American Journal of Theoretical and Applied Statistics
2020; 9(4): 90-100
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20200904.12
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

SciencePG
Science Publishing Group

Generalized Regression Control Chart for Monitoring Crop Production

Olatunji Taofik Arowolo, Matthew Iwada Ekum[*]

Department of Mathematics & Statistics, Lagos State Polytechnic Ikorodu, Lagos, Nigeria

Email address:
saka_1972@yahoo.com (O. T. Arowolo), matekum@yahoo.com (M. I. Ekum)
[*]Corresponding author

To cite this article:
Olatunji Taofik Arowolo, Matthew Iwada Ekum. Generalized Regression Control Chart for Monitoring Crop Production. *American Journal of Theoretical and Applied Statistics*. Vol. 9, No. 4, 2020, pp. 90-100. doi: 10.11648/j.ajtas.20200904.12

Received: January 31, 2020; **Accepted**: April 7, 2020; **Published**: May 28, 2020

Abstract: Recently, Nigeria focused on Agriculture as a way to diversify her economy. Crop production, which is a proxy to measure agricultural output is considered very important. So, controlling crop production (output) among states in Nigeria is very key. In this study, the generalized regression control chart was used rather than the conventional control chart. The conventional control chart does not put into consideration factor(s) that affect crop production. The generalized regression control chart considers the factor (independent variable) that affect crop production (dependent variable). The normal distribution is a special case of the generalized regression control chart. The possibility of using Weibull regression and other non-normal models were considered. In this research, Gaussian distribution was used as the underlying distribution because it fitted the crop production data. The cost of seed/seedling was selected from a set of independent variables, because it is most significant among other independent variables. The data were collected from secondary sources, precisely National Bureau of Statistics (NBS). All the 36 states in Nigeria, including the Federal Capital Territory (FCT) were involved in the study. The result of the generalized regression control chart showed that crop production is not in control in Nigeria, which was traced to assignable cause of variation in FCT, Abuja. This implied that FCT, Abuja produced below the lower control limit of crop production, despite the relative cost of seed/seedlings.

Keywords: Conventional Control Chart, Crop Production, Exponential Family, Gaussian Regression Model, Generalized Regression Control Chart

1. Introduction

In industry, the quality of the expected product and the actual goods manufactured should be the same, but sometimes some variations are found, thus producing deviations that are random or assignable. Statistical quality control and Six Sigma are ways of controlling the quality of products by reducing such deviations from standard.

Shewhart [1] was the first author to propose control charts and since then a lot of charts have been established in monitoring and controlling different production processes. A *conventional* Shewhart control chart is plotted with the mean of process observations at different points with a pair of control limits. In developing a Shewhart control chart, one of the important assumptions is that the distribution function of the underlying process data is normal and the other assumption is that process data are independently distributed.

Statistical quality control (SQC) was defined by Montgomery [2] as a technique of analysing the process, setting standards, comparing performance, verify and study deviations, to seek and implement solutions, analyse the process again after the changes, seeking the best performance of machinery and or persons. In statistics, *control charts* are statistical process control tools used to monitor and control a process. The process is said to be in control if all the points plotted fall within the upper and lower control limits.

Alwan and Roberts [3] showed that about 85% of a sample of 235 control charts displayed incorrect control limits, and Karaoglan and Bayhan [4] mentioned that more than half of these displacements were due to violation of the independence assumption, this implies that the remaining half of these displacements could be due to violation of normality assumption. So, for a conventional control chart to display a correct control limits, the process data must be normally distributed. The conventional control chart does not put into consideration factors that may affect variable to be controlled. The generalized regression control chart considers the factor that can affect the variable of interest. This is enough reason why the conventional control chart will not be appropriate for modelling such variable as crop production, which is not just dependent on time but also on other factors. It might even increase or decrease with time, with varying mean and variance. So, the need for a generalized regression control chart for such data is necessary. Regression will take care of the factor that affects the response variable. So, combining regression model and the conventional control chart will give a regression control chart.

The regression control chart was first published in 1955, in a book titled "Statistics: a new approach" by Wallis. and Roberts [5]. Mandel [6] popularised it and in 1969 applied it to monitor and control man hours spent in dispatching of mails in post office, regressed on the pieces of mail handled [7]. Mandel [7] mentioned that the regression control chart has proved useful for a variety of postal management problems and offers possibilities for more widespread applications in government, business, industry and agriculture.

Statistical model is a description of the probability distribution of random variables which can be assumed to represent a real world phenomenon [8]. A linear regression model describes the relationship of covariate x and a continuous response variable Y [8]. One important assumption of linear regression model is that the distribution of the response variable (y) and the error term are normal.

Some examples of the application of regression control chart to autocorrelated processes were given by Karaoglan [4]. The regression control chart considers the factor that can affect the dependent variable but assumes the normal distribution as a default distribution. The generalized regression control chart however, assumes any distribution, which a Gaussian (normal) distribution is a special case.

Thus, this paper, however, focused on applying generalized regression control chart as a means of setting standards in the controlling and monitoring crop production among states in Nigeria, to guide against over or under production. This combination of the conventional control chart and generalized regression model is an improvement to the work of Mandel [9] by selecting an independent variable that mostly affected the variation in the dependent variable, and also opening ground for generalized regression control chart (Weibull regression, Gamma regression, Rayleigh regression, Exponential regression and so on).

The remaining part of this paper is organized as follows. Section two comprises control charts. In section three, generalized control chart was presented as well as parameter estimation for the generalized regression control chart, and establishing the control chart limits. Section four consist of the application of the generalized regression control chart to regression of crop production on cost of seed/seedlings. Section five contains the concluding remarks.

2. Control Charts

2.1. Conventional Control Chart

The control chart was invented by Walter A. Shewhart, while working for Bell Labs in the 1920s. What makes the control chart such a useful tool is the fact that the chart can reveal the amount of variation by time, thus enabling the user to observe patterns for interpretation and the discovery of changes in the process. Grant and Leavenworth [10] showed an example of the use of Stewarts, use as the tool of the analysis on the tolerance of rheostat. In addition, conducting a control chart analysis prior to conducting a six sigma calculation allows the six sigma calculation to reveal the true inherent process capability [11], while Woodall et al. [12] stated that statistical quality control is a collection of tools that are essential in quality improvement activities.

An example of the conventional control chart is depicted in Figure 1. The average characteristic ($\bar{x}$) is plotted against time. This conventional control chart is useful if a large variation is not suspected to be caused by another variable. If the characteristic variable is affected by another variable, then the conventional control chart will not be appropriate, hence, the need for a generalized regression control chart, of which Gaussian regression control chart is a special case.

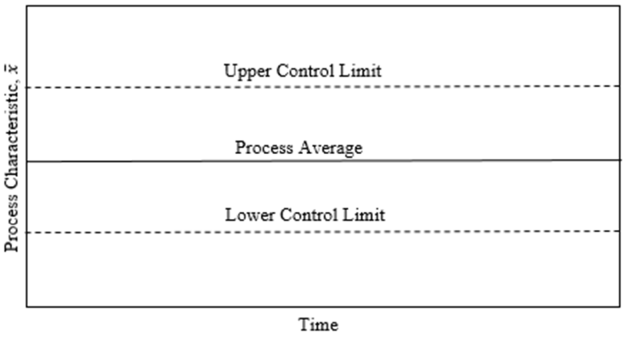

Figure 1. *Conventional Control Chart.*

2.2. Regression Control Chart

The conventional control chart uses a line of average performance with control limits parallel this central line. The upper control, lower control and central lines all parallel to the horizontal axis, implying that a single average is being controlled [9]; and Mandel [9] stated that the regression control chart has the following elements, which distinguished it from the conventional control chart.

1. It is a model that controls a varying average rather than a constant average. The central line is the regression line.
2. The control limits are parallel to the regression line rather than to the horizontal axis. The scatter plot is very useful here. Three lines are drawn on the scatter plot, the central line (line of best fit), upper control limit and lower control limit. The three lines are expected to slant upward or downward.
3. The computation for the construction of the regression

control chart is time consuming compared to the conventional control chart, but with the help of modern high speed computers, the problem of computation is solved. The standard deviation of the regression control chart is the standard error estimate of the regression line. It is the standard deviation estimate based on the deviation of the observed values about the regression line. It is quite different from the standard error of a predicted value of the dependent variable.

4. The regression control chart is appropriate for a number of applications, which the conventional control chart does not readily applies. It provides the basis of measuring the gains or loss in the response variable, for predicting and forecasting the response variable and scheduling the covariate resources.

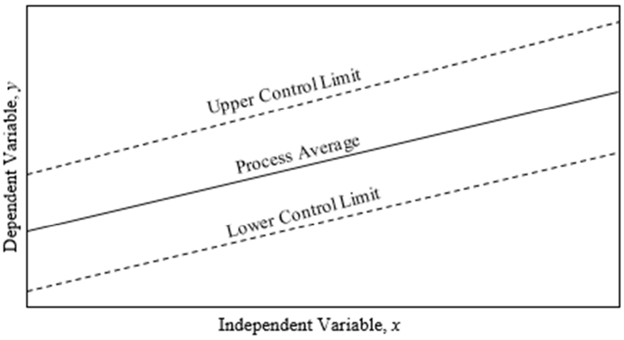

Figure 2. *Regression Control Chart.*

The schematic representation of the conventional control chart and the regression control chart are shown in Figure 1 and 2 respectively. The two charts look alike but are different. The conventional control chart is univariate, while the regression control chart is bivariate. The two figures are a replica of the one in [9].

3. Research Methodology

3.1. Data Description

The data collected initially are panel data consisting of 37 cross-sections (the states in Nigeria including FCT, Abuja) and 10 periods (10 years from 2006 60 2015). The average of the 10 years was computed for each cross-section, reducing the panel data to cross sectional data. The data was collected from the administrative records and publications of National Bureau of Statistics (NBS), through the two data collection infrastructure; National Integrated Survey of Household (NISH) and National Integrated Survey of Establishment (NISE). NISH Master Sample was constructed from the frame of EAs of 2006 Housing and Population Census by National Population Commission (NpopC). The household listing of the EAs were stratified into farming and non-farming household and the sample size is taken from the farming through randomization. (See [13], [14]). The data collected are crop production (Y), total area cultivated (X_1), fertilizer usage (X_2), rural employment in crop production (X_3) and cost of seed/seedlings (X_4). Generalized regression line is fitted to this historical data, establishing limits around the regression line.

3.2. Generalized Regression Control Chart

The generalized regression control chart has all the attributes of the regression control chart of Mendel [9]. The difference between these two charts is the difference between the ordinary regression model and generalized regression model. The formal assumes normality of the response variable and the error term, while the later assumes any distribution other than the normal distribution. So, the regression control chart is a special case of the generalized regression control chart. In the ordinary regression control chart by Mendel [9], it is assumed that the response variable, y values are linearly related to the covariate, x values. For each specific x value, it is assumed that the y values are normally and independently distributed with a mean value estimated from the regression line, and with a standard error, which is independent of the values of x and it is estimated from the deviations of the actual observations, Y from the $\hat{Y}$ estimated from the regression line. The generalized regression control chart also assumed that the y values are independently distributed with a mean value estimated from the regression line, but are not necessary normally distributed. A good example is the beta regression control chart (BRCC) by Bayer et al. [15].

3.2.1. Linear Model

In statistics, a multiple linear regression model describes the relationship of a continuous response variable, Y, and a covariate, X. This model is defined as

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i \quad (1)$$

The model in equation (1), if k =1, we have the simple linear regression model given by.

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad (2)$$

where β_0 is the intercept term, β_1 is the regression coefficient for variable X and e_i is the error term. Assume that the error terms are random, independent and normally distributed with mean 0 and variance σ^2, i.e $e_i \sim N(0, \sigma^2)$, $i = 1, 2, ..., n$. Note that the variance is independent of x. The error term, e_i, in equation (1) is written explicitly. It is also possible to write the model in equation (2) without explicitly specifying the error term, e_i.

$$E(Y_i \mid x_i) = \mu_i = \beta_0 + \beta_1 x_i. \quad (3)$$

The model in equation (3) specifies the expected value of Y conditional on x. Equation (3)

does not specify how the values of Y vary around the expected value $E(Y_i \mid x_i)$. By defining the $\mathrm{Var}(Y_i) = \sigma^2$, we obtain a model equivalent to model specified in equation (3).

If Y_i is normal, then $Y_i \sim$ N($\beta_0 + \beta_1 x_i$, σ^2).

The linear model in (3) is transformed to a generalized linear model by letting $\mu_i = g(\mu_i)$, so that equation (3) becomes

$$g(\mu_i) = \beta_0 + \beta_1 x_i = \eta_i, \tag{4}$$

where $g(.)$ is the link function, which is a real-valued monotonic and differentiable function and the term η_i is the linear predictor. Canterle and Bayer [16] presented several possible choices for link functions such as logit, probit, log-log, complement log-log, Cauchy, and also parametric links. It is obvious that μ_i is the expected value of y, η_i is a linear combination of the predictors, and $g(.)$ defines the relationship between μ_i and η_i. Since $g(.)$ is monotonic, then the relationship of μ_i and η_i is monotonic as well. Thus, the inverse of $g(.)$ is given as

$$\mu_i = g^{-1}(\eta_i), \tag{5}$$

which is an alternative to the linear model. Thus, the linear model is a special case of the generalized linear model, if $g(\mu_i) = \mu_i$. If the independent variables are more than one, then equation (4) becomes

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{6}$$

For equations (4) and (6) to be possible, some assumptions must hold for Y_i in the model. The distribution of Yi must belong to the exponential class of family, they must be mutually independent, and have expected value $\mu_i = E(Y_i)$, which depends on a linear predictor $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$ through a monotonic and differentiable link function g(.) such that $\mu_i = g^{-1}(\eta_i)$. The exponential class of family has a probability density function given by

$$f(y_i;\theta_i,\phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c\left(y_i, \frac{\phi}{a_i} \right) \right\}, \tag{7}$$

where θ_i and ϕ are location and scale parameters respectively, and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i,\phi)$ are known functions. Since the variation in *Yi* is distribution with exponential family of distribution, then it has mean and variance given by

$$E(y_i) = \mu_i = b'(\theta_i) \tag{8}$$

and

$$Var(y_i) = \sigma_i^2 = b''(\theta_i) a_i(\phi), \tag{9}$$

where equations (8) and (9) are the mean and variance respectively of random variable y. Also, $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$ respectively. With $a_i(\phi) = \frac{\phi}{a_i}$, the variance in equation (9) becomes (10).

$$Var(y_i) = \sigma_i^2 = b''(\theta_i)\frac{\phi}{a_i}. \tag{10}$$

As mentioned earlier, the second aspect of the generalization is that instead of modeling the mean, as μ_i, we use a one-to-one continuous differentiable transformation $g(\mu_i)$ given as

$$\eta_i = g(\mu_i). \tag{11}$$

The function $g(\mu_i)$ is called the link function. It is further assumed that the transformed mean follows a linear model, so that equations (4) and (6), which is equated to (11) is written in matrix form as

$$\eta_i = X_i'\beta. \tag{12}$$

Since the link function is one-to-one, we can invert equation (12) to obtain equation (5), making μ_i the subject of the formula. It should be noted that the response variable Y_i was not transformed but rather its expected value μ_i.

3.2.2. Gaussian Regression Model

Recall from equation (1), if Y follows a normal distribution with mean, μ and variance, σ^2. then its pdf is given by

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \right], -\infty \leq y \leq \infty \tag{13}$$

The same procedure used here for Gaussian (normal), can be used to achieve everything other distribution belonging to the class of exponential family.

Equation (13) can be rewritten as

$$f(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-ln\sigma - \frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \right), \tag{14}$$

Take the log of (14) to have

$$lnf(y) = ln\left(\frac{1}{\sqrt{2\pi}}\right) - ln\sigma - \frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2}, \tag{15}$$

Take back the exponential of (15) to have the desired exponential class of distribution, given by

$$f(y,\mu,\sigma) = exp\left[ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{y^2}{2\sigma^2} + \frac{\mu y}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - ln\sigma \right]$$

$$f(y,\mu,\sigma) = exp\left[ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{\mu y - \frac{1}{2}y^2 - \left(\frac{1}{2}\mu^2 + \sigma^2 ln\sigma\right)}{\sigma^2} \right] \tag{16}$$

By comparing equation (16) to (7), we have that

$$b(\theta_i) = \frac{1}{2}\mu^2 + \sigma^2 ln\sigma$$

$$\theta_i = \mu$$

$$a_i(\phi) = \sigma^2$$

Mean and Variance of Y

Recall from equation (8), we have $b(\theta_i) = \frac{1}{2}\mu^2 + \sigma^2 ln\sigma$, so that $b(\theta_i) = \frac{1}{2}(\theta_i)^2 + \sigma^2 ln\sigma$

So,

$$b'(\theta_i) = \theta_i$$

Since, $\mu_i = \theta_i$

$$E(y) = b'(\theta_i) = \mu_i \tag{17}$$

Also, recall from equation (8) that $Var(y_i) = b''(\theta_i)a_i(\phi)$.

But $a_i(\phi) = \sigma^2$ and $b''(\theta_i) = 1$. So that

$$Var(y) = \sigma^2 \tag{18}$$

Thus, equations (17) and (18) are the mean and variance of *Y* respectively, where *Y* is normally distributed.

The link function of Normal distribution is given by equation (19)

$$g(\mu_i) = \mu_i \tag{19}$$

where $g(\mu_i)$ is the link function, and μ_i is the mean.

The generalized linear model of Normal distribution is given by equation (20)

$$g(\mu_i) = \mu_i = \beta_0 + \beta_1 x_i = \eta_i \tag{20}$$

Maximum Likelihood Estimation for Normal Regression Parameters

From the pdf in equation (16), the log-likelihood is given as

$$l = lnL(y,\mu,\sigma) = \sum_{i=1}^{n}\left[ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{\mu y - \frac{1}{2}y^2 - \left(\frac{1}{2}\mu^2 + \sigma^2 ln\sigma\right)}{\sigma^2}\right] \tag{21}$$

Re-write equation (21) in terms of β to have

$$l = lnL(y_i,\theta_i,\sigma) = \sum_{i=1}^{n}\left[ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{\beta_0 y_i + \beta_1 x_i y_i - \frac{1}{2}y_i^2 - \frac{1}{2}(\beta_0 + \beta_1 x_i)^2 - \sigma^2 ln\sigma}{\sigma^2}\right] \tag{22}$$

Differentiate equation (22) partially with respect to β_0 and β_1 to have equations (23) and (24) respectively.

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i - \frac{n\beta_0}{\sigma^2} - \frac{\beta_1}{\sigma^2}\sum_{i=1}^{n} x_i \tag{23}$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i y_i - \frac{\beta_0}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{\beta_1}{\sigma^2}\sum_{i=1}^{n} x_1^2 \tag{24}$$

Equate (23) to zero and solve to have

$$\hat{\beta}_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i\right) \tag{25}$$

Also, equate (24) to zero and solve to have

$$\hat{\beta}_1 \sum_{i=1}^{n} x_1^2 = \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_1^2 + \left(\sum_{i=1}^{n} x_i\right)^2} \tag{26}$$

Thus, equations (25) and (26) are the unbiased estimates of β_0 and β_1 respectively.

This process can be used to derive the parameter estimates of other member of exponential family. However, in a situation where the differentiation looks difficult or not in close form, we can use the equation defined by [8] to obtained the first derivative of the log-likelihood function of the exponential family defined in equation (6) in terms of β as

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{27}$$

The regression parameters of other distributions that are member of the exponential family can also be derived using (27). This will set the pace for generalized regression control chart. Other examples could be Gamma regression control chart, Weibull regression control chart, Rayleigh regression control chart, Exponential regression control charts.

3.3. Establishing the Regression Control Chart

Using the generalized regression line derived using maximum likelihood method, and twice the standard error of estimate (i.e, $2S_e$), the control chart, with control limits set at 2 standard deviations above and below the generalized regression line are given by

$$\text{The Upper Control Limit (UCL)} = \hat{Y} + 2\sigma \tag{28}$$

$$\text{The Lower Control Limit (LCL)} = \hat{Y} - 2\sigma \tag{29}$$

The value of σ is unknown but is estimated with S_e. The use of 2σ or 3σ in equations (28) and (29) is a management decision on the level of quality desired. To construct the generalized regression control chart used in this research, we estimated for the following.

N = number of pairs of values ($x_i y_i$) (i = 1 to 37)

S_e = standard error of estimate of the regression based on history-period data

r = Correlation coefficient between x and y.

r^2 = Coefficient of determination (explained variation)

$\overline{y}$ = Average response from observed values

S_y = Standard deviation of response variable

$\overline{x}$ = Average independent variable from observed values

S_x = Standard deviation of the independent variable

$\frac{S_e}{\overline{y}}$ = Percentage coefficient of variation of the observed data

β_0 = Intercept on y axis

β_1 = Slope of the regression line

The generalized regression control chart is then set as follows.

The central line is E(y) = $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$, where the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed from equations (25) and (26) respectively.

The upper control limit (UCL) = $\hat{Y} + 2\sigma$

The Lower Control Limit (LCL) =$\hat{Y} - 2\sigma$,

where σ is estimated by equations S_e, which is standard error derived from the regression line.

3.4. Measuring Progress

The difference between expected and actual crop production ($\hat{y}_i - y_i$) can be plotted against time like the conventional control chart. Also, the cumulative crop production in excess or less at the end of a given period can be easily determined and tested for significance. The points that are out of control are not used in this calculation, because they represented assignable causes.

Test for Significant Gain or Loss in Crop Production

Statement of Hypothesis

$$H_0 = \sum_{i=1}^{N}\left(\hat{y}_i - y_i\right) = 0$$

$$H_1 = \sum_{i=1}^{N}\left(\hat{y}_i - y_i\right) \neq 0$$

Level of significance, α = 5% = 0.05

Test Statistic:

$$t = \frac{cumulative\ crop\ production\ increase\ (or\ decrese)\ to\ date}{standard\ error\ of\ cumulative\ crop\ production\ increase\ (or\ decrese)}$$

$$t = \frac{\sum_{i=1}^{m}\left(\hat{y}_i^{'} - y_i^{'}\right)}{S_e\sqrt{\frac{m^2}{n} + m + \frac{\left[\sum_{i=1}^{m}\left(x_i^{'} - \bar{x}\right)\right]^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}} \quad (30)$$

where

$x_i^{'}$ = explanatory variable observations made after the regression control was established

$y_i^{'}$ = cash crop production corresponding to the explanatory variable x_i'

$\hat{y}_i^{'}$ = predicted cash crop production corresponding to the explanatory variable x_i'

x = explanatory variable used to establish the regression control chart

$\bar{x}$ = mean of the x_i

m = number of x_i' values (points within control)

n = number of pairs of values ($x_i y_i$) (i = 1 to 37)

S_e = standard error of estimate of the regression based on history-period data

S_e is the standard error based on the estimate on the deviations of the observed values about the regression line. It should not be confused with the standard error of a predicted value of the dependent variable.

It should be noted that when n is large relative to m, then equation (30) can approximated by (31).

$$t \approx \frac{\sum_{i=1}^{n}\left(\hat{y}_i^{'} - y_i^{'}\right)}{S_e\sqrt{\frac{n^2}{N} + n}} \quad (31)$$

Decision Rule: Reject the null hypothesis, H_0, if the calculated t-value is greater than the critical t-value ($t_{\alpha/2, N-1}$). Note that N-1 is the degrees of freedom.

4. Result and Discussion

4.1. Exploratory Data Analysis

Table 1. *Panel Data of Key Variables.*

I	State	Year	Prod	Area	Fertilizer	Employ	Cost
1	Abia	2006	1607.4	225.36	378.61	1685	1504.67
2	Abia	2007	1529.5	226.84	367.37	1743	1868.96
3	Abia	2008	1142.8	220.1	356.13	1750	1645.87
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
68	Benue	2013	12368.3	2129.81	3937.44	1573	17229.72
69	Benue	2014	13023.7	2144.87	4331.18	1730	18952.69
70	Benue	2015	13688.2	2162.72	4764.3	1903	20847.96
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
369	FCT Abuja	2014	231.9	81.63	295.35	366	629.62
370	FCT Abuja	2015	243.8	82.31	324.88	403	692.58

Table 1 is a longitudinal data with 37 cross-sections (states) and 10 time periods, spanning 370 data points. The yearly average data for each state is used to construct the regression control chart.

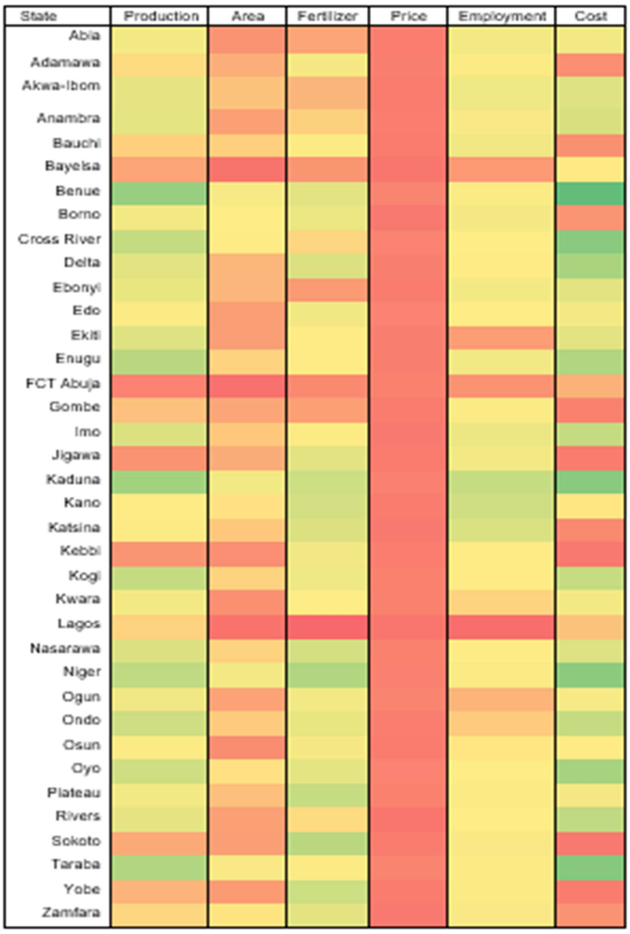

Figure 3. *Heat Map of the Data.*

The heat map displayed in Figure 3 shows that cost has a high correlation with production. The green colour indicates a very high value, while red indicates a very low value. So the variation from green to red shows how the values reduce from highest to lowest. If you look at the heat map very well, you will discover that the states with green cells for production also have green cells for cost, and the ones with red cells for production also have red cells for cost as compared with other variables.

Table 2. *Summary Data.*

	Production	Area	Fertilizer	Employment	Cost
Min.	284.00	63.00	13.21	52.19	120.40
1st	2218.00	366.80	756.89	914.20	304.60
Median	3887.00	522.10	2097.30	1266.59	1808.90
Mean	4310.00	641.10	2508.71	1562.32	3947.70
3rd	5394.00	698.80	3617.58	1800.40	6045.90
Max.	10331.00	1972.30	7928.01	5849.78	14981.80

Table 2 shows the summary statistics of the data collected for the analysis. One of these independent variables will be used to construct the regression control chart. The variable that contributes most to the variation in the dependent variable is selected. This can be determined from the multiple linear regression model.

4.2. Normality test of the Dependent Variable, Y (Crop Production)

Table 3. *Measure of Skewness and Kurtosis.*

Mean	Median	Std Dev	Skewness	Kurtosis
4309.67	3886.6	2636.77	0.466	-0.715

Table 3 shows that the skewness of the dependent variable (crop production) is 0.466 and the kurtosis is -0.715, which shows that the variable is non-Gaussian. Also, the histogram, QQ plot and boxplot all show that the variable is non-Gaussian. It is necessary that we subject the data to confirmatory test otherwise, the Gaussian regression model will not be relied upon, rather, the generalized regression model is appropriate. See also Figure 4.

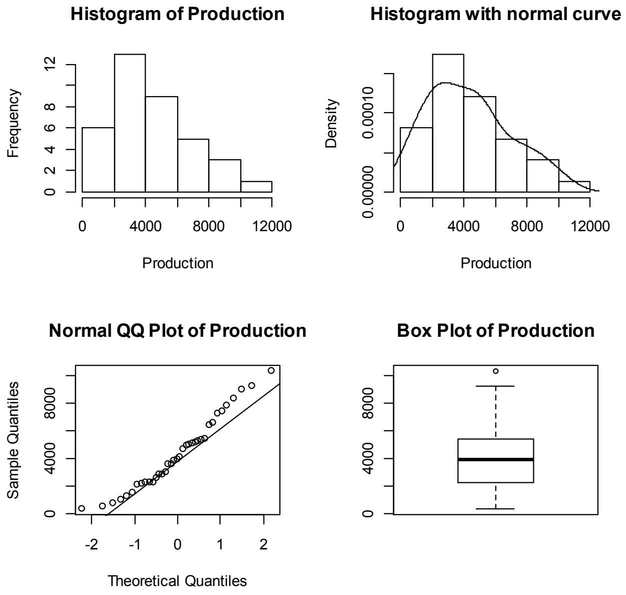

Figure 4. *Normality Test using plots.*

Table 4. *Gaussian Confirmatory Test for Crop production.*

	Kolmogorov - Smirnov	Shapiro-Wilk	Jarque – Bera	D' Agostino
Statistic	1	0.9595	1.9854	2.1652
P-value	0	0.1941	0.3706	0.3387

The result of the confirmatory test in Table 4 shows that Gaussian distribution adequately fit the data. This implies that the data is normally distributed as against the results from the exploratory data analysis, which earlier suggested that the data might not follow a Gaussian distribution. If the data is not Gaussian, then other non-Gaussian distributions like Gamma, Weibull, Rayleigh, Exponential and so on would be used.

4.3. Linear Regression Analysis

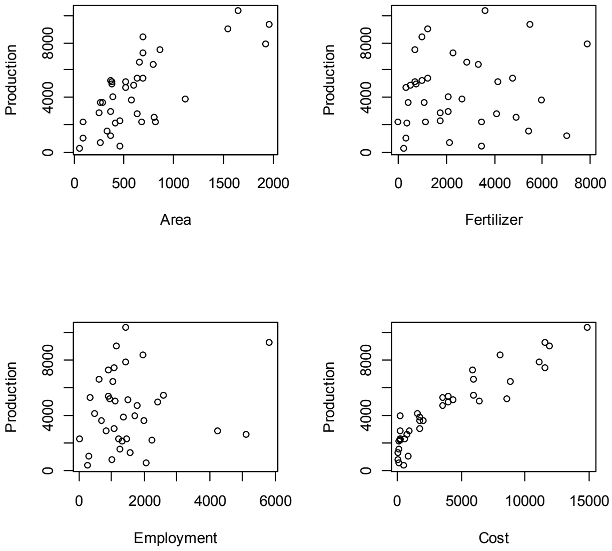

Figure 5. *Relationship between Crop Production and the Independent Variables.*

Table 5. *Parameter Estimate of Multiple Linear Regression.*

	Estimate	Std. Error	t-stat	P-value
Intercept	1910.628	317.879	6.011	0.000
Area	1.216	0.696	1.747	0.090
Employment	0.027	0.161	0.170	0.866
Fertilizer	-0.135	0.095	-1.421	0.165
Cost	0.485	0.063	7.651	0.000
Residual std. error: 980.9 on 32 DF, R^2: 0.877, Adjusted R^2: 0.862				
F-statistic: 57.04 on 4 and 32 DF, p-value: 0.000				

Table 5 shows the least squares parameter estimates of the multiple linear regression model. The multiple linear equation model is given by

$$\hat{y}_i = 1910.628 + 1.216x_{1i} + 0.027x_{2i} - 0.135x_{3i} + 0.485x_{4i} \quad (32)$$

where y is crop production, x_1 is area, x_2 is employment, x_3 is fertilizer and x_4 is cost. It is very obvious from Table 5 that x_4, that is, cost of seed/seedling is the most significant independent variable. Thus, x_4, will be used to control the variability in crop production (y). See also the scatter plots in Figure 5 for pictorial explanation.

Table 6. *Regression Parameter for Simple Linear Regression.*

	Estimate	Std. Error	t-stat	P-value
Intercept	2045	225.7	9.058	0.000
Cost	0.5737	0.03914	14.657	0.000
Residual std. error: 1001 on 35 DF, R^2: 0.8599, Adjusted R^2: 0.8559				
F-statistic: 214.8 on 1 and 35 DF, p-value: 0.000				

Table 6 shows the least square parameter estimates of the simple linear regression model. The table shows that both the intercept and the slope are significant. From Table 6, the simple linear regression model is given by

$$\hat{y}_i = 2045 + 0.5737x_{4i} \quad (33)$$

where 2045 is the intercept, meaning that the value of crop production when cost of seed/seedling is equal to zero is 2,045 thousand tons; and 0.5737 is the slope of the regression model and it implies that for each unit increase in cost of seed/seedling, crop production will increase by 0.5737 thousand tons (573.7 tons). The analysis shows that 85.99% of the variation in crop production can be explained by the variation in the cost of seed/seedling. Thus, there is a significant linear relationship between crop production and cost of seed/seedlings. Note that equation (32) cannot be used for the regression control chart because, the chart is a 2-dimensional plot, containing only a dependent variable on the vertical axis and an independent variable on the horizontal axis. So, equation (33) is appropriate.

4.4. Conventional Control Chart

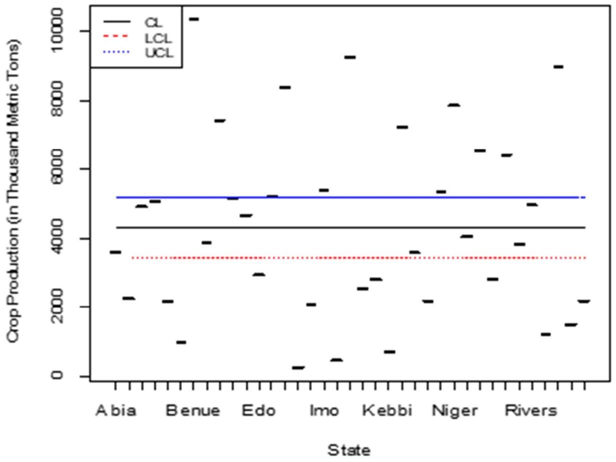

Figure 6. *Conventional Control Chart of Crop Production.*

It is obvious from Figure 6 that many points fall outside the control limits, which implies that inputs are also obviously not the same. Most of the states spent different amount on cost of seed/seedlings, which is not captured by the conventional control chart. Here the $CL = \bar{y}$, $LCL = \bar{y} - 2Se_y$, $CL = \bar{y} + 2Se_y$, where Se_y is the standard error of y.

4.5. Regression Control Chart

To establish a regression control chart in this study, data from agricultural data collected from [13] through the two data collection infrastructure; National Integrated Survey of Household (NISH) and National Integrated Survey of Establishment (NISE), which was first collected in 2006 census. The data is displayed in Table 1. Since a 2-dimensional plot involves only two variables, cost of seed/seedlings is selected as an independent variable among other independent variable as a result of its contribution and relationship with the dependent variable, crop production.

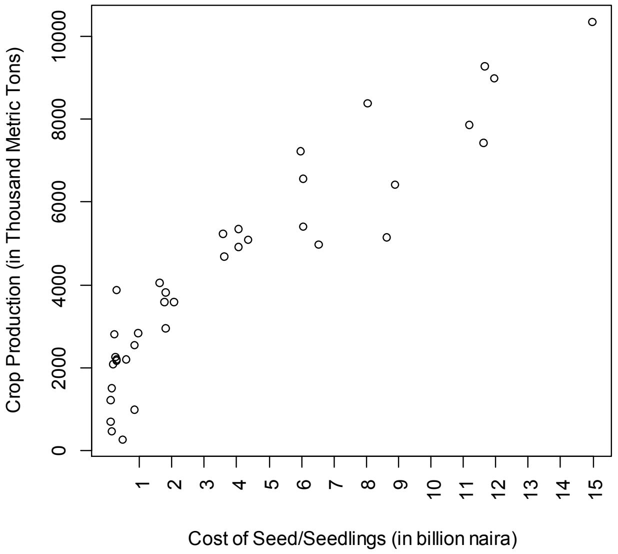

Figure 7. *Scatter Plot of Crop production on Cost of Seed/Seedlings.*

The first step is to use the data in Table 1 to plot the production against cost on a scatter diagram, which is shown in Figure 7. This scatter diagram is needed to primarily check on the linearity of the relationship and to detect atypical points. It should be noted that crop production depends on many other factors other than cost of seed/seedling, such as area crop cultivated, fertilizer consumption, employment in crop farming and so on, some of these factors vary and some are stable, but among the ones used in this study, cost of seed/seedling has most variability and explains the variation in production more than other variables. The points that depart from linear pattern are not easily detected. These points can be due to assignable causes of variation.

So, a good way to detect these points is through the regression control chart. It should be noted that each point is traceable to each state of the federation. A defaulted state can easily be detected and controlled.

The following values were computed using the data in Table 1, considering only production and cost.

N = 37 locations (states)

r = 0.9273

r^2 = 0.8599

S_e = 1001 thousand metric tons

$\overline{y}$ = 4309.7 thousand metric tons

S_y = 2636.768 thousand metric tons

$\overline{x}$ = 4309.667 million naira

S_x = 4261.814 million naira

$$cvy = \frac{S_e}{\overline{y}}\times 100 \;=\; cvy = \frac{1001}{4309.667}\times 100 = 23.2\%$$

It is now easy to compute the regression control chart. In this case, the centre line (CL) is the regression line in equation (33). The lower and upper control limits are CL-2S$_e$ and CL+2S$_e$ respectively. The narrower the limits, the higher the risk of false alarms. However, in this study, two-sigma is used.

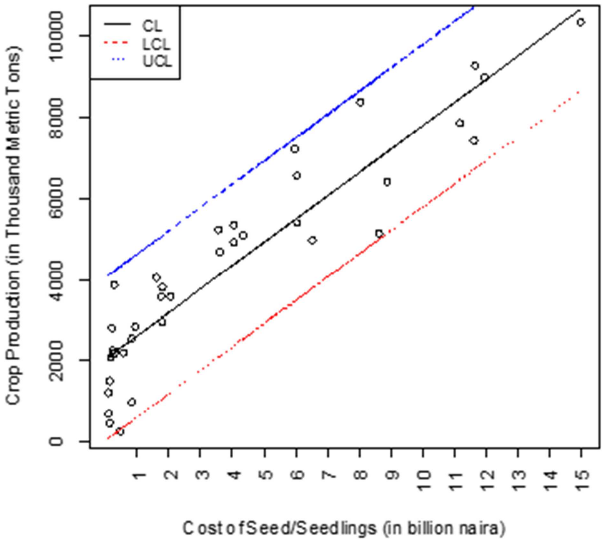

Figure 8. *Generalized Regression Control Chart of Crop Production on Cost of Seed/Seedlings.*

This control chart depicted in Figure 8 can be used in variety of applications. The first use of this regression control chart is to maintain control over performance of crop production in each state in Nigeria on continuous basis. For instance, if the cost of seed/seedling in 9 billion naira, what is the justification on the crop production performance for such state of the federation. Is there a gain or loss in productivity and is this performance acceptable? If the crop produced at this cost fall outside the control limits, then the performance is not acceptable, it can be counted as assignable causes of variation but if the point is within the control limits, then it is an acceptable performance and such variation can be attributed to chance. When performance is not acceptable, it is the duty of the management or people in authority to decide whether or not to investigate the cause of the variation.

It should be noted that in this study, points that fall below the lower control limit signifies under production, meaning that the cost of seed/seedling was not justified. On the other hand, if the points fall above the upper control limit, it signifies very high performance. This high performance should also be investigated due to the following reasons. Firstly, Other states can learn from them to see how they can have such a high performance, secondly it could be regarded as over production, if the demand is lower than the supply or if there is no good storage facilities or good market for the export of the excess production. The regression control chart is shown in Figure 4 for viewing.

4.6. Measuring Progress

The difference between the predicted $\hat{y}$ and observed y is a way to simplify regression control chart and getting additional information from it. These difference can be plotted against time just like the convention control chart. In this case, runs and trends can easily be observed, overcoming analysing results from cluttered scatter diagram. The cumulative production gain or loss can also easily be determined and tested for significance using *t*-test.

Table 7. *Gain or Loss Table for Points within control Limits.*

	Observed Production	**Predicted Production**	**Gain or Loss**	**Cum. Gain or Loss**	**Cum. Deviation**
State	(y_i')	$(\hat{y}_i')$	$(\hat{y}_i' - y_i')$	$\sum_{i=1}^{m}(\hat{y}_i - y_i')$	$(x_4' - \bar{x}_4)$
Jigawa	483.524	2124.754	1641.230	1641.230	-3808.301
Kebbi	701.825	2113.811	1411.986	3053.216	-7635.677
Bayelsa	1003.626	2536.301	1532.675	4585.891	-10726.677
Sokoto	1233.535	2116.076	882.541	5468.432	-14550.081
Yobe	1503.943	2127.797	623.854	6092.286	-18353.080
Gombe	2093.291	2147.913	54.622	6146.908	-22121.016
Bauchi	2187.851	2202.364	14.514	6161.422	-25794.043
Lagos	2198.734	2392.033	193.299	6354.720	-29136.478
Zamfara	2217.938	2202.905	-15.033	6339.687	-32808.563
Adamawa	2259.972	2193.041	-66.931	6272.756	-36497.839
Kano	2540.794	2522.616	-18.178	6254.578	-39612.667
Katsina	2806.883	2173.070	-633.813	5620.765	-43336.755
Osun	2851.752	2586.452	-265.300	5355.465	-46340.318
Edo	2962.887	3071.428	108.541	5464.007	-48498.568
Abia	3581.422	3237.640	-343.782	5120.224	-50367.109
Kwara	3590.656	3063.131	-527.525	4592.700	-52539.820
Plateau	3807.244	3082.577	-724.667	3868.033	-54678.638
Borno	3886.597	2219.527	-1667.070	2200.963	-58321.750
Ogun	4064.352	2967.553	-1096.799	1104.164	-60661.054
Ebonyi	4674.469	4122.835	-551.634	552.529	-60986.701
Akwa-Ibom	4906.204	4375.400	-530.804	21.725	-60872.128
Rivers	4971.138	5782.630	811.492	833.217	-58304.753
Anambra	5098.631	4550.876	-547.755	285.462	-57884.325
Delta	5139.580	6992.090	1852.510	2137.972	-53208.862
Ekiti	5223.452	4103.157	-1120.295	1017.677	-53568.807
Nasarawa	5346.320	4381.128	-965.192	52.485	-53444.250
Imo	5394.490	5513.452	118.962	171.447	-51346.052
Oyo	6404.884	7144.139	739.255	910.702	-46405.568
Ondo	6565.337	5512.423	-1052.914	-142.212	-44309.164
Kogi	7223.913	5466.072	-1757.841	-1900.053	-42293.550
Cross-River	7433.572	8717.602	1284.030	-616.023	-34610.520
Niger	7857.299	8461.275	603.977	-12.046	-27374.269
Enugu	8378.132	6666.066	-1712.066	-1724.112	-23267.066
Taraba	8974.789	8893.737	-81.052	-1805.164	-15277.034
Kaduna	9273.836	8729.668	-544.168	-2349.332	-7572.974
Benue	10330.816	10640.180	309.364	-2039.968	3461.107

Where y_i', $\hat{y}_i'$, x_4' and $\bar{x}_4$ are the values of the observed production for points in control, and their corresponding predicted y, cost, and average cost respectively. Table 7 contains 36 data points because 1 data point is below the lower control limit (LCL). This point is FCT-Abuja, and it is deleted from the table as the assignable cause of variation. The cumulative gain or loss can be used to determine whether the regression line and control chart limits need revision.

At the end of a particular period, or within a fiscal year, the monitoring team can check the level of performance to determine if it has changed significantly. This can be determined by a student's t test given in equation (30).

$$t = \frac{\sum_{i=1}^{m}\left(\hat{y}_i' - y_i'\right)}{S_e\sqrt{\frac{m^2}{n} + m + \frac{\left[\sum_{i=1}^{m}\left(x_i' - \bar{x}\right)\right]^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}}$$

$$t = \frac{-2039.968}{1001\sqrt{\frac{36^2}{37} + 36 + \frac{3461.107^2}{653870064}}} = \text{-0.24180}$$

The formula for the t-test above can be approximated to the one in equation (31) if *m* and *n* are close. The results obtained from the two formulas are equal when rounded up to 4 decimal places. Thus, the formula below is a good approximation for the t-test.

$$t \approx \frac{\sum_{i=1}^{m}\left(\hat{y}_i' - y_i'\right)}{S_e\sqrt{\frac{m^2}{n} + m}}$$

$$t = \frac{-2039.968}{1001\sqrt{\frac{36^2}{37} + 36}} = \text{-0.24183}$$

This calculated value of *t* can be compared with 2 or to the critical value of *t* checked on the tables at *n*-2 degrees of freedom. Alternative, the p-value can be derived from the R code and it is given by dt(-0.2418, 35) = 0.38438. Since the p-value is greater than the level of significance (α = 0.05), then we cannot reject the null hypothesis, and conclude that the cumulative net gain is not statistically different from zero. Note the following recommendations. If the t-test is significant, a new control chart would have been drawn based on current's year data, showing new performance level. This shows that there is gain but the gain is not significant, it could be due to chance. It is a gain because actual productivity is greater than the expected productivity.

5. Concluding Remarks

The generalized regression control chart is a combination of generalized regression model and control charts. The regression line is the central line, which is applicable to linear and non-linear models as well as generalized regression model, depending on the shape of the data under consideration. The crop production data used in this work appeared to be non-Gaussian from the histogram and boxplot, but the confirmatory test shows that it is Gaussian, so it will not be necessary to consider other distributions since Gaussian shows a good fit.

Based on the result of the analysis, we conclude that there is a significant relationship between crop production and the independent variable (cost of seed/seedlings). The result shows that among the four independent variables, cost of seed/seedling is the most significant. The regression line is fitted and the regression control chart fitted using the regression line as the central line (CL), and CL±2Se as the control limits.

The regression control chart is out of control as a result of a point just a little below the lower control limit. This point is FCT, Abuja. This shows that crop production in FCT, Abuja does not measure up to the cost incurred in seed/seedlings. To make adjustment and use the control chart for monitoring crop production subsequently, this point out of control (FCT-Abuja) was deleted from the table, since it is an assignable cause of variation. The cumulative gain or loss table developed can be used to determine whether the regression line and control chart limits need revision. This model will capture the data during production process and gives alarm at every deviation (variation) in the production line at the end of each year.

Major stakeholders and policy makers should work with the available statistical models to monitor the expected crop production in Nigeria by conscious effort. Cost of seed/seedling is a very important factor to be considered, when measuring crop production level at any point.

References

[1] Shewhart, W. A. (1931). Statistical Method from an Engineering Viewpoint. *Journal of the American Statistical Association*, 26, 262-269.

[2] Montgomery, D. C., and Woodall, W. H. (1997). A Discussion of Statistically-Based Process Monitoring and Control.

[3] Alwan, L. C. and Roberts, H. V. (1988). Time Series Modeling for Statistical Process Control. *Journal of Business and Economic Statistics*, 6(1), 87–95.

[4] Karaoglan, A. D. and Bayhan, G. M. (2014). A regression control chart for autocorrelated processes. *Int. J. Industrial and Systems Engineering*, 16(2), 238-256.

[5] Wallis, W. A. and Roberts, H. V. (1956). *Statistics: A New Approach*. The Free Press, Chicago, III, 549-553.

[6] Mendel, B. J. (1967). The Regression Control Chart – A Multipurpose Tool of Management, Universal Postal Union, International Bureau, Bern, Switzerland, September, 1967.

[7] Mendel, B. J. (1967). Statistical Programs of the United States Post Office Department. *Industrial Quality Control*, 23(11), 535-538.

[8] Karvanen, J. (2009). *Generalized linear models*. Available online at www.wiki.helsinki.fi. University of Helsinki, spring.

[9] Mendel, B. J. (1969). The Regression Control Chart. *Journal of Quality Technology*, 1(1), 1-9.

[10] Grant, E. L., and Leavenworth, R. S. (1980). *Statistical Quality Control*, 5th ed., McGraw-Hill, New York.

[11] Juran, J. M. and Gryna, F. M. (1988). *Quality control handbook*, (4th. ed.), McGraw-Hill, New York.

[12] Woodall, W. H., Spitzner, D. J., Montgomery, D. C. and Gupta, S. (2004). Using Control Charts to Monitor Process and Product Quality Profiles. *Journal of Quality Technology*, 36, 309–320.

[13] NBS (2017). National Bureau of Statistics Agricultural Data. Retrieved 22 June 2017.

[14] Arowolo, O. T., Aribike, E. and Ekum, M. I. (2017). Panel Predictive Modeling of Agricultural Production Among States in Nigeria. *IOSR Journal of Mathematics (IOSR-JM)*, 13(5), 76-89.

[15] Bayer, F. M., Tondolo, C. M. and Muller, F. M. (2018). Beta regression control chart for monitoring fractions and proportions. Preprint submitted to Computers & Industrial Engineering.

[16] Canterle, D. R. and Bayer, F. M. (2017). Variable dispersion beta regressions with parametric link functions. Statistical Papers (Forthcoming), DOI: 10.1007/s00362-017-0885-9.