Computers in Biology and Medicine 91 (2017) 191–197

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Precise confidence intervals of regression-based reference limits: Method comparisons and sample size requirements

Gwowen Shieh

Department of Management Science, National Chiao Tung University, Hsinchu, 30010, Taiwan, ROC

ARTICLE INFO

Keywords:
Percentile
Precision
Quantile
Reference limit
Sample size

ABSTRACT

Covariate-dependent reference limits have been extensively applied in biology and medicine for determining the substantial magnitude and relative importance of quantitative measurements. Confidence interval and sample size procedures are available for studying regression-based reference limits. However, the existing popular methods employ different technical simplifications and are applicable only in certain limited situations. This paper describes exact confidence intervals of regression-based reference limits and compares the exact approach with the approximate methods under a wide range of model configurations. Using the ratio between the widths of confidence interval and reference interval as the relative precision index, optimal sample size procedures are presented for precise interval estimation under expected ratio and tolerance probability considerations. Simulation results show that the approximate interval methods using normal distribution have inaccurate confidence limits. The exact confidence intervals dominate the approximate procedures in one- and two-sided coverage performance. Unlike the current simplifications, the proposed sample size procedures integrate all key factors including covariate features in the optimization process and are suitable for various regression-based reference limit studies with potentially diverse configurations. The exact interval estimation has theoretical and practical advantages over the approximate methods. The corresponding sample size procedures and computing algorithms are also presented to facilitate the data analysis and research design of regression-based reference limits.

1. Introduction

Percentiles provide informative reference values for determining the substantial magnitude and relative importance of quantitative measurements in biology and medicine. An important application is the 95% limits of agreement of Refs. [1,2] for comparing different measurement methods of the same medical quantity. To incorporate the influential covariate effects such as age, height, and weight, various parametric and nonparametric procedures for estimating covariate-dependent reference limits have been presented. In particular, regression analysis has been recommended to improve the precision for the estimation of covariate-dependent reference limits in Refs. [3–5]; among others. In view of these practical applications, this paper focuses on the versatile framework of multiple linear regression for its recognized usefulness in accommodating covariate measurements. Comprehensive reviews and general guidelines of the most commonly used methods for modeling percentile curves can be found in Refs. [6,7].

Because an individual estimate does not reflect the inherent variation, it is usually recommended to construct confidence intervals for the target parameters. General expositions and comprehensive guidelines of interval estimation are available in Refs. [8–11]. However, the problem of interval estimation of regression-based reference limits has received little attention and there are only a few pertinent discussions in the literature. Within the context of simple linear regression, an exact procedure was presented in Virtanen et al. [5] for obtaining confidence intervals of reference limits. But its use is impeded by the difficult computation of critical values associated with special algorithm not readily available in standard statistical packages. Alternatively, Virtanen et al. [5] proposed a normal-based approximation for its ease of implementation. In addition, following the same line of normal approximation, Bellera and Hanley [12] suggested a further simplified confidence interval formula.

Apparently, the approximate interval methods of Refs. [5,12] are computationally simple because they only involve standard mathematical equations and the quantiles of a normal distribution. Moreover, the resulting confidence intervals carry the symmetry property of a normal distribution and the confidence interval endpoints are equidistant about the primary statistic. It is essential to note that unlike a normal distribution, a noncentral t distribution is generally skewed, especially when sample size is small and noncentrality considerably deviates from 0.

E-mail address: gwshieh@mail.nctu.edu.tw.

https://doi.org/10.1016/j.compbiomed.2017.10.015
Received 25 August 2017; Received in revised form 29 September 2017; Accepted 14 October 2017

Accordingly, the actual value of normal percentiles has a noncentral t distribution and the associated exact confidence intervals are not symmetric around the desired point estimate [9,10]. Because regression-based reference limits are direct generalizations of normal percentiles, this suggests that an ideal interval procedure should also adopt asymmetric confidence intervals for the regression-based reference limits. However, this fundamental issue was not addressed in Refs. [5,12]. It is prudent to elucidate the underlying behavior and discrepancy of their normal-based methods to be accepted as a reliable technique.

In view of the absence of methodological clarification and supportive technique for interval estimation and research design of regression-based reference limits, this paper has two goals. The first is to describe exact confidence intervals for regression-based reference limits under the general context of multiple linear regression with one or more covariate variables. Accordingly, the suggested exact confidence interval approach improves the current popular methods by relaxing the model assumption of a single covariate and by correcting the theoretical deficiency of a normal approximation. Comprehensive numerical investigations are provided to illustrate the accuracy of the proposed technique and the disadvantages of the approximated methods. The second goal is to provide sample size procedures for precise interval estimation of regression-based reference limits. The required precision of a confidence interval is evaluated with respect to the ratio of the widths of a confidence interval and the reference limits under the considerations of the magnitude of expected ratio and the tolerance probability of ratio within a designated threshold. In view of the general availability of statistical software packages SAS and R, computer algorithms are developed to facilitate the implementation of the suggested confidence interval and sample size computations.

2. Confidence interval estimation

Consider the multiple linear regression model for associating the response variable Y with the covariate variables X_1, …, X_P:

$$Y_i = \beta_0 + \sum_{k=1}^{K} X_{ik}\beta_k + \varepsilon_i \tag{1}$$

where Y_i is the observed value of the ith subject on the response variable; X_{i1}, …, X_{iK} are the observed values of the ith subject on the continuous covariate variables; β_0, β_1, …, β_K are unknown coefficient parameters; and ε_i are iid $N(0, \sigma^2)$ random errors for $i = 1, \ldots, N$. The mean response $\mu = E[Y|\{X_{01}, \ldots, X_{0K}\}] = \beta_0 + \sum_{k=1}^{K} X_{0k}\beta_k$ at the specified set of values $\{X_{01}, \ldots, X_{0K}\}$ of the covariate variables is estimated by

$$\hat{\mu} = \hat{\beta}_0 + \sum_{k=1}^{K} X_{0k}\hat{\beta}_k, \tag{2}$$

where $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_K$ are the least squares estimators of the regression coefficients β_0, β_1, …, β_K. Fundamental principles and technical explications on regression analysis have been well documented in the standard texts of Refs. [13,14], among others.

In some practical applications, it may be more useful to consider the regression-based 100pth percentile or reference limit θ of the distribution $Y|\{X_{01}, \ldots, X_{0K}\} \sim N(\mu, \sigma^2)$, where

$$\theta = \mu + z_p\sigma \tag{3}$$

and z_p is the 100pth percentile of the standard normal distribution $N(0, 1)$. Evidently, the mean response μ is the 50th percentile of the normal distribution $N(\mu, \sigma^2)$ and therefore μ is a special case of θ. To estimate the percentile θ, an intuitive and simple formula is

$$\hat{\theta}_B = \hat{\mu} + z_p\hat{\sigma}, \tag{4}$$

where $\hat{\sigma}^2 = SSE/\nu$ is the usual unbiased estimator of σ^2, SSE is the error sum of squares, and $\nu = N - K - 1$. Note that $\hat{\theta}_B$ is a biased estimator because $E[\hat{\sigma}] = \sigma/c$ where $c = (\nu/2)^{1/2}\Gamma(\nu/2)/\Gamma\{(\nu+1)/2\} > 1$ for $\nu > 0$. Alternatively, the minimum variance unbiased estimator of θ is

$$\hat{\theta}_{MU} = \hat{\mu} + z_p c\hat{\sigma}. \tag{5}$$

Moreover, it follows from the standard results in Rencher and Schaalje [14] that

$$\hat{\mu} \sim N(\mu, W\sigma^2), \tag{6}$$

where $W = 1/N + (\mathbf{X}_0 - \overline{\mathbf{X}})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{X}_0 - \overline{\mathbf{X}})$, $\mathbf{X}_0 = (X_{01}, \ldots, X_{0K})^{\mathrm{T}}$, $\overline{\mathbf{X}} = \sum_{i=1}^{N}\mathbf{X}_i/N$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{iK})^{\mathrm{T}}$, and $\mathbf{A} = \sum_{i=1}^{N}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^{\mathrm{T}}$. Note that the variance and mean square error of $\hat{\theta}_B$ are $Var[\hat{\theta}_B] = \{W + z_p^2(1-1/c^2)\}\sigma^2$ and $MSE[\hat{\theta}_B] = \{W + 2z_p^2(1-1/c)\}\sigma^2$, respectively. Also, it can be shown that $Var[\hat{\theta}_{MU}] = MSE[\hat{\theta}_{MU}] = \{W + z_p^2(c^2 - 1)\}\sigma^2$. These results reveal that $Var[\hat{\theta}_B] < Var[\hat{\theta}_{MU}]$ and $MSE[\hat{\theta}_B] < MSE[\hat{\theta}_{MU}]$ because $c > 1$ and $W > 0$. Therefore, despite $\hat{\theta}_B$ is a biased estimator of θ, it outperforms $\hat{\theta}_{MU}$ under the variance and mean square error considerations.

2.1. Exact approach

To obtain confidence intervals for regression-based reference limit θ, standard derivations show that

$$T* = \frac{\hat{\mu} - \theta}{\left(W\hat{\sigma}^2\right)^{1/2}} \sim t\left(\nu,\ -z_p/W^{1/2}\right), \tag{7}$$

where $t(\nu, -z_p/W^{1/2})$ is a noncentral t distribution with degrees of freedom ν and noncentrality parameter $-z_p/W^{1/2}$ ([15], Chapter 31). Hence, T^* provides a pivotal quantity for constructing confidence intervals of the regression-based 100pth percentile θ. Let $t_{1-\alpha}(\nu, \Delta)$ denote the $100(1-\alpha)$th percentile of the distribution $t(\nu, \Delta)$ and it is important to note that $t_{1-\alpha}(\nu, -\Delta) = -t_{\alpha}(\nu, \Delta)$ for $0 < \alpha < 1$. Thus, a $100(1-\alpha)\%$ two-sided confidence interval of θ with equal tail probability is readily obtained as $\{\hat{\theta}_{EL}, \hat{\theta}_{EU}\}$ where

$$\begin{aligned}\hat{\theta}_{EL} &= \hat{\mu} + t_{\alpha/2}\left(\nu, z_p/W^{1/2}\right)W^{1/2}\hat{\sigma} \text{ and } \hat{\theta}_{EU} \\ &= \hat{\mu} + t_{1-\alpha/2}\left(\nu, z_p/W^{1/2}\right)W^{1/2}\hat{\sigma}.\end{aligned} \tag{8}$$

For the sake of completeness, the one-sided confidence intervals are described here as well. An upper $100(1-\alpha)\%$ one-sided confidence interval of θ can be expressed as $\{\hat{\theta}_{EL}, \infty\}$ and the lower confidence limit is

$$\hat{\theta}_{EL} = \hat{\mu} + t_{\alpha}\left(\nu, z_p/W^{1/2}\right)W^{1/2}\hat{\sigma}. \tag{9}$$

Moreover, a lower $100(1-\alpha)\%$ one-sided confidence interval of θ is $\{-\infty, \hat{\theta}_{EU}\}$ and the upper confidence limit has the form

$$\hat{\theta}_{EU} = \hat{\mu} + t_{1-\alpha}\left(\nu, z_p/W^{1/2}\right)W^{1/2}\hat{\sigma}. \tag{10}$$

For the special case of simple linear regression with $K = 1$ Virtanen et al. [5], presented exact confidence intervals of θ without referring to a noncentral t distribution. Therefore, their formulation differs from $\{\hat{\theta}_{EL}, \hat{\theta}_{EU}\}$ given in Equation (8) and requires specific critical values for attaining the actual confidence intervals. Because the computation of critical values for the exact confidence limits of Virtanen et al. [5] involves special algorithm not readily available in standard statistical packages, it appears that the particular formula has been disregarded. In contrast, the proposed exact confidence interval procedure is conceptually transparent and can be readily implemented with the embedded functions of modern software systems as presented in the supplemental SAS/IML and R algorithms.