Statistics and Its Interface Volume 9 (2016) 147–158

A hybrid parametric and empirical likelihood model for evaluating interactions in case-control studies*

Jing Qin†, Hong Zhang, Maria Landi, Neil Caporaso, and Kai Yu

The case-control design provides an effective way to collect covariate information conditioning on subjects' disease status. The standard logistic regression model can be used to model the interaction between two covariates under such a design, but the prospective logistic regression method might not be the most efficient one when certain appropriate constraints can be imposed on the covariate distribution. We develop a hybrid approach for the statistical inference of the interaction under the case-control design. We use a parametric model to characterize the conditional distribution of one covariate given the another covariate in the control population, while leaving the distribution of the later covariate to be fully nonparametric. A maximum hybrid parametric and empirical likelihood method is adopted for the evaluation of all parameters. The estimator and the associated test derived from the proposed semiparametric model are suitable for evaluating the interaction between two covariates of various types (discrete or continuous). Asymptotic results for both the estimators and the test statistics were established, and the advantages of the proposed method over the existing ones are demonstrated through simulation results and a real data example.

AMS 2000 subject classifications: Primary 62J12; secondary 62P10.
Keywords and phrases: Case-control, Genetic association studies, Hybrid parametric and empirical likelihood, Interaction test.

1. INTRODUCTION

In epidemiologic studies, it is often of interest to assess whether there is any interaction between two variables once they have been established as risk factors (covariate) for the disease under study. For example, due to the recent advances in high-throughput genomic technologies that allow the simultaneous measurement of a large volume of biological markers (such as DNA variants and mRNA markers), there is growing enthusiasm for detecting gene-gene and gene-environment interactions. An interaction exists if the effect of one covariate depends on the level of another one. A formal statistical definition for the interaction effect depends on the underlying statistical model. Under the framework of generalized linear model, the interaction effect is typically measured by the coefficient of the product of the two covariates.

For epidemiological studies of rare diseases, the case-control study design provides a cost-effective way of collecting covariate information conditioning upon subjects' disease status. Although samples are collected retrospectively, it has been shown that inference based on the prospective logistic regression model is asymptotically efficient if there is no information on the distribution of covariates [1, 2, 3, 4]. However, if certain constraints derived from auxiliary information can be imposed on the covariate distribution, the standard logistic regression is still valid, but might not be the most efficient one. For example, in genetic epidemiologic studies of gene-environment interaction, sometimes it is appropriate to assume that in the source population a subject's environmental exposure status is independent of his/her genetic makeup. Under such a gene-environment independence condition, a more efficient estimate for the interaction coefficient can be derived, such as the class of case-only estimates that only use information collected among cases [5, 6], and the more general semiparametric maximum likelihood estimate that integrates information from both cases and controls [7, 8].

Although the interaction estimates (or tests) derived under the independence assumption tend to be more efficient than the one based on the standard logistic regression, their validity relies on the critical independence assumption. It has been demonstrated that the case-only estimate (test) can be highly distorted when the independence assumption is violated [9]. An empirical Bayes-type estimate was proposed in [10], which was a weighted average of two interaction estimates, one derived under the gene-environment

*This research was supported by the State Key Development Program for Basic Research of China (Grant No. 2012CB316500) (HZ), the National Natural Science Foundation of China (Grant No. 11371101) (HZ), and the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (HZ, KY).
†Corresponding author.

independence assumption and the other more robust one derived under the standard logistic regression without the independence assumption. When there is an evidence of association between the two covariates, the empirical Bayes-type estimate puts more weight on the robust estimate, and vice versa.

To relax the independence assumption between X and Y, Zhang et al. [11] proposed a copula function to relating X and Y. In this paper, we develop an alternative approach for estimating and testing the interaction between two covariates (X and Y) in case-control studies. Instead of assuming an independent relationship between X and Y, we characterize the relationship between X and Y in the control population by specifying a parametric model for the conditional distribution of Y given X, while leaving the distribution for the other components of the joint covariate distribution to be fully nonparametric. The estimate and the associated test derived from this proposed semiparametric model are suitable for evaluating the interaction between two covariates of various types, i.e., discrete by discrete, discrete by continuous, and continuous by continuous. Therefore, our method can be applied to evaluating gene by gene and gene by environment interactions in genetic association studies. It is also applicable to the study of interaction between two continuous covariates in other epidemiology studies.

In the proposed method, we make an extra distribution assumption compared with the standard logistic regression. This method is a hybrid in the sense that the resulting log-likelihood function is the summation of a parametric log-likelihood and an empirical log-likelihood, as will be shown in the next section. Such a hybrid method lies between the traditional logistic regression for case-control design (corresponding to an empirical likelihood) and the fully parametric method (corresponding to a parametric likelihood). Since the hybrid method incorporates an additional assumption, it could be intuitively more efficient than the standard logistic regression method provided the extra information is accurate.

This paper is organized as follows. In Section 2, using an empirical likelihood method [12], we describe a new approach for assessing the interaction, and present some asymptotic results as the basis for statistical inference. In Section 3, we conduct some simulation studies to evaluate the performance of the proposed procedure. In Section 4, we demonstrate the application of the proposed procedure through a real data example. We conclude this paper with some discussions in Section 5.

2. MAIN RESULTS

Let $D = 1$ or 0 be the indicator of case/control status. Let Y and X be the two covariates under investigation. The common risk model for the binary outcome is the logistic regression model:

(1)
$$P(D=1|Y=y,X=x)=\frac{\exp(\alpha^*+y\beta+\gamma x+\xi xy)}{1+\exp(\alpha^*+y\beta+\gamma x+\xi xy)},$$

where α^* is an intercept, β and γ are main effects and ξ is an interaction effect. We are interested in testing the null hypothesis of no interaction, i.e., $H_0 : \xi = 0$.

Instead of prospectively collecting (D, Y, X), typically one collects (Y, X) by conditioning on the status of D in case-control studies. This is the so called retrospective sampling or case-control sampling. Let

$$\{(Y_{i1}, X_{i1}), i = 1, 2, \ldots, n_1\} \text{ and } \{(Y_{i0}, X_{i0}), i = 1, 2, \ldots, n_0\}$$

be the covariate data for cases ($D_i = 1$) and controls ($D_i = 0$), respectively, where n_1 and n_0 are the number of cases and controls. Using Bayes' formula, one obtains the density functions of (Y, X) for cases and controls:

$$f(y,x|D=1)=\frac{P(D=1|y,x)f(y,x)}{P(D=1)},$$
$$f(y,x|D=0)=\frac{P(D=0|y,x)f(y,x)}{P(D=0)},$$

where $f(y, x)$ is the marginal density function of (Y, X) (in the general population). The case and control density functions can be linked by the exponential tilting model [13]

(2) $$f(y,x|D=1)=\exp(\alpha+y\beta+\gamma x+\xi xy)f(y,x|D=0),$$

where $\alpha = \alpha^* + \log\{P(D = 0)/P(D = 1)\}$. Hereafter, we use $f_0(y, x)$ and $f_1(y, x)$ to denote $f(y, x|D = 0)$ and $f(y, x|D = 1)$, respectively.

[14] showed that the semiparametric likelihood estimation of the baseline distribution function $\int_{-\infty}^{y}\int_{-\infty}^{x} f_0(t,s)dsdt$ based on the exponential tilting model (2) had an asymptotic Bahadur representation and was more efficient than the empirical distribution function estimation based on control data only.

Without any auxiliary information on $f_0(y, x)$, one may perform a prospective logistic likelihood analysis with the log-likelihood function $\sum_{i=1}^{n} \log P(D_i|X_i, Y_i)$ equal to

(3)
$$\ell_P := \sum_{i=1}^{n_1}(\alpha+\beta Y_{i1}+\gamma X_{i1}+\xi X_{i1}Y_{i1}) - \sum_{i=1}^{n}\log[1+\exp(\alpha+\beta Y_i+\gamma X_i+\xi X_iY_i)].$$

Here we use $\{(X_i, Y_i), i = 1, 2, \ldots, n\ (= n_0 + n_1)\}$ to denote the pooled data $\{(X_{i1}, Y_{i1}), i = 1, 2, \ldots, n_1; (X_{i0}, Y_{i0}), i = 1, 2, \ldots, n_0\}$, and ":=" is used for a definition. Then one may use the score or likelihood ratio test to test $H_0 : \xi = 0$.

In the following, we model the dependence between X and Y in the control population through a parametric model $f_0(y|x) = f_0(y|x, \boldsymbol{\eta})$. Then the joint density function of Y and X for controls is $f_0(y, x) = f_0(y|x)f_0(x)$, where $f_0(x)$ is the marginal density function of X for controls. According

to the tilting model (2), the marginal density function for cases (defined as $\int f_1(y,x)dy$) is

$$f_1(x) := \exp(\alpha+\gamma x)\int \exp(\beta y+\xi xy)f_0(y|x,\boldsymbol{\eta})dy f_0(x) = \exp(\alpha+\gamma x)\mu_1(x,\beta,\xi,\boldsymbol{\eta})f_0(x), \tag{4}$$

where

$$\mu_1(x,\beta,\xi,\boldsymbol{\eta}) := \int \exp(\beta y+\xi xy)f_0(y|x,\boldsymbol{\eta})dy. \tag{5}$$

For convenience, let

$$\phi(x,\beta,\gamma,\xi,\boldsymbol{\eta}) := \gamma x+\log\mu_1(x,\beta,\xi,\boldsymbol{\eta}),$$

then

$$f_1(x) = \exp\{\alpha+\phi(x,\beta,\gamma,\xi,\boldsymbol{\eta})\}f_0(x).$$

Thus, similar to model (2), the marginal density functions $f_1(x)$ for cases and $f_0(x)$ for controls are linked again by an exponential tilting model, but through a different link function. Moreover, the conditional density function of Y given X for cases (defined as $f_1(y,x)/f_1(x)$) is

$$f_1(y|x) = \exp(\beta y+\xi xy)f_0(y|x,\boldsymbol{\eta})/\mu_1(x,\beta,\xi,\boldsymbol{\eta}), \tag{6}$$

which is known up to parameters $(\beta,\xi,\boldsymbol{\eta})$. Therefore, the full data log-likelihood (defined as $\sum_{i=1}^n \log P(Y_i,X_i|D_i)$) is

$$\ell = \ell_c+\ell_M,$$

where

$$\ell_c := \sum_{i=1}^{n_1}[\beta Y_{i1}+\xi Y_{i1}X_{i1} - \log\mu_1(X_{i1},\beta,\xi,\boldsymbol{\eta})] + \sum_{i=1}^{n}\log f_0(Y_i|X_i,\boldsymbol{\eta}) \tag{7}$$

is the conditional parametric log-likelihood and

$$\ell_M := \sum_{i=1}^{n_1}\{\alpha+\phi(X_{i1},\beta,\gamma,\xi,\boldsymbol{\eta})\}+\sum_{i=1}^{n}\log f_0(X_i) \tag{8}$$

is the marginal empirical likelihood.

We use the theory of empirical likelihood [12] to profile out the high-dimensional parameters $\{f_0(X_i), i = 1,2,\ldots,n\}$. Let p_i be the jump size of $\int_0^x f_0(s)ds$ at X_i, we seek to maximize the marginal likelihood with $f_0(X_i)$ replaced by p_i. Note that $p_1,\ldots,p_n$ satisfy the constraints $\sum_{i=1}^n p_i = 1$ and $\sum_{i=1}^n p_i\exp(\alpha+\phi(X_{i1},\beta,\gamma,\xi,\boldsymbol{\eta})) = 1$. Applying the result of [13], one obtains the maximizers

$$\hat{p}_i = \frac{1}{n}\frac{1}{1+\rho\exp(\alpha+\phi(X_{i1},\beta,\gamma,\xi,\boldsymbol{\eta}))}, \quad i=1,\ldots,n,$$

where $\rho := n_1/n_0$. As a result, one has the log marginal profile empirical likelihood

$$p\ell_M := \sum_{i=1}^{n_1}\{\alpha+\gamma X_{i1}+\log\mu_1(X_{i1},\beta,\xi,\boldsymbol{\eta})\} - \sum_{i=1}^{n}\log[1+\rho\exp(\alpha+\phi(X_{i1},\beta,\gamma,\xi,\boldsymbol{\eta}))\}]. \tag{9}$$

Define $\boldsymbol{\omega} := (\alpha,\beta,\gamma,\boldsymbol{\eta},\xi)$. Then, one may make inference for $\boldsymbol{\omega}$ based on the log hybrid parametric and empirical likelihood

$$\ell_H(\boldsymbol{\omega}) := \ell_c(\boldsymbol{\omega})+p\ell_M(\boldsymbol{\omega}), \tag{10}$$

where ℓ_c and $p\ell_M$ are given in (7) and (9), respectively. Let $\boldsymbol{\omega}_0$ be the true value of $\boldsymbol{\omega}$ and $\hat{\boldsymbol{\omega}}$ be the maximum hybrid parametric and empirical likelihood estimator. We have the following asymptotic result for $\hat{\boldsymbol{\omega}}$.

Theorem 1. *Under some regularity conditions specified in Appendix, $\sqrt{n}(\hat{\boldsymbol{\omega}}-\boldsymbol{\omega}_0)$ converges in distribution to a multivariate normal distribution with expectation 0 and variance-covariance matrix given in Appendix.*

The hybrid likelihood ratio test statistic for $H_{01}:\xi=0$ is defined as

$$R_1 = 2\left(\max_{\alpha,\beta,\gamma,\xi,\boldsymbol{\eta}}\ell_H(\boldsymbol{\omega}) - \max_{\alpha,\beta,\gamma,\xi=0,\boldsymbol{\eta}}\ell_H(\boldsymbol{\omega})\right).$$

Theorem 2. *Under some regularity conditions specified in Appendix and under the null hypothesis H_{01}, the hybrid likelihood ratio test statistic R_1 converges in distribution to the chi-square distribution with 1 degree of freedom.*

Suppose that the conditional density function $f_0(\cdot|x,\boldsymbol{\eta})$ depends on x and $\boldsymbol{\eta}$ $(=(\eta_1,\eta_2))$ only through $\eta_1+\eta_2 x$, i.e.,

$$f_0(y|x,\boldsymbol{\eta}) = f_0(y|\eta_1+\eta_2 x),$$

we can test the null hypothesis $H_{02}:\xi=0,\ \eta_2=0$ (i.e., X and Y are independent in both case and control groups). The hybrid likelihood ratio test statistic for testing H_{02} takes the form

$$R_2 = 2\left(\max_{\alpha,\beta,\gamma,\xi,\eta_1,\eta_2}\ell_H(\boldsymbol{\omega}) - \max_{\alpha,\beta,\gamma,\xi=0,\eta_1,\eta_2=0}\ell_H(\boldsymbol{\omega})\right).$$

Theorem 3. *Under the null hypothesis H_{02}, the hybrid likelihood ratio test statistic R_2 converges in distribution to the chi-square distribution with 2 degrees of freedom.*

The proofs of Theorems 1–3 are postponed to Appendix.

3. SIMULATION STUDIES

In this section, we conduct simulation studies to evaluate the performance of the proposed method. The density ratio

model (2), in general, is not convenient for generating observations for cases. Instead, we will use an approximation method as described in the following. The basic idea comes from the unequal weight sampling method used in survey sampling. Noting that the joint density functions of (X, Y) for cases and controls are linked by the exponential tilting model

$$f_1(y,x) = \exp\{\alpha + \phi(y,x,\beta)\} f_0(y,x),$$

one can first generate a large number of random vectors $\{(X_i, Y_i), i = 1, 2...., N\}$ from $f_0(y,x)$. Then, one may generate

$$f(Y_i = y, X_i = x | D_i = 1) := \frac{\exp\{\phi(Y_i, X_i, \beta)\}}{\sum_{i=1}^{N} \exp\{\phi(Y_i, X_i, \beta)\}}.$$

If N is large enough, then $f(Y_i = y, X_i = x|D_i = 1)$ can be a good approximation of

$$f_1(y,x) = \frac{\exp\{\phi(y,x,\beta)\} f_0(y,x)}{\int \exp\{\phi(y,x,\beta)\} f_0(y,x) dy dx}.$$

More details on the unequal weight sampling or biased sampling model can be found in [15, 16, 17, 18, 19].

We consider two scenarios, one for the interaction between two continuous variables, the other for the interaction between a discrete variable and a continuous variable.

For comparison, we consider three estimation methods. The first one is to estimate α, β, γ, and ξ based on the prospective likelihood function. In this approach, the conditional information of $f_0(y|x) := f_0(y|x, \boldsymbol{\eta})$ in the control group is not used, and $\boldsymbol{\eta}$ is estimated by maximizing the likelihood function for the control data alone. The second one is the proposed method in this paper with $(\alpha, \beta, \gamma, \xi, \boldsymbol{\eta})$ being estimated by maximizing the hybrid likelihood function $\ell_H = \ell_c + p\ell_M$. The third one is the Chatterjee and Carroll method as implemented in the R package CGEN, which assumes that the gene and environment are independent in the general population and Hardy-Weinberg equilibrium holds for the studied marker.

For testing $H_0 : \xi = 0$, we considered four test statistics: 1) the hybrid likelihood ratio test given in Theorem 2, 2) the likelihood ratio test based on logistic regression without using the conditional density information of $f_0(y|x, \boldsymbol{\eta})$, 3) the likelihood ratio test of Chatterjee and Carroll method, and 4) Pearson's correlation coefficient test between two covariates based on case-only data.

3.1 Study 1

We evaluate the interaction between two continuous variables.

For given covariates X and Y, the disease status D satisfies the logistic regression model (1). Let $\boldsymbol{\eta} := (\eta_1, \eta_2, \eta_3)$, and assume that the conditional distribution of Y given $X = x$ for the control population is the normal distribution with expectation $\eta_1 + \eta_2 x$ and variance η_3^2. Therefore, the marginal density function of X for cases is

$$f_1(x) = f_0(x) \exp\{\alpha + \gamma x + (\eta_1 + \eta_2 x)(\beta + \xi x) + 0.5(\beta + \xi x)^2 \eta_3^2\},$$

and the conditional density function of Y given $X = x$ for cases is

$$f_1(y|x) = \exp[\beta y + \xi x y - \{(\eta_1 + \eta_2 x)(\beta + \xi x) + 0.5(\beta + \xi x)^2 \eta_3^2\}] \times \frac{1}{\sqrt{2\pi\eta_3^2}} \exp\left\{-0.5(y - \eta_1 - \eta_2 x)^2 / \eta_3^2\right\}.$$

Assuming that $f_0(x)$ is the standard normal density function, we have a closed form for the conditional distribution of Y given X in the case population, that is,

$$Y|X = x \sim N\left(\beta + \xi x + (\eta_1 + \eta_2 x)/\eta_3^2, \eta_3^2\right).$$

Moreover, the marginal distribution of X in the case population is $N(\mu, \sigma^2)$, where

$$\sigma^2 = (1 - 2\eta_2\xi - \xi^2\eta_3^2)^{-1}, \quad \mu = \sigma^2(\gamma + \eta_2\beta + \eta_1\xi + \beta\xi\eta_3^2).$$

The above reasoning shows that, if the two covariates are jointly normally distributed in the control population, then the joint distribution in the case population is also bivariate normal, provided that the logistic regression model holds true. Furthermore, the data generation is straightforward under the above model.

We fixed the main effects at $\beta = \gamma = 1$, and $\eta_1 = 0$, $\eta_3 = 1$. For different choices of interaction effect ξ and η_2 that characterized the correlation between two covariates, we generated the covariate data for 500 cases and 500 controls, and applied to the simulated data the logistic regression method, the hybrid method (assuming a normal conditional distribution), and the case-only method. The estimation results and sizes/powers for testing interaction effect are reported in Table 1 and Table 2, respectively.

We have the following observations. In all simulation situations, both logistic regression method and the hybrid method produce little estimation bias, and the hybrid method has smaller variance than the logistic regression method, especially for the interaction effect. If the two covariates are independent in the control population ($\eta_2 = 0$), then the case-only method has type one error rate around the nominal levels, and it has greater powers than the logistic regression method and the hybrid method have ($\xi = 0.2$ and $\eta_2 = 0$). The case-only method, however, is very sensitive to the independence assumption, which could have type one error rate close to 1 ($\xi = 0$ and $\eta_2 = -0.5$). This result shows that one must be very cautious when using the case-only method. This observation is consistent with the

*Table 1. Estimation bias (standard error) in the continuous situation**

			logistic regression			hybrid method		
distribution**	ξ	η_2	γ	β	ξ	γ	β	ξ
normal	0	0	.008(.096)	.013(.098)	-.013(.090)	.001(.089)	.006(.091)	-.005(.064)
normal	0	-.5	.003(.091)	.007(.083)	-.008(.061)	.001(.089)	.004(.082)	-.005(.053)
normal	.2	0	.008(.107)	.015(.109)	-.012(.102)	.000(.095)	.006(.096)	-.003(.064)
normal	.2	-.5	.002(.095)	.007(.086)	-.008(.073)	.000(.092)	.005(.084)	-.006(.060)
mixed	0	0	.001(.077)	.003(.053)	-.001(.053)	.002(.075)	.011(.055)	-.002(.046)
mixed	0	-.5	.001(.078)	.002(.051)	.000(.044)	.005(.077)	.010(.053)	.000(.041)
mixed	.2	0	.000(.081)	.002(.056)	-.001(.057)	.026(.077)	.027(.060)	-.030(.046)
mixed	.2	-.5	.000(.081)	.002(.053)	.001(.054)	.020(.080)	.014(.056)	-.018(.046)
t_6	0	0	.002(.072)	.001(.073)	.001(.072)	-.003(.071)	-.097(.051)	.010(.045)
t_6	0	-.5	.000(.078)	.001(.070)	.000(.055)	-.054(.074)	-.095(.049)	.015(.042)
t_6	.2	0	.002(.075)	-.001(.075)	.000(.076)	.008(.074)	-.121(.045)	-.036(.041)
t_6	.2	-.5	-.001(.078)	.000(.071)	.001(.064)	-.071(.072)	-.118(.043)	-.003(.044)
t_{10}	0	0	.001(.073)	.002(.072)	-.001(.073)	.002(.072)	-.008(.068)	-.008(.061)
t_{10}	0	-.5	.000(.076)	.002(.069)	.000(.056)	-.005(.076)	-.011(.065)	-.002(.051)
t_{10}	.2	0	.000(.074)	.003(.074)	-.002(.076)	-.002(.074)	-.016(.067)	-.012(.059)
t_{10}	.2	-.5	.000(.078)	.000(.07)	.002(.063)	-.014(.076)	-.016(.064)	.005(.056)
skewed	0	0	.003(.112)	.002(.068)	-.002(.069)	.003(.107)	-.001(.067)	-.002(.06)
skewed	0	-.5	.002(.100)	.001(.065)	-.002(.054)	.000(.099)	-.003(.064)	-.001(.05)
skewed	.2	0	.005(.123)	-.001(.075)	-.002(.077)	-.021(.107)	-.010(.070)	.010(.056)
skewed	.2	-.5	-.002(.108)	-.001(.069)	.002(.064)	-.031(.104)	-.007(.066)	.018(.057)

*The working conditional distribution for the hybrid method was normal.
**Underlying conditional distribution: "normal", standard normal distribution; "mixed", mixture of two standard normal distributions with equal weights; "t_6", t distribution on 6 df; "t_{10}", t distribution on 10 df; "skewed", skewed normal distribution with shape parameter 2.

results found in [9] for testing interaction between two binary covariates. On the other hand, the hybrid method has well controlled type one error rates ($\xi = 0$), and it is more powerful than the logistic regression method ($\xi = 0.2$). The power gain of the hybrid method over the logistic regression method can be very considerable. For example, when $\xi = 0.2$ and $\eta_2 = -0.5$, the power at 0.05 level of the hybrid method is 0.881, compared with 0.465 for the logistic regression method.

To study the robustness of the proposed method to the misspecification of the conditional distribution of Y given X, we considered three types of conditional distributions other than the normal one. The first one was the mixture of two standard normal distribution with equal weights, the second one was the t distribution with six or ten degrees of freedom, and the third one was the skewed normal distribution with shape parameter two [20]. For each of these three conditional distributions, we considered four parameter combinations as described in the normal situation. In the hybrid method, we specified the conditional distribution of Y given X to be normal. The estimation results and sizes/powers are again reported in Table 1 and Table 2, respectively.

Under the null hypothesis ($\xi = 0$), the hybrid method has nearly unbiased estimates and good control of type one error rates except in one situation (t_6 distribution, $\xi = \eta_2 = 0$), with the type one error rate being slightly deflated in this situation. Under the alternatives ($\xi = 0.2$), the hybrid method is still more powerful than the logistic regression method, and the power gain over the logistic regression method is satisfying in most situations. On the other hand, the case-only method has an extremely distorted type one error rate when X and Y are correlated ($\eta_2 = -0.5$ and $\xi = 0$), and it can lose power dramatically in some situations (t_6 distribution, $\xi = 0.2$, $\eta_2 = -0.5$; mixed normal distribution, $\xi = 0.2$, $\eta_2 = -0.5$). The above simulation results show that the hybrid method performs pretty well except when the conditional distribution is seriously misspecified (t_6 distribution).

The method developed in [11] used a copula function to relate the distribution of X and Y. This method will be termed "copula" hereafter. We conducted additional simulations to compare the performance of the copula method and the method developed in this manuscript ("hybrid"). In the copula method, we used the gaussian copula function, and in the hybrid method, we specified the conditional function of Y given X to be normal. We generated two covariates X and Y that were jointly normal (the marginal distributions being standard normal) in the control population, so that both the hybrid method and copula method should work since the conditional distribution $Y|X$ in the controls is normal (work for the hybrid method) and the copula function for the joint distribution is Gaussian (work for the copula method). The true values of both β and γ in the underlying tilting model were fixed at 0.5, ξ was either 0

*Table 2. Size/power in the continuous situation**

distribution**	ξ	η_2	logist .01	logist .05	logist .1	hybrid .01	hybrid .05	hybrid .1	case-only .01	case-only .05	case-only .1
normal	0	0	.012	.047	.097	.012	.051	.102	.009	.049	.100
normal	0	-.5	.010	.054	.101	.010	.051	.102	1.000	1.000	1.000
normal	.2	0	.236	.465	.587	.717	.881	.930	.968	.994	.997
normal	.2	-.5	.541	.770	.859	.752	.907	.946	1.000	1.000	1.000
mixed	0	0	.008	.050	.099	.009	.047	.094	.012	.049	.097
mixed	0	-.5	.009	.044	.096	.010	.047	.100	1.000	1.000	1.000
mixed	.2	0	.829	.947	.972	.874	.960	.980	.998	.999	1.000
mixed	.2	-.5	.909	.976	.988	.919	.978	.989	.551	.776	.863
t_6	0	0	.011	.051	.100	.004	.029	.069	.002	.018	.044
t_6	0	-.5	.008	.052	.104	.009	.043	.096	1.000	1.000	1.000
t_6	.2	0	.557	.770	.855	.845	.964	.984	.976	.997	1.000
t_6	.2	-.5	.747	.898	.944	.959	.992	.997	.414	.721	.842
t_{10}	0	0	.011	.055	.106	.008	.045	.094	.007	.047	.094
t_{10}	0	-.5	.009	.051	.102	.009	.046	.097	1.000	1.000	1.000
t_{10}	.2	0	.535	.757	.850	.722	.887	.941	.962	.992	.999
t_{10}	.2	-.5	.746	.901	.946	.859	.951	.975	.997	1.000	1.000
skewed	0	0	.012	.053	.101	.009	.052	.101	.007	.047	.098
skewed	0	-.5	.010	.047	.102	.010	.045	.095	1.000	1.000	1.000
skewed	.2	0	.519	.757	.835	.861	.959	.982	.993	.998	.999
skewed	.2	-.5	.749	.897	.944	.903	.974	.988	.976	.997	.998

*The working conditional distribution for the hybrid method was normal. The sizes/powers for testing interactions ($H_0 : \xi = 0$) were calculated at nominal levels 0.01, 0.05, and 0.1.

**Underlying conditional distribution: "normal", standard normal distribution; "mixed", mixture of two standard normal distributions with equal weights; "t_6", t distribution on 6 df; "t_{10}", t distribution on 10 df; "skewed", skewed normal distribution with shape parameter 2.

or 0.25, and the correlation coefficient of X and Y was either 0, 0.2, or -0.2. We randomly draw 200 individuals from case population and 200 individuals from control population, and applied the hybrid method, the copula method, and the logistic regression method to the generated data. The simulation results based on 1,000 replications of simulations are reported in Table 3. Overall, the hybrid method has smaller standard errors of the estimates than the copula method, and estimation bias is also smaller for the hybrid method. As for the power for testing interaction, the hybrid method could have greater power than the copula method when the later overestimates the interaction, and vice versa otherwise. As expected, the standard logistic regression method is less efficient (in terms of standard error) and less powerful than both the copula method and the hybrid method.

3.2 Study 2

We evaluate interaction between a continuous covariate X and a categorical covariate $Y = 0, 1, 2$. To be consistent with notation used for gene-environment interaction study, we replace Y by G and X by E.

Again, for given covariates (G, E), we assume that the logistic regression (1) holds. Let

$$\boldsymbol{\eta} := (\eta_1, \eta_2, \eta_3, \eta_4), \tag{11}$$

and we consider the following polytomous regression model for the controls:

$$w_0(e, \boldsymbol{\eta}) := P_0(G = 0|E = e) = \frac{1}{1 + \exp(\eta_1 + \eta_2 e) + \exp(\eta_3 + \eta_4 e)} \tag{12}$$

$$w_1(e, \boldsymbol{\eta}) := P_0(G = 1|E = e) = \frac{\exp(\eta_1 + \eta_2 e)}{1 + \exp(\eta_1 + \eta_2 e) + \exp(\eta_3 + \eta_4 e)} \tag{13}$$

$$w_2(e, \boldsymbol{\eta}) := P_0(G = 2|E = e) = \frac{\exp(\eta_3 + \eta_4 e)}{1 + \exp(\eta_1 + \eta_2 e) + \exp(\eta_3 + \eta_4 e)} \tag{14}$$

Let $f_0(e|g)$ denote the conditional density function of E given $G = g$, then the above model is equivalent to

$$f_0(e|1) = f_0(e|0) \exp(\eta_1^* + \eta_2 e),$$
$$f_0(e|2) = f_0(e|0) \exp(\eta_3^* + \eta_4 e),$$

where $\eta_1^* = \eta_1 + \log\{P(G = 1)/P(G = 0)\}$, $\eta_3^* = \eta_3 + \log\{P(G = 2)/P(G = 0)\}$. In other words, given genotype $G = i, i = 0, 1, 2$, the "environment variable" E has density functions linked by the "exponential tilting model" with the baseline density function $f_0(e|0)$ totally unspecified. There-

Table 3. Estimation and test results with two normal covariates

		logist		hybrid		copula		case-only
ξ	θ^*	bias (se)**	power***	bias (se)**	power***	bias (se)**	power	power***
0	0	.005 (.116)	.049	.003 (.104)	.054	.007 (.104)	.062	.058
0	.2	-.002 (.115)	.046	-.007 (.099)	.048	.018 (.102)	.050	.783
0	-.2	.001 (.112)	.044	.001 (.101)	.044	-.014 (.102)	.045	.823
.25	0	.004 (.126)	.576	.022 (.107)	.776	.057 (.120)	.789	.959
.25	.2	.004 (.123)	.573	.002 (.099)	.765	.070 (.117)	.832	1.000
.25	-.2	.005 (.128)	.553	.000 (.109)	.664	-.010(.108)	.627	.084

*The two covariates X and Y were jointly normal and the common marginal distribution was standard normal. The true values of β and γ in the tilting model (2) were 0.5.
**The correlation coefficient of two standard normal covariates in the control population.
***Estimation bias (standard error) for the interaction effect ξ.
****Size/power for testing interaction effect ($H_0 : \xi = 0$) at nominal level 0.05.

fore, it follows from (4) that the marginal distribution functions $f_1(e)$ and $f_0(e)$ of E for cases and controls satisfy the relationship

$$f_1(e) = \exp\{\alpha + \gamma e + \log \mu_1(e)\} f_0(e),$$

where

$$\begin{aligned}\mu_1(e) :=& \sum_{g=0}^{2} \exp\{(\beta + \xi e)G\} P_0(G = g|E = e) \\ =& w_2 \exp(2\beta + 2\xi e) + w_1 \exp(\beta + \xi e) + w_0,\end{aligned}$$

which is derived according to (5). In virtue of (6), the conditional mass functions $P_1(G = g|E = e)$ and $P_0(G = g|E = e)$ for cases and controls satisfy the relationship

$$\begin{aligned}P_1(G = g|E = e) =& \exp\{\beta g + \xi e g - \log \mu_1(e)\} \\ & \times P_0(G = g|E = e).\end{aligned}$$

As a result, the conditional log-likelihood function (7) is

$$\begin{aligned}\ell_c =& \sum_{i=1}^{n_1} \{G_{i1}\beta + \xi G_{i1} E_{i1} - \log \mu_1(E_{i1})\} \\ &+ \sum_{i=1}^{n} \{I(G_i = 0) \log w_0(E_i, \boldsymbol{\eta}) \\ &\quad + I(G_i = 1) \log w_1(E_i, \boldsymbol{\eta}) \\ &\quad + I(G_i = 2) \log w_2(E_i, \boldsymbol{\eta})\}.\end{aligned}$$

To make sure that the Hardy-Weiberg equilibrium holds true in the control population, we chose the parameters $(\eta_1, \eta_2, \eta_3, \eta_4) = (1.54, 0, 1.195, 1)$. This is equivalent to the choice that in the control population $E|G = 0 \sim N(0, 1)$, $E|G = 1 \sim N(0, 1)$ and $E|G = 2 \sim N(1, 1)$, and $P_0(G = 0) = 0.09$, $P_0(G = 1) = 0.42$, $P_0(G = 2) = 0.49$. The choice of genotype frequencies corresponded to a dominant genetic model. For the parameters of interest (β, γ, ξ), we chose $(1, 1, 0)$. We randomly generated 200 case data and 200 control data and applied to the data the standard logistic regression method, the hybrid method (assuming a polytomous regression), Chatterjee and Carroll's method implemented in the R package CGEN (this method is called CGEN hereafter), and the case-only method. The bias and variance of the estimates and sizes/powers at levels of 0.01, 0.05, and 0.1 based on 1,000 replications of simulations are reported in Table 4. Compared with the logistic regression method without using the auxiliary model information $P_0(G = g|E = e)$, the hybrid method has smaller variances for all parameters. On the other hand, the CGEN method has inflated type one error rates and biased estimations of (β, γ, ξ). This is not surprising as the conditions required by the CGEN method do not hold to be true under our setup.

To test the robustness of the proposed test statistic, we also conducted simulations by assuming a misspecified model $P_0(G = g|E = e)$. The true underlying model for $P_0(G = g|E = e)$ was assumed to be either the proportional odds ratio model or the multi-probit model, but the "working model" is still the polytomous regression model. For the proportional odds ratio model, we used

$$\begin{aligned}P_0(G = 0|E = e) &= \frac{\exp(0 + 1 \times e)}{1 + \exp(0 + 1 \times e)}, \\ P(G \le 1|E = e) &= \frac{\exp(1 + 1 \times e)}{1 + \exp(1 + 1 \times e)},\end{aligned}$$

$$P_0(G = 2|E = e) = 1 - P_0(G = 0|E = e) - P_0(G = 1|E = e).$$

For the choice of $(\beta, \gamma, \xi) = (1, 1, 0)$ or $(1, 1, 0.5)$, the simulation results with sample sizes of $n_1 = n_0 = 200$ based on 1,000 replications are reported in Table 5. We can observe that for the estimation of (β, γ, ξ), the bias due to the misspecification of $P_0(G = g|E = e)$ is small. The type one error rates for testing no interaction between G and E are also close to the nominal levels. Again the CGEN method produces biased results for the estimation of (β, γ, ξ) and inflated type one error rates. This is also due to the fact that the conditions required by CGEN were not satisfied. Similar result was observed when the true underlying model

*Table 4. Estimation and testing results for gene-environment interaction**

method**	result	β	γ	ξ	η_1	η_2	η_3	η_4	power*** .01	.05	.1
	true	1	1	0	1.540	0	1.195	1			
hybrid	bias	.024	.005	.007	.051	.003	.046	.016	.008	.036	.088
	var.	.129	.199	.059	.069	.062	.084	.080			
logist	bias	.050	.066	-.025	.049	-.007	.046	.013	.008	.037	.082
	var.	.147	.270	.081	.079	.081	.090	.092			
CGEN	bias	-.122	-.387	.242	-	-	-	-	.054	.192	.309
	var.	.098	.143	.042							
case-only	bias	-	-	-	-	-	-	-	.939	.985	.992

*In the hybrid method, the true conditional model was the polytomous model determined by (12)–(14) and it was correctly specified.
** "hybrid", the proposed hybrid method; "logist", logistic regression method; "CGEN", the method implemented in the R package CGEN.
***Size/power for testing interaction between gene and environment ($H_0 : \xi = 0$) at nominal levels 0.1, 0.05, 0.01.

*Table 5. Estimation and testing results with misspecified conditional model**

method**	result	β	γ	ξ	power*** .01	.05	.1
	true	1	1	0			
hybrid	bias	.002	-.006	.019	.015	.056	.106
	var.	.023	.034	.018			
logist	bias	.014	.018	.002	.009	.046	.108
	var.	.025	.039	.022			
CGEN	bias	.534	.191	.029	.031	.092	.155
	var.	.051	.056	.032			
method	result	β	γ	ξ	.01	.05	.1
	true	1	1	.5			
hybrid	bias	-.029	-.042	.064	.925	.985	.993
	var.	.032	.042	.021			
logist	bias	.015	.021	.014	.574	.797	.881
	var.	.041	.057	.035			
CGEN	bias	.517	-.204	.369	.986	.993	.996
	var.	.073	.077	.047			

*In the hybrid method, the true conditional model was the proportional odds model, but it was misspecified to be the polytomous regression model determined by (12)–(14).
** "hybrid", the proposed hybrid method; "logist", logistic regression method; "CGEN", the method implemented in the R package CGEN.
***Size/power for testing interaction between gene and environment ($H_0 : \xi = 0$) at nominal levels 0.1, 0.05, 0.01.

$P_0(G = g|E = e)$ was generated from the multinomial probit model (result not shown here).

4. A LUNG CANCER EXAMPLE

It is well known that cigarette smoking is a major risk factor for lung cancer. Recent genome-wide association studies (GWAS) also identified a few chromosome regions (e.g., chromosome 15q25, 515, and 6p21) harboring genetic variants underlying the susceptibility for lung cancer [21, 22, 23, 24, 25, 26, 27]. In particular, the chromosome 15q25 region has been shown to be associated with both lung cancer and smoking behavior. It is of great interest to test whether there is any interaction between the genetic variants in 15q25 and smoking on the risk of lung cancer. We applied the proposed method to evaluating a potential gene-smoking interaction in the Environment and Genetics in Lung Cancer Etiology Study (EAGLE) [27]). In particular, we used the average intensity of cigarettes smoking (in terms of the averaged number of packs per day) defined for ever-smoking subjects in the EAGLE study to represent the smoking behavior, and denoted it as CPD. We evaluated the interaction between CPD and each of the 39 relatively common bi-allelic genetic variants called SNPs (single-nucleotide polymorphisms) within the 15q25 region. Genotypes on these 39 SNPs were obtained from the GWAS on lung cancer [27].

We focused on ever-smokers and removed subjects with missing genotypes. There were a total of 1,738 lung cancer cases, and 1,336 controls in the final dataset. First, we applied the standard logistic regression to evaluating the interaction between CPD and each of the 39 SNPs. The interaction with the SNP rs12912946 had the lowest p-value (p-value = 0.042). Then, we applied our new method to studying the interaction by modeling the relationship between CPD and a SNP with a polytomous regression model. Again, the interaction with the SNP rs12912946 turns out to have the smallest p-value (p-value = 0.014), which is very close to the one produced by CGEN (p-value = 0.012). Noting that the correlation between the SNP rs12912946 and the smoking intensity is not significant at level 0.05 (p-values for η_2 and η_4 are 0.052 and 0.373, respectively), and the Hardy-Weinberg equilibrium holds in the controls (p-value = 0.121), it is not a surprise to see that CGEN performs similarly to ours since the conditions for the validity

*Table 6. Interaction effect of rs12912946 × smoking_intensity on lung cancer**

method	parameter	est.	s.e.	z value	$P(>\|z\|)$
logist	(Intercept)	-.805	.110	-7.293	3.02e-13
	gene	-.268	.138	-1.948	.051
	smoking	1.143	.108	10.544	5.42e-26
	gene × smoking	.276	.136	2.032	.042
hybrid	(intercept)	-1.062	.106	-10.066	7.82e-24
	gene	-.280	.121	-2.320	.020
	smoking	1.138	.103	11.035	2.59e-28
	gene × smoking	.288	.117	2.450	.014
	η_1	-.094	.102	-.925	.355
	η_2	-.209	.108	-1.945	.052
	η_3	-1.882	.199	-9.454	3.27e-21
	η_4	-.190	.213	-.891	.373
CGEN	(intercept)	-.798	.105	-7.596	3.06e-14
	gene	-.283	.119	-2.370	.018
	smoking	1.134	.103	11.019	3.11e-28
	gene × smoking	.293	.117	2.508	.012

*In the hybrid method, the conditional model was specified to be the polytomous regression model determined by (12)–(14).

of CGEN nearly hold. More detailed results are reported in Table 6. Even though the p-value generated by our method is smaller than that from the standard logistic regression model, it is still not significant after the multiple comparison adjustment. This could be due to the limited power, or the possibility that there is no interaction between smoking and genetic variants in 15q25. Further investigations are needed to better understand how the smoking behavior and genes in 15q25 are interacted.

5. DISCUSSIONS

In this paper we have proposed a hybrid approach for testing the interactions between two covariates in case-control studies. By appropriately modelling the conditional distribution of one covariate conditioning on the other one in the control population, we are able to obtain a more powerful test than the one derived from the standard logistic regression model. We choose to model the conditional distribution in the control population instead of in the general population, in order to avoid the thorny issue in the estimation of disease prevalence probability as it is not estimable in a standard case-control study. By the symmetric property, one may also model the conditional distribution in the case population.

Chatterjee and Carroll [7] assumed that in the general population the gene and environment are independent of each other. They showed that the disease prevalence is estimable under this extra assumption. Lin and Zeng [28] assumed a parametric model for Y given X in the general population. If the disease prevalence is low, then approximately our method and Lin and Zeng's method are equivalent. On the other hand, if the disease prevalence is high, both Chatterjee and Carroll's and Lin and Zeng's methods might have a convergence problem unless the true disease prevalence is known.

Furthermore, if the sampling is unbiased (the true disease prevalence rate matches the sampling fraction of $n_1/(n_1+n_0)$), then both Chatterjee and Carroll's method and Lin and Zeng's method have no improvement over the standard logistic regression estimation provided that the distribution $f(x,y)$ does not carry any information on interested parameters. We can demonstrate this as follows.

Suppose there are n individuals and the sampling design is prospective. It ends up to n_1 cases and n_0 controls. Then one takes all cases and controls from n_1 cases and n_0 controls. The full likelihood can be decomposed either prospectively or retrospectively as

$$\begin{aligned}&\prod_{i=1}^{n}\left[\frac{\exp\{d_i(\alpha^*+\beta y_i+\gamma x_i+\xi x_i y_i)\}}{1+\exp(\alpha^*+\beta y_i+\gamma x_i+\xi x_i y_i)}f(y_i,x_i)\right]\\ =&\left\{\prod_{i=1}^{n_1}f(y_{1i},x_{1i}|D=1)\right\}\left\{\prod_{j=1}^{n_0}f(x_{0j},y_{0j}|D=0)\right\}\\ &\times\left\{P^{n_1}(D=1)P^{n_0}(D=0)\right\}\end{aligned}$$

Usually the density function $f(y,x)$ is unrelated to the parameters $(\alpha^*,\beta,\gamma,\xi)$ and the corresponding likelihood is factored out from the prospective likelihood. Therefore, the full likelihood

$$\prod_{i=1}^{n}\left[\frac{\exp\{d_i(\alpha^*+\beta y_i+\gamma x_i+\xi x_i y_i)\}}{1+\exp(\alpha^*+\beta y_i+\gamma x_i+\xi x_i y_i)}f(y_i,x_i)\right]$$

is exactly equivalent to the prospective likelihood even if $f(y,x)$ is completely known. As a result, the retrospective likelihood

$$\left\{\prod_{i=1}^{n_1}f(y_{1i},x_{1i}|D=1)\right\}\left\{\prod_{j=1}^{n_0}f(x_{0j},y_{0j}|D=0)\right\}$$

is usually less informative than the prospective likelihood, and one cannot expect any improvement over the prospective likelihood by using the retrospective likelihood and auxiliary information on $f(y,x)$.

Since in the proposed model we have made a parametric assumption for the conditional density function of Y given X directly in the case or control population, our method could produce improved estimation for the underlying parameters when the sampling is biased. This fact is observed in our simulation studies.

Moreover, the proposed method has potential applications in secondary outcome analysis where one may be interested in studying the relationship between Y and X in the control (or case) population. The proposed method can

utilize information from both cases and controls effectively, even though the focus is on the model in controls.

In practice, the underlying conditional model for a covariate given another one is usually unknown. One may adopt a two-step approach. In the first step, a model of most likelihood for the covariates is identified using some kind of model checking technique; in the second step, the hybrid method can be applied by assuming the identified conditional model.

APPENDIX

We give proofs for the theorems in Section 2. First, we present some regularity conditions for Theorems 1 to 3:

1) $f_0(y|x,\boldsymbol{\eta})$ and $f_1(y|x,\boldsymbol{\omega}_1)$ satisfy the regularity conditions given by Lehmann (1983, Chapter 6) on the normality of the maximum likelihood estimator in fully parametric models.
2) The regression model $P(D=0|x,y)=\{1+\exp(\alpha+\beta y+\gamma x+\xi xy)\}^{-1}$ satisfies regularity conditions for the standard logistic regression model.
3) $\min(n_0,n_1)\to\infty$ and $n_1/n_0\to\rho$, where $0<\rho<1$.

Proof of Theorem 1. The parameters of interested are $\boldsymbol{\omega}=(\alpha,\beta,\gamma,\boldsymbol{\eta},\xi)$. Denote $\Omega_1=(\beta,\gamma,\boldsymbol{\eta},\xi)$. Notice that the log hybrid likelihood can be decomposed as

$$\ell_H=\ell_c+p\ell_M, \tag{15}$$

where

$$\ell_c=\sum_{i=1}^{n_1}\log f_1(y_{i1}|x_{i1},\boldsymbol{\omega}_1)+\sum_{i=1}^{n_0}\log f_0(y_{i0}|x_{i0},\boldsymbol{\eta})$$

is the conditional log likelihood and

$$p\ell_M=\sum_{i=1}^{n_1}\{\alpha+\phi(x_{i1},\boldsymbol{\omega}_1)\}-\sum_{i=1}^{n}\log[1+\rho\exp\{\alpha+\phi(x_i,\boldsymbol{\omega}_1)\}]$$

is the profile marginal log likelihood. Differentiating ℓ_H with respect to $\boldsymbol{\omega}$, we have

$$\frac{\partial\ell_H}{\partial\boldsymbol{\omega}}=\frac{\partial\ell_c}{\partial\boldsymbol{\omega}}+\frac{\partial p\ell_M}{\partial\boldsymbol{\omega}},$$

where

$$\frac{\partial\ell_c}{\partial\boldsymbol{\omega}}=\sum_{i=1}^{n_1}\frac{\partial\log f_1(y_{i1}|x_{i1},\boldsymbol{\omega}_1)}{\partial\boldsymbol{\omega}}+\sum_{i=1}^{n_0}\frac{\partial\log f_0(y_{i0}|x_{i0},\boldsymbol{\eta})}{\partial\boldsymbol{\omega}}.$$

Let

$$g=\frac{\partial\ell_H}{\partial\boldsymbol{\omega}}=\frac{\partial\ell_c}{\partial\boldsymbol{\omega}}+\frac{\partial p\ell_M}{\partial\boldsymbol{\omega}}$$

be the score estimating equation. Since $E[\ell_c|x_1,..,x_n]=0$, the two terms in g are orthogonal to each other. By the standard result for the parametric likelihood, one has that

$$n^{-1/2}\frac{\partial\ell_c(\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\to N(0,V_c)\quad\text{in distribution}$$

where

$$V_c=\rho_1E\left(\frac{\partial\log f_1(y|x,\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\frac{\partial\log f_1(y|x,\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}^T}\right)+\rho_0E\left(\frac{\partial\log f_0(y|x,\eta_0)}{\partial\boldsymbol{\omega}}\frac{\partial\log f_0(y|x,\eta_0)}{\partial\boldsymbol{\omega}^T}\right).$$

On the other hand, using the results in [13], we have

$$n^{-1/2}\frac{\partial p\ell_M(\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}}\to N(0,V_M)\quad\text{in distribution},$$

where

$$V_M=\frac{\rho}{1+\rho}A-\rho\begin{pmatrix}A_0\\A_1^T\end{pmatrix}(A_0,A_1),\quad A=\begin{pmatrix}A_0&A_1\\A_1^T&A_2\end{pmatrix},$$

and

$$A_0=\int\frac{\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}{1+\rho\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}dF_0(x),$$

$$A_1=\int\frac{\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}{1+\rho\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}\frac{\partial\phi(x,\boldsymbol{\omega}_1)}{\partial\boldsymbol{\omega}_1}dF_0(x),$$

$$A_2=\int\frac{\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}{1+\rho\exp\{\alpha+\phi(x,\boldsymbol{\omega}_1)\}}\frac{\partial\phi(x,\boldsymbol{\omega}_1)}{\partial\boldsymbol{\omega}_1}\frac{\partial\phi(x,\boldsymbol{\omega}_1)}{\partial\boldsymbol{\omega}_1^T}dF_0(x).$$

Furthermore, by using the information identity for the parametric likelihood model, we have

$$n^{-1}\frac{\partial^2\ell_c}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\to-V_c\quad\text{in probability.}$$

Similar to [13], we can derive that

$$n^{-1}\frac{\partial^2 p\ell_M}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\to-\frac{\rho}{1+\rho}A\quad\text{in probability.}$$

Write

$$g(\boldsymbol{\omega})=\frac{1}{\sqrt{n}}\frac{\partial\ell_H(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}.$$

Finally by expanding $g(\hat{\boldsymbol{\omega}})$ at $\boldsymbol{\omega}_0$, we have

$$n^{1/2}(\hat{\boldsymbol{\omega}}-\boldsymbol{\omega}_0)=-\left(\frac{1}{n}\frac{\partial^2\ell_H(\boldsymbol{\omega}_0)}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^T}\right)^{-1}g(\boldsymbol{\omega}_0)+o_p(1). \tag{16}$$

Easily we can show that

$$n^{1/2}(\hat{\boldsymbol{\omega}}-\boldsymbol{\omega}_0)\to N(0,\Sigma)\quad\text{in distribution},$$

where

$$\Sigma=(V_c+\rho A/(1+\rho))^{-1}(V_c+V_M)(V_c+\rho A/(1+\rho))^{-1}. \tag{17}$$

□

Proof of Theorem 2. Denote $\boldsymbol{\omega}_2=(\alpha,\beta,\gamma,\boldsymbol{\eta})$ and $\hat{\boldsymbol{\omega}}=(\hat{\boldsymbol{\omega}}_2,\hat{\xi})$. Let $\tilde{\boldsymbol{\omega}}_2$ be the constrained maximum hybrid likelihood estimation of $\boldsymbol{\omega}_2$, i.e. it maximizes $\ell_H(\boldsymbol{\omega}_2,\xi_0)$, where

ξ_0 is the true interaction effect between Y and X. Denote $\tilde{\boldsymbol{\omega}} = (\tilde{\boldsymbol{\omega}}_2, \xi_0)$.

Expand $\ell_H(\tilde{\boldsymbol{\omega}})$ at $\hat{\boldsymbol{\omega}}$, we have

$$\ell_H(\tilde{\boldsymbol{\omega}}) - \ell_H(\hat{\boldsymbol{\omega}}) = \frac{1}{2}(\tilde{\boldsymbol{\omega}} - \hat{\boldsymbol{\omega}})^T \frac{\partial^2 \ell_H(\hat{\boldsymbol{\omega}})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T}(\tilde{\boldsymbol{\omega}} - \hat{\boldsymbol{\omega}}) + o_p(1).$$

By using regularity conditions for uniform convergence, one can show that

$$\frac{1}{n}\frac{\partial^2 \ell_H(\hat{\boldsymbol{\omega}})}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^T} \to U = -V_c - \rho A/(1+\rho) =: \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$$

in probability, which together with (16) follow that

$$\sqrt{n}(\hat{\boldsymbol{\omega}} - \boldsymbol{\omega}_0) = -U^{-1} g(\boldsymbol{\omega}_0) + o_p(1), \tag{18}$$

Expanding $\partial \ell_H(\tilde{\boldsymbol{\omega}}_2, 0)/\partial \boldsymbol{\omega}_2 = 0$ at $\boldsymbol{\omega}_2^0 = (\alpha_0, \beta_0, \gamma_0, \eta_0)$, one has

$$\sqrt{n}(\tilde{\boldsymbol{\omega}}_2 - \boldsymbol{\omega}_2^0) = -U_{11}^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell_H(\boldsymbol{\omega}_0)}{\partial \boldsymbol{\omega}_2} + o_p(1).$$

This can be written as

$$\begin{pmatrix} \sqrt{n}(\tilde{\boldsymbol{\omega}}_2 - \boldsymbol{\omega}_2^0) \\ 0 \end{pmatrix} = - \begin{pmatrix} U_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} g(\boldsymbol{\omega}_0) + o_p(1). \tag{19}$$

Taking the difference between (18) and (19), one has

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\omega}} - \tilde{\boldsymbol{\omega}}) &= U^{-1} \left[I - U \begin{pmatrix} U_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right] g(\boldsymbol{\omega}_0) \\ &:= U^{-1} B g(\boldsymbol{\omega}_0), \end{aligned}$$

where I is an identity matrix and

$$B = I - U \begin{pmatrix} U_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ -U_{21} U_{11}^{-1} & I \end{pmatrix}.$$

Therefore, the hybrid likelihood ratio statistic is

$$R_1(\xi_0) = -2[\ell_H(\tilde{\boldsymbol{\omega}}) - \ell_H(\hat{\boldsymbol{\omega}})] = -g^T(\boldsymbol{\omega}_0) B^T U^{-1} B g(\boldsymbol{\omega}_0).$$

Let $W = -B^T U^{-1} B$. In order to show Theorem 2, we only need to prove the conditions in Ogasawara-Takahashi's Theorem [29] (page 188) hold true, i.e.

$$VWVWV = VWV, \ \ \text{rank}(WV) = p,$$

where $V = V_M + V_c$. This can be done by the standard matrix algebra method. □

Proof of Theorem 3. The proof is similar to that of Theorem 2, so we omit the details. □

Received 27 February 2014

REFERENCES

[1] Anderson J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**:19–35. MR0345332

[2] Prentice R. L. and Pyke R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **13**:403–411. MR0556730

[3] Robinowitz D. (1997). A note on efficient estimation from case-control data. *Biometrika* **84**:486–488.

[4] Breslow N. E., Robins J. M., and Wellner J. A. (2000). On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli* **6**:447–455. MR1762555

[5] Begg C. B. and Zhang Z. F. (1994). Statistical analysis of molecular epidemiology studies employing case-series. *Cancer Epidemiol Biomarkers Prevention* **3**:173–175.

[6] Piegorsch W. W., Weinberg C. R., and Taylor J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine* **13**:153–162.

[7] Chatterjee N. and Carroll R. J. (2005). Semiparametric maximum likelihoodestimation exploiting gene-environment independence in case-control studies. *Biometrika* **92**:399–418. MR2201367

[8] Shin J. H., McNeney B., and Graham J. (2007). Case-control inference of interaction between genetic and nongenetic risk factors under assumptions on their distribution. *Statistical Applications in Genetics and Molecular Biology* **6**:13. MR2306948

[9] Albert P. S., Ratnastingle D. T. J., and Wacholder S. (2001). Limitations of the case-only design for identifying gene-environment interaction. *American Journal of Epidemiology* **154**:687–693.

[10] Mukherjee B. and Chatterjee N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* **64**:685–694. MR2526617

[11] Zhang H., Qin J., and Landi M., et al. (2013). A copula-model based semiparametric interaction test under the case-control design. *Statistica Sinica* **23**:1505–1521. MR3222807

[12] Owen A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC: Noca Raton.

[13] Qin J. and Zhang B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika* **84**:609–618. MR1603924

[14] Chen J. and Liu Y. (2013). Quantile and quantile-function estimations under density ratio model. *Annals of Statistics* **41**:1055–1692. MR3113825

[15] Cox D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, 506–527, (eds, Johnson N. L., Smith H.), Wile-Interscience: New York.

[16] Patil G. P. and Rao C. R. (1977). The weighted distributions: A survey of their applications. In *Applications of Statistics*, 383–405, (ed, Krishnaiah P. R.), North-Holland Publishing Company: Amsterdam.

[17] Patil G. P. and Rao C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34**:179–189. MR0507202

[18] Vardi Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics* **10**:616–620. MR0653536

[19] Nair V. N. and Wang P. C. C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31**:423–436. MR1041563

[20] Azzalini A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**:171–178. MR0808153

[21] Hung R. J., McKay J. D., and Gaborieau V., et al. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**:633–637.

[22] Amos C. I., Wu X., and Broderick P., et al. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature Genetics* **40**:616–622.

[23] Thorgeirsson T. E., Geller F., and Sulem P., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**:638–642.

[24] McKay J. D., Hung R. J., and Gaborieau V., et al. (2008). Lung cancer susceptibility locus at 5p15.33. *Nature Genetics* **40**:1404–1406.

[25] Wang Y., Broderick P., and Webb E., et al. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature Genetics* **40**:1407–1409.

[26] Rafnar T., Sulem P., and Stacey S. N., et al. (2009). Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nature Genetics* **41**:221–227.

[27] Landi M. T., Chatterjee N., and Yu K., et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *American Journal of Human Genetics* **85**:679–691.

[28] Lin D. and Zeng D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology* **33**:256–265.

[29] Rao R. C. (1973). *Linear Statistical Inference and Its Applications*. John Wiley: New York. MR0346957

Jing Qin
Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases, NIH
Bethesda, MD
USA
E-mail address: jingqin@niaid.nih.gov

Hong Zhang
Institute of Biostatistics
School of Life Sciences
Fudan University
Shanghai
PRC
E-mail address: zhanghfd@fudan.edu.cn

Maria Landi
Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH
Bethesda, MD
USA
E-mail address: landim@mail.nih.gov

Neil Caporaso
Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH
Bethesda, MD
USA
E-mail address: caporasn@mail.nih.gov

Kai Yu
Division of Cancer Epidemiology and Genetics
National Cancer Institute, NIH
Bethesda, MD
USA
E-mail address: yuka@mail.nih.gov