Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/100201

Impact of Magnitude of Zero Inflation of Covariates on Statistical Inference and Model Selection

Milan Bimali[1,2,3,*], *Songthip T. Ounpraseuth*[1] *and David Keith Williams*[1,2]

[1]Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, AR, USA
[2]Arkansas Children's Nutrition Center, Arkansas Children's Hospital, Little Rock, AR, USA
[3]Winthrop P. Rockefeller Cancer Institute, Arkansas Children's Hospital, Little Rock, AR, USA

Received: 8 Oct. 2020, Revised: 20 Jan. 2021, Accepted: 28 Jan. 2021
Published online: 1 Jul. 2021

Abstract: Modeling approaches that do not consider zero-inflation are inappropriate in modelling the relationship between zero-inflated outcomes and covariates. Similarly, the association between zero-inflated covariate and outcome using the aforementioned approaches is also prone to estimation and inferential errors. The case of zero-inflated covariate despite being observed in a wide variety of scenarios has attracted little attention. While the need to develop and implement specialized approach to model the association between zero-inflated covariate and outcome is indisputable, a more fundamental question that needs to be explored is whether the magnitude of zero inflation is large enough to warrant concern and whether the degree of this concern depends on the overall size of the data and the analysis objective. The present paper employs extensive simulation-based approach to assess the effect of magnitude of zero-inflated covariate on a number of statistical metrics, such as error rates and variable selection rates across a wide spectrum of sample size in the context of two commonly used modeling approach – logistic regression, and linear regression.

Keywords: Error rates, Variable selection, Zero-inflated covariates.

1 Introduction

Zero-Inflated data have a unusually higher proportion of zeros than expected under assumed standard distributions. Zero-inflated data are commonly seen in various situations such as counts of contraceptive use during sexual intercourse, consumption of specific fruits and vegetables, and alcohol intake in teetotaler communities. In the presence of zero-inflated variables, traditional modeling approach such as linear regression is inappropriate [1-2]. Variable transformation is not an appealing approach as the issue inflation stays intact. Categorizing the zero-inflated variables will lead to several issues, such as loss of information and loss of power [3-6].

Zero-inflated variable may manifest itself as an outcome or covariate. Modeling approaches, such as zero-inflated Poisson, hurdle model, Tobit model, and negative binomial modeling are developed specifically for the scenarios, where the outcomes exhibit larger frequency of zero observations [7-10]. The effect of zero-inflated covariates in the context of modeling has attracted relatively little attention. This may be due to the fact that statistical modeling approaches typically do not make distributional assumptions on the covariates. Royston et al. suggested that modeling zero-inflated covariates using fractional polynomials (power transformation) may lead to a better model fit than the untransformed covariates [11].

Although the statistical literature has a dedicated history of developing modeling approaches for zero-inflated variable as an outcome and to some extent as a covariate, a more fundamental question that needs to be explored is whether the magnitude of zero inflation is large enough to warrant concern and whether the degree of this concern depends on the overall size of the data and the analysis objective. For example, does a covariate whose 50% of the values are 0 consistently exhibit inflated type I error rate across different spectrum of sample sizes? There are no specific guidelines on what percentage of zeroes in the covariate is large enough to preclude the use of conventional inferential and modeling approaches. It is also unclear whether this threshold bears any relation to the overall size of the data. The notion of

[*]Corresponding author e-mail: mbimali@uams.edu

magnitude of zero-inflation and subsequent analysis approach, as seen in literature, appears to be largely based on the discretion of investigator and analyst.

To the authors' knowledge, there have been no prior studies that evaluate the effect of magnitude of zero inflation (relative to sample size) on different aspects of inferential statistics. The present paper aims at filling this gap in literature. We employ an extensive simulation-based approach to assess the effect of zero-inflated covariates on a number of statistical metrics across a wide spectrum of sample size. We evaluate the effect of zero-inflated covariate in terms of error rates, stability of parameter estimates, and in the context of variable selection within the framework of two commonly used regression models - logistic regression and linear regression. For the purpose of our study, the 0's are true continuous values. The rest of this paper is divided into 3 sections. The methods section provides details on data simulation and computation of different metrics (error rates, parameter stability, and variable selection rate). The results section will summarize the findings based on the simulated data. We have discussed the results and made recommendations in the discussion section.

2 Research Methods

2.1 Simulation Outline

We carried out Monte Carlo simulations to evaluate the effect of the magnitude of zero inflation on error rates, stability of parameter, and variable selection rate under different scenarios. The simulation approach for assessing the effect on error rates and stability of parameter involved a simple setting with 2 covariates, of which one was zero-inflated covariate of interest and the other was a confounder. To evaluate the effect of the magnitude of zero inflation on the rate of variable (zero-inflated covariate) selection, we increased the number of covariates to 4, of which one was zero-inflated covariate. The covariates were generated from different statistical distributions and are uncorrelated. Zero-inflation in simulated covariate was introduced by setting a randomly selected subset of it to zero.

The binary and continuous outcomes were simulated from binomial and normal distributions respectively. The sample sizes were set at: 50, 100, 200, 300, 400, and 500. The percentage of zero inflation ranges from 50% to 90% in the increment of 10%. Table 1 shows on distribution of covariate, values of true parameters and distribution of outcomes used in simulation. We simulated 10,000 datasets. Logistic and linear regression models were fitted, as follows:
Scenario 1: $logit(\pi(y)) = \beta_1 x_1 + \beta_2 x_2 + \epsilon$
Scenario 2: $logit(\pi(y)) = \beta_2 x_2 + \epsilon$
Scenario 3: $y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$
Scenario 4: $y = \beta_2 x_2 + \epsilon$

We recorded the value of log-odds for logistics regression and slope estimates for linear regression together with the corresponding p-value. The variable selection was done using the penalized LASSO algorithm. LASSO model is used primarily for the purpose of variable selection and as such we have not focused on other aspects, such as parameter estimation and MSE. The shrinkage parameter was chosen based on minimization of cross-validated mean squared error. The following model was fitted:
Scenario 5: $y = \Sigma_{j=1}^{4} \beta_j x_j + \lambda \Sigma_{j=1}^{4} |\beta_j|$

Table 1: Summary of Parameters and Distributions used in Simulation.

Scenario	Covariates	Covariate Distribution	Parameter	Outcome
Scenario: 1/3	$(X_1, \boldsymbol{X_2})$	$X_1 \sim Poisson(\lambda = 3)$ $X_2 \sim N(0,1)$	$(\beta_1, \beta_2) = (1,2)$	Logistic Regression: $y \sim Bin\left(n, p = \frac{1}{1+\exp(-\Sigma_{j=1}^{2}\beta_j X_j)}\right)$ Linear Regression: $y \sim N(\mu = \Sigma_{j=1}^{2}\beta_j X_j, \sigma = 2)$
Scenario: 2/4	$(X_1, \boldsymbol{X_2})$	$X_1 \sim Poisson(\lambda = 3)$ $X_2 \sim N(0,1)$	$(\beta_1, \beta_2) = (1,0)$	
Scenario: 5	$(X_1, \boldsymbol{X_2}, X_3, X_4)$	$X_1 \sim Poisson(\lambda = 3)$ $X_2 \sim N(0,1)$ $X_3 \sim Binomial(p = 0.3)$ $X_4 \sim N(5,2)$	$(\beta_1, \beta_2, \beta_3, \beta_4)$ $= (1,2,0.5,-1)$	Logistic Regression: $y \sim Bin\left(n, p = \frac{1}{1+\exp(-\Sigma_{j=1}^{4}\beta_j X_j)}\right)$ Linear Regression: $y \sim N(\mu = \Sigma_{j=1}^{4}\beta_j X_j, \sigma = 2)$
Note: X_2 is zero-inflated covariates.				

2.2 Computation of Error and Variable Selection Rate

Error Rate 1: This error rate is computed based on simulation scenarios 1 and 3. It is the proportion of “non-significant association between zero-inflated covariate and outcome”, based on a pre-specified alpha-level, under the simulation scenarios where association exists between outcome and the zero-inflated covariate. It is computed as follows:

Error Rate 1 (E1) $= \frac{\Sigma_{j=1}^{k} I\ (P-value>\alpha)}{k}$, where k is total number of simulations and p-value corresponds to zero-inflated covariate.

This error rate estimates the effect of zero-inflation in terms of the ability of the regression models to detect (or equivalently miss) the existing underlying association between zero-inflated covariate and the outcome.

Error Rate 2: This error rate is computed based on simulation scenarios 2 and 4. It is the proportion of “significant association between zero-inflated covariate and outcome”, based on pre-specified alpha-level, under the scenarios where there is no association between outcome and the zero-inflated covariate. It is computed, as follows:

Error Rate 2 (E2) $= \frac{I(P-value<\alpha)}{k}$, where k is the total number of simulation and p-value corresponds to zero-inflated covariate.

The error rate 2 essentially estimates the type I error rate. The threshold for α for error rates 1 and 2 was set at 0.05.

Stability of Parameter Estimates: The stability of parameter estimates is computed based on simulation scenarios 1 and 3. It is the deviation of parameter estimate for zero-inflated covariate from its true value. It is calculated, as follows:

Stability $= \frac{\Sigma_{j=1}^{k}\left(\hat{\beta}_{2(j)}-\beta_2\right)^2}{k}$

The stability metric estimates the mean squared error of the estimated parameter.

Variable Selection Rate: The variable selection rate is computed based on simulation scenario 5. It is the rate of selection of “zero-inflated covariate” and is computed, as follows:

Variable Selection Rate $= \frac{\Sigma_{j=1}^{k} I\left(\left|\hat{\beta}_{2(j)}\right|<\lambda\right)}{k}$, where λ is the shrinkage parameter.

The variable selection rate estimates the effect of zero inflation on the ability of LASSO regression model to correctly select the zero-inflated covariate.

3 Analyses of Results

3.1 Error Rates

The relation between error rates and magnitude of zero inflation across different samples sizes are summarized, as Fig 1. As the magnitude of zero inflation increases, the error rate (E1) also increases. However, as the sample size increases, E1 decreases in magnitude. Compared to logistic regression, lower error rates are observed in case of linear regression. The findings from linear regression modeling suggest that E1 converges to 0 faster than in the context of logistic regression modeling. In the context of logistic regression E1 drops to 0 when the sample size ≥ 300 and zero-inflation percentage ≤ 80. The error rates in linear regression drops to 0 when the sample size ≥ 200 and zero-inflation percentage ≤ 80. The results suggest that the likelihood of observing non-significant association when an underlying association exists are slim even with zero-inflation as high as 80%, when the sample size ≥ 200 for linear regression; and ≥ 300 for logistic regression. E2 (type I error rate) was fairly robust to zero-inflated covariate. In case of logistic regression, the value of type I error rates were observed to be within 4.5%-7.5% across all combination of sample size and magnitude of zero-inflation. As the sample size increases, the error rates were observed to be closer to the expected value of 5%. For linear regression, type I error rates were between 4.5%-6.0%. The results suggest that the effect of zero-inflation on type I error rate was minimal.

3.2 Parameter Stability and Variable Selection Rates

The relation between mean square error (log scale) and magnitude of zero inflation across different sample sizes for logistic and linear regression are summarized in Fig 2. The value of log-MSE increased as the magnitude of zero inflation increased. As the sample size increased, the value of log-MSE decreased. The plot of log-MSE vs magnitude of zero inflation tends to flatten as the sample size increases suggesting that as the sample size increases, increase in magnitude of zero inflation tends to have less effect on the stability of parameter estimates. The aforementioned trends are observed in both linear and logistic regression. The variable selection rate estimates the proportion of time zero-inflated covariate is selected in the final model. LASSO algorithm was used in variable selection approach. The shrinkage parameter was chosen based on minimization of mean squared error. The relation between variable selection rate and magnitude of zero inflation across different sample sizes has been summarized in Fig 2. For the sample size of 50, the selection rate for zero-inflated covariate decreases as the magnitude of zero inflation increases in case of logistic as well as linear regression. As the sample size increased, the variable selection percentage improved. For sample sizes ≥ 300, the variable selection rate was approximately 100% even with zero inflation as high as 80% for logistic. The variable selection rate of nearly 100% (>99%) across all magnitude of zero inflation was observed with much smaller sample sizes (≥ 200) in case of linear regression. The results suggest that in the context of variable selection using LASSO, zero inflation rate ≤ 60% is less problematic even with sample sizes as low as 50. For higher magnitude of zero inflation (≥80%), sample sizes ≥ 300 for logistic regression and sample sizes ≥ 200 for linear regression will almost always include zero-inflated covariate in final model.

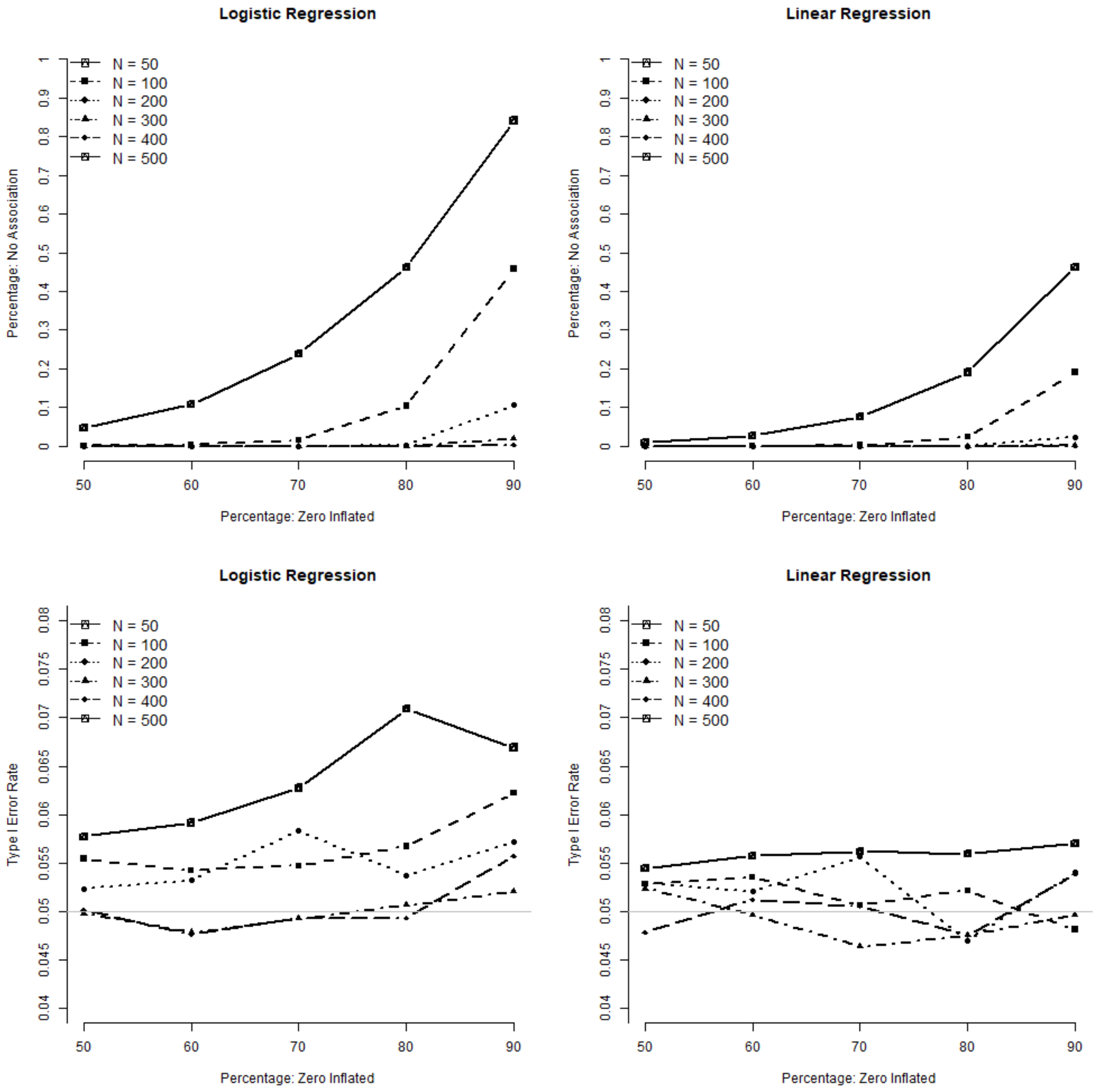

Fig 1: Error Rates.

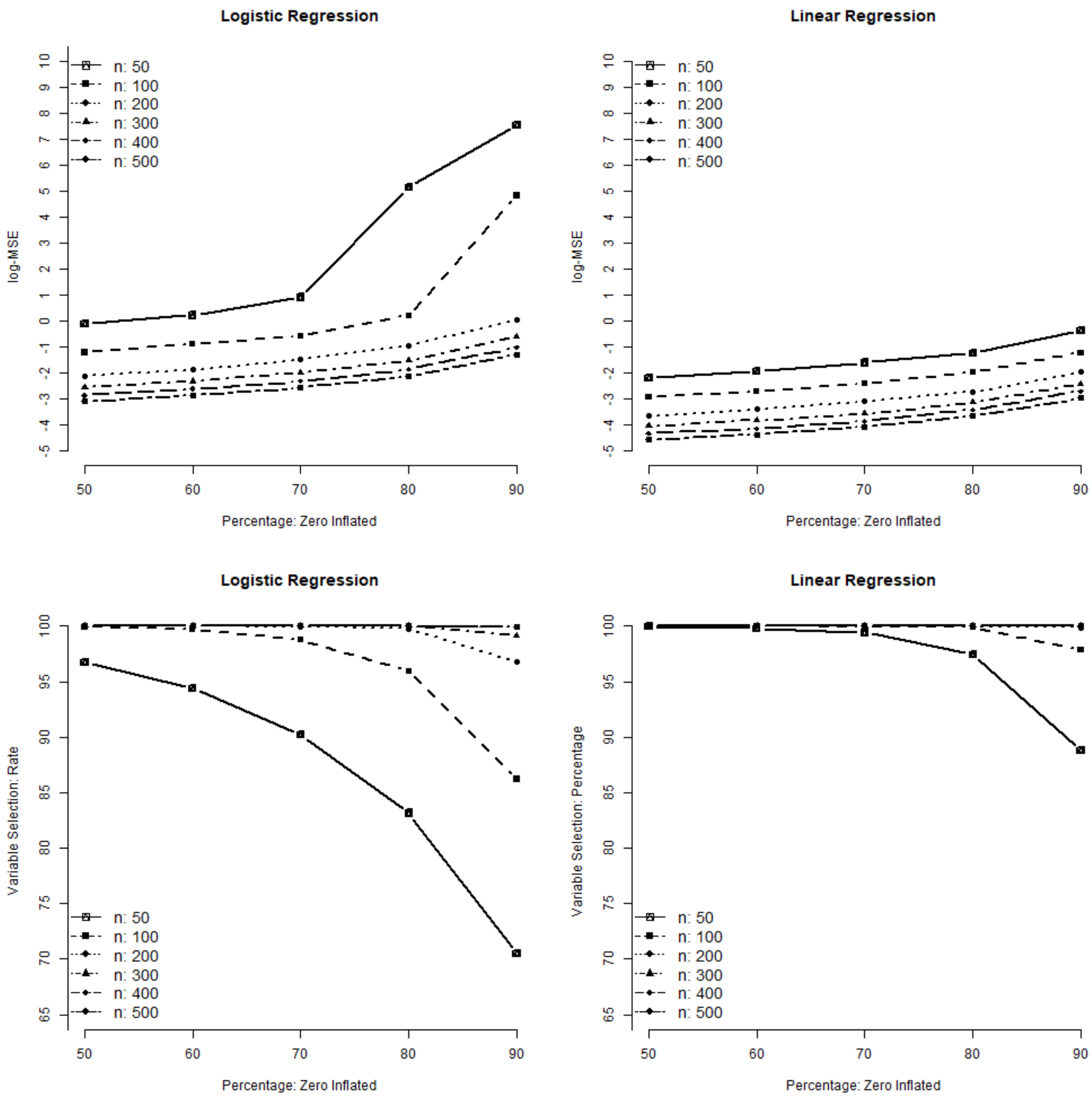

Fig.2: Parameter Stability and Variable Selection.

4 Discussion

Zero-inflated covariates occur in different phases of data collection and subsequent analysis. The effect of zero-inflated covariate in terms of drawing inferences, stability of parameter estimates, and model building may be of interest to practitioners of statistics. Our work has utilized a simulation-based approach to address the effect of zero-inflated covariates across different aspects of inferential statistics and model building. Although we have focused on the inflation of 0’s, results can be extended to scenarios where inflation of values other than 0 can be observed as a simple shift in data (subtracting the inflated value).

We observed that the error rates were well controlled for sample sizes ≥ 300 and 200 (for linear regression). Similarly, the simulation findings based on LASSO modeling showed the inclusion of zero-inflated covariate in the final model >99% of the times for sample sizes ≥ 200 using linear regression for sample sizes ≥ 300 using logistic regression even in scenarios with magnitude of zero inflation as high as 80%. For smaller sample sizes (≤100), the type I error rate was consistently higher across all the magnitudes of zero inflation. Similarly, sample sizes ≤ 100 with zero inflation ≥ 80% demonstrated poor performance in term of error rate and variable selection for linear as well as logistic regression. This could be due to the fact that for large sample even with higher magnitude of zero inflation, the absolute number of remaining non-zero values is large to permit statistical inference. For small sample sizes, even with modest magnitude of zero inflation, the absolute number of non-zero values tend to be fairly low and could lead to collinearity-related issues.

There are limitations associated with this study. Since the data are simulated from known statistical distribution, they may not necessarily represent the actual data obtained in real world setting. This limits the generalizability of the study findings. For smaller sample sizes (≤100) used in this study, the number of successes (events) per covariate is below the commonly recommended threshold of 10-20 events per covariate for logistic regression and variable selection may not be recommended in such settings [12]. In light of the fact that zero-inflated covariates can occur and conventional approaches to statistical inference and model building may not be applicable, we believe that our work has provided some generic sample size guidelines that will be helpful to researchers in employing the commonly used linear and logistic regression models to draw statistical inference and in using LASSO approach for covariate selection. Zero-inflated covariates are less problematic for large sample sizes (≥200 for linear regression and ≥ 300 for logistic regression). However caution must be exercised with smaller sample sizes and alternative modeling approaches that address zero inflation should be considered.

Acknowledgement

The authors are strongly grateful to the anonymous reviewers for their reviews and Department of Biostatistics, University of Arkansas for Medical Sciences for the support.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] M. Yongyi & A. Agresti. Modeling Nonnegative Data with Clumping at Zero: A Survey. JIRSS, 1(1-2), 7-33, (2002).

[2] J. L. Vives. Count data in psychological applied research, Psychological Reports, 98(3), 821–835, (2006).

[3] S. Greenland. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend. Epidemiology, 6(4), 450-454, (1995).

[4] B. Caroline, & V. Andrew. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Medical Research Methodology, 12(21), 1-5, (2012).

[5] J. M. Taylor, & M. Yu,. Bias and Efficiency Loss Due to Categorizing an Explanatory Variable. Journal of Multivariate Analysis, 83, 248-263, (2002).

[6] P. Royston, D. G. Altman, & W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine, 25, 127-141, (2006).

[7] L. Diane. Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing, Technometrics, 34(1), 1-14, (1992).

[8] J. Tobin. Estimation of relationships for limited dependent variables, Econometrica, 26, 24-36, (1958).

[9] M. C. Hu, M. Pavlicova, & E. V. Nunes. Zero-inflated and hurdle models of count data with extra zeros: examples from an HIV-risk reduction intervention trial, The American journal of drug and alcohol abuse., 37(5), 367–375(2011).

[10] K. K. Yau, K. Wang & A. H. Lee. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. Biometrical Journal., 45 (4), 437-452, (2003).

[11] P. Royston, G. Ambler, & W. Sauerbrei. The use of fractional polynomials to model continuous risk factors in epidemiology. International Journal of Epidemiology, 28, 964-974, (1999).

[12] G. Heinze, C. Wallisch, & D. Dunkler. Variable selection – A review and recommendations, Biometrical Journal, 60, 431–449, (2017).