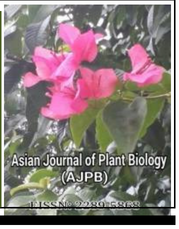
HIBISCUS

ASIAN JOURNAL OF PLANT BIOLOGY

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



Diagnostic of Residuals from the Buchanan Three-phase Model Used in the Fitting of the Growth of *Chlorella vulgaris* Cultivated in Microfluidic Devices

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

HISTORY

Received: 5th December 2014
Received In Revised Form: 24th December 2014
Accepted:27th December 2014

KEYWORDS

Buchanan- three phase model
ordinary least squares method
Grubbs test
normal distribution
homoscedastic
autocorrelation

ABSTRACT

Nonlinear regression of a data and its subsequent statistical analysis relies on the facts that the residuals (difference between observed and predicted data) followed a normal or Gaussian distribution, no autocorrelation and are free of outliers. Previously, we demonstrated that the Buchanan- three phase growth kinetics is the best model using the ordinary least squares method for the growth of the algae *Chlorella vulgaris* 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. If all of these assumptions are satisfied, the test is said to be robust. In this work we perform statistical diagnostics to the residuals and discovered the presence of an outlier that allows the residuals to be normally distributed and satisfy other diagnostic tests after its removal.

INTRODUCTION

The growth of algae is similar to microbial growth in that they often show growth with several phases. The initial phase is where the specific growth rate starts at the value of zero. Then this is followed by what is called the lag time (λ) that lasts for a given period of time before accelerating to a maximal value (μ_{max}). Finally the growth curves exhibit a final phase where the rate decreases and eventually reaches zero or an asymptote (A) . A mathematical model could help in understanding the basis behind kinetics of the algae growth and predicting yield and other secondary kinetics evaluation . The sigmoidal curve can be fitted by various mathematical functions such as Gompertz, logistics, von Bertalanffy, Buchanan and Baranyi-Roberts [6–9]. Previously we successfully modelled the sigmoidal growth profile of the algae *Chlorella vulgaris* using the non-linear regression model of Buchanan- three phase to fit the experimental data and obtain parameter constants. 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 [10]. In addition the residuals must be tested first for the presence of outliers (at 95 or 99% of confidence) using the Grubb's test in order for these assumption to be met [11]. 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 and Durbin-Watson test for detecting autocorrelation.

METHOD

Data were acquired from the works of Dewan et al. [1] from Figure 6A showing *Chlorella vulgaris* growth profile starting from one cell per droplet. The growth of *Chlorella vulgaris* was modelled using the Buchanan- three phase model (**Fig.** 1) as before to obtain residuals for the regression.

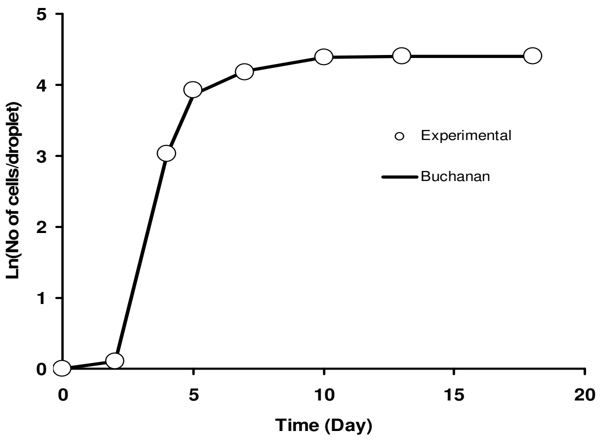


Fig 1. Growth curves of *Chlorella vulgaris* fitted by the Buchanan three-phase model.

Grubbs' Statistic

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. Grubbs test is used to detect outlier in univariate environment and the data is assumed to be normally distributed [11]. The test can be applied to the maximal or minimal observed data from a Student's t distribution (Equation 1) and to test for both data simultaneously (Equation 2).

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

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

$$p_{\mathrm{G}}=2n.p_t\left(G\frac{\sqrt{n(n-2)}}{n-1},n-2,1\right)$$

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

$$p_{\mathrm{G}}=n.p_t\left(G\frac{\sqrt{n(n-2)}}{n-1},n-2,2\right) \qquad (2)$$

Normality test

Residuals from the Buchanan- three phase model were subjected to three normality tests- Kolmogorov-Smirnov [12,13], Wilks-Shapiro [14] and the D'Agostino-Pearson omnibus K2 test [15]. 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. The detail mathematical basis of these normality test statistics is extensive and is available in the literature [10]. The normality tests were carried out using the GraphPad Prism® 6 (Version 6.0, GraphPad Software, Inc., USA).

Runs test

This test was carried out to the residuals of the regression in order to detect nonrandomness [16]. This could detect a systematic deviation of over or under estimation sections of the curve when using a specific model [10]. 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

$$Z=\frac{R-\bar{R}}{sR} \qquad (3)$$

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;

$$\bar{R}=\frac{2n_1.n_2}{n_1+n_2}+1 \qquad (4)$$

$$s^2R=\frac{2n_1.n_2\left(2n_1.n_2-n_1-n_2\right)}{\left(n_1+n_2\right)^2\left(n_1+n_2-1\right)} \qquad (5)$$

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

The Durbin–Watson statistic calculates the level of significance according to the method outlined by Draper and Smith [16]. 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 [17].

$$d=\frac{\sum_{t=2}^{T}\left(\hat{e}_t-\hat{e}_{t-1}\right)^2}{\sum_{t=1}^{T}\hat{e}_t^2} \qquad (6)$$

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 the analysis showed that the data were not randomly distributed for all tests (**Table** 1). This could indicate the presence of an outlier in the residual. 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 [11]. The Grubbs' test identify an outlier for the residual data 0.05521 at a significance level of 5% (α=5%). The presence of this outlier was graphically indicated using the residual plot (**Fig**. 2).

Table 1. Numerical normality test for the residual from the Buchanan- three phase model.

Normality test	Analysis
D'Agostino & Pearson omnibus normality test	
K2	21.06
P value	< 0.0001
Passed normality test (alpha=0.05)?	No
P value summary	****
Shapiro-Wilk normality test	
W	0.5707
P value	< 0.0001
Passed normality test (alpha=0.05)?	No
P value summary	****
KS normality test	
KS distance	0.4351
P value	< 0.0001
Passed normality test (alpha=0.05)?	No
P value summary	****

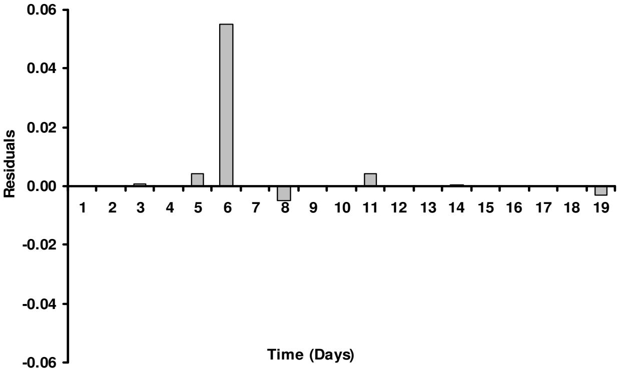


Fig. 2. Residual plot for the Buchanan- three phase model.

The outlier identified by the Grubb's test was and the same tests were again applied in order to assess the normality. The results are presented in **Table** 2. The removal of this outlier as indicated using the residual plot shows uniform random distribution (**Fig**. 3).

Table 2. Numerical normality test for the residual from the Buchanan- three phase model after removal of an outlier.

Normality tests	Diagnostic
D'Agostino & Pearson omnibus normality test	
K2	N too small
P value	
Passed normality test (alpha=0.05)?	
P value summary	
Shapiro-Wilk normality test	
W	0.9230
P value	0.4932
Passed normality test (alpha=0.05)?	Yes
P value summary	ns
KS normality test	
KS distance	0.1873
P value	> 0.1000
Passed normality test (alpha=0.05)?	Yes
P value summary	ns
Skewness	-0.2936
Kurtosis	-0.8828

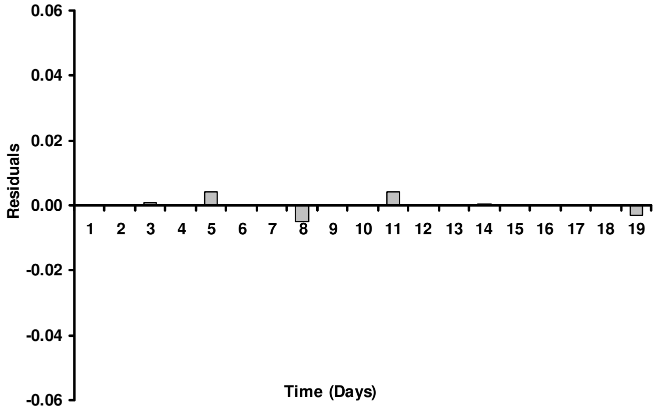


Fig. 3. Residual plot for the Buchanan- three phase model after removal of an outlier.

The normal probability Q-Q plot of residuals for the Buchanan- three phase model was almost in a straight line and appears to show no underlying pattern (**Fig**. 4). The resulting histogram overlaid with the resulting normal distribution curve (**Fig**. 5) indicates the residuals were truly random and the model used was appropriately fitted.

Graphical diagnostic of residuals normality

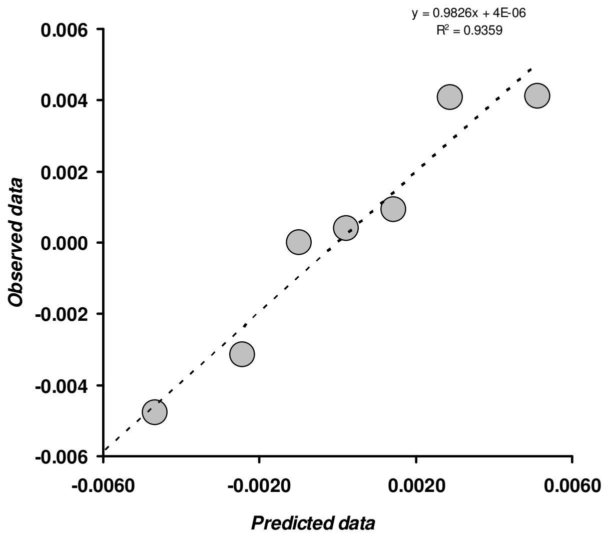


Fig 4. Normal Q-Q plot for the observed sample against theoretical quantiles.

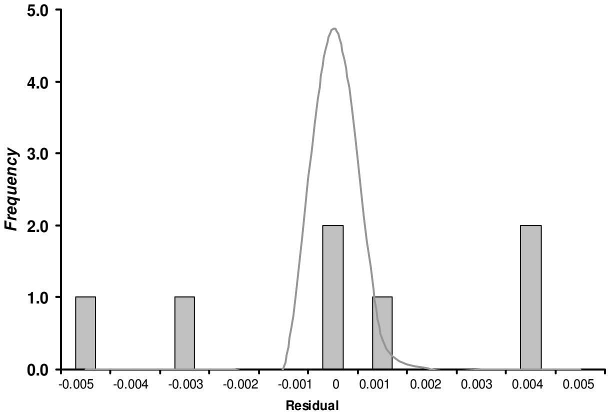


Fig. 5. Histogram of residual for the Buchanan- three phase model overlaid with a normal distribution (mean 0.000227 and standard deviation 0.003344).

After the removal of the outlier, all of the normality tests used showed that the residuals were normally distributed (**Table** 2). Number of bins and samples examined determined the shape of the distribution. 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 [14]. 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 [12,13]. 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 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 7, while the expected number of runs under the assumption of randomness was 4.428 (**Table** 3), 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 *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 [10].

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. 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 [16]. 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 [10].

Table 3. Runs test for randomness.

Runs test	Residual data set
Observations	5
Below mean	3
Above mean	4
No of runs	7
E(R)	4.428571
Var(R)	1.387755
StDev(R)	1.17803
Z-value	0.485071
p-value	0.686187

Durbin-Watson test of autocorrelation

The Durbin–Watson statistic (DW) can calculate for the presence of serial correlation of residuals. Autocorrelation, also known as serial correlation, is the cross-correlation of a signal with itself. The DW is used to test whether a model has been successful in describing the underlying trend. 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. This is 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 [10,16,17].

The value of the Durbin-Watson statistics d = 0.000195/0.000067=2.893. 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 2.102 while the lower critical value d_L is 0.229. 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 Buchanan- three phase model in fitting of the growth curve of an algae shows adequate statistics strength based on the diagnostics of the residuals. It is reported that many publications did not elaborate further on the use of statistical diagnosis of the residuals from the model used. This could results in data violating the Gaussian or normal distribution. This assumption is an important requirement for many of the parametric statistical evaluation methods used in non linear regression. Methods such as the Pearson's correlation coefficient either normal or adjusted, root mean square analaysis, F-test and t-test rely on the residuals to be normally distributed. These assumptions could avoid errors of the Type I and II errors. Furthermore, in the event that the dignostic tests shows that the residuals violated some of the assumptions various nonparametric treatments could be used or changing to a different model can in practice remedy the situation.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Dewan A, Kim J, Mclean RH, Vanapalli SA, Karim MN.Growth kinetics of microalgae in microfluidic static droplet arrays. Biotechnol Bioeng. 2012;109(12):2987–96.

[2] 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):1–6.

[3] Pinto G, Pollio A, Previtera L, Temussi F. Biodegradation of phenols by microalgae. Biotechnol Lett. 2002;24(24):2047–51.

[4] Halmi MIE, Shukor MS, Johari WLW, Shukor MY. Modeling the growth kinetics of Chlorella vulgaris cultivated in microfluidic devices. Asian J Plant Biol. 2014;2(1):7–10.

[5] 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):24–9.

[6] 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. PhilosTransRSocLondon. 1825;115:513–85.

[7] Bertalanffy L von. heoretische Biologie, Zweiter Band: Stoffwechsel,Wachstum. A FranckeAG Verlag, Bern, Switzerland; 1951. 418 p.

[8] Buchanan RL. Predictive food microbiology. Trends Food Sci Technol. 1993;4(1):6–11.

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

[10] 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.

[11] Grubbs F. Procedures for detecting outlying observations in samples. Technometrics. 1969;11(1):1–21.

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

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

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

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

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

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