Statistical Error Model Comparison for Logistic Growth of Green Algae (*Raphidocelis subcapitata*)

H.T. Banks, Elizabeth Collins, Kevin Flores, Prayag Pershad, Michael Stemkovski, Lyric Stephenson

Center for Research in Scientific Computation
Center for Quantitative Sciences in Biomedicine
Department of Mathematics
North Carolina State University
Raleigh, NC, 27695-8212

August 30, 2016

**Abstract**

We validate a model for the population dynamics, as they occur in a chemostat environment, of the green algae *Raphidocelis subcapitata*, a species that is often used as a primary food source in toxicity experiments for the fresh water crustacean *Daphnia magna*. We collected longitudinal data from 4 replicate population experiments with *R. subcapitata*. This data was fit to a logistic growth model to reveal patterns of the algae growth in a continuous culture. Overall, our results suggest that a proportional error statistical model is the most appropriate for logistic growth modeling of *R. subcapiala* continuous population growth.

**Key Words:** Algae growth models, parameter estimation, uncertainty quantification, asymptotic theory
1 Introduction

Population dynamic studies of the predator-prey relationship between zooplankton and phytoplankton in a lentic environment have revealed a variety of patterns including but not limited to low and high amplitude seasonal oscillations and stable year-round plankton populations. A given lake may sustain a stable population one year and an oscillating population the next year [7]. Some suggested explanations for such irregularities include temperature fluctuations, predator-prey interactions, and external effectors of nutrient concentration [5, 8, 9].

The aim of the present study is to develop and parameterize models with reasonable levels of parameter uncertainty for the growth of green algae (Raphidocelis subcapitata) in the absence of predation by Daphnia magna. In the future, we aim to pair this with a Daphnia magna growth model to analyze the predator-prey relationships in a controlled chemostat ecosystem. We tested the frequently-used logistic population growth model (a model previously discussed in this context in [12]) with both constant and proportional measurement error statistical models. Modeling the growth of these algae can be difficult in a field setting since many real life variables are out of the control of the experimenter and even harder to model. A first step to mathematically model microbial growth is through a chemostat, or continuous culture. This allows one to model a very simple “lake” with mathematics that can be carefully analyzed. The chemostat is a system that allows for continuous input and output of nutrients, which simulates an ecological setting [11]. The chemostat consists of three vessels: a feeding vessel, culture vessel, and collection vessel. First, the feeding vessel feeds nutrients into the culture vessel. The culture vessel is connected to a collection vessel, which outputs the nutrient, organisms, and anything that the organisms have produced. The flow of input (feeding vessel to culture vessel) and output (culture vessel to collection vessel) remain at a fixed rate. The inputs and outputs mimic an ecological setting, where the input corresponds to the flow of nutrients into an ecosystem, and the output corresponds to non-specific death, predation, or emigration [6]. This system is normally modeled by the chemostat equations [6] for steady-state operation; however, our focus here is the non-steady-state growth rates. In previous efforts [12] we compared a number of models (Gompertz and Logistic population models, as well as a Discrete Logistic and a Bernoulli model which is a generalization of the basic continuous Logistic model). In that study we compared asymptotic theory with bootstrapping to quantify best fits to chemostat data under the tacit assumption of an absolute error measurement statistical model and found use of the continuous logistic was the preferred model. Here we revisit this same logistic mathematical model under alternative statistical model assumptions (a family of proportional error models) in the context of new additional experimental data. Our focus is thus on the most appropriate statistical model to be used with the more extensive new data sets in conjugation with the logistic model previously studied in [12]. The data sets of our focus along with the related methodology for the experiments are detailed in [2].
2 Parameter Estimation: Mathematical and Statistical Models

The goal of this paper is to provide mathematical and statistical models for algae growth in the absence of consumption. The parameters of the model that best fit the data can be quantified via asymptotic theory \[3, 4\]. In this section we provide the theory behind our asymptotic theory methodology. The estimation of parameters as pursued here using asymptotic theory requires mathematical models of the form

\[
\frac{dx}{dt}(t) = g(t, x(t), q) = rx(t)(1 - \frac{x(t)}{K})
\]

and the corresponding observation process \(f(t, \theta) = x(t, \theta)\), and \(\theta = (q, x_0) \in \mathbb{R}^\kappa\) is the vector of unknown parameters, \(q\) is a vector of model parameters \((r, K)\), and \(x_0\) is the initial condition that is unknown. We consider the initial condition to be unknown because of measurement error. In (1) the state \(x(t)\) is the population as a function of time, \(K\) is the carrying capacity for the selected population, and \(r\) is the growth rate. The parameters of the logistic model, \(r\) and \(K\), as well as the initial population, \(x_0\), are estimated to achieve a logistic growth curve that best fits the data. We used the SimBiology package in MATLAB to estimate parameters for experimental data for each population. The package produced estimates with corresponding standard error values for each parameter; for Population 1 these are presented in Sections 3.2 and 3.3, while similar values for the other populations are given in \[2\].

Due to the discrete nature of our experimental data, the observations for our statistical error model occur at \(N = 23\) discrete times \(t_j\). To account for measurement error, we use statistical models of the form

\[
Y_j = f(t_j, \theta_0) + f(t_j, \theta_0)^\gamma \varepsilon_j, j = 1, 2, 3, ..., N,
\]

for our observations, where \(\varepsilon_j\) is a zero mean random variable representing identically, independently distributed (i.i.d.) noise that causes our observed data to deviate from our model solution (perhaps in a observation weighted manner), and \(\theta_0\) is the hypothesized “true” or “nominal” parameter vector that generates the observations \(\{Y_j\}_{j=1}^N\). The existence of this “true” parameter vector \(\theta_0\) is a standard assumption in frequentist statistical formulations. The i.i.d. nature of the error in our model implies that \(E(\varepsilon_j) = 0\) for each \(j\), and that \(\varepsilon_j, j = 1, \ldots, N\), are identically distributed with variance \(\sigma^2_0\).

Since \(\varepsilon_j\) is a random variable, \(Y_j\) is a random variable with corresponding realizations \(y_j\). In asymptotic theory we seek to estimate \(\theta_0\) by creating a random variable \(\hat{\theta}\) whose realizations for a given data set \(\{y_j\}\) will be estimates \(\hat{\theta}\) of \(\theta_0\).

A special case for \(\gamma = 0\) is called the Ordinary Least Squares (OLS) estimator and is denoted by \(\theta_{\text{OLS}}\). A more general methodology motivated by the weighted least squares (as we have presented it in \[3\]) involves the so-called Generalized Least Squares (GLS) estimator. To define the random vector \(\theta_{\text{GLS}}\) (see \[10\, p. 69\], \[3\, Chapter 3\], \[4\] and the extensive references therein) we take \(\theta_{\text{GLS}}\) given
by:
\[
\sum_{j=1}^{N} f^{-2}\gamma(t_j; \theta_{GLS})[Y_j - f(t_j; \theta_{GLS})] \nabla f(t_j; \theta_{GLS}) = 0_{\kappa_\theta},
\] (3)
where in this case \(\kappa_\theta = 3\) (for \(\theta = (r, K, x_0)\) ) and \(Y_j\) satisfies
\[
Y_j = f(t_j; \theta_0) + f'(t_j; \theta_0) E_j,
\]
and
\[
\nabla f(t_j; \theta) = \left( \frac{\partial f(t_j; \theta)}{\partial \theta_1}, \ldots, \frac{\partial f(t_j; \theta)}{\partial \theta_{\kappa_\theta}} \right)^T.
\]

The quantity \(\theta_{GLS}\) is a random vector, hence if \(\{y_j\}_{j=1}^{N}\) is a realization of \(\{Y_j\}_{j=1}^{N}\), then solving
\[
\sum_{j=1}^{N} f^{-2}\gamma(t_j; \theta)[y_j - f(t_j; \theta)] \nabla f(t_j; \theta) = 0_{\kappa_\theta}
\] (4)
for \(\theta\) will provide an estimate for \(\theta_{GLS}\).

The GLS equation (4) can be motivated by examining the special weighted least squares estimate
\[
\hat{\theta}_{WLS} = \arg \min_{\theta \in \Omega} \sum_{j=1}^{N} w_j[y_j - f(t_j; \theta)]^2.
\] (5)
for a given \(\{w_j\}\). If we differentiate the sum of squares in (5) with respect to \(\theta\) and then choose \(w_j = f^{-2}\gamma(t_j; \theta)\), an estimate \(\hat{\theta}_{GLS}\) is obtained by solving
\[
\sum_{j=1}^{N} w_j[y_j - f(t_j; \theta)] \nabla f(t_j; \theta) = 0_{\kappa_\theta}
\]
for \(\theta\), i.e., solving (4). However, we note the GLS relationship (4) does not follow from minimizing the weighted least squares with weights previously chosen as \(w_j = f^{-2}\gamma(t_j; \theta)\) (see p. 89 of [10]).

The GLS estimator \(\theta_{GLS} = \theta_{GLS}^N\) has the following asymptotic properties [3, 10]:
\[
\theta_{GLS} \sim \mathcal{N}(\theta_0, \Sigma_0^N),
\] (6)
where
\[
\Sigma_0^N \approx \sigma_0^2 \left( F_{\theta}^T(\theta_0) W(\theta_0) F_{\theta}(\theta_0) \right)^{-1},
\] (7)
the sensitivity matrix is given by
\[
F_{\theta}(\theta) = F_{\theta}^N(\theta) = \left( \begin{array}{ccc}
\frac{\partial f(t_1; \theta)}{\partial \theta_1} & \frac{\partial f(t_2; \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(t_1; \theta)}{\partial \theta_{\kappa_\theta}} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial f(t_N; \theta)}{\partial \theta_1} & \frac{\partial f(t_N; \theta)}{\partial \theta_2} & \cdots & \frac{\partial f(t_N; \theta)}{\partial \theta_{\kappa_\theta}}
\end{array} \right),
\]
and the matrix $W(\theta)$ is defined by $W^{-1}(\theta) = \text{diag}(f^2(t_1; \theta), \ldots, f^2(t_N; \theta))$.

Note that because $\theta_0$ and $\sigma^2_0$ are unknown, the estimates $\hat{\theta} = \hat{\theta}_{\text{GLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{GLS}}$ will again be used in (7) to calculate

$$\Sigma_0^N \approx \hat{\Sigma}^N = \hat{\sigma}^2 \left( F^T_{\hat{\theta}} W(\hat{\theta}) F_{\hat{\theta}} \right)^{-1},$$

where we take the approximation

$$\sigma^2_0 \approx \hat{\sigma}^2_{\text{GLS}} = \frac{1}{N - \kappa_\theta} \sum_{j=1}^{N} f^2(t_j; \hat{\theta}) \left[ y_j - f(t_j; \hat{\theta}) \right]^2.$$

We can then approximate the standard errors of $\hat{\theta}_{\text{GLS}}$ by taking the square roots of the diagonal elements of $\hat{\Sigma}^N$.

### 2.1 Calculation of Residuals and Modified Residuals

Plots of the residuals versus model or modified model were made for each population and analyzed to determine the validity of an attempted model. Standard residuals were calculated based on the logistic growth constant error statistical model. Modified residuals were calculated by accounting for the proportional error term in the statistical model. Error is always a confounding variable in in vivo experiments that must be accounted for to obtain reliable uncertainty quantification. The type of error in the observations determines how to account for it. For this particular mathematical model, we analyzed the data using both constant and proportional error statistical models to determine which more accurately quantifies potential observation error in the data.

### 3 Results

In this section, we compare and present the results of logistic growth estimation models using both standard and proportional statistical error models. In the figures, “model” refers to the constant error model, while “modified model” refers to the proportional error model. The initial parameter guesses $(r, K, x_0) = (7.0, 4500, 200)$ were used for all computations for both the constant and proportional error statistical models.

#### 3.1 Data Fit to Logistic Growth Model

Algae data counts were plotted and then fit to the logistic growth mathematical model with an OLS formulation. Shown below in Figure 1 are the data fit graphs for OLS estimation with evaluation of the algae population over time (hours). The data fit was evaluated for all four populations of algae and an OLS estimation of the parameters $r$, $K$, and $x_0$, where $x_0$ is the initial algae population, was carried out. Estimates for these parameters for Population 1 are given in Table 1 below along with their respective confidence intervals. Similar results for the other populations can be found in [2]. The corresponding results using proportional error (GLS) estimation are given in Figure 2 and Table 2, with the results for other populations given in [2].
3.2 Parameter Estimates using Constant Error Model

Figure 1: Plots of the constant error based (OLS) logistic model fitted to the data for populations 1, 2, 3, and 4.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>3.67</td>
<td>[2.99, 4.35]</td>
</tr>
<tr>
<td>$K$</td>
<td>4503.18</td>
<td>[4323.72, 4682.64]</td>
</tr>
<tr>
<td>$x_0$</td>
<td>414.74</td>
<td>[234.35, 586.13]</td>
</tr>
</tbody>
</table>

Table 1: Constant error parameter (OLS) estimates for Population 1.
3.3 Parameter Estimates using Proportional Error Model

Figure 2: Plots of the proportional error (GLS) based logistic model fitted to the data for populations 1, 2, 3, and 4.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r$</td>
<td>5.75</td>
<td>[3.75, 7.75]</td>
</tr>
<tr>
<td>$K$</td>
<td>4184.1</td>
<td>[4030.53, 4337.66]</td>
</tr>
<tr>
<td>$x_0$</td>
<td>194.07</td>
<td>[131.55, 256.59]</td>
</tr>
</tbody>
</table>

Table 2: Proportional error (GLS) parameter estimates for Population 1.
3.4 Constant Error Residuals

In this section, the constant error (OLS) model was adapted to the data collected during experimentation. For populations 1, 2, 3, and 4, the residuals are plotted in Figure 5 along the y-axis and compared to model predictions on the x-axis.

![Figure 3: Plots of (OLS) residuals versus the constant error (OLS) model for populations 1, 2, 3, and 4.](image)

We can analyze whether the constant error model is the best fit for evaluating this data by plotting residuals against the model [1]. The constant error model describes the statistical error of the data at $\gamma = 0$. Analyzing patterns in these plots can help reveal the validity (or lack thereof) of assumptions about the variance of the residual errors [4]. In all of the populations, the data shows an outward fanning pattern from left to right, which suggests that constant variance is suspect for this data. This is an indication that the statistical model is wrong, which can lead to incorrect uncertainty conclusions. The goal is to have a model for our data that does not reveal a pattern in the plot of residuals against model predictions. Thus, proportional error is the next model we considered.
3.5 Proportional Error Residuals

We obtained an array of plots of residuals versus the modified statistical model with gamma values varying from 0.8 to 1.2 for each population to determine the best value for the proportional error term. These are given in Figure 4 for Population 2, with the corresponding figures for Populations 1, 3, and 4 given in Figures 6, 8, and 9 of [2], respectively. The plot with the most independent, identically distributed residuals is accepted as the best suited to account for statistical error.

Figure 4: Plots of residuals versus the proportional error model for Population 2. Gamma values from top to bottom are 0.8, 0.9, 1.0, 1.1, and 1.2, respectively.
4 Difference-Based Method for Determining Statistical Model

All of the results presented above depended on first carrying out the inverse problem prior to further analysis with the resulting residuals. These residuals were analyzed to suggest the appropriate form of the statistical model. In this section we further investigate the appropriate form of the statistical model without requiring the prior inverse problem calculations. For this we will use the 2nd order difference approach applied directly to the data realizations \( \{ y_j \} \) for statistical models of the form (2) as developed and tested in [1].

Utilizing pseudo measurements and the experimental data and accounting for the proportional error factor, modified residuals are calculated in hopes of confirming the findings of the analysis of the previous sections to determine the best statistical model using residuals.

Assuming the statistical model \( Y_j \) as described in equation (2), we calculate the pseudo measurement errors \( \hat{\varepsilon}_j \) given by

\[
\hat{\varepsilon}_j = \frac{1}{\sqrt{6}}(y_{j-1} - 2y_j + y_{j+1}), \quad j = 2, ..., n - 1,
\]

(8)

where \( y_j \) is a single measurement (data point) at time \( t_j \). The corresponding modified residuals taking into account the proportional error are calculated by

\[
\frac{\hat{\varepsilon}_j}{|y_j - \hat{\varepsilon}_j|^\gamma}, \quad \gamma = 0, 1, ...
\]

(9)

We may choose \( \gamma \) as any value \( \geq 0 \), but for our report here we will test \( \gamma \) at 0 and 1. Plotting the modified residuals versus time \( t_j \) will help confirm the appropriate value for gamma and thus the correct statistical model.

4.1 Difference-Based Method Residuals

Modified residuals using the pseudo measurement errors \( \hat{\varepsilon}_j \) versus time \( t_j \) are found below for Population 2 and for the other populations in Figures 10-11 of [2]. Residuals were calculated for populations 1 to 4 at gamma values of 0 and 1. (We remark that the residual at \( t_j = 32 \) for \( \gamma = 1 \) was removed due to it being an extreme outlier).

From these results and others in [2], the modified residual plots for \( \gamma = 1 \) appear to have the most desirable distribution, whereas plots for \( \gamma = 0 \) have a non-ideal outward fanning pattern. These results can be considered as a non-inverse problem based confirmation of our findings in Section 3.
5 Discussion

As we can see from the results section, both the constant and proportional error statistical models provide visually reasonable estimation curves for algae growth in a steady-state environment. On initial observation of the logistic model fit to the data, the constant error model appears to closely mirror the data. Comparison of parameter estimate confidence intervals and residual plots of the model and modified model estimates against population data suggests that the proportional error model with $\gamma = 1$ provides the most accurate estimates of algae growth in a steady-state environment. It provided consistent parameter estimates across replicates with significantly tighter 95 percent confidence intervals than did the constant error model. One notable inconsistency occurs at the upper bound of the confidence interval for the growth rate parameter of Population 1 using the proportional error model (see Table 2 and Tables 1-8 of [2]).

With respect to the plots of residuals versus model for the constant error logistic model, we see a characteristic pattern of fanning (Figure 3) as noted above, which indicates the characterization of error is not appropriately accounted for in the statistical model. When plotting the residuals against the modified model used with the proportional error statistical model, there was no pattern of fanning. We assume the absence of a pattern in the modified residuals indicates that most sources of error are accounted for within the statistical model. In order to find the best fit, we plotted residuals vs. the modified model while letting $\gamma$ vary at the following values: 0.8, 0.9, 1, 1.1, and 1.2 (Figures 6-9 of [2]). When studying the patterns at each proportional error exponent, $\gamma$, we conclude that setting $\gamma = 1$ most accurately accounts for statistical error in the observations for the logistic growth mathematical model. These findings are strongly supported by the difference-based results presented in Section 4.

Based on our finding, we suggest that the population growth of *Raphidocelis subcapitata* in a controlled chemostat ecosystem is best and most accurately modeled using a proportional error statistical model with the logistic growth mathematical model. It is important to recognize that our findings are limited to a tightly regulated laboratory environment with constant temperature, fixed photoperiod, and steady nutrient flow and thus should not be unduly applied to lake populations. In order to form a more representative model of population dynamics in the field, we must consider fluctuations in temperature, nutrient
availability, and photoperiod as well as predation by zooplankton, competition with other phytoplankton, and abiotic factors. Nonetheless, this study is another step towards developing a more comprehensive mathematical model.

Acknowledgements

This research was supported in part by the National Science Foundation under NSF Undergraduate Biomathematics grant number DBI-1129214, NSF grant number DMS-0946431, NSF grant number DMS-1514929 and in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-15-1-0298. The authors are grateful to Robert Baraldi and Brittany McLawhorn for their assistance in collection of the data used in this report.

References


