# Enhancing Model Monitoring and Quality Control: The Integration of Control Chart for Binomial Regression under the Principal Component Framework


## Abstract

The aimof this study is uses a new approach called the Shewhart control chart with binomialregression, but under this condition, there was a multicollinearity problem in binomial regression model, and solving this problem by principal component, binomialregression with control chart is a statistical method used in quality control to monitor and improve the quality of products or services. This method involves using binomial regression to model the relationship between an input variable and output variable, and then integrating control charts to detect any changes in the process over time.


## 1.Introduction

Model monitoring and quality control are vital in statistical analysis and predictive modelling, as they are responsible for guaranteeing the dependability, precision, and resilience of the developed models. These procedures encompass ongoing assessment, validation, and enhancement of the models to uphold their efficacy in making forecasts and informed choices.However,Bilen et al., (2007)have developed a comprehensive framework for controlling multivariate auto-correlated processes that rely on principal component analysis and multiple regression control charts. The framework begins by identifying critical process variables using principal component analysis. Next, the autoregressive error correction model's white noise series is used to create a control chart that can be used for ongoing process monitoring. So, organizations must use quality control methods to guarantee the quality of their goods and services.The quality characteristic to be monitored in the second stage of two-stage processes whose distribution is binomial is the subject of this study's proposed control chart .The deviance residual, which is simply the generalized log-likelihood ratio statistic derived from the generalized linear model, forms the basis of the suggested control chart. And suggest applying a new link function in a generalized linear model framework to determine the relationship between the first- and second-stage quality features. Regarding the average run length requirement, the performance of the suggested control chart with the new link function is contrasted with that under the conventional logit link function.

Furthermore, a comparison is made between the suggested control chart's performance and that of the chart created using the original residuals under the new link function, as well as the conventional np-chart used to track the binomial quality characteristic during the second stage (Amiri et al., 2016).

This study concentrated on the statistical method of creating a regression residual control chart using regression analysis to track students' academic progress in higher education institutions. While the variability was being tracked using the Moving Range Residual Control Chart, in e moving range chart, there were three spots out of control and one out of control in the individual control chart. As a result, the out-of-control locations were eliminated, and new, updated control boundaries were determined ((Rashid et al., 2013)). But in Yassin and Mohamed, (2022)used one useful and popular method for creating associations between pairs of endogenous and exogenous variables regression modelling. They typically experience multicollinearity, though. This work employs residual control charts on count data (Poisson regression) following the application of the ridge technique to address multicollinearity issues. But (Filho et al., 2016) created residual control charts with Poisson regression. And used the principal component to solve the multicollinearity issue and then created a control chart. We used average run length as the statistic to assess the control chart.To improve quality, control charts are crucial statistical quality control instruments. Nancy et al., (2023)explained the method and graphical procedure known as statistical quality control aided process monitoring and control. For the desired level of quality. Standard Shewhart control charts and their variations and control charts that are impacted by associated independent variables for the process shift are two distinct ways to look at control charts. In statistical process control, control charts are effective tools for monitoring processes. When examining a process based on an exponential family distributed response variable (such as binary outcomes) along with a single explanatory variable, the generalized linear model (GLM) provides better estimates.

The generalized linear mode (GLM) l is introduced by Nelder and Wedderburn, so the binomial regression model is a specific state of a generalized linear model. In other words, the statistical relationship between one or more independent variables and dependent variables can be verified. Furthermore,three elementary parts create the

basis of a generalized linear model: a systematic component created from predictor variables that yield a linear predictor; a link function that connects the random and systematic parts; and a random component represented by the binary response variable, an average vector, and an exponential distribution ((Soares et al., (2006)).

This paper aims to discuss the binomial regression with multicollinearity problem thatexists, solve this problem by PCA draw the residual-basedShewhart control chart, and use the ARL as a performance measure.

The remaining sections of the paper provide a concise overview of the Binomial Regression Model and onlyone typeofresiduals, namely ordinary raw residual, in Section 2. Section 3 introduces our proposed methodology, which utilizes two approaches: (i) the application of the Principal Component Formula to address multicollinearity issues, and (ii) the use of a Residual Control Chart (Shewhart). Additionally, control charts for the binomial Model are presented. Section 4 consists of simulation studies describing the algorithmic approach and basic program for generating the models. We encourage further discussion on these topics. Lastly, Section 5 serves as the conclusion.

**2. Binomial Regression Model**

In binomial regression, the dependent variable represents the percentage of success in $n$ independent essays, each with a probability of occurring p. According to Pardo et al., (2007), the Regression model follows a binomial distribution with index n and parameter P, i.e., n $y^{\circ} \sim B(n, p)$. The density function of binomialdistribution is given by:

$$f(y^{o}, p) = \binom{n}{ny^{\circ}} p^{ny^{\circ}} (1-p)^{n-ny^{\circ}} \; where \; 0 < p, y^{\circ} < 1. \qquad (1)$$

Amiri et al., (2016)stated that the logit link function for binomial Regression is

$$Logit(p_i) = log\left(\frac{p_i}{(1-p_i)}\right)$$

Soares et al., (2006)stated that the statistical regression model for binomial Regression is:

$$y^{\circ}_{i} = \varphi_0 + \varphi_1 x_{i1} + \varphi_2 x_{i2} + \cdots + \varphi_k x_{ik}. \quad (2)$$

Where $= 1, \dots, I$ , and independent variable $x_i = (x_{i0}, x_{i1}, \dots, x_{ik})$ ,
$\varphi = (\varphi_0, \varphi_1, \dots, \varphi_k)^t$.

Furthermore,

$$p_i = p(x_i^t \varphi) = \frac{exp\left(\sum_{j=0}^{k} \varphi_j x_{ij}\right)}{1+exp\left(\sum_{j=0}^{k} \varphi_j x_{ij}\right)}, where\ j = 1, \dots, k. \quad (3)$$

A multicollinearity problem is known as when we have a high correlation between independent variables $(x)$. The least squares estimates are unbiased whenmulticollinearityexists, but because of their enormous variances, they could not be very close to the true value. Principal components regression lowers the standard errors by biasing the regression estimates to some extent.
We have used in this study an ordinary raw residualgiven from (Dunn and Smyth (2018)).

The ordinary raw residualis as follows:

$$\Re_o = y^{\circ} - \hat{\mu} (4)$$

where the test of the null hypothesis is $H_0 : p = p_0$ against the alternative hypothesis $H_a : p \neq p_0$ (See the details in (Amiri et al., 2016)).

**3. Methodology**

**(I) Principal Component Analysis:**

The definition of the principal component analysis refers to Filho and Sant'Anna (2016). So, Principal component analysis (PCA) is a multivariate statistical method that aims to summarize information about the linear correlation structure in a group of variables under examination. The formula for the $i^{th}$ Principal Component (PC) score referring to observation x is given by:

$$z_i = x * u_i \quad (5)$$

Where $z_i$presents the$i^{th}$ PC score, $x$ is the observation $u_i$ is the eigenvector corresponding to the $i^{th}$ PC.

**(ii) Control Charts (CCs)**

The control charts are an important and most widely used tool for statistical quality control. Furthermore, the statistical quality control approach is used to placement the

limits of control chart and realization the modifications for product or process refinement. In applicationcases, a lot of scenarios where the control charts are used for monitoring the product or process**.**

**Shewhart Control Chart (SCC)for a Binomial Regression Model**

The Shewhart Control Chart is a useful graphic statistical tool. Yassin and Mohamed (2022) used to the Shewhart based on residual control chart (RCC), furthermore (SCC) is used to detect process shifts or changes reverse to EWEMA control chart. The control chart limits (lower and upper) are given by:

$$CL_{\Re} = E(\Re_n) \pm \upsilon\sqrt{Var(\Re_n)} \cong \pm\upsilon. \quad (6)$$

Where$\Re \sim N(0,1)$, the constant $\upsilon$refers to the amplitude between control limits that depend on the false alarm probability $\tau$. So, we have used the Shewhart Control Chart to show the performance of the Principal Component.

**The Average Run Length:**

Average Run Length (ARL) is used asa statistical measurementfor the number of observations required before control chart signals an out-of-control process. It is used to evaluate the effectiveness of control charts to detect process shifts or changes. The ARL is based on the chart design and the parameters used to establish the control chart limits, and a longer ARL means a chart is slower to signal a process shift. A shorter ARL indicates that a chart is faster at identifying process changes, but may result in more false alarms.

According to Yassin and Mohamed (2022), ifthe control chart is in control the (ARL) is equal to $ARL_0 = 1/\hat{\alpha}$ , but if control chart is out of control the (ARL) is equal to$ARL_1 = 1/(1-\hat{\beta})$, where $\hat{\alpha}$ is the probability of false alarm (type I error) and $\hat{\beta}$ is the probability of true alarm (type two).

**4. SimulationStudies**

By using the R program version 4.3.3we draw the Shewhart control chart for binomial regression under the condition multicollinearity problem wase exists**,** we will use it to solve our problem by principal component.And also, used the deviance residual to

draw a chart, we are also going to employ the Shewhart Control Chart to show the performance of the principal component in different phases and phase II.

**The steps of generating data:**

1. Put $N = 200, p = n,$ andgenerate$z$ where $z_{ij}$followinga standard normal distribution.
2. Put the degree of correlation$\delta = 0.95\ in\ equation$ generating independent variables $x$ by equation $x_{ij} = (1-\delta^2)^{(\frac{1}{2})} z_{ij} + \rho z_{ij},$ where $i = 1, \dots, n, j = 1, \dots, m.$
3. Then choosing $\varphi$ under condition $\sum_{j=1}^{m} \varphi$=1 and taking $\varphi_0 = 1.5$ from (Filho and Sant'Anna, 2016).
4. Generating dependent variable $y^{\circ}$ of binomial regression model following n $y^{\circ} \sim B(n, p)$.
5. Finally, generating$p_i$, furthermore, we using it in step 4. Where p is given by:$p_i = \frac{\exp(x_i\theta)}{1+\exp(x_i\theta)}$ , $i = 1, 2, \dots, n.$

**Table (1) shows that thecorrelation matrix**

| Variable | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| $x_1$ | 1.0000000 | 0.9973866 | 0.9973168 | 0.9986863 |
| $x_2$ | 0.9973866 | 1.0000000 | 0.9965585 | 0.9984274 |
| $x_3$ | 0.9973168 | 0.9965585 | 1.0000000 | 0.9986005 |
| $x_4$ | 0.9986863 | 0.9984274 | 0.9986005 | 1.0000000 |

**Table (2) shows that the estimated coefficient of model**

| terms | Estimate of $\varphi$ | SE Coef | VIF | Z-value | P-value |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| **constant** | -0.9178 | 0.1635 | | -5.615 | 1.97e-08 |
| $x_1$ | -2.5532 | 3.2164 | 381.6762 | -0.794 | 0.42732 |
| $x_2$ | -9.4354 | 3.4138 | 328.4978 | -2.764 | 0.00571 |
| $x_3$ | -0.3039 | 3.2785 | 365.6465 | -0.92614 | 0.92614 |
| $x_4$ | 11.3533 | 5.2941 | 1096.3540 | 2.145 | 0.03199 |
| **AIC** | 242.9 | | | | |
| **CI** | 78.58528 | | | | |

Table (2)explain the estimated coefficient of modeland show that the VIF is statically measure the degree of multicollinearity, so in this table VIF greater than 10 indicates a high degree of multicollinearity, CI is equal to 78.58528, then we found $x_2\ and\ x_4$ is significant but $x_1\ and x_3$is not significant at the 5% level.

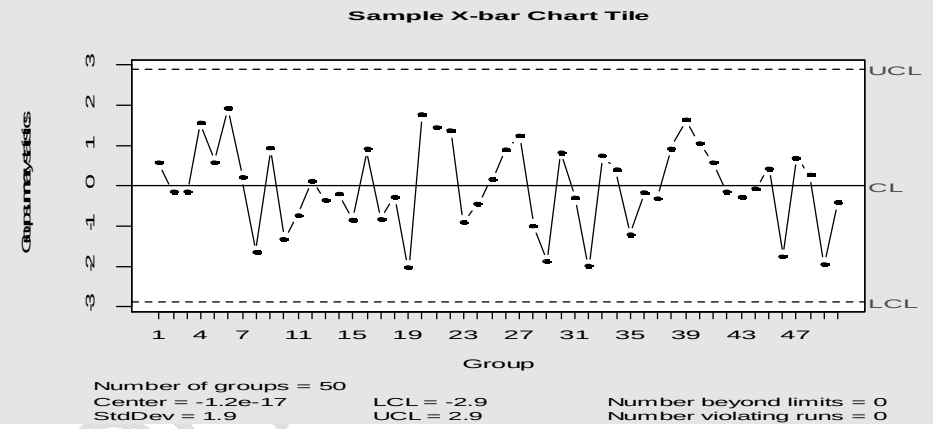


Figure (1) phase one with sample size $n = 4$ and number of sample m=50 for in control process.

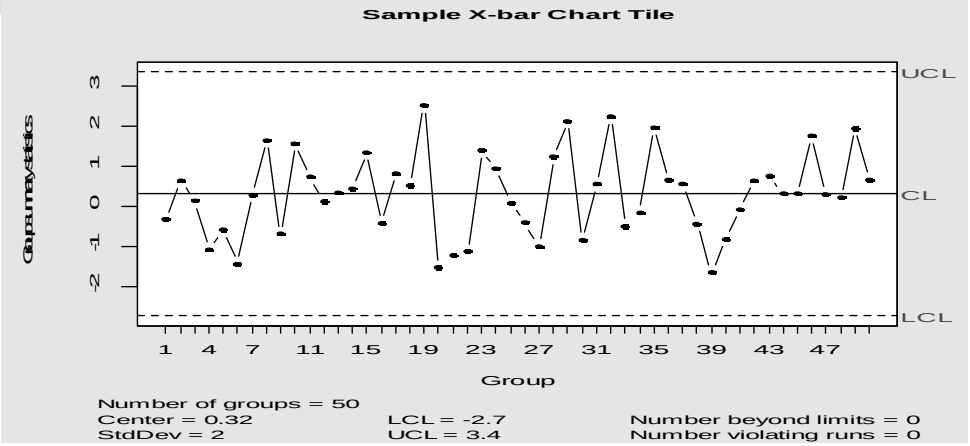

Figure (2) phase one for binomial principal component residual based on control chart.

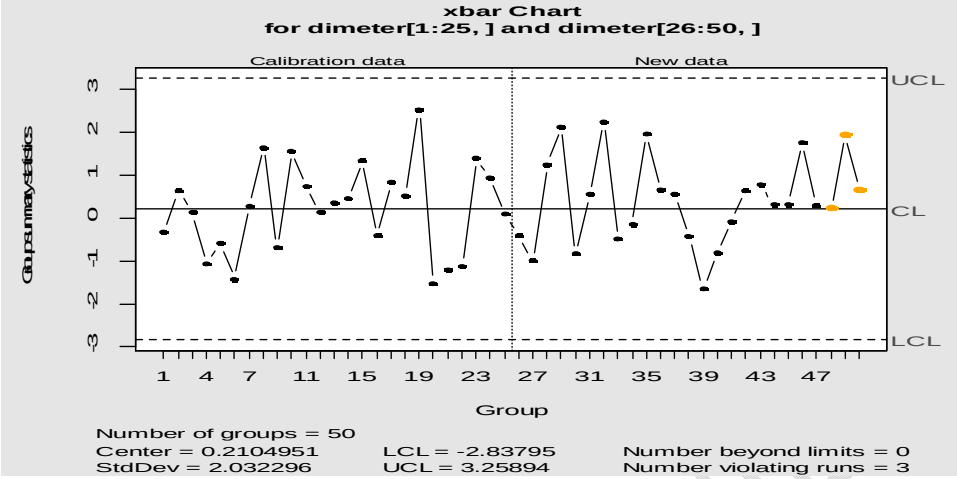


Figure (3) phase two for binomial principal component residual based on control chart.

**Analysis of Simulation Study**

In this section, we clarify the proposed technique of monitoring through a simulation study, so we check our data by condition index, variance inflation factor, and correlation matrix to see if this data has multicollinearity problem. So, the subgroups of the simulation study are in control, and Table 2 shows the value of the estimated coefficient of the model. Furthermore, in Figure (1) phase one obtained from the principal component, and the average run length($ARL_0$)is equal to7.94. And in Figure (2) is obtained from residuals, and$ARL_0$is equal to 7.86, and Figure (3) is the phase two control chart, so $ARL_1$ value is equal to 1.16. Where the sample size is N=200 and p is the number of independent variables, so p=4, then we divide N into 50 subgroup and subgroups, and the sample size is equal to 4. So, if we do a comparison by ARL measurement between Figure (1) and Figure (2), we find that the $ARL_0$ value of Figure (1) is better than Figure (2) because the $ARL_0$ value of the principal component is bigger than the binomial principal component residual based on the control chart.

**8. CONCLUSIONS**

All of binomial control chartsare in control.Furthermore, we present a new planning combining betweenbinomial regression and principal

componentanalysis. Furthermore, the binomial principal component residual based on control chart (BPCR) is used to monitor data processes. A lot of research has used the Shewhart control chart, but Yassin and Mohamed used Shewhart based on a residual control chart with Poisson regression, so the current paper proposes a different strategy interms of choosing a different model with a different treatment method.In future work, can use another control chart with a new approach and use another mythology to treatmulticollinearity problems, or we can alsouse missing values and draw control charts. A lot of ideascan be usedin this field.

**Reference**


1. Amiria, A., Yeh, B. A., and Asgari, A. (2016). Monitoring Two-Stage Processes with Binomial Data using Generalized Linear Model-Based Control Charts. Quality Technology & Quantitative Management, Vol. 13, No. 3, P. 241–262.
2. Bilen, C., Chen, X., Khan, A., and Yadav, P.O. (2007). Multiple Regression Control Chart Integrated with Principal Component Analysis. Industrial Engineering Research Conference G. Bayraksan.
3. Dunn, K. B., and Smyth, K. G. (2018). Generalized Linear Models with Examples in R. Springer Science+Business Media, LLC.
4. Filho, M. D. and O. M. A. Sant'Anna, (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.
5. Nancy, M., Joshi, H., and Dhandra, V, B. (2023). Regression Control Charts-A Survey. Journal of Pharmaceutical Negative Results, Vol. 14, No. 3, P. 1079- 1086.
6. Pardo, A. J., Pardo, L., and Pardo, C. D. M. (2007). A simulation study of a nested sequence of binomial regression models. Statistics, Vol. 41, No. 3, P. 253–267.
7. Rashid, A. N., Mokhtar. F.S., Wan Hassan, S.W., and Che Hussin, E.W., (2013). Regression Residual Control Chart for Monitoring Academic

Performance of Students in Higher Learning Institution. International Conference on Computing, Mathematics and Statistics.

8. Soares, G., Gomes, S., and Ludermir, B, T. (2006). Feature Selection for Neural Networks through Binomial Regression. Neural Information Processing, 13th International Conference, Hong Kong, China, October 3-6, Proceedings, Part I, Springer-Verlag Berlin Heidelberg.
9. Yassin, S. M. and 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, P. 2301–2326.