Test of the Randomness of Residuals and Detection of Potential Outliers for the Buchanan-3-phase Model Used in the Fitting of the Effect of nZVI/Pd and SiO₂-nZVI/Pd Nanoparticles on the Growth of *P. putida*

Garba Uba¹, Murtala Ya'u² and Ibrahim A. Allamin³*

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.
³Department of Microbiology, Faculty of Science, University of Maiduguri P.M.B 1069 Maiduguri, Nigeria.

*Corresponding author:
Ibrahim A. Allamin,
Department of Microbiology,
Faculty of Science,
University of Maiduguri
P.M.B 1069 Maiduguri,
Nigeria.
Email: ibnallamimmcb@unimaid.edu.ng

INTRODUCTION

Revolutions in biotechnology and information technology have generated massive volumes of data, hastening the process of discovering biological systems' knowledge. These developments are altering the way biomedical research, development, and applications are carried out. Clinical data supplements biological data, allowing for full descriptions of both healthy and sick states, as well as illness development and response to treatments. The availability of data reflecting multiple biological states, processes, and their time dependencies allows for the study of biological systems at many levels of the organization, from molecules to organisms and even populations. High-throughput genomics and proteomics research are increasingly using mathematical and computational models to assist understand biological data. The use of sophisticated computer models that allow for the modelling of complicated biological processes creates hypotheses and recommends experiments. Text mining and knowledge discovery methodologies are being used to utilize the quantity of data held in biomedical databases by computational models.

The availability of data reflecting multiple biological states, processes, and their time dependencies allows for the study of
biological systems at many levels of an organization, from molecules to organisms and even populations [1–6]. A biological system, for example, might be a collection of diverse cellular compartments (e.g., cell types) that are specialized for a certain biological purpose (e.g., white and red blood cells have very different commitments). An object is some elemental unit that may be observed but whose interior structure is unknown or does not exist. The system's depiction scale is defined by the elemental unit chosen. The availability of data reflecting multiple biological states, processes, and their time dependencies allows for the study of biological systems at many levels of an organization, from molecules to organisms and even populations. A model is a description of a system in terms of constituent components and their interactions, where the description is decodable or easily decipherable by researchers in general.

However, in nonlinear regression, the residuals of the curve must be naturally dispersed, as opposed to the standard least squares approach, which needs the residues to be normally distributed in linear regression. More importantly, the residuals must be random and have the same variance (homoscedastic distribution). The Wald-Wolfowitz runs test determines whether or not randomization has occurred [7]. More often than not, the residuals must be tested for the presence of outliers (at 95 or 99% confidence). This is normally done using the Grubbs’ test. The subject of this study is to test for the randomness of the residuals for the Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO₂-nZVI/Pd nanoparticles on the growth of P. putida and whether outliers are present.

**METHODOLOGY**

**Residual data**

The residual information can be utilized to measure the accuracy of any model fitting a curve in nonlinear regression can be achieved by evaluating the residuals [8]. In the statistical meaning, residual data is calculated by the difference between observed and predicted data, the latter obtained using a suitable model and usually carried out using nonlinear regression (Eqn. 1);

\[
e_i = y_i - f(x_i, \beta)
\]

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 Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO₂-nZVI/Pd nanoparticles on the growth of P. putida was obtained from previous work.

**Grubbs’ 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. Grubbs’s test assumes that the data is regularly distributed. The test is used to discover outliers in a univariate context [9]. 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).

The ROUT method can be employed if there are more than one outliers [10]. 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. The assumption that all data has a Gaussian distribution is mandatory.

**Runs test**

In nonlinear regression, the residuals of the curve must have a natural distribution, as opposed to the least squares technique, which requires the residues to be regularly distributed. Furthermore, residuals must be random and have identical variance (homoscedastic distribution). The Wald-Wolfowitz test is used to detect whether or not randomization has been attained. Biological systems are genuinely random and thus the model is statistically accurate to be utilized [11–13]. This test was applied to the regression residuals 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, which 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 runs sign may identify the presence of negative serial correlation [7,14].

The test statistic is

\[
H_0 = \text{the sequence was produced randomly} \\
H_a = \text{the sequence was not produced randomly}
\]

\[
Z = \frac{R - R_{0.05}}{sR}
\]
Where $Z$ is the test statistic, $R$ indicates the anticipated number of runs, $s$ 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 sign) is as follows.

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

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

As an example

Test statistic: $Z = 3.0$

Significance level: $\alpha = 0.05$

Critical value (upper tail): $Z_{1-\alpha/2} = 1.96$

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**

In the statistical analysis of nonlinear regression, residual data, which represent the difference between observed and anticipated data, play a crucial role. Residuals are the disparities between the values predicted by a mathematical model and the actual values observed in the data. Statistical tests must be conducted to determine whether the residuals are sufficiently random, do not contain any outliers, adhere to the normal distribution, and do not exhibit autocorrelation. Residues data are frequently in the form of positive and negative values, which is essential for indicating a balance of the data; this can be viewed visually before performing any tests. In nonlinear regression, residual tests are frequently disregarded. As a general rule, a model's quality is regarded to be lower when the difference between anticipated and observed values is greater. This is because the correlation between the two data sets is weaker [15]. The residuals for the Buchanan-3-phase model are shown in Table 1.

**Table 1.** Residuals for the Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO$_2$-nZVI/Pd nanoparticles on the growth of *P. putida.*

<table>
<thead>
<tr>
<th>Time (h)</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0040</td>
</tr>
<tr>
<td>4</td>
<td>-0.1360</td>
</tr>
<tr>
<td>8</td>
<td>0.1461</td>
</tr>
<tr>
<td>12</td>
<td>-0.0278</td>
</tr>
<tr>
<td>16</td>
<td>0.0724</td>
</tr>
<tr>
<td>20</td>
<td>0.0254</td>
</tr>
<tr>
<td>24</td>
<td>-0.0845</td>
</tr>
<tr>
<td>28</td>
<td>-0.0594</td>
</tr>
<tr>
<td>32</td>
<td>-0.0012</td>
</tr>
<tr>
<td>36</td>
<td>0.0042</td>
</tr>
<tr>
<td>40</td>
<td>0.0162</td>
</tr>
<tr>
<td>44</td>
<td>0.0201</td>
</tr>
<tr>
<td>48</td>
<td>0.0175</td>
</tr>
</tbody>
</table>

Using Grubbs' test on the previously reported data revealed that there was no evidence of an outlier. This indicates that the model adequately represented the data. A large degree of error can be introduced while fitting a nonlinear curve if either the mean is altered by a single data point or a single data point from a triple is warped. The Grubbs test can detect a single outlier at any given period. Checking for outliers is a necessary component of curve fitting. This data point is deleted from the collection because it was considered to be an outlier, and the analysis is repeated until there are no more outliers. Multiple repetitions can alter the likelihood of detection, and sample sizes of six or less should not be utilised because the test consistently detects the majority of points as outliers. The Grubbs' test statistic identifies the sample value with the largest absolute deviation from the sample mean, as measured by the sample's standard deviation. If the test statistic $g$ exceeds the critical value, the value in question is considered an outlier. This is due to the fact that the critical value is the lowest allowable value [9]. The Grubbs' test indicated that there was an outlier (Table 2). This outlier must be removed and the modelling is redone.

**Table 2.** Outlier for the residuals for the Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO$_2$-nZVI/Pd nanoparticles on the growth of *P. putida.*

<table>
<thead>
<tr>
<th>Row</th>
<th>Value</th>
<th>$Z$</th>
<th>Significant Outlier?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0040</td>
<td>0.0609</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.1360</td>
<td>1.9501</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.1461</td>
<td>2.1016</td>
<td>Furthest from the rest, but not a significant outlier ($P &gt; 0.05$).</td>
</tr>
<tr>
<td>4</td>
<td>-0.0278</td>
<td>0.3960</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.0724</td>
<td>1.0429</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.0254</td>
<td>0.3678</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.0845</td>
<td>1.2107</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.0594</td>
<td>0.8493</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>-0.0012</td>
<td>0.0135</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0042</td>
<td>0.0643</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.0162</td>
<td>0.2357</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.0201</td>
<td>0.2918</td>
<td></td>
</tr>
</tbody>
</table>

An example of a probable outlier is an extreme data point that the investigator classifies as implausible because it does not satisfy several certain parameters. A lot more precisely, an outlier in a sample is a number that is unreasonably high and extreme. For instance, the maximum is considered an outlier when it is statistically substantially greater than the distribution expected for the maximum based on the population model [16]. In engineering, Chauvenet's criterion, the 3-sigma criterion, and the $Z$-score are used to identify probable measurement outliers. The $Z$-score is utilised in conjunction with the 3-sigma criterion in chemometrics. A boxplot is an easy tool for identifying potential measurement outliers. Although these methods are simple, quick, and pass eye inspections, a statistical test is preferable for evaluating whether or not an outlier exists in the data set.

The Dixon's Q-test and the Grubbs' ESD-test are two specific tests that can be used to identify an outlier. The expected number of outliers, $k$, must be given precisely for the Grubbs test to be considered legitimate. This is the test's principal limitation. The test results will likely be altered if $k$ is not precisely represented. In situations where there are multiple outliers or the exact number of outliers cannot be identified, Rosner's generalised Extreme Studentized Deviate, frequently referred to as the ESD-test, or the ROUT approach may be utilised [10] are recommended [17]. Of the two, the ROUT method, which combines robust regression and outlier removal is increasingly being employed in the removal of multiple outliers [18–22].

The runs test revealed that there were 7 total runs, whereas 7.15 runs were expected based on the assumption of randomness (Table 3). This indicates that the series of residuals includes suitable runs. The $Z$-value indicates the number of standard errors by which the actual number of runs deviates from the expected number of runs, while the accompanying p-value indicates the severity of this $z$-value. The interpretation is the same as any other p-value statistic. If the p-value is less than 0.05, it is possible to reject the null hypothesis and infer that the residuals are not truly random. Since the p-value was more than 0.05, the null hypothesis is not rejected; this suggests that there is no convincing evidence of the non-randomness of the

---

This work is licensed under the terms of the Creative Commons Attribution (CC BY) (http://creativecommons.org/licenses/by/4.0/).
residuals; rather, the residuals represent noise. When there are too many instances of a specific run sign, it may indicate a negative serial correlation; when there are too few runs, it may indicate a clustering of residuals with the same sign or the existence of a systematic bias [14]. The runs test can identify a systematic departure of the curve, such as an overestimation or underestimation of the sections, when applying a given model. This is possible by comparing actual values.

Table 3. Runs test for randomness for the Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO₂-nZVI/Pd nanoparticles on the growth of P. putida.

<table>
<thead>
<tr>
<th>Runs test</th>
<th>Residual data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-</td>
<td>7</td>
</tr>
<tr>
<td>n₀</td>
<td>5</td>
</tr>
<tr>
<td>n₁</td>
<td>8</td>
</tr>
<tr>
<td>n₂</td>
<td>13</td>
</tr>
<tr>
<td>E(R)+</td>
<td>7.15</td>
</tr>
<tr>
<td>Var(R)+</td>
<td>2.64</td>
</tr>
<tr>
<td>StDev(R)+</td>
<td>1.63</td>
</tr>
<tr>
<td>Z+</td>
<td>-0.09</td>
</tr>
<tr>
<td>p-value</td>
<td>0.4623</td>
</tr>
</tbody>
</table>

The runs test is used to assess whether there is an excessive number or an insufficient number of signs runs. The runs test was used to the regression residuals to determine whether nonrandomness was evident. It is possible to design a model with an ordered variance of the curve that is either more or smaller than the estimate. This can be accomplished in either direction. To evaluate whether or not a substance is dangerous, the run test compares a drug's generally negative sequence of residues versus its generally positive sequence. A noteworthy outcome is often distinguished by a change or combination of shifts [7]. 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 a negative serial correlation, but it is also possible that residues are related to the same sign or that there are systemic biases that are influencing the results [14].

The run approach is commonly used to test for the presence of autocorrelation in time-series regression models. Run-time test results in unequal error rates in both tails, according to Monte Carlo simulation experiments. Run-time autocorrelation research may not be stable, and the Durbin-Watson methodology will be negative serial correlation, but it is also possible that residues are more likely, the run's test is used. Run signs may imply a negative serial correlation, but it is also possible that residues are related to the same sign or that there are systemic biases that are influencing the results [14].

In the course of this inquiry, the Wald–Wolfowitz runs test for randomness for the Buchanan-3-phase model used in the fitting of the effect of nZVI/Pd and SiO₂-nZVI/Pd nanoparticles on the growth of P. putida was carried out. The runs test revealed that there were 7 total runs, whereas 7.15 runs were expected based on the assumption of randomness. This indicates that the series of residuals includes suitable runs. Since the p-value was more than 0.05, the null hypothesis is not rejected; this suggests that there is no convincing evidence of the non-randomness of the residuals; rather, the residuals represent noise. This suggests that there is no requirement for additional intervention to detect prospective outliers. Due to the absence of an outlier, the data will not need to be reanalyzed as a result of the findings of Grubb's test. In total, the residual analyses suggest that no remodelling of the data is needed.

REFERENCES