Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2012, Article ID 639124, 12 pages
doi:10.1155/2012/639124

Research Article

Let Continuous Outcome Variables Remain Continuous

Enayatollah Bakhshi,[1] Brian McArdle,[2] Kazem Mohammad,[3] Behjat Seifi,[4] and Akbar Biglarian[1]

[1] *Department of Statistics and Computer, University of Social Welfare and Rehabilitation Sciences, Tehran 1985713834, Iran*
[2] *Department of Statistics, The University of Auckland, Private Bag 92010, Auckland, New Zealand*
[3] *Department of Biostatistics, School of Public Health and Institute of Public Health Research, Tehran University of Medical Sciences, Tehran, Iran*
[4] *Department of Physiology, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran*

Correspondence should be addressed to Enayatollah Bakhshi, bakhshi@razi.tums.ac.ir

Received 8 November 2011; Revised 21 February 2012; Accepted 29 February 2012

Academic Editor: Alberto Guillén

The complementary log-log is an alternative to logistic model. In many areas of research, the outcome data are continuous. We aim to provide a procedure that allows the researcher to estimate the coefficients of the complementary log-log model without dichotomizing and without loss of information. We show that the sample size required for a specific power of the proposed approach is substantially smaller than the dichotomizing method. We find that estimators derived from proposed method are consistently more efficient than dichotomizing method. To illustrate the use of proposed method, we employ the data arising from the NHSI.

1. Introduction

Recently, logistic regression has become a popular tool in biomedical studies. The parameter in logistic regression has the interpretation of log odds ratio, which is easy for people such as physicians to understand. Probit and complementary log-log are alternatives to logistic model. For a covariate X and a binary response variable Y, let $\pi(X) = P(Y = 1 \mid X = x)$. A related model to the complementary log-log link is the log-log link. For it, $\pi(x)$ approaches 0 sharply but approaches 1 slowly. When the complementary log-log model holds for the probability of a success, the log-log model holds for the probability of a failure [1].

These models use a categorical (dichotomous or polytomous) outcome variable. In many areas of research, the outcome data are continuous. Many researchers have no hesitation in dichotomizing a continuous variable, but this practice does not make use of within-category information. Several investigators have noted the disadvantages of dichotomizing both independent and outcome variables [2–10]. Ragland [11] showed that the magnitude of odds ratio and statistical power depend on the cutpoint used to dichotomize the response variable. From a clinical point of view, binary outcomes may be preferred for some reasons such as (1) setting diagnostic criteria for disease, (2) offering a simpler interpretation of common effect measures from statistical models such as odds ratios and relative risks. However, all advantages come at the lost information. From a statistical point of view, this loss of information means more samples which are required to attain prespecified powers.

Moser and Coombs [12] provided a closed-form relationship that allows a direct comparison between the logistic and linear regression coefficients. They also provided a procedure that allows the researcher to analyze the original continuous outcome without dichotomizing. To date, a method that applies the complementary log-log model without dichotomizing and without loss of information has not been available.

We aim to (a) provide a method that allows the researcher to estimate the coefficients of the complementary log-log model without dichotomizing and without loss of information, (b) show that the coefficient of the complementary log-log model can be interpreted in terms of the regression coefficients, (c) demonstrate that the coefficient estimates from

this method have smaller variances and shorter confidence intervals than the dichotomizing method.

2. Methods

2.1. Model. Let $y_1, y_2, \ldots, y_n$ be n independent observations on y, and let $x_1, x_2, \ldots, x_{p-1}$ be $p-1$ predictor variables thought to be related to the response variable y. The multiple linear regression model for the ith observation can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{ip-1} + E_i \quad i = 1, 2, \ldots, n, \tag{1}$$

or

$$y_i = x_i \beta + E_i \quad i = 1, 2, \ldots, n, \tag{2}$$

where

$$x_i = \left(1, x_{i1}, x_{i2}, \ldots, x_{ip-1}\right). \tag{3}$$

To complete the model, we make the following assumptions:

(1) $E(E_i) = 0$ for $i = 1, 2, \ldots, n$,

(2) $\operatorname{var}(E_i) = \sigma^2$ for $i = 1, 2, \ldots, n$,

(3) the independent E_i follows an extreme value distribution for $i = 1, 2, \ldots, n$.

Writing the model for each of the n observations, in matrix form, we have

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 x_{11} & x_{12} & \ldots & x_{1p-1} \\ 1 x_{21} & x_{22} & \ldots & x_{2p-1} \\ . & & & \\ . & & & \\ 1 x_{n1} & x_{n2} & \ldots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ . \\ . \\ E_n \end{bmatrix}, \tag{4}$$

or

$$y = X\beta + E. \tag{5}$$

The preceding three assumptions on E_i and y_i can be expressed in terms of this model:

(1) $E(E) = 0$,

(2) $\operatorname{cov}(E) = \sigma^2 I$,

(3) the E_i is extreme value $(0, \sigma^2)$ for $i = 1, 2, \ldots, n$.

2.2. (Largest) Extreme Value Distribution. The PDF and CDF of the extreme value distribution are given by

$$\begin{aligned} f(y \mid x\beta, \sigma) &= \frac{\pi}{\sigma\sqrt{6}} \\ &\quad \times \exp\left(-\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}} \right. \\ &\quad \left. - \exp\left(\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\ &\quad - \infty \langle x \langle \infty, \ \sigma \rangle 0, \\ P(y \le c) &= \exp\left(-\exp\left(-\frac{c - x\beta + k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\ &\quad - \infty \langle x \langle \infty, \sigma \rangle, \ k \approx 0.45. \end{aligned} \tag{6}$$

It is easy to check that

$$\begin{aligned} \omega_j = \frac{\ln \pi_1}{\ln \pi_2} &= \frac{\ln(p(y \le c \mid x))}{\ln\left(p\left(y \le c \mid x_{(-1,j)}\right)\right)} \\ &= \frac{-\exp(-((c - x'\beta + k\sigma)/\sigma) \times \pi/\sqrt{6})}{-\exp\left(-\left(\left(c - x'_{(-1,j)}\beta + k\sigma\right)/\sigma\right) \times \pi/\sqrt{6}\right)} \\ &= \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) \Longrightarrow \pi_1 = \pi_2^{\exp((\pi/\sqrt{6}) \cdot (\beta_j/\sigma))}, \end{aligned} \tag{7}$$

where

$$\begin{aligned} x &= \left(1, x_1, \ldots, x_j, \ldots, x_{p-1}\right), \\ x_{(-1,j)} &= \left(1, x_1, \ldots, x_j - 1, \ldots, x_{p-1}\right), \\ \beta &= \left(\beta_0, \ \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}\right)'. \end{aligned} \tag{8}$$

To return to a random sample of observations $(y_1, y_2, \ldots, y_n)$, we conclude that the PDF and CDF of each independent y_i are given by (6), and the corresponding equality (7) is given by

$$\frac{\ln \hat{\pi}_1}{\ln \hat{\pi}_2} = \exp\left(\frac{\pi}{\hat{\sigma}\sqrt{6}}\hat{\beta}_j\right), \tag{9}$$

where the estimate $\hat{\beta}_j$ is the $(j+1)$th element of vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_{p-1})'$. It is readily shown that the results also hold true for the smallest extreme value distribution (Appendix A).

2.3. The Proposed Confidence Intervals. Let

$$\begin{aligned}\hat{\beta} &= \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_{p-1}\right)' \\ &= (X'X)^{-1}X'Y \quad j = 0, \ldots, p-1, \\ \hat{\sigma}^2 &= \frac{Y'\left(I_n - X(X'X)^{-1}X'\right)Y}{(n-p)}.\end{aligned} \tag{10}$$

According to the preceding three assumptions on E_i and y_i, we obtain

$$\begin{aligned}E\left(\hat{\beta}\right) &= E\left[(X'X)^{-1}X'Y\right] \\ &= (X'X)^{-1}X'EY = (X'X)^{-1}X'X\beta = \beta, \\ E(\hat{\sigma}^2) &= \frac{1}{n-p}E\left(Y'\left(I_n - X(X'X)^{-1}X'\right)Y\right) \\ &= \frac{1}{n-p}\left\{\operatorname{tr}\left[\left(I_n - X(X'X)^{-1}X'\right)\sigma^2 I\right]\right. \\ &\qquad \left. + E(Y')\left[I_n - X(X'X)^{-1}X'\right]E(Y)\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2 \operatorname{tr}\left[I_n - X(X'X)^{-1}X'\right]\right. \\ &\qquad \left. + \beta'X'\left[I_n - X(X'X)^{-1}X'\right]X\beta\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2\left[n - \operatorname{tr}\left(X(X'X)^{-1}X'\right)\right]\right. \\ &\qquad \left. + \beta'X'X\beta - \beta'X'X(X'X)^{-1}X'X\beta\right\} \\ &= \frac{1}{n-p}\left\{\sigma^2\left[n - \operatorname{tr}\left(X(X'X)^{-1}X'\right)\right]\right. \\ &\qquad \left. + \beta'X'X\beta - \beta'X'X\beta\right\} \\ &= \frac{1}{n-p}\sigma^2[n - \operatorname{tr}(I_P)] = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2.\end{aligned} \tag{11}$$

Therefore, $\hat{\beta}$ and $\hat{\sigma}^2$ are unbiased estimators of β and σ^2.

We have assumed that E_i is distributed as an extreme value, and we use the approximation of the extreme value distribution of the errors E_i by the normal distribution. For normally distributed observations, $\hat{\beta}_j/(\hat{\sigma}\sqrt{\delta_j})$ follows a noncentral t distribution with $n-p$ degree of freedom and noncentrality parameter $-\infty < \beta_j/(\sigma\sqrt{\delta_j}) < \infty$,

$$\begin{aligned}1 - \alpha = P\left\{t_{1-(\alpha/2)}\left[n-p, \frac{\beta_j}{\left(\sigma\sqrt{\delta_j}\right)}\right]\right. \\ \left. < \frac{\hat{\beta}_j}{\left(\hat{\sigma}\sqrt{\delta_j}\right)} < t_{\alpha/2}\left[n-p, \frac{\beta_j}{\left(\sigma\sqrt{\delta_j}\right)}\right]\right\},\end{aligned} \tag{12}$$

where $t_{\alpha/2}[r, s]$ represents the $100(1-(\alpha/2))$ percentile point of a noncentral t distribution with r degrees of freedom and noncentrality parameter $-\infty < s < \infty$, and δ_j is the $(j+1)$st diagonal element of $(X'X)^{-1}$. We use the approximation of the percentiles of the noncentral t distribution by the standard normal percentiles [13], then

$$\begin{aligned}1 - \alpha = P\left\{\frac{\beta_j/\left(\sigma\sqrt{\delta_j}\right) - z_{\alpha/2}\left[1 + \left(\beta_j^2/\left(\sigma^2\delta_j\right) - z_{\alpha/2}^2\right)/2(n-p)\right]^{1/2}}{1 - \left(z_{\alpha/2}^2/2(n-p)\right)}\right. < \\ \left. \frac{\hat{\beta}_j}{\left(\hat{\sigma}\sqrt{\delta_j}\right)} < \frac{\beta_j/\left(\sigma\sqrt{\delta_j}\right) + z_{\alpha/2}\left[1 + \left(\left(\beta_j^2/\left(\sigma^2\delta_j\right) - z_{\alpha/2}^2\right)/2(n-p)\right)\right]^{1/2}}{1 - \left(z_{\alpha/2}^2/2(n-p)\right)}\right\},\end{aligned} \tag{13}$$

$$\left(\frac{\beta_j}{\sigma}\right)^U = \left\{\frac{\hat{\beta}_j}{\hat{\sigma}}\left[1 - \frac{z_{\alpha/2}^2}{2(n-p)}\right] + z_{\alpha/2}\left[\delta_j\left(1 + \left(\frac{\left(\hat{\beta}_j^2/\hat{\sigma}^2\delta_j\right) - z_{\alpha/2}^2}{2(n-p)}\right)\right)\right]^{1/2}\right\},$$

$$\left(\frac{\beta_j}{\sigma}\right)^L = \left\{\frac{\hat{\beta}_j}{\hat{\sigma}}\left[1 - \frac{z_{\alpha/2}^2}{2(n-p)}\right] - z_{\alpha/2}\left[\delta_j\left(1 + \left(\frac{\left(\hat{\beta}_j^2/\hat{\sigma}^2\delta_j\right) - z_{\alpha/2}^2}{2(n-p)}\right)\right)\right]^{1/2}\right\},$$

Thus, we obtain an approximate $100(1-\alpha)$ percent confidence interval for ω_j

$$\left\{\exp\left[\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^L\right], \exp\left[\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^U\right]\right\}. \tag{14}$$

3. Comparison of the Two Methods

Let Y_i be a continuous outcome variable. For fixed value of C, we define Y_i^* such that

$$Y_i^* = \begin{cases}1 & \text{if } Y_i \geq C, \\ 0 & \text{if } Y_i < C.\end{cases} \tag{15}$$

Suppose that $Y_1^*, \ldots, Y_n^*$ form a random sample of observations, and we fit a complementary log-log model

$$\begin{aligned} \pi_{i1} &= P(Y_i^* = 1 \mid x_i) = \exp(-\exp(x_i\theta)), \\ \pi_{i2} &= P(Y_i^* = 1 \mid x_{(-1,i)}) = \exp(-\exp(x_{(-1,i)}\theta)), \end{aligned} \tag{16}$$

where $x_i = (1, x_{i1}, \ldots, x_{i,p-1})'$ is the $P \times 1$ vector of covariates for the ith observation, and $\theta = (\theta_0, \ldots, \theta_{p-1})'$ is the $P \times 1$ vector of unknown parameters. The dichotomized ω_j^* parameter corresponding to the effect θ_j is

$$\begin{aligned} \omega_j^* &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \\ &= \frac{\ln(P(Y^* = 1 \mid x))}{\ln\left(P\left(Y^* = 1 \mid x_{(-1,j)}\right)\right)} \\ &= \frac{(\exp(x\theta))}{\left(\exp\left(x_{(-1,j)}\theta\right)\right)} \\ &= \exp\left(\theta_j\right) \quad j = 0, \ldots, p-1. \end{aligned} \tag{17}$$

In general, maximum likelihood estimation (MLE) can be used to estimate the parameter $\theta = (\theta_0, \ldots, \theta_{p-1})$. Let $\hat{\theta} = (\hat{\theta}_0, \ldots, \hat{\theta}_{p-1})'$ be the $P \times 1$ ML estimate of θ, and let $\mathrm{COV}(\hat{\theta})$ be the $P \times P$ covariance matrix of $\hat{\theta}$. Using $\mathrm{COV}(\hat{\theta})$ from (23), one can construct confidence intervals. This matrix has as its diagonal the estimated variances of each of the ML estimates. The $(j+1)$th diagonal element is given by $\sigma^2_{\hat{\theta}_j}$. Therefore,

$$\hat{\omega}_j^* = \exp\left(\hat{\theta}_j\right), \tag{18}$$

and for large samples, $(\hat{\theta}_j^L, \hat{\theta}_j^U) = (\hat{\theta}_j - z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_j}, \hat{\theta}_j + z_{\alpha/2}\hat{\sigma}_{\hat{\theta}_j})$ is a $100(1-\alpha)$ percent confidence interval for the true θ_j. Then $(\exp(\hat{\theta}_j^L), \exp(\hat{\theta}_j^U))$ is a $100(1-\alpha)$ percent confidence interval for the true ω_j^*.

We now compare the ω_j from (7) with the ω_j^* from (17)

$$\begin{aligned} \omega_j &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \\ \omega_j^* &= \frac{\ln(\pi_1)}{\ln(\pi_2)} \end{aligned} \Longrightarrow \omega_j^* = \omega_j$$

$$\Longrightarrow \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) = \exp\left(\theta_j\right)$$

$$\Longrightarrow \frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma} = \theta_j \quad \forall \beta_j, \theta_j, \sigma. \tag{19}$$

This show that the coefficient of the complementary log-log model, θ_j, can be interpreted in terms of the regression coefficients, β_j. Note that β are related to the responses through the general linear regression model

$$y_i = x_i\beta + E_i \quad i = 1, \ldots, n, \tag{20}$$

where the independent E_i are distributed as an extreme value with mean 0 and variance $\sigma^2 > 0$.

4. Covariance Matrix of Model Parameter Estimators

4.1. Derivation of $\mathrm{var}(\omega_j^*)$ for Large n. The information matrix of generalized linear models has the form $\int = X'WX$ [1], where W is the diagonal matrix with diagonal elements $w_i = (\partial\mu_i/\partial\eta_i)^2/(\mathrm{var}(y_i))$, y is response variable with independent observations $(y_1, \ldots y_n)$, and x_{ij} denote the value of predictor j,

$$\mu_i = E(y_i), \qquad \eta_i = g(\mu_i) = \sum_j \theta_j x_{ij}, \quad j = 0, 1, \ldots, p-1. \tag{21}$$

The covariance matrix of $\hat{\theta}$ is estimated by $(X'\widehat{W}X)^{-1}$.

Maximum likelihood estimation for the complementary log-log model is a special case of the generalized linear models. Let

$$\begin{aligned} \mu_i = \pi_i &= \exp\left(-\exp\left(\sum_j \theta_j x_{ij}\right)\right) \\ &\Longrightarrow \pi_i = \exp(-\exp(\eta_i)), \\ \frac{\partial\mu_i}{\partial\eta_i} &= (-\exp(\eta_i))' \exp(-\exp(\eta_i)) = \pi_i \ln\pi_i, \\ w_i &= \frac{(\pi_I \ln\pi_i)^2}{\pi_i(1-\pi_i)} = \frac{\pi_i(\ln\pi_i)^2}{1-\pi_i}, \end{aligned} \tag{22}$$

then

$$X'WX = \begin{bmatrix} \sum_{i=1}^n \frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{i1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \\ \sum_{i=1}^n x_{i1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{i1}^2\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_1 x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \\ \vdots & & \vdots & \\ \sum_{i=1}^n x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \sum_{i=1}^n x_{1i}x_{i,p-1}\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} & \cdots & \sum_{i=1}^n x_{i,p-1}^2\frac{\pi_i(\ln\pi_i)^2}{1-\pi_i} \end{bmatrix}. \tag{23}$$

It is readily shown that the results hold true for the largest extreme value distribution (Appendix A).

In large samples, $\operatorname{var}(\hat{\theta}_j)$ approaches $\sigma^2_{\theta_j}|_{\theta=\hat{\theta}}$ [14] which equals the $(j+1)$th diagonal element of $(X'WX)^{-1}$.

By applying the delta method, let $f(\hat{\theta}_j) = \exp(\hat{\theta}_j)$, then

$$\begin{aligned}\operatorname{var}\left(\hat{\omega}_j^*\right) \longrightarrow \operatorname{var}\left(\exp\left(\hat{\theta}_j\right)\right) &= \operatorname{var}\left(f\left(\hat{\theta}_j\right)\right)\\ &= \left(\frac{\partial f\left(\hat{\theta}_j\right)}{\partial \hat{\theta}_j}\Big|_{\hat{\theta}_j=\theta_j}\right)^2 \left(\operatorname{var}\left(\hat{\theta}_j\right)\right)\\ &= \left(\exp\left(\theta_j\right)\right)^2 \times \sigma^2_{\hat{\theta}_j}.\end{aligned} \tag{24}$$

4.2. Derivation of $var(\hat{\omega}_j)$ for Large n. In large samples, from (10) $\hat{\sigma}^2 \to \sigma^2$ [15]. Therefore,

$$\operatorname{var}\left(\hat{\omega}_j\right) = \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\hat{\sigma}\sqrt{6}}\right)\right) \longrightarrow \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\sigma\sqrt{6}}\right)\right). \tag{25}$$

In addition, $\operatorname{var}(\hat{\beta}_j) = \sigma^2\delta_j$.

By applying the delta method, let $g(\hat{\beta}_j) = \exp(\pi\hat{\beta}_j/(\sigma\sqrt{6}))$, then

$$\begin{aligned}\operatorname{var}\left(\hat{\omega}_j\right) &\longrightarrow \operatorname{var}\left(\exp\left(\frac{\pi\hat{\beta}_j}{\sigma\sqrt{6}}\right)\right)\\ &= \operatorname{var}\left(g\left(\hat{\beta}_j\right)\right)\\ &= \left(\frac{\partial g\left(\hat{\beta}_j\right)}{\partial \hat{\beta}_j}\Big|_{\hat{\beta}_j=\beta_j}\right)^2 \times \operatorname{var}\left(\hat{\beta}_j\right)\\ &= \left(\frac{\pi}{\sigma\sqrt{6}}\exp\left(\frac{\pi\beta_j}{\sigma\sqrt{6}}\right)\right)^2 \sigma^2\delta_j\\ &= \frac{\pi^2}{\sqrt{6}}\delta_j\left(\exp\frac{\pi\beta_j}{\sigma\sqrt{6}}\right)^2.\end{aligned} \tag{26}$$

5. Sample Sizes Saving

5.1. The Power for the Dichotomized Method. In large samples, $\hat{\sigma}_{\hat{\theta}_j}$ converges to $\sigma_{\hat{\theta}_j}$ almost surely [14]. Therefore, for a given value of $\omega_j = \exp\theta_j$ (i.e., $\ln\omega_j = \theta_j$), the power is given by

$$\begin{aligned}p\left(\omega_j\right) &= p\left\{\text{rejection of } \omega_j = 1 \mid \omega_j \neq 1\right\}\\ &= p\left\{\exp\left(\theta_j^L\right) > 1 \mid \theta_j\right\} + p\left\{\exp\left(\theta_j^U\right) < 1 \mid \theta_j\right\}\\ &= p\left\{\hat{\theta}_j > z_{\alpha/2}\sigma_{\hat{\theta}_j} \mid \theta_j\right\} + p\left\{\hat{\theta}_j < -z_{\alpha/2}\sigma_{\hat{\theta}_j} \mid \theta_j\right\}\\ &= p\left\{Z > \frac{z_{\alpha/2}\sigma_{\hat{\theta}_j} - \ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &\quad + p\left\{Z < \frac{-z_{\alpha/2}\sigma_{\hat{\theta}_j} - \ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &= p\left\{Z > z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\} + p\left\{Z < -z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\right\}\\ &= P\{Z > z_1^\bullet\} + P\{Z < -z_2^\bullet\},\end{aligned} \tag{27}$$

where

$$\left\{\begin{aligned}z_1^\bullet &= z_{\alpha/2} - \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\\ z_2^\bullet &= z_{\alpha/2} + \frac{\ln\omega_j}{\sigma_{\hat{\theta}_j}}\end{aligned}\right\}. \tag{28}$$

5.2. The Power for the Proposed Method. In large samples, $\hat{\sigma}$ converges to σ almost surely [15]. Therefore, for a given value of $\omega_j = \exp(\pi\beta_j/\sigma\sqrt{6})$ (i.e., $\beta_j = \sigma(\ln\omega_j\sqrt{6}/\pi)$), the power is given by

$$\begin{aligned}p\left(\omega_j\right) &= p\left\{\exp\left(\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^L\right) > 1 \mid \omega_j\right\}\\ &\quad + p\left\{\exp\left(\frac{\pi}{\sqrt{6}}\left(\frac{\beta_j}{\sigma}\right)^U\right) < 1 \mid \omega_j\right\}\\ &= P\left\{\beta_J^L > z_{\alpha/2}\sigma\sqrt{\delta_j} \mid \beta_j = \frac{\sigma\ln\omega_j\sqrt{6}}{\pi}\right\}\\ &\quad + P\left\{\beta_J^U < -z_{\alpha/2}\sigma\sqrt{\delta_j} \mid \beta_j = \frac{\sigma\ln\omega_j\sqrt{6}}{\pi}\right\}\\ &= p\left\{Z > \frac{z_{\alpha/2}\sigma\sqrt{\delta_j} - \left(\sigma\ln\omega_j\sqrt{6}/\pi\right)}{\sigma\sqrt{\delta_j}}\right\}\\ &\quad + p\left\{Z < \frac{-z_{\alpha/2}\sigma\sqrt{\delta_j} - \left(\sigma\ln\omega_j\sqrt{6}/\pi\right)}{\sigma\sqrt{\delta_j}}\right\}\end{aligned}$$

$$= p\left\{Z > z_{\alpha/2} - \frac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}}\right\} + p\left\{Z < -z_{\alpha/2} - \frac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}}\right\} = p\{Z > z_1\} + p\{Z < -z_2\}, \tag{29}$$

where

$$\left\{\begin{array}{l} z_1 = z_{\alpha/2} - \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \\ z_2 = z_{\alpha/2} + \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \end{array}\right\}. \tag{30}$$

Our proposed method, since it is based on continuous data rather than dichotomized, is likely to be more powerful. We show that the proposed method can produce substantial sample size saving for a given power. Let

(i) the number of parameters $p = 2$ (i.e., $\theta = (\theta_0, \theta_1)$),

(ii) $x_i = (1, x_{i1})', x_{i1} \in \{-a+(2an/(g-1)) \mid n = 0, \ldots, g-1\}$, that is, x_{i1} follows a discrete uniform distribution with range $(-a,\ a)$. For simplicity, $a = 2$.

(iii) Total samples are n and n^* for the proposed and dichotomized methods, respectively. These samples included k and k^* set of these g uniformly distributed points for the proposed and dichotomized methods, respectively. That is, $n = gk$ and $n^* = gk^*$, then

$$\delta_j = \left[k\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2\right]^{-1}, \quad j = 1, \tag{31}$$

and from (23),

$$\sigma^2_{\hat{\theta}_j} = \frac{\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(k^*)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}. \tag{32}$$

We consider the same power for two methods:

$$\begin{array}{l} z_1 = z_1^* \\ z_2 = z_2^* \end{array} \Longrightarrow \left\{\begin{array}{l} z_{\alpha/2} - \dfrac{\ln \omega_j}{\sigma_{\hat{\theta}_j}} = z_{\alpha/2} - \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \\ z_{\alpha/2} + \dfrac{\ln \omega_j}{\sigma_{\hat{\theta}_j}} = z_{\alpha/2} + \dfrac{\ln \omega_j \sqrt{6}}{\pi\sqrt{\delta_j}} \end{array}\right. \Longrightarrow \frac{\pi}{\sqrt{6}}\sqrt{\delta_j} = \sigma_{\hat{\theta}_j}, \quad j = 1 \Longrightarrow \frac{\pi}{\sqrt{6}}\sqrt{\left[k\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2\right]^{-1}} \tag{33}$$

$$= \sqrt{\frac{\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(k^*)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}}$$

relative sample size

$$\frac{n^*}{n} = \frac{k^*}{k} = \frac{6\sigma^2_{\hat{\theta}_j}}{\pi^2\delta_j} = \frac{\sum_{i=1}^{g}(x_{1i} - \bar{x}_{1.})^2 \times \sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)}{(\pi^2/6)\left\{\sum_{i=1}^{g} x_{1i}^2\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\sum_{i=1}^{g}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right) - \left[\sum_{i=1}^{g} x_{1i}\left((\pi_i)(\ln(\pi_i))^2/\ln(1-\pi_i)\right)\right]^2\right\}}. \tag{34}$$

Table 1: Relative sample sizes required to attain any power for the dichotomizing method versus the proposed method.

$\omega^* = \exp(\theta)$	Average proportion of successes ($\overline{\pi}$)				
	0.1	0.2	0.3	0.4	0.5
0.25	23.7166	9.5092	7.4954	7.1996	6.8575
0.50	10.6719	5.4176	3.4215	2.5209	2.1784
0.75	7.7088	3.8713	2.5171	1.9380	1.5841

That is, (34) is independent of σ^2 and applies for any power, and any test size α.

Table 1 presents relative sample sizes n^*/n for a given fixed parameter ω_j^* and an average proportion of success $\overline{\pi}$. We consider the situations in which $\overline{\pi} = \sum_{i=1}^{g}(\pi_i/g) = 0.1, 0.2, 0.3, 0.4, 0.5$, $g = 9$, $\omega_j^* = 0.25, 0.50, 0.75$.

For given fixed ω_j^* and $\overline{\pi}$, the relative sample sizes in Table 1 can be computed by the following step:

(i) compute the value θ_j via the equation $\theta_j = \ln(\omega_j^*)$,

(ii) calculate the cut-off point C iteratively such that $\overline{\pi}$ attained the specified value for the values x_{i1}, using the value of θ_j in (i).

As can be seen from Table 1, all values are greater than 1. The values of n^*/n increase as the ω_j^* moves farther away from 1. Values of Table 1 immediately highlight the improvement accomplished by the proposed method.

6. Relative Efficiency of $\widehat{\omega}_j$ with $\widehat{\omega}_j^*$

Here, we examine the relative efficiency of the estimate $\widehat{\omega}_j$ to the estimate $\widehat{\omega}_j^*$.

Using (24) and (26), the relative efficiency is given by

$$\begin{aligned}\text{r.e.}\left(\widehat{\omega}_j, \widehat{\omega}_j^*\right) &= \frac{\text{var}\left(\widehat{\omega}_j^*\right)}{\text{var}\left(\widehat{\omega}_j\right)} \\ &= \frac{6\left(\exp\left(\theta_j\right)\right)^2 \times \sigma_{\widehat{\theta}_j}^2}{\pi^2\delta_j\left(\exp\left(\lambda\beta_j/\sigma\right)\right)^2} = \frac{6\sigma_{\widehat{\theta}_j}^2}{\pi^2\delta_j}.\end{aligned} \tag{35}$$

Note that the relative efficiency is independent of n and σ^2 and converges to a constant. Comparing (34) and (35), the relative efficiency equals the relative sample sizes. Therefore, as in Table 1, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies.

It should be noted that these results hold true under the following assumptions:

(1) the responses y_i and β are related through the equation $y_i = x_i\beta + E_i$ where the independent E_i are distributed as an extreme value with mean 0 and variance $\sigma^2 > 0$,

(2) the independent variables x_i follow a discrete uniform distribution.

7. Odds Ratio

For values of π larger than 0.90, $-\ln(\pi)$ and $\pi/(1-\pi)$ are very close. Hence, for large values of π,

$$\frac{\ln(\pi_1)}{\ln(\pi_2)} \cong \frac{\pi_1/1-\pi_1}{\pi_2/1-\pi_2} = \text{OR}. \tag{36}$$

And from (7), odds ratio is given by

$$\text{OR} = \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right). \tag{37}$$

The parameters estimated from the linear regression can be interpreted as an odds ratio.

8. Simulation Study

It should be noted that, as in Table 1, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies. These results hold true under the assumption that predictor variable has a discrete uniform distribution and that the random variables E_i follow an extreme value distribution. To demonstrate the robustness of this conclusion to changes in the distributions of predictor variables, simulations were run under different distributional conditions. The data were sampled 10000 times for three sample sizes $\{n = 250, 500, 1000\}$, three average proportions of successes $\{\overline{\pi} = 0.10, 0.50, 0.95\}$, and seven $\omega_j\{\omega_j = 0.75, 0.90, 1.1, 1.2, 1.3, 1.4, 1.5\}$. The simulated data are generated using the following algorithm

(1) Generate y_i, where $y_i = \beta_0 + \beta_1 x_i + E_i$, $\beta_1 = \sqrt{6}\ln\omega_j/\pi$ through (7) to produce the correct ω_j, and for simplicity $\beta_0 = 0$, $\sigma^2 = 1$.

(2) For fixed $\overline{\pi}$, generate cutoff point C using (15).

We simulated the data for two scenarios based on the distribution of the explanatory variable. In the first scenario, the independent variable follows a continuous uniform distribution and range $(-2, 2)$, and in the second, the independent variable follows a truncated normal distribution with mean 0 and range $(-2, 2)$. The relative mean square errors, relative interval lengths, absolute biases, and the probability of coverage were calculated.

Results of the simulations addressing the validity of the proposed method are displayed in Tables 2 and 3.

The simulations show that the relative mean square errors are all greater than 1, increasing with the average proportion of successes and when the ω_j moves farther away

Table 2: Simulated relative mean square errors, relative intervals lengths, coverage probabilities, and absolute biases for the proposed and dichotomizing methods (using a continuous uniform distribution for the explanatory variable and an extreme value distribution for the errors).

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.15[a]	1.07	1.09	1.14	1.24	1.47	1.71
		1.10[b]	1.03	1.03	1.07	1.14	1.23	1.35
	0.10	0.943[c]	0.948	0.949	0.949	0.945	0.938	0.933
		0.948[d]	0.947	0.949	0.947	0.951	0.947	0.953
		0.05[e]	0.04	0.12	0.14	0.10	0.15	0.11
		0.07[f]	0.01	0.17	0.13	0.24	0.34	0.58
		1.23	1.26	1.27	1.28	1.27	1.24	1.26
		2.16	1.13	1.23	1.14	1.15	1.17	1.19
1000	0.50	0.940	0.951	0.951	0.945	0.942	0.937	0.934
		0.951	0.949	0.951	0.950	0.948	0.947	0.948
		0.04	0.01	0.08	0.10	0.05	0.09	0.04
		0.05	0.04	0.15	0.12	0.09	0.12	0.13
		12.75	12.44	13.22	12.68	13.14	12.91	12.79
		3.67	3.57	3.58	3.63	3.69	3.76	3.84
	0.95	0.943	0.951	0.952	0.944	0.944	0.938	0.929
		0.952	0.954	0.952	0.952	0.951	0.951	0.951
		0.04	0.07	0.11	0.10	0.10	0.17	0.10
		0.75	0.68	0.86	1.01	1.21	1.45	1.24
		1.30	1.08	1.07	1.17	1.24	1.54	1.95
		1.16	1.03	1.04	1.08	1.15	1.25	1.39
	0.10	0.942	0.950	0.951	0.95	0.944	0.941	0.936
		0.951	0.950	0.949	0.951	0.954	0.954	0.953
		0.12	0.07	0.24	0.25	0.21	0.18	0.29
		0.23	0.08	0.33	0.39	0.41	0.73	1.21
		1.35	1.10	1.27	1.26	1.26	1.25	1.26
		1.26	1.03	1.13	1.14	1.16	1.17	1.20
500	0.50	0.940	0.949	0.947	0.948	0.943	0.940	0.933
		0.952	0.951	0.949	0.949	0.954	0.950	0.951
		0.23	0.34	0.27	0.23	0.26	0.25	0.38
		0.48	0.11	0.17	0.18	0.31	0.26	0.42
		13.04	13.17	13.8	13.90	14.45	14.48	14.47
		3.72	3.65	3.68	3.73	3.82	3.91	3.99
	0.95	0.942	0.947	0.951	0.949	0.947	0.938	0.935
		0.953	0.952	0.954	0.955	0.955	0.953	0.954
		0.05	0.11	0.08	0.08	0.24	0.32	0.27
		0.94	1.38	1.78	1.92	2.52	3.00	2.90
		13.41	14.46	1.12	1.28	1.52	1.96	2.33
		3.78	3.73	1.04	1.09	1.18	1.30	1.45
	0.10	0.942	0.949	0.949	0.945	0.942	0.942	0.933
		0.957	0.954	0.948	0.949	0.952	0.957	0.953
		0.02	0.20	0.38	0.33	0.42	0.41	0.66
		2.11	2.74	0.42	0.84	1.18	1.78	2.24

Table 2: Continued.

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.27	1.25	1.32	1.28	1.30	1.30	1.29
		1.16	1.13	1.13	1.14	1.16	1.18	1.20
250	0.50	0.941	0.948	0.952	0.947	0.945	0.943	0.933
		0.951	0.951	0.951	0.950	0.951	0.951	0.951
		0.12	0.13	0.35	0.44	0.41	0.53	0.55
		0.11	0.22	0.39	0.47	0.51	0.74	0.59
		12.98	14.6	15.64	15.46	17.05	16.89	18.33
		3.75	3.72	3.82	3.88	4.01	4.12	4.29
	0.95	0.945	0.955	0.946	0.948	0.940	0.937	0.932
		0.959	0.955	0.955	0.959	0.958	0.957	0.952
		0.02	0.16	0.39	0.22	0.46	0.47	0.51
		1.22	2.75	3.97	3.98	4.99	5.19	6.19

a: Relative mean square errors, b: Relative intervals lengths, c: Coverage probability (proposed), d: Coverage probability (dichotomized), e: % bias (proposed), f: % bias (dichotomized).

from 1. The results in Tables 1 and 2 demonstrate that the proposed method provides confidence intervals which successfully maintain their nominal 95 percent coverage. For the proposed method in first scenario, 51 out of 63 coverage probabilities fell within (0.94, 0.96), and all 63 coverage probabilities are greater than 0.93 and, in the second scenario, almost all coverage probabilities fell within (0.94, 0.96). The absolute biases for proposed method are never greater than a few percent. The proposed method is less biased than the dichotomizing method in 6 of 63 simulations in both two scenarios.

9. An Example

To illustrate the application of the proposed method presented in the previous section, we utilize the data arising from the National Health Survey in Iran. The other analyses using this data appear in many places [16].

In this study, 14176 women aged 20–69 years were investigated. BMI (body mass index), our dependent variable, was calculated as weight in kilograms divided by height in meters squared (kg/m^2). Independent variables included place of residence, age, smoking, economic index, marital status, and education level. The independent variables considered were both categorical and continuous. At first, BMI was treated as a continuous variable, and $\widehat{\omega}_j$ and 95 percent confidence intervals were calculated using the proposed linear regression method. Then subjects were classified into obese (BMI $\geq$ 30 kg/m^2) and nonobese (BMI <30 kg/m^2). A complementary log-log model was used for the binary analysis, with obese or nonobese used as the outcome measure. The $\widehat{\omega}_j^*$ and 95 percent confidence intervals were calculated using the dichotomized method. Table 4 presents the coefficient estimates, estimated confidence intervals, and relative confidence interval lengths. The proposed and dichotomizing methods produced different confidence intervals, although the $\widehat{\omega}_j$ and $\widehat{\omega}_j^*$ were similar only varying slightly. The $\widehat{\omega}_j$ estimate from the proposed method had smaller variances and shorter confidence intervals than the dichotomizing method. All relative confidence interval lengths were greater than 2.58.

10. Discussion

When assuming the errors E_i are distributed as an extreme value distribution, as noted before, the method has several advantages. First, the method allows the researcher to apply the complementary log-log model without dichotomizing and without loss of information. Second, the $\widehat{\omega}_j^*$ from the dichotomizing method is dependent on the chosen cutoff point C and will vary with c. However, the proposed $\widehat{\omega}_j$ is independent of the c since $\widehat{\omega}_j$ is a function of the continuous Y_i and not a function of the dichotomized Y_i^* defined through C. Third, we show that the coefficient of the complementary log-log model, θ_j, can be interpreted in terms of the regression coefficients, β_j. Fourth, when the independent variables x_i follow a discrete uniform distribution, the proposed method is a consistent improvement over the dichotomizing method with respect to relative efficiencies. The proposed method can provide sample size saving, smaller variances, and shorter confidence intervals than the dichotomized method. Fifth, when π is large, the parameters estimated from the linear regression can be interpreted as odds ratios.

Our results were consistent with the findings by Moser and Coombs [12] and Bakhshi et al. [16] showing the greater efficiency of parameter estimates from the regression method that avoids dichotomizing in comparison with a more traditional dichotomizing method using the logistic regression.

Our main recommendation is to let continuous response remain continuous. Do not throw away information by transforming the data to binary. This means that if the objective is to estimate and/or test coefficients when responses

Table 3: Simulated relative mean square errors, relative intervals lengths, coverage probabilities, and absolute biases for the proposed and dichotomizing methods (using a truncated normal distribution for the explanatory variable and an extreme value distribution for the errors).

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.17[a]	1.02	1.08	1.13	1.19	1.28	1.36
		1.11[b]	1.03	1.03	1.06	1.10	1.25	1.22
	0.10	0.942[c]	0.948	0.948	0.952	0.944	0.942	0.940
		0.951[d]	0.951	0.950	0.952	0.949	0.951	0.951
		0.08[e]	0.06	0.03	0.14	0.13	0.14	0.16
		0.10[f]	0.11	0.15	0.23	0.30	0.39	0.39
		1.26	1.24	1.26	1.28	1.28	1.25	1.28
		1.24	1.13	1.13	1.14	1.14	1.15	1.17
1000	0.50	0.944	0.948	0.952	0.947	0.947	0.944	0.941
		0.948	0.951	0.949	0.949	0.947	0.950	0.949
		0.02	0.09	0.08	0.07	0.18	0.16	0.13
		0.03	0.06	0.12	0.16	0.20	0.16	0.14
		12.33	13.12	13.03	12.71	12.86	12.55	12.88
		3.62	3.59	3.61	3.62	3.64	3.68	3.71
	0.95	0.944	0.951	0.948	0.948	0.945	0.945	0.946
		0.952	0.948	0.95	0.949	0.949	0.951	0.952
		0.10	0.04	0.11	0.04	0.16	0.16	0.20
		1.26	1.05	1.56	1.36	1.43	1.80	1.94
		1.18	1.09	1.06	1.75	1.23	1.32	1.58
		1.11	1.03	1.03	1.06	1.11	1.16	1.23
	0.10	0.945	0.95	0.951	0.951	0.949	0.943	0.944
		0.953	0.953	0.953	0.950	0.949	0.951	0.950
		0.04	0.13	0.31	0.18	0.33	0.36	0.37
		0.21	0.08	0.37	0.50	0.62	0.69	0.96
		1.25	1.27	1.27	1.29	1.27	1.29	1.25
		1.14	1.13	1.13	1.14	1.15	1.16	1.17
500	0.50	0.944	0.948	0.949	0.947	0.948	0.944	0.935
		0.951	0.951	0.951	0.948	0.951	0.948	0.949
		0.13	0.22	0.35	0.37	0.35	0.30	0.44
		0.16	0.19	0.39	0.48	0.44	0.41	0.54
		13.11	14.02	14.02	13.5	13.54	13.80	14.32
		3.73	3.71	3.73	3.75	3.77	3.81	3.86
	0.95	0.944	0.95	0.951	0.950	0.947	0.944	0.944
		0.954	0.95	0.951	0.953	0.948	0.956	0.953
		0.15	0.10	0.24	0.38	0.32	0.33	0.43
		2.50	2.70	2.92	3.10	2.92	3.36	3.89
		1.28	1.11	1.12	1.19	1.33	1.54	1.76
		1.11	1.03	1.04	1.08	1.13	1.19	1.28
	0.10	0.947	0.951	0.950	0.947	0.950	0.950	0.942
		0.951	0.950	0.950	0.952	0.954	0.952	0.951
		0.40	0.34	0.37	0.64	0.69	0.58	0.81
		0.26	0.06	0.69	1.08	1.30	1.55	2.22

Table 3: Continued.

Sample size	ω Cut off	.75	.9	1.1	1.2	1.3	1.4	1.5
		1.32	1.30	1.27	1.33	1.31	1.33	1.31
		1.15	1.13	1.13	1.14	1.18	1.17	1.18
250	0.50	0.951	0.95	0.953	0.951	0.940	0.945	0.940
		0.949	0.951	0.952	0.948	0.948	0.950	0.948
		0.22	0.43	0.57	0.69	0.66	0.58	0.66
		0.38	0.53	0.64	0.89	0.91	0.82	0.91
		14.09	14.51	16.27	15.91	15.89	15.73	15.60
		3.86	3.87	3.93	3.92	3.98	4.04	4.11
	0.95	0.943	0.95	0.951	0.951	0.947	0.944	0.937
		0.953	0.95	0.953	0.956	0.953	0.956	0.952
		0.30	0.37	0.57	0.68	0.42	0.62	0.75
		4.98	5.52	6.547	5.91	6.17	6.88	7.72

a: Relative mean square errors, b: Relative intervals lengths, c: Coverage probability (proposed), d: Coverage probability (dichotomized), e: % bias (proposed), f: % bias (dichotomized).

Table 4: Adjusted $\widehat{\omega}_j^*$, $\widehat{\omega}_j$ for obesity and confidence intervals using two methods for the National Health Survey.

Covariates	$\widehat{\omega}_j(\widehat{\omega}_j^*)$	95% CI[a] (proposed)	95% CI (dichotomized)	Relative[b] length of CI
Place of residence	1.65 (1.97)[c]	1.58–1.74	1.79–2.18	2.43
Age	1.021 (1.019)	1.018–1.022	1.015–1.022	1.75
Years of education	0.99 (0.98)	0.985–0.997	0.971–0.994	1.92
Smoking	0.76 (0.68)	0.66–0.90	0.51–0.92	1.71
Marital status	1.16 (1.42)	1.10–1.22	1.27–1.58	2.58
Lower-middle economy index	1.24 (1.32)	1.14–1.32	1.18–1.48	1.67
Upper-middle economy index	1.21 (1.26)	1.14–1.29	1.12–1.42	2.0
High economy index	1.20 (1.21)	1.11–1.30	1.08–1.36	1.47

[a]Confidence interval, [b]dichotomized/proposed, [c]proposed (dichotomized).

are continuous, please resist dichotomizing your response variable.

Appendix

A. Largest Extreme Value Distribution

(a) *The PDF and CDF are Given by*

$$
\begin{aligned}
f(y \mid x\beta, \sigma) &= \frac{\pi}{\sigma\sqrt{6}} \\
&\quad \times \exp\left(-\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right. \\
&\quad \left. - \exp\left(\frac{y - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\
&\quad - \infty \langle x \langle \infty, \sigma \rangle 0, \\
P(y \le c) &= 1 - \exp\left(-\exp\left(-\frac{c - x\beta - k\sigma}{\sigma} \times \frac{\pi}{\sqrt{6}}\right)\right) \\
&\quad - \infty \langle x \langle \infty, \sigma \rangle 0,
\end{aligned}
\tag{A.1}
$$

where Y is a continuous outcome variable, $x = (1, x_1, \ldots, x_{p-1})$ is the $p \times 1$ vector of known independent variables, $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is the $p \times 1$ vector of unknown parameters, and $k \approx 0.45$.

It is easy to check that

$$
\begin{aligned}
\omega_j &= \frac{\ln(1 - \pi_1)}{\ln(1 - \pi_2)} = \frac{\ln(1 - p(y \le c \mid x))}{\ln\left(1 - p\left(y \le c \mid x_{(-1,j)}\right)\right)} \\
&= \frac{-\exp(-((c - x'\beta - k\sigma)/\sigma) \times (\pi/\sqrt{6}))}{-\exp\left(-\left(\left(c - x'_{(-1,j)}\beta - k\sigma\right)/\sigma\right) \times (\pi/\sqrt{6})\right)} \\
&= \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\beta_j}{\sigma}\right) \Longrightarrow 1 - \pi_1 \\
&= (1 - \pi_2)^{\exp((\pi/\sqrt{6}) \cdot (\beta_j/\sigma))},
\end{aligned}
\tag{A.2}
$$

where

$$
\begin{aligned}
x &= \left(1, x_1, \ldots, x_j, \ldots, x_{p-1}\right), \\
x_{(-1,j)} &= \left(1, x_1, \ldots, x_j - 1, \ldots, x_{p-1}\right), \\
\beta &= \left(\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_{p-1}\right)'.
\end{aligned}
\tag{A.3}
$$

(b) *Suppose that E_i is distributed as a largest extreme value with mean 0 and variance $\sigma^2 > 0$.* We conclude that the PDF and CDF of each independent Y_i are given by (A.1), and the corresponding equality (A.2) is given by

$$\widehat{\omega}_j = \frac{\ln(1-\widehat{\pi}_1)}{\ln(1-\widehat{\pi}_2)} = \exp\left(\frac{\pi}{\sqrt{6}} \cdot \frac{\widehat{\beta}_j}{\widehat{\sigma}}\right). \tag{A.4}$$

(c) *Similar to largest extreme value distribution*

$$\begin{aligned} \mu_i = \pi_i &= 1 - \exp\left(-\exp\left(\sum_j \theta_j x_{ij}\right)\right) \\ \Longrightarrow \pi_i &= 1 - \exp(-\exp(\eta_i)), \end{aligned} \tag{A.5}$$

then

$$\begin{aligned} \frac{\partial \mu_i}{\partial \eta_i} &= -(-\exp(\eta_i))' \exp(-\exp(\eta_i)) \\ &= -(1-\pi_i)\ln(1-\pi_i) \\ w_i &= \frac{((1-\pi_i)\ln(1-\pi_i))^2}{\pi_i(1-\pi_i)} \\ &= \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i}, \end{aligned}$$

$$X'WX = \begin{bmatrix} \sum_{i=1}^{n} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \\ \sum_{i=1}^{n} x_{i1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} x_{i1}^2 \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_1 x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \\ \vdots & & & \vdots \\ \sum_{i=1}^{n} x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \sum_{i=1}^{n} x_i x_{i,p-1} \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} & \cdots & \sum_{i=1}^{n} x_{i,p-1}^2 \frac{(1-\pi_i)(\ln(1-\pi_i))^2}{\pi_i} \end{bmatrix}. \tag{A.6}$$

Conflict of Interests

The authors have declared no conflict of interests.

References

[1] A. Agresti, *Categorical Data Analysis*, Wiley, New York, NY, USA, 2nd edition, 2002.

[2] L. P. Zhao and L. N. Kolonel, "Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies," *American Journal of Epidemiology*, vol. 136, no. 4, pp. 464–474, 1992.

[3] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the practice of dichotomization of quantitative variables," *Psychological Methods*, vol. 7, no. 1, pp. 19–40, 2002.

[4] J. Cohen, "The cost of dichotomization," *Applied Psychological Measurement*, vol. 7, no. 3, pp. 249–253, 1983.

[5] S. Greenland, "Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis.," *Epidemiology*, vol. 6, no. 4, pp. 450–454, 1995.

[6] P. C. Austin and L. J. Brunner, "Inflation of the type I error rate when a continuous confounding variable is categorized in logistics regression analyses," *Statistics in Medicine*, vol. 23, no. 7, pp. 1159–1178, 2004.

[7] A. Vargha, T. Rudas, H. D. Delaney, and S. E. Maxwell, "Dichotomization, partial correlation, and conditional independence," *Journal of Educational and Behavioral Statistics*, vol. 21, no. 3, pp. 264–282, 1996.

[8] S. E. Maxwell and H. D. Delaney, "Bivariate median splits and spurious statistical significance," *Psychological Bulletin*, vol. 113, no. 1, pp. 181–190, 1993.

[9] D. L. Streiner, "Breaking up is hard to do: the heartbreak of dichotomizing continuous data," *Canadian Journal of Psychiatry*, vol. 47, no. 3, pp. 262–266, 2002.

[10] H. Chen, P. Cohen, and S. Chen, "Biased odds ratios from dichotomization of age," *Statistics in Medicine*, vol. 26, no. 18, pp. 3487–3497, 2007.

[11] D. R. Ragland, "Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint," *Epidemiology*, vol. 3, no. 5, pp. 434–440, 1992.

[12] B. K. Moser and L. P. Coombs, "Odds ratios for a continuous outcome variable without dichotomizing," *Statistics in Medicine*, vol. 23, no. 12, pp. 1843–1860, 2004.

[13] N. L. Johnson, H. Welch, and C. Z. Wei, "Application of the non-central t distribution," *Biometrika*, vol. 31, no. 3-4, pp. 362–389, 1940.

[14] R. J. Serfling, *Approximation Theory of Mathematical Statistics*, Wiley, New York, NY, USA, 1980.

[15] T. L. Lai, H. Robbins, and C. Z. Wei, "Strong consistency of least squares estimates in multiple regression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 75, no. 7, pp. 3034–3036, 1978.

[16] E. Bakhshi, M. R. Eshraghian, K. Mohammad, and B. Seifi, "A comparison of two methods for estimating odds ratios: results from the National Health Survey," *BMC Medical Research Methodology*, vol. 8, article 78, 2008.