Communications for Statistical Applications and Methods
2024, Vol. 31, No. 4, 409–425

https://doi.org/10.29220/CSAM.2024.31.4.409
Print ISSN 2287-7843 / Online ISSN 2383-4757

Effect of zero imputation methods for log-transformation of independent variables in logistic regression

Seo Young Park[1,a]

[a]Department of Statistics and Data Science, Korea National Open University, Korea

Abstract

Logistic regression models are commonly used to explain binary health outcome variable using independent variables such as patient characteristics in medical science and public health research. Although there is no distributional assumption required for independent variables in logistic regression, variables with severely right-skewed distribution such as lab values are often log-transformed to achieve symmetry or approximate normality. However, lab values often have zeros due to limit of detection which makes it impossible to apply log-transformation. Therefore, preprocessing to handle zeros in the observation before log-transformation is necessary. In this study, five methods that remove zeros (shift by 1, shift by half of the smallest nonzero, shift by square root of the smallest nonzero, replace zeros with half of the smallest nonzero, replace zeros with the square root of the smallest nonzero) are investigated in logistic regression setting. To evaluate performances of these methods, we performed a simulation study based on randomly generated data from log-normal distribution and logistic regression model. Shift by 1 method has the worst performance, and overall shift by half of the smallest nonzero method, replace zeros with half of the smallest nonzero method, and replace zeros with the square root of the smallest nonzero method showed comparable and stable performances.

Keywords: log transformation, zero imputation, skewed distribution, logistic regression, limit of detection

1. Introduction

In studies in medical science or public health, the health outcome we want to predict or explain with independent variables is often binary. For example, the outcome could be whether a patient has a specific disease or not (e.g. diabetic or not), or it could be whether a patient develops a new condition within a certain period of time (e.g. recurrence of cancer within 3 years since surgery). This makes logistic regression one of the most commonly used methods among others in health research.

Logistic regression assumes that the response variable follows the Bernoulli distribution and the probability of "success" is equal to the logistic function of linear combination of independent variables. There is no distributional assumption required for the independent variables in logistic regression. However, when an independent variable has a very skewed distribution to the right, it is common to apply log transformation of such variables in medical science and public health (Ekwaru and Veugelers, 2018; Feng *et al.*, 2014). A typical example of variables with right-skewed distribution includes lab values such as Creatinine, CRP (C-Reactive Protein), or CA19-9 (Carbohydrate antigen 19-9). These variables tend to have values within a "normal" range when the patient is healthy, but

This research was supported by Korea National Open University Research Fund.

[1]Corresponding author: Department of Statistics and Data Science, Korea National Open University, Main Building 320, 86(Dongsung-Dong), Daehak-ro, Jongno-Gu, Seoul 03087, Korea. E-mail: biostat81@gmail.com

Published 31 July 2024 / journal homepage: http://csam.or.kr

can have much higher values, sometimes multiple hundred times the normal value, when the patient is in an abnormal condition. This makes the distribution severely skewed to the right. Although logistic regression does not require independent variables to have symmetric distribution, these variables with skewed distribution tend to explain the response variable better when they are transformed into a more symmetric distribution. This is because extremely high, abnormal lab values work as influential points and thus tend to ruin the overall model fit. Also, it is unlikely that having multiple hundred times the normal lab value really has hundred times the impact that a patient with a normal lab value would have on a logit of "success" probability.

Log-transformation is simply replacing a variable X with $\log(X)$. Due to its simplicity and popularity, a lot of statistical software such as SAS and SPSS includes built-in log-transformation. It is a special case of the Box-Cox transformation (Box and Cox, 1964) and tends to make rightly skewed distribution less skewed. The downside of log-transformation is that it can only be applied to positive variables because logarithm is not defined at zero or negatives.

Although lab values are usually nonnegative, often times they have zeros. This is because in many cases, due to technical restrictions, there is a limit of detection. That is, if the amount of the substance of interest goes smaller than a certain value, it is not detected and is recorded as zero. Therefore, to log-transform a lab value, it is wise to check if there are zeros, and if so, preprocessing to handle zeros is necessary before log-transformation.

The most commonly used method to handle zeros before log-transformation is the so called "started logarithm" (Rocke and Durbin-Johnson, 2003). This method is adding a small positive constant before log-transformation. That is, replacing X with $\log(X + c)$, where c is a very small, positive constant. For c, 1, or half of the minimum non-zero values, or arbitrary numbers such as 0.01 or 0.001 are often used. There have been studies on how to select optimal value for c in the literature (Ekwaru and Veugelers, 2018; Rocke and Durbin-Johnson, 2001, 2003; Durbin and Rocke, 2004). Also, there have been suggestions for new methods of zero imputation as well, but it is challenging to use such methods in practice due to their complexity (Bellégo *et al.*, 2006).

Park (2023) introduced five zero imputation methods commonly used in practice including "started logarithm" and compared their performances in linear regression setting. However, the impact of use of these methods in logistic regression remains unclear.

In this study, we investigate the effect of zero imputation methods for log-transformation of independent variables in logistic regression setting using a simulation study. The rest of this paper is organized as follows: In Chapter 2, we introduce the logistic regression setting with a right-skewed independent variable. In Chapter 3, the five most common zero imputation methods used in medical science are introduced. Chapter 4 presents the details and the results of the simulation study, and we conclude with recommendations for zero imputation methods in Chapter 5.

2. Problem setting

We are interested in the situation where the logistic regression model (2.1) explains the relationship between the binary response variable and the continuous, nonnegative independent variable as,

$$P(Y = 1) = \frac{1}{1 + \exp[-\{\beta_0 + \beta_1 \log(X)\}]}. \tag{2.1}$$

The model above reflects the common situation where X has right-skewed distribution and it explains $P(Y = 1)$ better when X is log-transformed. We assume that $\log(X)$ has normal distribution (thus X follows log-normal distribution which is skewed to the right) and we cannot observe the exact

value of X if $X < d$, where d is the limit of detection and its value is unknown. That is, we observe the realization of $\tilde{X}$ instead of X, where

$$\tilde{X} = \begin{cases} 0, & \text{if } X < d \\ X, & \text{otherwise.} \end{cases}$$

Setting up $\tilde{X}$ this way, we are reflecting the nature of lab values which have skewed distribution and have zeros due to limit of detection.

Let $\tilde{x}_1, \ldots, \tilde{x}_n$ and $y_1, \ldots, y_n$ be the observed values of $\tilde{X}$ and Y in a sample with size n. Using these observations one would try to fit the model in the following:

$$P(y_i = 1) = \frac{1}{1 + \exp[-\{\beta_0 + \beta_1 \log(\tilde{x}_i)\}]}, \quad i = 1, \ldots, n. \tag{2.2}$$

However, $\log(\tilde{x}_j)$ cannot be computed for $\tilde{x}_j = 0$. Therefore, we need to use transformation $z_i = f(\tilde{x}_i)$ to replace $\log(\tilde{x}_j)$ in (2.2). Here, $f(\tilde{x}_i)$ needs to be close to $\log(\tilde{x}_j)$ and should be defined at zero.

3. Zero imputation methods

We consider five different choices for $z_i = f(\tilde{x}_i)$ to replace $\log(\tilde{x}_j)$ in model equation (2.2) as follows:

3.1. Shift by one

This method adds one to the whole observations, and then takes the logarithm. This can be written as follows:

$$z_i = \log(\tilde{x}_i + 1) \quad \text{for} \quad i = 1, \ldots, n.$$

The rationale behind this method is that it transforms zero to zero. That is, z_i becomes zero when $\tilde{x}_i = 0$. This method only involves arithmetic operations and does not require conditional statements, which makes the use of this method simple and easy. The downside of this method is that when $\tilde{X}$ is distributed too closely to 1, the impact of shifting by one can be too big so that z_i can be very different from $\log(x_i)$. In practice, sometimes any other arbitrary number such as 0.01 or 0.001 is added instead of 1, but then it does not transform zero to zero anymore and the rationale is lost.

3.2. Shift by half of the smallest nonzero

In this method, half of the minimum value of nonzero observations is added to $\tilde{x}_i = 0$ before log-transformation. In particular, we use z_i as follows:

$$z_i = \log\left(\tilde{x}_i + \frac{1}{2} \min_{\tilde{x}_i > 0} \{\tilde{x}_1, \ldots, \tilde{x}_n\}\right) \quad \text{for} \quad i = 1, \ldots, n.$$

This method is different from the 'shift by one' method in the sense that the added constant is determined based on the distribution of the observed data. Thus using this method prevents z_i from becoming too far from $\log(x_i)$. If we think of $\min_{\tilde{x}_i > 0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ as an estimate of d, $\hat{d}$, then this method is equivalent to shifting all the observations by arithmetic mean of zero and $\hat{d}$ before log-transformation.

Table 1: Distribution of the independent variable X

Scenario	1, 4	2, 5	3, 6
Distribution of X	Lognormal$(0, 1^2)$	Lognormal$(3, 1^2)$	Lognormal$(3, 3^2)$
Mean	1.65	33.12	1808.04
Median	1.00	20.09	20.09
1 percentile	0.10	1.96	0.02
3 percentile	0.15	3.06	0.07
5 percentile	0.19	3.88	0.14
10 percentile	0.28	5.58	0.43
99.5 percentile	13.14	263.97	45592.02

3.3. Shift by the square root of the smallest nonzero

We can consider adding the square root of $\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ instead of halving it, as follows:

$$z_i = \log\left(\tilde{x}_i + \sqrt{\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}}\right) \quad \text{for} \quad i = 1, \ldots, n.$$

Here, we can think of $\sqrt{\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}}$ as the geometric mean of one and $\hat{d}$. When $\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\} < 4$, this method adds a slightly bigger constant to the observed values than 'shift by half of the smallest nonzero' method, but otherwise it adds a smaller constant resulting in z_i being closer to $\log(x_i)$.

3.4. Replace zeros with half of the smallest nonzero

This method keeps nonzero observations the same and only replaces zeros with a half of the minimum of nonzero observations. That is, we use transformation as follows:

$$z_i = \begin{cases} \log\left(\frac{1}{2}\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}\right), & \text{if } \tilde{x}_i = 0 \\ \log(\tilde{x}_i), & \text{if } \tilde{x}_i > 0 \end{cases} \quad \text{for} \quad i = 1, \ldots, n.$$

For positive observations, this method does not change anything before log-transformation, resulting in z_i being pretty close to $\log(x_i)$ overall. However, it requires a conditional statement in the analysis and this can be challenging in some software.

3.5. Replace zeros with the square root of the smallest nonzero

Similar to the previous method, this method only replaces zeros with a positive constant and keeps nonzeros as they are. The difference is that the square root of the minimum of nonzero observations is used as the positive constant rather than half of the smallest nonzero. This can be written as follows:

$$z_i = \begin{cases} \log\left(\sqrt{\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}}\right), & \text{if } \tilde{x}_i = 0 \\ \log(\tilde{x}_i), & \text{if } \tilde{x}_i > 0 \end{cases} \quad \text{for } i = 1, \ldots, n.$$

When the minimum of nonzero observations is less than 1, $\sqrt{\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}} > \min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ holds. This means that zeros can be replaced by a constant that is greater than the minimum of nonzero observations, so that the order of the observations may not be preserved.

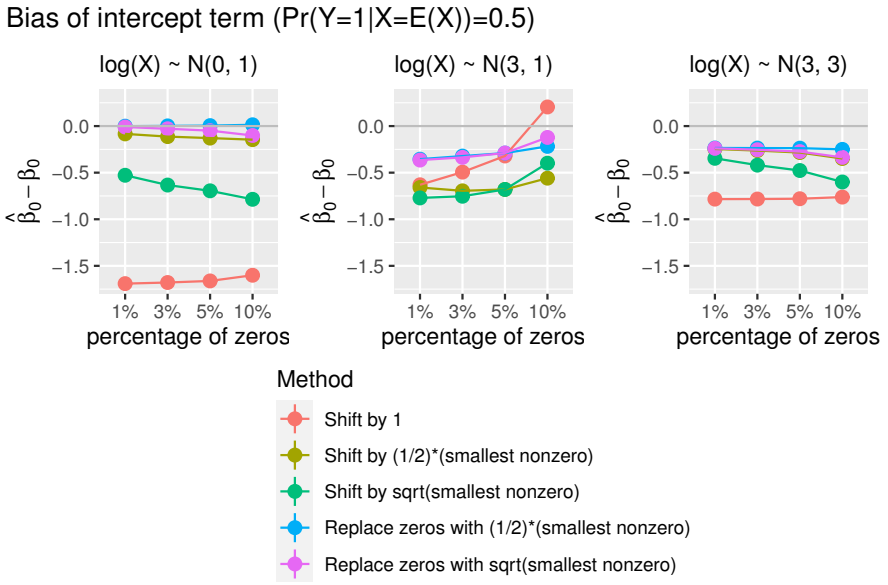

Figure 1: *Bias of intercept term β_0 averaged in 1,000 simulated datasets (Scenario 1, 2, and 3).*

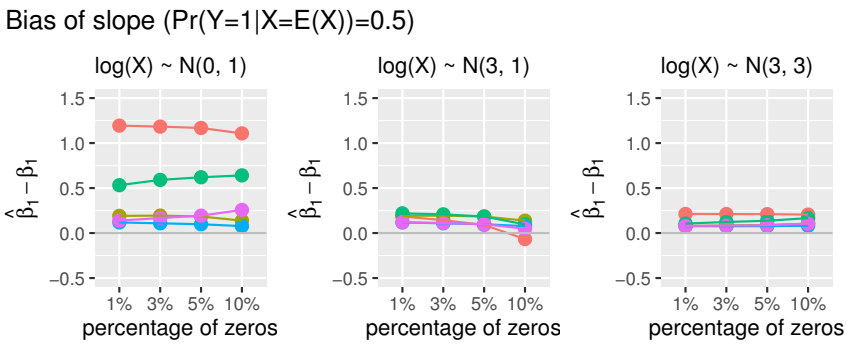

Figure 2: *Bias of slope β_1 averaged in 1,000 simulated datasets (Scenario 1, 2, and 3). Refer to Figure 1 for legend.*

4. Simulation study

To evaluate the performance of the five zero imputation methods outlined in the previous chapter, we carried out a simulation study and report the results here. Here, we outline the steps to generate simulated data. First, to simulate a continuous, right-skewed independent variable, we generated random numbers that follow a log-normal distribution. We considered three different kinds of log-normal distribution: Lognormal$(0, 1^2)$, Lognormal$(3, 1^2)$, Lognormal$(3, 3^2)$, which are similar to the distributions of commonly used lab values CEA (carcinoembryonic antigen), CRP, and CA19-9, respectively. To help grasp the overall distribution, mean, median, 1, 3, 5, 10, and 99.5 percentile of X is listed in Table 1.

Second, the lowest 1, 3, 5, or 10% of the generated values were replaced by zero to reflect the

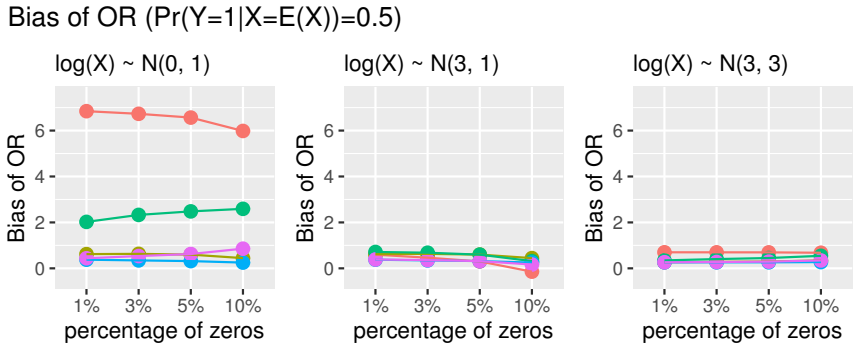

Figure 3: *Bias of OR* ($\exp(\beta_1)$) *averaged in 1,000 simulated datasets (Scenario 1, 2, and 3). Refer to Figure 1 for legend.*

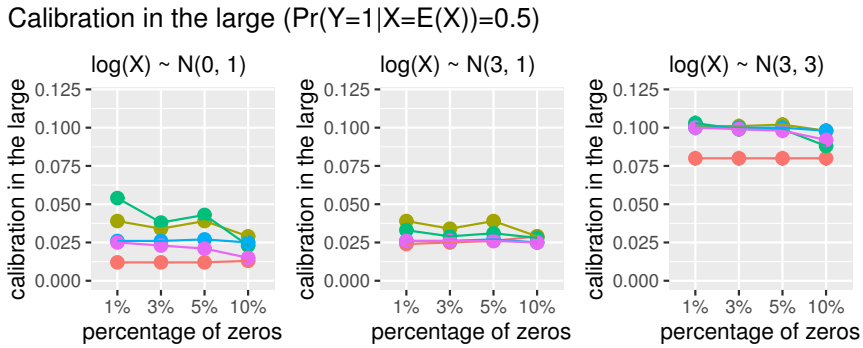

Figure 4: *Calibration in the large averaged in 1,000 simulated datasets (Scenario 1, 2, and 3). Refer to Figure 1 for legend.*

measurement error due to limit of detection.

Third, the values of the response variable were generated from the Bernoulii distribution with the success probability calculated by equation (2.2) based on the simulated data of the independent variable. We set $\beta_1 = 1$ for all scenarios. For β_0, we set the values such that the "success" probability becomes 0.5 or 0.2 when $X = E(X)$. That is, we simulated data for two types of outcome distribution: $P(Y = 1|X = E(X)) = 0.5$ and $P(Y = 1|X = E(X)) = 0.2$. The first type represents the case where the distribution of the binary response variable is balanced, and the second type shows the case where the response variable has an unbalanced distribution.

In summary, we considered 6 different scenarios. Scenario 1, 2, and 3 are for the cases where $P(Y = 1|X = E(X)) = 0.5$ and Scenario 4, 5, and 6 are for the cases where $P(Y = 1|X = E(X)) = 0.2$. In Scenario 1 and 4, the independent variable X follows Lognormal$(0, 1^2)$. Scenario 2 and 5 have Lognormal$(3, 1^2)$, and Scenario 3 and 6 have Lognormal$(3, 3^2)$ for the distribution of X.

We generated 1000 sets of samples, each with a sample size of 200. We repeated the same experiment with sample size of 100 and 300, and the results were similar, but are not reported here. For measure of performance, we calculated bias of MLE estimates of β_0, β_1, and $\exp(\beta_1)$ which is the OR (odds ratio) of X. We also generated a separate test set with a size of 1000 for each scenario so that we

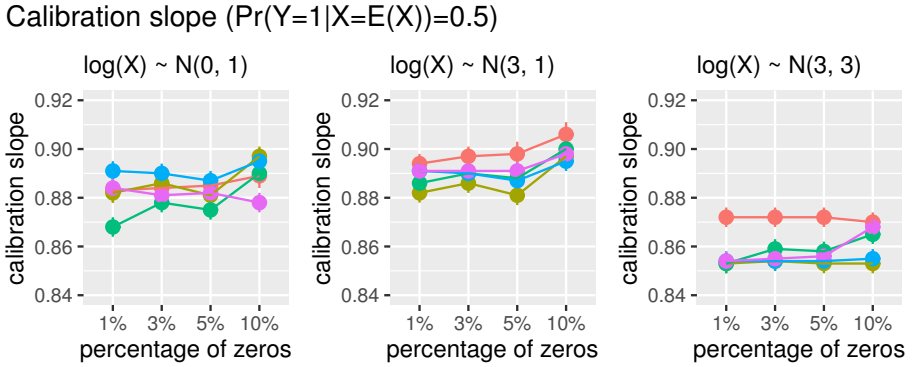

Figure 5: *Calibration slope averaged in 1,000 simulated datasets (Scenario 1, 2, and 3). Refer to Figure 1 for legend.*

could evaluate the prediction performance of the resulting logistic regression model. The c-index of the model was calculated and reported. In addition, we calculated calibration in the large and calibration slope based on the test set so that we can understand the effect of the zero imputation method on calibration of the resulting logistic regression model. Calibration means the agreement between the observed outcomes and the predicted response. In the logistic regression context, calibration in the large stands for the mean difference between estimated vs. actual logit of the success probability, and the calibration slope refers to the slope of the recalibration model: $\text{Logit}(p) = \alpha + \beta \cdot \text{logit}(\hat{p})$ where $p = Pr(Y = 1|X)$ (Steyerberg, 2019). If the fitted model has perfect calibration, the calibration in the large would be 0 and the calibration slope would be 1. The performance measures were calculated for each of the 1000 simulated data sets, and we report their mean and standard deviation in Tables 2–7 and Figures 1–10.

First, let's take a look at the scenario 1 to 3 where the response variable has balanced distribution. Figure 1 shows that the bias in the intercept term tends to be negative. This is because, all five zero imputation methods tend to make the value of independent variable greater than or equal to the originally observed value. That is, $z_i \geq \log(\tilde{x}_i)$ holds for all cases except for rare cases where $\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\} < \sqrt{\min_{\tilde{x}_i>0}\{\tilde{x}_1, \ldots, \tilde{x}_n\}} < z_i < 1$ holds and 'replace zeros with square root of the smallest nonzero' method is used. Hence, $\beta_0 + \beta_1 z_i$ would be greater than $\beta_0 + \beta_1 \log(x_i)$ for most cases when $\beta_1 > 0$. To compensate the increase in the value of $\beta_0 + \beta_1 X$ in the logistic regression model, the estimate of $\hat{\beta}_0$ tends to become lower than β_0. Bias of $\hat{\beta}_0$ is the greatest (in terms of absolute value) when the 'shift by 1' method was used in scenario 1 because using this method in scenario 1 the difference between the observed value $\log(\tilde{x}_i)$ and the imputed value z_i is the greatest.

Bias in the slope is presented in Figure 2. With all five zero imputation methods, the range of z_i's becomes narrower than that of the original observations $\log(x_i)$'s. Therefore, $\hat{\beta}_1$ tends to be greater than β_1 to make $\beta_0 + \beta_1 z_i$ cover the range of $\beta_0 + \beta_1 \log(x_i)$, which is why the bias of $\hat{\beta}_1$ tends to be positive. Shrinkage in the range is the greatest in scenario 1 using 'shift by 1' method, which is reflected in the magnitude of the bias of $\hat{\beta}_1$. Bias in the OR is directly related with bias in the slope because $\text{OR} = \exp(\beta_1)$ holds. Figure 3 shows that the bias of OR can be quite dramatic because of exponentiation.

The c-indices in the training set and testing set were almost the same for all five methods because the different zero imputation methods do not change the order of observations of the independent

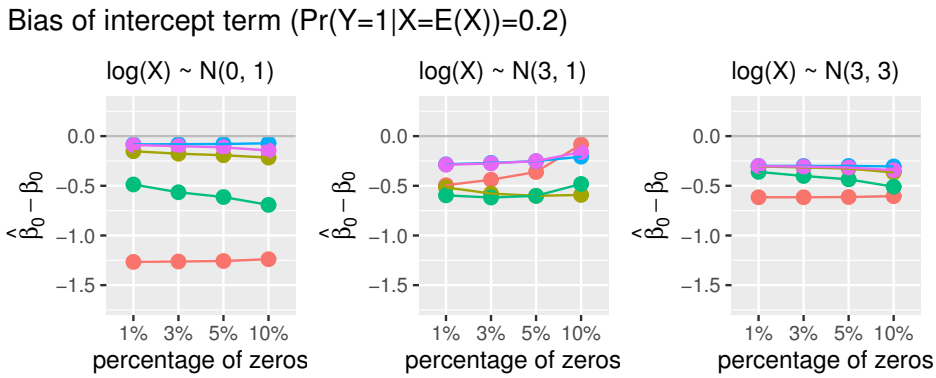

Figure 6: *Bias of intercept term β_0 averaged in 1,000 simulated datasets (Scenario 4, 5, and 6). Refer to Figure 1 for legend.*

variable most of the time.

Figure 4 shows calibration in the large is all positive, which means that the success probability is underestimated overall using z_i. Figure 5 shows that the calibration slopes are less than 1, which implies that the estimated success probabilities are too extreme, that is, they are close to either zero or one. This seems to be due to the fact that the bias of $\hat{\beta}_1$ is positive. That is, because β_1 is overestimated, effect of independent variable is exaggerated.

Now, we consider the results of scenario 4 to 6, where the response variable has an unbalanced distribution. From Figures 6, 7, and 8, we can see similar patterns to those in scenario 1 to 3, where the bias of the intercept term tends to be negative and the bias of the slope is more likely to be positive.

Figure 9 shows calibration in the large is all positive, which is also similar to scenario 1 to 3, except that the magnitude of calibration in the large is much bigger in scenario 4. The calibration slope was less than 1 in scenario 5 and 6, but tends to be greater than 1 in scenario 4. These results imply that when the success probability is low and the range of the independent variable is narrow, the success probability tends to be greatly underestimated and the effect of the independent variable is also underestimated. Overall, shift by 1 method remains to be the worst method, similarly to scenario 1 to 3.

5. Conclusion

We investigated the impact of zero imputation methods for nonnegative independent variables with skewed distribution in logistic regression. We evaluated and compared the performance of five zero imputation methods commonly used in clinical research using a simulation study. Overall, bias in the estimated coefficients tends to be high when the range of independent variable is narrow, especially with the 'shift by 1' method. Bias can be dramatically high in OR estimate and thus, caution is necessary when selecting zero imputation method especially when the range of the observed values of independent variable is rather narrow. The effect of the zero imputation method in the discrimination of the resulting logistic regression model is minimal, but calibration suffers somehow, although there is not so much of a difference among the zero imputation methods. Overall, it is recommended to avoid the shift by 1 method because it can yield too high a bias in the parameter estimation despite its ease of use. Shift by half of the smallest nonzero, replace zeros with half of the smallest nonzero, and replace zeros with the square root of the smallest nonzero methods showed comparable and stable

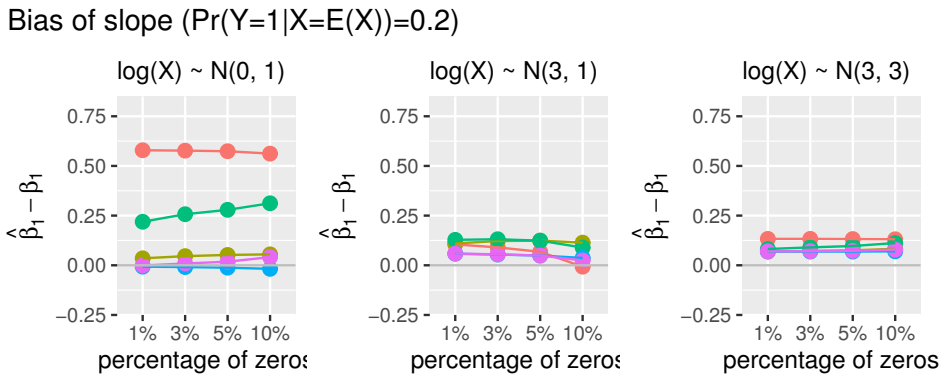

Figure 7: *Bias of slope β_1 averaged in 1,000 simulated datasets (Scenario 4, 5, and 6). Refer to Figure 1 for legend.*

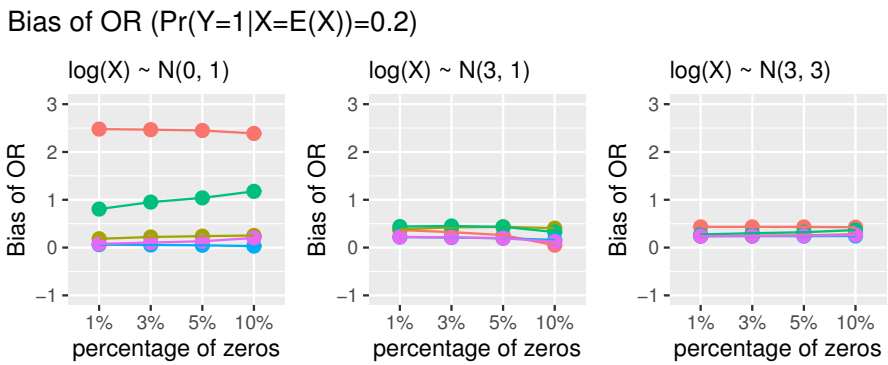

Figure 8: *Bias of OR* ($\exp(\beta_1)$) *averaged in 1,000 simulated datasets (Scenario 4, 5, and 6). Refer to Figure 1 for legend.*

performance. In the future, investigation of the effect of the zero imputation methods in the Cox proportional hazards model, which is another frequently used method in clinical research, is warranted. Another possible research subject includes investigation of the effect of the zero imputation methods on agreement studies, where a new technique to measure a quantity is compared to the gold standard.

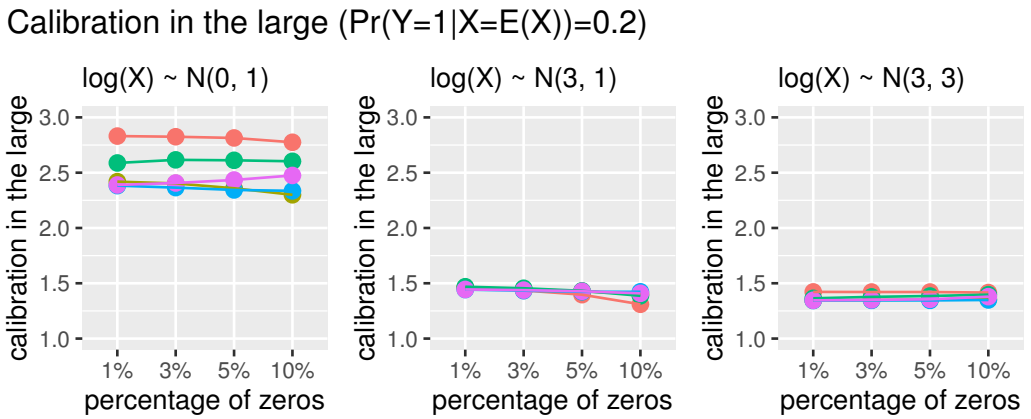

Figure 9: *Calibration in the large averaged in 1,000 simulated datasets (Scenario 4, 5, and 6). Refer to Figure 1 for legend.*

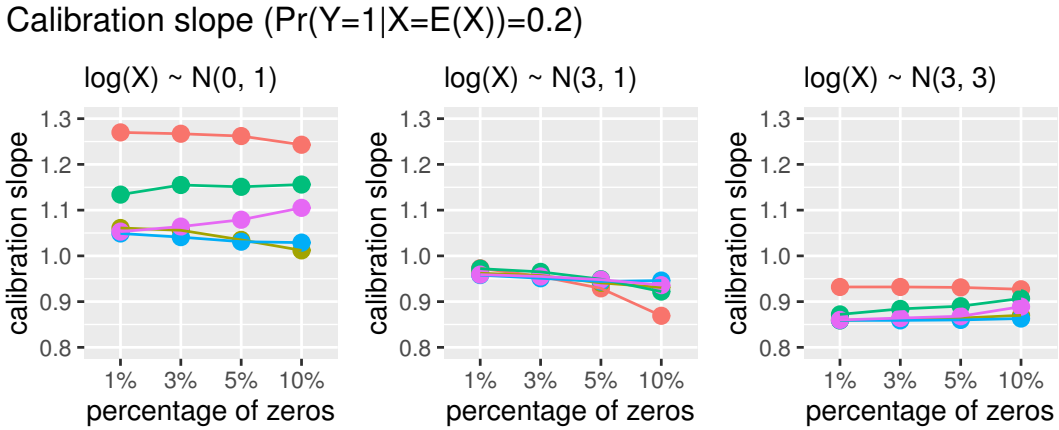

Figure 10: *Calibration slope averaged in 1,000 simulated datasets (Scenario 4, 5, and 6). Refer to Figure 1 for legend.*

Table 2: Results of the simulation study for scenario 1 where $P(Y = 1|X = E(X)) = 0.5$, $X \sim \text{Lognormal}(0, 1^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−1.691(0.273)	−0.084(0.147)	−0.528(0.165)	−0.001(0.145)	−0.010(0.145)
	$\hat{\beta}_1 - \beta_1$	1.194(0.351)	0.191(0.164)	0.532(0.219)	0.118(0.152)	0.138(0.153)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	6.844(3.610)	0.619(0.559)	2.023(1.084)	0.376(0.476)	0.439(0.490)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.737(0.000)
	Calibration in the large	0.012(0.137)	0.039(0.135)	0.054(0.139)	0.026(0.135)	0.025(0.134)
	Calibration slope	0.883(0.146)	0.882(0.126)	0.868(0.127)	0.891(0.126)	0.884(0.124)
3%	$\hat{\beta}_0 - \beta_0$	−1.679(0.271)	−0.113(0.147)	−0.633(0.168)	0.002(0.145)	−0.029(0.144)
	$\hat{\beta}_1 - \beta_1$	1.183(0.348)	0.194(0.171)	0.591(0.230)	0.108(0.152)	0.167(0.158)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	6.729(3.535)	0.631(0.586)	2.327(1.215)	0.345(0.473)	0.534(0.521)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.758(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.735(0.000)
	Calibration in the large	0.012(0.137)	0.034(0.135)	0.038(0.139)	0.026(0.135)	0.023(0.133)
	Calibration slope	0.884(0.146)	0.886(0.132)	0.878(0.131)	0.890(0.128)	0.881(0.124)
5%	$\hat{\beta}_0 - \beta_0$	−1.662(0.269)	−0.129(0.147)	−0.695(0.170)	0.005(0.145)	−0.049(0.144)
	$\hat{\beta}_1 - \beta_1$	1.168(0.344)	0.184(0.173)	0.620(0.236)	0.098(0.151)	0.193(0.164)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	6.565(3.419)	0.598(0.587)	2.480(1.283)	0.315(0.465)	0.624(0.560)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.756(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.734(0.000)
	Calibration in the large	0.012(0.137)	0.039(0.134)	0.043(0.137)	0.027(0.134)	0.021(0.133)
	Calibration slope	0.885(0.145)	0.881(0.135)	0.875(0.131)	0.887(0.128)	0.882(0.126)
10%	$\hat{\beta}_0 - \beta_0$	−1.601(0.261)	−0.146(0.147)	−0.786(0.174)	0.014(0.145)	−0.101(0.142)
	$\hat{\beta}_1 - \beta_1$	1.108(0.328)	0.138(0.166)	0.641(0.237)	0.077(0.146)	0.257(0.183)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	5.983(3.048)	0.445(0.536)	2.591(1.313)	0.250(0.439)	0.856(0.671)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.751(0.027)
	c-index in test set	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.728(0.000)
	Calibration in the large	0.013(0.137)	0.029(0.135)	0.023(0.137)	0.025(0.135)	0.015(0.133)
	Calibration slope	0.889(0.143)	0.897(0.138)	0.890(0.133)	0.895(0.128)	0.878(0.133)

Table 3: Results of the simulation study for scenario 2 where $P(Y = 1|X = E(X)) = 0.5$, $X \sim \text{Lognormal}(3, 1^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−0.630(0.511)	−0.659(0.518)	−0.771(0.526)	−0.355(0.472)	−0.365(0.470)
	$\hat{\beta}_1 - \beta_1$	0.184(0.163)	0.191(0.164)	0.219(0.167)	0.118(0.152)	0.121(0.151)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.594(0.547)	0.619(0.559)	0.713(0.583)	0.376(0.476)	0.386(0.476)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	0.024(0.136)	0.039(0.135)	0.033(0.135)	0.026(0.135)	0.026(0.135)
	Calibration slope	0.894(0.128)	0.882(0.126)	0.886(0.126)	0.891(0.126)	0.891(0.125)
3%	$\hat{\beta}_0 - \beta_0$	−0.494(0.530)	−0.695(0.545)	−0.753(0.545)	−0.322(0.475)	−0.336(0.472)
	$\hat{\beta}_1 - \beta_1$	0.143(0.168)	0.194(0.171)	0.208(0.171)	0.108(0.152)	0.112(0.152)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.463(0.542)	0.631(0.586)	0.680(0.594)	0.345(0.473)	0.358(0.473)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	0.025(0.136)	0.034(0.135)	0.029(0.135)	0.026(0.135)	0.026(0.135)
	Calibration slope	0.897(0.139)	0.886(0.132)	0.890(0.132)	0.890(0.128)	0.891(0.127)
5%	$\hat{\beta}_0 - \beta_0$	−0.321(0.542)	−0.680(0.557)	−0.680(0.552)	−0.290(0.473)	−0.289(0.472)
	$\hat{\beta}_1 - \beta_1$	0.090(0.170)	0.184(0.173)	0.184(0.173)	0.098(0.151)	0.098(0.151)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.300(0.519)	0.598(0.587)	0.597(0.582)	0.315(0.465)	0.315(0.464)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	0.026(0.136)	0.039(0.134)	0.031(0.134)	0.027(0.134)	0.026(0.135)
	Calibration slope	0.898(0.150)	0.881(0.135)	0.888(0.136)	0.887(0.128)	0.891(0.129)
10%	$\hat{\beta}_0 - \beta_0$	0.205(0.523)	−0.559(0.545)	−0.398(0.535)	−0.218(0.459)	−0.122(0.459)
	$\hat{\beta}_1 - \beta_1$	−0.067(0.159)	0.138(0.166)	0.096(0.163)	0.077(0.146)	0.048(0.145)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	−0.144(0.415)	0.445(0.536)	0.314(0.502)	0.250(0.439)	0.165(0.422)
	c-index in training set	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)	0.759(0.026)
	c-index in test set	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)
	Calibration in the large	0.029(0.138)	0.029(0.135)	0.028(0.136)	0.025(0.135)	0.025(0.136)
	Calibration slope	0.906(0.166)	0.897(0.138)	0.900(0.142)	0.895(0.128)	0.898(0.131)

Table 4: Results of the simulation study for scenario 3 where $P(Y = 1|X = E(X)) = 0.5$, $X \sim \text{Lognormal}(3, 3^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−0.784(0.591)	−0.246(0.571)	−0.346(0.567)	−0.235(0.571)	−0.239(0.570)
	$\hat{\beta}_1 - \beta_1$	0.212(0.183)	0.079(0.172)	0.105(0.172)	0.076(0.172)	0.078(0.171)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.699(0.654)	0.268(0.533)	0.346(0.548)	0.260(0.532)	0.264(0.531)
	c-index in training set	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	0.080(0.157)	0.101(0.154)	0.103(0.154)	0.100(0.154)	0.100(0.154)
	Calibration slope	0.872(0.133)	0.853(0.138)	0.853(0.134)	0.854(0.139)	0.854(0.138)
3%	$\hat{\beta}_0 - \beta_0$	−0.783(0.591)	−0.263(0.571)	−0.419(0.570)	−0.236(0.571)	−0.254(0.565)
	$\hat{\beta}_1 - \beta_1$	0.211(0.183)	0.083(0.172)	0.123(0.173)	0.076(0.172)	0.082(0.170)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.697(0.654)	0.282(0.536)	0.403(0.562)	0.261(0.532)	0.275(0.529)
	c-index in training set	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	0.080(0.157)	0.101(0.154)	0.099(0.155)	0.100(0.154)	0.099(0.154)
	Calibration slope	0.872(0.133)	0.854(0.138)	0.859(0.134)	0.854(0.138)	0.855(0.136)
5%	$\hat{\beta}_0 - \beta_0$	−0.780(0.591)	−0.284(0.573)	−0.477(0.573)	−0.237(0.570)	−0.274(0.561)
	$\hat{\beta}_1 - \beta_1$	0.210(0.183)	0.089(0.173)	0.137(0.175)	0.077(0.172)	0.087(0.169)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.695(0.653)	0.297(0.541)	0.449(0.575)	0.262(0.531)	0.291(0.529)
	c-index in training set	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	0.080(0.157)	0.102(0.154)	0.099(0.155)	0.100(0.154)	0.098(0.154)
	Calibration slope	0.872(0.133)	0.853(0.138)	0.858(0.133)	0.854(0.138)	0.856(0.135)
10%	$\hat{\beta}_0 - \beta_0$	−0.762(0.591)	−0.348(0.575)	−0.600(0.580)	−0.250(0.565)	−0.336(0.552)
	$\hat{\beta}_1 - \beta_1$	0.205(0.183)	0.104(0.174)	0.166(0.178)	0.081(0.170)	0.104(0.168)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.678(0.649)	0.344(0.553)	0.545(0.604)	0.272(0.528)	0.343(0.534)
	c-index in training set	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)	0.924(0.020)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	0.080(0.157)	0.098(0.154)	0.088(0.156)	0.098(0.154)	0.092(0.156)
	Calibration slope	0.870(0.133)	0.853(0.136)	0.865(0.133)	0.855(0.137)	0.868(0.134)

Table 5: Results of the simulation study for scenario 4 where $P(Y = 1|X = E(X)) = 0.2$, $X \sim \text{Lognormal}(0, 1^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−1.266(0.438)	−0.152(0.255)	−0.487(0.302)	−0.083(0.247)	−0.089(0.247)
	$\hat{\beta}_1 - \beta_1$	0.579(0.365)	0.035(0.246)	0.219(0.281)	−0.007(0.238)	−0.001(0.238)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	2.479(2.134)	0.184(0.748)	0.805(1.060)	0.061(0.693)	0.077(0.696)
	c-index in training set	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.740(0.048)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.737(0.000)
	Calibration in the large	2.831(0.738)	2.419(0.592)	2.588(0.628)	2.383(0.590)	2.389(0.597)
	Calibration slope	1.270(0.361)	1.061(0.300)	1.134(0.313)	1.049(0.301)	1.053(0.304)
3%	$\hat{\beta}_0 - \beta_0$	−1.262(0.438)	−0.177(0.259)	−0.565(0.314)	−0.081(0.247)	−0.101(0.248)
	$\hat{\beta}_1 - \beta_1$	0.577(0.364)	0.046(0.250)	0.257(0.289)	−0.009(0.238)	0.009(0.239)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	2.467(2.124)	0.221(0.771)	0.951(1.138)	0.055(0.691)	0.105(0.708)
	c-index in training set	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.739(0.048)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.735(0.000)
	Calibration in the large	2.825(0.742)	2.403(0.635)	2.616(0.654)	2.365(0.607)	2.407(0.580)
	Calibration slope	1.267(0.363)	1.056(0.320)	1.155(0.326)	1.041(0.309)	1.064(0.297)
5%	$\hat{\beta}_0 - \beta_0$	−1.257(0.437)	−0.192(0.263)	−0.613(0.321)	−0.079(0.247)	−0.113(0.249)
	$\hat{\beta}_1 - \beta_1$	0.574(0.364)	0.052(0.253)	0.279(0.295)	−0.012(0.238)	0.018(0.241)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	2.450(2.115)	0.239(0.788)	1.039(1.190)	0.048(0.691)	0.132(0.719)
	c-index in training set	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.738(0.048)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.734(0.000)
	Calibration in the large	2.814(0.733)	2.359(0.599)	2.612(0.645)	2.344(0.583)	2.434(0.595)
	Calibration slope	1.262(0.359)	1.035(0.305)	1.151(0.321)	1.031(0.299)	1.079(0.302)
10%	$\hat{\beta}_0 - \beta_0$	−1.239(0.436)	−0.216(0.269)	−0.692(0.334)	−0.072(0.247)	−0.144(0.251)
	$\hat{\beta}_1 - \beta_1$	0.562(0.362)	0.055(0.260)	0.312(0.303)	−0.018(0.239)	0.041(0.246)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	2.388(2.080)	0.253(0.814)	1.176(1.271)	0.030(0.687)	0.201(0.749)
	c-index in training set	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.740(0.048)	0.735(0.049)
	c-index in test set	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.728(0.000)
	Calibration in the large	2.775(0.705)	2.300(0.572)	2.603(0.637)	2.336(0.570)	2.475(0.633)
	Calibration slope	1.243(0.347)	1.012(0.296)	1.156(0.317)	1.029(0.295)	1.105(0.318)

Table 6: Results of the simulation study for scenario 5 where $P(Y = 1|X = E(X)) = 0.2$, $X \sim \text{Lognormal}(3, 1^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−0.494(0.747)	−0.516(0.750)	−0.596(0.758)	−0.284(0.720)	−0.288(0.719)
	$\hat{\beta}_1 - \beta_1$	0.105(0.205)	0.110(0.205)	0.128(0.207)	0.058(0.198)	0.059(0.198)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.367(0.672)	0.382(0.677)	0.440(0.696)	0.220(0.619)	0.223(0.619)
	c-index in training set	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	1.463(0.252)	1.460(0.249)	1.469(0.251)	1.443(0.247)	1.444(0.247)
	Calibration slope	0.973(0.185)	0.962(0.183)	0.972(0.184)	0.958(0.184)	0.959(0.183)
3%	$\hat{\beta}_0 - \beta_0$	−0.439(0.759)	−0.578(0.772)	−0.618(0.774)	−0.269(0.722)	−0.275(0.720)
	$\hat{\beta}_1 - \beta_1$	0.090(0.208)	0.122(0.211)	0.131(0.211)	0.054(0.199)	0.055(0.198)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.323(0.673)	0.423(0.704)	0.451(0.712)	0.209(0.618)	0.213(0.618)
	c-index in training set	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	1.435(0.253)	1.448(0.250)	1.455(0.252)	1.433(0.246)	1.438(0.246)
	Calibration slope	0.955(0.189)	0.957(0.185)	0.965(0.185)	0.951(0.184)	0.955(0.184)
5%	$\hat{\beta}_0 - \beta_0$	−0.361(0.775)	−0.601(0.789)	−0.601(0.786)	−0.252(0.723)	−0.251(0.723)
	$\hat{\beta}_1 - \beta_1$	0.068(0.213)	0.124(0.215)	0.124(0.215)	0.049(0.199)	0.049(0.199)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.261(0.674)	0.433(0.721)	0.433(0.719)	0.195(0.617)	0.195(0.617)
	c-index in training set	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)
	c-index in test set	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)	0.738(0.000)
	Calibration in the large	1.397(0.253)	1.428(0.249)	1.432(0.251)	1.423(0.245)	1.428(0.246)
	Calibration slope	0.929(0.192)	0.941(0.185)	0.949(0.186)	0.944(0.183)	0.948(0.184)
10%	$\hat{\beta}_0 - \beta_0$	−0.084(0.808)	−0.592(0.812)	−0.481(0.804)	−0.207(0.720)	−0.160(0.725)
	$\hat{\beta}_1 - \beta_1$	−0.007(0.221)	0.114(0.220)	0.089(0.218)	0.037(0.198)	0.024(0.199)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.050(0.654)	0.406(0.734)	0.326(0.711)	0.160(0.606)	0.124(0.603)
	c-index in training set	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)	0.749(0.036)
	c-index in test set	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)	0.737(0.000)
	Calibration in the large	1.310(0.251)	1.402(0.247)	1.387(0.247)	1.423(0.244)	1.409(0.244)
	Calibration slope	0.869(0.205)	0.931(0.188)	0.922(0.191)	0.946(0.185)	0.936(0.186)

Table 7: Results of the simulation study for scenario 6 where $P(Y = 1|X = E(X)) = 0.2$, $X \sim \text{Lognormal}(3, 3^2)$

Percentage of zeros	Measure of performance	Shift by 1	Shift by (1/2)(smallest nonzero)	Shift by sqrt(smallest nonzero)	Replace zeros with (1/2)(smallest nonzero)	Replace zeros with sqrt(smallest nonzero)
1%	$\hat{\beta}_0 - \beta_0$	−0.615(0.717)	−0.305(0.727)	−0.360(0.723)	−0.299(0.728)	−0.301(0.727)
	$\hat{\beta}_1 - \beta_1$	0.133(0.171)	0.070(0.171)	0.082(0.170)	0.069(0.171)	0.070(0.171)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.433(0.564)	0.242(0.531)	0.275(0.535)	0.239(0.531)	0.240(0.531)
	c-index in training set	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	1.422(0.208)	1.346(0.186)	1.364(0.190)	1.344(0.186)	1.346(0.186)
	Calibration slope	0.932(0.140)	0.859(0.136)	0.872(0.136)	0.859(0.137)	0.860(0.137)
3%	$\hat{\beta}_0 - \beta_0$	−0.615(0.717)	−0.315(0.727)	−0.401(0.721)	−0.300(0.728)	−0.307(0.724)
	$\hat{\beta}_1 - \beta_1$	0.133(0.171)	0.072(0.171)	0.090(0.170)	0.069(0.171)	0.071(0.170)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.433(0.564)	0.248(0.532)	0.300(0.539)	0.239(0.531)	0.244(0.530)
	c-index in training set	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	1.421(0.208)	1.349(0.187)	1.377(0.193)	1.345(0.186)	1.350(0.187)
	Calibration slope	0.932(0.140)	0.862(0.137)	0.884(0.137)	0.859(0.137)	0.864(0.136)
5%	$\hat{\beta}_0 - \beta_0$	−0.613(0.717)	−0.327(0.726)	−0.435(0.720)	−0.300(0.727)	−0.316(0.721)
	$\hat{\beta}_1 - \beta_1$	0.132(0.171)	0.075(0.171)	0.097(0.170)	0.069(0.171)	0.073(0.169)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.432(0.564)	0.255(0.534)	0.321(0.543)	0.239(0.531)	0.249(0.529)
	c-index in training set	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	1.421(0.208)	1.352(0.187)	1.385(0.195)	1.345(0.186)	1.355(0.188)
	Calibration slope	0.931(0.140)	0.864(0.136)	0.890(0.137)	0.860(0.137)	0.868(0.136)
10%	$\hat{\beta}_0 - \beta_0$	−0.604(0.719)	−0.366(0.726)	−0.509(0.719)	−0.305(0.725)	−0.343(0.714)
	$\hat{\beta}_1 - \beta_1$	0.131(0.171)	0.083(0.171)	0.112(0.170)	0.070(0.170)	0.079(0.168)
	$\exp(\hat{\beta}_1) - \exp(\beta_1)$	0.426(0.564)	0.279(0.538)	0.366(0.552)	0.242(0.530)	0.266(0.527)
	c-index in training set	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)	0.925(0.019)
	c-index in test set	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)	0.908(0.000)
	Calibration in the large	1.416(0.207)	1.355(0.189)	1.397(0.200)	1.349(0.187)	1.379(0.193)
	Calibration slope	0.927(0.140)	0.870(0.137)	0.907(0.138)	0.863(0.137)	0.889(0.138)

Acknowledgement

This research was supported by the Korea National Open University Research Fund.

References

Bellégo C, Benatia D, and Pape L (2022). Dealing with logs and zeros in regressinon models, Available from: *arXiv* eprint 2203.11820

Box GEP and Cox DR (1964). An analysis of transformations, *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**, 211–243.

Durbin BP and Rocke DM (2004). Variance-stabilizing transformations for two-color microarrays, *Bioinformatics*, **20**, 660–667.

Ekwaru JP and Veugelers PJ (2018). The overlooked importance of constants added in log transformation of independent variables with zero values: A proposed approach for determining an optimal constant, *Statistics in Biopharmaceutical Research*, **10**, 26–29.

Feng C, Hongyue W, Lu N, Chen T, He H, Lu Y, and Tu X (2014). Log-transformation and its implications for data analysis, *Shanghai Archives of Psychiatry*, **26**, 105–109.

Park SY (2023). Zero imputation methods for log-transformation of independent variables, *Journal of the Korean Data Analysis Society*, **25**, 79–90.

Rocke DM and Durbin-Johnson B (2001). A model for measurement error for gene expression arrays, *Journal of Computational Biology*, **8**, 557–569.

Rocke DM and Durbin-Johnson B (2003). Approximate variance-stabilizing transformations for gene-expression microarray data, *Bioinformatics*, **19**, 966–972.

Steyerberg EW (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* (2nd ed), Springer, Berlin.

Received December 12, 2023; Revised January 16, 2024; Accepted February 06, 2024