Published in final edited form as:
Comput Stat Data Anal. 2008 December 15; 53(2): 546–553. doi:10.1016/j.csda.2008.09.021.

Confidence Intervals for A Common Mean with Missing Data with Applications in AIDS Study

Hua Liang[1], **Haiyan Su**[1], and **Guohua Zou**[1,2]

[1]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY 14642, USA, hliang@bst.rochester.edu

[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Summary

In practical data analysis, nonresponse phenomenon frequently occurs. In this paper, we propose an empirical likelihood based confidence interval for a common mean by combining the imputed data, assuming that data are missing completely at random. Simulation studies show that such confidence intervals perform well, even the missing proportion is high. Our method is applied to an analysis of a real data set from an AIDS clinic trial study.

Some key words

Empirical likelihood; jackknife; missing completely at random; ratio imputation; regression imputation

1 Introduction

In biomedical and epidemiologic researches, data are often missing because subjects fail to report at clinical centers or refuse to answer some questions, or technicians may lose data. Simply excluding the missing data, known as complete case analysis, may waste useful information, because other observed variables associated with the missing variables are also excluded. More seriously, this simple exclusion may result in an inefficient estimation (see, for example, Liang et al., 2004; Wu, 2004) and may even lead to a false conclusion although the implementation of the complete case method is much simpler and it is a default method in most statistical software. In the literature on missing data, common approaches have been described: maximum likelihood (Ibrahim, Chen, and Lipsitz, 1999; Ibrahim, Lipsitz, and Horton, 2001), weighting adjustment (c.f., Cochran, 1977), single imputation (c.f., Rao and Sitter, 1995), and multiple imputation (Rubin, 1987; Little and Rubin, 2002). This paper will focus on single imputation.

Single imputation is meant to fill a single value for the missing data. It includes mean imputation, ratio imputation, regression imputation, and hot deck imputation etc. When such an imputation is utilized to construct a confidence interval, a normal or *t* approximation is usually used. This may not be a very good approximation in practice. In this paper, we will

propose an empirical likelihood based confidence interval by combining single imputation approach for missing data. This work was motivated by the analysis of an AIDS clinical trial data set (see Section 5). CD4+ cell count is an important biomarkers in AIDS research (Liang, Wu, and Carroll, 2003; Wu, 2004), and have commonly been used to investigate the treatment effects, which may help clinicians more deeply understand AIDS pathogenesis and improve therapy. Although antiretroviral therapy for HIV-1 infected patients has been greatly improved in recent years, and administration of drug cocktails consisting of three or more drugs can reduce and maintain the viral load below the detection limit for many patients, it is unlikely that combination therapy can eradicate HIV in infected patients because of the existence of long-lived infected cells and sites within the body where drugs may not be effective. With the success of highly active antiretroviral therapy (HAART) against HIV infection, CD4+ cell counts can come back, and the infection is considered chronic. Clinicians and patients are therefore interested in monitoring the immunologic system (measured by CD4+ cell counts). However, there is a common challenge that CD4+ cell count is often missing because CD4+ cell counts and the viral load are measured at different time points. As discussed above, simply exclusion for the missing data is not wise. For the imputation of the missing CD4+ cell counts, the use of auxiliary information like treatment time is helpful. We will utilize the methods discussed in this paper to impute the missing CD4+ cell counts and then give the confidence intervals of the mean of CD4+ cell counts (Wu, Wong, and Wei, 2006). Careful investigation of this quantity is biologically and clinically important because it is a good biomarker for anti-HIV treatment and may be used to evaluate antiretroviral therapies.

The article is organized as follows. In Section 2, we briefly introduce two existing methods in the literature on the imputation for missing data using auxiliary information. In Section 3, we propose to use an empirical likelihood based confidence interval incorporating the imputed data. We illustrate the methods with intensive simulation experiments in Section 4, and analyze a data set from an AIDS study in Section 5. A discussion is provided in Section 6. The proof of the theoretical result is put in the appendix.

2 Jackknife-based Confidence Intervals

Assume that a group of subjects with the characteristic values (y, x) are independently observed n times, where y is the variable of interest with the mean θ, and x is an auxiliary variable. Let (y_i, x_i) be available for r_1 subjects, whose set is denoted by A_1, $\bar{y}_1$ and $\bar{x}_1$ be their sample means; only x_i be available for r_2 subjects, whose set is denoted by A_2, and $\bar{x}$ be the sample mean of the auxiliary variable over the sample set $s = A_1 + A_2$.

2.1 RS Method

For the missing data, Rao and Sitter (1995) used a ratio imputation approach to impute their values in the finite population inference: For $i \in A_2$, define $\widehat{y}_i = \frac{\bar{y}_1}{\bar{x}_1} x_i$. Correspondingly, an estimator of θ can be given by

$$\widehat{\theta}_{\mathrm{RS}} = \frac{\bar{y}_1}{\bar{x}_1} \bar{x}.$$

Under the following ratio model

$$y_i = \beta x_i + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad E(\varepsilon_i \varepsilon_j) = 0, \quad i \neq j, \quad V(\varepsilon_i) = \sigma^2 x_i,$$

$\hat{\theta}_{RS}$ is an unbiased estimator of θ regardless of missing mechanism. By using Jackknife approach, a variance estimator of $\hat{\theta}_{RS}$ can be obtained as

$$\mathrm{var}(\widehat{\theta}_{RS})=\left(\frac{\bar{x}}{\bar{x}_1}\right)^2\cdot\frac{1}{r_1}A+2\left(\frac{\bar{x}}{\bar{x}_1}\right)\cdot\frac{1}{n}B+\cdot\frac{1}{n}C,$$

where

$$A=\frac{1}{r_1-1}\sum_{j\in A_1}\left(y_j-\frac{\bar{y}_1}{\bar{x}_1}x_j\right)^2,$$

$$B=\frac{\bar{y}_1}{\bar{x}_1}\cdot\frac{1}{r_1-1}\sum_{j\in A_1}\left(y_j-\frac{\bar{y}_1}{\bar{x}_1}x_j\right)x_j,$$

and

$$C=\left(\frac{\bar{y}_1}{\bar{x}_1}\right)^2\cdot\frac{1}{n-1}\sum_{j\in s}(x_j-\bar{x})^2.$$

The reader is referred to Haziza and Picard (2008) for the good properties of the Jackknife variance estimators in the presence of imputed data. The confidence interval with the level of $1-\alpha$ is given by

$$\left(\widehat{\theta}_{RS}-z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_{RS})},\widehat{\theta}_{RS}+z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_{RS})}\right), \quad (1)$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

2.2 LZ Method

Recently, aiming to reduce the possible design-bias of the estimator $\hat{\theta}_{RS}$, Liu et al. (2006) developed a mean-of-ratio imputation approach to impute the missing data for finite population: For $i\in A_2$, let $\widehat{y_i}=\left(\frac{1}{r_1}\sum_{i\in A_1}\frac{y_i}{x_i}\right)x_i$. Their estimator for θ is

$$\widehat{\theta}_{LZ}=\bar{u}_1\bar{x}+\frac{(n-1)r_1}{(r_1-1)n}(\bar{y}_1-\bar{u}_1\bar{x}_1),$$

where $\bar{u}_1=\frac{1}{r_1}\sum_{j\in A_1}u_j$ with $u_j=y_j/x_j$.

Similarly, $\hat{\theta}_{\mathrm{LZ}}$ is an unbiased estimator under the ratio model regardless of missing mechanism.

Write

$$\begin{aligned} s_x^2(n) &= \tfrac{1}{n-1}\sum_{j\in s}(x_j-\bar{x})^2, \\ s_x^2(r_1) &= \tfrac{1}{r_1-1}\sum_{j\in A_1}(x_j-\bar{x}_1)^2, \\ s_{yx}(r_1) &= \tfrac{1}{r_1-1}\sum_{j\in A_1}(y_j-\bar{y}_1)(x_j-\bar{x}_1), \end{aligned}$$

and similarly for $s_y^2(r_1)$, $s_u^2(r_1)$, $s_{yu}(r_1)$, and $s_{xu}(r_1)$. The Jackknife variance estimator of $\hat{\theta}_{\mathrm{LZ}}$ is given by

$$\begin{aligned} \mathrm{var}(\widehat{\theta}_{\mathrm{LZ}}) = {} & \tfrac{\bar{u}_1^2 s_x^2(n)}{n-1} + \tfrac{(\bar{x}-\bar{x}_1)^2 s_u^2(r_1)}{r_1-1} + \tfrac{s_y^2(r_1)}{r_1-1} + \tfrac{\bar{u}_1^2 s_x^2(r_1)}{r_1-1} + \tfrac{2\bar{u}_1 s_{yx}(r_1)}{r_1-1} \\ & + \tfrac{2\bar{u}_1(\bar{x}-\bar{x}_1)s_{xu}(r_1)}{n-1} - \tfrac{2\bar{u}_1^2 s_x^2(r_1)}{n-1} + \tfrac{2(\bar{x}-\bar{x}_1)s_{yu}(r_1)}{r_1-1} \\ & - \tfrac{2\bar{u}_1(\bar{x}-\bar{x}_1)s_{xu}(r_1)}{r_1-1} - \tfrac{2\bar{u}_1 s_{yx}(r_1)}{r_1-1}. \end{aligned}$$

Correspondingly, the confidence interval with the level of $1-\alpha$ is given by

$$\left(\widehat{\theta}_{\mathrm{LZ}} - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_{\mathrm{LZ}})}, \widehat{\theta}_{\mathrm{LZ}} + z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{\theta}_{\mathrm{LZ}})}\right). \tag{2}$$

3 Empirical Likelihood-based Confidence Interval

The confidence intervals introduced in the previous section are based on a normal approximation. Such confidence intervals, although intuitive and easy to calculate, may not be true and in a consequence their numerical performance may not be optimistic in small sample sizes. In this section, we propose an alternative choice, empirical likelihood based confidence intervals, which have systematically been studied by Owen (2001). The basic idea is to give an empirical likelihood ratio and then prove that the ratio has a limiting chi-squared distribution, which assures one to obtain confidence intervals for a variety of settings. The empirical likelihood method has many advantages over its competitors, such as the normal-approximation-based method and the bootstrap method (see Hall and La Scala, 1990). The most appealing features of the empirical likelihood method include the improvement of confidence region, increase of accuracy of coverage because of using auxiliary information, and easy implementation. The method has been applied in a variety of topics, for example, linear models (Owen, 1991; Chen, 1993, [1994]), generalized linear models (Kolaczyk, 1994), and general estimating equation (Qin and Lawless, 1994).

In this section we use the empirical likelihood method combining the imputation for the missing data to construct a confidence interval for θ. This process for independent data without missingness was actually studied by Owen (1991), who introduced empirical likelihood ratio confidence regions and ascertained that the confidence intervals for a one dimensional mean are less adversely affected by skewness than those based on t statistic. Empirical likelihood procedure has been used to develop confidence intervals of the means of the response variables in linear and partially linear models under the setting of missing data; see Wang and Rao (2002) and Wang, Linton, and Härdle (2004). The topic we study here is similar to that in

Wang and Rao (2002). But unlike Wang and Rao (2002), we do not make any model assumption.

Let $(p_1, \ldots, p_n)$ be a probability vector. Our empirical likelihood ratio function for the mean θ is defined as

$$\mathscr{R}_n(\theta) = \max \left\{ \prod_{i=1}^{n} np_i : \sum_{i=1}^{n} p_i y_i^* = \theta,\ p_i \geq 0,\ \sum_{i=1}^{n} p_i = 1 \right\},$$

where $y_i^* = y_i$ for $i \in A_1$, and $= \hat{y}_i$ which is defined in Section 2.1 for $i \in A_2$. Under the missing completely at random (MCAR) structure, the resulting empirical likelihood based confidence interval for the mean in large sample sense is given by (3) below whose proof is provided in the appendix:

$$\{\theta : -2Q_n(\theta)\log\mathscr{R}_n(\theta) \leq \chi_1^2(\alpha)\}, \tag{3}$$

where $\chi_1^2(\alpha)$ is the $(1-\alpha)$– quantile of the chi-squared distribution with one degree of freedom, and $Q_n(\theta) = U_n(\theta)/V_n(\theta)$ with

$$U_n(\theta) = \frac{1}{n}\sum_{i\in A_1}(y_i - \theta)^2 + \frac{n-r_1}{n}\frac{1}{r_1}\sum_{i=1}^{n}\left(y_i - \frac{\theta x_i}{\bar{x}_i}\right)^2,$$
$$V_n(\theta) = \frac{1}{r_1}\sum_{i\in A_1}(y_i - \theta)^2 + \theta^2\frac{n-r_1}{r_1}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

4 Simulation Study

To compare the performances of the three confidence intervals introduced in Sections 2 and 3, we conduct intensive simulation studies. We generate the data with the sample sizes 15 and 30 from the following model:

$$\text{Model (I)}: y_i = 3x_i + \varepsilon_i,$$

where the distribution of the auxiliary variable x_i varies from the normal $N(2, 3)$, Gamma (20, 4) to Uniform (2, 5), and the distribution of the error term ε_i is normal $N(0, 0.5^2)$ or $0.25(\chi_2^2 - 2)$ with χ_2^2 being the chi-squared random variable with the degrees of freedom of 2. The missing rate for the variable of interest y varies from 50%, 25%, 10% to 0%.

We set the nominal level of confidence interval to be $1 - \alpha = 95\%$ and use formulae (1)-(3) to calculate the coverage rate, lower bound and upper bound of interval estimation, where for the empirical likelihood method, we combine the imputed data obtained by the ratio imputation approach (see Section 2.1, because the results are similar when we use the mean-of-ratio imputation approach in Section 2.2). We repeat our simulations 25,000 times. As a comparison, we also do our calculation using naive method, based on normal approximation, which uses the information only from the observed values of the variable of interest. The calculation results are presented in Tables 1-2 for model (I). We show the confidence intervals (CI) obtained by

the averages of lower and upper limits, the average CI lengths (LEN), and the coverage probabilities (COV). It is seen from Tables 1-2 that the naive method often provides low coverage probability especially for the serious missingness ($\geq 25\%$) or small sample size ($n = 15$). Overall, RS, LZ and empirical likelihood methods are superior to the naive method. Both of the LZ and empirical likelihood methods perform better than RS method. Among the three methods, RS generally gives the lowest coverage probability together with the shortest interval lengths, and LZ and empirical likelihood often present the coverage probabilities close to the nominal level. Further, LZ and empirical likelihood methods are comparable, and the latter is slightly better. On the other hand, the missing rate has no large effect on the three methods but the sample size has a relatively larger effect for RS method than for LZ and empirical likelihood methods. In addition, we have also conducted our simulations based on large samples ($n = 200$) which show that all of the four methods tend to have similar behaviors (data not shown).

5 Data Analysis of an AIDS Clinic Trial Group Study

It is well-known that the patients who have been tested positive for the HIV antibody test develop AIDS as HIV destroys their immune system. The immune system damage is evaluated by the CD4+ cell count. If a patient with HIV infection has a CD4+ cell count less than 200, he/she is said to have AIDS. Appropriately estimating CD4+ cell counts is therefore very helpful for evaluating treatment effects. In this section, we present an analysis of an AIDS clinical trial group (ACTG 315) study. Our purpose is to investigate the immunologic response in AIDS clinical trials. In this study, CD4+ cell counts were measured after initiation of an anti-viral therapy. There are a total of 514 observations with 13.8% of CD4+ cell counts missing from 53 subjects, with number of observations per subject ranging from 3 to 11. Most of the missing values of the CD4+ cell count occurred because it was measured at the time points different from those for the viral load measurements. In other words, the missingness does not depend on the values being missing, and it is reasonable to regard to be MCAR. The range of CD4+ cell counts is from 17.28 to 773.76 with mean 267.72 and median 268.96. Figure 1 presents a scatter plot of CD4+ cell count versus treatment time with a simple linear regression (solid line) and a horizontal line of $y = 200$ (dashed line) in the left panel, and a boxplot of CD4 + cell counts in the right panel.

When we simply discard the data corresponding to missing CD4+ cell counts, we obtain that the mean of the CD4+ cell counts is 266.54, and the associated 95% confidence interval is [256.52, 276.56]. We apply the methods presented in the previous sections to analyze the data set. The estimated means of the CD4+ cell counts based on RS and LZ methods are 270.51 and 270.75, and the corresponding confidence intervals are [260.45, 280.58] and [260.75, 280.76], respectively, while an empirical likelihood based confidence interval is [263.89, 281.31]. Three estimated values are similar. However, the interval estimation based on the empirical likelihood is the shortest among the four methods in which we place confidence on the basis of our simulation studies. It is also seen that the lower bounds of all confidence intervals are larger than 200, a critical value as we pointed out at the beginning of this section. Therefore the empirical likelihood based confidence interval provides the most accurate information.

6 Discussion

In this paper, we proposed to use an empirical likelihood based confidence interval by combining the imputed data. Simulation studies show that such confidence intervals perform well and are not largely affected by the missing rates. Our method was also applied to an analysis of a real data set from an AIDS clinic trial group study.

Generally, treating the imputed data as if they were true and using the estimator and its variance estimator based on full data would underestimate the true variance of the estimator. This would lead to a shorter confidence interval and so a lower actual coverage probability. Interestingly, it has been observed from our simulations using model (I) and its version with intercept term that the similar phenomenon does not occur for the empirical likelihood based confidence interval, even the missing rate is high (data not shown). This may owe to the sufficient auxiliary information and appropriate imputations for the missing data. Although showing the robustness in a sense, the empirical likelihood based confidence interval with complete data has been adjusted to theoretically attain the nominal level in the presence of missing data in this paper. Such an adjustment has been taken into account under the given models in literature (see, for example, Wang and Rao 2002). But we have made no assumption on the underlying models.

A fairly restrictive assumption with our proposed method is that data are MCAR. We can relax it by forming multiple imputation classes based on the auxiliary variable, as Rao and Shao did (1992). This article also assumed that the auxiliary information is complete, that is, the observation on the auxiliary variable x is available for all subjects. Clearly, this may not be true in practical problems. In other words, the auxiliary information itself may also be missing. In this scenario, the imputation methods for missing data can be found in Sitter and Rao (1997) and Liu et al. (2006). Correspondingly, our idea here can be extended to this case. On the other hand, for the missing data, we have used the ratio imputation and the mean-of-ratio imputation methods in the empirical likelihood based confidence interval. Naturally, some other imputations such as regression imputation and hot deck imputation can also be combined with the empirical likelihood methods. When the underlying model is unavailable, the modification of the usual empirical likelihood method under these imputations and its efficiency investigation warrant our further work.

Acknowledgments

The authors are grateful to the two referees for their insightful comments and suggestions which substantially improved an earlier version of this paper. This research was partially supported by two grants AI62247 and AI59773 from the National Institute of Allergy and Infectious Diseases. Zou's research was also partially supported by the grants 70625004, 10721101 and 70221001 from the National Natural Science Foundation of China.

Appendix: The derivation of formula (3)

To derive the empirical likelihood based confidence interval (3), we consider the following incomplete observations:

$$(y_i, x_i, \delta_i), \quad i=1, \ldots, n,$$

where the auxiliary variable x_i is observed completely, and $\delta_i = 0$ if y_i is missing and = 1 otherwise. We assume that these observations are i.i.d. and the missingness of the variable of interest, y, is MCAR. Also, assume that y_i and x_i have finite two moments. Define $z_i=y_i^* - \theta$. Then we have

$$\sum_{i=1}^{n} z_i = \sum_{i=1}^{n} (\delta_i y_i - \theta) + \sum_{i=1}^{n} (1 - \delta_i) x_i \cdot \frac{\bar{y}_1}{\bar{x}_1}.$$

Write

$$\zeta_i=\left[\delta_i y_i-\theta E(\delta),(1-\delta_i)x_i-E\{(1-\delta)X\},\delta_i\left\{y_i-\frac{\theta}{E(X)}x_i\right\}\right]'$$

and

$$a_n=\left\{1,\frac{\bar{y}_1}{\bar{x}_1},\frac{nE(X)}{n_1\bar{x}_1}\cdot E(1-\delta)\right\}'.$$

It can be shown that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_i=a_n^{'}\cdot\sqrt{n}\bar{\zeta}_n, \tag{A.1}$$

where $\bar{\zeta}_n=\frac{1}{n}\sum_{i=1}^{n}\zeta_i$.

It is clear that ζ_i are i.i.d. with the mean $E(\zeta_i) = 0$ and the variance-covariance matrix, say Σ, as follows.

$$\begin{pmatrix} E(\delta)E(Y^2)-\{E(\delta)\}^2\theta^2 & -E(\delta)E(1-\delta)\theta E(X) & E(\delta)E(Y^2)-E(\delta)\frac{\theta E(YX)}{E(X)} \\ -E(\delta)E(1-\delta)\theta E(X) & E(1-\delta)E(X^2)-\{E(1-\delta)\}^2(EX)^2 & 0 \\ E(\delta)E(Y^2)-E(\delta)\frac{\theta E(YX)}{E(X)} & 0 & E(\delta)E\left(Y-\frac{\theta}{EX}X\right)^2 \end{pmatrix}.$$

So from the central limit theorem, we obtain $\sqrt{n}\bar{\zeta}_n\to N(0,\sum)$ when $n\to\infty$. Noting that $a_n\to\left\{1,\frac{\theta}{EX},\frac{E(1-\delta)}{E(\delta)}\right\}'\equiv a$ in probability, we have $a_n'\cdot\sqrt{n}\bar{\zeta}_n\to N(0,a'\sum a)$, which together with (A.1) implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_i\to N(0,a'\sum a).$$

After some algebraic calculations, we see that

$$\begin{aligned} a'\sum a&=E(Y-\theta)^2+\tfrac{E(1-\delta)}{E(\delta)}E\left(Y-\tfrac{\theta}{EX}X\right)^2 \\ &\equiv V(\theta). \end{aligned}$$

Furthermore, it is readily shown that

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 \to E(\delta)E(Y-\theta)^2+\theta^2(E\delta-1)+E(1-\delta)\theta^2\frac{E(X^2)}{(EX)^2} \equiv U(\theta)$$

in probability. Also, it is easy to verify that $U_n(\theta) \to U(\theta)$ and $V_n(\theta) \to V(\theta)$ in probability. Thus,

$$Q_n(\theta)\cdot\left(\frac{1}{n}\sum_{i=1}^{n} z_i^2\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i\right)^2 \to \chi_1^2 \tag{A.2}$$

in distribution.

On the other hand, mimicking the proof of Theorem 3.2 of Owen (2001), we have $|\lambda| = O_p(n^{-1/2})$. A direct simplification yields

$$\begin{aligned} -2\log\{R_n(\theta)\} &= 2\sum_{i=1}^{n}\log(1+\lambda z_i) \\ &= 2\sum_{i=1}^{n}\left\{\lambda z_i - \tfrac{1}{2}(\lambda z_i)^2\right\}+o_p(1), \end{aligned} \tag{A.3}$$

where λ satisfies $\frac{1}{n}\sum_{i=1}^{n} z_i/(1+\lambda z_i)=0$. Noting that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{z_i}{1+\lambda z_i}=\frac{1}{n}\sum_{i=1}^{n} z_i(1-\lambda z_i)+\frac{1}{n}\sum_{i=1}^{n}\frac{(\lambda z_i)^2 z_i}{1+\lambda z_i},$$

we have

$$\lambda=\left(\sum_{i=1}^{n} z_i^2\right)^{-1}\sum_{i=1}^{n} z_i+o_p(n^{-1/2}). \tag{A.4}$$

Further,

$$0=\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda z_i}{1+\lambda z_i}=\frac{1}{n}\sum_{i=1}^{n}\lambda z_i-\frac{1}{n}\sum_{i=1}^{n}(\lambda z_i)^2+o_p(n^{-1}).$$

So

$$\sum_{i=1}^{n}\lambda z_i=\sum_{i=1}^{n}(\lambda z_i)^2+o_p(1). \tag{A.5}$$

From (A.3)-(A.5), we obtain

$$\begin{aligned}-2\log\{R_n(\theta)\} &= \sum_{i=1}^{n}\lambda z_i+o_p(1)\\ &= \left(\frac{1}{n}\sum_{i=1}^{n}z_i^2\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_i\right)^2+o_p(1).\end{aligned}$$

Combining this and formula (A.2), we obtain the empirical likelihood based confidence interval (3).

References

Chen SX. On the accuracy of empirical likelihood confidence regions for linear regression model. Ann Inst Statist Math 1993;45:621–637.

Chen SX. Empirical likelihood confidence intervals for linear regression coefficients. J Mult Anal 1994;49:24–40.

Cochran, WG. Sampling Techniques. Vol. 3rd. New York: John Wiley & Sons; 1977.

Hall P, La Scala B. Methodology and algorithms of empirical likelihood. Int Statist Rev 1990;58:109–127.

Haziza, D.; Picard, F. Jackknife variance estimation in the presence of imputed data. The Proceedings of the Workshop on Calibration and Estimation in Surveys; 2008.

Ibrahim JG, Chen MH, Lipsitz SR. Monte Carlo EM for missing covariates in parametric regression models. Biometrics 1999;55:591–596. [PubMed: 11318219]

Ibrahim JG, Lipsitz SR, Horton N. Using auxiliary data for parameter estimation with nonignorable missing outcomes. Appl Statist 2001;50:361–373.

Kolaczyk ED. Empirical likelihood for generalized linear models. Statist Sinica 1994;4:199–218.

Liang H, Wang SJ, Robins J, Carroll R. Estimation in partially linear models with missing covariates. J Am Statist Assoc 2004;99:357–367.

Liang H, Wu HL, Carroll RJ. The relationship between virologic and immunologic responses in AIDS clinical research using mixed-effect varying-coefficient semiparametric models with measurement error. Biostatistics 2003;4:297–312. [PubMed: 12925523]

Little, RJA.; Rubin, DB. Statistical Analysis with Missing Data. New York: John Wiley & Sons; 2002.

Liu L, Tu Y, Li Y, Zou G. Imputation for missing data and variance estimation when auxiliary information is incomplete. Model Assisted Statist Appl 2006;1:83–94.

Owen AB. Empirical likelihood for linear models. Ann Statist 1991;19:1725–1747.

Owen, AB. Empirical Likelihood. London: Chapman and Hall; 2001.

Qin J, Lawless J. Empirical likelihood and general estimating equations. Ann Statist 1994;22:300–325.

Rao JNK, Shao J. Jackknife variance estimation with survey data under hot deck imputation. Biometrika 1992;79:811–822.

Rao JNK, Sitter RR. Variance estimation under two-phase sampling with application to imputation for missing data. Biometrika 1995;82:453–460.

Rubin, DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons; 1987.

Sitter RR, Rao JNK. Imputation for missing values and corresponding variance estimation. Can J Statist 1997;25:61–73.

Wang QH, Linton O, Härdle W. Semiparametric Regression Analysis with Missing Response at Random. J Am Statist Assoc 2004;99:334–345.

Wang QH, Rao JNK. Empirical likelihood-based inference in linear models with missing data. Scan J Statist 2002;29:563–576.

Wu JR, Wong ACM, Wei W. Interval estimation of the mean response in a log-regression model. Statist Med 2006;25:2125–2135.

Wu L. Exact and approximate inferences for nonlinear mixed-effects models with missing covariates. J Am Statist Assoc 2004;99:700–709.

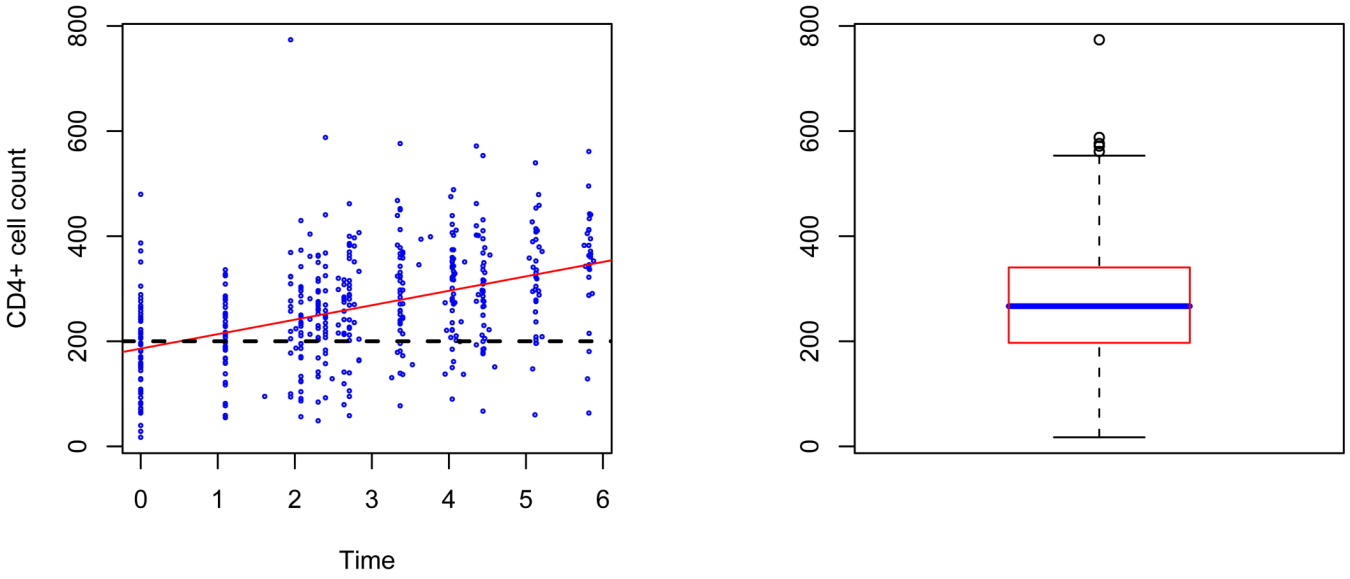

Figure 1.
Scatter plot of CD4+ cell counts versus treatment time (left panel) and boxplot of CD4+ cell counts.

Table 1

The 95% confidence intervals based on the naive, RS, LZ, and the empirical likelihood (EL) methods, and the associated coverages (COV) and the lengths (LEN) under the different configurations with missing rate (MR) for the simulated data generated from model (II) when ε follows $N(0, 0.5^2)$.

			Naive			RS			LZ			EL		
X	**MR**	***n***	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**
norm	0.5	15	(2.74, 14.99)	12.25	89.60	(0.44, 15.12)	14.68	93.50	(0.59, 15.06)	14.47	95.00	(0.97, 14.41)	13.44	93.70
		30	(4.18, 13.07)	8.89	94.00	(4.42, 11.43)	7.01	94.90	(3.38, 12.65)	9.27	95.60	(3.88, 11.90)	8.02	94.40
	0.25	15	(3.13, 13.48)	10.34	93.00	(2.70, 13.32)	10.62	93.50	(2.14, 13.90)	11.76	94.00	(2.88, 13.05)	10.17	93.10
		30	(4.43, 11.81)	7.38	94.40	(4.78, 11.33)	6.55	94.30	(4.07, 11.99)	7.92	95.40	(4.67, 11.42)	6.75	93.90
	0.1	15	(3.34, 12.81)	9.47	93.00	(2.25, 13.84)	11.59	93.50	(2.77, 13.32)	10.55	94.00	(3.14, 12.93)	9.80	94.00
		30	(4.65, 11.38)	6.74	94.10	(4.79, 11.28)	6.49	94.30	(4.32, 11.69)	7.37	94.90	(4.67, 11.39)	6.72	95.00
	0	15	(3.52, 12.47)	8.96	93.10	(3.52, 12.47)	8.96	93.10	(3.36, 12.63)	9.27	94.00	(3.16, 12.83)	9.66	94.50
		30	(4.80, 11.19)	6.39	94.00	(4.80, 11.19)	6.39	94.00	(4.74, 11.24)	6.50	94.30	(4.64, 11.35)	6.71	94.90
gamm	0.5	15	(14.74, 19.40)	4.66	88.40	(15.27, 18.71)	3.44	92.80	(15.19, 18.77)	3.58	93.80	(14.52, 19.82)	5.29	96.30
		30	(15.41, 18.79)	3.39	92.20	(15.77, 18.22)	2.46	93.90	(15.74, 18.24)	2.50	94.40	(15.34, 18.88)	3.54	95.30
	0.25	15	(15.10, 18.97)	3.88	91.70	(15.31, 18.70)	3.39	92.80	(15.24, 18.76)	3.51	93.70	(15.01, 19.24)	4.23	95.00
		30	(15.64, 18.42)	2.78	93.50	(15.79, 18.21)	2.42	93.90	(15.77, 18.23)	2.46	94.20	(15.65, 18.51)	2.86	94.30
	0.1	15	(15.23, 18.78)	3.55	92.40	(15.31, 18.68)	3.37	92.80	(15.25, 18.74)	3.49	93.70	(15.20, 19.00)	3.80	94.60
		30	(15.74, 18.27)	2.54	93.90	(15.79, 18.20)	2.41	94.10	(15.77, 18.23)	2.46	94.50	(15.76, 18.35)	2.59	94.50
	0	15	(15.32, 18.69)	3.36	92.60	(15.32, 18.69)	3.36	92.60	(15.27, 18.75)	3.48	93.50	(15.29, 18.91)	3.62	94.10
		30	(15.79, 18.19)	2.40	93.70	(15.79, 18.19)	2.40	93.70	(15.77, 18.21)	2.44	94.20	(15.79, 18.31)	2.52	94.80
unif	0.5	15	(10.75, 14.43)	3.68	88.10	(11.11, 13.87)	2.76	92.30	(11.04, 13.93)	2.89	93.30	(10.49, 14.50)	4.01	96.80
		30	(11.26, 13.92)	2.66	92.10	(11.51, 13.47)	1.95	93.70	(11.48, 13.48)	2.00	94.10	(11.17, 13.81)	2.64	95.10
	0.25	15	(10.99, 14.05)	3.06	91.50	(11.15, 13.84)	2.69	92.70	(11.10, 13.89)	2.79	93.50	(10.88, 14.12)	3.24	95.50
		30	(11.43, 13.61)	2.18	93.60	(11.54, 13.45)	1.91	94.20	(11.52, 13.46)	1.94	94.50	(11.41, 13.57)	2.16	94.50
	0.1	15	(11.11, 13.90)	2.80	92.20	(11.16, 13.83)	2.67	92.70	(11.12, 13.88)	2.76	93.60	(11.05, 13.95)	2.90	95.20
		30	(11.51, 13.50)	1.99	93.90	(11.55, 13.45)	1.89	93.80	(11.54, 13.46)	1.93	94.20	(11.52, 13.48)	1.97	94.80
	0	15	(11.17, 13.83)	2.65	92.80	(11.17, 13.83)	2.65	92.80	(11.12, 13.87)	2.75	93.60	(11.11, 13.88)	2.77	95.60
		30	(11.56, 13.44)	1.89	93.70	(11.56, 13.44)	1.89	93.70	(11.54, 13.46)	1.92	94.10	(11.54, 13.46)	1.93	95.30

Table 2

The 95% confidence intervals based on the naive, RS, LZ, and the empirical likelihood (EL) methods, and the associated coverages (COV) and the lengths (LEN) under the different configurations with missing rate (MR) for the simulated data generated from model (II) when ε follows $0.25(\chi^2_2 - 2)$.

			Naive			RS			LZ			EL		
X	**MR**	***n***	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**	**CI**	**LEN**	**COV**
norm	0.5	15	(2.70, 14.94)	12.25	89.80	(0.52, 14.91)	14.39	93.40	(0.05, 15.32)	15.26	94.80	(0.86, 14.38)	13.52	93.80
		30	(4.18, 13.09)	8.91	94.30	(4.43, 11.47)	7.04	95.20	(3.46, 12.66)	9.20	96.10	(3.90, 11.93)	8.03	94.80
	0.25	15	(3.17, 13.50)	10.33	93.00	(1.67, 14.38)	12.70	93.80	(2.27, 13.83)	11.56	94.40	(2.83, 13.12)	10.29	93.40
		30	(4.43, 11.81)	7.38	94.70	(4.78, 11.31)	6.52	94.70	(4.05, 11.98)	7.93	95.40	(4.66, 11.41)	6.74	94.30
	0.1	15	(3.33, 12.81)	9.47	93.40	(2.90, 13.19)	10.29	93.80	(2.78, 13.25)	10.47	94.20	(3.17, 12.90)	9.73	94.10
		30	(4.63, 11.38)	6.75	94.00	(4.80, 11.23)	6.43	94.30	(4.32, 11.68)	7.36	95.00	(4.66, 11.38)	6.72	95.20
	0	15	(3.51, 12.48)	8.97	92.90	(3.51, 12.48)	8.97	92.90	(3.36, 12.64)	9.28	93.80	(3.16, 12.82)	9.67	94.40
		30	(4.80, 11.19)	6.39	94.20	(4.80, 11.19)	6.39	94.20	(4.75, 11.25)	6.50	94.50	(4.65, 11.35)	6.70	95.10
gamm	0.5	15	(14.75, 19.41)	4.65	88.50	(15.28, 18.72)	3.44	92.40	(15.20, 18.78)	3.58	93.30	(14.54, 19.82)	5.29	96.20
		30	(15.40, 18.79)	3.38	92.20	(15.76, 18.21)	2.45	94.10	(15.73, 18.23)	2.50	94.60	(15.33, 18.88)	3.55	95.50
	0.25	15	(15.10, 18.98)	3.88	91.50	(15.31, 18.70)	3.39	92.70	(15.25, 18.76)	3.52	93.60	(15.02, 19.24)	4.22	94.80
		30	(15.64, 18.42)	2.78	93.60	(15.79, 18.20)	2.42	94.00	(15.76, 18.23)	2.46	94.40	(15.65, 18.49)	2.84	94.30
	0.1	15	(15.24, 18.79)	3.55	92.70	(15.31, 18.69)	3.38	92.90	(15.25, 18.75)	3.50	93.70	(15.21, 19.01)	3.80	94.30
		30	(15.75, 18.28)	2.54	93.90	(15.80, 18.21)	2.41	94.00	(15.78, 18.23)	2.45	94.50	(15.77, 18.36)	2.59	94.50
	0	15	(15.32, 18.69)	3.37	92.80	(15.32, 18.69)	3.37	92.80	(15.26, 18.75)	3.49	93.80	(15.29, 18.92)	3.64	94.40
		30	(15.81, 18.21)	2.40	93.70	(15.81, 18.21)	2.40	93.70	(15.79, 18.23)	2.44	94.00	(15.80, 18.32)	2.52	94.60
unif	0.5	15	(10.72, 14.42)	3.70	87.70	(11.11, 13.88)	2.76	92.30	(11.04, 13.93)	2.89	93.30	(10.49, 14.51)	4.02	96.50
		30	(11.26, 13.92)	2.65	92.00	(11.51, 13.46)	1.96	93.70	(11.48, 13.48)	2.00	94.10	(11.17, 13.80)	2.63	95.10
	0.25	15	(11.00, 14.05)	3.06	91.50	(11.16, 13.84)	2.69	92.70	(11.10, 13.89)	2.79	93.60	(10.89, 14.12)	3.23	95.50
		30	(11.44, 13.62)	2.18	93.40	(11.54, 13.45)	1.91	93.90	(11.52, 13.46)	1.94	94.30	(11.42, 13.57)	2.15	94.50
	0.1	15	(11.11, 13.91)	2.80	92.70	(11.17, 13.83)	2.66	92.80	(11.12, 13.88)	2.76	93.60	(11.05, 13.96)	2.91	95.20
		30	(11.51, 13.50)	1.99	94.00	(11.55, 13.44)	1.89	94.00	(11.53, 13.46)	1.92	94.30	(11.52, 13.48)	1.96	94.70
	0	15	(11.18, 13.84)	2.66	93.00	(11.18, 13.84)	2.66	93.00	(11.13, 13.88)	2.75	93.70	(11.12, 13.89)	2.77	95.60
		30	(11.55, 13.44)	1.89	94.00	(11.55, 13.44)	1.89	94.00	(11.54, 13.46)	1.92	94.40	(11.54, 13.46)	1.93	95.60