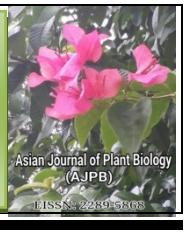
HIBISCUS

ASIAN JOURNAL OF PLANT BIOLOGY

Website: http://journal.hibiscuspublisher.com/index.php/AJPB



Statistical Diagnosis Test of the Growth Kinetics Model for the Algae *Dunaliella tertiolecta*

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

[1]Department of Biochemistry, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, UPM 43400 Serdang, Selangor, Malaysia.

[2]Snoc International Sdn Bhd, Lot 343, Jalan 7/16 Kawasan Perindustrian Nilai 7, Inland Port, 71800, Negeri Sembilan, Malaysia.

*Corresponding author:
Assoc. Prof. Dr. Mohd. Yunus Abd. Shukor
Department of Biochemistry,
Faculty of Biotechnology and Biomolecular Sciences,
Universiti Putra Malaysia, 43400 UPM Serdang,
Selangor, Malaysia.
Email: yunus.upm@gmail.com

HISTORY

Received: 3th December 2014
Received In Revised Form: 26th December 2014
Accepted:29th December 2014

KEYWORDS

Baranyi-Roberts
ordinary least squares method
normal distribution
homoscedastic
autocorrelation

ABSTRACT

Mathematical modeling of physical, chemical or biological data could help the investigator to explain a phenomenon observed based on physical, chemical or biological mechanisms. The model could also be used to predict or forecast future behavior, simulate a hypothetical event or input and design better experiments. Previously, we demonstrated that the Baranyi-Roberts growth kinetics is the best model using the ordinary least squares method for the growth of the algae *Dunaliella tertiolecta* compared to other models such as modified logistic, modified Gompertz, modified Richards, modified Schnute, Baranyi-Roberts, Von Bertalanffy, Huang and the Buchanan three-phase linear model. The ordinary least squares method relies heavily on several important assumptions such as residuals conformation to normal distribution, does not have outliers, is truly random, of equal variance (homoscedastic) and does not show autocorrelation. If all of these assumptions are satisfied, the test is said to be robust. In this work we perform statistical diagnosis test for the adequacy of the model to satisfy these requirements.

INTRODUCTION

In biotechnology, a mathematical model could help in understanding the basis behind a biological process and predicting yield and cellular growth kinetics during a bioprocess event. A mathematically-based model is not equivalent to a theory or a hypothesis since models could not be verified directly through experiments [1]. Algae, similar to microbial growth often shows growth with several phases where the specific growth rate starts at the value of zero. This is followed by acceleration to a maximal value (μ_{max}) for a given period of time, resulting in what is called the lag time (λ). Finally the growth curves exhibit a final phase where the rate decreases and eventually reaches zero or an asymptote (a). The growth phases usually resulted in a sigmoidal curve [2]. The sigmoidal curve can be fitted by various mathematical functions such as Gompertz, logistics, von Bertalanffy, Buchanan and Baranyi-Robetrs. Previously we successfully modelled the sigmoidal growth profile of the algae *Dunaliella tertiolecta* using the non-linear regression model of Baranyi-Roberts to fit the experimental data and obtain parameter constants [3]). The method of mathematically fitting the non linear curve is through the use of ordinary least squares method that relies heavily on the residuals for the curve to be normally distributed of equal variance (homoscedastic), random and does not show autocorrelation [4–8]. In order for these assumption to be met we perform statistical diagnosis tests such as the Kolmogorov-Smirnov, wilks-shapiro and d'agostino-Pearson tests for normality (normal or gaussian distribution), the Wald–

Wolfowitz runs test for detecting residual nonrandomness, Durbin-Watson test for detecting autocorrelation.

METHODOLOGY

Data were acquired from the works of Chen et al. from Figure 4a [9]. the effect of different light intensity on the growth of *Dunaliella tertiolecta* was modelled using the Baranyi-Roberts model (**Fig**. 1) as before to obtain residuals for the regression.

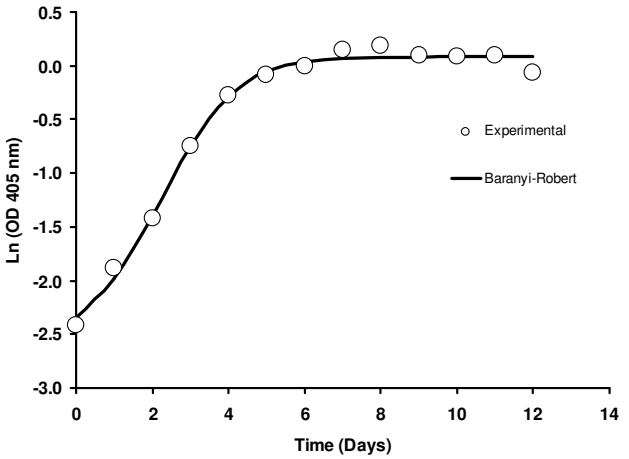


Fig 1. Growth curves of *Dunaliella tertiolecta* fitted by the Baranyi-Roberts model. The optical density was transformed into natural logarithm.

Normality test

Residuals from the Baranyi-Roberts model were subjected to the normality tests. Two ways to check for normality are through graphical and numerical means. 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 normality tests- Kolmogorov-Smirnov [7,10] Wilks-Shapiro [11] and the D'Agostino-Pearson omnibus K2 test [12] were used in assessing 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 [13] was carried out to the residuals of the regression in order to detect nonrandomness. This could detect a systematic deviation of over or under estimation sections of the curve when using a specific model [14]. 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. The runs test calculates the probability for the presence of too many or too few runs of sign. 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 test statistic is

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

$$Z=\frac{R-\bar{R}}{sR} \quad (1)$$

Where Z is the test statistic, $\bar{R}$ is the expected number of runs, R is the observed number of runs and sR is the standard deviation of the runs. The computation of the values of $\bar{R}$ and sR (n_1 is positive while n_2 is negative signs) is as follows;

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

$$s^2R=\frac{2n_1.n_2\left(2n_1.n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)} \quad (3)$$

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 $|Z| > 1.96$
Since the test statistic value (Z) is larger than the critical value then the null hypothesis is rejected at the 0.05 significance level or the sequence was produced in a non random manner.

The Durbin-Watson test

Nonlinear regression normally uses the assumption that data points do not depend on each other or the value of a data point is not dependent on the value of preceding or proceeding data points. Autocorrelation amongst data can occur due to events such as temperature drift during time measurements or an overused tungsten lamp in a spectrophotometer. If one were to count the number of animals per year in a given area the data would be highly autocorrelated and nonindependence as the number of animals in a current year would be highly dependent upon the number of animals in the previous year [15]. The Durbin–Watson statistic calculates the level of significance according to the method outlined by Draper and Smith [13].

$$d=\frac{\sum_{t=2}^{T}\left(\hat{e}_t-\hat{e}_{t-1}\right)^2}{\sum_{t=1}^{T}\hat{e}_t^2} \quad (4)$$

As usual the hypothesis H_0: $\rho = 0$ versus the alternative $H1$: $\rho > 0$ is tested. The statistic is approximately equal to $2(1- p)$. The Durbin-Watson test statistic equals 2 when the ρ value is zero while a ρ value of one equals a Durbin-Watson test statistic of 0. Non-autocorrelation is indicated by a d value near 2 while a value

towards 0 indicates positive autocorrelation. Negative autocorrelation is indicated by d values nearing 4.

The null hypothesis should be rejected for a low value of the Durbin-Watson test statistic indicating significant autocorrelation. Unlike the t- or z-statistics, the distribution of the Durbin-Watson test statistic is not available for ρ-value associated with d and tables must be used in the hypothesis testing.

The decision rule for the Durbin-Watson bounds test is
• if d > upper bound, fail to reject the null hypothesis of no serial correlation,
• if d < lower bound, reject the null hypothesis and conclude that positive autocorrelation is
present,
• if lower bound < d < upper bound, the test is inconclusive.

RESULTS

The fit of a statistical model can be diagnosed accurately using tests that use residuals. Residuals are the difference between a predicted and observed quantity using a particular mathematical model. The rule of thumb is that the larger the differenced between the predicted and observed values, the poorer the model.

Plot of residuals (observed-predicted) were checked and there were no evidence of a trend and the residuals appears to be randomly distributed. The normal probability Q-Q plot of residuals for the Baranyi-Roberts model was almost in a straight line and appears to show no underlying pattern (**Fig.** 2). The residual plot (**Fig.** 3) and the resulting histogram overlaid with the resulting normal distribution curve (**Fig.** 4) indicated the residuals were truly random and the model used was appropriately fitted.

Graphical diagnostic of residuals normality

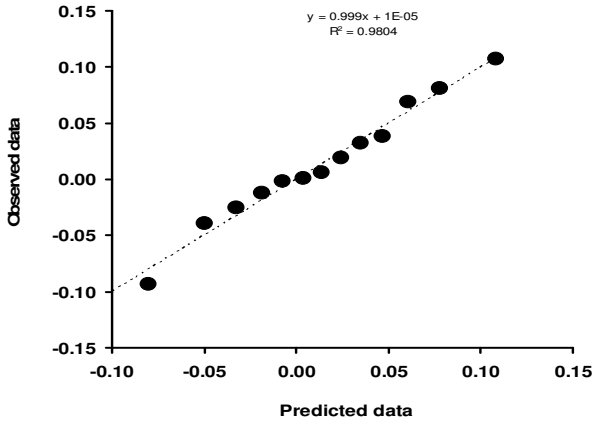


Fig 2. Normal Q-Q plot for the observed sample against theoretical quantiles.

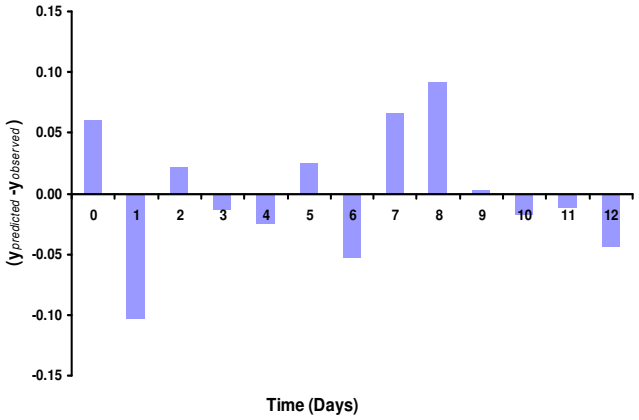


Fig. 3. Residual plot for the Baranyi-Roberts model.

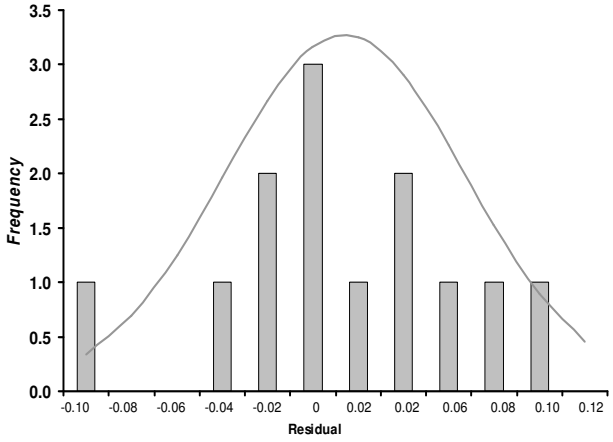


Fig. 4. Histogram of residual for the Baranyi-Roberts model overlaid with a normal distribution (mean 0.0140 and standard deviation 0.053).

All of the normality tests used showed that the residuals are normally distributed (**Table** 1). The shape of the distribution calculated is dependent upon the number of bins and samples examined. 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 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 [7,10]. In the Wilks-Shapiro test 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, the skewness and kurtosis of the distribution is computed as a method to quantify the difference between the sample distributions to a normal distribution. 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.

Table 1. Numerical normality test for the residual from the Baranyi-Roberts model.

Normality tests	Diagnostic
D'Agostino & Pearson omnibus	
K2	0.2541
P value	0.8807
Passed normality test (alpha=0.05)?	Yes
P value summary	ns
Shapiro-Wilk	
W	0.9842
P value	0.9939
Passed normality test (alpha=0.05)?	Yes
P value summary	ns
Kolmogorov-Smirnov	
KS distance	0.09633
P value	> 0.1000
Passed normality test (alpha=0.05)?	Yes
P value summary	ns
Skewness	-0.1046
Kurtosis	0.2944

Runs test

From **Table** 2, the number of runs was 13, the expected number of runs under the assumption of randomness was 7.461538, indicating the series of residuals had 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 o-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.

Table 2. Runs test for randomness.

Runs test	Residual data set
Observations	8
Below mean	7
Above mean	6
No of runs	13
E(R)	7.461538
Var(R)	2.940828
StDev(R)	1.714884
Z-value	0.313993
p-value	0.623237

The runs test calculates the probability for the presence of too many or too few runs of sign. 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 is an important tool in nonlinear regression to detect nonrandomness of the residuals [13]. The runs test could detect systematic deviation of the curve such as over or under estimation of the sections when using a specific model. 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 [14].

Durbin-Watson test of autocorrelation

Serial correlation of residuals was examined further with the Durbin–Watson statistic (DW) [13]. The DW is used to test whether a model has been successful in describing the underlying trend. Autocorrelation, also known as serial correlation, is the cross-correlation of a signal with itself. Informally, it is the similarity between observations as a function of the time lag between them. It is a mathematical tool for finding repeating patterns, such as the presence of a periodic signal obscured by noise. Because most regression problems involving time series data exhibit positive autocorrelation. Autocorrelation amongst data can occur due to events such as temperature drift during time measurements or an overused tungsten lamp in a spectrophotometer. If one were to count the number of animals per year in a given area the data would be highly autocorrelated and nonindependence as the number of animals in a current year would be highly dependent upon the number of animals in the previous year [15].

The value of the Durbin-Watson statistics d = 0.0760/0.0367=2.0694. As usual the hypothesis H_0: ρ= 0 versus the alternative $H1$: $\rho > 0$ is tested. The statistic is approximately equal to $2(1- p)$. The Durbin-Watson test statistic equals 2 when the ρ value is zero while a ρ value of one equals a Durbin-Watson test statistic of 0. Non-autocorrelation is indicated by a d value near 2 while a value towards 0 indicates positive autocorrelation. Negative autocorrelation is indicated by d values nearing 4. The null hypothesis should be rejected for a low value of the Durbin-Watson test statistic indicating significant autocorrelation. Unlike the t- or z-statistics, the distribution of the Durbin-Watson test statistic is not available for ρ-value associated with d and tables must be used in the hypothesis testing. The upper critical value d_U is 1.826 while the lower critical value d_L is 0.294. Since d was larger than the upper critical value then the null hypothesis is not rejected i.e. there appears to be no evidence of autocorrelation.

In conclusion, various tests for the residuals used in this work has indicated that the use of the Baranyi-Roberts model in fitting of the growth curve of an algae shows adequate statistics strength based on the diagnostics of the residuals. Many publications negate statistical diagnosis of the model they used and the data may have violated normal distribution- an important requirement for all of the parametric statistical evaluation methods to chose a test such as Pearson's correlation coefficient either normal or adjusted, root mean square analysis, F-test and t-test etc.

Statistical diagnosis allows a model used and the underlying statistical assumptions used to be robust. Checking these assumptions would allow the researcher to make sure that an analysis meets the associated assumptions in avoidance of the Type I and II errors. In the event that the dignostic tests shows that the residuals violated normality, shows autocorrelation or the residuals indicate a trend, then various treatments such as nonparametric analysis or changing to a different model should remedy the problem.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Halmi MIE, Ahmad SA, Syed MA, Shamaan NA, Shukor MY. Mathematical modelling of the molybdenum reduction kinetics in *Bacillus pumilus* strain Lbna. Bull Environ Sci Manag. 2014;2(1).

[2] Baranyi J. Mathematics of predictive food microbiology. Int J Food Microbiol. 1995;26(2):199–218.

[3] Halmi MIE, Shukor MS, Johari WLW, Shukor MY. Evaluation of several mathematical models for fitting the growth of the algae *Dunaliella tertiolecta*. Asian J Plant Biol.
2014;2(1).

[4] Durbin J, Watson GS. Testing for serial correlation in least squares regression. I. Biometrika. 1950;37(3-4):409–28.

[5] Dyer AR. Comparisons of tests for normality with a cautionary note. Biometrika. 1974;61(1):185–9.

[6] Jäntschi L, Bolboacă SD, Sestras RE. Meta-heuristics on quantitative structure-activity relationships: Study on polychlorinated biphenyls. J Mol Model. 2010;16(2), 377-386.(2):377–86.

[7] Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. G Dell' Ist Ital Degli Attuari. 1933;4:83–91.

[8] Pearson ES. Note on tests for normality. Biometrika. 1931;22(3/4):423–4.

[9] Chen M, Mertiri T, Holland T, Basu AS. Optical microplates for high-throughput screening of photosynthesis in lipid-producing algae. Lab Chip. 2012;12(20):3870–4.

[10] Smirnov N. **Table** for estimating the goodness of fit of empirical distributions. Ann Math Stat. 1948;19:279–81.

[11] Royston P. Wilks-Shapiro algorithm. Appl Stat. 1995;44(4):R94.

[12] D'Agostino RB. Tests for Normal Distribution. In: D'Agostino RB, Stephens MA, editors. Goodness-Of-Fit Techniques. Marcel Dekker; 1986.

[13] Draper NR, Smith H. Applied Regression Analysis. Wiley, New York; 1981.

[14] 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(5):365–74.

[15] Mcdonald JH, Dunn KW. Statistical tests for measures of colocalization in biological microscopy. J Microsc. 2013;252(3):295–302.