Shewhart Control Chart of Poisson Regression Under Ridge Regression

https://www.doi.org/10.56830/WRBA11202303

Shaimaa Y. Mohammed
Accounting and Auditing Department, Bader Institutes for Information's technology, Badr City, Egypt
shaimaayassin917@yahoo.com
Salah M. Ramadan
Department of Applied Statistics, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Abstract
This study focuses on the significance of control charts in various fields. Specifically, it introduces a residual control chart as a graphical and statistical tool for monitoring processes or products. A novel 2k estimator, which employs only a k estimator, is utilized and compared in this investigation, as described by (Yassin & Mohamed, 2022) The study is divided into two main parts. The first part addresses the generation of real data, employing Poisson regression to address multicollinearity issues, and utilizing ridge regression as a solution. In the second part, a residual-based Shewhart control chart is constructed, and the Average Run Length is calculated. To support the analysis, a water sample is obtained post-treatment and control charts are prepared after resolving the multicollinearity problems in the data through ridge regression. Overall, this study provides insights into the practical implementation of control charts and their application in addressing statistical challenges.
Keywords: Control charts, Multicollinearity, Ridge regression

1. Introduction
(Yassin & Mohamed, 2022) used a common technique for addressing multicollinearity issues in regression models is ridge regression and draw residual control chart. (Filho & Sant'Anna, 2016) using the principal component to solve this problem for Poisson regression model (PRM) and added a new Methodology to see the performance of the new approach after solving the multicollinearity problem, then where thy used a deviance residual control chart. (Biswas, Masud, & Kabir, 2016) introduced a popular quality control tool in the industrial sector is called the control chart. To establish the control limits and investigate the variations that call for process improvement, statistical quality control techniques are used. There are numerous instances in real life where a single sample is used for process tracking and a control chart is then used for individual measurement. Control charts also referred to as Shewhart charts after Walter A. Shewhart or process-behavior charts are tools used in

CC BY

statistical process control to assess whether an industrial or commercial process is under statistical control (Montgomery, 2009). (Osei-Aning & Riaz, 2017) Monitoring of serially correlated processes using residual control charts. They also discussed the different types of residual control charts, including Shewhart-EWMA – CUSUM residual control charts and so on. In his explanation of the meaning and applications of water quality in 2019, (Roy, 2019) addressed the various uses for it, including drinking, swimming, farming, and manufacturing. Additionally, in order to accomplish each of these specific uses' goals, several set chemical, physical, and biological conditions must be met. For instance, there are stricter regulations governing water used for drinking or swimming than for use in industry or agriculture. Because we need to see the impact of some properties in the formation of algae from which there is no negative or positive effect on the drinking water, we took the data for the water after treating it to see if there was an effect or not. As a result, we took all of the data from the Holding Company for Water and Wastewater in Egypt (Benha water station).

This study uses residual control charts (RCCs) on PRM after solving the multicollinearity problem by ridge regression, followed by the ordinary and Pearson residual control charts from the fitted model, and then drawing the residual control chart. We use the average run length (ARL) as a measure of the RCCs. Simulation and real data on water quality. Finally, we compare one k of (Yassin & Mohamed, 2022) with two different k's of the ridge regression estimator proposed.

In this paper, we introduced Poisson Regression Model in Section 2. In Section 3, we introduce ridge regression (RR). In Section 4, we present our suggested mythology consists of two parts: (i) the Ridge Estimator Formulas, and (ii) Residual Control Chart. In Section 5, shows the control charts of Poisson Model and ARL. In Section 6, we conduct simulation studies for Poisson model. In Section 7, we introduce a case study using real data as Poisson regression. An algorithmic approach to basic programs for generating models is also discussed in this section. Finally, the study is concluded in Section 8.

2. The Poisson Regression Model (PRM)

In the Poisson regression model, the dependent variable is account data, so the PRM is useful for modelling response variables Y. Hence, the probability mass function is given by

$$f(X = x) = {\lambda^x e^{-\lambda}}/{x!}\,, x = 0,1,2, \dots$$

Where $f(X = x)$ denotes the probability, and $x! = x(x-1)\dots 3.2.1$denot the independent variables (Filho & Sant'Anna, 2016).

The regression model for the Poisson regression (PR), as presented in (Yang & Berdine, 2015), is as follows.

$$log_e(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \quad (1)$$

According to (Yassin & Mohamed, 2022), there are two types of residual: ordinary raw residual and Pearson residual.

The ordinary raw residuals are obtained by

$$r_o = y - \hat{\mu}\,, \quad (2)$$

Print ISSN 2771-1161
Online ISSN 2771-1153
CC BY

Therefore, the Pearson residuals could be rewritten as follows:

$$r_p = \frac{y-\hat{\mu}}{\sqrt{\hat{\mu}}}. \quad (3)$$

3. Ridge Regression(RR)

In count data models, ridge regression is an important and common tool to solve multicollinearity. Therefore, the multiple ridge regression could be written as follows:

$$\hat{\beta}_{rr} = (x`x + kI)^{-1}x`y, \qquad k \geq 0. \quad (4)$$

Where I is a symbol of the identity matrix and k denotes a positive integer known as the ridge parameter. According to (Rashad & Algamal, 2019), the Poisson ridge regression model is given as follows:

$$\hat{\beta}_{prr} = \left(x`\widehat{W}x + kI\right)^{-1}x`\widehat{W}\, x\, \hat{\beta}_{pML} \;\; , \;\; k \geq 0.$$

Where the Maximum likelihood is given by

$$\hat{\beta}_{ML} = \left(x`\widehat{W}x\right)^{-1}x`\widehat{W}\hat{z}. \quad (6)$$

4. METHODOLOGY

4.1 The Ridge Parameter

RR is a common method to solve multicollinearity problem in regression models.

(Månsson & Shukur., 2011) suggested using the proposed K Ridge Estimator of (Hoerl & Kennard, 1970) as follows:

$$K_1 = \hat{k}_{HK1} = \frac{\hat{s}^2}{\hat{\alpha}^2_{max}}\,, \quad (4)$$

since $\hat{\alpha}^2_{max}$ is the maximum element of $\hat{\alpha}_i^2$, where $\hat{\alpha}_i^2 = (\gamma\hat{\beta}_{ML})^2$ and $\hat{s}^2 = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_i)^2}{n-p-1}$.

The next K Ridge estimator, suggested by (Yassin & Mohamed, 2022), are obtained this way:

$$k_2 = \frac{p}{\sum_{i=1}^{n}\left(\hat{\alpha}_i^2/\left(1+\left(1+\lambda_i\sqrt{\hat{\alpha}_i^2}\right)\right)\right)}, \quad (5)$$

(Zaldivar, 2018) suggested using the proposed k ridge estimators from Alkhamisi which is as follows:

$$k_3 = median\left\{\sqrt{\lambda_i\hat{\alpha}_i^2}\right\} \quad (6)$$

4.2 Residual Control Chart

For the statistical control of multiphase processes or products, Residual Control Charts (RCCs) are useful instruments. (Souza, Zanini, & B. Reichert, 2015)showed how the ability of RCCs to monitor the stability of production variables using a single chart to simultaneously validate mean and variation is significantly impacted by the choice of appropriate forecasting models.

CC BY

5. Control Chart for Poisson Regression Model

The Shewhart control chart's ability to identify changes in processes or procedures, but a drawback is that it takes longer to identify minor flaws. According to (Filho & Sant'Anna, 2016), the residuals Shewhart Control upper and lower limits are as follows:

$$CL_r = E(r_n) \pm w\sqrt{Var(r_n)} \cong \pm w. \quad (7)$$

Where r is defined as the residuals and $r \sim N(0,1)$, the constant w is defined as the amplitude between control limits and depends on the false alarm probability α. Since the Average Run Length (ARL) is a widely used metric to illustrate how well a procedure or product is performing while the control chart is in control, the if the control chart in control the average run length is $ARL_0 = 1/\hat{\alpha}$, but if the control chart is out of control the $ARL_1 = 1/(1-\hat{\beta})$, where $\hat{\alpha}$ is the probability of false alarm (type I error) and $\hat{\beta}$ is the probability of true alarm (type II error) (Montgomery, 2009).

6. Simulation Study

We have used the Shewhart control charts for the Poisson Regression Model to deal with that multicollinearity problem. We are going to use the Ridge Regression technique to solve this problem. We are also going to employ the Residual Control Chart to show the performance of the k estimator in different phases and phase II by using the R program version 3.5.3. Therefore, we are generating the data on a multicollinearity problem where the sample size is equal to 100. So we suggested a new 2 k estimator with only a k estimator from (Yassin & Mohamed, 2022) and makes a comparison between them.

6.1 Simulation Study for Poisson Regression

Our simulation algorithm follows the following steps.

1. Set $N = 100$, $p = 4$ and generate a standard normal variate z.
2. Generate an independent variables X by an equation $X_{ij} = (1-\rho^2)^{(\frac{1}{2})} z_{ij} + \rho z_{ip}$ where $i = 1,2,\dots,n$, $j = 1,2,\dots,p$ and $\rho = 0.85$. With ρbeing the correlation between the independent variables.
3. Choosing $\boldsymbol{\beta}$but by condition $\sum_{i=1}^{p} \beta_j^2 = 1$ and taking $\beta_0 = 1.5$.
4. Generate satisfying Y following the Poisson regression model.
5. Generate $po(\mu)$ and use it in (4) where $\boldsymbol{\mu}$ is given by $\mu_i = exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$

Table 1 shows the estimates of k estimators and Poisson ridge parameters.

Ridge Parameter	Value of ridge parameter	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{\beta_2}$	$\boldsymbol{\beta_3}$	$\boldsymbol{\beta_4}$
$\boldsymbol{k_1}$	2.657954e-18	1.44127	-0.4587619	-0.298932	-0.315431	-0.8922906
$\boldsymbol{k_2}$	13.73599	1.40075	-0.53034	-0.414035	-0.41937	-0.65834
$\boldsymbol{k_3}$	11.96738	1.40567	-0.5292823	-0.4080983	-0.413473	-0.6673269

Table 2 shows the values of the control limits, ARL_0, and ARL_1 corresponding to ordinary raw residuals.

Print ISSN 2771-1161
Online ISSN 2771-1153

k	m	n	Phase One		Phase Two		ARL_0	ARL_1
			LCL	UCL	LCL	UCL		
k_1	25	4	-4.24	4.19	-2.23	0.528	385	370
k_2	25	4	-1.9	-0.36	-1.93	-0.299	434.8	1.001
k_3	25	4	-9.11	9.71	-6.71	10.02	400	417

Table 3 shows the values of control limits, ARL_0, and ARL_1 for corresponding to Pearson residuals.

k	m	n	Phase One		Phase Two		ARL_0	ARL_1
			LCL	UCL	LCL	UCL		
k_1	25	4	-3.227	0.899	-4.249	1.379	151	103
k_2	25	4	-2.8	1.3	-3.06	2.08	621	1.002
k_3	25	4	-2.156	0.128	-2.39	0.5	125	357

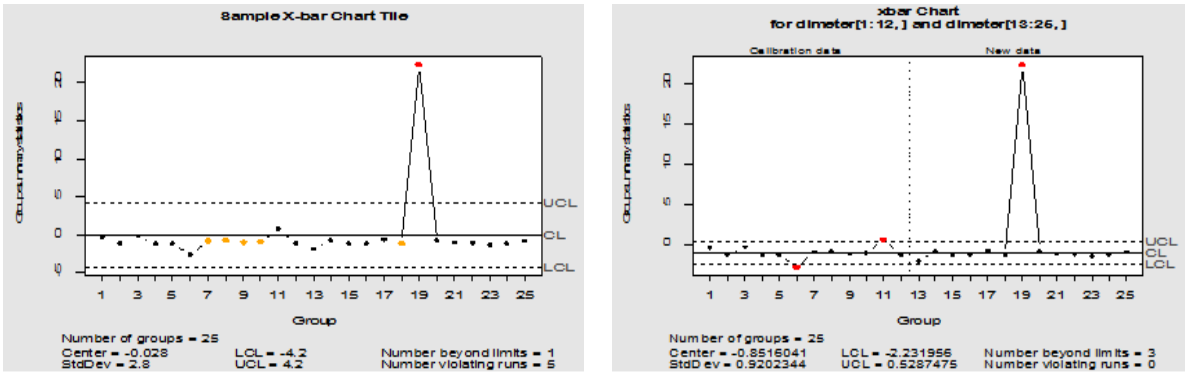


Figure (1) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_1

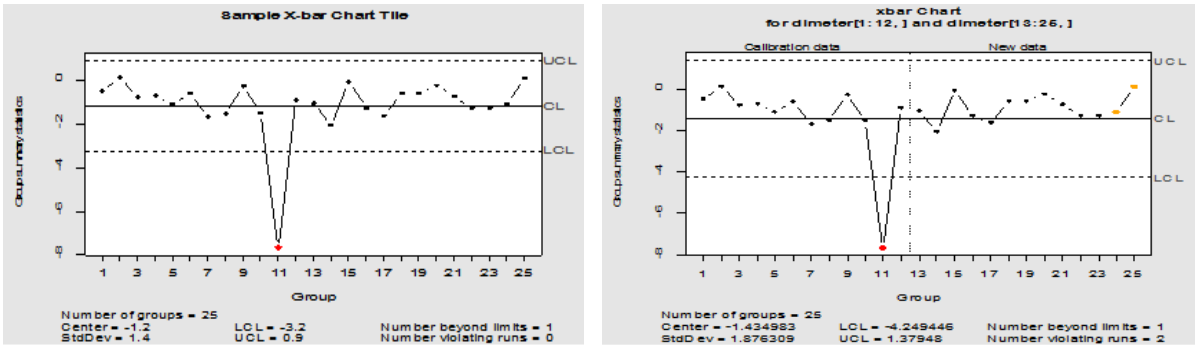


Figure (2) is the Control Chart (Phase I and Phase II) for Pearson Residuals of k_1.

CC BY

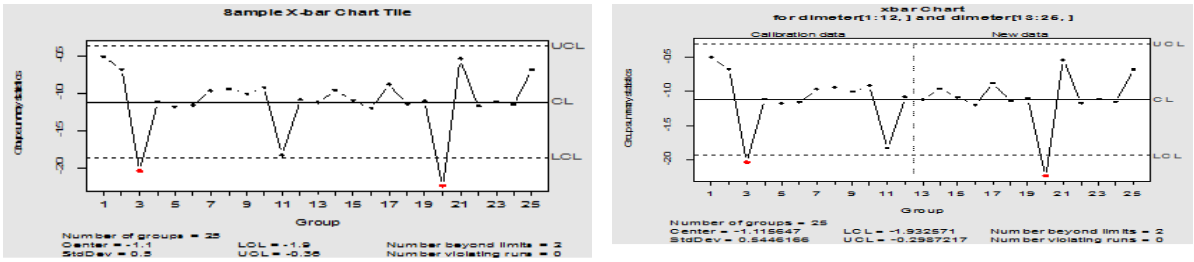


Figure (3) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_2.

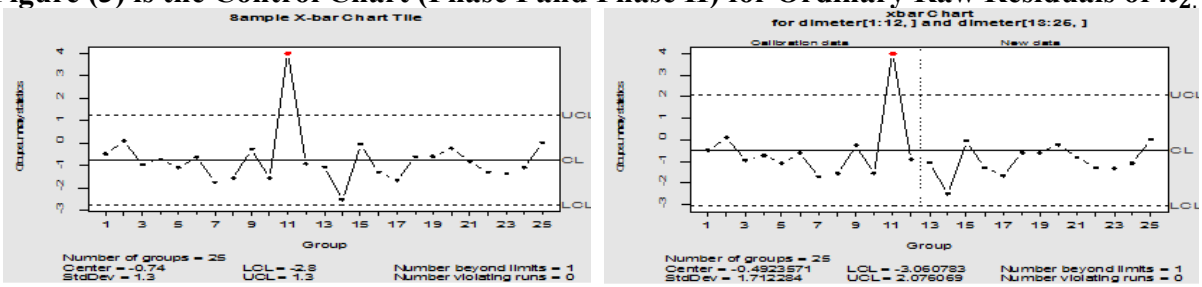


Figure (4) is the Control Chart (Phase I and Phase II) for Pearson Residuals of k_2.

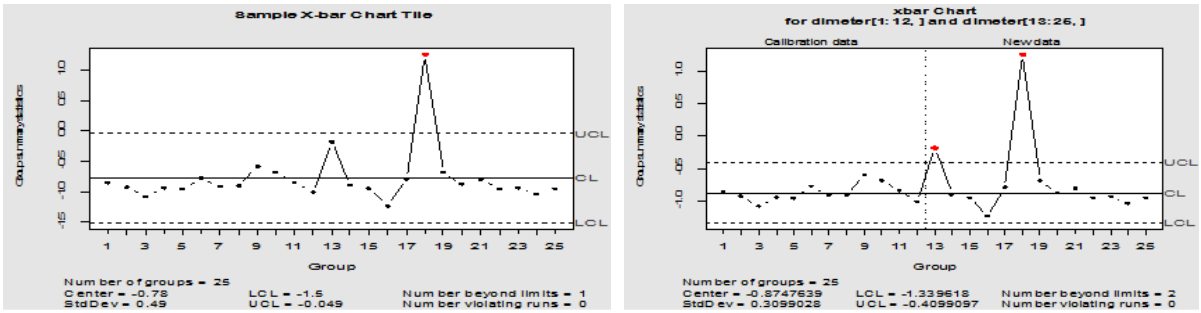


Figure (5) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_3.

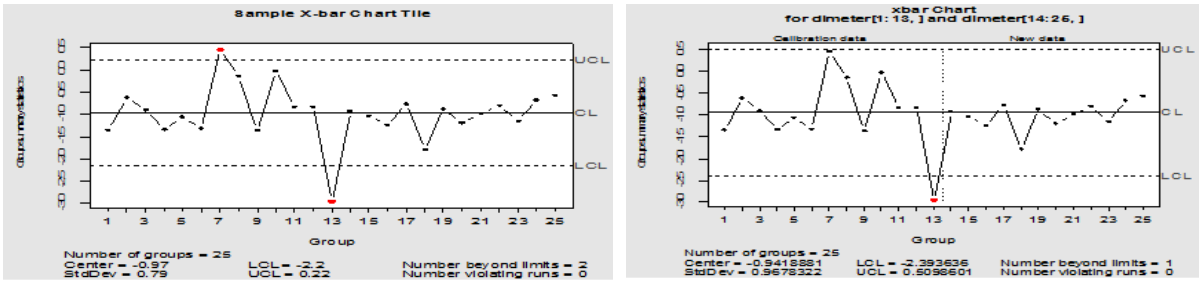


Figure (6) is the Control Chart (Phase I and Phase II) for Pearson Residuals of k_3.

Analysis of the simulation study

The subgroups of the sample from the simulation are out of control, and the control chart is unstable. Figures (1) and (2) show Phase I and Phase II control charts that correspond to the ordinary raw and Pearson residuals of k_1, respectively. In terms of the out-of-control-limit samples, there are a few variations between the situations of ordinary raw residuals and Pearson residuals. As shown by the related Phase I Control Chart, Sample 11 is beyond the control limits for Pearson residuals but Samples 3 and

Print ISSN 2771-1161
Online ISSN 2771-1153

20 are outside the control limits for ordinary raw residuals of k_2. But in ordinary raw case for k_1 it is the sample number 19 where the shift in the mean is 9, 10 but in the case of Pearson it is the number 11 for the stage of control chart for the Phase I, when the defect is discovered in the control chart, those who have reasons are deleted Special reasons for variance, and the new data is used in quality control charts for the Phase II according to the new data. But in Figure (5) the control chart is out of control, and the points outside the boundaries of the control chart are 18 for the ordinary residuals, but in Figure (6) the points outside the limits of the control chart are (13,7) for Pearson's residuals. Thus, we remove all outside points for those charts and using the new data (Phase II), while Table (2) and Table (3) show the ARL value of the ordinary and Pearson residual control chart, in the case of ordinary raw residuals are k_2 is equal to 434.8 with respect to ARL_0 and k_2 is equal to 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_2 is equal to 621 with respect to ARL_0 and $k_2$is equal =1.002 and k_2 is equal to 368 with respect to ARL_1. Therefore, the best value here for the two cases is k_2 taken from (Yassin & Mohamed, 2022), then k_1 , finally k_3 , and for this, we find that the value of k_2 is the best k in the simulation study.

7. Application with real data for Poisson regression model

In this section we illustrated the suggested strategy monitoring by real study of some properties of Egyptian Water. So the dependent variable is a Total alga a count (y), and the independent variables are Temperature (x_1), Electrical Conductivity (x_2), Residual Chlorine (x_3), Excess salts(x_4). So we doing Condition Index, variance inflation factor to show if the data have a multicollinearity problem or not and make fitting for data to obtained on $\hat{\beta}_{ml}$ from the model to include in parameter of k ridge estimator and beta ridge parameter, next make fitting for model to obtained on residuals (raw ordinary and Pearson), then we draw the residual based Shewhart control charts for all of k's ridge estimator, where the sample size is N=115 and p is number of independent variables, so p=4.

Table 4 Estimated coefficient of model (Yassin & Mohamed, 2022)

terms	Estimate of β	SE Coef	VIF	Z- value	P-Value
Constant	5.164673	0.392		13.172	<0.001
Temperature	-0.09565	0.00966	1.20	-9.898	<0.001
Electrical Conductivity	0.012714	0.00499	569.66	2.548	0.002
Residual Chlorine	-0.02154	0.134	1.00	-0.161	0.872
Excess salts	-0.02023	0.00752	565.24	-2.691	0.001
AIC	148.8				
CI	707.984				

CC BY

Table 5 shows the value of k estimator and Poisson ridge parameter from 115 observations (y,x) .

Ridge Parameter	Value of ridge parameter	β_0	β_1	β_2	β_3	β_4
k_1	1.128356e-20	5.164673	-0.0956497	0.01271406	-0.0215408	-0.0202346
k_2	9.58673	2.187355	-0.0445185	0.0204936	0.6701919	-0.0309523
k_3	851.9038	0.0802308	0.06605027	0.03410772	0.160257	-0.0491641

Table 6 shows the values of Limit Control and ARL_0 , ARL_1 because case one is about Ordinary Raw Residuals.

k	m	n	Phase one		Phase two		ARL_0	ARL_1
			LCL	UCL	LCL	UCL		
k_1	23	5	-7.53	7.53	-6.94	10.83	370	1.001
k_2	23	5	-7.8	7.6	-6.58	12.35	389	370
k_3	23	5	-8.87	7.15	-6	12.5	91.7	357

Table 7 shows the values of Limit Control and ARL_0 , ARL_1 as case two is about the Pearson Residuals.

k	m	n	Phase one		Phase two		ARL_0	ARL_1
			LCL	UCL	LCL	UCL		
k_1	23	5	2.64	16.11	4.09	21.41	344.8	833
k_2	23	5	2.6	16	4.17	21.59	365	854
k_3	23	5	2.47	16.03	4.24	21.61	0.991	370

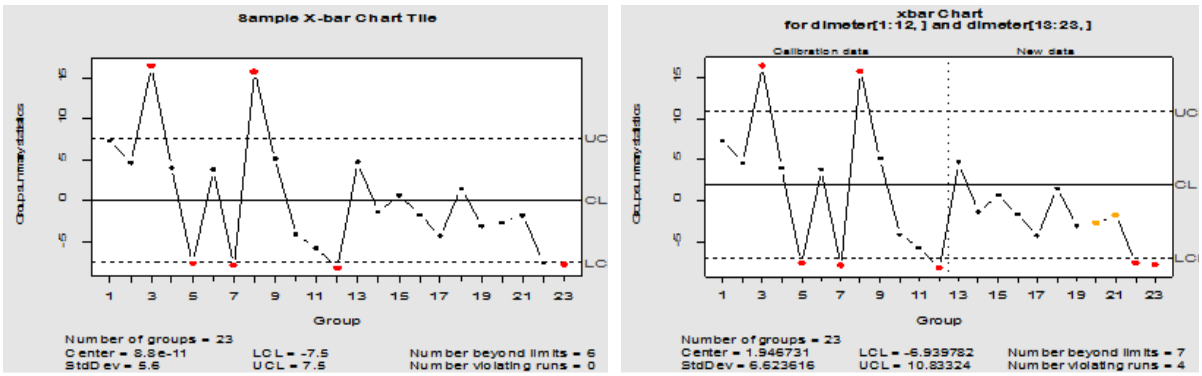


Figure (7) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_1.

Print ISSN 2771-1161
Online ISSN 2771-1153

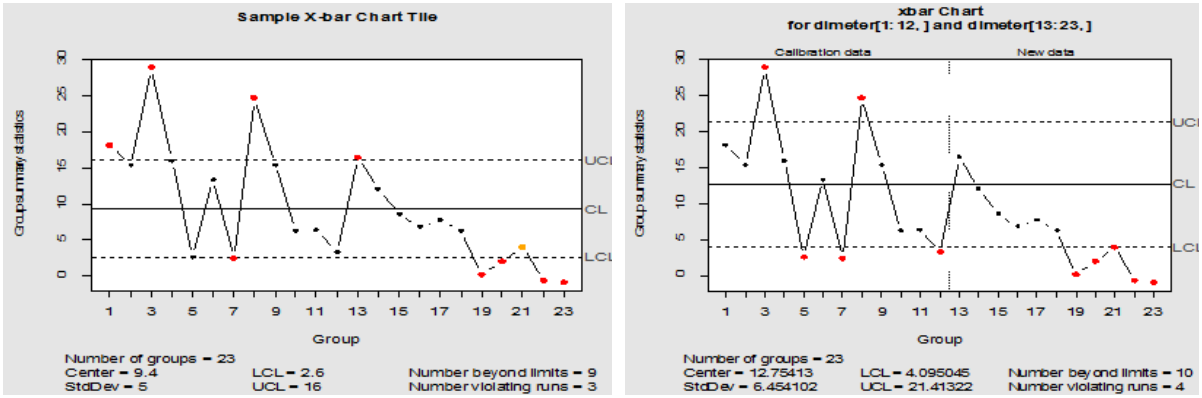


Figure (8). Control Chart (Phase I and Phase II) corresponding to Pearson residuals of k_1.

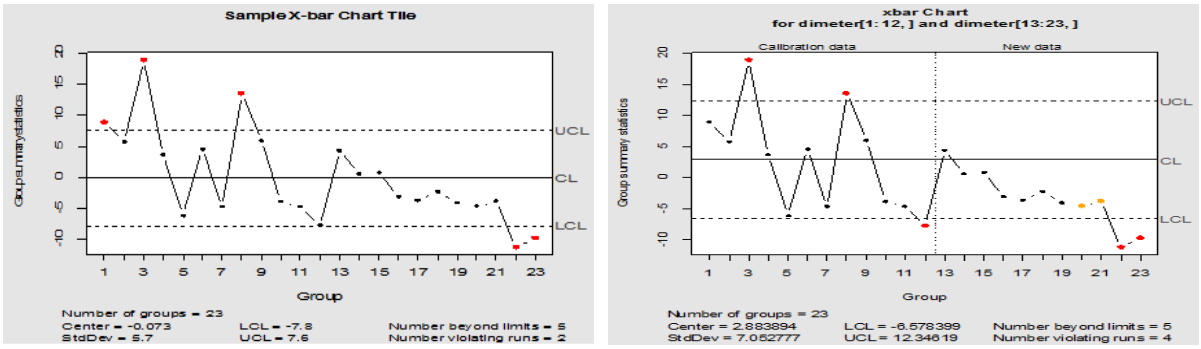


Figure (9) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_2.

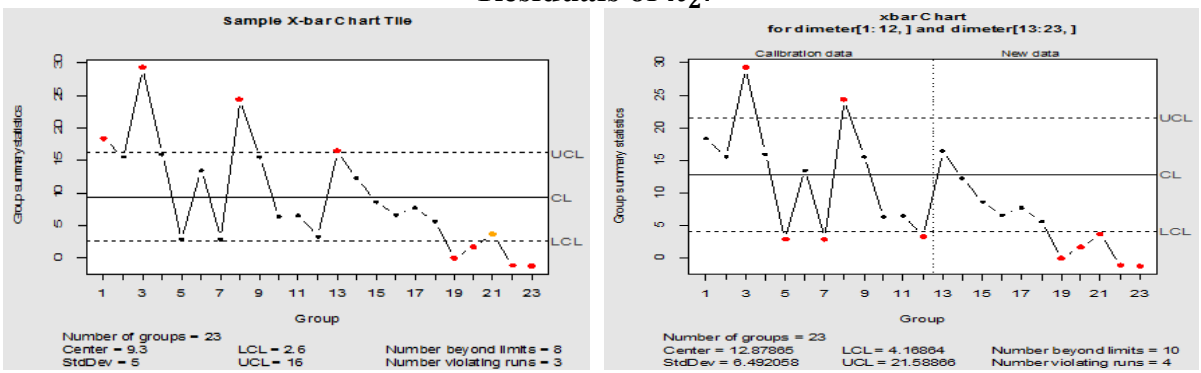
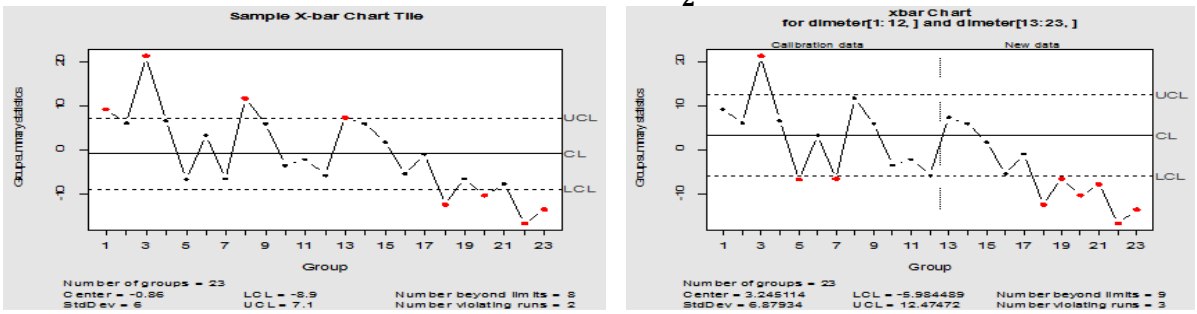


Figure (10). Control Chart (Phase I and Phase II) corresponding to Pearson residuals of k_2.



CC BY

Figure (11) is the Control Chart (Phase I and Phase II) for Ordinary Raw Residuals of k_3.

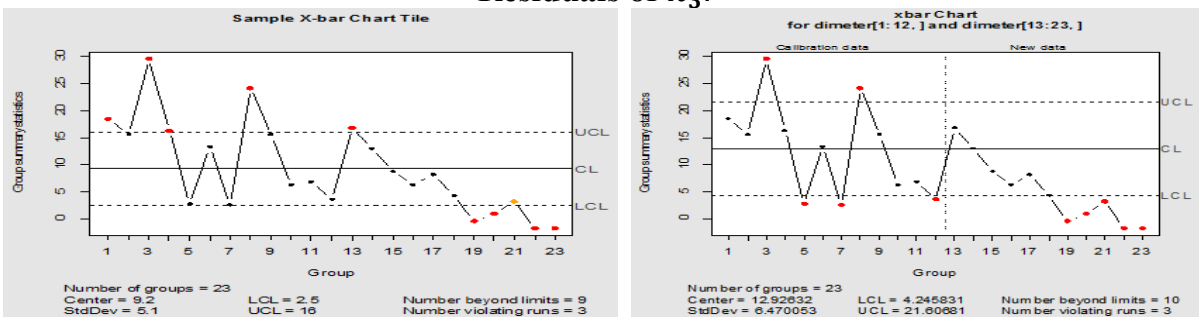


Figure (12). Control Chart (Phase I and Phase II) corresponding to Pearson residuals of k_3.

So all of the control charts are out of control for two types of residuals (ordinary raw and Pearson), and the best performance is measured by the average run length of the control charts. In Table (6), the ARL_0 of ordinary raw is k_2 which is equal to 389, and in ARL_1 the best performance of good k which is k_1 equal to 1.001, but in the Pearson residuals of Table (7), the best performance of k is k_2 in ARL_0 and in ARL_1 is k_3. However, as depicted in Figure (7), Samples 3,5,7,8,14,19,21,22, and 23 lie outside the control limit in the case of ordinary raw residuals, but in Figure (9), Samples 1,3,8,12, and 23 lie outside the control limit in the case of ordinary raw residuals, but in Figure (8), Samples 3,5,7,8,14,19,20,21,22, and 23 lie outside the control limit in the case of Pearson residuals. So, we found that the k-ridge estimator that was taken from (Yassin & Mohamed, 2022) had the best performance in average running length in the simulation study. However, in the real data, the situation was different. Hence, the best value is k_2, which was estimated by (Yassin & Mohamed, 2022) in ARL_0. As for ARL_1, it was completely different in the ordinary raw residuals; it was k_1, and in the Pearson residuals, it was k_3.

8. CONCLUSIONS

1. for the ridge regression, the values of k's estimators are good because the condition of the k estimator is $k \geq 0$, and this condition has been investigated in this study.
2. All of the charts in the account data regression model are out of control for Pearson and ordinary raw residuals, so when we detect the control chart, we omit those with special causes of the calculations and then draw the control chart for phase II with new data. We also recommend using k_2.
3. As for the real data, we find that the treated water is still polluted, and therefore, we recommend that the company use other treatments with the treatments that were used in order to reduce the presence of harmful algae in the water because of its danger to life because it can cause **Algal Toxin**.
4. In future work we shall use different methods to solve such problem and draw the residual control chart to show the effect of the performance of these methods under the existence of multicollinearity and make a comparison between different methods and different residual control charts.

Print ISSN 2771-1161
Online ISSN 2771-1153

Reference

Biswas, R. K., Masud, M. S., & Kabir, B. E. (2016). Shewhart control chart for individual measurement: an application in a weaving mill. *Australasian Journal of Business, Social Science and Information Technology, Vol. 2, No. 2*, 89-100.

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Non orthogonal problems. *Technometrics, Vol. 12, No. 1*, 55-67.

Månsson, K., & Shukur., G. (2011). On Ridge Parameters in Logistic Regression. *Communications in Statistics -Theory and Methods, 40(18)*, 3366-3381.

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

Osei-Aning, R. A., & Riaz, M. (2017). Monitoring of Serially Correlated Processes Using Residual Control Charts. *Scientia Iranica,Vol. 24, No. 3*, 1603-1614.

Rashad, N. K., & Algamal, z. y. (2019). A New Ridge Estimator for the Poisson Regression Model. *Iranian Journal of Science and Technology, Vol. 43*, 2921–2928.

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

Souza, M. A., Zanini, M. F., & B. Reichert, V. A. (2015). Applications Residual Control Charts Based on Variable Limits. *Journal of Engineering Research and Applications, 5(5)*, 44–50. ISSN: 2248-9622.

Yang, S., & Berdine, G. (2015). Poisson Regression. *The Southwest Respiratory and Critical Care Chronicles, 3(9)*, DOI: 10.12746/swrccc2015.0309.125.

Yassin, S. M., & Mohamed, S. M. (2022). Performance Comparison of Residual Control Charts for a Count Data Based on Ridge Regression. *Information Sciences Letters, Vol. 11, No.1*, 2301–2326.

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