HIBISCUS PUBLISHER

NANOBIO BIONANO
www.journal.hibiscuspublisher.com

Statistical Diagnostic Tests of Residuals From the Gompertz Model Used in the Fitting of the Growth of *E. coli* Measured Using a Real-Time Impedimetric Biosensor

Shukor, M.S.[1] and Shukor, M.Y.*[1,2]

[1]Snoc International Sdn Bhd, Lot 343, Jalan 7/16 Kawasan Perindustrian Nilai 7, Inland Port, 71800, Negeri Sembilan, Malaysia.
[2]Department of Biochemistry, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, UPM 43400 Serdang, Selangor, Malaysia.

History
Received: Dec, 2014
Revised: Dec, 2014
Accepted: Dec, 2014

Keyword
Biosensor
Impedance
E. coli
Gompertz
Buchanan three-phase
Statistics

The development of in situ sensor for measuring bacterial concentrations in biotechnology and the health sciences would allow real-time monitoring of the concentration of bacteria. Kim et al [1] has developed such a method using impedance spectroscopy, and was able to measure in real-time the concentration of *E. coli* at 0.01 MHz frequency using impedance changes. We modeled the growth kinetics using several nonlinear regression methods and discovered that the modified Gompertz model is the best model for the growth of the bacterium [2]. It is well known that nonlinear regression of a data and further statistical analysis to find the best model relies on the facts that the residuals (difference between observed and predicted data) followed a normal or Gaussian distribution and that the data must be free of outliers. If all of these assumptions are satisfied, the test is said to be robust. In this work we perform statistical diagnostics to the residuals to satisfy the requirements above and found that removal of an outlier allows the residuals to conform to all of the requirements above. The results indicated that remodelling of the Gompertz model using the new set of data should be carried out.

Monitoring bacterial growth has been traditionally carried out using plate count agar or through counting on a haemocytometer. These methods are time consuming, require trained personnel and cannot be carried out in real-time. Due to this, several biosensor-based methods have been developed to overcome these hurdles including impedemetric biosensor. Impedance spectroscopy utilizes electrical properties of materials and their interfaces with electronically conducting electrodes. It is a relatively novel and powerful method [1,3,4]. The use of this method by Kim et al. [1] for monitoring bacterial growth has been explored and showed promising results. The resultant bacterial growth showed a unique sigmoidal characteristics of bacterial growth including a lag time (λ) followed by an acceleration to a maximal value (μ_{max}) or exponential phase culminating in a final phase in which the rate decreases and finally reaches zero, so that an asymptote (A) is reached [5]. Of several of the models we used such as the modified Logistic [5,6], modified Gompertz [5,7], modified Richards [5,8], modified Schnute [5,9], Baranyi-Roberts [10] and Von Bertalanffy [11], Buchanan three-phase [12] and Huang model [13], the modified Gompertz was fund to be the best [14]. The nonlinear regression method used for choosing this model relies on the Levenberg–Marquardt algorithm (LMA), also known as the damped least-squares (DLS) method [15]. However, the subsequent statistical tests used such as F-test, t-test, Chi-square test and Pearson correlation coefficient rely heavily on the residuals for the curve to be normally distributed and random [15]. Furthermore, the residuals must be tested first for the presence of outliers (at 95 or 99% of confidence) normally using the Grubb's test in order for these assumption to be met. Data distortions by a single data point either the mean or a single data point from a triplicate can lead to gross error in the fitting of a nonlinear curve. Checking for outlier is thus an important part of curve fitting. In this work we perform statistical diagnosis tests such as the Kolmogorov-Smirnov, Wilks-Shapiro and D'Agostino-Pearson tests for normality (normal or Gaussian distribution) and the Wald–Wolfowitz runs test for detecting residual randomness.

Methods

Data were acquired from the works of [2]. The reduction kinetics using the modified Gompertz model (Fig. 1) was used as before to obtain residuals for the regression. Visual observation of the data indicated that data at hour 7 was probably an outlier, and Grubb's test will be used to assess this [16].

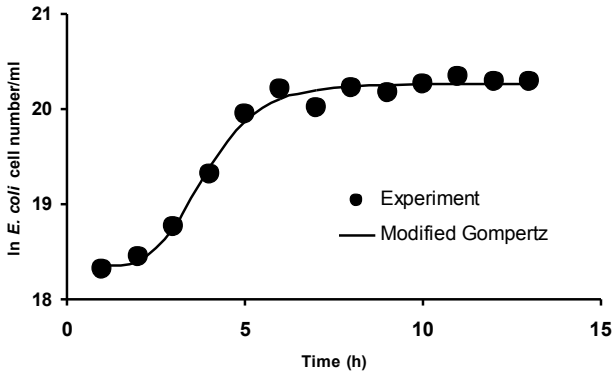

Figure 1. Growth curve of *E. coli* fitted with the modified Gompertz growth model. The number of cells/ml was transformed into natural logarithm.

Grubbs' Statistic

The test is a statistical test used to detect outliers in a univariate data set which is assumed to originate from a population of Gaussian or normal distribution. Grubbs test assume that the data is normally distributed. The test is used to detect outlier in univariate environment [16]. The test can be applied to the maximal or minimal observed data from a Student's t distribution (Eq. 1) and to test for both data simultaneously (Eq. 2).

$$G_{\min}=\frac{\bar{X}-\min(X)}{s} \quad (1)$$

$$G_{\max}=\frac{\max(X)-\bar{X}}{s} \quad (2)$$

$$p_G=2n.p_t\left(G\frac{\sqrt{n(n-2)}}{n-1},n-2,1\right) \quad (3)$$

$$G_{\text{all}}=\frac{\max(\bar{X}-\min(X),\max(X)-\bar{X}}{s} \quad (4)$$

$$p_G=n.p_t\left(G\frac{\sqrt{n(n-2)}}{n-1},n-2,2\right) \quad (5)$$

Normality test

There is two ways to check for normality of residual and is normally carried out through graphical and numerical means. Of the two, graphical methods such as the normal quantile–quantile (Q-Q) plots, histograms or box plots are the simplest and easiest way to assess normality of data. Three of the most reported normality tests were used in this work. They are the Kolmogorov-Smirnov [17,18], Wilks-Shapiro [19] and the D'Agostino-Pearson omnibus K2 [20] test. These tests were used to test for the normality of the residuals. The detail mathematical basis of these normality test statistics is extensive and is available in the literature. The normality tests were carried out using the GraphPad Prism® 6 (Version 6.0, GraphPad Software, Inc., USA).

Runs test

The runs test is also called Wald–Wolfowitz test, after Abraham Wald and Jacob Wolfowitz. It is a non-parametric statistical test that checks for the randomness hypothesis. The runs test could detect a systematic deviation of over or under estimation sections of the curve when using a specific model. This test was carried out to the residuals of the regression in order to detect randomness in the residuals. The number of runs of sign is usually expressed in the form of a percentage of the maximum number possible. The runs test look at the sequence of the residuals that are usually positive and negative. A good runs is usually signifies by alternating or a balance number of positive and negative residual values. The runs test calculates the probability for the presence of too many or too few runs of sign (Eq. 3). The presence of too few runs could indicate a clustering of residuals with the same sign or the presence of systematic bias while the presence of too many of a run sign could indicate the presence of negative serial correlation [15,21].

The test statistic is

H_0 = the sequence was produced randomly
H_a = the sequence was not produced randomly

$$Z=\frac{R-\bar{R}}{sR} \quad (6)$$

$$\bar{R}=\frac{2n_1.n_2}{n_1+n_2}+1 \quad (7)$$

$$s^2R=\frac{2n_1.n_2\left(2n_1.n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)} \quad (8)$$

Results and Discussions

Residual is the difference between measured and predicted values of a regression model either linear or nonlinear. Residuals could show how accurate a mathematical function in the form of a curve in representing sets of data. Residuals are the difference between predicted and observed values of a mathematical model and statistical tests should be carried out to test for the adequacy of the residuals in obeying normality, randomness and does not contain outlier. The rule of thumb is that the larger the difference between the predicted and observed values, the poorer the model.

Plot of residuals (observed-predicted) were checked and the analysis showed that the data were randomly distributed for all tests. In addition all normality tests carried out shows the residual conforming to the normal distribution (Table 1). The Grubbs' test was applied in order to identify the outlier(s). The Grubbs' test statistic identifies the largest absolute deviation from the sample mean in units of the sample standard deviation. The Grubbs' test did nor indicates the presence of any outlier. Residuals are very important in assessing the health of a curve from a particular used model. Mathematically, residual for the i^{th} observation in a given data set can be defined as follows;

$$e_i=y_i-f\left(x_i;\hat{\beta}\right) \quad (9)$$

where y_i denotes the i^{th} response from a given data set while x_i is the vector of explanatory variables to each set at the i^{th} observation corresponding values in the data set.

The Q-Q plot could be used as a visual indication for normality. The residuals data when plotted on the normal probability Q-Q plot of residuals for the modified Gompertz model showed an almost straight line and indicates no underlying pattern (Fig. 3). The resulting histogram of the residuals showed at first a non Gaussian distribution but the normality tests showed that the residuals were indeed conforming to normality. The histogram was then overlaid with the resulting normal distribution curve (Fig. 4).

Number of bins and samples examined determined the shape of the distribution. The Kolmogorov-Smirnov statistic is a non-parametric numerical test that compares the cumulative frequency of residuals. It calculates the agreement between the model and observed values. It

Table 1. Numerical normality test for the residual from the modified Gompertz model.

Normality test Analysis	
D'Agostino & Pearson omnibus test	
K2	2.775
P value	0.2497
Passed normality test (alpha = 0.05)?	Yes
P value summary	ns
Shapiro-Wilk test	
W	0.9442
P value	0.5130
Passed normality test (alpha = 0.05)?	Yes
P value summary	ns
Kolmogorov-Smirnov test	
KS distance	0.1297
P value	> 0.1000
Passed normality test (alpha = 0.05)?	Yes
P value summary	ns
Skewness	-0.8819
Kurtosis	0.7503

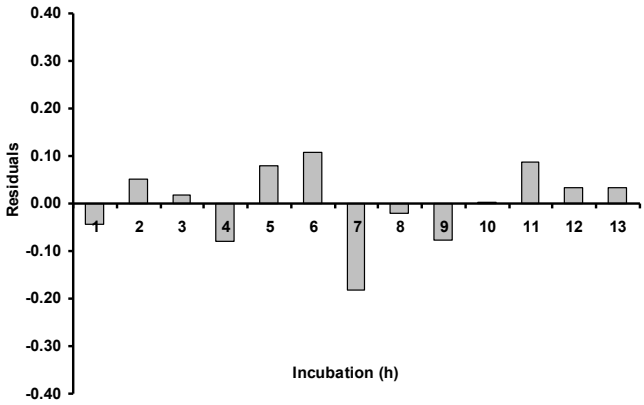

Figure 2. Residual plot for the modified Gompertz model.

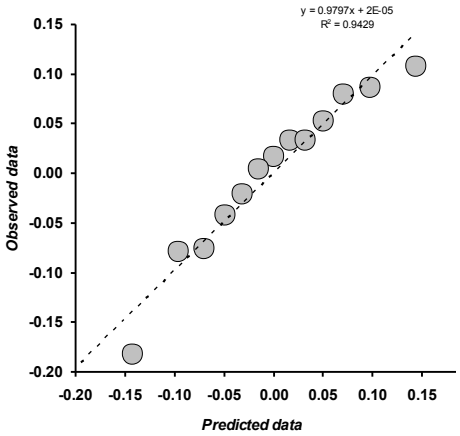

Figure 3. Normal Q-Q plot for the observed sample against theoretical quantiles.

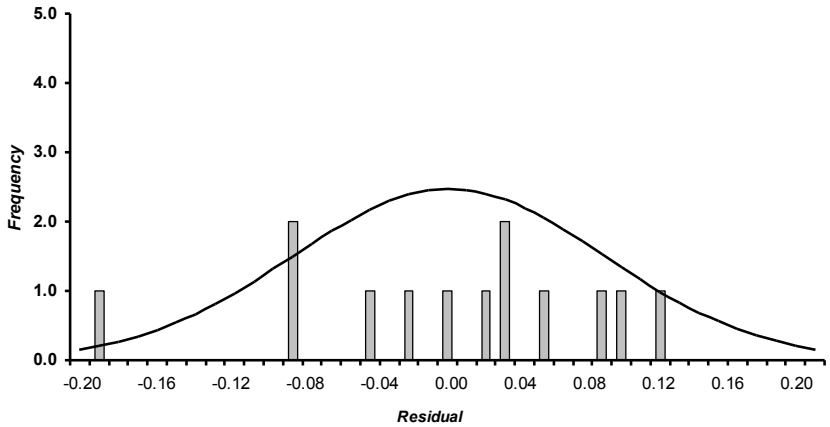

Figure 4. Histogram of residual for the modified Gompertz model overlaid with a normal distribution (mean 0.00078 and standard deviation 0.08084).

calculates the agreement between the model and observed values. It could also be used as a measure between two series of observation. The p value is calculated for the difference between two cumulative distributions and sample size [17,18]. The skewness and kurtosis of the distribution is computed as a method to quantify the difference between the sample distributions to a normal distribution. In the Wilks-Shapiro test [19], a W^2 statistic is calculated based on the expected values of the order statistics between identically-distributed random variables and their independent covariance and the standard normal distribution, respectively. If the test statistics value-W^2 is high, then the agreement is rejected. In the D'Agostino-Pearson normality test method, a p-value from the sum of these discrepancies is then computed. The most often form of the D'Agostino-Pearson normality tests is the omnibus K2 test as D'Agostino developed several normality tests [15].

Runs test

The runs test showed that the number of runs was 6, while the expected number of runs under the assumption of randomness was 7.46 (Table 2), indicating the series of residuals had marginally adequate runs. The Z-value indicates how many standard errors the observed number of runs is below the expected number of runs, the corresponding p-value indicate how extreme this z-value is. The interpretation is the same like other p-values statistics. If the p-value is less than 0.05 then the null hypothesis that the residuals are indeed random can be rejected. Since the p-value was greater than 0.05, therefore the null hypothesis is not rejected indicating no convincing evidence of non-randomness of the residuals and they do represent noise. The presence of too many of a run sign could indicate the presence of negative serial correlation whilst the presence of too few runs could indicate a clustering of residuals with the same sign or the presence of systematic bias. The runs test could detect systematic deviation of the curve such as over or under estimation of the sections when using a specific model [15]. The runs test calculates the probability for the presence of too many or too few runs of sign. The runs test is an important tool in nonlinear regression to detect nonrandomness of the residuals [21]. The runs test look at the sequence of the residuals that are usually positive and negative. A good runs is usually signifies by alternating or a balance number of positive and negative residual values. The number of runs of sign is usually expressed in the form of a percentage of the maximum number possible.

Table 2. Runs test for randomness.

Runs test	Residual data set
Observations	6
Below mean	6
Above mean	7
No. of runs	13
E (R)	7.4615
Var (R)	2.9408
StDev (R)	1.7149
Z - value	-0.8523
2 - sided p - value	0.3941

Conclusion

In conclusion, the various statistical tests for the residuals indicated that the use of the modified Gompertz model in fitting of the reduction curve for this bacterium is adequate. The tests statistics carried out in this work is important since if the results obtained violated Gaussian or normal distribution, than non parametric methods such as the Pearson's correlation coefficient either normal or adjusted, root mean square analysis, Kruskal-Wallis (nonparametric ANOVA) test should be used. Another remedy that can be used in the event of nonconformity includes changing to a different model that obeys or fulfills the above robust requirement. These assumptions could avoid errors of the Type I and II errors.

Acknowledgement

This project was supported by a grant from Snoc International Sdn Bhd.

References

1. Kim YH, Park JS, Jung HI. An impedimetric biosensor for real-time monitoring of bacterial growth in a microbial fermentor. Sensor Actuat B-Chem. 2009; 138:270–277.
2. Shukor MS, Shukor MY. Modeling the growth kinetics of *E. coli* measured using real-time impedimetric biosensor. Nanobio Bionano. 2014; 1:52-57.
3. Dweik M, Stringer RC, Dastider SG, Wu Y, Almasri M, Barizuddin S. Specific and targeted detection of viable *Escherichia coli* O157:H7 using a sensitive and reusable impedance biosensor with dose and time response studies. Talanta. 2012; 94:84–89.
4. Ward AC, Connolly P, Tucker NP. Pseudomonas aeruginosacan be detected in a polymicrobial competition model using impedance spectroscopy with a novel biosensor. PLoS ONE. 2014; 9.
5. Zwietering MH, Wit JCD, Cuppers HGAM, Riet KV. Modeling of bacterial growth with shifts in temperature. Appl Environ Microb. 1994; 60:204–213.
6. Ricker WE. 1979. 11 Growth Rates and Models. p. 677.
7. Gompertz B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos Trans R Soc London. 1825; 115:513–585.
8. Richards FJ. A flexible growth function for empirical use. J Exp Bot. 1959; 10:290–300.
9. Schnute J. A versatile growth model with statistically stable parameters. Can J Fish Aquat Sci. 1981; 38:1128–40.
10. Baranyi J. Mathematics of predictive food microbiology. Int J Food Microbiol. 1995; 26:199–218.
11. Bertalanffy LV. 1951. Heoretische Biologie, Zweiter Band: Stoffwechsel,Wachstum. A FranckeAG Verlag, Bern, Switzerland; p. 418.
12. Buchanan RL, Golden MH. Model for the non-thermal inactivation of Listeria monocytogenes in a reduced oxygen environment. Food Microbiol. 1995; 12:203–212.
13. Huang L. Optimization of a new mathematical model for bacterial growth. Food Control. 2013; 32:283–288.
14. Abd Rachman AR, Halmi MIE, Shukor MY. Amplification of new isolated luciferase gene from marine Photobacterium strain MIE by using specific PCR. J Environ Microbiol Toxicol. 2014; 2:35–7.
15. Motulsky HJ, Ransnas LA. Fitting curves to data using nonlinear regression: a practical and nonmathematical review. FASEB J Off Publ Fed Am Soc Exp Biol. 1987; 1:365–374.
16. Grubbs F. Procedures for detecting outlying observations in samples. Technometrics. 1969; 11:1–21.
17. Kolmogorov A. Confidence limits for an unknown distribution function. Ann Math Stat. 1941; 12:461–463.
18. Smirnov N. Table for estimating the goodness of fit of empirical distributions. Ann Math Stat. 1948; 19:279–281.
19. Royston P. Wilks-Shapiro algorithm. Appl Stat. 1995; 44:R94.
20. D'Agostino RB. 1986. Tests for Normal Distribution. In: D'Agostino RB, ed. Stephens MA, Goodness-Of-Fit Techniques. Marcel Dekker
21. Draper NR, Smith H. 1981. Applied Regression Analysis. Wiley, New York;