ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 1, *pp.* 127–151

Semiparametric estimation in the secondary analysis of case–control studies

Yanyuan Ma

University of South Carolina, Columbia, and Texas A&M University, College Station, USA

and Raymond J. Carroll

Texas A&M University, College Station, USA

[Received April 2013. Final revision October 2014]

Summary. We study the regression relationship between covariates in case–control data: an area known as the secondary analysis of case–control studies. The context is such that only the form of the regression mean is specified, so that we allow an arbitrary regression error distribution, which can depend on the covariates and thus can be heteroscedastic. Under mild regularity conditions we establish the theoretical identifiability of such models. Previous work in this context has either specified a fully parametric distribution for the regression errors, specified a homoscedastic distribution for the regression errors, has specified the rate of disease in the population (we refer to this as the true population) or has made a rare disease approximation. We construct a class of semiparametric estimation procedures that rely on none of these. The estimators differ from the usual semiparametric estimators in that they draw conclusions about the true population, while technically operating in a hypothetical superpopulation. We also construct estimators with a unique feature, in that they are robust against the misspecification of the regression error distribution in terms of variance structure, whereas all other non-parametric effects are estimated despite the biased samples. We establish the asymptotic properties of the estimators and illustrate their finite sample performance through simulation studies, as well as through an empirical example on the relationship between red meat consumption and hetero-cyclic amines. Our analysis verified the positive relationship between red meat consumption and two forms of hetro-cyclic amines, indicating that increased red meat consumption leads to increased levels of MeIQx and PhIP, both being risk factors for colorectal cancer. Computer software as well as data to illustrate the methodology are available from http://www.stat.tamu.edu/~carroll/matlab_programs/software.php.

Keywords: Biased samples; Case–control study; Heteroscedastic regression; Secondary analysis; Semiparametric estimation

1. Introduction

Population-based case–control designs, hereafter called simply case–control designs, are popularly used for studying risk factors for rare diseases, such as cancers. The idealized set-up of such designs is as follows. At a given time, there is an underlying base population, which we refer to as the *true population* throughout the paper. Within the true population, there are two subpopulations: those with the disease, called cases, and those without the disease, called controls. Separately, a random sample is taken from the case subpopulation, and a random

Address for correspondence: Yanyuan Ma, Department of Statistics, University of South Carolina, Columbia, SC 29208,USA.
E-mail: yanyuanma@gmail.com

1369–7412/16/78127

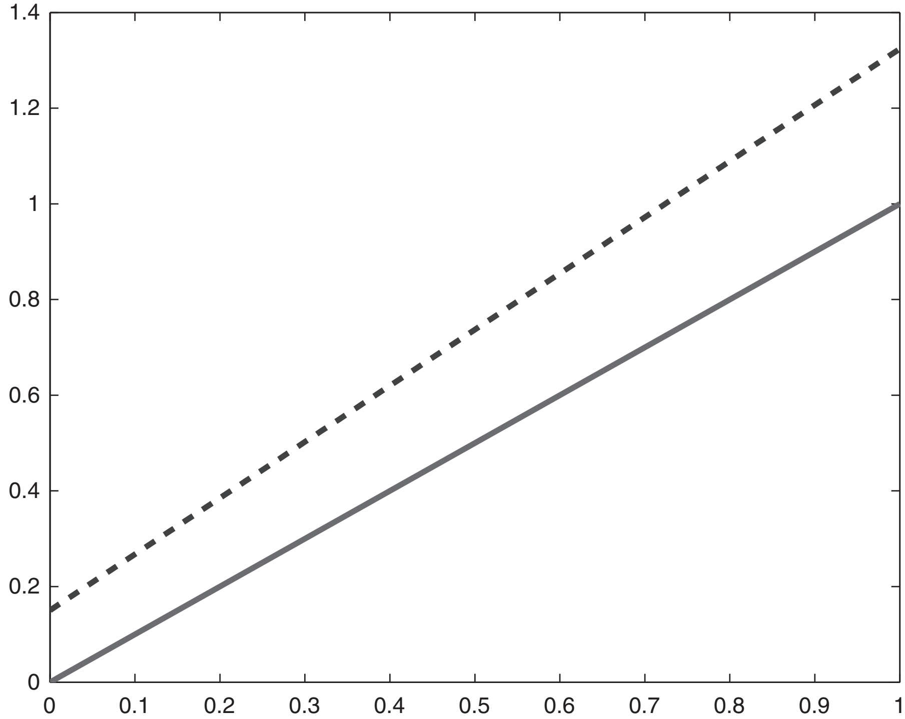

Fig. 1. Illustration of the bias induced by the case–control sampling scheme: ——, true regression function; – – –, regression function when using all the data and ignoring the case–control sampling scheme

sample is taken from the control subpopulation. Data on various covariates are then collected in a retrospective fashion, so that they reflect history before the disease. Nested case–control studies and case–cohort or case–base studies are variations of the retrospective case–control design.

The primary purpose of case–control designs is to understand the relationship between disease occurrence and the covariates. The *secondary analysis* of such case–control data (Jiang *et al.*, 2006; Lin and Zeng, 2009; Li *et al.*, 2010; Wei *et al.*, 2013; He *et al.*, 2012) is based on the realization that the data further provide information about the relationship between the covariates. The relationships between covariates are often of interest as well, as they can reveal associations between various covariates such as gene–environment, gene–gene and environment–environment associations. These analyses become especially important when, as for retrospective sampling, a random sample from the true population is not available; see the secondary analysis literature mentioned above for more examples. If we seek to understand the regression relationship between covariates Y and $\mathbf{X}$ in the true population, we generally cannot use the case–control data set as if it were a random sample from the true population. Indeed, unless disease is independent of Y given $\mathbf{X}$, the regression of Y on $\mathbf{X}$ based on the case–control sample will lead to a relationship which is different from that in the true population.

To see this numerically, we first define our notation. There are N_0 cases and N_1 controls, with $N = N_0 + N_1$. Suppose that $N_0 = N_1 = 500$, and that disease status D is related to covariates $(Y, \mathbf{X})$ in the true population through the linear logistic model

$$\mathrm{pr}(D=d|\mathbf{X}=x, Y=y)=f^{\mathrm{true}}_{D|X,Y}(d,\mathbf{x},y)=H(d,\mathbf{x},y,\boldsymbol{\alpha})=\frac{\exp\{d(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)\}}{1+\exp(\alpha_c+\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}_1+y\alpha_2)}, \qquad (1)$$

where, for this illustration, $\boldsymbol{\alpha}=(\alpha_c,\boldsymbol{\alpha}_1,\alpha_2)=(-5.5,1.0,0.5)$. Suppose further that the regression relationship in the true population is that $Y=\beta_c+X\beta+\varepsilon$, with $\beta_c=0$, $\beta=1$ and $\varepsilon\sim N(0,1)$. In addition, in the true population, $X\sim\mathrm{uniform}(0,1)$. In this set-up, suppose that the disease is rare, with $\mathrm{pr}(D=1)\approx 0.01$. Thus, although controls are 99% of the true population, they are only 50% of the case–control study. To understand the bias that is induced by ignoring the case–control sampling scheme, we generated 3000 case–control studies with intercept $\beta_c=0$ and slope $\beta=1$, and computed the intercept and slope estimates by using all the data. Simply regressing Y on X and ignoring the case–control sampling scheme, the mean estimated intercept and slope across the 3000 simulated data sets were 0.150 and 1.174 respectively, reflecting considerable bias, which leads to a coverage rate of only 67% for a nominal 95% confidence interval. Fig. 1 shows the attained regression function compared with the true regression function. Using the method that we develop in this paper, our method yields the average intercept and slope estimates of 0.0024 and 1.0035, thus eliminating the bias that is caused by ignoring the case–control sampling scheme.

The bias in the secondary analysis is in stark contrast with what happens in the primary analysis, where estimating $(\boldsymbol{\alpha}_1,\alpha_2)$ is of interest. It is well known that $\boldsymbol{\alpha}_1$ and α_2 can be estimated consistently via ordinary logistic regression of D on $(Y,\mathbf{X})$ by treating the case–control sample as if it were a random sample of the true population (Prentice and Pyke, 1979).

Our goal is to estimate the regression of Y on $\mathbf{X}$ in the true population, using case–control data, where, for a function $m(\cdot)$ known up to a parameter $\boldsymbol{\beta}$,

$$Y=m(\mathbf{X},\boldsymbol{\beta})+\varepsilon, \qquad (2)$$

where we make only the assumption that $E(\varepsilon|\mathbf{X})=0$. Two solutions to estimating $\boldsymbol{\beta}$ have been proposed in the literature. Lin and Zeng (2009) and, obliquely, Chen *et al.* (2008) proposed a particular fully parametric distribution for ε and then performed a semiparametric efficient analysis, where the distribution of $\mathbf{X}$ is non-parametric. There is excellent software for this problem in the case that $\varepsilon=N(0,\sigma^2)$, i.e. homoscedastic and normally distributed (`http://www.bios.unc.edu/~lin/software/SPREG/`). To implement this software, however, one must either specify the disease rate $\mathrm{pr}(D=1)$ in the true population or one must make a 'rare disease' assumption, which is implemented by assuming that $\mathrm{pr}(D=1)<0.01$. When the disease rate is known, reweighting the observations also corrects the biases (Scott and Wild, 2002). Wei *et al.* (2013) dispensed with the normality assumption but still assumed a homoscedastic distribution for ε independent of $\mathbf{X}$ and made a rare disease approximation.

In practice, the disease rate in the population being sampled is not known. In addition, it might not be rare. As an example, in Section 6, we use data from a case–control study of colorectal adenoma, a precursor to colorectal cancer, relating measures of hetero-cyclic amines (HCAs) to red meat consumption. Whereas colorectal cancer is rare, colorectal adenomas are not, being of the order of 7% or more depending on the population being sampled (Yamaji *et al.*, 2004; Corley *et al.*, 2014). In this data set, one of the regressions is also heavily heteroscedastic. We shall demonstrate that both approaches mentioned above have problems when some of the assumptions, such as the rare disease assumption, the known disease rate assumption and the known error distribution assumption, are violated (Tables 1–6).

To relax such assumptions, novel methods are needed. In this paper, we do not assume any distributional form for ε or $\varepsilon|\mathbf{X}$, we do not assume that the regression is homoscedastic, we do not require the disease rate to be known and we do not make a rare disease approximation. We do this by adopting the concept of a superpopulation (Ma, 2010): a similar idea was called an alternative characterization of the case–control study by Chen *et al.* (2009).

The main idea behind a superpopulation is to enable us to view the case–control sample as a sample of independent and identically distributed observations from the superpopulation. Conceptually, a superpopulation is simply a proportional expansion of the case–control sample to ∞. Why a superpopulation constructed through such expansion achieves the purpose of viewing the case–control sample as an independent and identically distributed sample was studied carefully by Ma (2010). The ability to view the case–control sample as a random sample permits us to use classical semiparametric approaches (Bickel *et al.*, 1993; Tsiatis, 2006), regardless of whether the disease rate in the real population is rare or not, or is known or not.

We derive a class of semiparametric estimators and identify the efficient member. We further construct a member of the family that is relatively simple to compute, and we illustrate how to construct the efficient estimator, applicably to both rare and common diseases. The derivation of semiparametric estimators in this context is challenging because the calculations must use quantities that are defined in the unknown true population to perform analysis in the superpopulation, since the models under the true population and the superpopulation share common parameters. In addition, as established in Ma (2010), the resulting semiparametric estimators further retain asymptotic consistency, a root n rate of convergence, asymptotic normality and semiparametric efficiency with respect to the true population as well. For example, our efficient estimator has the usual property that its asymptotic variance cannot be further reduced by any other device or by taking into account the case–control sampling structure.

The rest of the paper is organized as follows. Under conditions, we first establish the technical identifiability of our problem in Section 2. In Section 3, we formulate the problem as a classic semiparametric problem by using the superpopulation notion and we carry out analytic calculations to prepare for the estimation procedure. In Section 4, we describe details of implementation and the asymptotic theory. Simulation studies are performed in Section 5 to illustrate the finite sample performance of the procedure, showing that our method is robust, efficient and maintains nominal coverage for confidence intervals. An empirical analysis is provided in Section 6. Section 7 contains a short discussion. Technical details are given in Appendix A, as well as in the on-line supplementary material. Computer code and data to illustrate our method are available from `http://www.stat.tamu.edu/~carroll/matlab_programs/software.php`.

2. Superpopulation model framework

The primary disease model is the linear logistic model (1), with $\boldsymbol{\alpha} = (\alpha_c, \boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}}$. Here and throughout the text, we use the superscript 'true' to represent quantities or operations related to the underlying true population, and also to distinguish it from a superpopulation that will be formally introduced later. In addition, in this underlying true population, Y is believed to be related to $\mathbf{X}$ through equation (2), which we rewrite as the regression model

$$f_{Y|X}^{\text{true}}(\mathbf{x}, y) = \eta_2\{y - m(\mathbf{x}, \boldsymbol{\beta}), \mathbf{x}\}, \qquad (3)$$

where $m(\cdot)$ is the regression mean function known up to the parameter $\boldsymbol{\beta}$ and η_2 is an unknown probability density function that has mean 0 given $\mathbf{X}$. Defining $\varepsilon = Y - m(\mathbf{X}, \boldsymbol{\beta})$, then $E(\varepsilon|\mathbf{X}) = 0$. The distribution of ε, whether conditionally on $\mathbf{X}$ or marginally, is left unspecified. In particular, heteroscedasticity is allowed. Making the identification $\eta_2(\varepsilon, \mathbf{X}) = \eta_2\{Y - m(\mathbf{X}, \boldsymbol{\beta}), \mathbf{X}\}$, this means that $\eta_2 \geqslant 0$ satisfies $\int \varepsilon \eta_2(\varepsilon, \mathbf{x}) \mathrm{d}\mu(\varepsilon) = 0$ and $\int \eta_2(\varepsilon, \mathbf{x}) \, \mathrm{d}\mu(\varepsilon) = 1$, but its form is unknown. Here and throughout the text, we use $\mu(\cdot)$ to denote a Lebesgue measure for a continuous random variable and a counting measure for a discrete random variable. The distribution of the covariate $\mathbf{X}$ in the underlying true population is also unspecified, and its density or mass function is $f_{\mathbf{X}}^{\text{true}}(\mathbf{x}) = \eta_1(\mathbf{x})$, where $\eta_1 \geqslant 0$ satisfies $\int \eta_1(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = 1$.

The superpopulation framework of Ma (2010) is that we can think of the case–control sample as a random sample from an imaginary infinite superpopulation, in which the disease to non-disease ratio is N_1/N_0. Let $N_d = N_0$ when $d = 0$ and $N_d = N_1$ when $D = 1$. Define the true probability that $D = d$ as

$$p_D^{\text{true}}(d, \boldsymbol{\alpha}, \boldsymbol{\beta}, \eta_1, \eta_2) = \int \eta_1(\mathbf{x})\,\eta_2(\varepsilon, \mathbf{x})\,H(d, \mathbf{x}, y, \boldsymbol{\alpha})\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y).$$

The density of $(D, Y, \mathbf{X})$ in the superpopulation is defined as

$$f_{X,Y,D}(\mathbf{x}, y, d) = \frac{N_d}{N}\,\frac{\eta_1(\mathbf{x})\,\eta_2(\varepsilon, \mathbf{x})\,H(d, \mathbf{x}, y, \boldsymbol{\alpha})}{p_D^{\text{true}}(d, \boldsymbol{\alpha}, \boldsymbol{\beta}, \eta_1, \eta_2)}. \tag{4}$$

Although $\boldsymbol{\beta}$ appears in ε, for notational brevity, we do not explicitly write $\varepsilon(\boldsymbol{\beta})$. In the secondary analysis framework, the main interest is $\boldsymbol{\beta}$. However, we formally treat $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as the parameter of interest. We treat $\eta_1(\cdot)$ and $\eta_2(\cdot, \cdot)$ as the infinite dimensional nuisance parameters, thus bypassing the need to estimate them.

Remark 1. When no assumptions are made about the relationship between Y and X in the true population, the logistic intercept α_c is not identified (Prentice and Pyke, 1979); nor is the regression of Y on X. Thus, if consistency of estimation is desired, truly non-parametric regression in a case–control study of our type is not possible. We believe that the key to identification lies in placing a restriction on the joint distribution of (Y, X) in the base population. For example, Chatterjee and Carroll (2005) showed that, if Y and X are independent, then α_c is generally identified, and they showed this explicitly when one of the two is discrete. In our case, the restriction is a parametric model for $E(Y|X)$. It is a reasonable conjecture that such a restriction is enough for the identifiability of α_c: a conjecture that we confirm next.

2.1. Identifiability

We first establish identifiability of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the superpopulation. For greater generality, we consider the slightly more flexible model

$$H(d, \mathbf{x}, y) = \frac{\exp[d\{\alpha_c + u(\mathbf{x}, y, \boldsymbol{\alpha}_1, \alpha_2)\}]}{1 + \exp\{\alpha_c + u(\mathbf{x}, y, \boldsymbol{\alpha}_1, \alpha_2)\}},$$

where $u(0, 0, \boldsymbol{\alpha}_1, \alpha_2) = 0$ for all $\boldsymbol{\alpha}_1$ and α_2. Obviously, this model contains the original linear logistic model that we are studying. We assume that there is no $(\boldsymbol{\alpha}_1^{\mathrm{T}}, \alpha_2)^{\mathrm{T}} \neq (\tilde{\alpha}_1^{\mathrm{T}}, \tilde{\alpha}_2)^{\mathrm{T}}$ such that, for all $(\mathbf{x}, y)$, $u(\mathbf{x}, y, \boldsymbol{\alpha}_1, \alpha_2) = u(\mathbf{x}, y, \tilde{\boldsymbol{\alpha}}_1, \tilde{\alpha}_2)$. These are natural minimal conditions that are usually satisfied automatically as long as the parameterizations of u and m are not redundant. We also assume the following two conditions.

Assumption 1. Assume that the second moment of ε is bounded marginally and η_2 is a bounded function, i.e. $E(\varepsilon^2) < \infty$ and $\sup_{\mathbf{x},\varepsilon} \eta_2(\varepsilon, \mathbf{x}) < \infty$. For any fixed parameters $\boldsymbol{\alpha}_1$, α_2 and $\boldsymbol{\beta}$, and any $\delta > 0$, there is a constant vector $\mathbf{c}_1$, a constant $c_2 \in [0, 1]$ and a region $\mathcal{D}$ with complement $\mathcal{D}^{\mathrm{c}}$ such that, when $\mathbf{x} \to \mathbf{c}_1$,

$$\sup_{\varepsilon \in \mathcal{D}^{\mathrm{c}}} \lim_{\mathbf{x} \to \mathbf{c}_1} |(1 + \exp[\alpha_c + u\{\mathbf{x}, m(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon, \boldsymbol{\alpha}_1, \alpha_2\}])^{-1} - c_2| = 0,$$

and $\lim_{\mathbf{x} \to \mathbf{c}_1} \mathrm{pr}(\varepsilon \in \mathcal{D} | \mathbf{X} = \mathbf{x}) < \delta$. In addition, for any element $e \in \mathcal{D}$, $|e| \geqslant 1$. Typically we expect

$\mathcal{D}^c = [-K, K]$ for some large K, $\mathbf{c}_1 = \infty$ or $c_1 = -\infty$ or contains $\pm\infty$ as components, and $c_2 = 0$ or $c_2 = 1$, although this is not required.

Assumption 2. $c(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) = \lim_{\mathbf{x}\to\mathbf{c}_1}\{m(\mathbf{x}, \tilde{\boldsymbol{\beta}}) - m(\mathbf{x}, \boldsymbol{\beta})\} \neq 0$ for $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$.

Remark 2. Assume that $\text{pr}(|\varepsilon| > K|\mathbf{X} = \mathbf{x}) \to 0$ as $K \to \infty$ uniformly in $\mathbf{x}$. We can easily verify that when both m and u are linear functions, where we write $m(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1 + \beta_c$, both assumptions are satisfied except when $\boldsymbol{\alpha}_1 + \boldsymbol{\beta}_1\alpha_2 = 0$. When this happens, $u\{m(\mathbf{x}, \boldsymbol{\beta}), \boldsymbol{\alpha}_1, \alpha_2\}$ degenerates to a constant, and we can verify that, although $\boldsymbol{\beta}_1$ is still identifiable, β_c and α_c are no longer identifiable; see the on-line supplementary material for details of verification of both the identifiability and the non-identifiability.

We state the identifiability result in proposition 1 and provide the proof in Appendix A.1.

Proposition 1. Make assumptions 1 and 2. Also assume that there are constants (C_1, C_2) such that $0 < C_1 < N_0/N_1 < C_2 < \infty$. Then the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are identifiable.

Remark 3. Identifiability under some specific situations has been considered in the literature. For example, Chatterjee and Carroll (2005), Chatterjee *et al.* (2006) and Chen *et al.* (2009) considered the case that $\mathbf{X}$ and Y are independent, whereas Chen *et al.* (2008) and Lin and Zeng (2009) explicitly studied the identifiability issue when the disease rate model is linear logistic and the secondary model is fully parametric. The model that we consider here is more general, in that only a mean function is assumed for the secondary model. It is noted that although, in practice, it may be difficult to estimate α_c, estimation of the other parameters can still be performed effectively; see also Lobach *et al.* (2008).

3. Analytic derivations

3.1. True and conjectured models

The major point of our paper is that we propose a model for $E(Y|\mathbf{X})$ only, denoted $m(\mathbf{X}, \boldsymbol{\beta})$, and we specifically want to avoid positing a model for the density function of the regression errors $\varepsilon = Y - m(\mathbf{X}, \boldsymbol{\beta})$ conditionally on $\mathbf{X}$. We shall accomplish this by a two-step process. First, in Section 3.2, we shall derive the semiparametric efficient estimating equation in the superpopulation for estimating $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ when the density of Y given X in the true population is known. Recognizing that we do not want to make such an assumption, in Section 4, we shall show how to modify the estimating equation so that it has mean 0 asymptotically, even if the model conjectured for the regression errors is false, thus resulting in model robust consistent estimation.

3.2. Analysis under a true model

As described in Section 3.1, here we shall derive the form of the semiparametric efficient estimating equation when the model conjectured for the regression errors in equation (3) is true. Later, in Section 4, we shall modify the estimating function to make it model robust.

Viewing the observations as randomly sampled from the superpopulation, we can perform a conventional semiparametric analysis. Of course, all the calculations need to be done with respect to the superpopulation, and all the probability statements need to be with respect to Lebesgue measure for continuous random variables and counting measure for discrete random variables in the superpopulation, and they will be if not otherwise pointed out. The functions (η_1, η_2, H), which are probability density or probability mass functions in the true population, do not represent the corresponding probability density or probability mass functions in the superpopulation. They are merely functions that satisfy $\eta_1(\mathbf{x}) \geqslant 0$, $\int \eta_1(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}) = 1$, $\eta_2(\varepsilon,\mathbf{x}) \geqslant 0$, $\int \eta_2(\varepsilon,\mathbf{x})\,\mathrm{d}\mu(\varepsilon) = 1$, $\int \varepsilon\eta_2(\varepsilon,\mathbf{x})\,\mathrm{d}\mu(\varepsilon) = 0$, $H(d,\mathbf{x},y) \geqslant 0$ and $H(0,\mathbf{x},y) + H(1,\mathbf{x},y) = 1$. In fact, we introduced these expressions to discourage the mistake of automatically viewing them as the corresponding density or mass functions in the superpopulation.

Using model (4), calculating the partial derivative of the log-likelihood with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, it is easy to see that the score function has the form $\mathbf{S}_\theta(\mathbf{X},Y,D,\boldsymbol{\theta}) = \mathbf{S}(\mathbf{X},Y,D,\boldsymbol{\theta}) - E(\mathbf{S}|D)$, where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{S}_\theta = (\mathbf{S}_\alpha^{\mathrm{T}},\mathbf{S}_\beta^{\mathrm{T}})^{\mathrm{T}}$ and

$$\mathbf{S}(\mathbf{x},y,d,\boldsymbol{\theta}) = \left\{ \begin{array}{c} \partial\log\{H(d,\mathbf{x},y,\boldsymbol{\alpha})\}/\partial\boldsymbol{\alpha} \\ \partial\log\{\eta_2(\varepsilon,\mathbf{x})\}/\partial\boldsymbol{\beta} \end{array} \right\}. \tag{5}$$

Explicitly,

$$\begin{aligned} \mathbf{S}_\alpha(\mathbf{X},Y,D,\boldsymbol{\theta}) &= \partial\log\{H(D,\mathbf{X},Y,\boldsymbol{\alpha})\}/\partial\boldsymbol{\alpha} - E[\partial\log\{H(D,\mathbf{X},Y,\boldsymbol{\alpha})\}/\partial\boldsymbol{\alpha}|D], \\ \mathbf{S}_\beta(\mathbf{X},Y,D,\boldsymbol{\theta}) &= \partial\log\{\eta_2(\varepsilon,\mathbf{X})\}/\partial\boldsymbol{\beta} - E[\partial\log\{\eta_2(\varepsilon,\mathbf{X})\}/\partial\boldsymbol{\beta}|D]. \end{aligned}$$

In Appendix A.2, we further derive the nuisance tangent space Λ and its orthogonal complement space $\Lambda^\perp$ as

$$\begin{aligned} \Lambda &= \{\mathbf{g}(\varepsilon,\mathbf{X}) - E(\mathbf{g}|D) : E_{\text{true}}(\mathbf{g}) = E_{\text{true}}(\varepsilon\mathbf{g}|\mathbf{X}) = 0 \text{ almost surely}\}; \\ \Lambda^\perp &= [\mathbf{h}(D,\varepsilon,\mathbf{X}) : E(\mathbf{h}) = 0, E\{\mathbf{h} - E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\} \\ &\quad \times \sum_d (N_d/N)\,H(d,\mathbf{X},y)/p_D^{\text{true}}(d) = \varepsilon\mathbf{a}(\mathbf{X}) \text{ almost surely}], \end{aligned}$$

where $\mathbf{g}(\varepsilon,\mathbf{x})$ and $\mathbf{h}(D,\varepsilon,\mathbf{x})$ are arbitrary functions that satisfy their respective constraints described above, $\mathbf{a}(\mathbf{x})$ is an arbitrary function of $\mathbf{x}$ and ‘almost surely’ is with respect to the true superpopulation distribution.

Having obtained both the score function and the two spaces Λ and $\Lambda^\perp$, conceptually, we need only to project the score function onto $\Lambda^\perp$ to obtain the efficient score $\mathbf{S}_{\text{eff}}$. Doing this is, however, extraordinarily technical, and hence we defer the details to the on-line supplementary material. Here we merely state the result in proposition 2, which requires a series of definitions, as follows. Define

$$\begin{aligned} \pi_0 = p_D^{\text{true}}(0) &= \int \eta_1(\mathbf{x})\,\eta_2(\varepsilon,\mathbf{x})\,H(0,\mathbf{x},y)\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y), \\ \pi_1 = p_D^{\text{true}}(1) &= \int \eta_1(\mathbf{x})\,\eta_2(\varepsilon,\mathbf{x})H(1,\mathbf{x},y)\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y), \\ b_0 &= E\{f_{D|\mathbf{X},Y}(1,\mathbf{X},Y)|D=0\}, \\ b_1 &= E\{f_{D|\mathbf{X},Y}(0,\mathbf{X},Y)|D=1\}, \\ \mathbf{c}_0 &= E(\mathbf{S}|D=0) - E\{E(\mathbf{S}|\varepsilon,\mathbf{X})|D=0\}, \\ \mathbf{c}_1 &= E(\mathbf{S}|D=1) - E\{E(\mathbf{S}|\varepsilon,\mathbf{X})|D=1\}, \\ \kappa(\mathbf{x},y) &= \left[\sum_{d=0}^{1} \{N_d\,H(d,\mathbf{x},y)\}/(N\pi_d)\right]^{-1}, \\ t_1(\mathbf{X}) &= [E_{\text{true}}\{\varepsilon^2\kappa(\mathbf{X},Y)|\mathbf{X}\}]^{-1}, \\ \mathbf{t}_2(\mathbf{X}) &= E_{\text{true}}\,\{\varepsilon E(\mathbf{S}|\varepsilon,\mathbf{X})|\mathbf{X}\} - (\mathbf{c}_0/b_0)E_{\text{true}}\,\{\varepsilon\, f_{D|X,Y}(0,\mathbf{X},Y)|\mathbf{X}\}, \\ t_3(\mathbf{X}) &= -b_0^{-1}E_{\text{true}}\,\{\varepsilon f_{D|X,Y}(0,\mathbf{X},Y)|\mathbf{X}\}, \end{aligned}$$

$$
\begin{aligned}
\mathbf{a}(\mathbf{x}) &= t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x})+t_3(\mathbf{x})\mathbf{u}_0\},\\
\mathbf{u}_0 &= [1-E\{\varepsilon t_1(\mathbf{X})t_3(\mathbf{X})\kappa(\mathbf{X},Y)|D=0\}]^{-1}E\{\varepsilon t_1(\mathbf{X})\mathbf{t}_2(\mathbf{X})\kappa(\mathbf{X},Y)|D=0\},\\
\mathbf{u}_1 &= -(N_0/N_1)\mathbf{u}_0,\\
\mathbf{v}_0 &= (\pi_1/b_0)(\mathbf{u}_0+\mathbf{c}_0),\\
\mathbf{v}_1 &= -(\pi_0/b_0)(\mathbf{u}_0+\mathbf{c}_0),\\
\mathbf{g}(\varepsilon,\mathbf{x}) &= E(\mathbf{S}|\varepsilon,\mathbf{X}=\mathbf{x})-\varepsilon\mathbf{a}(\mathbf{x})\kappa(\mathbf{x},y)-\mathbf{v}_0 f_{D|X,Y}(0,\mathbf{x},y)-\mathbf{v}_1 f_{D|X,Y}(1,\mathbf{x},y).
\end{aligned}
\tag{6}
$$

Proposition 2. Make the definitions (6). In the superpopulation, the semiparametric efficient score function is $\mathbf{S}(\mathbf{X}_i,Y_i,D_i)-\mathbf{g}\{Y_i-m(\mathbf{X}_i,\boldsymbol{\beta}),\mathbf{X}_i\}-(N_0/N)\mathbf{v}_0-(N_1/N)\mathbf{v}_1$. The semiparametric efficient estimator is obtained by solving

$$\sum_{i=1}^{N}[\mathbf{S}(\mathbf{X}_i,Y_i,D_i)-\mathbf{g}\{Y_i-m(\mathbf{X}_i,\boldsymbol{\beta}),\mathbf{X}_i\}-(1-D_i)\mathbf{v}_0-D_i\mathbf{v}_1]=0. \tag{7}$$

We emphasize here that the estimator in proposition 2 is not only efficient with respect to the superpopulation; it is also efficient with respect to the true population. This is a direct consequence of the general result that, if an estimator is efficient with respect to the superpopulation, it is also efficient with respect to the true population. A careful justification of this claim was given in Ma (2010). Logically, this result can be understood because, if we could find a more efficient estimator with respect to the true population, this estimator would also be more efficient with respect to the superpopulation, which causes a contradiction. Intuitively, the special sampling strategy is in fact already absorbed into the formulation when we construct the superpopulation; hence no information has been lost during the conversion between populations.

4. Estimator construction

4.1. Basic calculations

The estimating equation (7) that was derived in proposition 2 is not useful, however, because it involves various calculations that rely on the unknown η_1 and η_2, which were assumed to be correctly conjectured in Section 3. If either is misspecified, the corresponding calculation will lead to inconsistent estimation of $\boldsymbol{\theta}$. The purpose of this section is to define estimators that are consistent for estimating $\boldsymbol{\theta}$ based on a posited score function, which we denote by $\mathbf{S}^*$. As it turns out, if the score function posited is correct, then, in addition to being consistent, the estimator of $\boldsymbol{\theta}$ has the additional property of being efficient. If the score function posited is incorrect, then the estimator of $\boldsymbol{\theta}$ is still consistent. So our method can be thought of as a locally efficient estimator.

A careful inspection of the estimation procedure given in proposition 2 and the definition of the related quantities suggests that the critical points lie in obtaining π_0 and π_1, in calculating $E(\mathbf{h}|\varepsilon,\mathbf{X})$ and $E(\mathbf{h}|D)$ for any function $\mathbf{h}(D,\mathbf{X},Y)$, and in calculating $E_{\text{true}}(\mathbf{h}|\mathbf{X})$ for any function $\mathbf{h}(D,\mathbf{X},Y)$.

Our algorithm is detailed as algorithm 1 in Table 1 and is based on the following considerations.

(a) First, we have that

$$
\begin{aligned}
N_d &= N\,p_D(d) = N\int f_{\mathbf{X},Y}(\mathbf{x},y)\,f_{D|X,Y}(d,\mathbf{x},y)\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y)\\
&= N\int f_{\mathbf{X},Y}(\mathbf{x},y)(N_dH/N\pi_d)\{\sum_d (N_dH)/(N\pi_d)\}^{-1}\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y).
\end{aligned}
$$

Table 1. Algorithm 1: computing the locally efficient score function†

Step 1: posit a model for $\eta_2(\varepsilon, \mathbf{x})$ which has mean 0, and calculate expression (5), calling the result $S^*(\mathbf{X}, Y, D)$; use $S^*(\cdot)$ in place of $S(\cdot)$ in expressions (6) and (7)

Step 2: estimate $f_{\mathbf{X}|D}(\mathbf{x}, d)$ by a kernel density estimate among the data with $D_i = d$, with result $\hat{f}_{\mathbf{X}|D}(\mathbf{x}, d)$

Step 3: solve $\hat{\pi}_0 = \Sigma_{i=1}^N H(0, \mathbf{X}_i, Y_i)\{N_0\, H(0, \mathbf{X}_i, Y_i)/\hat{\pi}_0 + N_1 H(1, \mathbf{X}_i, Y_i)/(1-\hat{\pi}_0)\}^{-1}$ to obtain $\hat{\pi}_0$ and set $\hat{\pi}_1 = 1 - \hat{\pi}_0$

Step 4: in the definition of $\kappa(\mathbf{x}, y)$ in expression (6), form $\hat{\kappa}(\mathbf{x}, y)$ by replacing π_d by $\hat{\pi}_d$; define $\hat{\kappa}_i = \hat{\kappa}(\mathbf{X}_i, Y_i)$

Step 5: define $\hat{f}_{di} = \hat{f}_{D|X,Y}(d, \mathbf{X}_i, Y_i) = N_d\, H(d, \mathbf{X}_i, Y_i)\hat{\kappa}_i/(N\hat{\pi}_d)$

Step 6: for any function $h(d, \mathbf{x}, y)$ in expression (6), estimate $E\{h(D, \mathbf{X}, Y)|\mathbf{X}, D = d\}$ by non-parametric regression between observations with $D_i = d$

Step 7: for any function $h(d, \mathbf{x}, y)$ in expression (6), estimate $E\{h(D, \mathbf{X}, Y)|D = d\}$ as $\hat{E}\{h(D, \mathbf{X}, Y)|D = d\} = \Sigma_{i=1}^N h(d, \mathbf{X}_i, Y_i)\hat{f}_{di}/\Sigma_{i=1}^N \hat{f}_{di}$

Step 8: for any function $h(d, \mathbf{x}, y)$ in expression (6), estimate $E\{h(D, Y, \mathbf{X})|\varepsilon, \mathbf{X}\}$ by $\hat{E}\{h(D, Y, \mathbf{X})|\varepsilon, \mathbf{X}\} = \Sigma_{d=0}^1 N_d\, H(d, \mathbf{X}, Y)\, h(d, Y, \mathbf{X})\, \hat{\kappa}(\mathbf{X}, Y)/(N\hat{\pi}_d)$

Step 9: for any function $h(d, \mathbf{x}, y)$ in expression (6), estimate $E_{\text{true}}\{h(D, \mathbf{X}, Y)|\mathbf{X}\}$ by $\hat{E}_{\text{true}}\{h(D, \mathbf{X}, Y)|\mathbf{X}\} = \Sigma_{d=0}^1 \hat{\pi}_d\, \hat{E}\{h(d, \mathbf{X}, Y)|\mathbf{X}, D = d\}\, \hat{f}_{\mathbf{X}|D}(\mathbf{X}, d)/\Sigma_{d=0}^1 \hat{\pi}_d \hat{f}_{\mathbf{X}|D}(\mathbf{X}, d)$

†Steps 1 and 2 are done only once. The rest of the steps are done iteratively in the estimation algorithm. Application to the terms in expression (6) yields $\hat{g}(\varepsilon_i, \mathbf{X}_i)$ and $\hat{\mathbf{v}}_d$, and we then form $\hat{\mathbf{S}}^*_{\text{eff}}(D, \mathbf{X}_i, Y_i) = \mathbf{S}^*(\mathbf{X}, Y, D) - \hat{\mathbf{g}}(\varepsilon, \mathbf{X}) - (1 - D)\hat{\mathbf{v}}_0 - D\hat{\mathbf{v}}_1$. We have described the algorithm when $\mathbf{X}$ is continuous. When $\mathbf{X}$ is discrete, one simply replaces the density estimators and various non-parametric regressions with the corresponding averages associated with the different $\mathbf{x}$-values.

If we estimate the last term by

$$\sum_{i=1}^N \frac{N_d\, H(d, \mathbf{X}_i, Y_i)}{N\pi_d} \left\{ \sum_d \frac{N_d\, H(d, \mathbf{X}_i, Y_i)}{N\pi_d} \right\}^{-1}$$

and remember that $\pi_0 + \pi_1 = 1$, we see that we can estimate π_0 by solving

$$\pi_0 = \sum_{i=1}^N H(0, \mathbf{X}_i, Y_i) \left[\frac{\sum_d N_d\, H(d, \mathbf{X}_i, Y_i)}{\{\pi_0^{1-d}(1-\pi_0)\}\pi_d} \right]^{-1}.$$

(b) Next we have that

$$\begin{aligned} E(\mathbf{h}|\varepsilon, \mathbf{X}) &= \sum_d \mathbf{h} f_{D|X,Y}(d, \mathbf{X}, Y) \\ &= \sum_d \frac{N_d\, H(d, \mathbf{X}, Y)\, \mathbf{h}(d, \mathbf{X}, Y)}{N\pi_d} \left\{ \sum_d \frac{N_d\, H(d, \mathbf{X}, Y)}{N\pi_d} \right\}^{-1}. \end{aligned}$$

(c) In addition,

$$\begin{aligned} E_{\text{true}}(\mathbf{h}|\mathbf{X}) &= \frac{\int \mathbf{h} \sum_d \pi_d\, f^{\text{true}}_{X,Y|D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)}{\int \sum_d \pi_d\, f^{\text{true}}_{X,Y|D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)} = \frac{\int \mathbf{h} \sum_d \pi_d\, f_{X,Y|D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)}{\int \sum_d \pi_d f_{X,Y|D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)} \\ &= \frac{\sum_d \pi_d \int \mathbf{h} f_{X,Y|D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)}{\sum_d \pi_d\, f_{\mathbf{X}|D}(\mathbf{X}, d)} = \frac{\sum_d \pi_d \int \mathbf{h} f_{Y|\mathbf{X},D}(\mathbf{X}, y, d)\, \mathrm{d}\mu(y)\, f_{\mathbf{X}|D}(\mathbf{X}, d)}{\sum_d \pi_d f_{\mathbf{X}|D}(\mathbf{X}, d)} \\ &= \sum_d \pi_d\, E(\mathbf{h}|\mathbf{X}, d)\, f_{\mathbf{X}|D}(\mathbf{X}, d) \Big/ \sum_d \pi_d\, f_{\mathbf{X}|D}(\mathbf{X}, d), \end{aligned}$$

where, in the last expression, both $f_{\mathbf{X}|D}(\mathbf{x},d)$ and $E(\mathbf{h}|\mathbf{x},d)$ need to be estimated non-parametrically.

(d) Finally, we have

$$E(\mathbf{h}|D=d)=\frac{\int f_{\mathbf{x},y}(\mathbf{x},y)\,\mathbf{h}(d,\mathbf{x},y)\,f_{D|X,Y}(d,\mathbf{x},y)\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y)}{\int f_{\mathbf{x},y}(\mathbf{x},y)\,f_{D|X,Y}(d,\mathbf{x},y)\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(y)},$$

which can be estimated as

$$\hat{E}(\mathbf{h}|D=d)=\sum_{i=1}^{N}\mathbf{h}(d,\mathbf{X}_i,Y_i)\,f_{D|X,Y}(d,\mathbf{X}_i,Y_i)\bigg/\sum_{i=1}^{N}f_{D|X,Y}(d,\mathbf{X}_i,Y_i).$$

4.2. Distribution theory

Because the locally efficient estimator is derived from well-established semiparametric procedures, while replacing the unknown quantities with non-parametric estimation in the model proposed, it is not surprising that it is asymptotically normally distributed with standard parametric rates of convergence. In addition, it achieves the semiparametric efficiency if the model proposed is correct. We describe the asymptotic properties of our estimator in theorem 1, and we provide a sketch of the proof for theorem 1 in Appendix A. We first list the set of regularity conditions that theorem 1 requires.

Condition 1. There are constants $0<C<\infty$ such that $\lim_{N\to\infty}N_1/N_2=C$. In addition, the identifiability assumptions 1 and 2 hold.

Condition 2. The univariate kernel function is a function that integrates to 1 and has support $(-1,1)$ and order r, i.e. $\int K(x)x^t\,\mathrm{d}x=0$ if $1\leqslant t<r$ and $\int K(x)x^r\,\mathrm{d}x\neq 0$. The d-dimensional kernel function, which is still represented by K, is a product of d univariate kernel functions, i.e. $K(\mathbf{x})=\Pi_{i=1}^{d}K(x_i)$ for a d-dimensional $\mathbf{x}$.

Condition 3. For $d=1,0$, $f_{X|D}(\mathbf{x}|D=d)$, $E(\varepsilon^2\kappa|\mathbf{X},D=d)$, $E(\varepsilon\mu_s|\mathbf{X},D=d)$ and $E(\varepsilon f_0|\mathbf{X},D=d)$ and $E(\varepsilon f_1|\mathbf{X},D=d)$ have compact support and have continuous rth derivatives.

Condition 4. The bandwidth $h=N^{-\tau}$ where $1/(2p)>\tau>1/(4r)$, with p the dimension of $\mathbf{x}$. This includes the optimal bandwidth $h=O(N^{-1/(2r+p)})$ as long as we choose a kernel of order $2r>p$.

Condition 1 ensures that there are a sufficient number of both cases and controls in the sample, which occurs in all case–control studies of the type that we are studying (see the introductory paragraph). Conditions 2 and 4 are standard requirements on an rth-order kernel function and on the bandwidth in the kernel smoothing literature (Ma and Zhu, 2013). Condition 3 is not the weakest possible. We impose this condition to simplify the technical proof. It can be replaced with weaker conditions in the region where $||\mathbf{x}||$ is large, at the expense of a more tedious technical treatment.

Theorem 1. We emphasize that, for any random vector $\mathbf{S}(D,Y,X)$, expectation and covariance in the superpopulation are linked to expectation and covariance in the case–control sampling scheme (conditional on disease status) through

$$E\{\mathbf{S}(D,Y,X)\}=\sum_{d=0}^{1}(N_d/N)\,E\{\mathbf{S}(D,Y,X)|D=d\},$$

$$\mathrm{cov}\{\mathbf{S}(D,Y,X)\}=\sum_{d=0}^{1}(N_d/N)\,\mathrm{cov}\{\mathbf{S}(D,Y,X)|D=d\}.$$

Under the regularity conditions 1–4, in the case–control study, as $N\to\infty$, the estimator $\hat{\boldsymbol{\theta}}$ that is obtained from solving the estimating equation $\Sigma_{i=1}^{N}\hat{\mathbf{S}}^{*}_{\mathrm{eff}}(D_i,\mathbf{X}_i,Y_i,\hat{\boldsymbol{\theta}})=0$ satisfies

$$N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to N\{\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\mathrm{T}}\}$$

where $\mathbf{A}=E\{\partial\mathbf{S}^{*}_{\mathrm{eff}}(D,\mathbf{X},Y,\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}^{\mathrm{T}}\}$ and $\mathbf{B}=\mathrm{cov}\{\mathbf{S}^{*}_{\mathrm{eff}}(D,\mathbf{X},Y,\boldsymbol{\theta}_0)\}$.

5. Simulations

5.1. Set-up

We performed a series of simulation studies to evaluate the finite sample performance of the various methods. In total, we considered 72 different cases. First, we considered a balanced design, where $N_0=N_1=500$, and an imbalanced design with $N_0=666$ and $N_1=334$, i.e. two controls for every case. Second, we considered three disease rates: a relatively rare disease rate of 4.5%, an extremely rare disease rate of 0.5% and a common disease rate of 10%. The balanced design in rare or extremely rare disease cases is representative of a typical case–control study.

Third, we considered three settings for the logistic regression. We generated X from a uniform(0,1) distribution. The logistic regression model was $\mathrm{pr}(D=1|Y,X)=H(\alpha_c+\alpha_1X+\alpha_2Y)$, where $\alpha_1=1$ and we varied $\alpha_2=0.00,0.25,0.50$. The regression model for Y given X is $Y=\beta_1+\beta_2X+\varepsilon$, with $\beta_1=0$ and $\beta_2=1$.

Finally, we varied the distribution of the regression errors and whether they were homoscedastic or not, as follows.

(a) In the first set of simulations, we generated homoscedastic errors ε. The distribution of ε was either $N(0,\sigma^2)$ with $\sigma^2=1$ or is a centred and standardized gamma distribution with shape parameter 0.4, normalized to have mean 0 and variance $\sigma^2=1$. To achieve an approximate 4.5% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-3.6, -3.8, -4.0)$. To achieve an approximate 0.5% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-5.8, -6.0, -6.2)$. To achieve an approximate 10% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-2.7, -2.9, -3.1)$.

(b) In the second set of simulations, we generated heteroscedastic errors as follows. The same distributions for ε were used, except that ε was multiplied by $(1+X^2)^{3/4}/2$ in all the cases, so that $\mathrm{var}(\varepsilon|X)=(1+X^2)^{3/2}/4$. To achieve an approximate 4.5% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-3.60, -3.75, -3.95)$. To achieve an approximate 0.5% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-5.8, -5.95, -6.2)$. To achieve an approximate 10% disease rate, for $\alpha_2=(0.00,0.25,0.50)$ we set $\alpha_c=(-2.7, -2.9, -3.1)$.

With respect to the method that was described in Section 4.1, we mention the following details. The posited model η_2^* is a standard normal model in step 1. This yields the second component in S^* as $(y-\beta_1-\beta_2x)(1,x)^{\mathrm{T}}$. In performing the many non-parametric calculations in steps 4, 5, 6 and 7, we used a kernel estimate with the same bandwidth h throughout. We set the bandwidth at $h=cn_0^{-1/3}$ and experimented with different values c between $c=0.5$ and $c=2.0$, with little change in the results. To assess variability, we used the asymptotic results in theorem 1, with the **A**- and **B**-matrices replaced by their corresponding sample averages evaluated at the estimated parameter values.

We compared our method with three others. The first was ordinary least squares among the controls, with sandwich standard errors: the sandwich method is used to adjust confidence intervals for possible heteroscedasticity. The second was the semiparametric efficient method that assumes normality and homoscedasticity, with standard errors obtained by inverting the Hessian of the log-likelihood (Lin and Zeng, 2009). The third was the method of Wei *et al.* (2013) that assumes homoscedasticity but otherwise does not specify any particular error distribution model: we used the bootstrap to obtain standard errors for this method.

A striking conclusion of these simulations is that our method, which assumes none of rare disease, normal errors or homoscedasticity, uniformly has coverage probabilities that achieve the nominal rates.

5.2. *Homoscedastic case*

Results for the homoscedastic case are given in Tables 2–4. We display the mean estimate, the standard deviation across the simulations, the mean estimated standard deviation, coverage probabilities for nominal 90% and 95% confidence intervals and the mean-squared error efficiency of the methods relative to using only the controls.

Table 2. Results of the simulation study with $n_1 = 500$ cases and $n_0 = 500$ controls and a disease rate of approximately 4.5%, with homoscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
$\alpha_2 = 0.00$								
Mean	0.998	0.998	1.001	1.008	0.996	1.000	0.997	1.001
sd	0.151	0.110	0.114	0.109	0.155	0.110	0.120	0.110
Estimated sd	0.155	0.110	0.122	0.130	0.154	0.110	0.122	0.116
90%	0.903	0.900	0.921	0.910	0.898	0.894	0.910	0.912
95%	0.952	0.955	0.957	0.959	0.958	0.954	0.956	0.959
MSE efficiency		1.878	1.734	1.909		1.966	1.663	1.987
$\alpha_2 = 0.25$								
Mean	0.980	0.983	0.976	0.998	0.977	0.962	0.961	0.993
sd	0.151	0.113	0.116	0.113	0.151	0.139	0.115	0.093
Estimated sd	0.154	0.111	0.119	0.115	0.148	0.140	0.120	0.103
90%	0.906	0.878	0.895	0.900	0.895	0.902	0.895	0.912
95%	0.947	0.939	0.953	0.966	0.939	0.948	0.943	0.963
MSE efficiency		1.785	1.663	1.816		1.129	1.599	2.682
$\alpha_2 = 0.50$								
Mean	0.974	0.969	0.946	0.992	0.954	0.799	0.958	1.002
sd	0.146	0.106	0.119	0.116	0.139	0.179	0.133	0.099
Estimated sd	0.154	0.112	0.122	0.126	0.139	0.173	0.132	0.103
90%	0.918	0.909	0.884	0.915	0.885	0.681	0.892	0.917
95%	0.961	0.955	0.943	0.964	0.934	0.787	0.943	0.961
MSE efficiency		1.780	1.270	1.627		0.295	1.092	2.186

†Here 'normal' means that $\varepsilon = N(0, 1)$, whereas 'gamma' means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only ('Controls'), the semiparametric efficient method that assumes normality and homoscedasticity ('Param'), the method of Wei *et al.* (2013) ('Robust') and our method ('Semi'). Over 1000 simulations, we computed the mean estimated β ('Mean'), its standard deviation ('sd'), the mean estimated standard deviation ('Estimated sd'), the coverage for a nominal 90% confidence interval ('90%'), the coverage for a nominal 95% confidence interval ('95%') and the mean-squared error efficiency compared with using only the controls ('MSE efficiency').

Table 3. Results of the simulation study with $n_1 = 500$ cases and $n_0 = 500$ controls, $\alpha_2 = 0.5$ and homoscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
Disease rate 10%								
Mean	0.913	0.876	0.784	0.979	0.885	0.885	0.929	0.993
sd	0.120	0.121	0.159	0.117	0.124	0.124	0.108	0.109
Estimated sd	0.119	0.123	0.154	0.117	0.153	0.126	0.110	0.109
90%	0.806	0.746	0.600	0.893	0.870	0.792	0.847	0.897
95%	0.867	0.837	0.723	0.956	0.926	0.891	0.908	0.948
MSE efficiency		0.731	0.305	1.554		0.951	1.628	2.279
Disease rate 0.5%								
Mean	0.991	0.996	0.987	1.010	0.978	0.854	1.029	0.991
sd	0.165	0.114	0.118	0.121	0.148	0.231	0.155	0.097
Estimated sd	0.155	0.112	0.120	0.122	0.149	0.223	0.160	0.096
90%	0.876	0.893	0.904	0.898	0.902	0.830	0.904	0.895
95%	0.925	0.942	0.949	0.938	0.945	0.904	0.950	0.945
MSE efficiency		2.099	1.938	1.852		0.300	0.900	2.359

†Here 'normal' means that $\varepsilon = N(0, 1)$, whereas 'gamma' means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only ('Controls'), the semiparametric efficient method that assumes normality and homoscedasticity ('Param'), the method of Wei *et al.* (2013) ('Robust') and our method ('Semi'). Over 1000 simulations, we computed the mean estimated β ('Mean'), its standard deviation ('sd'), the mean estimated standard deviation ('Estimated sd'), the coverage for a nominal 90% confidence interval ('90%'), the coverage for a nominal 95% confidence interval ('95%') and the mean-squared error efficiency compared with using only the controls ('MSE efficiency').

The case $\alpha_2 = 0.00$ is interesting, because here Y is independent of D given X. Hence, all methods should achieve nominal coverage probabilities for estimating β, which is indeed seen in Table 2. Surprisingly, our method, which assumes neither normality nor homoscedasticity, is as efficient in terms of mean-squared error as the semiparametric efficient method that assumes both, and is of course much more efficient than using only the controls.

For $\alpha_2 \neq 0$, and when ε is normally distributed, our method remains comparably as efficient as the semiparametric efficient method which assumes both normality and homoscedasticity. However, when the errors were not normally distributed, our method has much smaller bias and is much more efficient. In addition, the semiparametric efficient method has poor coverage probabilities when $\alpha_2 = 0.50$. Although the method of Wei *et al.* (2013) maintains good coverage probabilities in all cases, our method also maintains coverage, has smaller bias and is much more efficient.

5.3. Heteroscedastic case

The results for the heteroscedastic case, with various disease rates and equal or unequal case–control ratios, are given in Tables 5–7.

The results are much in line with the homoscedastic case, with a few important exceptions. The semiparametric efficient method, which assumes both homoscedasticity and normality, has a noticeable loss of coverage probability when $\alpha_2 \neq 0$, largely caused by bias. Because they

Table 4. Results of the simulation study with $n_1 = 334$ cases and $n_0 = 666$ controls, $\alpha_2 = 0.5$ and homoscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
Disease rate 4.5%								
Mean	0.962	0.960	0.956	0.994	0.951	0.856	0.936	0.996
sd	0.133	0.106	0.108	0.113	0.128	0.153	0.123	0.101
Estimated sd	0.133	0.110	0.113	0.121	0.120	0.152	0.120	0.108
90%	0.892	0.884	0.893	0.901	0.845	0.751	0.844	0.916
95%	0.957	0.943	0.954	0.952	0.925	0.848	0.910	0.960
MSE efficiency		1.491	1.407	1.494		0.426	0.977	1.839
Disease rate 10%								
Mean	0.921	0.850	0.831	0.991	0.937	0.879	0.927	1.060
sd	0.106	0.114	0.134	0.082	0.129	0.117	0.107	0.082
Estimated sd	0.103	0.113	0.136	0.080	0.133	0.117	0.110	0.077
90%	0.797	0.621	0.673	0.900	0.872	0.739	0.840	0.908
95%	0.881	0.752	0.780	0.949	0.932	0.845	0.909	0.949
MSE efficiency		0.492	0.375	2.568		0.727	1.228	1.996
Disease rate 0.5%								
Mean	1	0.997	0.991	1.004	0.997	0.901	1.018	1.000
sd	0.133	0.107	0.113	0.110	0.129	0.191	0.134	0.100
Estimated sd	0.134	0.111	0.111	0.113	0.130	0.190	0.142	0.099
90%	0.904	0.911	0.894	0.904	0.890	0.858	0.925	0.897
95%	0.944	0.959	0.943	0.945	0.947	0.921	0.966	0.953
MSE efficiency		1.544	1.377	1.460		0.360	0.911	1.665

†Here ‘normal’ means that $\varepsilon = N(0, 1)$, whereas ‘gamma’ means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only (‘Controls’), the semiparametric efficient method that assumes normality and homoscedasticity (‘Param’), the method of Wei *et al.* (2013) (‘Robust’) and our method (‘Semi’). Over 1000 simulations, we computed the mean estimated β (‘Mean’), its standard deviation (‘sd’), the mean estimated standard deviation (‘Estimated sd’), the coverage for a nominal 90% confidence interval (‘90%’), the coverage for a nominal 95% confidence interval (‘95%’) and the mean-squared error efficiency compared with using only the controls (‘MSE efficiency’).

used a bootstrap to compute standard errors, the method of Wei *et al.* (2013) maintains good coverage probability except when $\alpha_2 = 0.50$, where the bias causes deterioration in the coverage rates. Our method maintains good coverage probabilities in all cases and, because of its lack of bias, noticeably increased mean-squared error efficiency.

6. Empirical example

Epidemiological studies have led to the general belief that HCAs, such as MeIQx and PhIP, are significant risk factors associated with various forms of cancers, including colorectal cancer and breast cancer (Barrett *et al.*, 2003; Sinha *et al.*, 2001; De Stefani *et al.*, 1997). One of the important food sources contributing to carcinogenic HCAs, among many other potential sources, is red meat, which produces the agents during the cooking process. In addition, red meat contains other nutrients such as saturated fat which is also believed to relate to the occurrence of cancer. Because of this link, epidemiological and nutritional studies of cancer

Table 5. Results of the simulation study with $n_1 = 500$ cases and $n_0 = 500$ controls and a disease rate of approximately 4.5%, with heteroscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
$\alpha_2 = 0.00$								
Mean	0.996	0.996	1.000	1.005	0.992	0.994	1.000	1.002
sd	0.099	0.071	0.071	0.076	0.099	0.070	0.073	0.077
Estimated sd	0.096	0.070	0.072	0.082	0.096	0.070	0.071	0.078
90%	0.887	0.892	0.895	0.898	0.887	0.903	0.893	0.898
95%	0.932	0.953	0.949	0.950	0.944	0.946	0.947	0.951
MSE efficiency		1.948	1.961	1.692		1.971	1.847	1.663
$\alpha_2 = 0.25$								
Mean	0.986	1.044	0.973	0.997	0.983	1.063	0.964	0.995
sd	0.100	0.072	0.066	0.077	0.094	0.082	0.069	0.071
Estimated sd	0.096	0.071	0.070	0.081	0.094	0.083	0.072	0.074
90%	0.880	0.838	0.907	0.912	0.894	0.825	0.863	0.904
95%	0.936	0.907	0.953	0.959	0.946	0.900	0.934	0.950
MSE efficiency		1.415	1.984	1.717		0.852	1.516	1.801
$\alpha_2 = 0.50$								
Mean	0.972	1.088	0.949	0.991	0.962	1.145	0.906	0.993
sd	0.099	0.072	0.068	0.083	0.095	0.096	0.076	0.082
Estimation sd	0.096	0.072	0.071	0.102	0.090	0.100	0.076	0.105
90%	0.877	0.664	0.842	0.897	0.857	0.591	0.655	0.900
95%	0.936	0.789	0.914	0.946	0.909	0.714	0.756	0.935
MSE efficiency		0.816	1.479	1.519		0.343	0.717	1.546

†Here 'normal' means that $\varepsilon = N(0, 1)$, whereas 'gamma' means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only ('Controls'), the semiparametric efficient method that assumes normality and homoscedasticity ('Param'), the method of Wei *et al.* (2013) ('Robust') and our method ('Semi'). Over 1000 simulations, we computed the mean estimated β ('Mean'), its standard deviation ('sd'), the mean estimated standard deviation ('Estimated sd'), the coverage for a nominal 90% confidence interval ('90%'), the coverage for a nominal 95% confidence interval ('95%') and the mean-squared error efficiency compared with using only the controls ('MSE efficiency').

often include both red meat consumption and HCA as covariates to assess the risk of developing cancer, while simultaneously studying the relationship between HCA amount and red meat consumption. Understanding this relationship helps to understand the health effect of red meat consumption and is important in formulating food consumption guidelines for the general public.

We implemented our method on a data set involving colorectal adenoma, with 640 cases and 665 controls. The cases and controls were defined by the occurrence of colorectal adenoma, D. In our analysis, X is red meat consumption in grams. We used two different versions of Y, namely the HCAs MeIQx and PhIP that are produced during the cooking of meat.

PhIP, MeIQx and red meat were transformed by adding 1.0 and taking logarithms to alleviate the heavy skewness of these measurements on the original scale. We also analysed the subset of the study who were smokers. For the controls-only analysis, standard errors of the slope estimate were computed by using the usual formula for least squares and also by the sandwich method. For our semiparametric analysis, we computed standard errors by the asymptotic formula of

Table 6. Results of the simulation study with $n_1 = 500$ cases and $n_0 = 500$ controls, $\alpha_2 = 0.5$ and heteroscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
Disease rate 10%								
Mean	0.905	0.897	1.078	0.990	0.950	0.931	1.065	1.001
sd	0.083	0.073	0.091	0.117	0.101	0.073	0.071	0.108
Estimated sd	0.083	0.072	0.089	0.115	0.100	0.072	0.072	0.111
90%	0.676	0.600	0.770	0.895	0.847	0.765	0.781	0.895
95%	0.766	0.698	0.850	0.947	0.914	0.851	0.859	0.955
MSE efficiency		0.998	1.107	1.154		1.258	1.370	1.089
Disease rate 0.5%								
Mean	0.997	1.113	0.973	1.007	0.991	1.296	0.890	0.995
sd	0.098	0.073	0.067	0.082	0.102	0.113	0.087	0.072
Estimated sd	0.101	0.072	0.070	0.088	0.098	0.112	0.084	0.071
90%	0.906	0.541	0.892	0.897	0.895	0.145	0.630	0.907
95%	0.951	0.663	0.957	0.942	0.937	0.231	0.745	0.941
MSE efficiency		0.531	1.842	1.419		0.104	0.533	2.013

†Here 'normal' means that $\varepsilon = N(0, 1)$, whereas 'gamma' means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only ('Controls'), the semiparametric efficient method that assumes normality and homoscedasticity ('Param'), the method of Wei *et al.* (2013) ('Robust') and our method ('Semi'). Over 1000 simulations, we computed the mean estimated β ('Mean'), its standard deviation ('sd'), the mean estimated standard deviation ('Estimated sd'), the coverage for a nominal 90% confidence interval ('90%'), the coverage for a nominal 95% confidence interval ('95%') and the mean-squared error efficiency compared with using only the controls ('MSE efficiency').

theorem 1 and by the bootstrap, with 1000 bootstrap samples. Given the results of the simulation, we do not expect any significant difference between these two estimates of standard errors for our method, with the asymptotic formula being much faster computationally.

We performed a preliminary analysis using only the controls. On the original data scale, all the covariates (PhIP, MeIPx and red meat consumption) are very skewed and heavy tailed; see Figs S.1 and S.2 in the on-line supplementary material. The transformed data were much better behaved; see Figs S.3 and S.4 in the supplementary material. Numerically, the skewnesses of MeIQx in the original and transformed data scales are 3.46 and −0.19 respectively. The skewnesses of PhIP on the original and transformed data scales are 7.93 and −0.20 respectively. Finally, the skewness of red meat on the original and transformed data scales are 1.78 and −0.58 respectively. These numbers and the plots indicate that the transformation did an acceptable to very good job of removing skewness.

Further preliminary analysis of the controls included scatter plots of the transformed data, both of which were reasonably well behaved and indicated an increasing trend for increasing red meat consumption, consistent with a linear trend; see Fig. S.5 in the on-line supplementary material. To check this, we fit a quadratic model to the transformed data: in both cases, the p-value for the quadratic term exceeded 0.20; see Fig. 2. Thus, we adopted a linear function for the mean $m(\cdot)$ in the subsequent secondary analysis. In addition, the regression of PhIP on red meat consumption is heavily heteroscedastic, whereas the regression of MeIQx on red meat is passably homoscedastic. This is shown in Fig. 3, where we fit a regression of the absolute

Table 7. Results of the simulation study with $n_1 = 334$ cases and $n_0 = 666$ controls, $\alpha_2 = 0.5$ and heteroscedastic errors†

	Results for the normal model and the following methods:				*Results for the gamma model and the following methods:*			
	Controls	*Param*	*Robust*	*Semi*	*Controls*	*Param*	*Robust*	*Semi*
Disease rate 4.5%								
Mean	0.977	1.052	0.961	0.996	0.961	1.085	0.926	0.994
sd	0.084	0.070	0.063	0.077	0.083	0.087	0.066	0.082
Estimated sd	0.087	0.072	0.064	0.083	0.08	0.087	0.067	0.090
90%	0.883	0.825	0.859	0.913	0.827	0.735	0.702	0.918
95%	0.939	0.892	0.930	0.952	0.905	0.831	0.806	0.954
MSE efficiency		0.998	1.382	1.276		0.568	0.855	1.244
Disease rate 10%								
Mean	0.911	0.909	1.021	1.001	0.956	0.937	1.027	1.000
sd	0.072	0.064	0.080	0.079	0.084	0.066	0.070	0.087
Estimated sd	0.072	0.065	0.080	0.076	0.087	0.065	0.072	0.094
90%	0.654	0.595	0.895	0.901	0.867	0.772	0.877	0.906
95%	0.749	0.700	0.949	0.951	0.927	0.851	0.933	0.952
MSE efficiency		1.058	1.915	2.099		1.080	1.597	1.188
Disease rate 0.5%								
Mean	0.997	1.073	0.979	1.007	0.994	1.189	0.920	0.997
sd	0.088	0.073	0.066	0.078	0.084	0.100	0.071	0.070
Estimated sd	0.087	0.072	0.063	0.086	0.085	0.099	0.073	0.069
90%	0.891	0.728	0.871	0.901	0.899	0.384	0.725	0.911
95%	0.950	0.820	0.929	0.952	0.953	0.539	0.829	0.960
MSE efficiency		0.727	1.616	1.264		0.155	0.620	1.445

†Here 'normal' means that $\varepsilon = N(0, 1)$, whereas 'gamma' means that ε is a centred and scaled gamma random variable with shape 0.4, mean 0 and variance 1. The analyses performed were using controls only ('Controls'), the semiparametric efficient method that assumes normality and homoscedasticity ('Param'), the method of Wei *et al.* (2013) ('Robust') and our method ('Semi'). Over 1000 simulations, we computed the mean estimated β ('Mean'), its standard deviation ('sd'), the mean estimated standard deviation ('Estimated sd'), the coverage for a nominal 90% confidence interval ('90%'), the coverage for a nominal 95% confidence interval ('95%') and the mean-squared error efficiency compared with using only the controls ('MSE efficiency').

residuals from a quadratic fit against red meat consumption (Davidian and Carroll, 1987): the plots from a linear regression are essentially the same.

The results of this secondary analysis are given in Table 8. For MeIQx, the ordinary least squares standard errors when using only the controls are roughly the same as that of the sandwich method, which makes sense since the regression is homoscedastic. In this case, as expected from the theory, our semiparametric approach has smaller standard errors, with the least squares standard errors being approximately 30% larger. For PhIP, where the regression is distinctly homoscedastic, the sandwich standard errors for ordinary least squares among the controls is roughly 30% larger than the standard error that assumes homoscedasticity, and roughly 40% larger than our semiparametric approach. As expected from the theory, where homoscedasticity is not assumed, the standard errors for our semiparametric approach are nearly the same by using either the asymptotic formula or the bootstrap.

As a comparison, we also implemented the parametric method of Lin and Zeng (2009) as well as the robust method of Wei *et al.* (2013). Standard errors of the former were assessed both

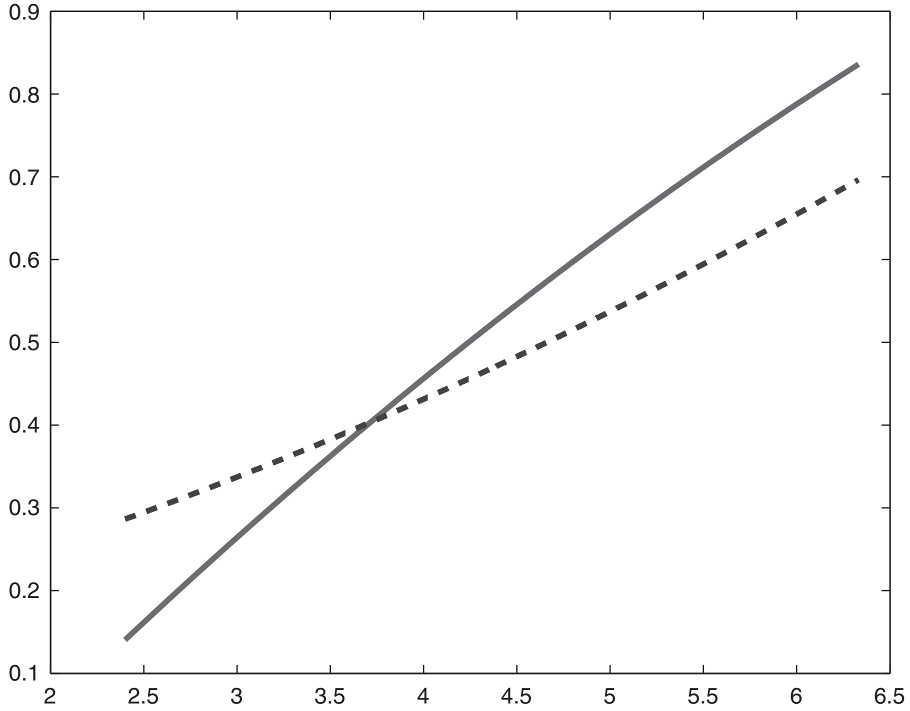

Fig. 2. Fitted curves from a regression of MeIQx (———) and PhIP (– – –) on red meat consumption, using the controls: the fitted values were normalized to fit on the same plot; neither has a statistically significant quadratic term

by using the inverse of the Hessian of the log-likelihood and by the bootstrap, whereas standard errors of the latter were assessed by the bootstrap alone. The parametric method's asymptotic standard error clearly underestimates the variability for PhIP when compared with the bootstrap: something that is expected because of the heteroscedasticity in PhIP. For MeIQx, where the error is homoscedastic, the parametric method, the robust method and our semiparametric approach are almost identical.

In summary, in analysing this data set, we verified the previous observation based on the control-only data that the regression error from MeIQx and red meat consumption is homoscedastic error, whereas that from the PhIP and red meat consumption is heteroscedastic error. Our analysis also verified the positive relationship between red meat consumption and these two forms of HCA, indicating that increased red meat consumption leads to increased levels of MeIQx and PhIP, both being risk factors for colorectal cancer. The first-order accuracy of the variability of the estimated slope for our method is validated though its nearly identical result with the bootstrap, and of course through the simulation results.

7. Discussion

We have developed a locally efficient semiparametric estimator for the secondary analysis of case–control studies, where only a mean model is specified to describe the relationship between the covariates. Despite this relatively weak assumption, we have shown that the problem is still identifiable under certain conditions. Through introducing the notion of a superpopulation, we can establish an estimation methodology via a conceptually tractable semiparametric procedure,

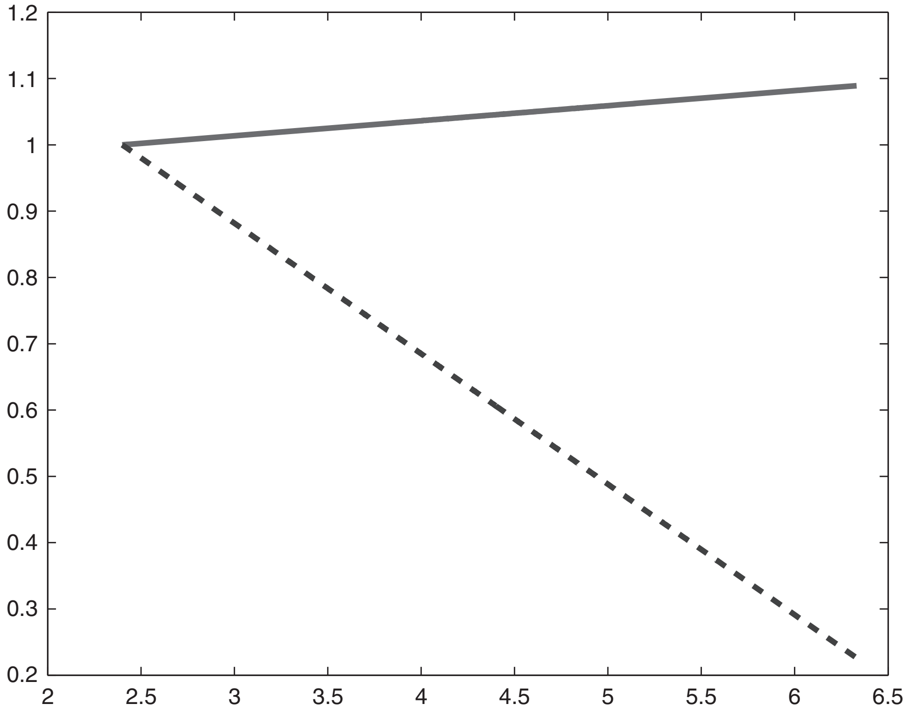

Fig. 3. Plots to diagnose heteroscedasticity, with the curves representing the relative standard deviation as a function of red meat consumption: plotted are the fitted curves from a linear regression of the absolute residuals of the regression of MeIQx (———) and PhIP (– – –) on red meat consumption, using the controls; the fitted values were normalized to be equal at the minimum value of red meat consumption; the essentially flat curve for MeIQx indicates homoscedasticity, whereas that for PhIP is very strongly heteroscedastic, which has implications for data analysis (see Table 7 and the discussion in Section 6)

although the derivation is highly non-standard and not trivial. The locally efficient estimator provides consistent estimation and can achieve optimal efficiency if a posited regression error model happens to be true. Although the analysis is performed under the superpopulation concept, the general statements of consistency and local efficiency are valid in the case–control sampling scheme (Ma, 2010). In addition, the general methodology is applicable even if the linear logistic model (1) is replaced by other parametric models such as a probit model, as long as identifiability can be established.

Implementing the locally efficient estimator via algorithm 1 requires several non-parametric regressions conditional on the covariates, which may be difficult when the dimension of the covariates increases. In such situations, dimension reduction techniques can be a good choice to achieve a balance between model flexibility and feasibility of parameter estimation and inference (Ma and Zhu, 2012). Further exploration of this is needed.

Acknowledgements

Ma's research was supported by National Science Foundation grant DMS-1206693 and National Institute of Neurological Disorders and Stroke grant R01-NS073671. Carroll's research was supported by National Cancer Institute grant R37-CA057030 and by National Institute of Neurological Disorders and Stroke grant R01-NS073671.

Table 8. Results of data analysis when Y is either MeIQx or PhIP†

HCA	*Results for controls-only method*			*Results for parametric method*		
	Estimate	*OLS se*	*Sandwich se*	*Estimate*	*Asymptotic se*	*Bootstrap se*
All data						
MeIQx	0.868	0.034	0.035	0.862	0.026	0.026
PhIP	0.742	0.064	0.080	0.751	0.046	0.056
	Results for robust method		*Results for semiparametric method*			
	Estimate	*Bootstrap se*	*Estimate*	*Asymptotic se*	*Bootstrap se*	
MeIQx	0.862	0.028	0.862	0.027	0.027	
PhIP	0.751	0.057	0.750	0.057	0.058	
	Results for controls-only method			*Results for parametric method*		
	Estimate	*OLS se*	*Sandwich se*	*Estimate*	*Asymptotic se*	*Bootstrap se*
Smokers only						
MeIQx	0.816	0.050	0.057	0.847	0.036	0.037
PhIP	0.619	0.095	0.132	0.737	0.063	0.080
	Results for robust method		*Results for semiparametric method*			
	Estimate	*Bootstrap se*	*Estimate*	*Asymptotic se*	*Bootstrap se*	
MeIQx	0.847	0.038	0.846	0.036	0.039	
PhIP	0.737	0.084	0.736	0.082	0.087	

†For the controls-only method, 'OLS se' is the ordinary least squares standard error estimate, whereas 'Sandwich se' is the sandwich method standard error estimate. For the parametric and semiparametric analysis, 'Asymptotic se' is the standard error estimate from asymptotic theory, whereas 'Bootstrap se' is the bootstrap standard error. For the robust analysis, only the bootstrap standard error is available. The regression of PhIP on red meat (X) is heteroscedastic, reflected in the difference between the ordinary least squares standard error and the sandwich standard error for the controls-only analysis, as well as the difference between the asymptotic standard error and the bootstrap standard error of the parametric estimator.

Appendix A: Sketch of technical arguments

A.1. *Proof of proposition 1*

Assume the contrary, i.e. assume that the problem is not identifiable. This means that we can find parameters $\alpha_c, \boldsymbol{\alpha}_1, \alpha_2, \boldsymbol{\beta}, \eta_2$ and η_1, and $\tilde{\alpha}_c, \tilde{\boldsymbol{\alpha}}_1, \tilde{\alpha}_2, \tilde{\boldsymbol{\beta}}, \tilde{\eta}_2$ and $\tilde{\eta}_1$ so that, denoting $\tilde{\varepsilon} = Y - m(\mathbf{x}, \tilde{\boldsymbol{\beta}})$,

$$\pi_d = \int \eta_1(\mathbf{x})\, \eta_2\{y - m(\mathbf{x}, \boldsymbol{\beta}), \mathbf{x}\} \frac{\exp\{\mathrm{d}\alpha_c + \mathrm{d}u(\mathbf{x}, y, \boldsymbol{\alpha}_1, \alpha_2,)\}}{1 + \exp\{\alpha_c + u(\mathbf{x}, y, \boldsymbol{\alpha}_1, \alpha_2)\}} \,\mathrm{d}\mu(\mathbf{x})\, \mathrm{d}\mu(y);$$

$$\tilde{\pi}_d = \int \tilde{\eta}_1(\mathbf{x})\, \tilde{\eta}_2\{y - m(\mathbf{x}, \tilde{\boldsymbol{\beta}}), \mathbf{x}\} \frac{\exp\{\mathrm{d}\tilde{\alpha}_c + \mathrm{d}u(\mathbf{x}, y, \tilde{\boldsymbol{\alpha}}_1, \tilde{\alpha}_2)\}}{1 + \exp\{\tilde{\alpha}_c + u(\mathbf{x}, y, \tilde{\boldsymbol{\alpha}}_1, \alpha_2)\}} \,\mathrm{d}\mu(\mathbf{x})\, \mathrm{d}\mu(y),$$

we have that

$$\frac{1}{\pi_d}\eta_1(\mathbf{x})\,\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}\frac{\exp\{\mathrm{d}\alpha_c+\mathrm{d}u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}$$
$$=\frac{1}{\tilde{\pi}_d}\tilde{\eta}_1(\mathbf{x})\,\tilde{\eta}_2\{y-m(\mathbf{x},\tilde{\boldsymbol{\beta}}),\mathbf{x}\}\frac{\exp\{\mathrm{d}\tilde{\alpha}_c+\mathrm{d}u(\mathbf{x},y,\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2)\}}{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2)\}} \tag{8}$$

for all $(\mathbf{x},y,d)$. Taking the ratio of this expression at $d=1$ and $d=0$, we obtain that, for all $(\mathbf{x},y)$,

$$\frac{\pi_0}{\pi_1}\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}=\frac{\tilde{\pi}_0}{\tilde{\pi}_1}\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2)\}.$$

This yields that $u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)-u(\mathbf{x},y,\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2)$ is a constant. Since the constant is 0 at $(\mathbf{x},y)=0$, hence we have $u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)-u(\mathbf{x},y,\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2)\equiv 0$. Thus, $\boldsymbol{\alpha}_1,\alpha_2=\tilde{\boldsymbol{\alpha}}_1,\tilde{\alpha}_2$, $\exp(\alpha_c)\pi_0/\pi_1=\exp(\tilde{\alpha}_c)\tilde{\pi}_0/\tilde{\pi}_1$ and

$$\frac{1}{\pi_0}\frac{\eta_1(\mathbf{x})\,\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}=\frac{1}{\tilde{\pi}_0}\frac{\tilde{\eta}_1(\mathbf{x})\,\tilde{\eta}_2\{y-m(\mathbf{x},\tilde{\boldsymbol{\beta}}),\mathbf{x}\}}{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}$$

for all $(\mathbf{x},y)$. This gives

$$\tilde{\eta}_1(\mathbf{x})\,\tilde{\eta}_2\{y-m(\mathbf{x},\tilde{\boldsymbol{\beta}}),\mathbf{x}\}=\frac{\tilde{\pi}_0}{\pi_0}\frac{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}\eta_1(\mathbf{x})\,\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}. \tag{9}$$

Integrating equation (9) and the product of equation (9) and y with respect to y, we obtain

$$\tilde{\eta}_1(\mathbf{x})=\frac{\tilde{\pi}_0}{\pi_0}\eta_1(\mathbf{x})\int\frac{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}\,\mathrm{d}y,$$
$$\tilde{\eta}_1(\mathbf{x})\,m(\mathbf{x},\tilde{\boldsymbol{\beta}})=\frac{\tilde{\pi}_0}{\pi_0}\eta_1(\mathbf{x})\int\frac{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}\,y\,\mathrm{d}y$$

respectively. Further taking ratios, we find

$$\int\frac{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}y\,\mathrm{d}y$$
$$=m(\mathbf{x},\tilde{\boldsymbol{\beta}})\int\frac{1+\exp\{\tilde{\alpha}_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}{1+\exp\{\alpha_c+u(\mathbf{x},y,\boldsymbol{\alpha}_1,\alpha_2)\}}\eta_2\{y-m(\mathbf{x},\boldsymbol{\beta}),\mathbf{x}\}\,\mathrm{d}y.$$

If $\alpha_c=\tilde{\alpha}_c$, then we obtain $m(\mathbf{x},\boldsymbol{\beta})=m(\mathbf{x},\tilde{\boldsymbol{\beta}})$; hence $\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}$. We also obtain $\tilde{\eta}_1(\mathbf{x})=\eta_1(\mathbf{x})\tilde{\pi}_0/\pi_0$. Since both $\tilde{\eta}_1(\mathbf{x})$ and $\eta_1(\mathbf{x})$ are valid density functions, we have $\tilde{\eta}_1(\mathbf{x})=\eta_1(\mathbf{x})$ and $\pi_0=\tilde{\pi}_0$ and $\pi_1=\tilde{\pi}_1$. This subsequently yields $\eta_2=\tilde{\eta}_2$, contradicting our assumptions. Thus we obtain that $\alpha_c\neq\tilde{\alpha}_c$.

Denote

$$r(\varepsilon,\mathbf{x})=\frac{1+\exp[\tilde{\alpha}_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\{\varepsilon-m(\mathbf{x},\tilde{\boldsymbol{\beta}})+m(\mathbf{x},\boldsymbol{\beta})\}$$
$$=\exp(\tilde{\alpha}_c-\alpha_c)\{\varepsilon-m(\mathbf{x},\tilde{\boldsymbol{\beta}})+m(\mathbf{x},\boldsymbol{\beta})\}+\{1-\exp(\tilde{\alpha}_c-\alpha_c)\}\frac{\varepsilon-m(\mathbf{x},\tilde{\boldsymbol{\beta}})+m(\mathbf{x},\boldsymbol{\beta})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}.$$

By definition, η_2 is a valid conditional density function and it satisfies $\int\varepsilon\,\eta_2(\varepsilon,\mathbf{x})\,\mathrm{d}\varepsilon=0$, and we have that

$$0=\int r(\varepsilon,\mathbf{x})\,\eta_2(\varepsilon,\mathbf{x})\,\mathrm{d}\varepsilon=-\exp(\tilde{\alpha}_c-\alpha_c)\{m(\mathbf{x},\tilde{\boldsymbol{\beta}})-m(\mathbf{x},\boldsymbol{\beta})\}$$
$$+\{1-\exp(\tilde{\alpha}_c-\alpha_c)\}\int\frac{\varepsilon-m(\mathbf{x},\tilde{\boldsymbol{\beta}})+m(\mathbf{x},\boldsymbol{\beta})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\eta_2(\varepsilon,\mathbf{x})\,\mathrm{d}\varepsilon$$

for all $\mathbf{x}$. This means that

$$
\begin{aligned}
\frac{\{m(\mathbf{x},\tilde{\boldsymbol{\beta}})-m(\mathbf{x},\boldsymbol{\beta})\}\exp(\tilde{\alpha}_c-\alpha_c)}{1-\exp(\tilde{\alpha}_c-\alpha_c)} = & \int \frac{\varepsilon\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \\
& -\int \frac{\{m(\mathbf{x},\tilde{\boldsymbol{\beta}})-m(\mathbf{x},\boldsymbol{\beta})\}\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon
\end{aligned}
$$

for all $\mathbf{x}$. If we let $\mathbf{x}\rightarrow\mathbf{c}_1$, then

$$
\begin{aligned}
\frac{c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\exp(\tilde{\alpha}_c-\alpha_c)}{1-\exp(\tilde{\alpha}_c-\alpha_c)} = & \, c_2\int_{\mathcal{D}^c}\varepsilon\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon - c_2 c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\int_{\mathcal{D}^c}\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon \\
& + \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{\varepsilon\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \\
& - \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \\
= & -c_2 c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}}) - c_2\int_{\mathcal{D}}\varepsilon\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon + c_2 c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\int_{\mathcal{D}}\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon \\
& + \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{\varepsilon\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \\
& - \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\left|\frac{\exp(\tilde{\alpha}_c-\alpha_c)}{1-\exp(\tilde{\alpha}_c-\alpha_c)}+c_2\right| = & \left| -c_2\int_{\mathcal{D}}\frac{\varepsilon\eta_2(\varepsilon,\mathbf{c}_1)}{c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})}\mathrm{d}\varepsilon + c_2\int_{\mathcal{D}}\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon \right. \\
& + \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{\varepsilon\eta_2(\varepsilon,\mathbf{x})/c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \\
& \left. - \lim_{\mathbf{x}\rightarrow\mathbf{c}_1}\int_{\mathcal{D}}\frac{\eta_2(\varepsilon,\mathbf{x})}{1+\exp[\alpha_c+u\{\mathbf{x},m(\mathbf{x},\boldsymbol{\beta})+\varepsilon,\boldsymbol{\alpha}_1,\alpha_2\}]}\mathrm{d}\varepsilon \right| \\
\leqslant & \frac{2}{|c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})|}\int_{\mathcal{D}}|\varepsilon|\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon + 2\int_{\mathcal{D}}\eta_2(\varepsilon,\mathbf{c}_1)\mathrm{d}\varepsilon \\
\leqslant & \frac{2}{|c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})|}[E\{\varepsilon^2 I(\varepsilon\in\mathcal{D})|\mathbf{c}_1\}\mathrm{pr}(\varepsilon\in\mathcal{D}|\mathbf{c}_1)]^{1/2} + 2\,\mathrm{pr}(\varepsilon\in\mathcal{D}|\mathbf{c}_1) \\
\leqslant & \frac{2}{|c(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})|}\{E(\varepsilon^2)\delta\}^{1/2} + 2\delta.
\end{aligned}
$$

We can make the upper bound of this expression arbitrarily small by choosing δ arbitrarily close to 0, whereas the quantity on the left-hand side is a constant. Hence we in fact have obtained

$$\frac{\exp(\tilde{\alpha}_c-\alpha_c)}{1-\exp(\tilde{\alpha}_c-\alpha_c)} = -c_2.$$

However, $-c_2$ is between -1 and 0; simple calculation shows that these two constants cannot be equal, and hence our problem is indeed identifiable.

A.2. Derivation of Λ and $\Lambda^{\perp}$

Considering the nuisance tangent space associated with η_1 and η_2 respectively, we have

$$
\begin{aligned}
\Lambda_1 &= \{\mathbf{g}(\mathbf{x}) - E(\mathbf{g}|d) : \forall\,\mathbf{g} \text{ such that } E_{\text{true}}(\mathbf{g}) = \mathbf{0}\}, \\
\Lambda_2 &= \{\mathbf{g}(\varepsilon,\mathbf{x}) - E(\mathbf{g}|d) : \forall\,\mathbf{g} \text{ such that } E_{\text{true}}(\mathbf{g}|\mathbf{X}) = E_{\text{true}}(\varepsilon\mathbf{g}|\mathbf{X}) = \mathbf{0} \text{ almost surely}\}.
\end{aligned}
$$

Hence $\Lambda=\Lambda_1+\Lambda_2=\{\mathbf{g}(\varepsilon,\mathbf{x})-E(\mathbf{g}|d):\forall\,\mathbf{g}\text{ such that }E_{\text{true}}(\mathbf{g})=E_{\text{true}}(\varepsilon\mathbf{g}|\mathbf{X})=0\text{ almost surely}\}$. It is easily seen that $\Lambda_1^{\perp}=[\mathbf{h}:E(\mathbf{h})=0,\,E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{X}\}=\mathbf{0}$ almost surely]. This is because from

$$\begin{aligned}0=&E[\mathbf{h}^{\mathrm{T}}\{\mathbf{g}(\mathbf{X})-E(\mathbf{g}|D)\}]\\=&E[(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}\{\mathbf{g}(\mathbf{X})-E(\mathbf{g}|D)\}]\\=&E\{(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}\mathbf{g}\}\\=&E[E\{(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}|\mathbf{X}\}\mathbf{g}],\end{aligned}$$

we obtain $E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{X}\}\,\Sigma_d\int f_{X,Y,D}(\mathbf{X},y,d)\,\mathrm{d}\mu(y)/\eta_1(\mathbf{X})=\mathbf{c}$ almost surely for some constant $\mathbf{c}$. Since $E[E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{X}\}]=\mathbf{0}$ almost surely, we obtain

$$\mathbf{0}=\int E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{x}\}\sum_d\int f_{X,Y,D}(\mathbf{x},y,d)\,\mathrm{d}\mu(y)\,\mathrm{d}\mu(\mathbf{x})=\int\mathbf{c}\,\eta_1(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x})=\mathbf{c}\qquad\text{almost surely.}$$

Hence $\mathbf{c}=\mathbf{0}$ and $E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{X}\}\,\Sigma_{d=0}^{1}\int f_{X,Y,D}(\mathbf{X},y,d)\,\mathrm{d}\mu(y)/\eta_1(\mathbf{X})=0$ almost surely, which yields $E\{\mathbf{h}-E(\mathbf{h}|D)|\mathbf{X}\}=\mathbf{0}$ almost surely.

Now we are in position to show that

$$\Lambda^{\perp}=\Lambda_1^{\perp}\cap\Lambda_2^{\perp}=[\mathbf{h}(d,\varepsilon,\mathbf{x}):E(\mathbf{h})=0,\,E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}\sum_d\frac{N_d}{N}\frac{H(d,\mathbf{X},Y)}{p_D^{\text{true}}(d)}=\varepsilon\,\mathbf{a}(\mathbf{X})\text{ almost surely}],$$

where $\mathbf{a}(\mathbf{x})$ is an arbitrary function of $\mathbf{x}$. This is because, for any $\mathbf{h}\in\Lambda_1^{\perp}$, $\mathbf{h}\in\Lambda_2^{\perp}$ is equivalent to

$$\begin{aligned}0=&E[\mathbf{h}^{\mathrm{T}}\{\mathbf{g}(\varepsilon,\mathbf{X})-E(\mathbf{g}|D)\}]\\=&E[(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}\{\mathbf{g}(\varepsilon,\mathbf{X})-E(\mathbf{g}|D)\}]\\=&E\{(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}\mathbf{g}\}\\=&E[E\{(\mathbf{h}-E(\mathbf{h}|D))^{\mathrm{T}}|\varepsilon,\mathbf{X}\}\mathbf{g}].\end{aligned}$$

Hence $E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}\,\Sigma_d f_{X,Y,D}(\mathbf{X},Y,d)/\{\eta_1(\mathbf{X})\eta_2(\varepsilon,\mathbf{X})\}=\varepsilon\,\mathbf{a}(\mathbf{X})+\mathbf{c}(\mathbf{X})$ almost surely. Because $\mathbf{h}\in\Lambda_1^{\perp}$, we have $E[E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}|\mathbf{X}]=0$ almost surely. Hence

$$\begin{aligned}\mathbf{0}=&\int E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}\frac{\sum_d f_{X,Y,D}(\mathbf{X},y,d)}{\int\sum_d f_{X,Y,D}(\mathbf{X},y,d)\,\mathrm{d}\mu(y)}\mathrm{d}\mu(y)\\=&\frac{\int\{\varepsilon\,\mathbf{a}(\mathbf{X})+\mathbf{c}(\mathbf{X})\}\,\eta_1(\mathbf{x})\,\eta_2(\varepsilon,\mathbf{X})\mathrm{d}\mu(y)}{\int\sum_d f_{X,Y,D}(\mathbf{X},y,d)\,\mathrm{d}\mu(y)}\\=&\frac{\mathbf{c}(\mathbf{X})\,\eta_1(\mathbf{X})}{\int\sum_d f_{X,Y,D}(\mathbf{X},y,d)\,\mathrm{d}\mu(y)}\qquad\text{almost surely;}\end{aligned}$$

hence $\mathbf{c}(\mathbf{X})=0$ almost surely and $E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}\Sigma_d f_{X,Y,D}(\mathbf{X},Y,d)/\{\eta_1(\mathbf{X})\,\eta_2(\varepsilon,\mathbf{X})\}=\varepsilon\,\mathbf{a}(\mathbf{X})$ almost surely. This means that $E\{\mathbf{h}-E(\mathbf{h}|D)|\varepsilon,\mathbf{X}\}\Sigma_d\,(N_d/N)\,H(d,\mathbf{X},Y)/p_D^{\text{true}}(d)=\varepsilon\,\mathbf{a}(\mathbf{X})$ almost surely.

A.3. Sketch of proof of theorem 1

For simplicity of proof, we split the N observations randomly into two sets. The first set contains $n_1=N-N^{1-\delta}$ observations and the second set contains $n_2=N^{1-\delta}$ observations, where δ is a small positive number. We form and solve the estimating equation by using data in the first set, while calculating all the hatted quantities described in the algorithm by using data in the second set. We use this only as a technical device, although in our simulations and empirical example we used all the data.

In the algorithm, the approximations involve either replacing expectation with averaging, or standard kernel regression estimation or kernel density estimation; hence the differences between the quantities with a hat and without a hat have either mean 0 and standard deviation $O(n_2^{-1/2})$ or mean $O(h^r)$ and standard deviation $O\{(n_2h^p)^{-1/2}\}$. In particular, $\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)$ has bias $O(h^r)$ and standard deviation $O\{(n_2h^p)^{-1/2}\}$. Recall the definition of expectation and covariance in the superpopulation explicitly written out in the statement of theorem 1. Then

$$\begin{aligned}\mathbf{0}&=n_1^{-1/2}\sum_{i=1}^{n_1}\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\hat{\boldsymbol{\theta}})\\&=n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)+n_1^{-1/2}\sum_{i=1}^{n_1}\{\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)\}\\&\quad+E\left\{\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\text{T}}}+o_p(1)\right\}n_1^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\&=n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)+E\left\{\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\text{T}}}\right\}n_1^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\&\quad+n_1^{-1/2}\sum_{i=1}^{n_1}\{\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)\}+o_p(1).\end{aligned}$$

We see that $\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)$ differs from $\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)$ in that all the unknown quantities, except $\mathbf{S}^*$, are estimated. This is equivalent to estimating the unknown functions $\eta_1(\mathbf{x})$ and $\eta_2(\varepsilon,\mathbf{x})$ in expression (4) and using the estimates $\hat{\eta}_1(\mathbf{x})$ and $\hat{\eta}_2(\varepsilon,\mathbf{x})$ in calculating $\mathbf{S}^*_{\text{eff}}$ from the posited $\mathbf{S}^*$. Thus, denoting $\hat{\eta}=(\hat{\eta}_1,\hat{\eta}_2)$, we can approximate

$$\begin{aligned}n_1^{-1/2}&\sum_{i=1}^{n_1}\{\hat{\mathbf{S}}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)\}\\&=n_1^{-1/2}\sum_{i=1}^{n_1}\{\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\hat{\eta})-\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\eta_0)\}\\&=\left\{n_1^{-1/2}\sum_{i=1}^{n_1}\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\eta_0)}{\partial\eta}\right\}(\hat{\eta}-\eta_0)+O_p\{n_1^{1/2}(\hat{\eta}-\eta_0)^2\}+o_p(1),\end{aligned}\qquad(10)$$

where $\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\eta_0)/\partial\eta$ is the pathwise derivative. However, $\mathbf{S}^*_{\text{eff}}$ is the projection of $\mathbf{S}^*$ to $\Lambda^{\perp}$ so $\mathbf{S}^*_{\text{eff}}\in\Lambda^{\perp}$. Thus, for any parametric submodel of η involving parameter γ, we have

$$\begin{aligned}E\left\{\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\gamma)}{\partial\gamma^{\text{T}}}\right\}&=\int\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\gamma)}{\partial\gamma^{\text{T}}}f_{X,Y,D}(x,y,d)\,\mathrm{d}\mu(x)\mu(y)\,\mathrm{d}\mu(d)\\&=-\int\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\gamma)\frac{\partial\log\{f_{X,Y,D}(x,y,d)\}}{\partial\gamma^{\text{T}}}f_{X,Y,D}(x,y,d)\,\mathrm{d}\mu(x)\,\mu(y)\mathrm{d}\mu(d)\\&=-E\{\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\gamma)\mathbf{S}_{\gamma}^{\text{T}}\}=\mathbf{0}.\end{aligned}$$

The last equality is because by definition $\mathbf{S}_{\gamma}\in\Lambda$ which is orthogonal to $\Lambda^{\perp}$ and $\mathbf{S}^*_{\text{eff}}\in\Lambda^{\perp}$. Here, $f_{X,Y,D}(x,y,d)$ is defined in expression (4). Because γ is a parameter of any arbitrary submodel of η, we actually have obtained

$$E\{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\eta_0)/\partial\eta\}=-E\{\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0,\eta_0)\mathbf{S}_{\eta}^{\text{T}}\}=0,$$

where $\mathbf{S}_{\eta}$ is the nuisance score function along the arbitrarily chosen specific path of the pathwise derivative. Thus, the first term of equation (10) is of order $o_p(1)$. In contrast, $O_p\{n_1^{1/2}(\hat{\eta}-\eta_0)^2\}=O_p\{n_1^{1/2}h^{2r}+n_1^{1/2}(n_2h^p)^{-1}\}=o_p(1)$. We therefore obtain

$$\mathbf{0}=n_1^{-1/2}\sum_{i=1}^{n_1}\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)+E\left\{\frac{\partial\mathbf{S}^*_{\text{eff}}(D_i,\mathbf{X}_i,Y_i,\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}^{\text{T}}}\right\}n_1^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+o_p(1).$$

This yields $n_1^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\rightarrow N\{\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\text{T}}\}$, and hence

$$N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\rightarrow N\{\mathbf{0},\mathbf{A}^{-1}\mathbf{B}(\mathbf{A}^{-1})^{\text{T}}\}$$

when $N\rightarrow\infty$.

References

Barrett, J. H., Smith, G., Waxman, R., Gooderham, N., Lightfoot, T., Garner, R. C., Augustsson, K., Wolf, C. R., Bishop, D. T., Forman, D. and the Colorectal Cancer Study Group (2003) Investigation of interaction between n-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectal cancer. *Carcinogenesis*, **24**, 275–282.

Bickel, P. J., Klassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Chatterjee, N. and Carroll, R. J. (2005) Semiparametric maximum likelihood estimation in case-control studies of gene-environment interactions. *Biometrika*, **92**, 399–418.

Chatterjee, N., Chen, J., Spinka, C. and Carroll, R. J. (2006) Comment on the paper Likelihood based inference on haplotype effects in genetic association studies by D. Y. Lin and D. Zeng. *J. Am. Statist. Ass.*, **101**, 108–110.

Chen, Y. H., Chatterjee, N. and Carroll, R. J. (2008) Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association. *Biostatistics*, **9**, 81–99.

Chen, Y. H., Chatterjee, N. and Carroll, R. J. (2009) Shrinkage estimators for robust and efficient inference in haplotype-based case-control studies. *J. Am. Statist. Ass.*, **104**, 220–233.

Corley, D. A., Jensen, C. D., Marks, A. R., Zhao, W. K., Lee, J. K., Doubeni, C. A., Zauber, A. G., de Boer, J., Fireman, B. H., Schottinger, J. E., Quinn, V. P., Ghai, N. R., Levin, T. R. and Quesenberry, C. P. (2014) Adenoma detection rate and risk of colorectal cancer and death. *New Engl. J. Med.*, **370**, 1298–1306.

Davidian, M. and Carroll, R. J. (1987) Variance function estimation. *J. Am. Statist. Ass.*, **82**, 1079–1092.

De Stefani, E., Ronco, A., Mendilaharsu, M., Guidobono, M. and Deneo-Pellegrini, H. (1997) Meat intake, heterocyclic amines, and risk of breast cancer: a case-control study in Uruguay. *Cancer Epidem. Biomark. Prevn*, **6**, 573–581.

He, J., Li, H., Edmondson, A. C., Rader, D. J. and Li, M. (2012) A gaussian copula approach for the analysis of secondary phenotypes in case–control genetic association studies. *Biostatistics*, **13**, 497–508.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006) Secondary analysis of case-control data. *Statist. Med.*, **25**, 1323–1339.

Li, H., Gail, M. H., Berndt, S. and Chatterjee, N. (2010) Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet. Epidem.*, **34**, 427–433.

Lin, D. Y. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epidem.*, **33**, 256–265.

Lobach, I., Carroll, R. J., Spinka, C., Gail, M. H. and Chatterjee, N. (2008) Haplotype-based regression analysis of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, **64**, 673–684.

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

Ma, Y. and Zhu, L. P. (2012) A semiparametric approach to dimension reduction. *J. Am. Statist. Ass.*, **107**, 168–179.

Ma, Y. and Zhu, L. P. (2013) Efficient estimation in sufficient dimension reduction. *Ann. Statist.*, **41**, 250–268.

Prentice, R. L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403–411.

Scott, A. and Wild, C. (2002) On the robustness of weighted methods for fitting models to case–control data. *J. R. Statist. Soc.* B, **64**, 207–219.

Sinha, R., Kulldorff, M., Chow, W.-H., Denobile, J. and Rothman, N. (2001) Dietary intake of heterocyclic amines, meat-derived mutagenic activity, and risk of colorectal adenomas. *Cancer Epidem. Biomark. Prevn*, **10**, 559–562.

Tsiatis, A. A. (2006) *Semiparametric Theory and Missing Data*. New York: Springer.

Wei, J., Carroll, R. J., Müller, U., Van Keilegom, I. and Chatterjee, N. (2013) Locally efficient estimation for homoscedastic regression in the secondary analysis of case–control data. *J. R. Statist. Soc.* B, **75**, 185–206.

Yamaji, Y., Mitsushima, T., Ikuma, H., Watabe, H., Okamoto, M., Kawabe, T., Wada, R., Doi, H. and Omata, M. (2004) Incidence and recurrence rates of colorectal adenomas estimated by annually repeated colonoscopies on asymptomatic Japanese. *Gut*, **53**, 568–572.

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this paper:

‘Supplementary material for Semiparametric estimation in secondary analysis of case-control studies’.