TWO-PART QUANTILE REGRESSION ANALYSIS WITH VARIABLE SELECTION FOR COMPLEX DATA AND ITS APPLICATION

by

Ting CHEN[a] and Min XIAO[b*]

[a]School of Mathematics and Statistics, Xinxiang University, Xinxiang, China
[b]School of Statistics and Mathematics, Zhejiang Gongshang University, Hangzhou, China

Original scientific paper
https://doi.org/10.2298/TSCI2503023C

Semi-continuous data, also known as zero-inflated non-negative continuous data, are commonly observed in various fields such as biomedicine, environmental science, and ecology. Such data exhibit a combination of zero values and positive continuous values that are right-skewed and heteroscedastic. In this study, we present a novel approach for analyzing complex semi-continuous data using a two-part quantile regression method. In addition, we investigate variable selection techniques using least absolute shrinkage and selection operator, smoothly clipped absolute deviation, and minimax concave penalty methods within the framework of two-part quantile regression. Simulation studies are then conducted to evaluate the effectiveness of the proposed methods. Finally, we apply these methods to examine the determinants of health care spending decisions in American households.

Key words: *semi-continuous distribution, two-part model, quantile regression, variable selection*

Introduction

Semi-continuous data is a typical type of complex data often encountered in biology, sociology, and so on. Some examples are alcohol consumption [1] and microbiome data [2]. The field of statistical theory on semi-continuous data has undergone many developments in the past decade, and there is much literature on statistical methods of semi-continuous data, see for example [3-5]. Meanwhile, the statistical analysis and application of semi-continuous data is also progressing steadily. For more details, see [6-8]. For the analysis of semi-continuous data, the two-part model plays an important role and can be extended to regression models by adding predictive factors to each component of the model [4], for example, the Bernoulli-lognormal two-part (mean) regression model. The typical framework of these models assumes that the covariates affect the mean of the conditional response variable distribution. It is well known that parameter estimation in regression analysis is affected by outliers and may be unreliable if the data distribution deviates from normality [9]. In addition, for the two-part model of semi-continuous data, it usually with a lot of predictive factors (explanatory variables), but most of the predictive factors are unrelated variables. In this case, it is necessary to select a subset of predictive factors and estimate them to obtain a simplified model.

* Corresponding author, e-mail: xiaomin90224@163.com

As we all know, the quantile regression model [10] is not limited by data error distribution, which can capture the heterogeneity of the influence of regression coefficients on different parts of the distribution and is robust to outliers. In terms of application, the quantile regression model has been applied in many different fields, both in frequency theory paradigms and in Bayesian theory paradigms, such as medicine [11], financial and economic studies [12, 13], and environmental modeling [14]. In addition, numerous variable selection techniques have been proposed, such as least absolute shrinkage and selection operator (LASSO), [15], smoothly clipped absolute deviation (SCAD) [16], and minimax concave penalty (MCP) [17].

This paper proposes a two-part quantile analysis method using quantile regression and variable selection methods to investigate the heterogeneity and model simplification challenges in two-part regression models for semi-continuous data. The method allows for both the heterogeneity and right skewness of the data and the complexity of the two-part model. Specifically, a mixed-effects logistic quantile regression model is proposed that includes: a logistic regression model with a non-zero outcome probability and a quantile regression model with a continuous positive outcome. In addition, variable selection based on LASSO, SCAD, and MCP penalty methods for the proposed model is further investigated.

Two-part quantile regression model and variable selection

For the semi-continuous data $\{Y_1, \cdots, Y_n\}$, let $Y_i^+ = \{Y_i \mid Y_i > 0\}$ denote the positive part of $\{Y_1, \cdots, Y_n\}$ and the probability of $Y_i, i = 1, \ldots, n$ being non-zero is π_i, that is $P(Y_i^+) = \pi_i$ and $1 - \pi_i = \Pr(Y_i = 0)$. Then, the quantile regression model for semi-continuous data in two-parts is expressed in the following manner.

$$\begin{aligned} g(\pi_i) = g[\Pr(Y_i > 0)] = X_{1i}^T \boldsymbol{\beta}_1, \quad i = 1, \ldots, n \\ Y_i^+ = Y_i \mid Y_i > 0 = \exp(X_{2i}^T \boldsymbol{\beta}_\tau) + \varepsilon_\tau \end{aligned} \tag{1}$$

where $g(\pi_i) = \ln[\ln(\pi_i)/(1 - \pi_i)]$, $X_{1i} = \{X_{1i0}, X_{1i1}, X_{1i2}, \ldots, X_{1ip}\}^T$ is p+1-D explanatory variable and $\boldsymbol{\beta}_1 = \{\beta_{1,0}, \beta_{1,1}, \ldots, \beta_{1,p}\}^T$ is the corresponding p+1-D parameter. The $X_{2i} = \{X_{2i0}, X_{2i1}, X_{2i2}, \ldots, X_{2iq}\}^T$ is q+1-D explanatory variable and $\boldsymbol{\beta}_\tau = \{\beta_{\tau,0}, \beta_{\tau,1}, \ldots, \beta_{\tau,q}\}^T$ is the corresponding q+1-D parameter under a given quantile $\tau \in (0,1)$, ε_τ is a random error. Note that logarithmic transformation is performed on the response variables to maintain the linearity of the regression model coefficients.

Next, the variable selection method of two-quantile regression models under LASSO, SCAD and MCP penalties will be studied. For the first part (binary part) of the model (1), each Y_i, $i = 1, \ldots, n$ of semi-continuous data $\{Y_1, \cdots, Y_n\}$ is 0 or 1 , and $\pi_i = P(Y_i = 1 \mid X_{1i})$, then $1 - \pi_i = P(Y_i = 0 \mid X_{1i})$. For n observation $\{Y_1, \cdots, Y_n\}$, the corresponding log likelihood function is:

$$L(\boldsymbol{\beta}_1) = \ln[l(\boldsymbol{\beta}_1)] = \sum_{i=1}^{n} y_i \times \ln(\pi_i) + (1 - y_i) \times \ln(1 - \pi_i) \tag{2}$$

The parameter $\boldsymbol{\beta}_1$ will be estimated by LASSO, CAD and MCP method, respectively. First of all, as defined by the LASSO, SCAD and MCP methods, $\boldsymbol{\beta}_1$ can be obtained by minimizing the following objective function:

$$Q_{L\lambda_1}(\boldsymbol{\beta}_1) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \times \ln(\pi_i) + (1-y_i)\times \ln(1-\pi_i)\} + \sum_{j=0}^{p} P_{\lambda_{1L}}\left(\left|\beta_{1j}\right|\right) \tag{3}$$

$$Q_{\lambda_{1S,M},\gamma_{1S,M}}(\boldsymbol{\beta}_1) = -\frac{1}{n}\sum_{i=1}^{n}\{y_i \times \ln(\pi_i) + (1-y_i)\times \ln(1-\pi_i)\} + \sum_{j=0}^{p} P_{\lambda_{1S,M}\gamma_{1S,M}}\left(\left|\beta_{1j}\right|\right) \tag{4}$$

where $P_{\lambda 1,L}(|\beta_{1j}|)$ is the penalty function of LASSO, $P_{\lambda 1_{S,M},\gamma 1_{S,L}}(|\beta_{1j}|)$ – the penalty function of SCAD or MCP method, λ_{1_L}, $\lambda_{1_{S,M}}$, and $\gamma_{1_{S,M}}$ are the hyperparameters for LASSO, CAD and MCP methods, respectively. For the solution of (3) and (4), the coordinate descent (CD) [18] algorithm to be used to obtain the optimal parameter estimation.

For the second part (quantile part) of model (1), the second part of the punishment parameters estimation rely mainly on $Y_i^+ = \{Y_i \mid Y_i > 0\}$, $I = 0, 1, \ldots, n^+$. According to the quantile loss function, the estimation $\boldsymbol{\beta}_\tau$ in LASSO SCAD and MCP penalty methods, the objective function can be expressed:

$$Q_{\lambda_{\lambda_L}}(\boldsymbol{\beta}_\tau) = \frac{1}{n^+}\sum_{i=1}^{n^+}\rho_\tau\left(\ln Y_i^+ - X_{2i}^T\boldsymbol{\beta}_\tau\right) + \sum_{j=0}^{q} P_{\lambda_{\tau_L}}\left(\left|\beta_{\tau,j}\right|\right) \tag{5}$$

$$Q_{\lambda\tau_{S,M}}(\boldsymbol{\beta}_{\tau_{S,M}}) = \frac{1}{n^+}\sum_{i=1}^{n^+}\rho_\tau\left(\ln Y_i^+ - X_{2i}^T\boldsymbol{\beta}_\tau\right) + \sum_{j=0}^{p} P_{\lambda_{\tau_{S,M}},\gamma_{\tau_{S,M}}}\left(\left|\beta_{\tau,j}\right|\right) \tag{6}$$

Similarly, tuning parameters λ_{τ_L}, $\lambda_{\tau_{S,M}}$, and $\gamma_{\tau_{S,M}}$ in the penalty function control the complexity of the model. In addition, for the values of $\gamma_{1_{S,M}}$ and $\gamma_{\tau_{S,M}}$, we refer to $\gamma_{1_{S,M}} = \gamma_{\tau_{S,M}} = 3.7$ suggested in [16]. For the solution of the objective function (5) and (6), the local linear approximation algorithm (LLA) was used for the parameter estimation [19]. In addition, the λ_{1_L}, $\lambda_{1_{S,M}}$, λ_{τ_L}, and $\lambda_{\tau_{S,M}}$ are chooed by using the 10-fold cross-validation to select variables.

Simulation

In the setting of simulation parameters, for the setting of the number of parameters, assumption $p = q = 17$ without losing the general. In addition, five quartiles were selected, which are 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. For the real value of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_\tau$, randomly set it:

$$\boldsymbol{\beta}_1 = (0.2, 0.4, 0.3, 0.4, 0.5, -0.6, 0.8, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\boldsymbol{\beta}_\tau = (0.5, 0.8, 0.7, -0.4, 0.6, 0.5, 0.7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

Assume that the covariates $\boldsymbol{X}_1 = \{X_{11}, X_{12}, \ldots, X_{1n}\}$ and $\boldsymbol{X}_2 = \{X_{21}, X_{22}, \ldots, X_{2m}\}$ as the first part and second part covariate, respectively. Suppose that $\boldsymbol{X}_1$ comes from the multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{X}_1 = \boldsymbol{X}_2$, where the elements of $\boldsymbol{\Sigma}$ are $\rho^{|i-j|}$, $i, j = 0, \ldots, p$, $\rho = 0, 0.5$. The $\rho = 0$ indicates that covariates are not correlated and $\rho = 0.5$ indicates that covariates are correlated to a certain extent.

The generation of response variables should be divided into two steps, the first step is to generate the mixture of 0 and non-0 data, and the proportion of 0 is:

$$\pi_i = \frac{\exp\left(X_{1i}^T\boldsymbol{\beta}_1\right)}{1+\exp\left(X_{1i}^T\boldsymbol{\beta}_1\right)}$$

The second step generates non-zero data, assuming $Y_i \mid Y_i > 0$ $\exp({X_{2i}}^T \boldsymbol{\beta}_\tau + \varepsilon_\tau)$. Depending on ε_τ, the corresponding response variables will follow different distributions, in this paper, we make $\varepsilon_\tau \sim N(\Phi^{-1}(\tau = 0),1)$ or $\varepsilon_\tau \sim t(3)$. Then the final response variable is $y_i = \text{Binomial}(\pi_i) \times [\exp({X_{2i}}^T \boldsymbol{\beta}_\tau + \varepsilon_{\boldsymbol{\tau}})], i = \cdots, n$. In the simulation analysis, the sample size to n = 150 for comparison. All simulations were conducted 500 times, and the displayed results are the average of these 500 results.

In order to evaluate the proposed model in terms of estimation error and model selection ability, three measures are given:

– Mean absolute Bias: $\text{Bias} = \frac{1}{p+1}\sum_{j=0}^{p}\left|\beta_j - \hat{\beta}_j\right|$.

– Mean square error: $\text{MSE} = \frac{1}{p+1}\sum_{j=0}^{p}\left(\beta_j - \hat{\beta}_j\right)^2$.

– Accuracy: $\text{Accuracy}\ (\boldsymbol{\beta}) = \frac{\#\{j:\beta_j \neq 0 \& \hat{\beta}_j \neq 0\} + \#\{j:\beta_j = 0 \& \hat{\beta}_j = 0\}}{p}$

where β_i is the true value and $\hat{\beta}_i$ is the predicted value of the respons. The Accuracy is defined based on the correct selection of the proportion of non 0 and 0 coefficients in the model: $\#\{j : A\}$ is the number of j satisfying A and p is the number of parameters.

Table 1. Results of normal distribution $p = q = 17$

n = 150		AC	Bias ρ = 0	MSE	AC	Bias ρ = 0.5	MSE
Binomial part	logit	0.7116	0.1337	0.0518	0.7448	0.1813	0.0911
Positive part	τ:10%	0.6522	0.1379	0.0475	0.6608	0.1620	0.0670
	τ:30%	0.6398	0.113	0.0295	0.6555	0.1348	0.0454
LASSO	τ:50%	0.6312	0.1061	0.0305	0.6463	0.1286	0.0415
	τ:70%	0.6476	0.1128	0.0304	0.6400	0.1367	0.0461
	τ:90%	0.6551	0.1378	0.052	0.6545	0.1677	0.0732
Binomial part	logit	0.7998	0.1371	0.0512	0.796	0.1751	0.0874
Positive part	τ:10%	0.7073	0.1347	0.0462	0.6792	0.1632	0.0659
	τ:30%	0.7194	0.1096	0.0300	0.7147	0.1342	0.0455
SCAD	τ:50%	0.7075	0.1019	0.0277	0.7043	0.1295	0.042
	τ:70%	0.7363	0.1110	0.0288	0.6947	0.1365	0.0459
	τ:90%	0.7004	0.1404	0.0506	0.6872	0.1677	0.0726
Binomial part	logit	0.8404	0.1277	0.0372	0.7828	0.1432	0.0506
Positive part	τ:10%	0.7014	0.1336	0.0451	0.6888	0.1537	0.0573
	τ:30%	0.7259	0.1092	0.0272	0.7152	0.1304	0.0394
MCP	τ:50%	0.7195	0.1017	0.0292	0.7122	0.1271	0.0367
	τ:70%	0.7335	0.1092	0.0272	0.7011	0.1346	0.0408
	τ:90%	0.7046	0.1385	0.0461	0.6884	0.1595	0.0636

Under the previous data parameter setting, the results of normal distributions are shown in tab. 1, the results of the *t*-distribution *t*(3) are shown in tab. 2, respectively. As can be seen from tabs. 1 and 2 that LASSO, SCAD, and MCP penalty methods performed comparably in term of three criterions for different quartiles. Comparatively, the MCP penalty method had the least bias, the highest degree of accuracy and the smallest mean square error. The SCAD method performed between the LASSO and MCP penalty methods. Overall, all three methods are better and in practical applications, these three methods can be used selectively.

Table 2. Results of *t* (3) *p* = *q* = 17

n = 150		AC ρ = 0	Bias	MSE	AC ρ = 0.5	Bias	MSE
Binomial part	logit	0.7032	0.1327	0.0512	0.7407	0.1828	0.0930
Positive part	τ:10%	0.6260	0.1939	0.0939	0.6365	0.2098	0.1128
	τ:30%	0.6300	0.246	0.0461	0.6366	0.1636	0.0657
LASSO	τ:50%	0.6225	0.1253	0.0386	0.6565	0.1480	0.0549
	τ:70%	0.638	0.1358	0.0464	0.6475	0.1704	0.0721
	τ:90%	0.6180	0.2133	0.1206	0.6300	0.2285	0.1385
Binomial part	logit	0.7895	0.093	0.0514	0.7836	0.1761	0.0856
Positive part	τ:10%	0.6387	0.1905	0.0921	0.6322	0.2117	0.1146
	τ:30%	0.6956	0.2137	0.0467	0.6889	0.1641	0.0664
SCAD	τ:50%	0.6898	0.1249	0.0382	0.7104	0.1448	0.0535
	τ:70%	0.7098	0.1351	0.0467	0.6726	0.1715	0.0728
	τ:90%	0.6253	0.2135	0.1201	0.6333	0.2298	0.1385
Binomial part	logit	0.8273	0.1191	0.0385	0.7852	0.1441	0.0512
Positive part	τ:10%	0.6499	0.1798	0.0778	0.6402	0.1926	0.0923
	τ:30%	0.6951	0.1568	0.0400	0.6871	0.1530	0.0537
MCP	τ:50%	0.6969	0.1247	0.0333	0.7049	0.1387	0.0441
	τ:70%	0.7047	0.1322	0.0393	0.6806	0.1573	0.0586
	τ:90%	0.6345	0.1997	0.1030	0.6393	0.2098	0.1176

Application

The RNHIE datasets in the RAND health insurance experiment are typical illustrations of semi-continuous data, as they contain a large proportion of zeros, with the non-negative portion having a pronounced right skew and heavy tail, and often exhibit a combination of different distribution shapes. Maruotti *et al.* [20] and other researchers have also examined the RAND health insurance experiment (RHIE) dataset to evaluate the impact of healthcare spending on patient utilization and quality of care. The information examined in this study represents the annual average for each person's insurance coverage. In addition, total medical expenditures (MED) include expenditures for outpatient visits, hospitalizations, prescription drugs, medical supplies, and mental health counseling [20]. Meanwhile, the predictor variables considered are consistent with those in the research [20].

Using the two-part quantile regression model with the MCP penalty method, the effects of covariates in the RHIE on whether U.S. households spend on health care and how covariates affect U.S. households health care spending at different quantile levels are examined. Table 3 reports the estimates of the bivariate partial coefficients, the estimates of the positive partial coefficients, and the absolute mean deviation of the penalty estimates at different quantile levels.

It can be seen from tab. 3 that the binary part of a two-part quantile regression model distinguishes two groups of individuals who are infrequent and frequent users of medical services. In addition, the influence of some covariate factors on each quantile is not consistent, and the sign and amplitude change with the change of quantile level. Considering the quantile aspect of a two-part model for quantiles, it is evident that not all factors affect medical utilization in a uniform manner. Variability exists in both direction and magnitude, with certain factors having no discernible effect at different quantile levels. This underscores the importance of using quantile techniques. In addition, the MCP penalty method effectively identifies crucial variables and reduce the complexity of the model.

Table 3. Parameter estimation (MCP)

Covariate	τ^{th} quantile					
	Binary	10	25	50	75	90
(Intercept)	2.347	1.274	2.389	3.408	4.222	5.443
LOGC	0.230	–0.144	–0.118	–0.107	–0.073	–0.036
LFAM		–0.083	–0.099	–0.132	–0.103	
LINC		0.084	0.111	0.102	0.093	0.067
XAGE		0.009	0.007			0.010
FEMALE	0.792	0.691	0.497	0.612	0.726	0.437
CHILD			–0.109	–0.332	–0.406	–0.293
FEMCHILD	–0.781	–0.742	–0.631	–0.739	–0.870	–0.563
BLACK	–1.668	–0.633	–0.576	–0.587	–0.426	–0.267
EDUCDEC	0.090	0.034	0.026			–0.031
PHYSLM	0.323	0.488	0.438	0.514	0.499	0.625
DISEA	0.005	0.032	0.022	0.021	0.017	0.004
HLTHG		–0.071		0.108	0.159	0.244
HLTHF		0.211	0.419	0.440	0.298	0.484
HLTHP	2.453	0.458	0.784	0.809	0.883	0.900
MHI		–0.004	–0.005	–0.002	–0.003	–0.003
PBISA	0.078	1.804	1.190	1.056	1.202	1.808

Conclusion

In this paper, we studied the complex semi-continuous data characterized by numerous zero values and right-skewed non-negative components. A two-part quantile regression

model was constructed to analyze the semi-continuous data. Meanwhile, LASSO, SCAD, and MCP penalty methods were introduced for variable selection in the two-part quantile regression model. Then, simulations were conducted to compare three penalty methods, and it was found that the MCP penalty method performed better than the LASSO and SCAD penalty methods. Finally, the applicability of the two-part quantile regression with variable selection method for semi-continuous data proposed in this paper was also verified by analyzing the RHIE sample.

Acknowledgment

This paper is supported by 2023 Henan Provincial High-level Talents Internationalization Cultivation Funding Project and Doctoral Scientific Research Start-up Project Fund of Xinxiang University.

Conflicts of interest

The authors declare no conflict of interest or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Han, D., *et al.*, Variable Selection for Random Effects Two-Part Models, *Statistical Methods in Medical Research*, *28* (2018), 9, pp. 2697-2709
[2] Chai, H., *et al.*, A Marginalized Two-Part Beta Regression Model for Microbiome Compositional Data. *PLOS Computational Biology*, *14* (2018), 7, pp. 1-16
[3] Neelon, B., *et al.*, Zero-Modified Count and Semicontinuous Data in Health Services Research Part 1: Background and Overview, *Statistics in Medicine*, *35* (2016), 27, pp. 5070-5093
[4] Neelon, B., *et al.*, Modeling Zero Modified Count and Semicontinuous Data in Health Services Research Part 2: Case Studies, *Statistics in Medicine*, *35* (2016), 27, pp. 5094-5112
[5] Liu, L., *et al.*, Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review, *Statistical Science*, *34* (2019), 2, pp. 253-279
[6] HCC de Souza., *et al.*, A Bayesian Approach for the Zero-Inflated Cure Model: An Application in A Brazilian Invasive Cervical Cancer Database, *Journal of Applied Statistics*, *49* (2022), 12, pp. 3178-3194
[7] Rustand, D., *et al.*, Bayesian Estimation of Two-Part Joint Models for A Longitudinal Semicontinuous Biomarker and A Terminal Event with R-INLA: Interests for Cancer Clinical Trial Evaluation, *Biometrical Journal*, *65* (2023), 4, pp. 1-22
[8] Lin, P., *et al*., Disease Mapping for Spatially Semi-Continuous Data by Estimating Equations with Application to Dengue Control, *Statistics in Medicine*, *42* (2023), 20, pp. 3636-3648
[9] Manning, W. G., Basu, A., Estimating Lifetime or Episode-of-Illness Costs under Censoring, *Social Science Electronic Publishing*, *19* (2020), 9, pp. 1010-1028
[10] Koenker, R., *Quantile Regression*, Cambridge University Press, New York, USA, 2000
[11] Waldmann, E., Quantile Regression: A Short Story on How and Why, *Statistical Modeling: Applications in Contemporary Issues*, *18* (2018), 3-4, pp. 203-218
[12] Bernardi, M., *et al.*, Bayesian Quantile Regression using the Skew Exponential Power Distribution, *Computational Statistics and Data Analysis*, *126* (2018), Oct., pp. 92-111
[13] Petrella, L., Raponi, V., Joint Estimation of Conditional Quantiles in Multivariate Linear Regression Models with An Application to Financial Distress, *Journal of Multivariate Analysis*, *173* (2019), Sept., pp. 70-84
[14] Reich, B. J., *et al.*, Bayesian Spatial Quantile Regression. *Journal of the American Statistical Association*, *106* (2011), 493, pp. 6-20
[15] Tibshirani, R., Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society, Series B*, *58* (1996), 1, pp. 267-288
[16] Zhang, C. H., Nearly Unbiased Variable Selection under Minimax Concave Penalty, *Annals of Statistics*, *38* (2010), 2, pp. 894-942
[17] Chowdhury, S., *et al.*, Group Regularization for Zero-Inflated Poisson Regression Models with An Application to Insurance Ratemaking, *Journal of Applied Statistics*, *46* (2018), 9, pp. 1567-1581

[18] Simon, N., *et al.*, Regularization Paths for Cox's Proportional Hazards Model *via* Coordinate Descent, *Journal of Statistical Software*, *39* (2011), 5, pp. 1-13
[19] Wang, L., *et al.*, Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension, *Journal of the American Statistical Association*, *107* (2012), 497, pp. 214-222
[20] Maruotti, A., *et al.*, A Two-Part Finite Mixture Quantile Regression Model for Semi-Continuous Longitudinal Data, *Statistical Modelling*, *22* (2020), 6, pp. 485-908

Paper submitted: March 1, 2024
Paper revised: July 7, 2024
Paper accepted: July 7, 2024