can suffer loss of efficiency for estimation of the interaction parameter. This is to be expected, since the conditional likelihood approach cannot exploit Assumption 3. Intriguingly, we have observed that in the absence of an interaction term in the standard logistic model, the semiparametric maximum likelihood, pseudolikelihood and conditional likelihood methods all produce numerically identical estimates of the parameters; see Table S1 in the Supplementary Material. In Appendix A5, we further investigate this phenomenon theoretically.

In additional simulation studies, the results of which are summarized in Tables S4 and S5 of the Supplementary Material, we investigate the impact of the misspecification of the two necessary and important assumptions, Assumptions 2 and 3. If the condition of gene-environment independence is violated, all methods can produce major bias in estimation of the gene-environment interaction parameter. These results are to be expected, given the known bias of the case-only and other related methods under violation of the underlying gene-environment independence assumption. When the distribution of \( X \) is related to \( S \), i.e., when Assumption 3 is violated, we observe that the proposed methods can yield bias in estimation of the main effect of \( X \). Such bias is also expected, because in this case \( S \) is a confounder for the effect of \( X \), but one cannot adjust \( S \) in the usual fashion since it is not observed for the internal controls. Further, violation of this assumption leads, somewhat unexpectedly, to bias also in estimation of the genetic main effect parameters for all of the methods except the working and conditional likelihood methods. Encouragingly, however, we observe that violation of Assumption 3 does not lead to major bias in estimates of the gene-environment interaction parameter for any of the methods.

5. Data example

We illustrate the application of our methods using data from a recent multi-centre genome-wide association study of bladder cancer. The original study (Rothman et al., 2011), which reported the discovery of several novel susceptibility SNPs for bladder cancer, was based on case-control samples drawn from five different sites in North America and Europe. In the current analysis, we investigate the interaction between rs11892031, one of the novel SNPs reported, and smoking status, which is a known risk factor for bladder cancer. We analyse the data from the New England site, pretending that only the cases in this study are genotyped and that the genotype data from the controls at another North American site, namely the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial cohort (Gohagan et al., 2000), are available for sharing. Based on the principal components analysis of approximately 12,000 ancestry-informative SNPs in the New England cases and external controls, we picked three eigenvectors for adjustment of possible population stratification between the two study sites.

Our analysis consists of 629 cases and 761 controls from New England and 1,879 external controls from the cancer screening trial. The SNP genotype data are coded as 0, 1 or 2, reflecting the count for the minor allele. Smoking status is coded as a binary variable that indicates whether a subject ever smoked or not. In addition, all models include the covariates age, coded by dummy variables associated with the age categories 55–65, 65–75 and >75, and gender, coded as a binary variable which indicates whether the subject is male or not. For all shared control analyses, we utilize the genotype data from the external controls and the covariate data from the internal New England controls. For comparison, we also analyse all of the data available from the New England cases and internal controls using standard logistic regression.
The methods proposed in §3 for the shared control analysis all produced similar point estimates for disease odds ratio parameters of the logistic regression model; see Table 2. None of the methods detected a statistically significant interaction effect between the SNP and smoking. When the interaction term was dropped from the model, in data not shown here, all the methods found the main effect of the SNP to be statistically significant.

Compared to the internal analysis of the New England study data using standard logistic regression, the shared control analysis produced attenuated estimates for the main effect of the SNP and the SNP × smoking interaction term. Although such variations are well within the limits of uncertainty, the possibility of bias due to violation of one or more of the underlying assumptions cannot be ruled out. Any bias in the genetic main effect is most likely due to violation of the underlying condition of gene-environment independence, Assumption 4. Any bias in the gene-environment interaction parameter will most likely be due to violation of the underlying condition of gene-environment independence, Assumption 2. Both of these biases, however, seem unlikely. First, both of the populations under study are American Caucasian populations, for which principal components methods are generally known to be adequate for adjusting for population stratification. Second, if the SNP under study is truly associated with smoking habits in the underlying population, such a relationship would have been detected by now in the extremely large genome-wide association study that has been conducted for various smoking phenotypes (Tobacco and Genetics Consortium, 2010).

### 6. Discussion

Sometimes there may exist auxiliary variables, such as geographic centres for the original case-control study, that could be related to $D$, $X$ and $S$. To account for such covariates $W$, our methods can be modified to allow for stratified analysis. Suppose that in the primary sample we have $K$ strata defined by categorized covariates, and suppose that in stratum $w$ ($w = 1, \ldots, K$) we have data $\{G_{wi}, X_{wi}, S_{wi}: i = 1, \ldots, n_{WP}^w\}$ on genotype, environmental exposure and population structure information for case subjects with $D_{wi} = 1$, as well as environmental exposure data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Standard</th>
<th>SPMLE</th>
<th>PL</th>
<th>WL</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_{SNP}$</td>
<td>$-0.58 (0.33)$</td>
<td>$-0.40 (0.30)$</td>
<td>$-0.41 (0.32)$</td>
<td>$-0.41 (0.30)$</td>
<td>$-0.40 (0.30)$</td>
</tr>
<tr>
<td>$\beta_{smoking}$</td>
<td>$0.92 (0.15)^*$</td>
<td>$0.94 (0.14)^*$</td>
<td>$0.94 (0.14)^*$</td>
<td>$0.94 (0.14)^*$</td>
<td>$-0.04 (0.14)^*$</td>
</tr>
<tr>
<td>$\beta_{gender}$</td>
<td>$0.02 (0.13)$</td>
<td>$0.03 (0.13)$</td>
<td>$0.03 (0.13)$</td>
<td>$0.03 (0.13)$</td>
<td>$-0.04 (0.14)^*$</td>
</tr>
<tr>
<td>$\beta_{age1}$</td>
<td>$0.11 (0.17)$</td>
<td>$0.16 (0.17)$</td>
<td>$0.12 (0.17)$</td>
<td>$0.12 (0.17)$</td>
<td>$0.17 (0.17)^*$</td>
</tr>
<tr>
<td>$\beta_{age2}$</td>
<td>$-0.03 (0.16)$</td>
<td>$-0.03 (0.16)$</td>
<td>$-0.03 (0.16)$</td>
<td>$-0.03 (0.16)$</td>
<td>$-0.04 (0.14)^*$</td>
</tr>
<tr>
<td>$\beta_{age3}$</td>
<td>$0.19 (0.18)$</td>
<td>$0.19 (0.18)$</td>
<td>$0.19 (0.18)$</td>
<td>$0.19 (0.18)$</td>
<td>$-0.04 (0.14)^*$</td>
</tr>
<tr>
<td>$\beta_{SNP \times smoking}$</td>
<td>$0.29 (0.37)$</td>
<td>$0.18 (0.32)$</td>
<td>$0.18 (0.33)$</td>
<td>$0.12 (0.32)$</td>
<td>$0.17 (0.32)$</td>
</tr>
<tr>
<td>$\theta_{EV1}$</td>
<td>$-5.99 (3.68)$</td>
<td>$5.99 (3.68)$</td>
<td>$5.94 (3.68)$</td>
<td>$5.97 (3.69)$</td>
<td>$-0.68 (2.42)^*$</td>
</tr>
<tr>
<td>$\theta_{EV2}$</td>
<td>$-13.04 (3.44)^*$</td>
<td>$-13.04 (3.11)^*$</td>
<td>$-12.98 (3.43)^*$</td>
<td>$-13.03 (3.44)^*$</td>
<td>$-0.68 (2.42)^*$</td>
</tr>
<tr>
<td>$\theta_{EV3}$</td>
<td>$-6.16 (4.48)$</td>
<td>$-6.16 (4.54)$</td>
<td>$-5.97 (4.81)$</td>
<td>$-6.18 (4.48)$</td>
<td>$-0.68 (2.42)^*$</td>
</tr>
<tr>
<td>$\theta_{01}$</td>
<td>$-1.72 (0.06)^*$</td>
<td>$-1.72 (0.06)^*$</td>
<td>$-1.72 (0.06)^*$</td>
<td>$-1.72 (0.06)^*$</td>
<td>$-0.68 (2.42)^*$</td>
</tr>
<tr>
<td>$\theta_{02}$</td>
<td>$-4.68 (0.24)^*$</td>
<td>$-4.68 (0.24)^*$</td>
<td>$-4.68 (0.24)^*$</td>
<td>$-4.68 (0.24)^*$</td>
<td>$-0.68 (2.42)^*$</td>
</tr>
</tbody>
</table>

Standard, standard logistic analysis of internal cases and controls; SPMLE, semiparametric maximum likelihood estimate; PL, pseudolikelihood estimate; WL, working likelihood estimate; CL, conditional likelihood estimate; age1–age3, three dummy variables for age group; EV1–EV3, three population stratification eigenvectors; *, statistical significance at the 5% level.
\{X_{wj} : j = 1, \ldots, n_{0\text{P}}^w\} for control subjects with \(D_{wj} = 0\). Let the disease model be

\[
\text{pr}(D = 1 | G, X, S, \text{stratum } w) = H \{ \beta_{0w} + r(S, \delta) + m(G, X, \beta) \} \quad (w = 1, \ldots, K);
\]

that is, we introduce saturated effects of the strata. Further, we modify Assumption 2 to become \(\text{pr}(G | X, S, \text{stratum } w, D = 0) = \text{pr}(G | S, D = 0)\) and change Assumption 3 to \(\text{pr}(X | S, \text{stratum } w, D = 0) = \text{pr}(X | \text{stratum } w, D = 0)\). We allow the distributions of \(W\) and \(S\) to be related. Figure 1(b) displays a causal diagram showing the assumed relationships between \(W, S, G, X\) and \(D\). As before, we model \(\text{pr}(G | S, D = 0)\) by (2). Then, following derivations parallel to those in the proof of Lemma 1, we obtain the profile likelihood

\[
\mathcal{L}_\text{prof} \propto \prod_{w=1}^{K} \left( \prod_{i=1}^{n_{\text{IP}}^w} \exp \left[ \eta^w \{ G_{wi}, S_{wi}, X_{wi}, \Omega^w, \psi^w(S_{wi}) \} \right] \right) \times \prod_{i=1}^{n_{\text{IP}}^w} D^w(X_i, \Omega^w, \psi^w) \times \prod_{j=1}^{n_{0\text{P}}^w} D^w(X_j, \Omega^w, \psi^w) \times \prod_{k=1}^{n_{0E}} Q(G_k | S_k, \theta).
\]

Here \(\Omega^w = (\kappa^w, \beta, \theta)\), where \(\kappa^w = \beta_{0w} - \log(\pi_i^w / \pi_0^w)\) with \(\pi_i^w = \text{pr}(D = i | \text{stratum } w)\) for \(i = 0, 1\), \(\eta^w \{ G, S, X, \Omega^w, \psi^w(S) \} = \kappa^w + \psi^w(S) + m(G, X, \beta) + \theta_{0G} + G\theta_1^f(S)\), \(\psi^w(S) = r(S, \delta) + h^w(S)\), where \(h^w(S)\) is defined as \(h(S)\) with \(\kappa\) replaced by \(\kappa^w\), and

\[
D^w(X, \Omega^w, \psi^w) = \left( n_{0\text{P}}^w + \sum_{g} \sum_{i=1}^{n_{\text{IP}}^w} \exp[\eta^w \{ g, S_{wi}, X, \Omega^w, \psi^w(S_{wi}) \}] \right)^{-1}.
\]

The semiparametric likelihood in this setting can thus be obtained in the same fashion as in the unstratified setting, if we treat each value of \(\psi^w(S)\) in each stratum as a separate nuisance parameter. The pseudolikelihood and working likelihood methods can be derived analogously.

**Umbach & Weinberg (1997)** described alternative designs and analyses for case-control studies under the gene-environment independence assumption. They showed that under a log-linear modelling framework for categorical covariates, the independence assumption can be exploited to make inference on all the parameters of a logistic model when genotype and exposure data for controls are unlinked or collected from different sets of people. A major advantage of our proposed framework is its ability to adjust for population stratification using standard genomic control methods, methods which are generally considered to be key to the validity of the analysis when external controls are used in genetic association studies. In addition, our framework is very general, allowing for the full flexibility of a general logistic regression analysis, such as incorporation of continuous covariates or modelling interaction on alternative scales.

Gene-environment studies using shared controls inevitably require reliance on certain assumptions. In this article, we explicitly identify the assumptions and define several methods that can utilize all available data to make the best possible inference. Although the assumptions are reasonable, they can be violated in some situations and are untestable from the dataset itself. Thus it is important that initial discoveries from such studies be followed up in subsequent replication studies, such as standard case-control studies, which do not rely on the same set of assumptions. It is
possible that once a genome-wide scan using the shared control design identifies some promising SNPs, a set of selected SNPs could be genotyped for all or some of the internal controls of the parent case-control study, assuming DNA samples are available. In the future, it will be interesting to explore methods that can efficiently combine data from both internal and external controls in such studies. Because genotype data on internal controls allow assessment of gene-environment association in the underlying population, it should be possible to develop methods that are robust to violation of the underlying gene-environment independence assumption.

Acknowledgement

We are grateful for the very helpful comments from the editor, associate editor and two referees, which have resulted in substantial improvements to this work. Chen’s research was supported by the National Science Council of Taiwan. Chatterjee’s research was supported by the Intramural Research Program of the National Cancer Institute. Carroll’s research was supported by the National Cancer Institute and the King Abdullah University of Science and Technology.

Supplementary Material

Supplementary material available at *Biometrika* online includes the results of additional simulations, such as those conducted using the model without gene-environment interaction, under conditions with stronger population stratification effects, and in cases where various assumptions are violated.

Appendix

A1. Proof of Lemma 1

First, consider maximizing the likelihood (4) with respect to the $\xi_\ell$. Let $n(x_\ell)$ be the number of subjects with $X_\ell = x_\ell$ in the primary case-control sample. The loglikelihood relevant for $\xi_\ell$ is

$$ n(x_\ell) \log \xi_\ell - n_{1P} \log \left[ \sum_{g, m} \exp\{\beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta)\} Q(g | s_m) \xi \alpha(s_m) \right] + \lambda \left( \sum_{p=1}^{L} \xi_p - 1 \right), $$

where the last term is the Lagrange multiplier argument. Differentiating with respect to $\xi_\ell$, we see that for $\ell = 1, \ldots, L$ we are solving

$$ 0 = \frac{n(x_\ell)}{\xi_\ell} - \frac{n_{1P} \sum_{g,m} \exp\{\beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta)\} Q(g | s_m) \alpha(s_m)}{\sum_{g,p,m} \exp\{\beta_0 + r(s_m, \delta) + m(g, x_p, \beta)\} Q(g | s_m) \xi_p \alpha(s_m)} + \lambda. $$

Multiply the right-hand side of (A1) by $\xi_\ell$ and then sum over $\ell$ to get $n_{1P} + n_{0P} - n_{1P} + \lambda \sum_\ell \xi_\ell$. This means that $\lambda = -n_{0P}$. We now solve (A1) to find that

$$ \xi_\ell = \left[ n_{0P} + n_{1P} \sum_{g,m} \exp\{\beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta)\} Q(g | s_m) \alpha(s_m) \right]^{-1} \sum_{g,p,m} \exp\{\beta_0 + r(s_m, \delta) + m(g, x_p, \beta)\} Q(g | s_m) \xi_p \alpha(s_m). $$

Now define $\kappa = \beta_0 - \log(\pi_1/\pi_0)$ and invoke (3) to get

$$ \xi_\ell = \left[ n_{0P} + n_{1P} \sum_{g,m} \exp\{\kappa + r(s_m, \delta) + m(g, x_\ell, \beta)\} Q(g | s_m, \theta) \alpha(s_m) \right]^{-1}. $$
Next, consider maximizing the likelihood (4) with respect to the \( \alpha(s_m) \). The loglikelihood relevant for \( \alpha_m = \alpha(s_m) \) is

\[
n(s_m) \log \alpha_m - n_{1P} \log \left[ \sum_{g, \ell, m} \exp \left\{ \beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \right] - \lambda \left( \sum_{j=1}^{n_{1P}} \alpha_j - 1 \right),
\]

where \( n(s_m) \) is the number of subjects with \( S_i = s_m \) for \( i = 1, \ldots, n_{1P} \). Upon differentiating, we solve

\[
0 = \frac{n(s_m)}{\alpha_m} - n_{1P} \frac{\sum_{g, \ell, m} \exp \left\{ \beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m)}{\sum_{g, \ell, m} \exp \left\{ \beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \xi_\ell} + \lambda.
\]

Multiplying by \( \alpha_m \) and then summing over \( m \), we see that \( \lambda = 0 \). Hence the solution is

\[
\alpha_m = \frac{n(s_m)}{n_{1P}} \frac{\sum_{g, \ell, m} \exp \left\{ \beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \xi_\ell \alpha_m}{\sum_{g, \ell, m} \exp \left\{ \beta_0 + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \xi_\ell}.
\]

Following the same basic steps as we did with \( \xi_\ell \), and invoking (3) and (5), we see that

\[
\alpha_m = \frac{n(s_m)}{n_{1P}} \left[ \sum_{g, \ell} \exp \left\{ \kappa + r(s_m, \delta) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \right]^{-1} = \frac{n(s_m)}{n_{1P}} \exp \left\{ h(s_m) \right\}. \tag{A3}
\]

Hence, recalling that \( \psi(s) = r(s, \delta) + h(s) \), we obtain

\[
1 = \sum_{g, \ell} \exp \left\{ \kappa + \psi(s_m) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m) \xi_\ell. \tag{A4}
\]

Now substitute the \( \alpha_m \) of (A3) into (A2), so that

\[
\xi_\ell = \xi(x_\ell) = \left[ n_{0P} + \sum_{g, m} n(s_m) \exp \left\{ \kappa + \psi(s_m) + m(g, x_\ell, \beta) \right\} Q(g \mid s_m, \theta) \right]^{-1}.
\]

The sum over \( m \) in the expression above is really a sum over the observed \( S \) data in the primary case sample. We therefore have that

\[
\xi(x_\ell) = \left[ n_{0P} + \sum_{g, m=1}^{n_{1P}} \sum \exp \left\{ \kappa + \psi(S_m) + m(g, x_\ell, \beta) \right\} Q(g \mid S_m, \theta) \right]^{-1} = \mathcal{D}(x_\ell, \kappa, \beta, \theta, \psi). \tag{A5}
\]

Now, substituting \( \xi(x_\ell) \) and \( \alpha(s_m) \) into (4) gives the profile likelihood in Lemma 1.

### A2. Proof of Proposition 1

To simplify expressions, as in (7) the subscripts \( u \) and \( p \) in the following expressions will index all the \( N = n_{1P} + n_{0P} + n_{0E} \) subjects in the whole sample, combining primary and external samples, and \( O_u = 1 \) indicates subject \( u \) being in the primary sample, with \( O_u = 0 \) otherwise.

For given \( X \) and a function \( c(G, S, X, \Omega^*) \) where \( \Omega^* = (\Omega, \psi) \) is as defined in § 3.3, write \( \hat{c}(\cdot, \Omega^*) = (\partial/\partial \Omega^*)c(\cdot, \Omega^*) \) and \( \check{c}(\cdot, \Omega^*) = (\partial^2/\partial \Omega^* \partial \Omega^*)c(\cdot, \Omega^*) \). With the notation \( c^{\partial_1} = c \) and \( c^{\partial_2} = cc^T \) for a
vector c, we define the quantities $E_O(Dc(G, X, \Omega^*)^{(a)} \mid X) (a = 1, 2)$ by

$$E_O(Dc(G, X, \Omega^*)^{(a)} \mid X) = \frac{\sum_{i=1}^{n_p} \sum_{g} c(g, S_u, X, \Omega^*)^{(a)} \exp[\eta(g, S, X, \psi, \Omega)]}{n_{op} + \sum_{i=1}^{n_p} \sum_{g} \exp[\eta(g, S, X, \psi, \Omega)]},$$

where in the second equality the notation used in (7) is adopted. Let $V_O[c(\cdot) \mid X] = E_O[c(\cdot)^{(2)} \mid X] - E_O[c(\cdot)^{(1)} \mid X]E_O[c(\cdot)^{(1)} \mid X]^T$. Similarly, for given S and a function $b(G, S, \theta)$,

$$E_O[b(G, S, \theta)^{(a)} \mid S] = \frac{\sum_{g=0}^{2} b(g, S, \theta)^{(a)} \exp[\theta_{0g} + g\theta_1 f(S)]}{\sum_{g=0}^{2} \exp[\theta_{0g} + g\theta_1 f(S)]} (a = 1, 2)$$

with $\theta_{0g} \equiv 0$ for $g = 0$, and $V_O[b(\cdot) \mid S] = E_O[b(\cdot)^{(2)} \mid S] - E_O[b(\cdot)^{(1)} \mid S]E_O[b(\cdot)^{(1)} \mid S]^T$.

The score function for the profile likelihood obtained by differentiating (6), or equivalently (7), is of the form $U_{\Omega^*} = \sum_{p=1}^{N} U_{\Omega^*, p}(\Omega^*)$ with

$$U_{\Omega^*, p}(\Omega^*) = O_p \left[ D_p \hat{\eta}(G_p, S_p, X_p, \Omega^*) - E_O [D_p \hat{\eta}(G, S, X, \Omega^*) \mid X = X_p] + (1 - O_p) \left[ \hat{\rho}(G_p, S_p, \theta) - E_O \{ \hat{\rho}(G, S, \theta) \mid S = S_p \} \right] \right],$$

where $\rho(G, S, \theta) = \theta_{0g} + G\theta_1 f(S)$. The observed information matrix, $I^* = -(\partial^2 \Omega^*) U_{\Omega^*}$, is

$$I^* = \sum_{p=1}^{N} O_p V_O \left[ D_p \hat{\eta}(G, S, X, \Omega^*) \mid X = X_p \right] + (1 - O_p) V_O \left[ \hat{\rho}(G, S, \theta) \mid S = S_p \right] - \sum_{p=1}^{N} O_p \left[ D_p \hat{\eta}(G_p, S_p, X_p, \Omega^*) - E [D \hat{\eta}(G, S, X, \Omega^*) \mid X = X_p] \right].$$

The last term in (A7) is of the order $o_p(N)$ and hence can be dropped in general; it vanishes in the standard logistic model where $m(G, X, \beta)$ is linear in $\beta$.

Before deriving the asymptotic results for $\hat{\Omega}$, we state a lemma.

**Lemma A1.** Let $n_p = n_{1p} + n_{0p}$ and $n_{1p}/n_p \rightarrow \mu_i > 0$ for $i = 0, 1$. For a function $c(G, S, X, \Omega^*)$ we have, for $a = 1, 2$,

$$n_p^{-1} \sum_{i=1}^{n_{1p} + n_{0p}} E_O \{ Dc(G, S, X, \Omega^*)^{(a)} \mid X = X_i \} \rightarrow \mu_1 E \{ Dc(G, S, X, \Omega^*)^{(a)} \mid D = 1 \},$$

where the convergence is in probability, provided that the expectation on the right-hand side of the above expression exists.

**Proof.** Let $\alpha(S) = pr(S \mid D = 0)$, $\alpha^+(S) = pr(S \mid D = 1)$, $\xi(X) = pr(X \mid D = 0)$ and $\xi^+(X) = pr(X \mid D = 1)$. Recall that, by our modelling assumptions and definitions for $\psi$ and $\eta$,

$$pr(G, X, S \mid D = 1) = \exp[\kappa + r(S, \delta) + m(G, X, \beta)] Q(G \mid S, \theta) \alpha(S) \xi(X) = \exp[\eta(G, S, X, \Omega, \psi(S))] \alpha^+(S) \xi(X).$$

Hence, summing both sides of (A8), we obtain

$$\hat{h}(X) \equiv \frac{\xi^+(X)}{\xi(X)} = \int_S \sum_{g} \exp[\eta(g, s, X, \Omega, \psi(s))] \alpha^+(s).$$
Then
\[
\frac{1}{n_p} \sum_{j=1}^{n_{ip}+n_{ip}} E_O \{ Dc(G, S, X, \Omega^*) \mid X = X_j \} = \frac{1}{n_p} \sum_{j=1}^{n_{ip}+n_{ip}} \sum_{g} c(g, S_j, X_j, \Omega^*) \exp[\eta(g, S_j, X_j, \psi_j, \Omega)]
\]
\[
= \int_x \frac{\sum_{g} c(g, s, x, \Omega^*) \exp[\eta(g, s, x, \psi(s), \Omega)]}{\mu_0/\mu_1} \alpha^+(s) \xi(x) + o_p(1).
\]

From (A8) and (A9), the limiting expression above then reduces to
\[
\mu_1 \int_x \int_s \sum_{g} c(g, s, x, \Omega^*) \exp[\eta(g, s, x, \psi(s), \Omega)] \alpha^+(s) \xi(x) = \mu_1 \int_x \sum_{g} \exp[\eta(g, s, x, \psi(s), \Omega)] \alpha^+(s) \xi(x) = E \{ Dc(g, s, x, \Omega^*) \mid D = 1 \}.
\]

From Lemma A1,
\[
N^{-1} \sum_{p=1}^{N} O_p \{ D_p \hat{\eta}(G, S_p, X_p, \Omega^*) - E_O \{ D \hat{\eta}(G, S, X, \Omega^*) \mid X = X_p \} \}
\]
\[
\rightarrow \tau \mu_1 \{ E \{ D \hat{\eta}(G, S, X, \Omega^*) \mid D = 1 \} - E \{ D \hat{\eta}(G, S, X, \Omega^*) \mid D = 1 \} \} = 0
\]
in probability, where \( \tau = \lim(n_{ip} + n_{ip})/N \). The zero expectation of \( (1 - O_p) [ \hat{\rho}(G, S_p, \theta) - E_O [ \hat{\rho}(G, S, \theta) \mid S = S_p] ] \) is obvious. We have thus established the asymptotic unbiasedness of the score function \( U_{\Omega^*} \), given in (A6). Further,
\[
N^{-1} I^* = \tau \int_x V_O \{ D \hat{\eta}(G, S, X, \Omega^*) \mid X = x \} \{ \mu_1 h(x) + \mu_0 \} \xi(x)
\]
\[
= (1 - \tau) \int_x V_O \{ \hat{\rho}(G, S, \theta) \mid S = s \} \alpha(s) + o_p(1)
\]
\[
= I^* + o_p(1).
\]

The assumption that \( I^* \) is positive definite ensures the existence of a unique consistent solution \( \hat{\Omega}^* \) to the score equation \( U_{\Omega^*(\hat{\Omega}^*)} = \{ U_{\Omega^*}^{\Omega^*}, U_{\Omega^*}^{\psi,\psi} \} = 0 \). Following similar calculations as in the proof of Lemma A1, we can verify that
\[
N^{-1} \sum_{p=1}^{N} U_{\Omega^*, \theta}(\Omega^*) U_{\Omega^*, \theta}(\Omega^*) = I^* + o_p(1).
\]

Write \( U_{\Omega^*, \theta} = (U_{\Omega^*, \theta}^{\Omega^*}, U_{\Omega^*, \theta}^{\psi,\psi})^T \), where \( U_{\Omega^*, \theta}^{\Omega^*} \) and \( U_{\Omega^*, \theta}^{\psi,\psi} \) are components of the score corresponding to \( \Omega \) and \( \psi \), respectively. By Taylor expansion, and writing \( I_{\Omega^*} = I_{\Omega^*}^{\Omega^*} = I_{\Omega^*}^{\psi,\psi} I_{\Omega^*}^{\psi,\psi} = I_{\Omega^*}^{\psi,\psi} I_{\Omega^*}^{\psi,\psi} \),
\]
\[
N^{1/2} (\hat{\Omega} - \Omega) = N^{-1/2} I_{\Omega^*} \left\{ \sum_{p=1}^{N} U_{\Omega^*, \theta}(\Omega^*) - I_{\Omega^*}^{\psi,\psi} I_{\Omega^*}^{\psi,\psi} \sum_{p=1}^{N} U_{\Omega^*, \theta}(\Omega^*) \right\} + o_p(1).
\]

Applying the central limit theorem and (A10), we then establish the asymptotic normality for \( \hat{\Omega} \).
A3. The estimating function and Hessian for the pseudolikelihood

Differentiating \( \log[L_{pse}(\Omega, \psi(\Omega))] \) with respect to \( \Omega = (\kappa, \beta, \theta) \) gives

\[
(\partial/\partial\Omega) \log[L_{pse}(\Omega, \psi(\Omega))] = \left[ (\partial/\partial\Omega) \log[L_{prof}(\Omega, \psi)] + J_{\Omega\psi} (\partial/\partial\psi) \log[L_{prof}(\Omega, \psi)] \right]_{\psi = \tilde{\psi}(\Omega)},
\]

where \( L_{prof} \) is the profile likelihood given in (7) and \( J_{\Omega\psi} \) is the \( q \times n_{1p} \) matrix with \( m \)-th column \( (m = 1, \ldots, n_{1p}) \) defined by \( J_{\Omega\psi,m} = (\partial/\partial\Omega) \tilde{\psi}_m(\Omega) \) where \( \tilde{\psi}_m(\Omega) = \psi(S_m, \Omega) \). The explicit expression for \( J_{\Omega\psi,m} \) is

\[
J_{\Omega\psi,m} = \frac{\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \phi(g, S_m, X_p, \Omega) \exp(\phi(g, S_m, X_p, \Omega))}{\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \exp(\phi(g, S_m, X_p, \Omega))},
\]

where \( \phi(g, s, x, \Omega) = \kappa + m(g, x, \beta) + \theta_0 g_1 f(s) \) and \( \dot{\phi}(\cdot, \Omega) = (\partial/\partial\Omega) \phi(\cdot, \Omega) \). Using the notation from Appendix A2, the right-hand side of (A11) is

\[
\sum_{p=1}^{N} [U_{\Omega,p}(\Omega, \tilde{\psi}(\Omega)) + J_{\Omega\psi} U_{\psi,p}[\Omega, \tilde{\psi}(\Omega)]].
\]

(A12)

It can be seen that \( \tilde{\psi}_m(\Omega) \) is a solution of \( \psi_m \) to 0 = \( \sum_{p=1}^{N} \tilde{U}_{\psi_m,p}(\Omega, \psi_m) \), where

\[
\tilde{U}_{\psi_m,p}(\Omega, \psi_m) = O_p \left[ D_p I(S_p = S_m) - \frac{1 - D_p}{n_{0p}} \sum_{g=0}^{2} \exp(\phi(g, S_m, X_p, \Omega) + \psi(S_m)) \right].
\]

Let \( \tilde{U}_{\psi,p}(\Omega, \psi) \) be the \( n_{1p} \)-vector \( \{\tilde{U}_{\psi_m,p}(\Omega, \psi_m) : m = 1, \ldots, n_{1p}\} \). Also, let \( T_{\Omega\psi}^* \) and \( T_{\psi\psi}^* \) be the submatrices of \( T^* \) as in Proposition 1, and let \( J_{\Omega\psi} \) and \( \tilde{J}_{\psi\psi} \) be the limiting matrices of \( J_{\Omega\psi} \) and \(-N^{-1}(\partial/\partial\psi) \sum_{p=1}^{N} \tilde{U}_{\psi,p}, \) respectively. By Taylor expansion with respect to \( \psi \), (A12) can be rewritten as \( \sum_{p=1}^{N} \tilde{U}_{\Omega,p} + o_p(N^{1/2}) \), where

\[
\tilde{U}_{\Omega,p} = U_{\Omega,p}(\Omega, \psi) + J_{\Omega\psi} U_{\psi,p}(\Omega, \psi) - (T_{\Omega\psi}^* + J_{\Omega\psi} T_{\psi\psi}^*) \tilde{J}_{\psi\psi}^{-1} \tilde{U}_{\psi,p}(\Omega, \psi).
\]

The negative Hessian matrix for the pseudolikelihood is

\[
\tilde{I}_{\Omega} = -(\partial^2/\partial\Omega \partial\Omega') \log[L_{pse}] = I_{\Omega\Omega}^* + I_{\Omega\psi}^* J_{\Omega\psi}^T + J_{\Omega\psi} I_{\psi\psi}^* + J_{\Omega\psi} I_{\psi\psi}^* J_{\Omega\psi}^T + J_{\Omega\Omega}.
\]

Here \( I_{\Omega\Omega}^* \) and \( I_{\psi\psi}^* \) are submatrices of the information matrix \( I^* \), derived in (A7), for the profile likelihood, and \( J_{\Omega\psi} = \sum_{m=1}^{n_{1p}} U_{\psi_m}(\Omega, \tilde{\psi}(\Omega)) K_m(\Omega) \), where \( U_{\psi_m} \) is the \( m \)-th component of \( U_{\psi} \) for \( m = 1, \ldots, n_{1p} \) and \( K_m(\Omega) = -(\partial^2/\partial\Omega \partial\Omega') \tilde{\psi}_m(\Omega) \) with explicit expression

\[
\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \phi(g, S_m, X_p, \Omega) \exp(\phi(g, S_m, X_p, \Omega))
\]

\[
\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \exp(\phi(g, S_m, X_p, \Omega))
\]

\[
+ \frac{\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \phi(g, S_m, X_p, \Omega)^2 \exp(\phi(g, S_m, X_p, \Omega))}{\sum_{p=1}^{N} O_p (1 - D_p) \sum_g \exp(\phi(g, S_m, X_p, \Omega))} - J_{\Omega\psi,m} J_{\Omega\psi,m}^T,
\]

where \( \phi(\cdot, \Omega) = (\partial^2/\partial\Omega \partial\Omega') \phi(\cdot, \Omega) \).
Shared genetic controls

A4. The score and information for the conditional likelihood

Here we redefine $\Omega = (\beta, \theta)$. For a function $c(G, X, S, \Omega)$ define, for $a = 1, 2,$

$$
E_S[c(G, S, X, \Omega)^{(a)} | S, X] = \frac{\sum_{g=0}^{2} c(g, X, S, \Omega)^{(a)} \exp[\theta_{0g} + \beta_{Gg} + g \sum_k \theta_{1k} f_k(S) + \beta_{GgX} X]}{\sum_{g=0}^{2} \exp[\theta_{0g} + \beta_{Gg} + g \sum_k \theta_{1k} f_k(S) + \beta_{GgX} X]},
$$

with $\theta_{0g} = 0$ for $g = 0$.

The score function for the conditional likelihood (9) is

$$
\hat{U}_\Omega = \sum_{p=1}^{N} O_p D_p \left[ \hat{\phi}(G_p, S_p, X_p, \Omega) - E_S \left\{ \hat{\phi}(G, S, X, \Omega) | S = S_p, X = X_p \right\} \right] + \sum_{p=1}^{N} (1 - O_p) \left[ \hat{\rho}(G_p, S_p, \theta) - E_Q \left\{ \hat{\rho}(G, S, \theta) | S = S_p \right\} \right],
$$

where $\phi(G, S, X, \Omega) = \theta_{0G} + \beta_{G} G + \theta_{1} f(S) + \beta_{G} G X \cdot \theta = (\partial / \partial \Omega) \phi,$ and $\hat{\rho}$ and $E_Q$ are as defined in Appendix A2. Let $V_S[c(\cdot) | S, X] = ES[c(\cdot)^{(2)} | S, X] - ES[c(\cdot)^{(1)} | S, X]ES[c(\cdot)^{(1)} | S, X]^T$. With $V_Q(\cdot)$ defined as in Appendix A2, the observed information matrix $\hat{I}_\Omega = -(\partial / \partial \Omega^T) \hat{U}_\Omega(\Omega)$ is given by

$$
\hat{I}_\Omega(\Omega) = \sum_{p=1}^{N} O_p D_p V_S \left\{ \hat{\phi}(G, S, X, \Omega) | S = S_p, X = X_p \right\} + \sum_{p=1}^{N} (1 - O_p) V_Q \left\{ \hat{\rho}(G, S, \theta) | S = S_p \right\}.
$$ (A13)

A5. Equivalence results under models without gene-environment interaction

In § 4, we found that the conditional likelihood method of § 3.5 is numerically identical to the semi-parametric maximum likelihood method of § 3.3 and the pseudolikelihood method of § 3.4, when it is assumed that $\beta_{G} = 0$, no gene-environment interaction, in the standard logistic regression model, i.e.,

$$
m(G, X, \beta) = \beta_{G} G + \beta_{X} X + \beta_{G} G X.
$$

Here is an explanation of the equivalence of the three methods in this setting. Since $\Pr(D, G, S | X)$ in the profile likelihood (7) is a function of $X$ only when $D = 0$, it contains no information on the parameters related to genotype. When $D = 1$,

$$
\Pr(D = 1, G, S | X) \propto \Pr(G | X, S, D = 1) v(S, X),
$$

where $\propto$ means equivalence up to a multiplicative term depending on $X$ only, $\Pr(G | X, S, D = 1)$ is as given in (10) under the standard logistic regression model, and

$$
v(S, X) = \exp[\kappa + \psi(S) + \beta_{X} X] w(S, X)
$$ (A14)

with $w(S, X) = \sum_g \exp(\beta_{G} G + \beta_{G} G X) Q(g | S, \theta).$ When $\beta_{G} = 0$, $w(S, X) = w(S)$ is a function of $S$ only. Therefore, if $\psi(S)$ is parameterized by a saturated model, as we did in § 3.3 and § 3.4, then the information in the primary case sample for parameters related to genotype can no longer be extracted from (A14) but only from $\Pr(G | X, S, D = 1)$. Since the three methods have the same likelihood for data from the external control sample, we can see the equivalence of the three methods for parameters related to genotype. Of course, this property pertains only to the standard logistic regression model.
REFERENCES


[Received July 2011. Revised October 2012]