Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/110129

Performance Comparison of Residual Control Charts for a Count Data Based on Ridge Regression

Shaimaa Mohamed Yassin * *and Salah M. Mohamed*

Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, 12613, Cairo University, Egypt

Received: 21 Feb. 2021, Revised: 22 June 2021, Accepted: 26 Oct. 2021
Published online: 1 Jan. 2022

Abstract: Regression models are important and widely used tools to design relationships between pairs of endogenous and exogenous variables. However, they usually suffer from multicollinearity. This study uses residual control charts on count data *(Poisson regression)*, after resolve multicollinearity problems by utilizing the ridge technique. Residual control charts with Poisson regression were introduced by Filho [1], who employed the principal component to treat the multicollinearity problem and, thereafter, prepared a control chart. However, we selected average run length as the metric for the evaluation of the control chart. We used simulated data and application using real data on water quality. The results corresponding to two types of residual values are consistent after being processed by the ridge method, and the joint-charts show the samples to be at out-of-control limits for count data.

Keywords: Count Data, Ridge Regression, Residual Control Chart , Average Run Length, Water Quality

1 Introduction

Zhou [2] and several other researchers reported the successful utilization of Residual Control Charts (RCCs) to monitor statistical processes or procedures of multiple operations. The monitoring performance was observed to be affected by the selection of the model fitting scheme. Areepong [3] used average run length (ARL) as a metric to evaluate certain RCCs used in industries, such as Shewhart $\bar{x}$ regression RCCs to evaluate student performance. Filho [1] performed a simulation and a real study on the Poisson regression model, which suffers from a multicollinearity problem, and investigated its principal component method. Yu and Liu [4] used logistic regression to construct a novel RCC. Roy [5] explained the definition of water quality and explained its use types, such as drinking, swimming, farming, or manufacturing. Furthermore, each of these specified uses has various established chemical, physical, and biological requirements, which are required to fulfill their respective purpose. For example, water used for drinking or swimming is subject to more strict requirements than that used in agriculture or industry. Water quality is impacted by the presence of algae, of which two types are relevant: those that are poisonous and those that are beneficial to the living organisms in water. We use data from the Holding Company for Water and Wastewater in Egypt to examine the effect of selected properties on the formation of algae, from which there have negative or positive effect on the drinking water. Therefore, we used the data for the water after treating it to observe if there was an effect or not.

This study uses RCCs on count data *(Poisson regression model)* after resolve their multicollinearity problems by utilizing the ridge technique. ARL is selected as the metric for the evaluation of the RCCs. Simulated data and application data using real data on water quality.

The remainder of this study is structured as follows. We investigate the Poisson regression model in Section 2. In Section 3, we introduce ridge regression. In Section 4, the ridge estimator formula and the RCC is discussed. The control charts of the Poisson model are presented in Section 5. In Section 6, we introduce ARL. In Section 7, we conduct simulation studies for count data for Poisson regression model. In Section 8, we introduce a case study using real data as count data *(Poisson regression model)*. An algorithmic approach to basic programs for generating models is also discussed in this section. Finally, the study is concluded in Section 9.

* Corresponding author e-mail: shaimaayassin917@yahoo.com

2 The Poisson Regression Model

It is assumed that an observed endogenous variable. It is characterized by the mean expected value [6]. The Poisson regression model is given by the following.

$$\log_e(\mu_i) = \varphi_0 + \varphi_1 y_{i1} + \varphi_2 y_{i2} + ... + \varphi_p y_{ip}. \quad (1)$$

Therefore, the Poisson Probability Function is given by

$$f(X = x) = \lambda^x \frac{e^- \lambda}{x!}, \quad (2)$$

where $f(X = x)$ denotes the probability, and $x! = x(x-1)\ldots 3.2.1$ denote the endogenous variable (for further details, please refer to [1]).

The two types of residuals used in this study are ordinary raw residuals and Pearson residuals introduced by Myers and Dobson [7,8].

The ordinary raw residuals are given by

$$r_p^* = x - \hat{\mu}, \quad (3)$$

and the Pearson residuals are given by

$$r_p^* = \frac{x - \hat{\mu}}{\sqrt{\hat{\mu}}}, \quad (4)$$

Where the residual r^* is asymptotical normality distributed with $(\mu = 0, \sigma^2 = 1)$ and independent (for further details, please consult [7]). In the aforementioned case, Famoye [9] defined the test of the null hypothesis to be $H_0 : \varphi_i = 0$ against an alternative hypothesis,$H_1 : \varphi_i \neq 0$.

3 Ridge Regression

Ridge regression is an important method to solve multicollinearity problems in count data models. In this context, multiple ridge regression is given by

$$\hat{\varphi}_{rr} = (y'y + kI)^{-1} y'x, \quad (5)$$

where $k \geq 0$ and, I denotes the identity matrix, and k denotes a positive integer known as the ridge parameter.

According to [10], the Poisson ridge regression is given as follows.

$$\hat{\varphi}_{prr} = (y'\hat{W}^p y + kI)^{-1} y'\, \hat{W}^p y \hat{\varphi}_{pML}, \qquad k \geq 0 \quad (6)$$

And the $\hat{\varphi}$ Maximum likelihood is obtained by

$$\hat{\varphi}_{pML} = (y'\hat{W}^p y + kI)^{-1} y'\hat{W}^p\, \hat{z}^p, \quad (7)$$

where y denotes exogenous variables,including an $n*(p+1)$ matrix, $\hat{W}^p = diag[\hat{\mu}_i]$ a vector, $\hat{z}^p = \log(\hat{\mu}_i) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$ is a vector; $\hat{\mu}_i = e^{y\hat{\varphi}}$ (for further details, please consult [11].

4 Methodology

4.1 The Ridge Parameter

Ridge regression is a popular method to solve multicollinearity problems in regression model.

Ridge Parameter Formulae for Poisson regression model

Zaldivar [10] proposed the following k-ridge regression estimators.

$$k_1 = \frac{p}{\sum_{i=1}^n \left(\frac{\hat{\tau}_i^2}{1+\left(1+\lambda_i \sqrt{\hat{\tau}_i^2}\right)} \right)}, \quad (8)$$

$$k_2 = \frac{2p}{\sum_{i=1}^p \lambda_{max} \hat{\tau}_i^2} \quad (9)$$

$$k_3 = max \left(\sqrt{\frac{(n-p) + \lambda_{max} \hat{\tau}_i^2}{\lambda_{max}}} \right) \quad (10)$$

where $\hat{\tau}_i^2 = (Y\hat{\varphi_{ML}})^2$ and λ_i denote the eigenvalues of $y'\hat{w}^1 y$.

4.2 Residual Control Chart

RCCs are effective tools for statistical control of multiphase processes or products. Souza [12] demonstrated that selection of appropriate forecasting models greatly affects the capabilities of RCCs in monitoring the stability of manufacturing variables using a single chart to simultaneously verify mean and variance.

5 Control Chart for Poisson Regression Model

Shewhart charts are used to detect large deviations in processes or products; therefore, Shewhart $\bar{x}$ and $\bar{R}$ charts that are biased to control charts are very simple and convenient. Given $r^* \sim N(0,1)$, where r is defined as residuals, and the Shewhart control limits of the residuals are given by

$$CL_{r^*} = E(r_n^*) \pm \omega \sqrt{Var(r_n^*)} \simeq \pm\omega, \quad (11)$$

where the constant ω was defined by Filho [1] to be the amplitude between the control limits based on the probability of the false alarm τ. Duttadeka and Gogoi [13] further clarified the properties of the control chart.

6 Average Run Length

ARL is a common metric used to evaluate control charts. Prajapat [14] defined ARL to be the average number of points that draw on the control chart, while an out-of-control signal is acquired.

1. If the control charts are in control, then ARL is given by

$$ARL_0 = \frac{1}{\hat{\tau}}.$$

2. If the control charts are out-of-control, then ARL is given by

$$ARL_1 = \frac{1}{1-\hat{\varphi}}.$$

Where $\hat{\tau}$ denotes the false alarm probability (type I error) and $\hat{\varphi}$ denotes the true alarm probability (type II error) [15].

7 Simulation Study

We use the ridge regression technique to address the multicollinearity problem in the Poisson Regression Model. Further, we utilize RCC to evaluate the performance of the k-estimator under two varying phases (Phase I and Phase II) using the R program. Therefore, we generate data onto the multicollinearity problem, where the sample size N is taken to be 100. We obtain the Ridge regression parameter and utilize two types of residuals—ordinary raw residuals and Pearson residuals. We divide the entire sample into 25 samples, calculate the mean corresponding to each sample of size $n = 4$, and construct the Shewhart Control Chart for residuals. Table (1) presents the values of the ridge estimators and those of the parameters for the poison Ridge Regression. Table (2) and Table (3) present Control limits, ARL_o ,and ARL_1, with respect to ordinary raw residuals and Pearson residuals, respectively.

The following algorithm is used to generate data:

1.Set $N = 100$ and $p = 4$, and assume that the generating random variable, z, follows the standard normal distribution.

2.Generate an exogenous variable Y using the equations, $y_{ij} = (1-\rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip}$, where $i = 1,2,\ldots,n, j = 1,2,\ldots,p$ and $\rho = 0.85$.

3.Choose φ satisfying $\sum_{i=1}^{p} \varphi_j^2 = 1$ and $\varphi_0 = 1.5$.

4.The generate $po(\mu_i)$, where μ_i is given by the set of equations,$\mu_i = \exp(\varphi_0 + \varphi_1 y_{i1} + \ldots.. + \varphi_p y_{ip})$.

5.Generate x following the Poisson regression model.

Analysis of the simulation study

The control chart is not stable and subgroups of our sample from the simulation are out-of-control. Figures (1) and (2) depict Phase I and Phase II control charts corresponding to two types of residuals of k_1 (ordinary raw and Pearson). However, there are several differences between the cases of ordinary raw residuals and Pearson residuals in terms of the out-of-control-limit samples.
For example, in the case of ordinary raw residuals of k_1, Samples 3 and 20 lie outside the control limits, but in the case of Pearson residuals, Sample 11 lies outside the control limits, as evidenced by the corresponding Phase I Control Chart. Sample 11 and Sample 7 are observed to lie outside the control limits of k_2 in Figure (3) and Figure (4), respectively. However, as depicted in Figure (5), Sample 3 and Sample 20 lie outside the control limit in the case of ordinary raw residuals, and Figure (6) bears a similar result to Figure (2), with Sample 11 lying outside the control limit.

We remove all of the aforementioned samples that lie outside the control limit from the respective charts and construct the Phase II charts using the remaining data. Table (2) and Table (3) present the values of ARL corresponding to the control charts using ordinary raw and Pearson residuals. As high quality charts correspond to high ARL_0 values and low ARL_1 values, the optimal values of k-estimators in the case of ordinary raw residuals are k_1 = 434.8 with respect to ARL_0 and k_1 = 1.001 with respect to ARL_1, as tabulated in Table (2). Similarly, the optimal values of k in the case of Pearson residuals is $k_1 = 621$ with respect to ARL_0 and $k_1 = 1.002$ and $k_2 = 368$ with respect to ARL_1.

Table 1: Values of the k-estimator and Poisson ridge parameters

Ridge Parameter	Value	φ_0	φ_1	φ_2	φ_3	φ_4
k_1	13.74	1.40	−0.53	−0.41	−0.42	−0.66
k_2	.0001	1.44	−0.46	−0.29	−0.32	−0.89
k_3	9.92	1.41	−0.53	−0.40	−0.41	−0.68

Table 2: Control limits, ARL_0, and ARL_1 with respect to ordinary raw residuals

			Phase I		Phase II			
k	m	n	LCL	UCL	LCL	UCL	ARL_0	ARL_1
k_1	25	4	−1.9	−0.36	−1.93	−0.29	434.8	1.001
k_2	25	4	−6.8	2.6	−12.53	5.42	1.6	3.75
k_3	25	4	−3.81	1.26	−5.4	1.3	364	398

Table 3: Control limits, ARL_0, and ARL_1 with respect to Pearson residuals

			Phase I		Phase II			
k	m	n	LCL	UCL	LCL	UCL	ARL_0	ARL_1
k_1	25	4	−2.8	1.3	−3.06	2.08	621	1.002
k_2	25	4	−1.9	0.4	−2.04	0.98	352	368
k_3	25	4	−2.9	1.6	−3.38	2.8	369	383

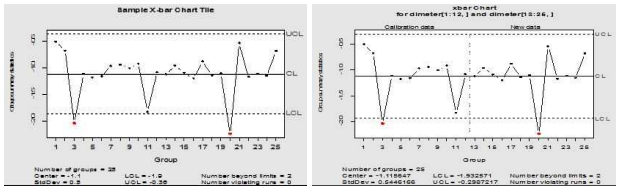

Fig. 1: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_1

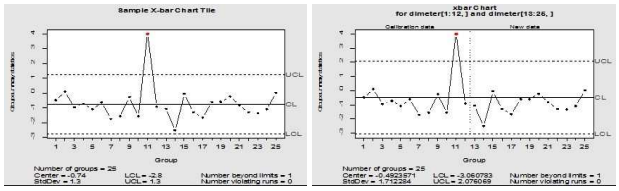

Fig. 2: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_1

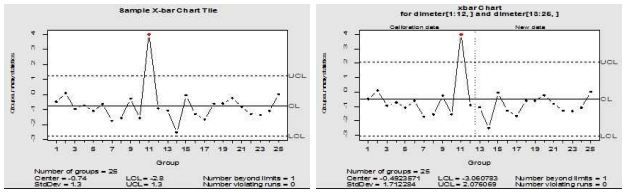

Fig. 3: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_2

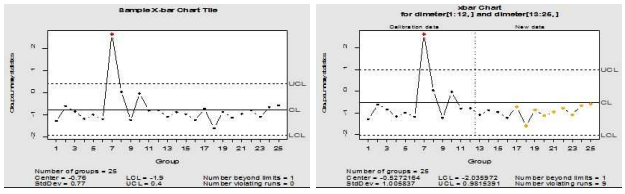

Fig. 4: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_2

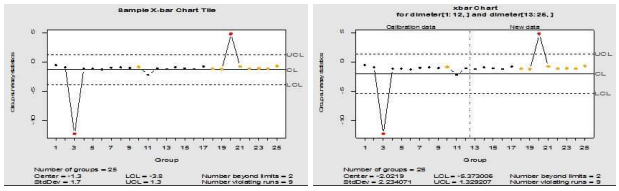

Fig. 5: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_3

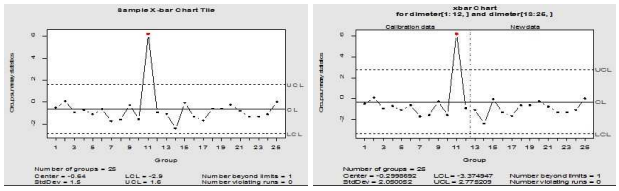

Fig. 6: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_3

8 Application Based on Real Data

In this section, we introduce a case study, explaining the techniques used to monitor five properties of the Holding Company for Water and Waste Water. Further, the endogenous variable is taken to be the total number of algae in the water (x), and the exogenous variables are taken to be Temperature (y_1), Electrical Conductivity (y_2), Residual Chlorine(y_3), and Excess Salts(y_4). First, the existence of multicollinearity is investigated using a correlation matrix in equation (12), Condition Index and variance inflation factor value in table (4). Then the data is fitted to determine significant and not significant correlations. The sample size is taken to be N = 115, and $p = 5$ and $k = 23$, where p denotes the number of exogenous variables and k denotes the number of samples. We proceed to solve the problem via ridge regression and construct the associated RCC [Table (5)]. Table (6) and Table (7) captures the Controlling limits, ARL_0, and ARL_1, corresponding to ordinary raw residuals and Pearson residuals, respectively.

The correlation matrix is given in equation(12)

$$Cor(x) = \begin{pmatrix} 1.000 & -0.329 & 0.109 & -0.300 \\ -0.329 & 1.000 & 0.049 & 0.994 \\ 0.109 & 0.049 & 1.000 & 0.051 \\ -0.300 & 0.994 & 0.051 & 1.000 \end{pmatrix} \qquad (12)$$

Table 4: Estimated coefficient of model

terms	Estimate of φ	SE Coef	VIF	Z- value	P-Value
Constant	5.164	0.392		13.172	0.00
Temperature	−0.096	0.009	1.20	−9.89	0.00
Electrical Conductivity	0.0127	0.005	569.66	2.548	0.002
Residual Chlorine	−0.0215	0.134	1.00	−0.161	0.872
Excess salts	−0.0202	0.0075	565.24	−2.691	0.001
AIC	148.8				
CL	707.984				

Analysis of the application residuals

Table 5: Values of k-estimator and Poisson ridge parameters based on 115 observations

Ridge Parameter	Value	φ_0	φ_1	φ_2	φ_3	φ_4
k_1	9.59	2.19	−0.045	0.02	0.67	−0.03
k_2	0.00	5.16	−0.09	0.013	−0.02	−0.021
k_3	4.31	3.15	−0.063	0.018	0.49	−0.03

Table 6: Control limits, ARL_0, and ARL_1 , corresponding to ordinary raw residuals

			Phase I		Phase II			
k	**m**	**n**	**LCL**	**UCL**	**LCL**	**UCL**	ARL_0	ARL_1
k_1	23	5	−7.8	7.6	−6.58	12.35	389	370
k_2	23	5	−7.5	7.5	−6.94	10.83	358	675
k_3	23	5	−8.9	7.1	−5.89	12	88.7	288

Table 7: Control limits, ARL_0, and ARL_1, corresponding to Pearson residuals

			Phase I		Phase II			
k	**m**	**n**	**LCL**	**UCL**	**LCL**	**UCL**	ARL_0	ARL_1
k_1	23	5	2.6	16	4.7	21.59	365	854
k_2	23	5	2.6	16	4.09	21.41	628	806
k_3	23	5	2.6	16	4.17	21.6	305	186

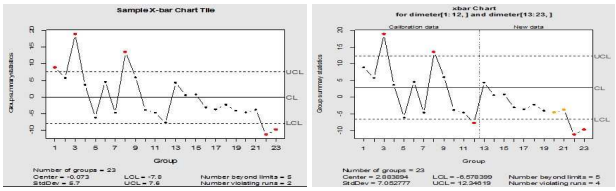

Fig. 7: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_1

As is evident from the figures, the control charts are not stable and the subgroups are out-of-control. Therefore, the optimal value of k-estimators corresponding to ordinary raw residuals is $k_1 = 389$ with respect to ARL_0, as depicted in Table (6). The value should be low. However, the optimal value of k is $k_3 = 288$ with respect to ARL_0 and $k_2 = 628$ and $k_3 = 186$ with respect to ARL_1 . While addressing the multicollinearity problem of data via ridge regression, the best ridge parameter, k_2, was found in the Poisson regression for the two types of residuals (see Table (5)). Table (6) and Table (7) show the Controlling limits, ARL_0, and ARL_1, corresponding to ordinary raw residuals and Pearson residuals, respectively. However, as depicted in Figure (7), Sample $1,3,8,22$ and Sample 23 lie outside the control limit in the case of ordinary raw residuals, but in Figure (8), Sample $1,3,8,22$ and Sample 23 lying outside the control limit in the case of Pearson residuals, but in Figure (9), Sample $3,5,7,12$ and Sample 23 lie outside the control limit in the case of Pearson residuals, and Figure (10) bears a similar result to Figure (11) and

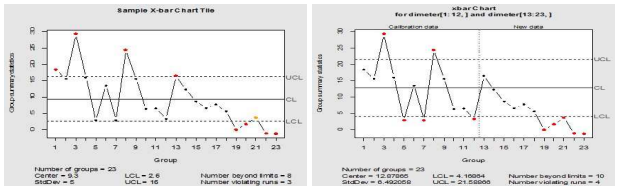

Fig. 8: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_1

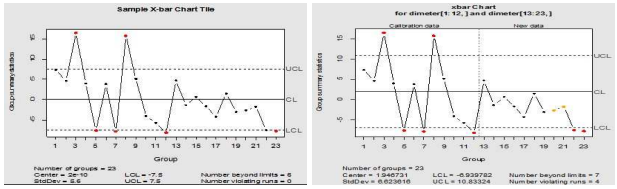

Fig. 9: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_2

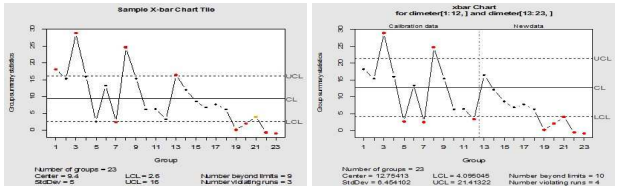

Fig. 10: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_2

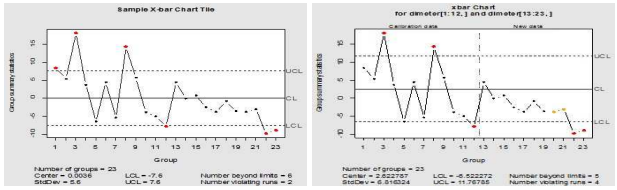

Fig. 11: Control Charts (Phase I and Phase II) corresponding to ordinary raw residuals of k_3

Figure (12), with Sample $1,3,8,22$ and Sample 23 lying outside the control limit.

9 Conclusions

Several researchers, including Månsson and Kibria [16], have utilized ridge regression alongside different regression models to address the multicollinearity problem. Other researchers, such as Filho and Sant'Anna [1], have included Control Charts within regression models and have used Poisson regression to address multicollinearity by using the principal component

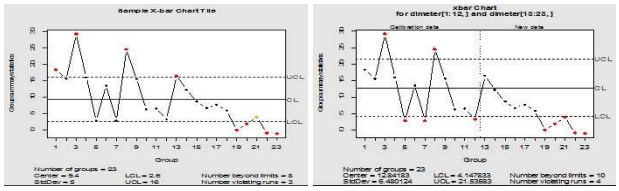

Fig. 12: Control Charts (Phase I and Phase II) corresponding to Pearson residuals of k_3

method and constructing RCCs for deviant residuals. Following their research, in our study, we used a method known as ridge regression, and two types of residuals (ordinary raw and Pearson) to construct RCCs corresponding to every k-estimator. Ordinary raw residuals and Pearson residuals for the control chart are out-of-control. While addressing the multicollinearity problem of the data via ridge regression, the best k-estimator was observed to be k_2 in Poisson regression.

All of the charts corresponding to this model were out-of-control for both types of residuals, and the best k-estimators of control charts of ordinary raw residuals, and Pearson residuals were k_1 and k_3, and all the charts in this model were also out-of-control for both types of residuals. However, in the real study, k_2 ,k_1,and k_3 all exhibited good performance. Thus, during the application of these methods and generation of data, multicollinearity problems were detected and different methods were used to solve the problems and construct control charts to evaluate the performances of these methods and the results from real data show that there is contamination in the treated water, and other treatment methods must be used besides the treatments in use now to be pure and clean for use. In future works, we intend to use further methods to solve such problems and construct RCCs to evaluate them even in the existence of multicollinearity.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

[1] M. D. Filho and O. M. A. Sant'Anna. Principal component regression-based control charts for monitoring count data. The International Journal of Advanced Manufacturing Technology, 85 (5–8), 1565–1574, (2016). DOI:10.1007/s00170-015-8054-6

[2] M. Zhou and T. N. Goh. Effects of Model Accuracy on Residual Control Charts. Quality and Reliability Engineering International, 32, 1785–1794, (2015). https://doi.org/10.1002/qre.1913

[3] Y. A. Areepong. Comparison of Performance of Residual Control Charts for Trend Stationary AR (p) Processes. International Journal of Pure and Applied Mathematics, 85(3), 583–592, (2013). http://dx.doi.org/10.12732/ijpam.v85i3.13

[4] J., Yu and J. Liu. LRProb Control Chart Based on Logistic Regression for Monitoring Mean Shifts of Auto-Correlated Manufacturing Processes. International Journal of Production Research, 49(8), 2301–2326. (2011). DOI: 10.1080/00207541003694803

[5] R. Roy. An Introduction to Water Quality Analysis. International Research Journal of Engineering and Technology, 06(01), 201–205, (2019).ISSN: 2395-0072.

[6] S, Yang and G. Berdine. Poisson Regression. The Southwest Respiratory and Critical Care Chronicles, 3(9), 61-64, (2015). DOI: 10.12746/swrccc2015.0309.125.

[7] R. H. Myers, D. C. Montgomery, G. G. Vining, and T. J. Robinson. Generalized Linear Models with Applications in Engineering and the Sciences. John Wiley and Sons, New York, 2nd Edition, 155-156, (2010).

[8] A. J. Dobson. An Introduction to Generalized Linear Models. Chapman and Hall CRC, London and New York, 2nd Edition, 158, (2002).

[9] F. Famoye, J. T. Wulu, and K. P. Singh. On the Generalized Poisson Regression with an application to Accident Data. Journal of Data Science, 2, 287–295, (2004).

[10] C. Zaldivar. On the Performance of Some Poisson Ridge Regression Estimators. Master of Science (MS), Florida International University FIU Digital Commons, FIU Electronic Theses and Dissertations, (2018). DOI: 10.25148/etd.FIDC006538

[11] K. Månsson and G. Shukur. A Poisson Ridge Regression Estimator. Economic Modelling, 28(4), 1475–1481, (2011). DOI: 10.1016/j.econmod.2011.02.030

[12] M. A. Souza, M. F. Souza, R. R. Zanini, B. Reichert and V. A. De Lima Junior. Applications Residual Control Charts Based on Variable Limits. Journal of Engineering Research and Applications, 5(5), 44–50, (2015). ISSN: 2248-9622.

[13] S. Duttadeka and B. Gogoi. A Study on Exponentially Weighted Moving Average Control Chart with Parametric and Nonparametric Approach. Journal of Agriculture and Life Sciences, 1(2), 2375-4222, (2014). ISSN: 2375-4214.

[14] D. R. Prajapati and S. Singh. Application of ANN to Monitor the Correlated Process using Higher Sample Size. International Journal of Performability Engineering, 11(4), 395–404, (2015).

[15] D. C. Montgomery. Introduction to Statistical Quality Control. John Wiley and Sons, New York, 6th edition, 249, (2009).

[16] K. Månsson, G. Shukur and B. G. Kibria. On Some Ridge Regression Estimators: A Monte Carlo Simulation Study Under Different Error Variances. Journal of Statistics, 17(1), 1–22, (2010).