Two-sample nonparametric likelihood inference based on incomplete data with an application to a pneumonia study

Albert Vexler, Jihnhee Yu*, **Lili Tian,** and **Shuling Liu**

Department of Biostatistics, University at Buffalo, State University of New York, Buffalo, NY, U.S.A.

Summary

The clinical pulmonary infection score (CPIS) and bronchoalveolar lavage (BAL) are important diagnostic variables of pneumonia for forcefully ventilated patients who are susceptible to nosocomial infection. Because of its invasive nature, BAL is performed for patients only if the CPIS is greater than a certain threshold value. Thus, CPIS and BAL are closely related, yet BAL values are substantially missing. In a randomized clinical trial, the control and oral treatment groups were compared based on outcomes from these procedures. Because of the relevance of both outcomes with respect to evaluating the efficacy of treatments, we propose and examine a nonparametric test based on these outcomes, which employs the empirical likelihood methodology. While efficient parametric methods are available when data are observed incompletely, performing appropriate goodness-of-fit tests in order to justify the parametric assumptions is difficult. Our motivation is to provide an approach based on no particular distributional assumption, which enables us to use all observed bivariate data, whether completed or not in an approximate likelihood manner. A broad Monte-Carlo study evaluates the asymptotic properties and efficiency of the proposed method based on various sample sizes and underlying distributions. The proposed technique is applied to a data-set from a pneumonia study demonstrating its practical worth.

Key words: Incomplete data, Nonparametric inference, Empirical likelihood, Two group comparison, Pneumonia study

* Corresponding author: e-mail: jinheeyu@buffalo.edu, Phone: +001 716 829 6029, Fax: +001 716 829 2200

1 Introduction

A recent clinical trial regarding ventilator associated pneumonia (VAP) at University at Buffalo, The State University of New York, investigated the relationship between the maintenance of the oral health and pneumonia among the patients staying in the intensive care unit (ICU). Many patients in the ICU are assisted to forcefully breathe through a mechanical ventilator because of their lack of the ability of breathing on their own. Because the mechanical ventilator is inserted into patients' airway directly, it causes VAP for approximately 10 - 20% of patients with more than 48 hours of its use [Porzecanski and Bowton, 2006]. An initial diagnosis of VAP can be carried out by a composite clinical score, called the clinical pulmonary infection score (CPIS) composed of fever, blood leukocyte counts, tracheal secretions, oxygenation index, chest x-ray, and tracheobronchial secretions. When the CPIS is greater than a threshold value, the diagnosis of VAP is further confirmed using the quantitative techniques such as bronchoalveolar lavage (BAL) fluid culture. BAL is an invasive procedure using a bronchoscope to collect the specimen directly from lung, and can cause some side effects such as systemic and sepsis-like effects [Sanchez et al., 1998]. Thus, it is difficult to carry out routinely for intubated patients. The CPIS is known to have a sensitivity of 72 - 77% [Koenig and Truwit, 2006] and a specificity of 42 - 85% while BAL has a sensitivity of 73% and a specificity of 82% [Dupont et al., 2004]. Pugin et al. [1991] first developed the CPIS procedure and showed that the CPIS and BAL values are highly correlated (Pearson's correlation coefficient of 0.84). Thus, using both the CPIS and BAL values is relevant to evaluating the treatment effect on pneumonia. However, values of BAL are substantially missing because performing BAL depends on the values of the CPIS.

The problem described above can be handled in the framework of censored or incomplete data [e.g., Qin, 2000, Zhao and Tian, 2001, Vexler et al., 2008]. In this article, we develop a method based on the empirical likelihood approach [e.g., Owen, 1988, 1990, 1991, 2001] to carry out two-group comparison utilizing these relevant outcomes. Note that, when data are observed incompletely, common powerful statistical tools are based on the parametric likelihood methodology. However with incomplete data, goodness-of-fit tests are difficult to apply for the purpose of confirming parametrical assumptions, which motivates us to

propose a methodology based on the nonparametric approach. Thomas and Grunkemeier [1975] initially proposed a method to maximize the likelihood function of the Kaplan and Meier estimator subject to some constraints for the confidence interval estimation of survival data. Owen [1988] introduced the empirical likelihood approach to construct confidence intervals for various parameters. The empirical likelihood (henceforth, referred to as EL) method does not require distributional assumptions as compared to the parametric likelihood based methods. The EL approach can achieve good coverage for the confidence intervals which are less affected by the underlying distributions [e.g., DiCiccio et al., 1991, Chen et al., 2003, Glenn et al., 2007]. Owen [1990] showed the better coverage rate of the confidence intervals using the empirical likelihood approach as compared with the t-statistic given skewed data. An advantage of the EL method is its capability to incorporate information for parameters into the likelihood functions [e.g., Chen and Qin, 1993]. In this setting, likelihood ratio tests can be constructed based on empirical likelihood functions estimated subject to restrictions on the parameters. Qin [2000] showed that the likelihood function can be approximated based on incomplete bivariate data by combining parametric and empirical likelihoods together. Some considerations of the EL method in the context of two-sample problems may also be found in the literature [Jing, 1995, Liu et al., 2008]. We propose a nonparametric likelihood ratio test for two-group comparison based on bivariate data that are observed incompletely. The proposed method in this article can be used as an alternative to the case when the distribution assumptions are questionable. To preserve the advantage of likelihood techniques, we pay full attention to the EL methodology in the context of the pneumonia study, where distribution functions of the diagnostic variables are assumed to be unknown. The rest of paper has the following structure. In Section 2, we construct the likelihood for incomplete bivariate data. For both the cases, complete and incomplete observations, we establish appropriate constraints that can be incorporated into the EL ratio test. In Section 3, we apply the proposed method to analyze data from an actual pneumonia study. In Section 4, we further investigate the performance of the proposed method through a broad Monte Carlo study. Section 5 is devoted to concluding remarks.

2 Main results

A semiparametric approach to combine parametric and empirical likelihoods developed by Qin [2000] provides the basic structure of the likelihood function in this article. We note that two-group comparison was not of interest in Qin's work. In this article, we take a step further with respect to developing a nonparametric method based on the EL approach for the purpose of two-group comparison. Let (X_i, Y_i) denote the bivariate random variable from group i ($i = 1, 2$). Let $(x_{11}, y_{11}), ..., (x_{1n_1}, y_{1n_1})$ and $(x_{21}, y_{21}), ..., (x_{2n_2}, y_{2n_2})$ denote independent bivariate observations from group 1 and group 2, respectively. It is assumed that, in each group, the observations have the same distribution. The value of y_{ij} is observed only when $x_{ij} \geq c$, where c is a certain threshold value. Following Qin [2000], we construct the likelihood function given by

$$\begin{aligned} L &= \prod_{i=1}^{2} \prod_{j=1, x_{ij}<c}^{n_i} dF_i(x_{ij}) \prod_{j=1, x_{ij}\geq c}^{n_i} dG_i(y_{ij}|x_{ij}, \theta) dF_i(x_{ij}) \\ &= \prod_{i=1}^{2} \prod_{j=1}^{n_i} dF_i(x_{ij}) \prod_{j=1, x_{ij}\geq c}^{n_i} dG_i(y_{ij}|x_{ij}, \theta), \end{aligned} \tag{1}$$

where $F(x)$ is the unknown distribution of X and $G(y|x)$ is the unknown conditional distribution of Y given X. Following the EL concept, we rewrite (1) as

$$L = \prod_{i=1}^{2} \prod_{j=1}^{n_i} p_{ij} \prod_{j=1, x_{ij}\geq c}^{n_i} q_{ij}, \tag{2}$$

where $dF(x_{ij})$ and $dG(y_{ij}|x_{ij})$ are replaced by the empirical probabilities p_{ij} and q_{ij}. Note that we do not assume specific underlying distributions.

The EL approach estimates p_{ij} and q_{ij} to maximize L subject to appropriate constraints regarding parameters considered under certain hypotheses. In this article, we focus on the restrictions under the null hypothesis H_0 that are

$$E[\Upsilon(X_1, \boldsymbol{\theta})] = E[\Upsilon(X_2, \boldsymbol{\theta})] \text{ and } E[\Psi(Y_1|X_1, \boldsymbol{\theta})] = E[\Psi(Y_2|X_2, \boldsymbol{\theta})], \tag{3}$$

where $\boldsymbol{\theta}$ is a vector of parameters defining the relationship between X and Y, Υ is a function of $(X_i, \boldsymbol{\theta})$ expressing the restriction summarized under H_0 and Ψ is a function of Y_i for the restriction of Y_i given

$(X_i, \boldsymbol{\theta})$ under H_0. Equations (3) have the empirical forms of

$$\sum_{j=1}^{n_1} \Upsilon(x_{1j}, \boldsymbol{\theta}) p_{1j} = \sum_{j=1}^{n_2} \Upsilon(x_{2j}, \boldsymbol{\theta}) p_{2j}, \tag{4}$$

and

$$\sum_{j=1, x_{1j} \geq c}^{n_1} \Psi(y_{1j}|x_{1j}, \boldsymbol{\theta}) q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} \Psi(y_{2j}|x_{2j}, \boldsymbol{\theta}) q_{2j}. \tag{5}$$

Suppose that $Y_i = \xi(X_i, \boldsymbol{\theta}) + \epsilon_i$, where ξ is a function defining the relationship between X_i and Y_i. Then, under H_0, constraint (5) can be expressed as

$$\sum_{j=1, x_{1j} \geq c}^{n_1} \epsilon_{1j} q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} \epsilon_{2j} q_{2j},$$

where $\epsilon_{ij} = y_{ij} - \xi(x_{ij}, \boldsymbol{\theta})$ indicates the residual for ij-th subject. Note that p_{ij} and q_{ij} satisfy restrictions (4) and (5) under H_0 as well as $\sum_j^{n_i} p_{ij} = \sum_{j, x_{ij} \geq c}^{n_i} q_{ij} = 1$. Obtaining p_{ij} and q_{ij} can be accomplished using the method of Lagrange multipliers. Once p_{ij} and q_{ij} are obtained, the empirical likelihood function (2) is used as a numerator of the likelihood ratio test. The denominator of the likelihood ratio test is based on evaluation of the empirical likelihood (2) without taking into account the constraints (4) and (5).

Without loss of generality, consider testing a hypothesis

$$E(Y_1) = E(Y_2).$$

Since Pugin et al. [1991] showed the strong linear relationship between the CPIS and BAL values (Pearson's correlation coefficient of 0.84), we assume a linear relationship between X_i and Y_i with an intercept α and slope β,

$$E(Y_i) = \alpha + \beta E(X_i). \tag{6}$$

Generally speaking, relationship (6) may be interpreted as a model restriction on the statement of problem; however, following standards within the empirical likelihood literature (e.g., Owen, 1991), we consider the proposed tests as a nonparametric decision rule. By (6), equation (4) can be expressed as

$$\beta_1 \sum_{j=1}^{n_1} x_{1j} p_{1j} + \alpha_1 = \beta_2 \sum_{j=1}^{n_2} x_{2j} p_{2j} + \alpha_2, \tag{7}$$

where α_i and β_i are the true parameters related to group i. Under H_0, we assume that $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$, thus equation (7) is equivalent to

$$\sum_{j=1}^{n_1} x_{1j} p_{1j} = \sum_{j=1}^{n_2} x_{2j} p_{2j}. \tag{8}$$

Equation (5) has the form of

$$\sum_{j=1, x_{1j} \geq c}^{n_1} (y_{1j} - \alpha_1 - \beta_1 x_{1j})\, q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} (y_{2j} - \alpha_2 - \beta_2 x_{2j})\, q_{2j}. \tag{9}$$

Since $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$ under H_0, (9) can be restated as follows using the estimations of the parameters,

$$\sum_{j=1, x_{1j} \geq c}^{n_1} \left(y_{1j} - \hat{\alpha}_1 - \hat{\beta}_1 x_{1j}\right) q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} \left(y_{2j} - \hat{\alpha}_2 - \hat{\beta}_2 x_{2j}\right) q_{2j},$$

$$\sum_{j=1, x_{1j} \geq c}^{n_1} \left(y_{1j} - \hat{\alpha}_2 - \hat{\beta}_2 x_{1j}\right) q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} \left(y_{2j} - \hat{\alpha}_1 - \hat{\beta}_1 x_{2j}\right) q_{2j},$$

and

$$\sum_{j=1, x_{1j} \geq c}^{n_1} \left(y_{1j} - \hat{\alpha} - \hat{\beta} x_{1j}\right) q_{1j} = \sum_{j=1, x_{2j} \geq c}^{n_2} \left(y_{2j} - \hat{\alpha} - \hat{\beta} x_{2j}\right) q_{2j}, \tag{10}$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are estimators based on the data from group i and $\hat{\alpha}$ and $\hat{\beta}$ based on the pooled data. To justify the method of using estimators of unknown parameters in the EL constructions, we refer the readers to the theoretical arguments of Qin and Lawless [1994]. In this article, for simplicity, we employ least square estimation of the parameters based on the observed data. Note that α_i and β_i can also be estimated through censored regression methods such as the Tobit approach [e.g., Tobin, 1958, Takeshi, 1984]. Our broad Monte-Carlo study showed that the third method (10) is consistently most powerful among the three methods.

Under H_0, $\prod_{j=1}^{n_i} p_{ij}$ is maximized subject to

$$\sum_{j=1}^{n_1} x_{1j} p_{1j} = \sum_{j=1}^{n_2} x_{2j} p_{2j}, \sum_{j=1}^{n_1} p_{1j} = 1, \sum_{j=1}^{n_2} p_{2j} = 1, 0 \leq p_{ij} \leq 1, \tag{11}$$

using (8). This is accomplished through the Lagrange multiplier method. We need to find the stationary points of

$$\sum_{j=1}^{n_1}\log p_{1j}+\sum_{j=1}^{n_2}\log p_{2j}+\lambda_1\left(1-\sum_{j=1}^{n_1}p_{1j}\right)$$
$$+\lambda_2\left(1-\sum_{j=1}^{n_2}p_{2j}\right)+\lambda_3\left(\sum_{j=1}^{n_1}x_{1j}p_{1j}-\sum_{j=1}^{n_2}x_{2j}p_{2j}\right),$$

where $\lambda_k, k=1,2,3$ are the Lagrange multipliers. Maximizing over p_{ij}, we have

$$p_{1j}=\frac{1}{\lambda_1-\lambda_3 x_{1j}}, j=1,2,...,n_1, \text{ and } p_{2j}=\frac{1}{\lambda_2+\lambda_3 x_{2j}}, j=1,2,...,n_2.$$

To satisfy the equations at (11), λ_k are given as the roots of

$$\lambda_1+\lambda_2=n_1+n_2,\ \sum_{j=1}^{n_1}\frac{n_2-\lambda_2-\lambda_3 x_{ij}}{n_1+n_2-\lambda_2-\lambda_3 x_{ij}}=0, \text{ and } \sum_{j=1}^{n_2}\frac{\lambda_2+\lambda_3 x_{2i}-n_2}{\lambda_2+\lambda_3 x_{2i}}=0.$$

Under the alternative hypothesis H_1, $\prod_{j=1}^{n_i}p_{ij}$ is maximized subject only to $\sum_{j=1}^{n_i}p_{ij}=1$, where $0\le p_{ij}\le 1$. Using (10), the value of $\prod_{j=1}^{n_i^*}q_{ij}$ is maximized subject to

$$\sum_{j=1,x_{1j}\ge c}^{n_1}\left(y_{1j}-\hat{\alpha}-\hat{\beta}x_{1j}\right)q_{1j}=\sum_{j=1,x_{2j}\ge c}^{n_2}\left(y_{2j}-\hat{\alpha}-\hat{\beta}x_{2j}\right)q_{2j},$$
$$\sum_{j=1,x_{1j}\ge c}^{n_1}q_{1j}=1,\ \sum_{j=1,x_{2j}\ge c}^{n_2}q_{2j}=1, 0\le q_{ij}\le 1, \tag{12}$$

where $n_i^*, i=1,2$ are the numbers of x_{ij} such that $x_{ij}\ge c$ for group i. Maximizing over q_{ij} subject to (12) in a similar way to obtaining p_{ij}, we have

$$q_{1j}=\frac{1}{\lambda_1^\epsilon-\lambda_3^\epsilon\epsilon_{1j}}, j=1,2,...,n_1^*, \text{ and } q_{2j}=\frac{1}{\lambda_2^\epsilon+\lambda_3^\epsilon\epsilon_{2j}}, j=1,2,...,n_2^*,$$

where λ_i^ϵ are the Lagrange multipliers. Under H_1, $\prod_{j=1,x_{ij}\ge c}^{n_i^*}q_{ij}$ is maximized subject only to $\sum_{j=1,x_{ij}\ge c}^{n_i^*}q_{ij}=1$ where $0\le q_{ij}\le 1$. Based on the likelihoods under both the null and alternative hypotheses, the likelihood ratio test, say R, is

$$R=\frac{\prod_{i=1}^{2}\prod_{j=1}^{n_i}p_{ij}\prod_{j=1}^{n_i^*}q_{ij}}{\prod_{i=1}^{2}\left(\frac{1}{n_i}\right)^{n_i}\left(\frac{1}{n_i^*}\right)^{n_i^*}}. \tag{13}$$

The test statistic, R, can be understood as a combination of test statistics for X and $Y|X$. Under the assumption that H_1 is true, the combined test gives rise to an increased power by utilizing information from both X and Y.

We can show that the Lagrange multipliers for x_{ij} can be asymptotically expressed as

$$\lambda_3 = \frac{\sum_{j=1}^{n_2} x_{2j}/n_2 - \sum_{j=1}^{n_1} x_{1j}/n_1}{\sum_{j=1}^{n_1}(x_{1j}-\mu_{X_1})^2/n_1^2 + \sum_{j=1}^{n_2}(x_{2j}-\mu_{X_2})^2/n_2^2} + o(\max[n_1^{-1/2}, n_2^{-1/2}]), \tag{14}$$

$\lambda_1 = n_1 + \lambda_3 \sum_{j=1}^{n_1} x_{1j}/n_1$ and $\lambda_2 = n_2 - \lambda_3 \sum_{j=1}^{n_2} x_{2j}/n_2$ where μ_{X_i} are the mean of X_i. Similarly, the Lagrange multipliers for ϵ_{ij} can be asymptotically expressed as

$$\lambda_3^{\epsilon} = \frac{\sum_{j=1}^{n_2^*} \epsilon_{2j}/n_2^* - \sum_{j=1}^{n_1^*} \epsilon_{1j}/n_1^*}{\sum_{j=1}^{n_1^*} \epsilon_{1j}^2/n_1^{*^2} + \sum_{j=1}^{n_2^*} \epsilon_{2j}^2/n_2^{*^2}} + o(\max[n_1^{*^{-1/2}}, n_2^{*^{-1/2}}]), \tag{15}$$

$\lambda_1^{\epsilon} = n_1^* + \lambda_3^{\epsilon} \sum_{j=1}^{n_1^*} \epsilon_{1j}/n_1^*$ and $\lambda_2^{\epsilon} = n_2^* - \lambda_3^{\epsilon} \sum_{j=1}^{n_2^*} \epsilon_{2j}/n_2^*$. Based on the approximations (14) and (15), we can directly show that $-2\log R$ has an asymptotic χ_2^2 distribution under H_0 given independently distributed x_{ij}, ϵ_{ij}, where we also assume $E(|x_{ij}|^3) < \infty$, $E(|\epsilon_{ij}|^3) < \infty$. Note that showing this fact as well as the proofs of (14) and (15) are technical and, in general, repeat the proofs of Wilk type theorems related to the EL ratio tests [e.g., Owen, 1988, 1990, 1991, 2001]. Since the closed form solutions for λ_3 and λ_3^{ϵ} do not exist, the leading terms in (14) and (15) may be used for the initial values to find numeric solutions of them. The numerical solutions can be obtained using available optimization programs such as 'optim' in R (`http://www/r-project.org`). Note that the attention to the initial values of λ_i appreciably reduces the complexity and bias of the numerical calculations when the two-sample comparison is carried out. In Section 4, we demonstrate a broad Monte Carlo study to investigate the asymptotic property of the proposed method with finite sample sizes.

Remark 1. Note that equation (10) can be substituted by a system of following constraints.

$$\begin{aligned}
&\sum_{j=1, x_{1j}\geq c}^{n_1} (1, x_{1j}^1, ..., x_{1j}^k)^T \left(y_{1j} - \hat{\alpha} - \hat{\beta} x_{1j}\right) q_{1j} = 0, \\
&\sum_{j=1, x_{2j}\geq c}^{n_2} (1, x_{2j}^1, ..., x_{2j}^k)^T \left(y_{2j} - \hat{\alpha} - \hat{\beta} x_{2j}\right) q_{2j} = 0,
\end{aligned} \tag{16}$$

where $k = 0, 1, ...$ For example, if $k = 1$, we have 4 constraints that substitute one constraint (10). These kinds of constraints are common when the EL methodology is applied in the simple linear model regression case for one sample given fully observed data (e.g., Owen 1991). In our case, this approach will complicate the testing procedure since additional numerical equations need to be considered. If we increase the number of constraints (i.e., k is 2 or more), the power of the test will increase so that it may be appealing in some cases. However, it is also clear that, by increasing k, we also increase the Type I error and complexity of testing, which may result in a greater difference between the theoretically expected Type I error and actual Type I error. In Section 4, we briefly compare the test utilizing constraints (16) with the proposed test based on constraint (10). We show the performance of the proposed method based on constraint (10) surpasses the test based on constraints (16) for analyzing the data with the characteristics close to those of interest (the data from the pneumonia study).

Remark 2. The proposed method in this article can be extended to multivariate cases, e.g., the linear association between Y and X can be substituted by the general condition $E(Y|X) = [\beta_0, \beta_1, \beta_2, \ldots, \beta_p] \times [1, X, X^2, \ldots, X^p]^T$, where β_i are parameters corresponding to the underlying model. Thus, following Taylor's theorem, we could investigate tests approximately, assuming $E(Y|X) = g(X)$, where g is some unknown function. This issue needs substantial mathematical details and is subject to future study.

3 Application

Subjects for the pneumonia study were recruited from patients admitted to an ICU in a local hospital, who were mechanically ventilated within 48 hours of admission. Eligible patients were randomly assigned to a control arm or oral topical treatment (once or twice daily). The primary outcome variables were the dental plaque index score and pathogen colonization of the oral cavity. No dose responding difference was observed for these primary outcomes, thus combining two oral topical treatment groups were considered. As secondary outcomes, surrogate variables to diagnose pneumonia such as the CPIS and BAL values were investigated. The CPIS was monitored daily until patients were discharged. If the CPIS was greater than

5, BAL was performed. The mean frequency of BAL was 0.65 as 41.4% of patients did not have BAL. Majority of patients have either no BAL or single BAL (83.4%). The group comparison was carried out using BAL values and its corresponding CPIS. When no BAL was performed, the maximum CPIS until discharge was used. When multiple BALs were performed, an earliest BAL was used. The majority of the data (74.9%) was observed within a one week period of hospitalization, thus the analysis may shed some light for the treatment effect on the early onset of pneumonia. A total of 175 subjects were enrolled in the study (116 and 59 patients in treatment and control groups, respectively). BAL results were provided for 24 subjects in the control group and 44 subjects in the treatment group. The outcomes of BAL (range: $0 \sim 1.6 \times 10^8$) were log-transformed after adding 1. The normal quantile plots of the CPIS for the two treatment groups are shown in Figure 1. The plots show that the distributions of the CPIS are close to a normal distribution. The residual plots and normal quantile plots with residuals of the simple linear regression between the CPIS and BAL values are depicted in Figure 2. The residual plots show no particular trends. The residual plots also exhibit some study irregularities, that is, some BAL was performed even if the CPIS was below the threshold value. One reason of this is that BAL can be independently ordered at a physician's discretion. *When $X_{ij} < c$, let an indicator MD_{ij} denote the status of BAL in terms of the observation of its value where $MD_{ij} = 1$ if BAL is observed, or otherwise $MD_{ij} = 0$. Utilizing the proposed methodology, the additional observations are incorporated into the likelihood function in the form of*

$$L = \prod_{i=1}^{2} \prod_{j=1, x_{ij}<c, MD_{ij}=0}^{n_i} dF_i(x_{ij}) \prod_{j=1, x_{ij}\geq c}^{n_i} dG_i(y_{ij}|x_{ij},\theta) dF_i(x_{ij}) \prod_{j=1, x_{ij}<c, MD_{ij}=1}^{n_i} dG_i(y_{ij}|x_{ij},\theta) dF_i(x_{ij}), \quad (17)$$

and (17) leads to

$$L = \prod_{i=1}^{2} \prod_{j=1}^{m_i} dF_i(x_{ij}) \prod_{j=m_i+1}^{n_i} dG_i(y_{ij}|x_{ij},\theta) dF_i(x_{ij}), \quad (18)$$

where $m_i (< n_i)$ is the number of incomplete observations (i.e., y_{ij} is not observed) in group i. Note that, in Qin's (2000) one sample problem, the likelihood function was obtained by combining the likelihoods

of the complete observations, (y_{ij}, x_{ij}) *and incomplete observations,* $x_{ij}(j \neq i)$. *Similarly to Qin's (2000) likelihood function, in the likelihood function (18), the complete observations,* (y_{ij}, x_{ij}) *provide an additional contribution to the likelihood function in a form of conditional density* $f(y_{ij}|x_{ij})$. The normal quantile plots reveal that the distributions of the residuals are not normally distributed. For the CPIS, the initial Lagrange multipliers λ_2 and λ_3 for the EL were 113.011 and 0.594, respectively. The numerical solutions of λ_2 and λ_3 were 113.006 and 0.597, respectively. For the EL based on the residuals, the initial Lagrange multipliers λ_2^ϵ and λ_3^ϵ were 43.839 and 0.569, respectively. The numerical solutions of λ_2^ϵ and λ_3^ϵ were 44.038 and 0.578, respectively. These numeric solutions demonstrate that the approximation of the initial values by (14) and (15) and their subsequent estimations of the other Lagrange multipliers are very close to the actual solutions. The likelihood ratio statistics using $-2\log R$ was 0.4910 which gives the p-value of 0.7823 based on χ_2^2 showing the lack of an evidence that the control and oral topical treatment groups are different.

[Figure 1 is about here.]

[Figure 2 is about here.]

4 Simulation study

The performance of the proposed method is investigated through the following simulation study. We conducted 10000 simulations per each scenario. We examined the study power for the pneumonia study using the parameter settings based on the actual data. To achieve the semblance of the pneumonia study, the sample sizes of $n_1 = 59$ for group 1 and $n_2 = 116$ for group 2 were used. To investigate the performance under the null hypothesis, the values of X were generated from $Normal(5, 3.6)$ and the values of Y were generated based on model (6) with $\alpha = 10$, $\beta = 0.05$, and the variance of the error term$= 25$ for both groups. The values of Y were generated only when the values of X were greater than 5. All tests were conducted at the significance level of 0.05. The value of Y was observed 29.5 times and 58.0 times on average for each group. The Monte Carlo Type I error was 0.0561, slightly greater

than the target significance level. We also examined the study power for the conditions emulating the direct parameter estimations and the sample sizes from the pneumonia study. The values of X_1 and X_2 were generated from $Normal(4.98, 3.24)$ and $Normal(5.03, 3.61)$, respectively. The values of Y_1 were generated using $\alpha = 9.69$ and $\beta = .12$ and the error term from $Normal(0, 23.04)$ and Y_2 were generated using $\alpha = 11.41$,$\beta = -.06$ and the error term from $Normal(0, 29.16)$. The value of Y was observed 29.3 times and 58.7 times on average for each group. The Monte Carlo power was 0.0795 which indicates that the study was largely a negative study. The simulations based on the normally distributed data with various sample sizes are shown in Table 1. For each sample size combination, the first case used the same parameters setting from the pneumonia data, and the second case used the same parameters except that the smaller variances were employed to investigate the power. The observation rates of Y_1 and Y_2 were in accordance with $P\{X_i > 5\}$. The power does not increase regardless of the sample size increase with the first cases, that is, the group difference is minuscule compared with the the data variability (close to the situation under H_0). When the variances become small (the second cases), the study power increases accordingly when the sample size increases demonstrating a desirable property for the test statistic.

[Table 1 is about here.]

The Type I errors and powers were further examined in various cases through the Monte-Carlo study as displayed in Tables 2, 3 and 4. All tests were conducted at the significance level of 0.05. The values of Y were generated only if the values of X were greater than given threshold values described in the captions for each table. Thus, Y was observed proportionally to the probability of X to be greater than the threshold. The simulations were carried out using the underlying distribution of normal (Table 2), lognormal (Table 3) and exponential distributions (Table 4). For the normal distribution, the values of X were generated with the mean μ_i and variance σ_i^2 for group i. The values of Y were generated based on model (6) with respective parameters and the variance for the error term ($\sigma^2_{\mathrm{res}_i}$) of 1. For each design, the combinations of equal or unequal variances for X and the error term of Y were investigated. For the lognormal distribution, the values of X were generated from $Lognormal(0, \nu) + \zeta$ where the parameters ν and ζ were adjusted to

preserve the same population mean and variance as the corresponding normal distribution in Table 2. The error terms for Y were generated in a similar manner to generating X following the lognormal distribution. The means of the error terms are 0 and the variances are same as the corresponding error terms in Table 2. The data for the exponential distribution were generated from $Exp(\nu) + \zeta$ preserving the same means and variances as the corresponding normal distribution in Table 2 by adjusting the parameters ν and ζ. The means and variances of the error terms for Y are same as the lognormal distribution. The first case for each sample size combination is to demonstrate the probability of rejection under H_0. When the sample sizes increase, the Type I errors for the normal and exponential distributions are closer to the nominal significance level. It is observed that the Type I error in the cases with the lognormal distribution is not as accurate as the other two distributions even though it becomes closer to the nominal value when the same sizes increase. When the difference between the two groups increases (larger differences toward Design 4), the power of the test also monotonely increases. Note that more values of Y for group 2 are observed as the magnitude of the parameters used in group 2 increases. The power increases are relatively small for the lognormal distribution. While improvements of the power are evident with the larger sample sizes, it seems that the performance of the EL method with the lognormal distributions may not be as efficient as with the normal or exponential distribution. When the variances (σ_2^2, $\sigma_{res_2}^2$) of one group increase, it is shown that the proposed test generally becomes more conservative. The study power is smaller as the more variation in the data results in the less distinction between two groups. However, these unequal variance cases exhibit the similar trends to equal variance cases in terms of Type I errors and powers.

Remark 1. According to Remark 1 in Section 2, we compare the proposed test with the test that utilizes constraints (16) with $k = 1$. Our Monte-Carlo simulation showed, in the cases of $n_1 = n_2 = 25$ with $\sigma_{\mathrm{res}_1} = 1$ and $\sigma_{\mathrm{res}_2} = 1.5$ in Table 2, that the Monte-Carlo Type I error of the test based on (16) was 0.196 instead of the expected significance level 0.05. Note that the simulated Type I error of the proposed method is 0.0527. When the sample size per group increases to 50 and 100, the Monte-Carlo Type I error improved to be 0.142 and 0.075, respectively. The simulated Type I errors for our proposed method was 0.0511 and

0.0484, respectively. Corresponding to this design, when the power was investigated (with Design 4 in Table 2), the test based on (16) demonstrated the power of 0.4413, 0.5800 and 0.8245 for the per-group sample sizes of 25, 50 and 100, respectively. The simulated powers of the proposed method were 0.2097, 0.4293, 0.7746. Note that, to carry out the test utilizing constraints (16), we used a more complicated and computationally extensive program comparing with the program that we used to carry out our proposed method. The relatively poor performance of the test based on constraints (16) in terms of the simulated Type I error may be due to the additional equations that were necessarily to solve and given the limited sample size. However, we expect that, for a large sample size, the approach utilizing (16) can provide a very powerful test. This issue, which needs substantial mathematical details, is under investigation. For the test based on (16), the effect of different values of k needs to be investigated using asymptotic propositions.

Remark 2. In order to compare the proposed method with a parametric technique, we adapted the results presented by Lyles et al. (2001). To this end, we assume that Y_{ij} and X_{ij} have a joint normal distribution, satisfying the linear regression model $Y_{ij} = \alpha + \beta X_{ij} + \epsilon_{ij}$, where ϵ_{ij} are independent and identically-distributed normal random variables with mean 0 and variance $\sigma^2_{res_i}$, $i = 1, 2$, $j = 1, \ldots, n_i$. For group i, the likelihood function, L_i, based on a sample with the size n_i takes the form

$$L_i(\alpha, \beta, \sigma, \sigma_{res}, \mu) = \left(\frac{1}{2\pi\sigma\,\sigma_{res_i}}\right)^{\sum_{j=1}^{n_i} I(x_{ij} \geq c)} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{\sum_{j=1}^{n_i} I(x_{ij} < c)} \times$$
$$\exp\left\{ - \sum_{j=1, x_{ij} \geq c}^{n_i} \frac{(y_{ij} - \alpha - \beta x_{ij})^2}{2\sigma^2_{res_i}} - \sum_{j=1}^{n_i} \frac{(x_{ij} - \mu_i)^2}{2\sigma^2} \right\}.$$

Let $L(\alpha, \beta, \sigma, \sigma_{res}, \mu)$ denote the likelihood function based on both groups 1 and 2, respectively. The parametric *maximum* likelihood ratio test, Λ, takes the form

$$\Lambda = \frac{\max_{\alpha,\beta,\sigma,\sigma_{res},\mu} L(\alpha, \beta, \sigma, \sigma_{res}, \mu)}{\max_{\alpha,\beta,\sigma,\sigma_{res},\mu} L_1(\alpha, \beta, \sigma, \sigma_{res}, \mu) \max_{\alpha,\beta,\sigma,\sigma_{res},\mu} L_2(\alpha, \beta, \sigma, \sigma_{res}, \mu)}. \tag{19}$$

Since the proposed parametric method is based on the *maximum* likelihood ratio, it follows from Wilks' theorem that the asymptotic distribution of $-2\log\Lambda$ takes the form of χ^2 distribution with 5 degrees of freedom, which in turn enables us to compute the critical value of the test. Note that, in order to execute the proposed parametric method, we have to estimate more parameters than we need to evaluate

for the proposed nonparametric approach (e.g., σ_i and σ_{res_i} do not need to be evaluated in our proposed method). The requirement of additional parameters can affect the accuracy and power of the test. Our Monte-Carlo simulation showed, in the cases of $\sigma_{\mathrm{res}_1} = 1$ and $\sigma_{\mathrm{res}_2} = 1.5$ in Table 2, that the simulated Type I errors of the parametric test were 0.0861, 0.0637 and 0.0577 for the per-group sample sizes of 25, 50 and 100, respectively. Comparing this with the results in Table 2, our proposed method shows better accuracy (0.0527, 0.0511 and 0.0484, respectively). The simulated powers based on Design 4 were 0.2857, 0.4809 and 0.7934 for per-group sample sizes of 25, 50 and 100. The corresponding powers for the proposed method (Table 2) were 0.2097, 0.4293 and 0.7746, respectively, which are slightly lower than those for the parametric test. *This result suggests that, in general, the parametric approach is more powerful when the underlying distribution is correctly specified; however, its accuracy and power can be compromised in case that too many parameters are need to be estimated even with the correct specification of the distribution.* We also carried out the test based on log-normal distribution based on scenarios similar to Table 3 ($\sigma_{\mathrm{res}_1} = 1$ and $\sigma_{\mathrm{res}_2} = 1.5$). For the parametric test, the simulated Type I errors with the sample sizes of 25, 50 and 100 were 0.1285, 0.0818 and 0.0772, respectively, and the simulated powers using Design 4 were 0.1988, 0.2624 and 0.4596, respectively. The corresponding simulated Type I errors for our proposed method were 0.1080, 0.0796 and 0.0645 and the corresponding simulated powers were 0.3059, 0.3933 and 0.5673. For both the Type I errors and powers, the proposed method (Table 3) outperforms to the parametric approach. This demonstrates that our proposed method provides generally more robust testing, which is certainly to be expected, considering our proposed method does not require the underlying parametric assumptions.

[Table 2 is about here.]

[Table 3 is about here.]

[Table 4 is about here.]

5 Concluding remarks

We proposed and examined the nonparametric likelihood ratio test for two-group comparison using the EL method. Among discussions of two-group comparisons using the EL method [e.g., Jing, 1995, Liu et al., 2008], our proposed approach is unique with respect to being applicable to incomplete bivariate data. When two outcomes are relevant surrogate variables of treatment effects, and closely related, while one outcome is substantially missing, the proposed method provides a way to test the difference between two groups utilizing all available observations in the approximate likelihood manner. When the parameters for both variables are truly different under the alternative hypothesis, incorporating complete and incomplete data into the test provides a more power compared with testing based on an individual variable as discussed in Section 2. The complete nonparametric approach enables the test statistic (13) to be applied to the actual data with various underlying distributions. In general, the proposed method may be applied to various forms of the linear functions; however, further investigations of such cases based on finite sample sizes are needed. An application to the actual data demonstrates that the method is easily applicable and useful for a situation where it is difficult to obtain all relevant variables for evaluating the treatment effect.

Acknowledgements The authors are very grateful to the reviewers for their comments and suggestions that have greatly helped us improve the manuscript. We would like to acknowledge that, owing to the reviewers' suggestion, comparisons between our proposed method and other possible approaches were carried out more in depth. We greatly appreciate all insightful suggestions.

References

Chen, J., Chen, S. Y., and Rao, J. N. K. (2003). Empirical likelihood confidence intervals for the mean of a population containing many zero values. *The Canadian Journal of Statistics* **31,** 53–68.

Chen J., Qin J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80,** 107–116.

DiCiccio, T. J., Hall, P., and Romano, J. P. (1989). Empirical likelihood is Bartlett-correctable. *The Annals of Statistics* **19,** 1053–1061.

Dupont, H., Chalhoub, V., Plantefève, G., Vaumas, C., Kermarrec, N., Paugam-Burtz, C., and Mantz, J. (2004). Variation of infected cell count in bronchoalveolar lavage and timing of ventilator-associated pneumonia. *Journal of Intensive Care Medicine* **30,** 1557–1563.

Glenn, N. L., Zhao, Y. (2007). Weighted empirical likelihood estimates and their robustness properties. *Computational Statistics & Data analysis* **51,** 5130–5141.

Jing, B. Y. (1995). Two-sample empirical likelihood method. *Statistics & Probability Letters* **24,** 315–319.

Koenig, S. M. and Truwit, J. D. (2006). Ventilator-associated pneumonia: diagnosis, treatment, and prevention .*Clinical Microbiology Reviews,* **19,** 637–657.

Liu, Y., Zou C., Zhang R. (2008). Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters* **78,** 548–556.

Lyles, R. H., Williams, J. K., Chuachoowong, R.(2001) Correlating two viral load assays with known detection limits. *Biometrics* **57,** 1238-1244.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75,** 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18,** 90–120.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics* **19,** 1725–1747.

Owen, A. (2001). *Empirical Likelihood* New York: Chapman & Hall.

Porzecanski, I., and Bowton, D. L. (2006). Diagnosis and treatment of ventilator-associated pneumonia. *Chest* **130,** 597–604.

Pugin, J., Auckenthaler, R., Mili, N., Janssens, J. P., Lew, P. D., and Suter, P. M. (1991). Diagnosis of ventilator-associated pneumonia by bacteriologic analysis of bronchoscopic and nonbronchoscopic ”blind” broncoalveolar lavage fluid. *American Review Of Respiratory Disease* **143,** 1121-1129.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika* **87,** 484–490.

Qin, J., and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22,** 300–325.

Sanchez-Nieto JM, Torres A, Garcia-Cordoba F, El-Ebiary M, Carrillo A, Ruiz J, Nuez ML, Niederman M. (1998). Diagnosis and treatment of ventilator-associated pneumonia. *American Journal of Respiratory and Critical Care Medicine* **157,** 371–376.

Takeshi, A. (1984). Tobit Models: A Survey. *Journal of Econometrics* **24,** 3–61.

Thomas, D. R., Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association* **70,** 865–871.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26,** 24–36.

Vexler, A., Liu, A., Eliseeva, E., and Schisterman, E. F. (2008). Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics* **64,** 895-903

Zhao, H., and Tian, L. (2001). On estimating medical cost and incremental cost-effectiveness ratios with censored data. *Biometrics* **57,** 1002–1008.

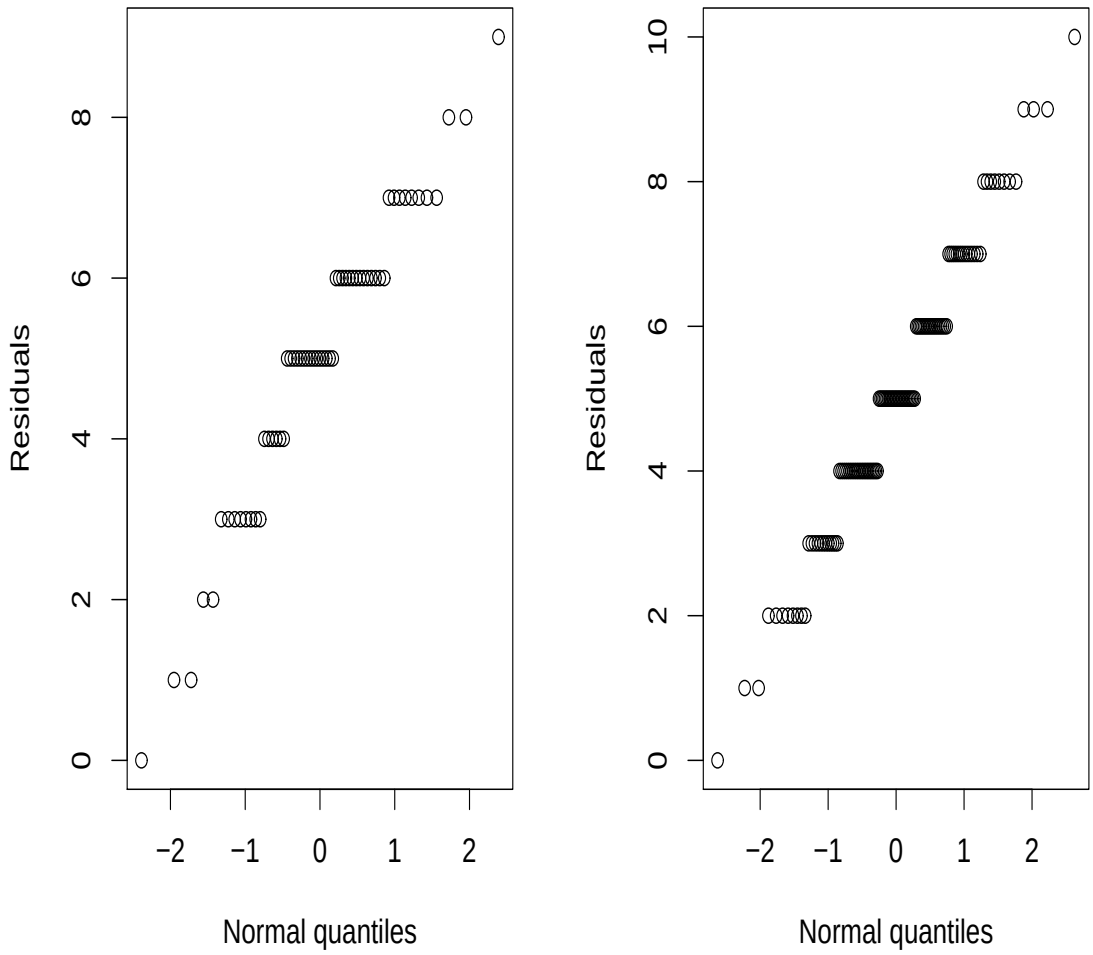

Fig. 1 Normal quantile plots of CPIS for control (left) and treatment groups (right).

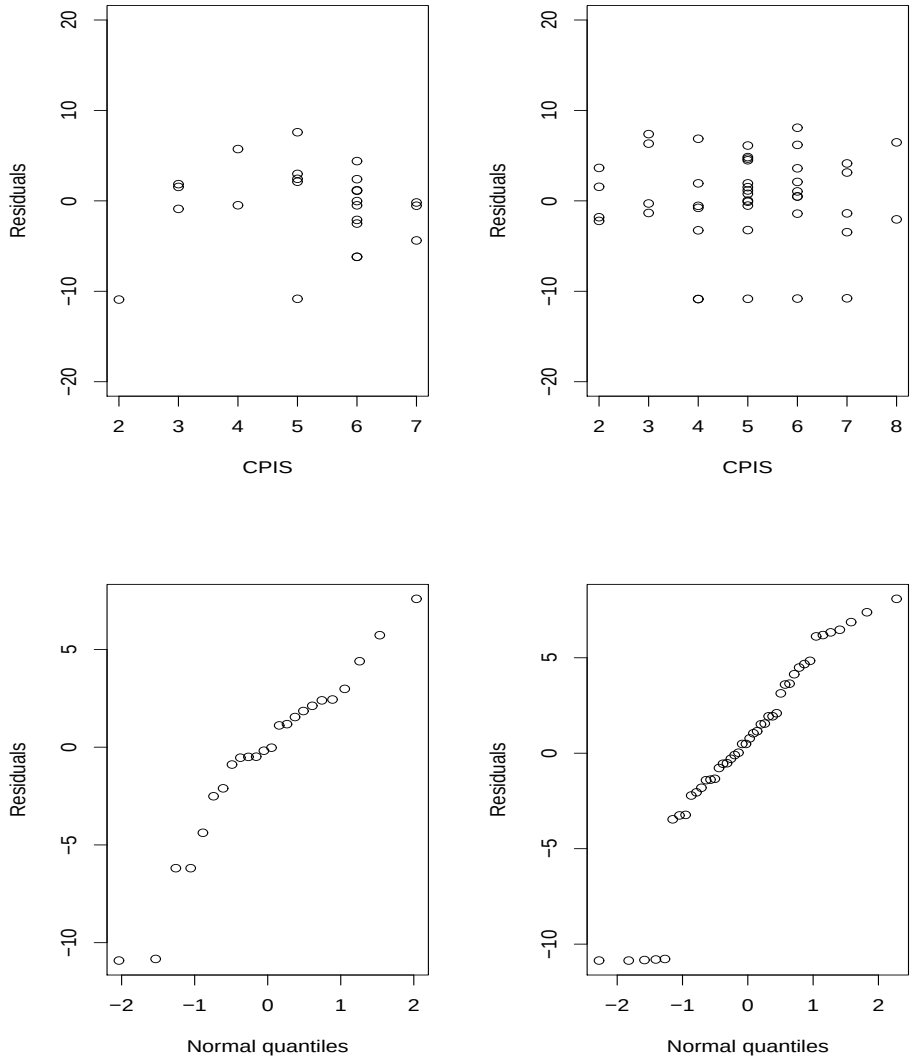

Fig. 2 Residual plots of BAL based on the model (6) using CPIS as a explanatory variable and quantile plots of the residuals. The figures in the left-side are for the control group and those in the right-side are for the treatment group.

Table 1 Simulations using the parameters from the pneumonia study to examine the power of the proposed method. The parameters of $\mu_1 = 4.98, \mu_2 = 5.03, \alpha_1 = 9.69, \alpha_2 = 11.41, \beta_1 = .12, \beta_2 = -.06$ were used. In the table, σ_1, and σ_2 indicate the standard deviations for X_1 and X_2. σ_{res_1} and σ_{res_2} indicate the standard deviations for the error term in the model of 6.

Parameters	Sample sizes n_1, n_2			
σ_1,σ_2,σ_{res_1},σ_{res_2}	25,25	50,50	50,100	100,100
$1.8, 1.9, 4.8, 5.4$	0.0791	0.0660	0.0732	0.0753
$0.15, 0.15, 1, 1$	0.4289	0.7427	0.8334	0.9390

Table 2 Simulations with the normal distribution to examine the power of the proposed method. σ_i are the standard deviation of X_i and σ_{res_i} are the standard deviation of the error term in the model (6). The threshold is 0 throughout the cases. For each design, the values of $(\mu_1, \mu_2, \alpha_1, \alpha_2, \beta_1, \beta_2)$ are (0,0,2,2,1,1) for Design 1, (0,0.1,2,2.1,1,1) for Design 2, (0,0.2,2,2.2,1,1.2) for Design 3 and (0,.3,2,2.3,1,1.3) for Design 4.

Sample size	Design	Standard deviations			
n_1,n_2		σ_1=1 σ_2=1		σ_1=1 σ_2=1.5	
		σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5	σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5
25,25	1	0.0612	0.0527	0.0491	0.0467
	2	0.0782	0.0689	0.0569	0.0596
	3	0.1497	0.1173	0.1125	0.0784
	4	0.2750	0.2097	0.2002	0.1432
50,50	1	0.0567	0.0511	0.0467	0.0478
	2	0.1105	0.0919	0.0828	0.0731
	3	0.2893	0.2171	0.2204	0.1592
	4	0.5682	0.4293	0.4450	0.3150
50,100	1	0.0488	0.0466	0.0447	0.0462
	2	0.1204	0.1042	0.1085	0.0948
	3	0.3620	0.3084	0.3000	0.2510
	4	0.6855	0.614	0.5825	0.5090
100,100	1	0.0501	0.0484	0.0464	0.0445
	2	0.1695	0.1306	0.1415	0.1118
	3	0.5306	0.4152	0.4433	0.3192
	4	0.8793	0.7746	0.8036	0.6481

Table 3 Simulations with the lognormal distribution to examine the power of the proposed method. σ_i are the standard deviation of X_i and σ_{res_i} are the standard deviation of the error term in the model (6). The threshold is the median of X_1. Refer to the text for how to generate the data. For each design, the values of $(\mu_1, \mu_2, \alpha_1, \alpha_2, \beta_1, \beta_2)$ are (0,0,2,2,1,1) for Design 1, (0,0.1,2,2.1,1,1) for Design 2, (0,0.2,2,2.2,1,1.2) for Design 3 and (0,.3,2,2.3,1,1.3) for Design 4.

Sample size	Design	Standard deviations			
n_1,n_2		σ_1=1 σ_2=1		σ_1=1 σ_2=1.5	
		σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5	σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5
25,25	1	0.0991	0.1080	0.1029	0.1074
	2	0.1207	0.1206	0.1172	0.1148
	3	0.2229	0.1993	0.1954	0.1789
	4	0.3688	0.3059	0.3262	0.2549
50,50	1	0.0741	0.0796	0.0755	0.0827
	2	0.1309	0.1206	0.1166	0.1067
	3	0.2843	0.2273	0.2547	0.1968
	4	0.4873	0.3933	0.4360	0.3401
50,100	1	0.0674	0.0712	0.0653	0.0666
	2	0.1429	0.1303	0.1305	0.1153
	3	0.2973	0.2544	0.2822	0.2288
	4	0.4169	0.3685	0.4492	0.3814
100,100	1	0.0605	0.0645	0.0568	0.0653
	2	0.1610	0.1348	0.1383	0.1108
	3	0.4267	0.3425	0.3905	0.2804
	4	0.6108	0.5673	0.5879	0.5088

Table 4 Simulations with the exponential distribution to examine the power of the proposed method. σ_i are the standard deviation of X_i and σ_{res_i} are the standard deviation of the error term in the model (6). The threshold is the mean of X_1. Refer to the text for how to generate the data. For each design, the values of $(\mu_1, \mu_2, \alpha_1, \alpha_2, \beta_1, \beta_2)$ are (0,0,2,2,1,1) for Design 1, (0,0.1,2,2.1,1,1) for Design 2, (0,0.2,2,2.2,1,1.2) for Design 3 and (0,.3,2,2.3,1,1.3) for Design 4.

Sample size	Design	Standard deviations			
n_1,n_2		σ_1=1 σ_2=1		σ_1=1 σ_2=1.5	
		σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5	σ_{res_1}=1 σ_{res_2}=1	σ_{res_1}=1 σ_{res_2}=1.5
25,25	1	0.0779	0.0783	0.0750	0.0726
	2	0.0976	0.0934	0.0835	0.0795
	3	0.1683	0.1438	0.1553	0.1239
	4	0.2955	0.2460	0.2554	0.2012
50,50	1	0.0585	0.0587	0.0558	0.0570
	2	0.0998	0.0838	0.0858	0.0725
	3	0.2656	0.2040	0.2210	0.1588
	4	0.5084	0.3998	0.4521	0.3233
50,100	1	0.0572	0.0498	0.0519	0.0515
	2	0.1238	0.1035	0.1126	0.0948
	3	0.3123	0.2617	0.3007	0.2438
	4	0.5693	0.5059	0.5706	0.4851
100,100	1	0.0528	0.0470	0.0449	0.0417
	2	0.1563	0.1273	0.1297	0.1031
	3	0.5019	0.3910	0.4311	0.3096
	4	0.8249	0.7188	0.7717	0.6158