Test of the Randomness of Residuals and Detection of Potential Outliers for the Modified Gompertz Model Used in the Fitting of the Growth of *Shigella flexneri*

Garba Uba* and Bilal Ibrahim Dan-Iya

1Department of Science Laboratory Technology, College of Science and Technology, Jigawa State Polytechnic, Dutse, PMB 7040, Nigeria.
2Pharmacy Technician Department, College of Health Sciences and Technology, Kano 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

One way to conceptualize a biological system is as a collection of several 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 an elemental unit of some kind that can be examined, but the inside structure of the thing is either unknown or it does not exist. The scale that is applied to the presentation of the system will be decided based on the elemental unit that is selected. The availability of data that represent a variety of biological states and processes, as well as the temporal dependencies of those activities, makes it feasible to conduct research on biological systems at a wide variety of various levels of organization. These levels start at the molecular level and go all the way up to the population level. In the field of high-throughput genomics and proteomics research, the application of the modified Gompertz model modelled the growth of *Shigella flexneri*. The changes that have taken place in the fields of biotechnology and information technology have both sped up the process of recognizing the knowledge that is held inside biological systems, which has resulted in a shortened amount of time needed to complete the procedure. Because of these breakthroughs, the approaches taken in the field of biomedicine to research, development, and application are undergoing significant transformation. With the inclusion of clinical data to biological data, it is feasible to offer detailed descriptions of both healthy and sick states, as well as the evolution of sickness and the body's reaction to treatments. This is made possible by the combination of the two sets of data [8–14].

The availability of data that represent a variety of biological states and processes, as well as the temporal dependencies of those activities, makes it feasible to conduct research on biological systems at a wide variety of various levels of organization. These levels start at the molecular level and go all the way up to the population level. In the field of high-throughput genomics and proteomics research, the application of
mathematical and computational models to help in the comprehension of biological data is becoming increasingly common. The formulation of hypotheses and the recommendation of experiments as the subsequent stages of the research process are both brought about as a result of the utilization of complicated computer models that make it possible to represent intricate biological processes.

Computational models are now employing knowledge discovery strategies in order to make use of the vast amounts of data that are recorded in biomedical databases. This is being done in order to make the most of the information that is available. The availability of data that represent a variety of biological states and processes, as well as the temporal dependencies of those activities, makes it feasible to conduct research on biological systems at a wide variety of various levels of organization. These levels start at the molecular level and go all the way up to the population level. In many instances, the relationship of observed phenomenon to time or concentration can be mathematically modelled using software via least square methods often used in nonlinear regression [8–14].

In spite of this, the residuals of the curve in a nonlinear regression need to have a natural dispersion. This is in contrast to the usual least square’s technique, which calls for the residuals of a linear regression to have a normal distribution. This is due to the fact that the principle of least squares forms the foundation of the traditional least square’s technique. The residuals, which are significantly more important, have to be random and have the same variance (homoscedastic distribution). It is possible to determine whether or not the randomization procedure was effective by employing the Wald-Wolfowitz runs test. [15]. 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 of the curve in a nonlinear regression need to have a natural dispersion, whereas in a linear regression, the 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 detecting whether or not randomization has been accomplished. Because biological systems are intrinsically unexpected, the model can be depended upon to be correct in terms of statistics. [19–21].

**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);

\[ e_i = y_i - f(x_i; \hat{\Theta}) \]  
(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 the residual data from the modified Gompertz model used in the Fitting of the growth of *Shigella flexneri* [16].

**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. 1) and to test for both data instantaneously (Eqn. 2).

\[ G_{min} = \frac{\bar{X} - \min(X)}{s} \]  
(Eqn. 2)

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

\[ p_{1} = 2n_{1}p_{1} \left( \frac{G_{1}(n-2)}{n-1} \right) \]

\[ p_{2} = n_{2}p_{2} \left( \frac{G_{2}(n-2)}{n-1} \right) \]

\[ G_{all} = \frac{\max(X) - \min(X)}{s} \]

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) recognising 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, it is necessary for the curve’s residuals to have a natural distribution. This is in contrast to the requirements of the least squares approach, which ask for the residues to have a regular distribution. Those requirements may be found here. In addition to this, residuals are needed to have the same variance and be random at the same time (homoscedastic distribution). The Wald–Wolfowitz test is utilised for the purpose of detecting whether or not randomization has been accomplished. Because biological systems are intrinsically unexpected, the model can be depended upon to be correct in terms of statistics. [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. 3). 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 [15,22].

This work is licensed under the terms of the Creative Commons Attribution (CC BY) (http://creativecommons.org/licenses/by/4.0/).
The test statistic is

$$H_0 = \text{the sequence was produced randomly}$$

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

$$Z = \frac{R - \bar{R}}{sR}$$  \hspace{1cm} (Eqn. 3)

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 (Eq. 4). The calculation of the respective values of $\bar{R}$ and $sR$ ($n_2$ is positive while $n_1$ is negative signs) is as follows.

$$\bar{R} = \frac{2n_1n_2}{n_1+n_2} + 1$$  \hspace{1cm} (Eqn. 4)

$$s^2R = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$  \hspace{1cm} (Eqn. 5)

As an example

Test statistic: $Z = 3.0$

Significance level: $\alpha = 0.05$

Critical value (upper tail): $Z_{0.05} = 1.96$

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

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. Data known as residuals play a crucial part in the statistical analysis of nonlinear regression, which is done using nonlinear regression. The discrepancy between the data that was actually collected and the data that was expected to be collected is denoted by the residuals.

The term "residuals" refers to the disparities that exist between the values that are predicted by a mathematical model and the values that are actually seen in the data. These disparities may be seen when comparing the model's predictions to the data. The residuals need to be analyzed statistically in order to determine whether or not they are sufficiently random, do not include any outliers, adhere to the normal distribution, and do not exhibit autocorrelation. This may be done by determining whether or not they do not have any outliers. It is common practice to display the data on residues in the form of positive and negative values. This is an essential component for establishing that the data are balanced, and it can be visually observed before any tests are carried out.

When carrying out nonlinear regression, it is standard practice to disregard the findings of any residual tests that are carried out. As a general rule, the quality of a model is deemed to be worse when there is a bigger difference between the values that were predicted and those that were actually observed. This is because there is a lesser degree of correlation between the two sets of data, which explains why this is the case [23]. The residuals for the modified Gompertz model used in the Fitting of the growth of *Shigella flexneri* are shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Residuals for the modified Gompertz model used in the Fitting of the growth of <em>Shigella flexneri</em>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (h)</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

Grubbs’ test indicated that there was no hint of an outlier when applied to previously released data. This indicates that the model was able to properly describe the data. When attempting to fit a nonlinear curve, it is easy to introduce considerable inaccuracies if either the mean or a single data point from a triple is warped. Both of these circumstances set the same result. The Grubbs test is capable of detecting a single abnormality across any specified time period. It is critical to seek for and delete any outlying data points when fitting curves [24–30]. Because this specific data point was considered to be an outlier, it was eliminated from the collection, and the analysis was repeated until no further outliers were found. Because the test consistently identifies the great majority of points as outliers, sample sizes of six or less are not recommended. Furthermore, repeating the test numerous times might vary the likelihood that it will detect something.

The Grubbs' test statistic zeroes in on the sample value with the largest absolute deviation from the sample mean, as measured by the sample's standard deviation, to select the winner. If the test statistic $g$ yields a result that is bigger than the critical value, the result is referred to as an outlier. This is due to the fact that the critical value is the lowest value that can be tolerated. The results of Grubbs's test suggested the absence of an outlier (Table 2). This suggests that the model was adequate.
A possible outlier is an extreme data point that the investigator believes is implausible because it does not meet a number of specific characteristics. An outlier is a figure that stands out as being notably different from the rest of the data in a sample. For example, an outlier is defined as a maximum that is statistically considerably larger than the maximum distribution anticipated by the population model employed in engineering. This criterion is used to assess if the maximum is an outlier or not. The Chauvenet’s criterion, the 3-sigma criterion, and the Z-score may all be used to identify potential measurement outliers. The Z-score is commonly used in chemometrics in conjunction with the 3-sigma criteria.

A boxplot is a simple way for identifying potential outliers in measurement data [26]. A statistical test is recommended for evaluating whether or not a data set contains an outlier, despite the fact that the methods in question are simple, quick, and can pass visual inspections. The Dixon’s Q-test and the Grubbs’ ESD-test are two specific tests that may be used to identify whether or not a person is an outlier. 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 test found that there were 5 total runs, although the randomness assumption predicted 7.46 runs (Table 3). This implies that the residuals collection only includes runs that are marginally relevant. The Z-value is the number of standard errors by which the actual number of runs differs from the projected number of runs, and the p-value denotes the degree to which this Z-value is significant. The interpretation is the same as for any other statistic that includes p-values. If the p-value is less than 0.05, which is the threshold at which the null hypothesis may be rejected, it is possible to conclude that the residuals are not truly random. The null hypothesis is not rejected since the p-value is marginally greater than 0.05; this suggests that there is no convincing evidence of the residuals’ non-randomness; rather, the residuals represent noise. When there are a large number of occurrences of a certain 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 [22]. When a specific model is used, the runs test can detect a systematic deviation from the curve, such as an overestimation or underestimation of the sections. This is performed by comparing the actual and projected values.

Table 2. Detection of outlier for the residuals for the modified Gompertz model used in the Fitting of the growth of Shigella flexneri.

<table>
<thead>
<tr>
<th>Row</th>
<th>Value</th>
<th>Z</th>
<th>Significant Outlier?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0365</td>
<td>1.207</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.0286</td>
<td>0.9652</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-0.0074</td>
<td>0.2574</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.0659</td>
<td>2.0601</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.0246</td>
<td>0.8347</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.0503</td>
<td>1.6578</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.0281</td>
<td>0.9466</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.0143</td>
<td>0.5060</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.0067</td>
<td>0.1670</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.0171</td>
<td>0.4999</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.0134</td>
<td>0.3803</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.0173</td>
<td>0.5056</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.0159</td>
<td>0.4610</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Runs test for randomness.

<table>
<thead>
<tr>
<th>Runs test</th>
<th>Residual data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>k=5</td>
<td></td>
</tr>
<tr>
<td>n0=6</td>
<td></td>
</tr>
<tr>
<td>n1=13</td>
<td></td>
</tr>
<tr>
<td>E(R)=7.46</td>
<td></td>
</tr>
<tr>
<td>Var(R)=2.94</td>
<td></td>
</tr>
<tr>
<td>SDxv(R)=1.71</td>
<td></td>
</tr>
<tr>
<td>Z=1.44</td>
<td></td>
</tr>
<tr>
<td>p-value=0.0756</td>
<td></td>
</tr>
</tbody>
</table>

This work is licensed under the terms of the Creative Commons Attribution (CC BY) (http://creativecommons.org/licenses/by/4.0/).
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

In the framework of this investigation, the Wald-Wolfowitz runs test was performed. The observations led to the conclusion that the residual series had an acceptable number of runs; this was the result of the observations. The p-value indicates how significant this z-value is, whereas the z-value discloses the number of typical errors that were detected when the number of runs actually found exceeded the number of runs predicted to be found. The runs test found that there were 5 total runs, although the randomness assumption predicted 7.46 runs. The null hypothesis is not rejected since the p-value is greater than 0.05; this suggests that there is no convincing evidence of the residuals’ non-randomness; rather, the residuals represent noise. In addition, the Grubb’s outlier test shows no indication of an outlier, further corroborate the scenario of the adequacy of the model used in fitting the data.

REFERENCES


