Jackknife empirical likelihood method for case-control studies with gene-environment independence on controls

BING-YI JING, ZHOUPING LI*, JING QIN AND WANG ZHOU

In this paper, we propose a jackknife empirical likelihood method to do inference for the interested parameters of the multiplicative-intercept risk models by taking into account the gene-environment independence on controls in case-control studies. It is shown that the proposed statistic is asymptotically chi-squared distributed. Simulation studies investigate the small-sample properties. A real example is also given.


Keywords and phrases: Case-control study, Gene-environment independence, Jackknife empirical likelihood, Logistic regression, Multiplicative-intercept risk model, U-statistics.

1. INTRODUCTION

The case-control study, where sampling is conditioned on the presence (cases) and absence (controls) of a disease, is an efficient way of studying the effect of risk factors to rare diseases. As the regression coefficients have a desired interpretation in terms of log-odds ratios which are estimable based on case-control samples, logistic regression models have been widely applied to analyze binary data arising from case-control studies in epidemiology (see, e.g., Breslow, 1996).

To be precise, suppose that $D$ is a binary response variable of presence ($D = 1$), or absence ($D = 0$) of a disease, and $X$ is the associated explanatory variables. We consider the general logistic regression model, also known as the multiplicative-intercept risk model:

$$\text{Pr}(D = 1|X = x) = \frac{\exp\{\alpha^* + m(x, \beta)\}}{1 + \exp\{\alpha^* + m(x, \beta)\}},$$

where $m(x, \beta)$ is a given function, $\alpha^*$ is a scale parameter, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a p-vector parameter, and the marginal distribution of $X$ is unspecified. As in Prentice and Pyke (1979), the case-control samples are two independent groups of samples based on model (1), i.e., $X_1, \ldots, X_{n_0}$ are a random sample (controls) from $P(x|D = 0)$ and $X_{n_0+1}, \ldots, X_n$ are another random sample (cases) from $P(x|D = 1)$.

Let $f(x|D = 0)$ and $f(x|D = 1)$ be the conditional density functions of the controls and cases. Under the logistic assumptions, it can be shown that

$$f(x|D = 1) = f(x|D = 0) \exp\{\alpha + m(x, \beta)\},$$

where $\alpha = \alpha^* + \log\{P(D = 0)/P(D = 1)\}$. Therefore, model (1) is equivalent to the following two-sample semiparametric density ratio model:

$$\begin{align*}
(2) & \quad X_1, \ldots, X_{n_0} \text{ are independent with density } f(x|D = 0), \\
& \quad X_{n_0+1}, \ldots, X_n \text{ are independent with density } \\
& \quad \exp\{\alpha + m(x, \beta)\} f(x|D = 0).
\end{align*}$$

Notice that model (2) is considerably flexible. For example, it is a biased sampling model with weight function $\exp\{\alpha + m(x, \beta)\}$, where $\alpha$ and $\beta$ are unknown (see, e.g., Vardi, 1982; Gill et al., 1988; Qin, 1993). If $m(x, \beta) = \beta^T x$ with $x$ being a $p \times 1$ random vector, model (2) reduces to the standard logistic model (see, e.g., Anderson, 1979; Breslow and Day, 1980). Model (2) is also related to the Cox proportional hazards model (see, Qin, 1998).

There is extensive literature on model (2) under the case-control sampling plans. For example, O’Neill (1980) studied the discrimination problem in (2); Cox and Ferry (1991) and Weinberg and Wacholder (1993) investigated this model mainly for discrete data; Qin (1998) employed empirical likelihood to model (2), derived the asymptotic normality and further showed the optimality in the sense of Godambe (1960). Related work also has been done for the standard logistic regression model, references include Prentice and Pyke (1979), Anderson and Blair (1982), Wang and Carroll (1993), Qin and Zhang (1997), and Breslow et al. (2000) among others.

Our research is motivated by epidemiological studies on gene-environment interaction problems. As the genetic susceptibility and environmental (or non-genetic) exposures play an interactive role in etiology of many diseases, epidemiologists often seek to examine how these two risk factors affect a disease through case-control studies. Intuitively, an individual’s genetic susceptibility is determined since birth, and may not relate to his/her environmental exposures. Recent studies showed that the assumption of independence between gene and environmental exposures may be reason-
able; see Hwang et al. (1994) for example. This has attracted some researchers’ attention, and they have developed new methods for analyzing case-control data in these occasions; see Piegorsch et al. (1994), Umbach and Weinberg (1997), Modan et al. (2001). However, all these methods have their own drawbacks, see Chatterjee and Carroll (2005) for more discussion. For instance, Chatterjee and Carroll (2005) considered the case-control problem when the underlying populations satisfy gene-environment independence but assume that the populations contain only finite values. Thus, seeking a novel method to this case becomes important.

As pointed out by Piegorsch et al. (1994), assumptions that genotype and exposure are independent in the population and that the disease is rare, in fact, imply that genotype and exposure are (approximately) independent in the non-diseased population. Naturally, a method that could make full use of such information will invariably increase efficiency.

It is well-known that empirical likelihood approach of Owen (1988, 1990, 2001) can easily incorporate such information in the model, in addition to its many other nice properties. In fact, the empirical likelihood to case-control studies has been investigated by Qin (1998) when there is no auxiliary information. When applying the procedure of Qin (1998) to profile out the nuisance parameters, unfortunately, this method does not produce a closed form for this problem.

To overcome the difficulties encountered by Owen’s direct application of empirical likelihood, we consider the jackknife empirical likelihood (JEL) method, proposed recently by Jing, Yuan and Zhou (2009). To do this, we define

\[ L(\alpha, \beta) = \log \left( \prod_{i=1}^{n_0} (p_{i,1}p_{i,2}) \prod_{k=n_0+1}^{n} q_k \right) \]

\[ = \sum_{j=1}^{2} \sum_{i=1}^{n_0} \log p_{i,j} + \sum_{k=n_0+1}^{n} \log q_k \]

subject to the constraints

\[ \sum_{i=1}^{n_0} p_{i,j} = 1 \quad \text{and} \quad \sum_{k=n_0+1}^{n} q_k = 1, \quad p_{i,j}, q_k \geq 0, \quad j = 1, 2, \]

\[ \sum_{i_1=1}^{n_0} \sum_{i_2=1}^{n_0} p_{i_1,1}p_{i_2,2} \{ \exp \{ \alpha + m(X_{i_1,1}, X_{i_2,2}, \beta) \} - 1 \} = 0, \]

\[ \sum_{k=n_0+1}^{n} q_k \{ \exp \{ -\alpha - m(X_{k,1}, X_{k,2}, \beta) \} - 1 \} = 0. \]

The constraints (5) and (6) come respectively from

\[ 1 = \int f(x_i|D=1)dx_i = \int \exp \{ \alpha + m(x_i, \beta) \} dF(x_i|D=0) \]

\[ = \int \exp \{ \alpha + m(x_i, \beta) \} dF(x_{i,1}|D=0)dF(x_{i,2}|D=0), \]

and

\[ 1 = \int f(x_i|D=0)dx_i = \int \exp \{ -\alpha - m(x_i, \beta) \} dF(x_i|D=1). \]

### 2.2 Jackknife empirical likelihood

Due to its highly nonlinear nature of the (5) and (6), direct application of Owen’s empirical likelihood will encounter some serious computational difficulties. To overcome this, we consider using the jackknife empirical likelihood method, proposed by Jing, Yuan and Zhou (2009). To do this, we define

\[ U_1(\alpha, \beta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} h_1(X_{i,1}, X_{j,2}; \alpha, \beta), \]

\[ U_2(\alpha, \beta) = \frac{1}{n_1} \sum_{k=1}^{n_1} h_2(X_{n_0+k}; \alpha, \beta) \]

with respective kernels

\[ h_1(X_{i,1}, X_{j,2}; \alpha, \beta) = \exp \{ \alpha + m(X_{i,1}, X_{j,2}, \beta) \} - 1, \]

\[ h_2(X_k; \alpha, \beta) = \exp \{ -\alpha - m(X_{k,1}, X_{k,2}, \beta) \} - 1. \]

Note that \( U_1(\alpha, \beta) \) is a two-sample \( U \)-statistic while \( U_2(\alpha, \beta) \) is a one-sample \( U \)-statistic.
However, $U_1(\alpha, \beta)$ can be rewritten expressed as a one-sample U-statistic:

$$U_1(\alpha, \beta) = \left(\frac{2n_0}{2}\right)^{-1} \sum_{1 \leq i < j \leq 2n_0} H_1(Z_{1,i}, Z_{1,j}; \alpha, \beta),$$

where

$$Z_{1,i} = \begin{cases} X_{i,1}, & i = 1, \ldots, n_0, \\ X_{i-n_0,2}, & i = n_0 + 1, \ldots, 2n_0, \\ \end{cases}$$

$$H_1(Z_{1,i}, Z_{1,j}; \alpha, \beta) = \begin{cases} (2n_0 - 1)h_1(Z_{1,i}, Z_{1,j}; \alpha, \beta) / n_0, & 1 \leq i \leq n_0 < j \leq 2n_0, \\ 0, & \text{otherwise}. \end{cases}$$

Define the jackknife pseudo-values by

$$\begin{align*}
\hat{V}_1(i, \alpha, \beta) &= 2n_0 U_1(i, \alpha, \beta) - (2n_0 - 1)U_1^{(-i)}(\alpha, \beta), \\
&= \begin{cases} (2n_0 - 1)h_1(Z_{1,i}, Z_{1,j}; \alpha, \beta) / n_0, & 1 \leq i \leq n_0 < j \leq 2n_0, \\ 0, & \text{otherwise}, \end{cases} \\
\hat{V}_2(i, \alpha, \beta) &= n_1 U_2(i, \alpha, \beta) - (n_1 - 1)U_2^{(-i)}(\alpha, \beta), \\
&= \begin{cases} (2n_0 - 1)h_1(Z_{1,i}, Z_{1,j}; \alpha, \beta) / n_0, & 1 \leq i \leq n_0 < j \leq 2n_0, \\ 0, & \text{otherwise}. \end{cases}
\end{align*}$$

where $U_1^{(-i)}$ is calculated from $U_1$ using $Z_{1,1}, \ldots, Z_{1,i-1}, Z_{1,i+1}, \ldots, Z_{1,2n_0}$ after deleting the $i$th data value $Z_{1,i}$, while $U_2^{(-i)}$ is computed from $X_{n_0+1}, \ldots, X_{n_0+k-1}, X_{n_0+k+1}, \ldots, X_n$ after deleting the $k$th data value $X_{n_0+k}$.

It can be seen that

$$U_1(\alpha, \beta) = \frac{1}{2n_0} \sum_{i=1}^{2n_0} \hat{V}_1(i, \alpha, \beta),$$

$$U_2(\alpha, \beta) = \frac{1}{n_1} \sum_{j=1}^{n_1} \hat{V}_2(j, \alpha, \beta).$$

Using the Lagrange multiplier method, we have that $L(\alpha, \beta)$ attains its maximum at

$$p_i = \frac{1}{2n_0} \frac{1}{1 + \lambda_1 \hat{V}_1(i, \alpha, \beta)}, \quad \text{and} \quad q_j = \frac{1}{n_1} \frac{1}{1 + \lambda_2 \hat{V}_2(j, \alpha, \beta)}.$$

This, in turn, gives the $-2 \times$ log-likelihood ratio as

$$l(\alpha, \beta) = 2 \sum_{i=1}^{2n_0} \log \{1 + \lambda_1 \hat{V}_1(i, \alpha, \beta)\} + 2 \sum_{j=1}^{n_1} \log \{1 + \lambda_2 \hat{V}_2(j, \alpha, \beta)\},$$

where $\lambda_1$ and $\lambda_2$ satisfy

$$Q_{1n_0n_1}(\alpha, \alpha, \beta) := \frac{1}{2n_0} \sum_{i=1}^{2n_0} \hat{V}_1(i, \alpha, \beta) = 0,$$

$$Q_{2n_0n_1}(\alpha, \alpha, \beta) := \frac{1}{n_1} \sum_{j=1}^{n_1} \hat{V}_2(j, \alpha, \beta) = 0.$$

Since we are interested in $\beta$, we minimize $l(\alpha, \beta)$ with respect to $\alpha$ and obtain the profile likelihood ratio

$$l(\beta) = 2 \sum_{i=1}^{2n_0} \log \{1 + \lambda_1 \hat{V}_1(i, \hat{\alpha}, \beta)\} + 2 \sum_{j=1}^{n_1} \log \{1 + \lambda_2 \hat{V}_2(j, \hat{\alpha}, \beta)\},$$

where $\hat{\alpha} = \arg\min_{\alpha} l(\alpha, \beta)$.

### 2.3 Main results

Before stating the main results, we list two assumptions:

**Assumption 1.** $\lim_{n_0, n_1 \to \infty} n_1/2n_0 = \rho < \infty$, and denote $\alpha_0$ and $\beta_0$ as the true values of $\alpha$ and $\beta$, respectively.

**Assumption 2.** $h_1(X_1, X_2; \alpha, \beta_0)$ and $h_2(X_1; \alpha, \beta_0)$ are bounded by some integrable function $G(X_1, X_2)$ for $|\alpha - \alpha_0| \leq n_0^{-1/3}$.

Our main results are as follows:

**Proposition 1.** Assume that Assumptions 1–2 hold. Then, with probability tending to one, $l(\hat{\alpha}, \alpha_0)$ attains its minimum at some point $\hat{\alpha}$ in the interior of $|\alpha - \alpha_0| \leq n_0^{-1/3}$. Moreover, $\hat{\alpha}$, $\lambda_1$ and $\lambda_2$ in (11) satisfy (9), (10) and

$$Q_{3n_0n_1}(\alpha, \alpha_1, \alpha_2) := \frac{1}{2n_0} \sum_{i=1}^{2n_0} \lambda_1 \hat{V}_1(i, \alpha, \beta_0) / \partial \alpha + \frac{1}{n_1} \sum_{j=1}^{n_1} \lambda_2 \hat{V}_2(j, \alpha, \beta_0) / \partial \alpha = 0.$$

**Theorem 1.** Assume that Assumptions 1–2 hold. We have $l(\hat{\alpha}, \beta_0) \Rightarrow \chi^2_2$, as $n \to \infty$, where $\hat{\alpha}$ is given in Proposition 1.
From Theorem 1, a confidence interval of $\beta_0$ with level $\gamma$ can be obtained as

$$I_\gamma(r) = \{\beta: l(\hat{\alpha}, \beta) \leq \chi^2_{\gamma, \gamma}\},$$

where $\chi^2_{\gamma, \gamma}$ is the $\gamma$th quantile of $\chi^2_\gamma$.

3. SIMULATIONS

In this section, we conduct a simulation study to evaluate the finite sample performances of the proposed jackknife empirical likelihood confidence interval for the interested parameter. We consider the following experiment, the components of the cases are not independent and follow a bivariate normal distribution, i.e., $(Y_{1,1}, Y_{1,2}) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. That is,

$$f(x_1, x_2|D = 0) = \frac{1}{2\pi\sigma_1\sqrt{1 - \rho^2}} \exp\left\{ \frac{(x_1 - \mu_1)^2}{2(1 - \rho^2)\sigma_1^2} \right\} \times \frac{1}{2\pi\sigma_2\sqrt{1 - \rho^2}} \exp\left\{ \frac{(x_2 - \mu_2)^2}{2(1 - \rho^2)\sigma_2^2} \right\},$$

where

$$f(x_1, x_2|D = 1) = \exp(\alpha + m(x, \beta)) f(x_1, x_2|D = 0),$$

where

$$\alpha = \log\sqrt{1 - \rho^2} + \frac{\rho \mu_1 \mu_2}{(1 - \rho^2)\sigma_1\sigma_2},$$

$$m(x_1, x_2, \beta) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$$= -\frac{\rho \mu_2}{(1 - \rho^2)\sigma_1\sigma_2} x_1 - \frac{\rho \mu_1}{(1 - \rho^2)\sigma_1\sigma_2} x_2 + \frac{\rho}{(1 - \rho^2)\sigma_1\sigma_2} x_1 x_2.$$

In our simulation, we consider $\rho = 0.2$ and 0.5. For each setup, we put $\mu_1 = 1$, $\mu_2 = 0.5$, $\sigma_1 = \sigma_2 = 1$, and generate 1,000 cases and 1,000 controls with sample sizes $n_0 = n_1 = 20, 40, 70$, respectively. We employ the package 'emplik' in the software $R$ to compute the coverage probabilities of the proposed jackknife empirical likelihood method.

For comparisons, we also computed the coverage probabilities of the logistic regression method not using the independent information in controls while they actually have, i.e., all of the setup are the same as before, and the cases and controls are corresponding to 1 and 0, respectively, we employ the $R$ package “glm” to do the logistic regression. Thus, the confidence region for $\beta = (\beta_1, \beta_2, \beta_3)$ with confidence level $\gamma$ is given by

$$\{\hat{\beta}: (\hat{\beta} - \beta_0)^T \hat{\Sigma}^{-1} (\hat{\beta} - \beta_0) \leq \chi^2_{\gamma, \gamma}\},$$

where $\chi^2_{\gamma, \gamma}$ is the $\gamma$th quantile of $\chi^2_\gamma$. $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ and the variance-covariance matrix $\hat{\Sigma}$ for $\hat{\beta}$ can be obtained through the package simultaneously.

In Table 1, we report the coverage probabilities of the 90% and 95% confidence regions for $\beta_0$ using JEL and logistic regression. From these simulation results, we observe:

- As sample size increases, the coverage probabilities in both cases converge to the nominal levels.
- The JEL method performs better than logistic regression method which does not use the independent among controls in most cases.
- The JEL tends to have under-coverage probabilities while the logistic regression tend to have over-coverage probabilities. However, over-coverage probabilities are usually achieved by too wide confidence regions, which are less desirable.

4. A REAL EXAMPLE

We apply our method to the data set of Cameron and Pauling (1978). See Table 33.1 in Andrew and Herzberg (1985). The survival times of 100 terminal cancer patients (cases) and 1,000 matched controls were recorded. To be specific, the treated group of 100 patients with terminal cancer of various kinds began ascorbate treatment, while the controls received the same treatment as the treated group except for the ascorbate. This data set comprises the survival times of the ascorbate-treated patients after the date of first hospital attendance for the cancer that became untreatable, their survival times measured from the dates of untreatability, and the corresponding mean values for the matched controls. We are interested in whether supplemental ascorbate prolongs the survival times of patients with terminal cancer, and we use JEL to estimate the parameters involved. We consider $m(x, \beta) = m(x_1, x_2, \beta)$ in (1) as

$$m(x, \beta) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2.$$ 

References on density ratio model selection include Qin and Zhang (1997), Fokianos (2007), among others. The survival times of patients of stomach cancer and breast cancer (Table 33.1) were selected for illustrating purposes. Here we transform the survival times in days to years, using the package simultaneous.

Table 1. Coverage probabilities of the 90% and 95% confidence intervals for $(\beta_1, \beta_2, \beta_3)$

<table>
<thead>
<tr>
<th>$(n_0, n_1, \rho)$</th>
<th>JEL</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(20, 20, 0.2)$</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>$(20, 20, 0.5)$</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>$(40, 40, 0.2)$</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>$(40, 40, 0.5)$</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>$(70, 70, 0.2)$</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>$(70, 70, 0.5)$</td>
<td>0.92</td>
<td>0.92</td>
</tr>
</tbody>
</table>

296 B.-Y. Jing et al.
### Table 2. 90% and 95% confidence intervals for \((\beta_1, \beta_2, \beta_3, \beta_4)\)

<table>
<thead>
<tr>
<th>(\beta)</th>
<th>Breast</th>
<th>Stomach</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\gamma = 0.90)</td>
<td>((-59.24426, -58.34373))</td>
<td>((-59.28645, -57.86389))</td>
</tr>
<tr>
<td>(\beta_1)</td>
<td>((70.54456, 71.44511))</td>
<td>((70.50234, 71.92042))</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>((-14.94648, -14.04593))</td>
<td>((-14.98869, -13.57061))</td>
</tr>
<tr>
<td>(\beta_3)</td>
<td>((17.56186, 18.46241))</td>
<td>((17.51964, 18.93772))</td>
</tr>
</tbody>
</table>

5. PROOFS

For simplicity, we give some notations. Put

\[
\begin{align*}
U_1(\alpha) &= U_1(\alpha, \beta_0), \quad l = 1, 2 \\
U_1^2(\alpha) &= \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha, \beta_0) \\
U_2^2(\alpha) &= \frac{1}{n_1} \sum_{j=1}^{n_1} \tilde{V}_2(\alpha, \beta_0) \\
V_1(\alpha) &= \text{the original statistics based on } X_1, \ldots, X_{n_0} \\
V_1(\alpha) &= n_0 U_1(\alpha) - (n_0 - 1) U_1^{l}(\alpha) \\
V_1(\alpha) &= \text{the statistics after deleting } X_{1,1} \\
V_1(\alpha) &= n_0 U_1(\alpha) - (n_0 - 1) U_1^{l}(\alpha) \\
V_1(\alpha) &= \text{the statistics after deleting } X_{1,2} \\
\sigma_1^2(\alpha) &= \text{Var}(g_{1,0}(X_{1,1}, \alpha, \beta_0)) \\
\sigma_2^2(\alpha) &= \text{Var}(g_{1,0}(X_{1,2}, \alpha, \beta_0)) \\
\sigma_3^2(\alpha) &= \text{Var}(g_{1,0}(X_{1,3}, \alpha, \beta_0)) \\
\sigma_4^2(\alpha) &= \text{Var}(g_{1,0}(X_{1,4}, \alpha, \beta_0)) \\
\end{align*}
\]

It is known that the consistent estimator of \(\text{Var}(U_1(\alpha))\) is (see Arvesen, 1969)

\[
\text{Var}_{\text{Jack}}(U_1(\alpha)) = \frac{1}{n_0(n_0 - 1)} \sum_{i=1}^{n_0} (V_1(\alpha) - \bar{V}_1(\alpha))^2 \\
+ \frac{1}{n_0(n_0 - 1)} \sum_{j=1}^{n_0} (V_1(\alpha) - \bar{V}_1(\alpha))^2
\]

where \(\bar{V}_1(\alpha)\) and \(\bar{V}_1(\alpha)\) are the averages of \(V_1(\alpha)\) and \(V_1(\alpha)\), respectively. From Schechtman and Schecht-

man (2002), we have

\[
\tilde{V}_{1,0}(\alpha) = U_1(\alpha), \quad \tilde{V}_{1,0}(\alpha) = U_1(\alpha)
\]

For \(i = 1, \ldots, 2n_0, j = 1, \ldots, n_1\), it can be verified that

\[
\frac{\partial \tilde{V}_{1,0}(\alpha)}{\partial \alpha} = \tilde{V}_{1,0}(\alpha) + 1, \quad \frac{\partial \tilde{V}_{2,0}(\alpha)}{\partial \alpha} = -\tilde{V}_{2,0}(\alpha) - 1.
\]

First we list some lemmas.

**Lemma 1.** Under the conditions of Proposition 1, and if \(\sigma_1^2(\alpha) > 0\), \(\sigma_2^2(\alpha) > 0\), for each \(\alpha\) such that \(|\alpha - \alpha_0| \leq n_0^{-1/3}\).

Then as \(n_0, n_1 \to \infty\), we have

\[
\begin{align*}
\left[\begin{array}{c}
\left(\frac{n_0}{n_0 - 1}\right)^{1/2} U_1(\alpha_0) \\
\left(\frac{n_0}{n_0 - 1}\right)^{1/2} U_2(\alpha_0)
\end{array}\right] \\
\left(\frac{n_0}{n_0 - 1}\right)^{1/2} U_1(\alpha_0) + \left(\frac{n_0}{n_0 - 1}\right)^{1/2} U_2(\alpha_0) - U_1(\alpha_0)
\end{align*}
\]

where \(S_1^2 = 2(\sigma_1^2(\alpha_0) + \sigma_2^2(\alpha_0))\), \(S_2^2 = 2(\sigma_1^2(\alpha_0) + \sigma_2^2(\alpha_0))\). \(S_2^2 = Eh_2^2(X_{n_0+1}; \alpha_0, \beta_0)\).

**Proof.** The first one is from Arvesen (1969), while the second one comes from the central limit theorem. Now, we prove that \(U_1^2(\alpha_0) = S_2^2 + \sigma_2^2(\alpha_0)\). For \(i = 1, \ldots, n_0\), it can be checked that

\[
\begin{align*}
\frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1^2(\alpha_0) \\
= \frac{1}{n_0} \sum_{i=1}^{n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0) \left[2n_0 - 1 \right] U_1(\alpha_0) \\
= \frac{1}{n_0} \sum_{i=1}^{n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0)
\end{align*}
\]

when \(i = n_0 + 1, \ldots, 2n_0\),

\[
\begin{align*}
\frac{1}{n_0} \sum_{i=n_0+1}^{2n_0} \tilde{V}_1^2(\alpha_0) \\
= \frac{1}{n_0} \sum_{i=n_0+1}^{2n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0) \\
= \frac{1}{n_0} \sum_{i=n_0+1}^{2n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0)
\end{align*}
\]

Thus,

\[
U_1^2(\alpha) = \frac{1}{2} \left(\frac{2n_0 - 1}{n_0} - 1\right) \left[\frac{1}{n_0} \sum_{i=1}^{n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0) \right] \\
+ \frac{1}{n_0} \sum_{i=n_0+1}^{2n_0} \left[2n_0 - 1 \right] \left[2n_0 - 1 \right] U_1(\alpha_0)
\]

**Case-control studies with gene-environment independence on controls** 297
Proof. We only prove the cases of $2^j$, where $2^j \leq n_0^{-1/3}$ (14)

$$\sum_{i=1}^{n_0} h_1(X_{i,1}, X_{i,2}; \alpha^*, \beta_0)$$

Note that $2 \n_0 \rightarrow \infty$.

Hence, we complete the proof.

**Lemma 2.** Under conditions of Proposition 1, if $\sigma^2_{1,10}(\alpha)$, $\sigma^2_{1,01}(\alpha) > 0$, for each $\alpha \in \Theta(\alpha_0)$, then as $n \rightarrow \infty$, we have

$$P \left( \min_{1 \leq i \leq n_0} \tilde{V}_1(\alpha) < 0 < \max_{1 \leq i \leq 2n_0} \tilde{V}_1(\alpha) \right) \rightarrow 1,$$

$$P \left( \min_{1 \leq i \leq n_1} \tilde{V}_2(\alpha) < 0 < \max_{1 \leq i \leq 2n_1} \tilde{V}_2(\alpha) \right) \rightarrow 1.$$

Proof. Note that

$$\tilde{V}_1(\alpha) = \frac{2n_0 - 1}{n_0} \{ V_{1,0}(\alpha) \} I(1 \leq i \leq n_0) + V_{1,0}(\alpha) I(n_0 + 1 \leq i \leq 2n_0) - \frac{n_0}{n_0 - 1} U_1(\alpha),$$

$$\tilde{V}_2(\alpha) = h_2(X_{n_0 + j}; \alpha_0, \beta_0) = \exp\{- \alpha_0 - m(X_{n_0 + j}, \beta)\} - 1.$$

Following similar arguments to those of Jing et al. (2009), we can reach the conclusion.

**Lemma 3.** Under the conditions of Proposition 1, we have

$$\left\{ \begin{array}{l}
U_1(\alpha) = O_p(n_0^{-1/3}) \\
U_2(\alpha) = O_p(n_0^{-1/3}) \\
\hat{U}_2(\alpha) = S^2 + o_p(1) \\
\hat{U}^2_2(\alpha) = S^2 + o_p(1),
\end{array} \right. \leq \frac{1}{n_0} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} G(X_{i,1}, X_{j,2})$$

$$E[G(X_{1,1}, X_{1,2})] < \infty,$$

this together with Lemma 1 leads to $U_1(\alpha) = O_p(n_0^{-1/3}).$ Next, let’s look at $\hat{U}^2_2(\alpha)$. Notice that

$$\begin{align*}
|\hat{U}^2_2(\alpha) - \hat{U}^2_2(\alpha_0)|
&= \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) [\tilde{V}_1(\alpha^*) + 1] (\alpha - \alpha_0) \right| \\
&\leq \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) [\tilde{V}_1(\alpha^*) + 1] (\alpha - \alpha_0) \\
&+ \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) [\tilde{V}_1(\alpha^*) + 1] (\alpha - \alpha_0) \\
&:= A_1 + A_2.
\end{align*}$$

Now,

$$A_1 = \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) \right| O_p(n_0^{-1/3}) + O_p(1)O_p(n_0^{-1/3}),$$

while

$$\begin{align*}
\left| \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) \right|
&= \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) \right\} \right|^2 \\
&= \left| \frac{1}{n_0} \sum_{i=1}^{n_0} \left\{ \frac{1}{n_0} \sum_{i=1}^{n_0} \tilde{V}_1(\alpha^*) \right\} \right|^2 \times \left| \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} h_1(X_{i,1}, X_{j,2}; \alpha^*, \beta_0) \right|^2 \\
&+ \left( \frac{2n_0 - 1}{2n_0} \right) \left[ \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} h_1(X_{i,1}, X_{j,2}; \alpha^*, \beta_0) \right]^2 \\
&\leq \frac{1}{n_0} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} h_1(X_{i,1}, X_{j,2}; \alpha^*, \beta_0) \\
&+ \frac{9}{n_0} \sum_{i=1}^{n_0} \sum_{j=1}^{n_0} h_1(X_{i,1}, X_{j,2}; \alpha^*, \beta_0).$$
\]
Lemma 4. We have
\[
\frac{4}{n_0^{m+1}} \sum_{i_1=1}^{n_0} \sum_{i_2=1}^{n_0} h_i^2(X_{i_1,1}, X_{i_2,2}; \alpha^*, \beta_0) + \frac{9}{n_0^{m+1}} \sum_{i_1=1}^{n_0} \sum_{i_2=1}^{n_0} h_i^2(X_{i_1,1}, X_{i_2,2}; \alpha^*, \beta_0) \leq C \frac{1}{n_0^{m+1}} \sum_{i_1=1}^{n_0} \sum_{i_2=1}^{n_0} G(X_{i_1,1}, X_{i_2,2}) \quad \text{P.s.}
\]
and
\[
CE(G(X_{i_1,1}, X_{i_2,2})) < \infty.
\]

By the Borel-Cantelli lemma, with probability one, we have
\[
\max_{1 \leq i \leq 2n_0} |\tilde{V}_{1i}(\alpha, \beta_0)| = o_p(n_0^{-1/3})
\]
\[
\frac{1}{2n_0} \sum_{i=1}^{2n_0} |\tilde{V}_{1i}(\alpha, \beta_0)|^3 = o_p(n_0^{-1/3})
\]

Combining Lemmas 3 and 4, we have
\[
\lambda_1 = \lambda_1(\alpha) = \lambda_2(\alpha)
\]
where
\[
\lambda_1 = o_p(n_0^{-1/3}), \quad \lambda_2 = o_p(n_0^{-1/3})
\]

Lemma 5. For \( \lambda_1 = \lambda_1(\alpha) \) and \( \lambda_2 = \lambda_2(\alpha) \) satisfying (9) and (10), under conditions of Proposition 1, we have

Proof. From (9), we have
\[
\lambda_1 \tilde{V}_{1i}(\alpha, \beta_0) = o_p(1), \quad \lambda_2 \tilde{V}_{2i}(\alpha, \beta_0) = o_p(1)
\]

Since (9) is equal to
\[
U_1(\alpha) = \lambda_1 U_1^2(\alpha) + \frac{1}{2n_0} \sum_{i=1}^{2n_0} \tilde{V}_{1i}(\alpha, \beta_0) = 0,
\]
we have
\[
\lambda_1 = [U_1^2(\alpha)]^{-1} U_1(\alpha) + o_p(n_0^{-1/3}).
\]

Similarly,
\[
\lambda_2 = [U_2^2(\alpha)]^{-1} U_2(\alpha) + o_p(n_0^{-1/3}).
\]

Denote
\[
l_1(\alpha, \beta_0) = \frac{1}{2n_0} \sum_{i=1}^{2n_0} \log \{1 + \lambda_1 \tilde{V}_{1i}(\alpha, \beta_0)\}, \quad l_2(\alpha, \beta_0) = \frac{1}{2n_0} \sum_{j=1}^{2n_0} \log \{1 + \lambda_2 \tilde{V}_{2j}(\alpha, \beta_0)\},
\]
and set \( \alpha = \alpha_0 + n_0^{-1/3} \). Applying Taylor expansion to \( l_1(\alpha, \beta_0) \), we have
\[
l_1(\alpha, \beta_0) = \lambda_1 \sum_{i=1}^{2n_0} \tilde{V}_{1i}(\alpha, \beta_0) - \frac{1}{2} \lambda_1^2 \sum_{i=1}^{2n_0} \tilde{V}_{1i}^2(\alpha, \beta_0) = \lambda_1 \sum_{i=1}^{2n_0} \tilde{V}_{1i}(\alpha, \beta_0) + o_p(n_0^{1/3}).
\]
that with probability tending to one. On the other hand, By Taylor expansion, we have

\[ \alpha = 2n_0 \{ U_2^2(\alpha) \}^{-1} U_1^2(\alpha) + o_p(1^{1/3}) \]
\[ = 2n_0 \{ U_1^2(\alpha) + o_p(1) \}^{-1} \times \{ U_1(\alpha_0) + U_1(\alpha^*) + 1(\alpha - \alpha_0)^2 \} + o_p(1^{1/3}). \]

By Lemma 3 and the fact that \( U_1(\alpha) \to 0 \) uniformly in \( \{ \alpha : |\alpha - \alpha_0| < n_0^{-1/3} \} \), we obtain

\[ l_1(\alpha, \beta_0) \geq C n_0^{1/3} \]

with probability tending to one. On the other hand,

\[ l_1(\alpha_0, \beta_0) = 2n_0 \{ U_2^2(\alpha_0) \}^{-1} U_1^2(\alpha_0) + o_p(1) = O_p(1). \]

Thus, as \( n_0 \to \infty \), with probability tending to one, it holds that \( l_1(\alpha_0 + n_0^{-1/3}, \beta_0) > l_1(\alpha_0, \beta_0) \). Similarly, we can show that \( l_1(\alpha_0 - n_0^{-1/3}, \beta_0) > l_1(\alpha_0, \beta_0) \). Also, we can prove \( l_2(\alpha_0 + n_0^{-1/3}, \beta_0) > l_2(\alpha_0, \beta_0) \) in the same manner.

Since \( l_1(\alpha, \beta_0) \) and \( l_2(\alpha, \beta_0) \) are continuous in \( [\alpha_0 - n_0^{-1/3}, \alpha + n_0^{-1/3}] \), \( l(\alpha, \beta_0) \) attains its minimum in the interior of \( [\alpha_0 - n_0^{-1/3}, \alpha + n_0^{-1/3}] \), denoted by \( \hat{\alpha} \).

**Proof of Theorem 1.** Put \( \hat{\lambda}_i = \lambda_i(\hat{\alpha}) \), \( i = 1, 2 \). Then

\[ Q_{j=0,1}(\hat{\alpha}, \hat{\lambda}_1, \hat{\lambda}_2) = 0, \quad j = 1, 2, 3. \]

By Taylor expansion, we have

\[ (18) \]
\[ 0 = Q_{j=0,1}(\alpha_0, 0, 0) \]
\[ + \frac{\partial}{\partial \lambda_1} Q_{j=0,1}(\alpha^*, \lambda_1^*, \lambda_2^*) \hat{\lambda}_1 + \frac{\partial}{\partial \lambda_2} Q_{j=0,1}(\alpha^*, \lambda_1^*, \lambda_2^*) \hat{\lambda}_2 \]
\[ + \frac{\partial}{\partial \alpha} Q_{j=0,1}(\alpha^*, \lambda_1^*, \lambda_2^*) (\hat{\alpha} - \alpha_0), \]

where \( \alpha^* \) lies between \( \hat{\alpha} \) and \( \alpha_0 \), \( \lambda_2^* \) lies between \( \hat{\lambda}_1 \) and \( \lambda_1 \). Using Lemma 5, it can be checked that

\[ \lim_{n_0 \to \infty} \frac{\partial}{\partial \lambda_1} Q_{1=0,1}(\alpha, \lambda_1, \lambda_2) \]
\[ = \lim_{n_0 \to \infty} \frac{1}{2n_0} \sum_{i=1}^{2n_0} -\hat{V}_2^2(\alpha, \beta_0) \]
\[ = \lim_{n_0 \to \infty} -\frac{1}{2n_0} \sum_{i=1}^{2n_0} \hat{V}_2^2(\alpha, \beta_0) \]
\[ = -S_2^2, \]
\[ \lim_{n_0 \to \infty} \frac{\partial}{\partial \lambda_2} Q_{1=0,1}(\alpha, \lambda_1, \lambda_2) = 0, \]
\[ \lim_{n_0 \to \infty} \frac{\partial}{\partial \alpha} Q_{1=0,1}(\alpha, \lambda_1, \lambda_2) \]
\[ = \lim_{n_0 \to \infty} \frac{1}{2n_0} \sum_{i=1}^{2n_0} \hat{V}_1(\alpha, \beta_0) + 1 \]
\[ = 1, \]

300 B.-Y. Jing et al.
\[ \hat{\alpha} - \alpha_0 = \frac{-1}{|S|} \left( S_2^T Q_{1n_0n_1} (\alpha_0, 0, 0) - \rho S_1^T Q_{2n_1n_0} (\alpha_0, 0, 0) \right) + o_p(1). \]

The above gives that \( \hat{\lambda}_1 = \rho \hat{\lambda}_2 + o_p(1) \). Furthermore,

\[ \sqrt{n_1} \left( \frac{Q_{1n_0n_1} (\alpha_0, 0, 0)}{Q_{2n_1n_0} (\alpha_0, 0, 0)} \right) \xrightarrow{d} N \left( \left( 0 \right), \left( \begin{array}{cc} \rho S_1^2 & 0 \\ 0 & S_2^2 \end{array} \right) \right), \]

which implies that

\[ \sqrt{n_1} \hat{\lambda}_2 \xrightarrow{d} N(0, |S|^{-1}). \]

\[ l(\beta_0) = 2 \left( \sum_{i=1}^{2n_0} \log(1 + \lambda_1 \hat{V}_1(i, \beta_0)) \right) 
+ \sum_{i=1}^{n_1} \log(1 + \lambda_2 \hat{V}_2(i, \beta_0)) \right) 
= 2n_1 \lambda_1 \hat{V}_1(i, \beta_0) - 2n_0 \lambda_1 \hat{V}_1(i, \beta_0) + o_p(1) 
+ 2n_1 \lambda_2 \hat{V}_2(i, \beta_0) - n_1 \lambda_2 \hat{V}_2(i, \beta_0) + o_p(1) 
= 2n_1 \lambda_2 \hat{U}_1^2(i, \beta_0) + n_1 \lambda_2 \hat{U}_2^2(i, \beta_0) + o_p(1) 
= \left( (\rho^2 + o(1)) \hat{U}_1^2(i, \beta_0) \right) + o_p(1) 
= \left( (\rho^2 + o(1)) S_1^2 + S_2^2 \right) \left( \sqrt{n_1} \lambda_2 \right)^2 + o_p(1) 
\]

\[ \hat{\beta}_0 \xrightarrow{d} \lambda_1. \]

\section*{Acknowledgements}

The authors would like to thank the Editor, an Associate Editor and referees for their suggestions and criticisms which helped to improve this paper. Jing’s research is partially supported by HK RGC grants HKUST6011/07P and HKUST6015/08P, HKUST6019/10P, the National Science Foundation of China (71071155), the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (Grant No. 10XNL07), and in part by NSFC (Grant No. 71071155). Li’s research is partially supported by the Fundamental Research Funds for the Central Universities (Grant No. lzujbky-2011-121). Zhou’s research is partially supported by a grant R-155-000-116-112 at the National University of Singapore.

\section*{Received 3 August 2011}

\section*{References}


Case-control studies with gene-environment independence on controls 301

Bing-Yi Jing
Department of Mathematics
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon
Hong Kong
E-mail address: majing@ust.hk

Zhouping Li
School of Mathematics and Statistics
Lanzhou University
Lanzhou 730000
China
E-mail address: lizhp@lzu.edu.cn

Jing Qin
Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases, NIH
Bethesda, MD 20892
USA
E-mail address: jinqin@niaid.nih.gov

Wang Zhou
Department of Statistics and Applied Probability
National University of Singapore
Singapore 117546
Singapore
E-mail address: stazw@nus.edu.sg