Test of the Randomness of Residuals and Detection of Potential Outliers for the Modified Logistics Used in the Fitting of the Growth Curve of Immobilized *Pseudomonas putida* on Phenol

Garba Uba1* and Murtala Ya’u2

1Department of Science Laboratory Technology, College of Science and Technology, Jigawa State Polytechnic, Dutse, PMB 7040, Nigeria.
2Department of Biochemistry, Faculty of Basic Medical Sciences, College of Health Sciences, Bayero University Kano, Kano, PMB 3001- Nigeria.

*Corresponding author:
Garba Uba,
Department of Science Laboratory Technology,
College of Science and Technology,
Jigawa State Polytechnic,
Dutse,
PMB 7040,
Nigeria.
Email: garbauba@jigpoly.edu.ng

INTRODUCTION

The process of identifying the knowledge contained inside biological systems has been sped up as a result of the revolutions that have taken place in both the field of biotechnology and the field of information technology. These advancements are changing the methods used for doing research, development, and applications in the field of biomedicine. The addition of clinical data to biological data makes it possible to provide comprehensive descriptions of both healthy and sick states, as well as the progression of illness and the body's reaction to therapies. The study of biological systems at many different levels of organization is possible thanks to the availability of data reflecting diverse biological states, processes, and the time dependencies of those activities. These levels range from molecules to organisms and even populations. In high-throughput genomics and proteomics research, mathematical and computational models are being used more frequently to assist in the understanding of biological data. The use of complex computer models that enable the modelling of intricate biological processes leads to the creation of hypotheses and the suggestion of experiments as next steps in the research process. Text mining and knowledge discovery approaches are currently being utilized by computational models in order to make use of the large amount of data that is stored in biomedical databases. The study of biological systems at many different levels of organization is possible thanks to the availability of data reflecting diverse biological states, processes, and the time dependencies of those activities. These levels range from molecules to organisms and even populations [1–7].
A collection of things that are linked together is called a system. For instance, a biological system could be thought of as a collection of several cellular compartments (such as cell types), each of which is specialized for a certain biological function (e.g. white and red blood cells have very different commitments). An elemental unit of some sort that can be observed but whose interior structure is either unknown or does not exist is referred to as an object. The chosen elemental unit will determine the scale that will be used to display the system. The study of biological systems at many different levels of organization is possible thanks to the availability of data reflecting diverse biological states, processes, and the time dependencies of those activities. These levels range from molecules to organisms and even populations. A representation of a system that can be deciphered or understood by researchers in general is called a model. A model is a description of a system in terms of the constituent components and the interactions between those components [8–15].

Nevertheless, in a nonlinear regression, the curve’s residuals need to have a natural dispersion, in contrast to the standard least squares approach, which requires the residuals to have a normal distribution in a linear regression. This is because the standard least squares approach is based on the principle of least squares. More importantly, the residuals must be random and have the same variance (homoscedastic distribution). The Wald-Wolfowitz runs test is used to establish whether or not the randomization process was successful [16]. On the other hand, the residuals of the curve in a nonlinear regression need to have a natural dispersion, whereas in a linear regression, the residuals need to have a normal distribution in order for the typical least squares approach to work well. More crucially, the residuals must be random and have the same variance (homoscedastic distribution). The residuals must also be random and outliers’ absence. The Wald-Wolfowitz runs test is used to establish whether or not the residuals for the modified logistics used in the Fitting of the growth curve of immobilized Pseudomonas putida on phenol is random whilst the Grubb’s test is applied to detect the presence of outliers.

### METHODOLOGY

One of the utilities of residual information is that it can be utilized to measure the accuracy of any model fitting a curve in nonlinear regression can be achieved by evaluating (D’Agostino, 1986). In the statistical meaning, residual data is calculated by the difference between observed and predicted data, the latter obtained using suitable model and usually carried out using nonlinear regression (Eqn. 1);

$$e_i = y_i - f(x_i; \beta)$$  \hspace{1cm} (Eqn. 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 the modified logistics used in the fitting of the growth curve of immobilized Pseudomonas putida on phenol.

**Grubb’s Statistic**

The test is a statistical test used to discover outliers in a univariate data set that is believed to have a Gaussian or normal distribution. Grubb’s test assumes that the data is regularly distributed. The test is used to discover outliers in a univariate context [17]. 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).

$$\begin{align*}
G_{\min} & = \frac{\bar{X} - \min(X)}{s} \\
G_{\max} & = \frac{\max(X) - \bar{X}}{s} \\
p_{ci} & = 2n_p \left( G_{\max} \right) \\
g_{ai} & = \frac{\max(X) - \bar{X}}{s} \\
p_{ci} & = n_p \left( G_{\max} \right)
\end{align*}$$  \hspace{1cm} (Eqn. 2 and 3)

The ROUT method can be employed in the event that there is more than one outliers [18]. 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**

The residuals of the curve in a nonlinear regression need to have a natural distribution. This differs from the requirements of the least squares method, which calls for the residues to have a regular distribution. In addition to this, residuals are required to be random and have the same variance (homoscedastic distribution). For the purpose of determining whether or not randomization has been achieved, the Wald–Wolfowitz test is utilized. Biological systems are inherently unpredictable, and as a result, the model may be relied upon to be statistically accurate. [19–21]. This test was applied to the regression residuals in order to find unpredictability in the residuals. The number of sign runs is often stated as a percentage of the greatest number possible.

The runs test examines the sequence of residuals, of which they are composed of positive and negative values. A successful run, after running the test, is often represented by the presence of an alternating or adequately balanced number of positive and negative residual values. The runs test computes the likelihood of the residuals data having too many or too few runs of sign (Eq. 4). Too few runs may suggest a clustering of residuals with the same sign or the existence of systematic bias, whereas too many of a run sign may identify the presence of negative serial correlation [16,22].

The test statistic is

$$Z = \frac{\bar{R} - R}{sR}$$  \hspace{1cm} (Eqn. 4)

Where $Z$ is the test statistic, $\bar{R}$ 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 $\bar{R}$ and $sR$ ($n_1$ is positive while $n_2$ is negative signs) is as follows.


\[
\begin{align*}
\bar{R} &= \frac{2n_1 n_2 + 1}{n_1 + n_2} \\
S^2 R &= \frac{2n_1 n_2 (2n_1 n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}
\end{align*}
\]

(Eqn. 5)  
( Eqn. 6)  

As an example

Test statistic: \( Z = 3.0 \)
Significance level: \( \alpha = 0.05 \)
Critical value (upper tail): \( Z_{\alpha/2} = 1.96 \)
Critical region: Reject \( H_0 \) 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**

It might be difficult to locate an appropriate model for biological and even chemical processes. The process of modelling is challenging in and of itself, and mistakes are not an extremely uncommon occurrence. The modelling technique is in and of itself a process that adheres to a loosely formalized set of guidelines. The process is based on the completion of four large phases. The first step is to get a solid grasp of the issue at hand, which involves precisely defining the queries that are posed to the model.

The second phase is to develop a strategy for addressing the problem, which entails outlining a sequence of activities that need to be carried out in order to locate an accurate model of the system that is the subject of the investigations. In this step, you will acquire knowledge and data from specialists in the field as well as from published works, model structure, model hypothesis, conceptual model, appropriate mathematical formalism selection, solving the formal model, obtaining the results, checking to see if the results of the model match the data that is available, and other similar tasks. The third phase is to put the plan into action, which involves doing the processes from the previous two steps, determining whether or not the solution is accurate, and finally refining the model. This last step is a significant test to examine the hypothesis that was developed prior to the setting of the model. Ultimately all models will need to be subjected to mathematical curve fitting and this is where nonlinear regression comes into place [1–7].

In the statistical analysis of nonlinear regression, the data known as residuals play an important role. Residuals indicate the difference between the data that was observed and the data that was anticipated. The differences that exist between the values that are predicted by a mathematical model and the values that are actually observed in the data are referred to as residuals. The residuals must be subjected to statistical analysis in order to assess whether or not they are sufficiently random, do not include any outliers, adhere to the normal distribution, and do not display autocorrelation.

The data on residues are frequently presented in the form of positive and negative values, which is vital for demonstrating that the data are balanced; this may be seen visually before any tests are carried out. When performing nonlinear regression, it is common practice to ignore the results of residual tests. When there is a larger gap between the values that were anticipated and those that were actually observed, a model's quality is considered to be lower, as a general rule. This is due to the lower degree of correlation that exists between the two sets of data [23]. The residuals for the modified logistics model are shown in Table 1.

**Table 1.** Residuals for the modified logistics used in the fitting of the growth curve of immobilized *Pseudomonas putida* on phenol.

<table>
<thead>
<tr>
<th>Time</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>-0.04703</td>
</tr>
<tr>
<td>8</td>
<td>0.063298</td>
</tr>
<tr>
<td>12</td>
<td>-0.05265</td>
</tr>
<tr>
<td>16</td>
<td>0.043047</td>
</tr>
<tr>
<td>20</td>
<td>-0.03325</td>
</tr>
<tr>
<td>24</td>
<td>0.012304</td>
</tr>
<tr>
<td>28</td>
<td>-0.00158</td>
</tr>
<tr>
<td>32</td>
<td>-0.00951</td>
</tr>
<tr>
<td>36</td>
<td>-0.00648</td>
</tr>
<tr>
<td>40</td>
<td>-0.00998</td>
</tr>
<tr>
<td>44</td>
<td>0.00701</td>
</tr>
<tr>
<td>48</td>
<td>0.011794</td>
</tr>
</tbody>
</table>

When applied to the data that had been previously published, Grubbs' test demonstrated that there was no indication of an outlier. This suggests that the model was successful in accurately representing the data. When trying to fit a nonlinear curve, it is possible to add a significant amount of inaccuracy if either the mean is changed by a single data point or a single data point from a triple is distorted. Both of these scenarios have the same effect. The Grubbs test has the ability to identify a single anomaly throughout any given time period. When fitting curves, it is essential to look for and eliminate any outlying data points.

Because it was determined that this particular data point was an outlier, it was removed from the collection, and the analysis was carried out again until there were no more outliers. Because the test reliably identifies the vast majority of points as outliers, it is not a good idea to utilise sample sizes of six or less. Additionally, doing several repeats of the test can change the chance that it will find something. The sample value that has the greatest absolute departure from the sample mean, as assessed by the sample's standard deviation, is the one that the Grubbs' test statistic zeroes in on to determine the winner. In the event that the test statistic g produces a number that is greater than the critical value, the result in question is referred to as an outlier. This is as a result of the fact that the critical value is the minimum value that may be tolerated (Grubbs 1969). The results of Grubbs's test suggested the absence of an outlier (Table 2).

**Table 2.** Outlier detection for the residuals for the modified logistics used in the fitting of the growth curve of immobilized *Pseudomonas putida* on phenol.

<table>
<thead>
<tr>
<th>Row</th>
<th>Value</th>
<th>Z</th>
<th>Significant Outlier?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.0470</td>
<td>1.3699</td>
<td>Furthest from the rest, but not a significant outlier (P &gt; 0.05).</td>
</tr>
<tr>
<td>2</td>
<td>0.0633</td>
<td>1.9213</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.0526</td>
<td>1.5374</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.0430</td>
<td>1.3172</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.0325</td>
<td>0.9371</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.0123</td>
<td>0.4001</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.0016</td>
<td>0.0141</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.0095</td>
<td>0.2565</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.0065</td>
<td>0.1603</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0010</td>
<td>0.0388</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.0070</td>
<td>0.2422</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.0118</td>
<td>0.3849</td>
<td></td>
</tr>
</tbody>
</table>

An extreme data point that the investigator deems to be improbable due to the fact that it does not fulfil a number of certain requirements is an example of a potential outlier. A figure that stands out as being significantly different from the rest of the data in a sample is known as an outlier [24–30]. For example, the maximum is regarded as an outlier when it is statistically significantly bigger than the distribution predicted for the maximum based on the population model used in engineering. This criterion is applied to determine whether or not the maximum is an outlier. Identifying possible measurement
outliers may be accomplished with the use of Chauvenet's criterion, the 3-sigma criterion, and the Z-score. In the field of chemometrics, the Z-score is typically applied in combination with the 3-sigma criteria. A boxplot is a straightforward method for pinpointing probable outliers in measurement data.

For the purpose of determining whether or not a data set contains an outlier, a statistical test is preferred, despite the fact that the procedures in question are uncomplicated, rapid, and can pass visual inspections. Two particular tests that may be utilised to determine whether or not an individual is an outlier are the Dixon's Q-test and the Grubbs' ESD-test. The specific value of the predicted number of outliers, denoted by k, needs to be provided before the Grubbs test may be accepted as valid. This is the most significant restriction of the exam. If k is not accurately reflected in the test, it is quite possible that the findings of the test will be changed. Rosner's generalised Extreme Studentized Deviate, also known as the ESD-test, or the ROUT methodology may be utilised in circumstances in which there are several outliers, or the precise number of outliers cannot be determined. Both of these methodologies are usually referred to as the ESD-test [18] are recommended [31]. Of the two, the ROUT method, which combines robust regression and outlier removal is increasingly being employed in removal of multiple outliers [24,32–35].

The runs discovered that there were 8 total runs, but the assumption of randomness led to the prediction of 7 runs (Table 3). This suggests that the collection of residuals is appropriate. The Z-value represents the number of standard errors by which the actual number of runs deviates from the predicted number of runs, and the p-value that accompanies it represents the degree to which this Z-value is significant. The null hypothesis is not rejected since the p-value was more than 0.05; this means that there is no persuasive evidence of the non-randomness of the residuals; rather, the residuals represent noise. When there are an excessive number of occurrences of a certain run sign, it may be an indication of a negative serial correlation; when there are an inadequate number of runs, it may be an indication of a clustering of residuals that have the same sign or the presence of a systematic bias [22].

### Table 3. Runs test for randomness.

<table>
<thead>
<tr>
<th>Runs test</th>
<th>Residual data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>R±</td>
<td>8</td>
</tr>
<tr>
<td>n±</td>
<td>6</td>
</tr>
<tr>
<td>n</td>
<td>12</td>
</tr>
<tr>
<td>E(R)</td>
<td>7.00</td>
</tr>
<tr>
<td>Var(R)</td>
<td>2.73</td>
</tr>
<tr>
<td>StdDev(R)</td>
<td>1.65</td>
</tr>
<tr>
<td>Z</td>
<td>0.61</td>
</tr>
<tr>
<td>p-value</td>
<td>0.7276</td>
</tr>
</tbody>
</table>

When using a particular model, the runs test can discover a systematic divergence from the curve, such as an overestimation or underestimating of the sections. This can be accomplished by comparing the actual values to the predicted values. This may be accomplished by contrasting the model's predictions with the values that actually occurred. The runs test is utilized to determine if an excessive amount of sign runs is present or whether there are inadequate runs overall. In order to assess whether or not there was evidence of nonrandomness, the runs test was applied to the regression residuals. It is possible to create a model with an ordered variance of the curve that is either larger or lower than the estimate. This is one of the many ways in which this is feasible.

A comparison is made between a drug's typically negative sequence of residues and its generally positive sequence as part of the run test, which is used to determine whether or not a chemical poses a risk to human health. A movement or combination of shifts between the negative and positive residual values is frequently what distinguishes a remarkable event from other possible outcomes. A transformation or series of shifts is frequently the defining characteristic of a noteworthy conclusion [16]. A popular practice in this area is to use the largest proportion of indicators that can be counted. To determine whether a large number of sign passes are likely or a low number of sign passes are more likely, the run's test is used. Run signs may imply negative serial correlation, but it is also possible that residues are related with the same sign or that there are systemic biases that are influencing the results [22]. It is usual practise to apply the run method when testing time-series regression models to determine whether or not autocorrelation is present. According to the results of Monte Carlo simulation experiments, run-time testing produces uneven error rates in both tails of the distribution. This conclusion suggests that run-time autocorrelation research might not be stable, and it predicts that the Durbin-Watson methodology will become the most used approach to evaluating autocorrelation in the foreseeable future [36]. It has been demonstrated that the methodology utilised in this investigation, which was derived from earlier research that investigated the unpredictability of the residuals, is reliable.

For example, modelling the growth curve of algae using the Baranyi-Roberts model, which demonstrates statistical sufficiency [37], Moraxella sp. B on monobromoaetic acid (MBA) [20] and the Buchanan-three-phase model used in the fitting the growth of Paracoccus sp. SKG on acetonitrile [38]. 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 [39]. It was found in a previous study that a runs test on the residual series of data from the pseudo-1st order kinetic modelling of the adsorption of the brominated flame retardant 4-bromophenyl ether onto biochar-immobilized Sphingomonas sp. showed that the residual series had sufficient runs after the test was carried out on the runs [40]. 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 [41–45].

### CONCLUSION

The Wald–Wolfowitz runs test was performed on the changed logistics that were utilized in the fitting of the growth curve of immobilized Pseudomonas putida on phenol. This test was carried out in order to determine whether or not the logistical changes were effective. The runs test revealed that there was a total of eight runs, despite the assumption that there would only be seven runs based on the unpredictability of the situation. This would seem to indicate that the collecting of residuals is something that should be done. The Z-value is a measure of the amount by which the actual number of runs deviates from the number of runs that was predicted, and the p-value that is associated with it is a measure of the extent to which this Z-value is statistically significant. The null hypothesis is not rejected since the p-value was greater than 0.05; this indicates that there is no convincing evidence of the non-randomness of the residuals; rather, the residuals represent noise in the data. Since the p-value was greater than 0.05, the null hypothesis is not rejected. As a result of the results of Grubb's test, which reveal that there is no outlier, the data do not need to be reanalyzed because the model that was utilized was sufficient.
REFERENCES


