HHS Public Access

Author manuscript

J Multivar Anal. Author manuscript; available in PMC 2019 November 01.

Published in final edited form as:

J Multivar Anal. 2019 September ; 173: 38–50. doi:10.1016/j.jmva.2019.01.006.

A semiparametric efficient estimator in case-control studies for gene–environment independent models

Liang Liang[a,*], **Yanyuan Ma**[b], **Raymond J. Carroll**[c,d]

Yanyuan Ma: yzm63@psu.edu; Raymond J. Carroll: carroll@stat.tamu.edu

[a]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

[b]Department of Statistics, Penn State University, University Park, PA 16802, USA

[c]Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843, USA

[d]School of Mathematical and Physical Sciences, University of Technology Sydney, PO Box 123, Broadway NSW 2007, Australia

Abstract

Case-controls studies are popular epidemiological designs for detecting gene–environment interactions in the etiology of complex diseases, where the genetic susceptibility and environmental exposures may often be reasonably assumed independent in the source population. Various papers have presented analytical methods exploiting gene–environment independence to achieve better efficiency, all of which require either a rare disease assumption or a distributional assumption on the genetic variables. We relax both assumptions. We construct a semiparametric estimator in case-control studies exploiting gene–environment independence, while the distributions of genetic susceptibility and environmental exposures are both unspecified and the disease rate is assumed unknown and is not required to be close to zero. The resulting estimator is semiparametric efficient and its superiority over prospective logistic regression, the usual analysis in case-control studies, is demonstrated in various numerical illustrations.

Keywords

Biased samples; Case-control study; Gene–environment independence; Gene–environment interaction; Semiparametric estimation

1. Introduction

The etiology of most complex diseases, such as cancers and cardiovascular diseases, is the joint effect of genetic susceptibility and environmental or non-genetic exposures, as well as their interactions. Even subtle differences in genetic factors between people, when exposed to the same environmental factors, can lead to dramatically different responses. In other words, people with certain genes may have a low risk of developing a disease whereas others

*Corresponding author. liliang@hsph.harvard.edu.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2019.01.006.

may be more vulnerable when exposed to an identical environmental agent. One common example is that sunlight exposure results in higher risk of developing skin cancer among fair-skinned individuals than people with dark skin [17,22]. Studying gene–environment interactions is thus of great importance to understand disease mechanisms and develop new treatments and prevention strategies.

The case-control study design is commonly used to investigate the intricate interplay of genetic susceptibility and environment effects. It is cost-efficient and convenient to implement compared to a cohort study, especially when dealing with relatively rare diseases [6]. Instead of taking a random sample from the underlying source population, the case-control design randomly draws a fixed number of cases (diseased subjects) and a comparable number of controls (non-diseased subjects) from the respective case and control subpopulations. Genetic and environmental factors are measured and recorded for these sampled subjects. The standard approach for the analysis of such a case-control study is prospective logistic regression, which ignores the underlying retrospective nature of the case-control design. Cornfield [10] showed the equivalence of prospective and retrospective odds ratios, which validates the prospective approach. Prentice and Pyke [24] further showed that prospective logistic regression analysis gives an efficient estimator, in the sense that it yields the maximum likelihood estimates of the odds ratio parameters under a semiparametric model that allows an arbitrary covariate distribution.

Despite this, prospective logistic regression treatment in a case control study can still require a large sample size to obtain adequate statistical power for detecting gene–environment interactions or testing other hypotheses of interest. As a consequence, epidemiological researchers often exploit the potential efficiency gain from further assuming certain parametric or semiparametric structures for the covariate distribution. For example, in practice, a common assumption is that genetic susceptibility and environmental exposure are independent in the underlying source population [23], possibly given strata. Under such a model, prospective logistic regression analysis is still valid but may not be efficient because it ignores gene–environment independence.

A growing number of articles have been published in the last two decades, proposing analytical methods that exploit gene–environment independence assumption [5,14,15,20,21,23]. Piegorsch et al. [23] showed that under gene–environment independence and a rare disease assumption, the multiplicative interaction odds-ratio parameter can be estimated by cases alone and the resulting estimator is more precise than the estimator from traditional prospective logistic regression analysis using both cases and controls. However, the misuse of a rare disease assumption in analyzing diseases with moderate prevalence or diseases with small marginal probability in the source population but high risk for certain combination of genetic and environmental exposures can lead to considerable bias in the estimation. Noting this fact, Chatterjee and Carroll [5] developed a semiparametric maximum likelihood estimator employing the gene–environment independence assumption but not requiring any rare-disease assumption. Their approach leaves the distribution of the environmental exposures totally unspecified but restricts genetic susceptibility to have a discrete distribution that takes values in a finite and fixed set. Ma [20] proposed a semiparametric efficient estimator in the same setting as Chatterjee and Carroll [5] except

the distribution of genetic susceptibility is allowed to be either discrete or continuous with a finite-dimensional parameter. The key ingredient of this approach is to construct a hypothetical population with infinite population size and a disease to non-disease ratio of n_1/n_0, where n_1 and n_0 are the numbers of cases and controls in the case-control sample. Section 2 of Ma [20] showed that the case-control sample can be viewed as a size $n = n_0+n_1$ random sample of independent and identically distributed observations from this hypothetical population, and hence classical semiparametric analysis is applicable. The validity and usefulness of such a hypothetical population was established in Ma [20]. Instead of assuming independence of gene and environment, there is a literature based on parametric modeling of the relationship between them [8,9,18,19]: we make no such parametric assumptions.

In this paper, we consider a more general setting which keeps the gene–environment independence assumption, while further allowing an unknown disease rate and completely nonparametric distributions for both the genetic susceptibility and the environmental exposure. Under such a model setting, we adopt the hypothetical population framework of Ma [20] and derive the semiparametric efficient estimator by employing a semiparametric approach, which links the efficient estimator with the efficient score function. Throughout our work, the underlying source population is referred to as the true population to emphasize the difference between the underlying source population and the hypothetical population. The inherent connection between the two populations allows us to transport parameter estimation and inference results derived in the hypothetical population directly to those in the true population, see Theorem 1. Although general semiparametric theory applies in the hypothetical population framework, computing the efficient estimator in this context is technically challenging because the efficient score does not have an explicit form and must be solved from an integral equation. We adopt a simple numerical approach to solve the integral equation by discretizing the distribution of the genetic susceptibility when it is continuous. The resulting estimator, when properly implemented, is asymptotically linear with optimal efficiency.

The rest of the paper is organized as follows. The specific model and the hypothetical population framework are presented in Section 2, with the corresponding identifiability conditions provided in Appendix A.1. In Section 3, we formulate the problem by using a conventional semiparametric approach. The analytic expression of our semiparametric efficient estimator as well as its detailed implementation are discussed in this section. Section 4 illustrates the asymptotic properties of the resulting estimator. Several simulation studies are conducted in Section 5 to demonstrate the numerical performance of our semiparametric efficient estimator compared with prospective logistic regression. A real data analysis is provided in Section 6, followed with a brief discussion in Section 7. Technical details and proofs are given in an Appendix and in the Online Supplement.

2. Model and framework

2.1. Background

It is useful to describe how the methods, referenced in Section 1, for exploiting a genetic-environmental relationship in an underlying source population have evolved from the earlier

work, a relatively simple case in Chatterjee and Carroll [5], which includes the following key ingredients:

- **a.** An underlying logistic regression for disease D as a function of genetic variables G and environmental exposures X.
- **b.** A parametric distribution assumption for G in the source population when G and X are independent.
- **c.** Writing out the retrospective likelihood of the observed case-control data.
- **d.** A profile likelihood argument that estimates the distribution of X in the source population using a Lagrange multiplier argument that places probability mass at each observed value of X. This leads to a pseudolikelihood that involves the distribution of G but not the distribution of X.
- **e.** The main technical difficulty is carrying out the algebra of the Lagrange multiplier argument and getting an explicit pseudolikelihood, where by explicit we mean that the resulting formula requires no numerical solutions to nonlinear equations.

In our case, however, we are not making the assumption of a parametric distribution for G in the source population. A profile likelihood method to remove the distribution of G and get a new, explicit, profile likelihood based on a Lagrange multiple argument does not appear to be possible, or at least it seems to be very difficult, because of the form of the pseudolikelihood.

To overcome these difficulties, there have been two main alternatives, and they are both based on the idea of relating the case-control study to some version of a prospective random sampling framework to derive a methodology, and to then show that this methodology is valid in the case-control study. Recall that n_0 is the number of controls in the sample and n_1 the number of cases. Define $\pi_d = \Pr(D = d)$.

- **i.** I. In Section 2.3.3 of [9], Chen et al. treat the case-control study as if it were a random sample from the source population but with data missing at random. They propose a prospective sampling scenario where each subject from the source population is observed with probability $1/\{1 + (n_{1-d}\pi_d)/(n_d\pi_{1-d})\}$, where $d = 1$ for cases and $d = 0$ for controls, respectively. They show that performing a missing data analysis for the distribution of (D, G) given X and the probability that the subject is observed yields the same pseudolikelihood as other papers have computed, but without having to do the Lagrange argument, and in a much easier way.
- **ii.** II. Ma [20] takes an entirely different approach, also without having to do the Lagrange argument. This approach, which she calls a hypothetical population approach, differs from that of Chen et al. [9] in that she aims to create a likelihood that (a) is equivalent to that of the case-control sample; and (b) is that of a simple random sample of size $n = n_0 + n_1$ from a hypothetical population.

  Because it is a random sample, rather than a sample with missing data, when we

use it this allows us to rely on the classic machinery of semiparametric methods as exemplified by Bickel et al. [4] and Tsiatis [27].

2.2. Basic calculations and likelihood

Assume that the prospective risk given the covariates (*G, X*) follows a logistic model, viz.

$$\begin{aligned}\Pr(D=d\,|\,G=g,X=x) &= f^{\text{true}}_{D|G,X}(d,g,x) = H(d,g,x,\boldsymbol{\theta})\\ &= \frac{\exp[d\{\alpha+m(g,x,\boldsymbol{\beta})\}]}{1+\exp\{\alpha+m(g,x,\boldsymbol{\beta})\}},\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}=\left(\boldsymbol{\beta}^{\top},\boldsymbol{\alpha}\right)^{\top}$ and m is a function known up to the parameter $\boldsymbol{\beta}$. Here and throughout the text, the superscript "true" is used to emphasize that those quantities are related to the true source population. In addition, in the true population, G and X are assumed to be independent so that the joint probability density/mass function of G, X can be written as $f^{\text{true}}_{G,X}(g,x)=f^{\text{true}}_{G}(g)f^{\text{true}}_{X}(x)=\eta_1(g)\eta_2(x)$. Here, for notational simplicity, we write $\left\{f^{\text{true}}_{G}(g),f^{\text{true}}_{X}(x)\right\}$ as $\left\{\eta_1(g),\eta_2(x)\right\}$. The problem stated above is identifiable in the case-control study under mild conditions, which are given in Appendix A.1, along with the proof of identifiability.

The hypothetical population study joint density/mass function of (*D, G, X*) is

$$\begin{aligned}f_{D,G,X}(d,g,x,\boldsymbol{\theta},\eta_1,\eta_2) &= (n_d/n)f_{G,X|D}(d,g,x) = (n_d/n)f^{\text{true}}_{G,X|D}(d,g,x)\\ &= \frac{n_d}{n}\frac{f^{\text{true}}_{G}(g)f^{\text{true}}_{X}(x)f^{\text{true}}_{D|G,X}(d,g,x,\boldsymbol{\theta})}{\int f^{\text{true}}_{G}(g)f^{\text{true}}_{X}(x)f^{\text{true}}_{D|G,X}(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g)}\\ &= \frac{n_d\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})}{n\int\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g)}\\ &= \frac{n_d}{n\pi_d}\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta}),\end{aligned} \tag{2}$$

where

$$\pi_d=\int\eta_1(g)\eta_2(x)H(d,g,x,\boldsymbol{\theta})d\mu(x)d\mu(g)\,. \tag{3}$$

We consider $\eta=\{\eta_1,\eta_2\}$ as the infinite-dimensional nuisance parameter. The approach of Ma [20] views this as a semiparametric problem, to be solved using techniques explained in Bickel et al. [4] and Tsiatis [27]. Here, the concept of hypothetical population and the corresponding distorted likelihood is used as a vehicle to allow us to transport the semiparametric tools for direct application. It enables us to construct consistent estimators without having to concern about the non-random sample issue in case-control study. Because the non-random sampling issue is already taken into account when we formulate the distorted likelihood, the resulting estimator is indeed automatically consistent under the

original case-control sampling framework, that is, if the case-control sample size grows to infinity while retaining the relative sample proportion of n_1/n_0, the estimator will converge to the true parameter value. We formally write out this result in Theorem 1.

Theorem 1. *Assume* $(d_1, g_1, x_1), \ldots, (d_n, g_n, x_n)$ *is a case-control sample with* n_1 *cases,* n_0 *controls, and with disease model* (1) and independence of X and G. Assume $(\tilde{d}_1, \tilde{g}_1, \tilde{x}_1), \ldots, (\tilde{d}_n, \tilde{g}_n, \tilde{x}_n)$ is a random sample of independent and identically distributed *observations with size n from model* (2). Then, if $\hat{\boldsymbol{\theta}}\{(\tilde{d}_1, \tilde{g}_1, \tilde{x}_1), \ldots, (\tilde{d}_n, \tilde{g}_n, \tilde{x}_n)\}$ is a $\sqrt{n}$ -consistent regular asymptotically linear estimator of $\boldsymbol{\theta}$ *and satisfies* $\mathrm{E}[\hat{\boldsymbol{\theta}}\{(\tilde{d}_1, \tilde{g}_1, \tilde{x}_1), \ldots, (\tilde{d}_n, \tilde{g}_n, \tilde{x}_n)\} \mid D] - \boldsymbol{\theta} = o_p(n^{-1/2})$, then so is $\hat{\boldsymbol{\theta}}\{(d_1, g_1, x_1), \ldots, (d_n, g_n, x_n)\}$.

Theorem 1 essentially says that if we can develop a $\sqrt{n}$ -consistent estimator based on a random sample from model (2), then we can simply apply this estimation procedure to the case-control sample and we will still get a $\sqrt{n}$ -consistent estimator. The proof of Theorem 1 is the entire content of Section 2 of Ma [20]. We take advantage of this property to generate an estimation procedure, which we will then show consistently estimates the parameters when using the case-control data. In particular, the procedure is not dependent on the hypothetical population study formalism.

3. Analytic derivations: Efficient score and algorithm

The outline of the semiparametric approach is to first construct a Hilbert space $\mathcal{H}$, consisting of all measurable functions with mean zero and finite variance. We next decompose $\mathcal{H}$ into nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$. The efficient estimator can then be obtained by solving

$$\sum_{i=1}^{n} S_{\mathrm{eff}}(D_i, G_i, X_i; \boldsymbol{\theta}) = 0,$$

where $\mathbf{S}_{\mathrm{eff}}$ is the projection of the score function $\mathbf{S}_{\boldsymbol{\theta}}$ onto $\Lambda^{\perp}$, and thus $\mathbf{S}_{\mathrm{eff}}$ is called efficient score function.

Careful calculation shows that the score function under the hypothetical population (2) takes the form $\mathbf{S}_{\boldsymbol{\theta}}(d, g, x) = \mathbf{S}(d, g, x) - \mathrm{E}(\mathbf{S} \mid d)$ where $\mathbf{S} = \{d - H(1, g, x, \boldsymbol{\theta})\}\{\mathbf{m}'_{\boldsymbol{\beta}}(g, x, \boldsymbol{\beta})^{\top}, 1\}^{\top}$ and $\mathbf{m}'_{\boldsymbol{\beta}}(g, x, \boldsymbol{\theta}) \equiv \partial m(g, x, \boldsymbol{\theta}) / \partial \boldsymbol{\beta}$. Let p denote the dimension of $\boldsymbol{\theta}$. The final form of the spaces Λ and Λ is listed below with the detailed derivation provided in Appendix A.2. Specifically,

$$\Lambda = \{\mathbf{a}_1(G) + \mathbf{a}_2(X) - \mathrm{E}\{\mathbf{a}_1(G) + \mathbf{a}_2(X) \mid D\} \text{ for all } \mathbf{a}_1(G), \mathbf{a}_2(X)\},$$

$$\Lambda^{\top} = \{\mathbf{f}(D, G, X) : \mathrm{E}(\mathbf{f} \mid G) = \mathrm{E}\{\mathrm{E}(\mathbf{f} \mid D) \mid G\}, \mathrm{E}(\mathbf{f} \mid X) = \mathrm{E}\{\mathrm{E}(\mathbf{f} \mid D) \mid X\}, \mathrm{E}(\mathbf{f}) = 0\}.$$

Define $\mathbf{S}_x(x) = \mathrm{E}(\mathbf{S}_{\boldsymbol{\theta}} | x) = \mathrm{E}(\mathbf{S}|x) - \mathrm{E}\{\mathrm{E}(\mathbf{S}|D)|x\}$ and $\mathbf{S}_g(g) = \mathrm{E}(\mathbf{S}_{\boldsymbol{\theta}} | g) = \mathrm{E}(\mathbf{S}|g) - \mathrm{E}\{\mathrm{E}(\mathbf{S}|D)|g\}$. Projecting the score function onto $\Lambda^{\perp}$ shows that

$$\mathbf{S}_{\mathrm{eff}}(d, g, x) = \mathbf{S}(d, g, x) - \mathbf{a}(g) - \mathbf{b}(x) - \mathrm{E}\{\mathbf{S}(d, G, X)|d\} + \mathrm{E}\{\mathbf{a}(G) + \mathbf{b}(X)|d\},$$

where

$$\mathrm{E}\{\mathbf{a}(G)|x\} + \mathbf{b}(x) - \mathrm{E}\{\mathrm{E}(\mathbf{a} + \mathbf{b}|D)|x\} = \mathbf{S}_x(x), \tag{4}$$

$$\mathbf{a}(g) + \mathrm{E}\{\mathbf{b}(X)|g\} - \mathrm{E}\{\mathrm{E}(\mathbf{a} + \mathbf{b}|D)|g\} = \mathbf{S}_g(g). \tag{5}$$

It is easy to check that $\mathrm{E}\{\mathbf{S}_{\mathrm{eff}}(d, G_i, X_i)|d\} = \mathbf{0}$.

In order to obtain the efficient score function, we need to solve **a** and **b** from the integral equations (4) and (5). The existence of the solution is automatically guaranteed by the identifiability of the problem, whereas the uniqueness is not. However, it is shown in Appendix A.3 that **a** and **b** are unique up to constant shifts. Thus, (4) and (5) have a unique solution under the constraints $\mathrm{E}(\mathbf{a}) = \mathrm{E}(\mathbf{b}) = \mathbf{0}$. It is further proved in Appendix A.4 that, under the mean zero constraint, (4) and (5) have an equivalent expression, which is given by Eqs. (A.1)–(A.3), in the Appendix. Such an equivalent expression allows us to separate **a** and **b** by introducing an intermediate variable $\mathbf{u}_0 = \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0)$. However, there is no explicit expression for **a** and **b**. We still need to solve the integral equation (A.1). In Appendix A.5, we propose an approximation to its solution in the spirit of Tsiatis and Ma [28], by discretizing X if X is continuous.

The detailed algorithm for constructing the efficient score function and computing the efficient estimator for $\boldsymbol{\theta}$ is given in Algorithm 1, where the disease rate is estimated during the procedure. Usually, the disease prevalence is not identifiable from a case-control sample [24]. However, the additional assumption we make on the relationship between G and X in the source population, i.e., gene–environment independence, leads to the technical identifiability [5,20].

Algorithm 1

1. Estimate $f_{X|D=d}$ the conditional density/mass function of X given disease status $D = d$, by nonparametric kernel density estimation among the data with $D_i = d$ for $d \in \{0, 1\}$.

$$\hat{f}_{X|D=d}(x) = \frac{1}{n_d h} \sum_{i, D_i = d} K\{(X_i - x)/h\},$$

for continuous X, and

$$\hat{f}_{X|D=d}(x) = \frac{1}{n_d}\sum_{i, D_i = d} 1(X_i = x),$$

for discrete X, where K is a univariate kernel function.

2. Estimate $f_{G|D=d}$ the conditional density/mass function of G given disease status $D = d$, by nonparametric kernel density estimation among the data with $D_i = d$ for $d \in \{0, 1\}$. similarly as for X. Denote the result by $\hat{f}_{G|D}$.

3. Define $\hat{\eta}_1(g, \pi_0) = \pi_0 \hat{f}_{G|D=0}(g) + (1-\pi_0)\hat{f}_{G|D=1}(g)$, $\hat{\eta}_2(x, \pi_0) = \pi_0 \hat{f}_{X|D=0}(x) + (1-\pi_0)\hat{f}_{X|D=1}(x)$, what we call a weighted nonparametric density/mass function estimate, being weighted by the (estimated) population probabilities.

4. When (π_0, π_1) is unknown, estimate them by solving the integral equation

$$\pi_0 = \int H(0, g, x)\hat{\eta}_1(g, \pi_0)\hat{\eta}_2(x, \pi_0) d\mu(g) d\mu(x),$$

and setting $\hat{\pi}_1 = 1 - \hat{\pi}_0, \hat{\eta}_1(g) = \hat{\eta}_1(g, \hat{\pi}_0), \hat{\eta}_2(x) = \hat{\eta}_2(x, \hat{\pi}_0)$.

5. Follow the method described in Appendix A.5 to obtain the solution of the integral equations (4) and (5), with result $\hat{\mathbf{a}}, \hat{\mathbf{b}}$, and approximate $\mathrm{E}(\hat{\mathbf{a}} + \hat{\mathbf{b}} | D)$ using nonparametric density estimates $\hat{f}_{X|D}$ and $\hat{f}_{G|D}$ with result $\hat{E}(\hat{\mathbf{a}} + \hat{\mathbf{b}} | D)$.

6. From $\hat{\mathbf{S}}_{\mathrm{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}) = \hat{\mathbf{S}}_{\boldsymbol{\theta}}(D_i, G_i, X_i) - \hat{\mathbf{a}}(G_i) - \hat{\mathbf{b}}(X_i) + \hat{E}\{\hat{\mathbf{a}}(G_i) + \hat{\mathbf{b}}(X_i) | D_i\}$, and estimate $\boldsymbol{\theta}$ by solving the estimating equation

$$\sum_{i=1}^{n} \hat{\mathbf{S}}_{\mathrm{eff}}(D_i, G_i, X_i, \boldsymbol{\theta}) = \mathbf{0}. \quad (6)$$

It is critical that we estimate $\mathrm{E}\{\hat{\mathbf{a}}(G_i) + \hat{\mathbf{b}}(X_i) | D_i\}$ and $\mathrm{E}(\mathbf{S} | D_i)$ involved in Steps 5 and 6 using $\hat{f}_{X|D}$ and $\hat{f}_{G|D}$ described in Steps 1 and 2 of the above algorithm, instead of simply taking a sample version of the expectations. This ensures that all the conditional expectations are computed using the same kind of approximation and the gene–environment independence assumption is fully employed.

4. Distribution theory

It is not surprising that the semiparametric estimator described in Algorithm 1 is asymptotically normal with a parametric convergence rate and optimal efficiency as it is formed by estimating all conditional expectations in the efficient score nonparametrically. The asymptotic properties of our estimator are described in Theorem 2 under regularity conditions C1–C2 listed below. The proof is provided in the Online Supplement.

C1 The univariate kernel function K has support (−1, 1) and satisfies $\int K(u)udu = 0$, $\int K(u)u^2du < \infty$. The bandwidth h satisfies $nh^2 \to \infty$ and $nh^8 \to 0$.

C2 Any discrete covariate has finitely many levels. Any continuous covariate has compact support and its density function is twice continuously differentiable.

Theorem 2. *Under the regularity conditions C1 and C2, the estimator* $\hat{\boldsymbol{\theta}}$ *obtained from solving the estimating* Eq. (6) is asymptotically normal with optimal efficiency, i.e., $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to \mathcal{N}\left[0,\ var\left(\mathbf{S}_{\text{eff}}\right)^{-1}\right]$, and is semiparametric efficient.

5. Simulation study

We performed simulations to understand the finite sample performance of the semiparametric efficient estimator described in Section 3 and demonstrate its superiority to prospective logistic regression method under the gene–environment independent model. Two scenarios are considered: (a) $\Pr(D = 1) = 0.045$ and (b) $\Pr(D = 1) = 0.10$, corresponding to cases with a relatively rare disease rate and a common disease rate, respectively. In each scenario, we generated X from the standard normal distribution $\mathcal{N}(0,\ 1)$ or the Gamma distribution with mean 20 and variance 20, $\mathcal{G}(20,\ 1)$, while the distribution of G is one of the following: (i) Bernoulli with success probability 0.6 $\mathcal{B}(0.6)$, where for example $G = 1$ or $G = 0$ corresponds to the presence or absence of a genetic mutation, and (ii) $\mathcal{N}(0,\ 1)$, which can be used to model gene expression levels or continuous traits, such as height and skin color, that are controlled by several genes. Given G and X, we generated disease status D from the logistic regression model

$$\Pr(D = 1 \mid G, X) = 1/\left[1 + \exp\left\{-\left(\alpha + \beta_1 G + \beta_2 X + \beta_3 GX\right)\right\}\right],$$

where $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3\right)^{\top} = (0.76,\ 0.36,\ -0.63)$ for both settings with normal X, and $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \beta_3\right)^{\top} = (3.577,\ 0.080,\ -0.141)$ for both settings with Gamma X. We varied the intercept β_0 in different simulations to get the desired disease rate. Specifically speaking, in the case of $X = \mathcal{N}(0,\ 1)$, we set $\alpha = -3.61$ and -3.465 for binary G and normal G respectively to achieve a disease rate of 4.5%, and we set $\alpha = -2.74$ and -2.538 for binary G and normal G respectively to achieve a disease rate of 10%. In the case of $X = \mathcal{G}(20,\ 1)$, we set $\alpha = -5.220$ and -5.086 for binary G and normal G respectively to achieve a disease rate of 4.5%, and we set $\alpha = -4.352$ and -4.158 for binary G and normal G respectively to achieve a disease rate of 10%. For each setting, we simulated 1000 data sets, each with $n_1 = 1000$ cases and $n_0 = 1000$ controls. The details of simulating the case-control data are provided in the Online Supplement. In the computation of the weighted nonparametric density/mass function estimates defined in Algorithm 1, we used the asymptotically justified bandwidth $h = cn^{-1/5}$, where $c \in [0.4,\ 1.2]$, and the results were insensitive to the choice of c.

The results are summarized in Tables 1–4. For 4.5% disease prevalence and normally distributed X (Table 1), it is clear that prospective logistic regression and our semiparametric

efficient estimator are both consistent, while the semiparametric estimator has smaller variance. Specifically, the semiparametric efficient estimator has a mean squared error efficiency gain as large as 57% (the interaction term between G and X) for binary G, and 46% (the interaction term between G and X) for normal G. For 4.5% disease prevalence and Gamma X (Table 3), when G follows a Bernoulli distribution, our semiparametric efficient estimator has a mean squared error efficiency gain between 31% (the main effect of X) and 56% (the interaction term between G and X); when G is normal, the corresponding efficiency gain of the interaction term is 44%.

The results for the 10% disease rate case (Tables 2 and 4) are similar. Both approaches are asymptotically valid, with our approach being superior to prospective logistic regression in the sense that our semiparametric efficient estimator has smaller mean squared error.

6. Example

Prostate cancer is a heterogeneous disease resulting from the complex interplay of genetic susceptibility and environmental exposures. It is the second leading cause of cancer death among men in the USA [1]. Prostate cells (both primary and cancer cells) were demonstrated to have 1α-OHase activity, whereas 1α-OHase is the enzyme responsible for converting [25(OH)D], the major circulating form of vitamin D that reflects both dietary and sunlight exposures, into 1,25-dihydroxy-vitamin D [1,25(OH)2D], the most active form of this vitamin that can induce cell-cycle regulation, apoptosis and differentiation in prostate cancer cells via the vitamin D receptor (VDR). Thus, (a) [25(OH)D] is hypothesized to have an anticancer effect, and (b) an important question is whether its relationship with the risk of developing prostate cancer is modified by genetic polymorphisms in the VDR gene.

In this section, we implemented our methodology in a case-control study of prostate cancer, using the same data set analyzed but in a different context by Chen et al. [9], see that reference for details about the study. Specifically, our analysis is based on a polygenic risk score, a single risk factor incorporating information from susceptibility SNPs, whereas Chen et al. [9] focused on haplotypes. The data consist of $n_1 = 690$ cases and $n_0 = 717$ controls randomly selected from the screening arm of a large population-based cohort study, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) at the National Cancer Institute. The PLCO cohort study recruited a total of 76,685 men aged 55–74 at 10 screening centers between November 1993 and July 2001, then randomly assigned 38,340 of them to the screening arm and the rest to the non-screening arm. In a 10-year follow-up period, in the study population, the cumulative incidence rate for prostate cancer in the screening arm was 108.4 per 10,000 person-years [3]. Apart from case-control status, [25(OH)D] level (nmol/L) and genotype data on 19 single-nucleotide polymorphisms (SNPS) are available for each subject involved in the case-control study. According to Chen et al. [9], these polymorphisms, our G, are unlikely to affect the [25(OH)D] level, our X, as the VDR gene plays a “downstream” role in the vitamin-D pathway. In other words, the gene–environment independence assumption is likely to be valid in this application. Detailed information about the design can be found in Andriole et al. [3],Hayes et al. [16],Prorok et al. [25].

One difficulty in investigating the genetic modification of the VDR gene to [25(OH)D] on the risk of prostate cancer is that the VDR gene contains multiple underlying susceptibility SNPs, where each individual SNP may only confer a small component of overall risk. In fact, running a logistic regression of case-control status on each of the 19 SNPs shows only three SNPs have p-values ≤ 0.10. Recently, it has been recognized that the polygenic risk score has the potential of improving risk prediction for some common diseases [2,7,11–13,26]. Therefore, we created a polygenic risk score for the prostate cancer data by weighting those 19 SNPs, where the weights are the effect sizes of separate logistic regressions applied to each SNP.

The results of prospective logistic regression and our semiparametric approach based on 1000 bootstrap samples are given in Table 5. The two sets of estimates are fairly consistent as expected. However, our semiparametric efficient estimator has smaller standard errors than does the prospective logistic regression, in accordance with theory and our simulations. This leads to a substantial difference in inference for the interaction between the polygenic risk score and the [25(OH)D] level. Specifically, both prospective logistic regression and our semiparametric efficient method show that the main effects of both the polygenic risk score and the [25(OH)D] level are statistically significant and positive. That is, if ignoring the interaction, men with higher polygenic risk scores or/and higher [25(OH)D] levels tend to have higher risk of developing prostate cancer.

Importantly, the estimates of the interaction parameter from the prospective logistic regression is not significant at the 5% level. However, our approach shows significant evidence of interaction, i.e., the effects of [25(OH)D] level on prostate cancer risk differ depending on the polygenic risk score.

In addition, our approach provides an estimated disease rate in the population of 10.6%, whereas the disease rate in the PLCO cohort study is 10.8% per person-year. This validation of our methodology suggests an additional use to which it can be applied.

7. Discussion

We have developed a semiparametric efficient estimator in case-control studies for the gene–environment independent model, where the distributions of genetic susceptibility and environmental exposure are allowed to be arbitrary and the disease rate is assumed completely unknown. We showed that in spite of these weak assumptions, the problem is identifiable in most cases. The proposed estimator is derived under the so-called hypothetical population framework, which enables us to view the case-control sample as a random sample from a hypothetical distribution and thus facilitates the application of a conventional semiparametric approach. Such an estimator is semiparametric efficient and its superiority over the prospective logistic regression was demonstrated in various simulations. The general methodology of our approach can be extended to parametric models other than the logistic model, such as the probit model, and it can be used to consider assumptions other than gene–environment independence, such as Hardy–Weinberg equilibrium, as long as the resulting model is identifiable.

The method hinges on the assumption of gene–environment independence. When they are in fact dependent, blindly applying this method will not lead to a consistent estimator. It is possible to further apply the empirical Bayes shrinkage method of [9] to improve robustness to the model assumptions. This method effectively uses our method when the assumption holds, and effectively uses logistic regression when the model assumption fails.

To handle the nuisance parameters in the estimation procedure, nonparametric density/mass function estimation is used. When the dimensions of genetic susceptibility or environmental exposures increase, such nonparametric estimation suffers from the curse of dimensionality. In such cases, dimension reduction techniques might be needed to maintain model flexibility as well as ensure computation feasibility. This will be pursued in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Ma's research was partially supported by the National Science Foundation, USA (DMS-1608540). Carroll and Liang's research was supported by a grant from the National Cancer Institute, USA (U01-CA057030). We thank Nilanjan Chatterjee and Alex Asher for many helpful comments.

Appendix. Sketch of technical arguments

A.1. Identifiability

A1 There exists c_x so that when $x \to c_x$, $m(g, x, \boldsymbol{\beta}) \to \infty$ or $m(g, x, \boldsymbol{\beta}) \to -\infty$ for any g.

A2 There exists g_1 and x_1, x_2 such that $m(g_1, x_1, \beta) \neq m(g_1, x_2, \beta)$.

A3 There exists c_g so that when $g \to c_g$, $m(g, x, \boldsymbol{\beta}) \to \infty$ or $m(g, x, \boldsymbol{\beta}) \to -\infty$ for any x.

A4 There exists x_1 and g_1, g_2 such that $m(g_1, x_1, \beta) \neq m(g_2, x_1, \beta)$.

Proposition 1. The problem stated in (2) is identifiable

- **i.** (i) If condition A1 holds, and at least one of the conditions A3 and A4 holds;
- **ii.** (ii) or if at least one of the conditions A1 and A2 holds, and condition A3 holds.

Remark 1. In practice, a widely used model is the one including main effects and two-way interaction, i.e., $\alpha + \beta_1 g + \beta_2 x + \beta_3 xg$. It can be easily verified that if g and x both have the support on $\mathbb{R}$ then this model satisfies conditions A1 and A3 described above and hence is identifiable.

Remark 2. Proposition 1 applies in the case where at most one of G and X is discrete. In the case where both G and X are discrete with levels ℓ_G and ℓ_X respectively, identifiability requires $\ell_G \ell_X \geq 2\ell_G + 2\ell_X - 2$ as a necessary condition. Additional conditions may be

needed. Although for a specific model with known ℓ_G and ℓ_X, it can be easy to derive the sufficient conditions for identifiability, such result is difficult to describe in general.

Proof of Proposition 1. From [24], $\boldsymbol{\beta}$ is identifiable. Thus, we aim at establishing the identifiability of η_1, η_2 and α. We first prove the result under A1 and A3. Assume there are α, η_1, η_2 and α^*, η_1^*, η_2^* so that

$$\frac{n_d}{n\pi_d}\eta_1(g)\eta_2(x)H(d,\, g,\, x,\, \boldsymbol{\beta},\, \alpha) = \frac{n_d}{n\pi_d^*}\eta_1^*(g)\eta_2^*(x)H\left(d,\, g,\, x,\, \boldsymbol{\beta},\, \alpha^*\right).$$

This yields

$$\frac{1}{\pi_1}\eta_1(g)\eta_2(x)H(1,\, g,\, x,\, \boldsymbol{\beta},\, \alpha) = \frac{1}{\pi_1^*}\eta_1^*(g)\eta_2^*(x)H\left(1,\, g,\, x,\, \boldsymbol{\beta},\, \alpha^*\right),$$

$$\frac{1}{\pi_0}\eta_1(g)\eta_2(x)H(0,\, g,\, x,\, \boldsymbol{\beta},\, \alpha) = \frac{1}{\pi_0^*}\eta_1^*(g)\eta_2^*(x)H\left(0,\, g,\, x,\, \boldsymbol{\beta},\, \alpha^*\right).$$

Taking the ratio of the above two and solving, we obtain $\exp\left(\alpha^*\right) = \exp(\alpha)\pi_0\pi_1^*/\left(\pi_1\pi_0^*\right)$. This leads to

$$\frac{\eta_2^*(x)\,\eta_1^*(g)}{\eta_2(x)\,\eta_1(g)} = \frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g,\, x,\, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g,\, x,\, \boldsymbol{\beta})\}}.$$

Under condition A1, letting $x \to c_x$, we obtain $\eta_1^*(g) = \eta_1(g)$. Similarly, under condition A3, letting $g \to c_g$, we obtain $\eta_2^*(x) = \eta_2(x)$. This in turn leads to $\pi_0^* = \pi_0, \pi_1^* = \pi_1$. Finally, these results lead to $\alpha^* = \alpha$.

We now prove the result under A1 and A4. Under condition A1 alone, the same derivation as before leads to

$$\frac{\eta_2^*(x)}{\eta_2(x)} = \frac{\pi_0^*/\pi_0 + \exp\{\alpha + m(g,\, x,\, \boldsymbol{\beta})\}\pi_1^*/\pi_1}{1 + \exp\{\alpha + m(g,\, x,\, \boldsymbol{\beta})\}}.$$

Thus A4 further implies

$$\frac{\pi_0^*/\pi_0 + \exp\left\{\alpha + m\left(g_1,\, x_1,\, \boldsymbol{\beta}\right)\right\}\pi_1^*/\pi_1}{1 + \exp\left\{\alpha + m\left(g_1,\, x_1,\, \boldsymbol{\beta}\right)\right\}} = \frac{\pi_0^*/\pi_0 + \exp\left\{\alpha + m\left(g_2,\, x_1,\, \boldsymbol{\beta}\right)\right\}\pi_1^*/\pi_1}{1 + \exp\left\{\alpha + m\left(g_2,\, x_1,\, \boldsymbol{\beta}\right)\right\}},$$

or equivalently, $\left(\pi_0^*/\pi_0 - \pi_1^*/\pi_1\right)\left[\exp\left\{\alpha + m\left(g_1,\, x_1,\, \boldsymbol{\beta}\right)\right\} - \exp\left\{\alpha + m\left(g_2,\, x_1,\, \boldsymbol{\beta}\right)\right\}\right] = 0$. Hence, $\pi_d^* = \pi_d$ for d = 0, 1. As a result, $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}$ and $\alpha^* = \alpha$ and $\eta_2^*(x) = \eta_2(x)$.

The result under A2 and A3 is symmetric to the one under A1 and A4 hence is omitted. □

The requirements in A1 and A3 are appropriate in the case where G and X are both continuous. The requirements in A1 and A4 are suitable in the case where G is discrete and X is continuous. The requirements in A2 and A3 are suitable in the case where X is discrete and G is continuous.

A.2. Nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$

The nuisance tangent space Λ is computed in two steps. First, replacing the nuisance parameter $\eta = \left(\eta_1, \eta_2\right)$ with a finite-dimensional parameter, say $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{\top},\, \boldsymbol{\gamma}_2^{\top}\right)^{\top}$, and taking the derivative of $\ln f_{D,G,X}(d,\, g,\, x;\, \boldsymbol{\beta},\, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ to get $\mathbf{S}_{\boldsymbol{\gamma}} = \left(\mathbf{S}_{\boldsymbol{\gamma}1}^{\top},\, \mathbf{S}_{\boldsymbol{\gamma}2}^{\top}\right)^{\top}$. Second, finding the mean squared closure that contains all such $\mathbf{S}\boldsymbol{\gamma}$, which is Λ.

For any finite-dimensional parameter $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_1^{\top},\, \boldsymbol{\gamma}_2^{\top}\right)^{\top}$, we have $\mathbf{S}_{\boldsymbol{\gamma}} = \left(\mathbf{S}_{\boldsymbol{\gamma}1}^{\top},\, \mathbf{S}_{\boldsymbol{\gamma}2}^{\top}\right)^{\top}$, where

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\gamma}1} &= \eta_1\left(g,\, \boldsymbol{\gamma}_1\right)^{-1} \partial\eta_1\left(g,\, \boldsymbol{\gamma}_1\right)/\partial\boldsymbol{\gamma}_1 - \pi_d^{-1}\int \partial\eta_1\left(g,\, \boldsymbol{\gamma}_1\right)/\partial\boldsymbol{\gamma}_1\eta_2(x)H(d,\, g,\, x,\, \boldsymbol{\theta})d\mu(x)d\mu(g)\\ &= \eta_1\left(g,\, \boldsymbol{\gamma}_1\right)^{-1} \partial\eta_1\left(g,\, \boldsymbol{\gamma}_1\right)/\partial\boldsymbol{\gamma}_1 - \mathrm{E}\left\{\eta_1\left(g,\, \boldsymbol{\gamma}_1\right)^{-1} \partial\eta_1\left(G,\, \boldsymbol{\gamma}_1\right)/\partial\boldsymbol{\gamma}_1 \middle| D\right\},\end{aligned}$$

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\gamma}2} &= \eta_2\left(x,\, \boldsymbol{\gamma}_2\right)^{-1} \partial\eta_2\left(x,\, \boldsymbol{\gamma}_2\right)/\partial\boldsymbol{\gamma}_2 - \pi_d^{-1}\int \eta_1(g)\partial\eta_2\left(x,\, \boldsymbol{\gamma}_2\right)/\partial\boldsymbol{\gamma}_2 H(d,\, g,\, x,\, \boldsymbol{\theta})d\mu(x)d\mu(g)\\ &= \eta_2\left(x,\, \boldsymbol{\gamma}_2\right)^{-1} \partial\eta_2\left(x,\, \boldsymbol{\gamma}_2\right)/\partial\boldsymbol{\gamma}_2 - \mathrm{E}\left\{\eta_2\left(x,\, \boldsymbol{\gamma}_2\right)^{-1} \partial\eta_2\left(X,\, \boldsymbol{\gamma}_2\right)/\partial\boldsymbol{\gamma}_2 \middle| D\right\}.\end{aligned}$$

It is easy to show the nuisance tangent spaces associated with η_1 and η_2 are respectively

$$\begin{aligned}\Lambda_1 &= \left\{\mathbf{a}(g) - \pi_d^{-1}\int \mathbf{a}(g)\eta_1(g)\eta_2(x)H(d,\, g,\, x,\, \boldsymbol{\theta})d\mu(x)d\mu(g) : \mathrm{E}^{\mathrm{true}}\{\mathbf{a}(G)\} = 0\right\}\\ &= \{\mathbf{a}(g) - \mathrm{E}\{\mathbf{a}(G)|d\} \text{for all } \mathbf{a}(g)\},\end{aligned}$$

$$\begin{aligned}\Lambda_2 &= \left\{\mathbf{a}(x) - \pi_d^{-1}\int \mathbf{a}(x)\eta_1(g)\eta_2(x)H(d,\, g,\, x,\, \boldsymbol{\theta})d\mu(x)d\mu(g) : \mathrm{E}^{\mathrm{true}}\{\mathbf{a}(X)\} = 0\right\}\\ &= \{\mathbf{a}(x) - \mathrm{E}\{\mathbf{a}(X)|d\} \text{for all } \mathbf{a}(x)\}.\end{aligned}$$

Then

$$\Lambda = \Lambda_1 + \Lambda_2 = \left\{\mathbf{a}_1(g) + \mathbf{a}_2(x) - \mathrm{E}\left\{\mathbf{a}_1(G) + \mathbf{a}_2(X)|d\right\} \text{for all } \mathbf{a}_1(g), \mathbf{a}_2(x)\right\}.$$

Define $\Lambda_1^{\perp,\,\mathrm{conj}} = \{\mathbf{f}(d,\,g,\,x):\mathrm{E}(\mathbf{f}) = 0,\,\mathrm{E}(\mathbf{f}|G) = \mathrm{E}\{\mathrm{E}(\mathbf{f}|D)|G\}\}$. Now consider $\mathbf{f} \perp \Lambda_1$. Then for any $\mathbf{a}(g) - \mathrm{E}\{\mathbf{a}(G)|d\} \in \Lambda_1$,

$$\begin{aligned} 0 = \mathrm{E}\left(\mathbf{f}^\top[\mathbf{a}(G) - \mathrm{E}\{\mathbf{a}(G)|D\}]\right) &= \mathrm{E}\left[\mathbf{f}^\top\mathbf{a}(G) - \mathbf{f}^\top\mathrm{E}\{\mathbf{a}(G)|D\}\right] \\ &= \mathrm{E}\left[\mathbf{f}^\top\mathbf{a}(G) - \mathrm{E}\left(\mathbf{f}^\top\middle|D\right)\mathrm{E}\{\mathbf{a}(G)|D\}\right] \\ &= \mathrm{E}\left\{\mathbf{f}^\top\mathbf{a}(G) - \mathrm{E}\left(\mathbf{f}^\top\middle|D\right)\mathbf{a}(G)\right\} = \mathrm{E}\left[\mathrm{E}\left\{\mathbf{f}^\top - \mathrm{E}\left(\mathbf{f}^\top\middle|D\right)\middle|G\right\}\mathbf{a}(G)\right]. \end{aligned}$$

Hence, $\mathrm{E}\{f - \mathrm{E}(f|D)|G\} = \mathbf{0}$ almost surely. Besides, $\Lambda_1^\perp$ need to be a subspace of the Hilbert space $\mathscr{H}$, hence $\mathrm{E}(f) = 0$. Thus, we have shown $\Lambda_1^\perp \subset \Lambda_1^{\perp,\,\mathrm{conj}}$. Furthermore, for any $\mathbf{f} \in \Lambda_1^{\perp,\,\mathrm{conj}}$,

$$\mathrm{E}\left[\mathbf{f}^\top\mathbf{a}(G) - \mathbf{f}^\top\mathrm{E}\{\mathbf{a}(G)|D\}\right] = \mathrm{E}\left\{\mathbf{f}^\top\mathbf{a}(G) - \mathrm{E}\left(\mathbf{f}^\top|D\right)\mathbf{a}(G)\right\} = \mathrm{E}\left[\mathrm{E}\left\{\mathbf{f}^\top - \mathrm{E}\left(\mathbf{f}^\top|D\right)|G\right\}\mathbf{a}(G)\right] = \mathbf{0},$$

hence $\Lambda_1^{\perp,\,\mathrm{conj}} \subset \Lambda_1^\perp$. Thus, we have obtained $\Lambda_1^\perp = \Lambda_1^{\perp,\,\mathrm{conj}}$. Similarly, we can prove

$$\Lambda_2^\perp = \{\mathbf{f}(d,\,g,\,x):\mathrm{E}(\mathbf{f}) = \mathbf{0},\,\mathrm{E}(\mathbf{f}|X) = \mathrm{E}\{\mathrm{E}(\mathbf{f}|D)|X\}\}$$

Hence,

$$\Lambda^\perp = \{\mathbf{f}(d,\,g,\,x):\mathrm{E}(\mathbf{f}|G) = \mathrm{E}\{\mathrm{E}(\mathbf{f}|D)|G\},\mathrm{E}(\mathbf{f}|X) = \mathrm{E}\{\mathrm{E}(\mathbf{f}|D)|X\},\mathrm{E}(\mathbf{f}) = \mathbf{0}\}.$$

A.3. Uniqueness of a and b up to constants

To prove that **a** and **b** defined in Eqs. (4) and (5) are unique up to constant shifts, we consider the following. If there exist $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$ such that

$$\begin{aligned} \mathbf{S}_{\mathrm{eff}}(d,\,g,\,x) &= \mathbf{S}(d,\,g,\,x) - \mathbf{a}_1(g) - \mathbf{b}_1(x) - \mathrm{E}\{\mathbf{S}(d,\,G,\,X)|d\} + \mathrm{E}\left\{\mathbf{a}_1(G) + \mathbf{b}_1(X)\middle|d\right\} \\ &= \mathbf{S}(d,\,g,\,x) - \mathbf{a}_2(g) - \mathbf{b}_2(x) - \mathrm{E}\{\mathbf{S}(d,\,G,\,X)|d\} + \mathrm{E}\left\{\mathbf{a}_2(G) + \mathbf{b}_2(X)\middle|d\right\}, \end{aligned}$$

then

$$\mathbf{a}_2(g) - \mathbf{a}_1(g) = \mathbf{b}_1(x) - \mathbf{b}_2(x) - \mathrm{E}\left\{\mathbf{a}_1(G) + \mathbf{b}_1(X)\middle|d\right\} + \mathrm{E}\left\{\mathbf{a}_2(G) + \mathbf{b}_2(X)\middle|d\right\}.$$

The left-hand side is a function of g while the right-hand side is a function of x and d. Hence $\mathbf{a}_1(g) - \mathbf{a}_2(g)$ is a constant. Similarly, $\mathbf{b}_1(x) - \mathbf{b}_2(x)$ is also a constant. □

A.4. Equivalent expression of Eqs. (4) and (5) *and the proof under the condition* $\mathrm{E}(a) = \mathrm{E}(b) = 0$

We claim under the mean zero constraint $\mathrm{E}(\mathbf{a}) = \mathrm{E}(\mathbf{b}) = \mathbf{0}$, (4) and (5) are equivalent to (A.1)–(A.3), below, namely

$$\mathbf{S}_g(g) - \mathrm{E}\{\mathbf{S}_x(X)|g\} = \mathbf{a}(g) + \mathbf{u}_0 c_g(g) - \mathrm{E}\{\mathrm{E}(\mathbf{a}|X)|g\} - \mathbf{u}_0 \mathrm{E}\{c_x(X)|g\}, \tag{A.1}$$

$$\mathbf{S}_x(x) = \mathrm{E}(\mathbf{a}|x) + \mathbf{b}(x) + \mathbf{u}_0 c_x(x), \tag{A.2}$$

$$\mathbf{u}_0 = \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0), \tag{A.3}$$

where $c_x(x) = \mathrm{E}[\{n_0 - nI(D = 0)\}/n_1|x]$, $c_g(g) = \mathrm{E}[\{n_0 - nI(D = 0)\}/n_1|g]$.

Proof. Suppose $\mathbf{a}$ and $\mathbf{b}$ are the solution of Eqs. (4) and (5). Let $\mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0) = \mathbf{u}_0$, $\mathrm{E}(\mathbf{a} + \mathbf{b}|D = 1) = \mathbf{u}_1$. Then (A.3) automatically holds. It is easy to verify that $\mathbf{u}_0 n_0 + \mathbf{u}_1 n_1 = n\mathrm{E}(\mathbf{a} + \mathbf{b}) = 0$. Hence (4) and (5) become

$$\mathrm{E}(\mathbf{a}|x) + \mathbf{b}(x) + \mathbf{u}_0\{(n_0/n_1)f_{D|X}(1, x) - f_{D|X}(0, x)\} = \mathbf{S}_x(x),$$

$$\mathbf{a}(g) + \mathrm{E}(\mathbf{b}|g) + \mathbf{u}_0\{(n_0/n_1)f_{D|G}(1, g) - f_{D|G}(0, g)\} = \mathbf{S}_g(g).$$

Further write

$$\begin{aligned} c_x(x) &= (n_0/n_1)f_{D|X}(1, x) - f_{D|X}(0, x) = \{n_0 - nf_{D|X}(0, x)\}/n_1 \\ &= \mathrm{E}[\{n_0 - nI(D = 0)\}/n_1|x] = \mathrm{E}[\{n_0/n - I(D = 0)\}/(n_1/n)|x], \end{aligned}$$

$$\begin{aligned} c_g(g) &= (n_0/n_1)f_{D|G}(1, g) - f_{D|G}(0, g) = \{n_0 - nf_{D|G}(0, g)\}/n_1 \\ &= \mathrm{E}[\{n_0 - nI(D = 0)\}/n_1|g] = \mathrm{E}[\{n_0/n - I(D = 0)\}/(n_1/n)|g]. \end{aligned}$$

Then

$$\mathrm{E}(\mathbf{a}|x) + \mathbf{b}(x) + \mathbf{u}_0 c_x(x) = \mathbf{S}_x(x), \tag{A.4}$$

$$\mathbf{a}(g) + \mathrm{E}(\mathbf{b}|g) + \mathbf{u}_0 c_g(g) = \mathbf{S}_g(g). \tag{A.5}$$

Note that (A.4) above is exactly (A.2) defined in Section 3. Taking conditional expectation of (A.4) given $G = g$, we obtain

$$\mathrm{E}\{\mathrm{E}(\mathbf{a}|X)|g\} + \mathrm{E}(\mathbf{b}|g) + \mathbf{u}_0\mathrm{E}\left\{c_{\mathcal{X}}(X)\middle|g\right\} = \mathrm{E}\left\{\mathbf{S}_{\mathcal{X}}(X)\middle|g\right\}.$$

Subtracting the above from (A.5), we obtain (A.1), namely

$$\mathbf{a}(g) + \mathbf{u}_0 c_g(g) - \mathrm{E}\{\mathrm{E}(\mathbf{a}|X)|g\} - \mathbf{u}_0\mathrm{E}\left\{c_{\mathcal{X}}(X)\middle|g\right\} = \mathbf{S}_g(g) - \mathrm{E}\left\{\mathbf{S}_{\mathcal{X}}(X)\middle|g\right\}.$$

From the above derivation, it is clear that any mean zero functions $\mathbf{a}(g)$, $\mathbf{b}(x)$ that solve (4) and (5) also satisfy (A.1)–(A.3). We now prove the other way around, that is any mean zero functions $\mathbf{a}(g)$, $\mathbf{b}(x)$ that satisfy (A.1)–(A.3) also satisfy (4) and (5).

Taking the expectation of (A.2) conditionally on $G = g$ and adding the resulting equation to (A.1), we obtain exactly (A.5).

Hence Eqs. (A.1) and (A.2) lead to Eqs. (A.2) and (A.5).

For preparation, note also that $c_g(g) = \left(n_0/n_1\right)f_{D|G}(1, g) - f_{D|G}(0, g)$. Hence under (A.3) and the condition $n_1\mathrm{E}(\mathbf{a} + \mathbf{b}|D = 1) + n_0\mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0) = n\mathrm{E}(\mathbf{a} + \mathbf{b}) = 0$, we can further write

$$\begin{aligned}
\mathbf{u}_0 c_g(g) &= \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0)\left\{\left(n_0/n_1\right)f_{D|G}(1, g) - f_{D|G}(0, g)\right\} \\
&= \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0)\left(n_0/n_1\right)f_{D|G}(1, g) - \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0)f_{D|G}(0, g) \\
&= -\mathrm{E}(\mathbf{a} + \mathbf{b}|D = 1)f_{D|G}(1, g) - \mathrm{E}(\mathbf{a} + \mathbf{b}|D = 0)f_{D|G}(0, g) \\
&= -\mathrm{E}\{\mathrm{E}(\mathbf{a} + \mathbf{b}|D|g\}.
\end{aligned}$$

Similarly, $\mathbf{u}_0 c_{\mathcal{X}}(x) = -\mathrm{E}\{\mathrm{E}(\mathbf{a} + \mathbf{b}|D)|x\}$. From (A.2), we obtain

$$\mathbf{S}_{\mathcal{X}}(x) = \mathrm{E}(\mathbf{a}|x) + \mathbf{b}(x) + \mathbf{u}_0 c_{\mathcal{X}}(x) = \mathrm{E}(\mathbf{a}|x) + \mathbf{b}(x) - \mathrm{E}\{\mathrm{E}(\mathbf{a} + \mathbf{b}|D)|x\},$$

which is exactly (4). Similarly, from (A.5), we obtain (5). □

Eq. (A.1) allows us to solve for $\mathbf{a}(g)$ as a function of $\mathbf{u}_0$ and other known quantities, say $\mathbf{a}(g) = \mathbf{F}_a(g, \mathbf{u}_0) - \mathrm{E}\left\{\mathbf{F}_a(G, \mathbf{u}_0)\right\}$, where $\mathbf{F}_a$ is a function that solves (A.1) which does not need to have mean $\mathbf{0}$. Then we can solve $\mathbf{b}$ from (A.2) as a function of $\mathbf{u}_0$ to obtain

$$\mathbf{b}(x) = \mathbf{S}_{\mathcal{X}}(x) - \mathbf{u}_0 c_{\mathcal{X}}(x) - \mathrm{E}\left\{\mathbf{F}_a(G, \mathbf{u}_0)\middle|x\right\} + \mathrm{E}\left\{\mathbf{F}_a(G, \mathbf{u}_0)\right\}.$$

Now

$$\begin{aligned}
\mathbf{u}_0 &= \mathrm{E}\{\mathbf{a}(G) + \mathbf{b}(X)|D = 0\} \\
&= \mathrm{E}\left[\mathbf{F}_a(G, \mathbf{u}_0) + \mathbf{S}_{\mathcal{X}}(X) - \mathbf{u}_0 c_{\mathcal{X}}(X) - \mathrm{E}\left\{\mathbf{F}_a(G, \mathbf{u}_0)\middle|X\right\}\middle|D = 0\right],
\end{aligned}$$

which allows us to solve for $\mathbf{u}_0$. Having obtained $\mathbf{u}_0$, we can then solve for all other quantities easily. Unfortunately, the integral equation (A.1) does not have an explicit solution. We propose an approximation to its solution in the spirit of Tsiatis and Ma [28], which is provided in Appendix A.5, by discretizing X if X is continuous.

The efficient score $\mathbf{S}_{\text{eff}}$, especially the procedure of solving for **a** and **b**, contains several expectations conditional on D, G, or X. To get estimations of these conditional expectations, we need density estimators of the nuisance parameter $\eta = (\eta_1, \eta_2)$.

If the disease rate π_1 or the non-disease rate $\pi_0 = 1 - \pi_1$ is known, then η can be approximated by

$$\hat{\eta}_1 = \pi_0 \hat{f}_{G|D=0} + (1-\pi_0)\hat{f}_{G|D=1}, \quad \hat{\eta}_2 = \pi_0 \hat{f}_{X|D=0} + (1-\pi_0)\hat{f}_{X|D=1},$$

where $\hat{f}_{G|D=d}$ and $\hat{f}_{X|D=d}$ are the nonparametric estimators of the conditional density/mass function $f_{G|D=d}$ and $f_{X|D=d}$ respectively for $d \in \{0, 1\}$. Of course, in practice, π_0 is typically unknown. However, we can get an estimate of π_0 through (3).

A.5. Solving the integral equation (A.1)

Define $\mathbf{Z} = \mathbf{S} - \mathrm{E}(\mathbf{S}|D) - \mathbf{u}_0\{n_0 - n\mathbf{1}(D=0)\}/n_1$. An equivalent expression of (A.1) is

$$\mathbf{a}(G) - \mathrm{E}[\mathrm{E}\{\mathbf{a}(G)|X\}|G] = E(\mathbf{Z}|G) - \mathrm{E}\{\mathrm{E}(\mathbf{Z}|X)|G\}. \tag{A.6}$$

For fixed $\mathbf{u}_0$, all the quantities in $\mathbf{Z}$ are known or have explicit form except $\mathrm{E}(\mathbf{S} \mid D)$. With the weighted kernel density $\hat{\eta}_1, \hat{\eta}_2$, estimated non-disease rate $\hat{\pi}_0$ and disease rate $\hat{\pi}_1$, we can estimate it by

$$\hat{E}(\mathbf{S}|D=d) = \hat{\pi}_d^{-1} \int \mathbf{S}(d, g, x)\hat{\eta}_1(g)\,\hat{\eta}_2(x)d\mu(g)d\mu(x).$$

A.5.1. Discrete G with finite number of levels

Assume G is discrete with mass at m_g points $g_1, \ldots, g_{mg}$. We computed each term in (A.6) under the weighted nonparametric densities $\hat{\eta}_1, \hat{\eta}_2$

$$\hat{E}\{\mathbf{a}(G)|x\} = \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)k(g_j, x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j, x)\hat{\eta}_1(g_j)},$$

$$\hat{E}\left[\hat{E}\{\mathbf{a}(G)|X\}g_k\right] = \int \left\{ \frac{\sum_{j=1}^{m_g} \mathbf{a}(g_j)k(g_j, x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j, x)\hat{\eta}_1(g_j)} \right\} \frac{k(g_k, x)\hat{\eta}_2(x)}{\int k(g_k, x)\hat{\eta}_2(x)d\mu(x)} d\mu(x).$$

Similarly, we have

$$\hat{E}\{\mathbf{Z}(D,G,X)|X\} = \frac{\sum_{j=1}^{m_g}\sum_{d=0}^{1} n_d/(n\pi_d)\mathbf{Z}(d,g_j,x)H(d,g_j,x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j,x)\hat{\eta}_1(g_j)},$$

$$\hat{E}\left[\hat{E}\{\mathbf{Z}(D,G,X)|X\}\,\middle|\,g_k\right] = \int \left\{\frac{\sum_{j=1}^{m_g}\sum_{d=0}^{1} n_d/(n\pi_d)\mathbf{Z}(d,g_j,x)H(d,g_j,x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j,x)\hat{\eta}_1(g_j)}\right\} \times \frac{k(g_k,x)\hat{\eta}_2(x)}{\int k(g_k,x)\hat{\eta}_2 d\mu(x)} d\mu(x), \tag{A.7}$$

and

$$\hat{E}\{\mathbf{Z}(D,G,X)|g_k\} = \sum_{d=0}^{1}\int \mathbf{Z}(d,g_k,x)\frac{n_d/(n\pi_d)H(d,g_k,x)\hat{\eta}_2(x)}{\int k(g_k,x)\hat{\eta}_2(x)d\mu(x)}d\mu(x). \tag{A.8}$$

Consequently, the integral equation (A.6) reduces to the linear equations $(I - B)A^{\top} = C^{\top}$, where A is the $(p+1) \times m_g$ matrix $\mathbf{a}(g_1), \ldots, \mathbf{a}(g_{mg})$, corresponding to the solution of the integral equation, I is an $m_g\, m_g$ identity matrix, B is an $m_g\, m_g$ matrix whose (i, j)th element is given by

$$B_{ij} = \int \left\{\frac{k(g_j,x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j,x)\hat{\eta}_1(g_j)}\right\}\frac{k(g_i,x)\hat{\eta}_2(x)}{\int k(g_i,x)\hat{\eta}_2(x)d\mu(x)}d\mu(x),$$

and C is a $(p+1) \times m_g$ matrix whose kth column is $\hat{E}\{\mathbf{Z}(D,G,X)|g_k\} - \hat{E}\left[\hat{E}\{\mathbf{Z}(D,G,X)|X\}|g_k\right]$ defined in (A.7) and (A.8). After obtaining $\mathbf{a}$, we set

$$\mathbf{b}(x) = \hat{E}(\mathbf{Z} - \mathbf{a}|x)$$

$$= \frac{\sum_{j=1}^{m_g}\sum_{d=0}^{1} n_d/(n\pi_d)\mathbf{Z}(d,g_j,x)H(d,g_j,x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j,x)\hat{\eta}_1(g_j)} - \frac{\sum_{j=1}^{m_g}\mathbf{a}(g_j)k(g_j,x)\hat{\eta}_1(g_j)}{\sum_{j=1}^{m_g} k(g_j,x)\hat{\eta}_1(g_j)}.$$

Then we compute $\mathbf{u}_0 = \hat{E}(\mathbf{a} + \mathbf{b}|D = 0)$, where

$$\hat{E}(\mathbf{a}|D=0)=\frac{\sum_{j=1}^{m_g}\mathbf{a}(g_j)\hat{\eta}_1(g_j)\int H(0,g_j,x)\hat{\eta}_2(x)d\mu(x)}{\int\sum_{j=1}^{m_g}H(0,g_j,x)\hat{\eta}_1(g_j)\hat{\eta}_2(x)d\mu(x)},$$

$$\hat{E}(\mathbf{b}|D=0)=\int\mathbf{b}(x)\frac{\sum_{j=1}^{m_g}H(0,g_j,x)\hat{\eta}_1(g_j)\hat{\eta}_2(x)}{\int\sum_{j=1}^{m_g}H(0,g_j,x)\hat{\eta}_1(g_j)\hat{\eta}_2(x)d\mu(x)}d\mu(x).$$

A.5.1. Continuous G or discrete G with infinite number of levels

When G is a continuous variable, we discretize it at a finite number of equally distributed points, say, $g_1 \le \cdots \le gm_g$ with $g_{i+1}-g_i \equiv \Delta_g$ for all $i \in \{1, \ldots, m_g - 1\}$, such that

$$\sum_{i=1}^{m_g} f_{G|D}(g_i)\Delta_g \approx 1.$$

Similarly, when G is discrete with infinite number of levels, we simply choose a sufficient number of points from its support to get an overall probability close to 1. Then the sequential procedures are exactly the same as that described in the case where G is discrete with finite number of levels.

References

[1]. American Cancer Society, Cancer Facts & Figures 2015, American Cancer Society, Atlanta, GA, 2015.

[2]. Aly M, Wiklund F, Xu J, Isaacs WB, Eklund M, D'Amato M, Adolfsson J, Grönberg H, Polygenic risk score improves prostate cancer risk prediction: Results from the Stockholm-1 cohort study, Eur. Urol 60 (2011) 21–28. [PubMed: 21295399]

[3]. Andriole GL, Crawford ED, Grubb RL, Buys SS, Chia D, Church TR, Fouad MN, Isaacs C, Kvale PA, Reding DJ, Weissfeld JL, Yokochi LA, O'Brien B, Ragard LR, Clapp JD, Rathmell JM, Riley TL, Hsing AW, Izmirlian G, Pinsky PF, Kramer BS, Miller AB, Gohagan JK, Prorok PC, PLCO Project Team, Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: Mortality results after 13 years of follow-up, J. Natl. Cancer Inst 104 (2012) 125–132. [PubMed: 22228146]

[4]. Bickel PJ, Klaassen CA, Ritov Y, Wellner JA, Efficient and Adaptive Estimation for Semiparametric Models, The Johns Hopkins University Press, Baltimore, MD, 1993.

[5]. Chatterjee N, Carroll RJ, Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies, Biometrika 92 (2005) 399–418.

[6]. Chatterjee N, Chen Y-H, Luo S, Carroll RJ, Analysis of case-control association studies: SNPs, imputation and haplotypes, Statist. Sci 24 (2009) 489–502.

[7]. Chatterjee N, Shi J, García-Closas M, Developing and evaluating polygenic risk prediction models for stratified disease prevention, Nature Rev. Genet 17 (2016) 392–406. [PubMed: 27140283]

[8]. Chen YH, Chatterjee N, Carroll RJ, Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association, Biostatistics 9 (2008) 81–99. [PubMed: 17490987]

[9]. Chen YH, Chatterjee N, Carroll RJ, Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies, J. Amer. Statist. Assoc 104 (2009) 220–233.

[10]. Cornfield J, A statistical problem arising from retrospective studies, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 4, pp. 135–148.

[11]. Dudbridge F, Power and predictive accuracy of polygenic risk scores, PLoS Genet 9 (2013) e1003348. [PubMed: 23555274]

[12]. Evans DM, Visscher PM, Wray NR, Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk, Hum. Mol. Gen 18 (2009) 3525–3531. [PubMed: 19553258]

[13]. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, et al., The genetic architecture of type 2 diabetes, Nature 536 (2016) 41–47. [PubMed: 27398621]

[14]. Gauderman WJ, Zhang P, Morrison JL, Lewinger JP, Finding novel genes by testing G × E interactions in a genome-wide association study, Genetic Epidemiol 37 (2013) 603–613.

[15]. Han SS, Rosenberg PS, Ghosh A, Landi MT, Caporaso NE, Chatterjee N, An exposure-weighted score test for genetic associations integrating environmental risk factors, Biometrics 71 (2015) 596–605. [PubMed: 26134142]

[16]. Hayes RB, Reding D, Kopp W, Subar AF, Bhat N, Rothman N, Caporaso N, Ziegler RG, Johnson CC, Weissfeld JL, Hoover RN, Hartge P, Palace C, Gohagan JK, et al., Etiologic and early marker studies in the Prostate, Lung, Colorectal and Ovarian, PLCO cancer screening trial, Controlled Clin. Trials 21 (2000) 349S–355S. [PubMed: 11189687]

[17]. Hunter DJ, Gene–environment interactions in human diseases, Nature Rev. Genet 6 (2005) 287–298. [PubMed: 15803198]

[18]. Jiang Y, Scott AJ, Wild CJ, Secondary analysis of case-control data, Stat. Med 25 (2006) 1323–1339. [PubMed: 16220494]

[19]. Lin D, Zeng D, Proper analysis of secondary phenotype data in case-control association studies, Genet. Epidemiol 33 (2009) 256–265. [PubMed: 19051285]

[20]. Ma Y, A semiparametric efficient estimator in case-control studies, Bernoulli 16 (2010) 585–603.

[21]. Murcray CE, Lewinger JP, Gauderman WJ, Gene-environment interaction in genome-wide association studies, Am. J. Epidemiol 169 (2009) 219–226. [PubMed: 19022827]

[22]. Ottman R, Gene-environment interaction: Definitions and study designs, Prev. Med 25 (1996) 764. [PubMed: 8936580]

[23]. Piegorsch WW, Weinberg CR, Taylor JA, Non-hierarchical logistic models and case-only designs for assessing susceptibility in population based case-control studies, Stat. Med 13 (1994) 153–162. [PubMed: 8122051]

[24]. Prentice RL, Pyke R, Logistic disease incidence models and case-control studies, Biometrika 66 (1979) 403–411.

[25]. Prorok PC, Andriole GL, Bresalier RS, Buys SS, Chia D, Crawford ED, Fogel R, Gelmann EP, Gilbert F, Hasson MA, Hayes RB, Johnson CC, Mandel JS, Oberman A, O'Brien B, Oken MM, Rafla S, Reding D, Rutt W, Weissfeld JL, Yokochi L, Gohagan JK, et al., Design of the Prostate, Lung, Colorectal and Ovarian, PLCO cancer screening trial, Controlled Clin. Trials 21 (2000) 273S–309S. [PubMed: 11189684]

[26]. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P, Ruderfer DM, McQuillin A, Morris DW, et al., Common polygenic variation contributes to risk of schizophrenia and bipolar disorder, Nature 460 (2009) 748–752. [PubMed: 19571811]

[27]. Tsiatis AA, Semiparametric Theory and Missing Data, Springer, New York, 2007.

[28]. Tsiatis AA, Ma Y, Locally efficient semiparametric estimators for functional measurement error models, Biometrika 91 (2004) 835–848.

Table 1

Simulation results from 1000 simulated case-control samples taken from a population with a disease rate of approximately 4.5%, and independent genetic and environmental variables, under the logistic model with gene–environment interaction. The results for $G \sim \mathcal{B}$ (0.6) and $X \sim$ N(0, 1) is displayed on the left whereas the results for $G \sim \mathcal{N}$ N(0, 1) and $X \sim$ N(0, 1) is on the right. Each replicate contains $N_1 = 1000$ cases and $N_0 = 1000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary *G*, Normal *X*			**Normal *G*, Normal *X***		
	$\boldsymbol{\beta}$	0.76	0.36	−0.63	0.76	0.36	−0.63
Logistic	Mean	0.761	0.363	−0.635	0.762	0.363	−0.634
	se	0.101	0.088	0.103	0.055	0.053	0.056
	est se	0.101	0.084	0.101	0.056	0.054	0.055
	95%	0.952	0.939	0.942	0.950	0.954	0.942
Semi	Mean	0.761	0.360	−0.630	0.761	0.362	−0.627
	se	0.101	0.077	0.082	0.054	0.051	0.046
	est se	0.100	0.073	0.079	0.053	0.051	0.041
	95%	0.953	0.939	0.941	0.949	0.953	0.921
	MSE Eff	1.003	1.325	1.566	1.068	1.112	1.457

Table 2

Simulation results from 1000 simulated case-control samples taken from a population with a disease rate of approximately 10%, and independent genetic and environmental variables, under the logistic model with gene–environment interaction. The results for $G \sim \mathcal{B}$ (0.6) and $X \sim$ N(0, 1) is displayed on the left whereas the results for $G \sim \mathcal{N}$ N(0, 1) and $X \sim$ N(0, 1) is on the right. Each replicate contains $N_1 = 1000$ cases and $N_0 = 1000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary *G*, Normal *X*			**Normal *G*, Normal *X***		
	β	0.76	0.36	−0.63	0.76	0.36	−0.63
Logistic	Mean	0.762	0.363	−0.638	0.762	0.363	−0.633
	Se	0.102	0.084	0.100	0.056	0.051	0.057
	est se	0.100	0.083	0.100	0.056	0.053	0.057
	95%	0.943	0.952	0.955	0.957	0.960	0.952
Semi	Mean	0.762	0.359	−0.628	0.761	0.363	−0.629
	se	0.102	0.077	0.087	0.055	0.050	0.053
	est se	0.100	0.074	0.081	0.055	0.052	0.050
	95%	0.944	0.932	0.936	0.953	0.960	0.934
	MSE Eff	1.004	1.180	1.325	1.032	1.065	1.145

Table 3

Simulation results from 1000 simulated case-control samples taken from a population with a disease rate of approximately 4.5%, and independent genetic and environmental variables, under the logistic model with gene–environment interaction. The results for $G \sim \mathcal{B}$ (0.6) and $X \sim \mathcal{G}$ (20, 1) is displayed on the left whereas the results for $G \sim \mathcal{N}$ N(0, 1) and $X \sim \mathcal{G}$ (20, 1) is on the right. Each replicate contains $N_1 = 1000$ cases and $N_0 = 1000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary *G*, Gamma *X*			**Normal *G*, Gamma *X***		
	β	3.577	0.080	−0.141	3.577	0.080	−0.141
Logistic	Mean	3.599	0.081	−0.142	3.592	0.080	−0.141
	se	0.456	0.018	0.022	0.269	0.012	0.012
	est se	0.462	0.018	0.022	0.259	0.012	0.012
	95%	0.957	0.953	0.949	0.937	0.950	0.942
Semi	Mean	3.586	0.080	−0.141	3.569	0.080	−0.140
	se	0.375	0.016	0.018	0.230	0.011	0.010
	est se	0.369	0.016	0.017	0.202	0.011	0.009
	95%	0.950	0.949	0.942	0.914	0.940	0.919
	MSE Eff	1.484	1.305	1.559	1.372	1.059	1.437

Table 4

Simulation results from 1000 simulated case-control samples taken from a population with a disease rate of approximately 10%, and independent genetic and environmental variables, under the logistic model with gene–environment interaction. The results for $G \sim \mathcal{B}$ (0.6) and $X \sim \mathcal{G}$ (20, 1) is displayed on the left whereas the results for $G \sim \mathcal{N}$ N(0, 1) and $X \sim \mathcal{G}$ (20, 1) is on the right. Each replicate contains $N_1 = 1000$ cases and $N_0 = 1000$ controls, and is analyzed through two approaches, (1) "Logistic" is ordinary logistic regression, and (2) "Semi" is our semiparametric efficient estimator. Here, we list the sample mean ("mean"), the sample standard error ("se"), the mean estimated standard error ("est se") and the coverage for the nominal 95% confidence intervals ("95%") for both methods. In addition, we computed the mean squared error efficiency of the "Semi" method compared to the "Logistic" approach.

		Binary *G*, Gamma *X*			**Normal *G*, Gamma *X***		
	β	3.577	0.080	−0.141	3.577	0.080	−0.141
Logistic	Mean	3.589	0.081	−0.141	3.600	0.081	−0.142
	se	0.459	0.018	0.022	0.274	0.012	0.013
	est se	0.460	0.018	0.022	0.269	0.012	0.012
	95%	0.949	0.950	0.947	0.950	0.934	0.944
Semi	Mean	3.565	0.080	−0.140	3.590	0.081	−0.142
	se	0.394	0.016	0.019	0.268	0.012	0.012
	est se	0.381	0.016	0.018	0.247	0.011	0.011
	95%	0.945	0.953	0.938	0.934	0.937	0.930
	MSE Eff	1.360	1.240	1.406	1.048	1.031	1.061

Table 5

Analysis of the case-control study on prostate cancer, containing $n_1 = 690$ cases and $n_0 = 717$ controls. Two approaches were implemented, (1) “Logistic” is ordinary logistic regression, and (2) “Semi” is our semiparametric efficient estimator. Displayed are the estimates, bootstrap standard error (“se, bootstrap”), mean estimated asymptotic standard error (“est se, asymptotic”), bootstrap p-value (“ p-value, bootstrap”), and asymptotic p-value (“ p-value, asymptotic”) of the coefficients for the standardized polygenic risk score (G), [25(OH)D] level (X), and the interaction between them (GX).

		β_G	β_X	β_{GX}
Logistic	Estimates	0.169	0.123	−0.101
	se, bootstrap	0.056	0.056	0.054
	est se, asymptotic	0.055	0.055	0.055
	p-value, bootstrap	0.002	0.028	0.064
	p-value, asymptotic	0.002	0.024	0.066
Semi	Estimates	0.168	0.124	−0.110
	se, bootstrap	0.056	0.056	0.049
	est se, asymptotic	0.055	0.054	0.042
	p-value, bootstrap	0.003	0.027	0.026
	p-value, asymptotic	0.002	0.021	0.009