Normality, Grubb’s and Runs Tests for the von Bertalanffy Model Used in the Fitting of the Growth of Bacillus cereus strain wwcp1 on Malachite Green Dye

Salihu Yahuza1* and Ibrahim Alhaji Sabo2

1Department of Microbiology and Biotechnology, Faculty of Science, Federal University Dutse, PMB 7156, Dutse, Jigawa State, Nigeria.
2Department of Microbiology, Faculty of Pure and Applied Sciences, Federal University Wukari, PMB 1020, Wukari, Taraba State, Nigeria.

*Corresponding author:
Salihu Yahuza,
Department of Microbiology and Biotechnology,
Faculty of Science,
Federal University Dutse,
PMB 7156, Dutse,
Jigawa State,
Nigeria.
Email: salihuyahuza66@gmail.com

INTRODUCTION

A biological system can be viewed as a collection of different cellular compartments (such as cell types), each of which is specialized for a certain biological purpose (e.g. white and red blood cells have very different commitments). An object is a type of elemental unit that can be investigated but whose internal structure is unknown or non-existent. The scale used to depict the system will be decided by the elemental unit chosen. The availability of data covering a wide range of biological states and processes, as well as the temporal interdependence of those activities, allows for biological system study at a variety of organizational levels. These stages start at the molecular level and work their way up to the population level. In high-throughput genomics and proteomics research, the use of mathematical and computer models to aid in the interpretation of biological data is becoming more common.

The later stages of the research process, hypothesis creation and experiment prescription, are both facilitated by the use of complicated computer models that allow for the depiction of intricate biological processes. Computational models are now using knowledge discovery approaches to make advantage of the huge amounts of data collected in biomedical databases. This is done to make the best use of the given information. The
availability of data covering a wide range of biological states and processes, as well as the temporal interdependence of those activities, allows for biological system study at a variety of organizational levels. These stages start at the molecular level and work their way up to the population level. In many circumstances, the observed phenomenon’s relationship to time or concentration can be statistically represented using least square techniques, which are often employed in nonlinear regression [1–4].

In any case, the residuals of a nonlinear regression must have natural dispersion. This contrasts from the usual least squares approach, which requires linear regression residuals to have a normal distribution. This is because the concept of least squares is the foundation of the traditional least square’s technique. The far more important residuals must be random and have the same variance (homoscedastic distribution). The Wald-Wolfowitz runs test can be used to determine the effectiveness of the randomization approach (Motulsky and Ransnas 1987). In order for the usual least squares technique to work effectively, the residuals of the curve in a nonlinear regression must have a natural dispersion, whereas in a linear regression, the residues must have a normal distribution. The residuals must be randomly distributed and have the same variance (homoscedastic distribution) (homoscedastic distribution). The Wald-Wolfowitz runs test is used to determine if the trial was randomized or not. In this study the von Bertalanffy model used in the fitting of the growth of Bacillus cereus strain wwcp1 on malachite green dye was assessed for its non-randomness of residuals and the potential presence of an outlier.

METHODOLOGY

Residual data
One of the benefits of residual information is that it may be used to assess the correctness of any model when fitting a curve in nonlinear regression. (D’Agostino, 1986). In statistics, residual data is estimated as the difference between observed and anticipated data, the latter generated using an appropriate model and typically performed using nonlinear regression. (Eqn. 1);

\[ e_i = y_i - f(x_i; \hat{\beta}) \]  

(1)

where \( y_i \) is the \( i \)th response from a particular data and \( x_i \) is the vector of descriptive variables to each set at the \( i \)th observation which corresponds to values from a particular data. Residual data from in this study, the von Bertalanffy model used in the fitting of the growth of Bacillus cereus strain wwcp1 on malachite green dye was assessed from a previous work [5].

Grubbs’ Statistic
The test is a statistical technique for identifying outliers in a univariate data set with a Gaussian or normal distribution. Grubbs’ test assumes that the data is spread uniformly. In a univariate context, the test is used to find outliers [6]. The test can be utilized to the maximal or minimal examined data from a Student’s t distribution (Eq. 2) and to test for both data instantaneously (Eq. 3).

\[ G_{\min} = \frac{\bar{x} - \min(X)}{s} \]  

\[ G_{\max} = \frac{\max(X) - \bar{x}}{s} \]  

\[ p_0 = 2n_p \left( G_\text{min} \frac{n(n-2)}{n-1} n-2.1 \right) \]  

\[ G_{\text{out}} = \frac{\max(X - \min(X), \max(X) - \bar{x})}{s} \]  

\[ p_0 = n_p \left( G_\text{out} \frac{n(n-2)}{n-1} n-2.2 \right) \]  

(2)  

(3)

The ROUT method can be employed in the event that there is more than one outliers [7]. The False Discovery Rate is the foundation of the approach (FDR). Q, a probability of (incorrectly) recognizing one or more outliers must be explicitly specified. It is the highest desired FDR. Q is fairly comparable to alpha in the absence of outliers. Assumption that all data has a Gaussian distribution is mandatory.

Runs test
When doing a nonlinear regression, the residuals of the curve must have a natural distribution. This is in contrast to the least squares approach, which requires that the residues have a normal distribution. These specs are available here. Furthermore, residuals must be random and have the same variance (homoscedastic distribution). The Wald-Wolfowitz test is used to assess whether or not randomization has been completed. The concept is statistically valid because biological systems are inherently unpredictable [8–10]. This test was run on the regression residuals to detect unpredictability.

The number of sign runs is sometimes expressed as a percentage of the maximum number possible. The runs test looks at a sequence of residuals that include both positive and negative values. Following the test, a good run is frequently described by the existence of an alternating or adequately balanced number of positive and negative residual data. The runs test determines if the residuals data contains an excessive or inadequate amount of runs of sign (Eq. 4). Too few runs may indicate systematic bias or a clustering of residuals with the same sign, but too many with the same run sign may indicate negative serial correlation [11,12]. The test statistic is

\[ H_0= \text{ the sequence was produced randomly} \]  

\[ H_a= \text{ the sequence was not produced randomly} \]  

\[ Z = \frac{R - \bar{R}}{sR} \]  

(4)
Where $Z$ is the test statistic, $R$ indicates the anticipated number of runs, $sR$ is the standard deviation of the runs and $R$ is the observed number of runs and (Eqs. 5 and 6). The calculation of the respective values of $R$ and $sR$ ($n_1$ is positive while $n_2$ is negative signs) is as follows.

$$R = \frac{2n_1n_2}{n_1+n_2} + 1$$ (5)

$$s^2R = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$ (6)

As an example
Test statistic: $Z = 3.0$
Significance level: $\alpha = 0.05$

Critical value (upper tail): $|Z| > 1.96$
Critical region: Reject H0 if $|Z| > 1.96$

If the test statistical value ($Z$) is greater than the critical value, then the rejection of the null hypothesis at the 0.05 significance level shows that the sequence was not generated randomly.

RESULTS AND DISCUSSION

Finding a good model for biological and even chemical processes may be difficult. The modelling approach is complex on its own, and mistakes are common. The modelling technique is a strategy, which includes outlining a set of actions that must be done in order to locate an accurate model of the system under examination. In this step, you will collect knowledge and data from experts in the field as well as published works, model structure, model hypothesis, conceptual model, appropriate mathematical formalism selection, solving the formal model, obtaining results, determining whether the model's results match the data available, and other similar tasks. The third part is to put the plan into action, which involves repeating the methods from the previous two stages, determining whether or not the solution is right, and finally enhancing the model. This final phase is a critical test for validating the hypothesis developed prior to the model's setup. Finally, all models must be mathematically fit, which is where nonlinear regression comes in.

The residuals data are crucial in the statistical analysis of nonlinear regression, which is performed using nonlinear regression. The residuals show the discrepancy between the data collected and the data expected to be collected. The gap between the values predicted by a mathematical model and the values observed in the data is referred to as "residuals." These discrepancies may be seen when comparing the model's predictions to the data. The residuals must be statistically analyzed to ensure that they are sufficiently random, lack outliers, have a normal distribution, and do not exhibit autocorrelation. This may be done by determining whether or not they have any outliers. It is common practice to show residue data autocorrelation. This may be done by determining whether or not outliers have a normal distribution, and do not exhibit

The Chauvenet's criteria, the 3-sigma criterion, and the Z-score are all approaches for determining if a measurement contains likely outliers. The 3-sigma criteria and the Z-score are two popular statistical methods in the field of chemometrics. A boxplot may be used to rapidly and readily identify outliers in measurement data. Despite the fact that the methods in question are simple, quick, and pass visual inspections, it is recommended that a statistical test be performed to establish whether or not a data set contains an outlier. Dixon's Q-test and Grubbs' ESD-test are two specific tests that may be used to assess whether or not a

| Time | Residuals |
|------|-----------|
| 0    | -0.3      |
| 10   | 0.4       |
| 18   | 0.17      |
| 30   | -0.19     |
| 39   | -0.23     |
| 49   | -0.25     |
| 59   | -0.26     |
| 69   | -0.2      |
| 79   | -0.12     |
| 90   | -0.09     |
| 100  | -0.08     |
| 110  | -0.03     |
| 120  | 0.02      |
| 129  | 0.05      |
| 140  | 0.05      |
| 150  | 0.06      |
| 159  | 0.07      |
| 169  | 0.09      |
| 180  | 0.1       |
| 189  | 0.1       |
| 199  | 0.11      |
| 209  | 0.12      |
| 219  | 0.13      |
| 229  | 0.13      |
| 239  | 0.13      |

Grubbs' test indicated no indication of an outlier when applied to previously available data. This demonstrates that the model correctly described the data. When attempting to fit a nonlinear curve, it is easy to introduce large errors if either the mean or a single data point from a triple is twisted. Both of these circumstances have the same outcome. The Grubbs test can identify a single abnormality throughout any given time period. It is critical to seek for and delete any outlying data points when fitting curves. The Grubbs test was repeated until no further outliers were found. Sample sizes of six or below are not recommended since the test consistently flags the vast majority of points as outliers. Furthermore, repeating the test may change the likelihood of it discovering something.

The Grubbs' test statistic zeroes in on the sample value with the largest absolute deviation from the sample mean, as indicated by the sample's standard deviation, to select the winner. If the result of the test statistic $g$ is larger than the critical value, the result is termed an outlier. This is due to the fact that the critical value is the lowest value that can be tolerated. Grubbs' test results revealed the absence of outlier (Table 2). An extreme data point that the investigator thinks implausible because it lacks a number of certain qualities is an example of a possible outlier. An outlier is a figure that stands out as being notably different from the bulk of the other data in a sample. An outlier is, for example, a maximum that is statistically significantly bigger than the maximum distribution predicted by the population model employed in engineering. Using these criteria, one may assess if the highest value is an anomaly.

The Chauvenet's criteria, the 3-sigma criterion, and the Z-score are all approaches for determining if a measurement contains likely outliers. The 3-sigma criteria and the Z-score are two popular statistical methods in the field of chemometrics. A boxplot may be used to rapidly and readily identify outliers in measurement data. Despite the fact that the methods in question are simple, quick, and pass visual inspections, it is recommended that a statistical test be performed to establish whether or not a data set contains an outlier. Dixon's Q-test and Grubbs' ESD-test are two specific tests that may be used to assess whether or not a
person is an outlier. Grubbs created both of these examinations [20].

**Table 2.** Outlier detection from the residuals of the von Bertalanffy model used in the fitting of the growth of *Bacillus cereus* strain wwc1 on malachite green dye.

| row | Value | Z     | Significant Outlier? |
|-----|-------|-------|----------------------|
| 1   | -0.3  | 1.7685| Furthest from the rest, but not a significant outlier (P > 0.05). |
| 2   | 0.4   | 2.369 |                      |
| 3   | 0.17  | 1.0965|                      |
| 4   | -0.19 | 1.1183|                      |
| 5   | -0.23 | 1.3547|                      |
| 6   | -0.25 | 1.4792|                      |
| 7   | -0.26 | 1.532 |                      |
| 8   | -0.2  | 1.1774|                      |
| 9   | -0.12 | 0.7045|                      |
| 10  | -0.09 | 0.5272|                      |
| 11  | -0.08 | 0.4681|                      |
| 12  | -0.03 | 0.1726|                      |
| 13  | 0.02  | 0.1229|                      |
| 14  | 0.05  | 0.3003|                      |
| 15  | 0.05  | 0.3003|                      |
| 16  | 0.06  | 0.3964|                      |
| 17  | 0.07  | 0.4185|                      |
| 18  | 0.09  | 0.5367|                      |
| 19  | 0.1   | 0.5958|                      |
| 20  | 0.1   | 0.5958|                      |
| 21  | 0.11  | 0.6549|                      |
| 22  | 0.12  | 0.714 |                      |
| 23  | 0.13  | 0.7731|                      |
| 24  | 0.13  | 0.7731|                      |
| 25  | 0.13  | 0.7731|                      |

Before the Grubbs test findings to be deemed valid, the specific value of the anticipated number of outliers, denoted by the letter k, must be specified. This is by far the most significant constraint imposed by the exam. It is very possible that the exam results will be tainted if k is not appropriately reflected in the test. This is not an unthinkable prospect. In circumstances when there are several outliers or the exact number of outliers cannot be identified, the Rosner's generalised Extreme Studentized Deviate test, often known as the ESD-test or the ROUT approach, might be used. The Extreme Studentized Deviation exam refers to both of these assessments. In common terminology, both of these procedures are known as the ESD-test [7] are recommended [21]. The ROUT approach, which combines robust regression and outlier removal, is becoming more popular in the elimination of numerous outliers [14,22–25]

The runs test revealed that there were 4 total runs, whereas the randomization assumption predicted 13 runs (Table 3). This implies that the runs included in the residual collection are just those that are marginally important. The Z-value is the number of standard errors by which the actual number of runs differs from the predicted number of runs, and the p-value indicates how important this Z-value is in terms of the difference between the two values. It should be evaluated similarly to any other statistic that employs p-values. If the p-value is less than 0.05, which is the cutoff point at which the null hypothesis may be rejected, it is possible to conclude that the residuals are not fully random. When the threshold is reached, this is the situation.

The fact that the p-value is smaller than 0.05 implies that the null hypothesis is rejected; this demonstrates that there is convincing evidence of the residuals’ non-randomness. When there are a large number of occurrences of a particular run sign, this may indicate a negative serial correlation; when there are a small number of runs, this may indicate a clustering of residuals with the same sign or the presence of a systematic bias; and when there are a large number of occurrences of a particular run sign, this may indicate a positive serial correlation; and when there are a large number of occurrences of a particular run sign, this may indicate [12]. The runs test can discover a systematic deviation from the curve, such as an overestimation or underestimating of the sections, when a specific model is applied. This is accomplished by contrasting the actual and anticipated values.

**Table 3.** Runs test for randomness.

| Runs test | Residual data set |
|-----------|-------------------|
| R=        | 4                 |
| n0=       | 10                |
| n1=       | 15                |
| n=        | 25                |
| E(R)=     | 13.00             |
| Var(R)=   | 5.50              |
| SDx(R)=   | 2.35              |
| Z=        | 3.34              |
| p-value=  | 0.0001            |

Note: E(R) Expected run

This may be performed by comparing the model's predictions to the actual outcomes. The runs test is used to detect if there are an excessive number of signs runs and whether there are insufficient runs in general. A runs test was performed on the regression residuals to assess whether or not there was evidence of nonrandomness. It is possible to build a model with an ordered variance of the curve that is more or less than the estimate. This is only one of the ways that something like this is possible. As part of the run test, which is used to establish whether or not a chemical poses a risk to human health, a comparison is done between a drug's normally negative sequence of residues and its typically positive sequence.

This test is used to determine whether a substance is hazardous to human health. A notable occurrence is usually differentiated from other conceivable outcomes by a change or combination of shifts between negative and positive residual values. The transition from one condition to another, or a succession of changes, is frequently what characterizes a remarkable conclusion [11]. The use of the largest number of indicators that may be counted is a frequent practice in this discipline. The run's test is used to determine whether a large number of signs passes, or a small number of sign passes are more likely. Run signs may imply negative serial correlation, but it is also possible that residues are connected with the same sign or that systemic biases are influencing the results [12].

When testing time-series regression models to identify the existence or absence of autocorrelation, it is usual practise to utilise the run technique. According to the conclusions of study that used Monte Carlo simulation, run-time testing is responsible for unequal error rates in both tails of the distribution. This study implies that run-time autocorrelation research may be unstable, and it predicts that in the not-too-distant future, the Durbin-Watson methodology will be the most often used method for analysing autocorrelation [26]. It has been demonstrated that the approach utilised in this analysis, which is based on earlier research on the unpredictability of residuals, is reliable. For example, modelling the growth curve of algae using the Baranyi-Roberts model, which provides statistical sufficiency [27], *Moraxella sp.* B on monobromoacetic acid (MBA) [9] and the Buchanan-three-phase model used in the fitting the growth of *Paracoccus* sp. SKG on acetonitrile [28]. For lead (II) absorption by alginate gel bead, the runs tests on the residuals for the Sips and Freundlich models were found to be sufficient [29].

It has been shown that the strategy used in this analysis, which was derived from previous research on the unpredictability of residuals, is dependable. One can anticipate the growth curve

This work is licensed under the terms of the Creative Commons Attribution (CC BY) (http://creativecommons.org/licenses/by/4.0/).
of algae using the Baranyi-Roberts model, which assesses whether or not statistical sufficiency has been fulfilled [30]. In the body of academic research, different applications of the runs test of residual may be found for the purpose of evaluating the validity of the nonlinear regression [31–35]. However, the test for the lack of an outlier does not demand reanalysis of the data, as indicated by Grubb's test findings. To resolve this discrepancy, either a different model needs to be used or more data needs to be added.

CONCLUSION

The Wald-Wolfowitz runs test was performed as part of the scope of this investigation. The observations revealed that the residual series had an acceptable number of runs; this was the result of the observations. The runs test found a total of four runs, although the randomization assumption expected thirteen runs. This means that the runs included in the residual collection are only slightly significant. The fact that the p-value is less than 0.05 suggests that the null hypothesis is rejected; this demonstrates that there is compelling evidence of non-randomness in the residuals. Furthermore, this illustrates the need of investigating potential outliers. This illustrates, in particular, that extra intervention is required to detect any outliers that may exist in the data. The results of Grubb's test suggest, however, that testing for the absence of an outlier does not necessitate reanalyzing the data. Either a different model or additional data is required to explain the discrepancy.

REFERENCES

1. Lee KB, Lee JM, Park TS, Lee SH. Construction of classical model of this investigation. The observations revealed that the results of Grubb's test suggest, however, that testing for the presence of an outlier does not necessitate reanalysis of the data. Miller JN. Using the Grubbs and Cochran tests to identify outliers. Anal Methods. 2015;7(19):7948–50.
2. Barnett V, Lewis T. Outliers in Statistical Data. 3rd ed. Chichester; New York: Wiley; 1994. 604 p.
3. Jain RB. A recursive version of Grubbs’ test for detecting multiple outliers in environmental and chemical data. Clin Biochem. 2010;43(12):1030–3.
4. Cohn TA, England JF, Berenbrock CE, Mason RR, Stedingr JR, Lamontagne JR. A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resour Res. 2013;49(8):5047–58.

- 25 -