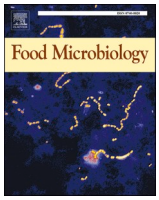
Food Microbiology 128 (2025) 104726

Contents lists available at ScienceDirect

Food Microbiology

journal homepage: www.elsevier.com/locate/fm



Variability of growth parameter estimates - The role of rescaling and reparametrization

Maha Rockaya, József Baranyi *

Doctoral School of Food Science and Nutrition, University of Debrecen, Hungary

ARTICLE INFO

Keywords:
Predictive microbiology
Regression
Rescaling
Reparametrization
Monte Carlo simulation
Variability

ABSTRACT

The focus of this paper is to analyze the reliability of the error estimation when well-known primary models of predictive microbiology are used to fit growth parameters. We also demonstrate the application of rescaling and reparametrization to improve this reliability. We highlight that the technique can be useful for achieving linearity and homoscedasticity, reducing the complexity of the model, generating initial parameter estimates when fitting experimental data by non-linear regression, and obtaining realistic standard errors for the parameter estimates, which are crucial for decision-making in food safety.

We classify the sources of the total variability and correlation of the parameter estimates as "wet" and "dry". We point out that, rescaling and reparametrization do not change the model in a mechanistic sense but they can reduce the variances of (and/or the correlation between) the parameter estimates, thus mitigate the effects of such “dry” (i.e. statistical) relationships.

We analyze the reliability of the error estimation when the model of Baranyi and Roberts (BRM) and the Gompertz function (GF) are used to fit data. The comparison is based on the distribution of the standard error of the maximum specific growth rate estimates. The results show that the error structure of the BRM-fit is closer to that of the linear regression, making BRM more reliable for constructing confidence intervals by conventional means, using the t-distribution assumption for the parameter estimates.

1. Introduction

Predictive microbiology focuses on modelling microbial dynamics in various environments typically related to food. Its goal is to quantify and model observations on microbial kinetics, enabling reliable predictions and quantifying the confidence in them. In this process, transforming the variables (rescaling) and the parameters (reparametrization) of the used mathematical models play a crucial role (Baranyi, 1992).

In regression analysis, the function to be fitted is frequently called “model”, even if the function itself has no physical/biological interpretation. In contrast, we use the term "model" in its mechanistic sense. In our concept, a model consists of definitions and assumptions, as well as some (possibly differential) equations, preferably grounded in fundamental science. Defining variables, their domains of interpretation and value sets, together with other conditions (e.g., certain constraints on the variables and/or the parameters of the model) is as integral to the modelling process as constructing its core equations. Referring to an equation alone as a “mathematical model” assumes that its interpretation and conditions are already known and understood.

When describing the temporal changes in a genetically homogeneous bacterial population, the logarithmic transformation is often applied to the population size. The primary reason is that, under typical conditions, bacteria proliferate exponentially, which appears linear when plotting the logarithm of cell concentration against time. Another reason is that cell concentrations are typically measured using sampling methods that include the dilution of the bacterial culture to a low level, producing easy-to-count colonies on agar, each originating from a single cell. The original cell concentration can then be estimated by means of the used dilution factor (Corry et al., 2007). This method ensures that the relative error of the estimated cell concentration remains constant, regardless of the magnitude of the cell concentration itself. This is a well-known reason why the logarithm- (a so-called link-) function is fequently applied to bacterial concentration. Such logarithmic rescaling is common in biochemistry; for example, see the definition of pH, which is also a logarithmically rescaled measure of hydrogen ion concentration.

Primary models describe the temporal change of cell concentration in a constant environment. The simplest primary model of bacterial growth assumes a purely exponential growth of cell concentration, whose

* Corresponding author.
E-mail address: baranyi.jozsef@med.unideb.hu (J. Baranyi).

https://doi.org/10.1016/j.fm.2025.104726
Received 11 October 2024; Received in revised form 5 January 2025; Accepted 7 January 2025
Available online 8 January 2025

logarithm exhibits a linear trend over time. The model, in its canonical form, is linear also in its parameters, which are (assuming that the link-function is the *natural* logarithm) the maximum specific growth rate and the natural logarithm of the initial cell concentration. These parameters can be estimated from the measured logcount values, using linear regression, which has well-explored mathematical and statistical properties.

In this paper, we investigate, which of two widely used sigmoid primary models is more reliable in terms of the variability of the parameter estimates, obtained via fitting simulated logcount data. As the properties of linear regression is well-analysed by means of t- and F-distributions (see standard textbooks on regression like Nau, 2016), and the error-estimation in non-linear regression is carried out by locally-linear approximations, our task is equivalent to establishing which of the two primary models presents error-distributions for the parameter estimates with properties closer to those generated by linear regression, the gold standard in regression analysis.

The two primary models are that of Baranyi and Roberts (1994), referred as BRM in what follows, and the reparametrized Gompertz Function, GF (Zwietering et al., 1990). The distribution of their parameter estimates (primarily that of the maximum specific growth rate) will be the gauge, regressing which model is closer to linear regression. We refer to GF as a "function" because, when the logarithm of the population size (and not the size itself) is described by the Gompertz equation, its connection to the original Gompertz model is lost: the Gompertz model had been developed for the population size and not for its logarithm. Despite this, the Gompertz function is useful for representing a measured growth curve with a smooth sigmoid function, even if that has no mechanistic interpretation. In contrast, BRM builds on the fundamental Linear Model, valid for the exponential phase, but with refinements for the preceding and following transition phases, derived from well-established models of population dynamics and biochemistry.

First, we clarify the concept of maximum specific growth rate, then summarize the discussed primary models, with their usual parameterization (Pérez-Rodríguez and Valero, 2013).

Specific growth rate as a slope. Let $x(t)$ denote the bacterial cell concentration at the time t, where the time zero is that of the inoculation, and rescale it by the natural logarithm transformation: $y(t) = \mathrm{Ln}\, x(t)$. Consider the

$$\frac{\Delta Ln(x)}{\Delta t} \approx \frac{\Delta x / x}{\Delta t} \tag{1}$$

relationship. As $\Delta t \to 0$, the above expression converges to the slope of the $y(t)$ curve, $\mu(t)$, called the (instantaneous) specific growth rate at the time t. One of its possible interpretations is that, for a small Δt time interval, the quantity $P(t,\Delta t) = \mu(t)\cdot\Delta t$ is approximately the probability of division in the $[t,\ t+\Delta t]$ interval. Another interpretation is that, depending on the distribution of the cells' individual generation time, $\mu(t)$ is a sort-of (not necessarily arithmetical) average of the number of divisions for single cells in a unit time. In a constant growth environment, $\mu(t)$ reaches a maximum, μ_{max}, called the maximum specific growth rate. It represents an inherent biological link between the organism and the environment.

Linear Model (LM) *and linear regression*. The simplest primary model describes the scenario when a homogeneous bacterial population is inoculated in a growth environment and start to proliferate immediately, at the μ_{max} specific growth rate. Then, for the $x(t)$ concentration of the cells:

$$\frac{dx}{x\,dt} = \frac{d\,\mathrm{Ln}(x)}{dt} = \frac{dy}{dt} = \mu_{max} \tag{2}$$

with $y(t) = \mathrm{Ln}\, x(t)$ and $y(0) = \mathrm{Ln}\, x(0) = \mathrm{Ln}\, x_0 = y_0$.

The algebraic solution of this differential equation for the $y(t)$ natural logarithm of the cell concentration is

$$y(t) = \mu_{max}\cdot t + y_0 \tag{3}$$

where y_0 is the natural logarithm of the initial cell concentration, i.e. the value of $y(t)$ at $t=0$.

Not only that the natural logarithm of the cell concentrations is a linear function of time, but it is linear in terms of the μ_{max} and y_0 parameters, too, so they can be estimated by linear regression. As mentioned in the Introduction, the Linear Model is a starting point for any growth model and the linear regression is a reference point for any cutve fitting.

The model of Baranyi and Roberts (BRM). In a finite space, the closer the bacterial concentration to the x_{max} carrying capacity of the environment, the more the specific growth rate is reduced. The logistic model (Turner et al., 1976) uses the inhibition factor $u(x)=(1-x/x_{max})$ to modulate the maximum specific growth rate. This $u(x)$ is between 0 and 1, and monotone converges to 1 as $x(t)$ increases. Apply the natural logarithm to the cell concentration, so the inhibition factor becomes $u(y) = 1-\exp(y-y_{max})$ and the model for the $y(t) = \mathrm{Ln}\, x(t)$ rescaled response variable becomes:

$$\frac{dy}{dt} = \mu_{max}(1 - e^{y-y_{max}}) \tag{4}$$

where $y_{max} = \mathrm{Ln}\, x_{max}$, and $y(0) = y_0 < y_{max}$.

This is the logistic growth model which is known to be sigmoidal for the $x(t)$ cell concentration (but not for its logarithm).

As the pre-inoculation environment was different from the new one, the cells undergo a transition before the exponential phase. Baranyi and Roberts (1994) introduced the $\alpha(t) = q(t)/(1+q(t))$ inhibition factor, where $q(t)$ increases exponentially from a q_0 initial value. So $\alpha(t)$ is always between 0 and 1, converging to 1 as $q(t)$ increases. Therefore, the specific growth rate, i.e. the slope of $y(t)$, converges to μ_{max} as the population increases. See the Glossary table below for explanations for all notations.

$$\frac{dy}{dt} = \frac{q(t)}{1+q(t)}\,\mu_{max} \tag{5}$$

Combining the two inhibition factors we arrive to the model of Baranyi and Roberts (1994):

$$\frac{dy}{dt} = \frac{q(t)}{1+q(t)}\,\mu_{max}(1 - e^{y-y_{max}}) \tag{6a}$$

$$\frac{dq}{dt} = \mu_{max} q \tag{6b}$$

The model has algebraic solution, which makes data fitting easier:

$$y(t) = y_0 + \mu_{max}A(t) - \mathrm{Ln}\left(1 + \frac{\exp(\mu_{max}\cdot A(t)) - 1}{\exp(y_{max} - y_0)}\right) \tag{7a}$$

where

$$A(t) = t - \lambda + \frac{1}{\mu_{max}}\mathrm{Ln}(1 - \exp(-\mu_{max}\cdot t) + \exp(-\mu_{max}\cdot(t-\lambda))) \tag{7b}$$

The Gompertz model and the reparametrized Gompertz Function (GF). The Gompertz model (Turner et al., 1976) states that, for $y(t)$, the logarithm of a growing population

$$\frac{dy}{dt} = -c\cdot(y - y_{max}) \tag{8}$$

where $y(0) < y_{max}$

This, just as its akin, the logistic model (4), is not sigmoidal for $y(t)$, as its slope is monotone decreasing as the population increases. However, it does give a sigmoid function for $x(t)$. Zwietering et al. (1990) reparametrized the equation in such a way that the maximum specific growth rate and the lag became the new parameters:

$$y(t) - y_0 = A \cdot \exp\left(- \exp\left(\frac{\mu_{max}}{A} \cdot e \cdot (\lambda - t) + 1\right)\right) \quad (9a)$$

with the interpretation

$$A = y_{max} - y_0 \quad (9b)$$

This 3-parameter equation only describes the increase of $y(t)$ from its initial level. If an estimate for this inoculum is also needed, then the 4-parameter version should be used:

$$y(t) = y_0 + (y_{max} - y_0) \cdot \exp\left(- \exp\left(\frac{\mu_{max}}{y_{max} - y_0} \cdot e \cdot (\lambda - t) + 1\right)\right) \quad (10)$$

with parameter interpretations analogous to those of Equation (7)

This function has been used to regress bacterial growth curves since the 80-s. Of course, via the process, we lost the original interpretation of the Gompertz model. However, Equation (10) is still useful for curve fitting, just it should not be called (modified) Gompertz model, rather a (reparametrized) Gompertz function or Gompertz equation.

Standard error of parameter. When regressing measured data, we estimate the parameters of the fitted model by some numerical method. Part of this method is a quantification of the confidence in the parameter estimates, which is most frequently expressed by their standard error. This is, however, another estimation from the same data, which depends on the regression algorithm. It is a feasible scenario that a model accurately describes the biological system but the test points to be fitted are poorly chosen or/and the regression algorithm is not robust enough, and the standard errors of the parameter estimates become unrealistic. The opposite of this is, when the regression is robust, such as the linear regression, but it does not reflect reality with sufficient accuracy; for example, when sigmoid growth curves are fitted by LM. This is why it is reasonable to ask, whether a non-linear regression, in some sense, is close to linear regression. This is the basic question we ask here for the BRM- and GF-regression. Each of them will be evaluated based on how similar the distribution of the standard error of the maximum specific growth rate estimates is to what we would expect from linear regression.

In most predictive microbiology studies, the 95% confidence interval for a $\widehat{p}$ estimate is often approximated by $\widehat{p} \pm 2\mathrm{se}(\widehat{p})$, where $\mathrm{se}(\widehat{p})$ is its standard error. This approach assumes that the t-score of the estimate $\widehat{p}/\mathrm{se}(\widehat{p})$, follows a distribution close to normal. However, the normality assumption often does not hold for non-linear regression. Suitable reparametrization can result in new parameters that are, at least approximately, normally distributed, as we will see in this paper.

When assessing confidence in estimates and predictions, the total error arises from various sources. Significant effort has been made (Corry et al., 2007; Akkermans et al., 2018; Nauta, 2000) to define and categorize these sources. Here, we chose to divide them into two broad categories only.

1 Wet Source: Like the term "*wet science*," this category encompasses all factors contributing to variability and uncertainty in observations related to biology and/or the conditions under which the observations were made. Examples include strain- or cell-to-cell variability, environmental randomness, the accuracy of laboratory methods or equipment, and even the expertise of the laboratory operator.
2 Dry Source: This category covers all factors that introduce errors or uncertainties in the algorithms used for data cleaning, inference and estimations, i.e. the entire process from data recording to processing and numerical calculations/approximations.

When fitting model parameters, part of the regression output is the standard error of the parameters, which are obviously from "dry" sources. The square of the standard error is the variance that can be seen as a special case of covariance, which leads to the idea that correlation between parameters also has "*dry*" and "*wet*" sources. As a shorthand, we use the terms "dry" and "wet" variability and correlation, depending which class of sources caused them. Textbooks on regression routinely discuss the variance and correlation for the estimates of the slope and the intercept in case of linear regression. These are examples for dry variability and correlation. It is also known that the correlation can be made zero, if the measured points for the explanatory and response variables are shifted in such a way that the centre of gravity of their measured values coincides with the origin of the coordinate system. For a detailed analyses, see Nau (2016). Generally, rescaling and reparametrization can only be used to decrease the *dry* (but not the wet) variability and correlation.

In this paper, we only focus on dry variabilities and correlation, more specifically in case of fitting logcount data using the above three primary models. We assess the distribution of the parameter estimates and draw conclusions about the reliability of the error estimations resulted from fitting the models via the standard Least Squares method.

2. Material and methods

Consider the linear model (3) with measurement error: $y(t) = \mu_{max} \cdot t + y_0 + \varepsilon$, where μ_{max} (the slope) and y_0 (the intercept with the y axis) are the parameters of the model and ε is a stochastic additive term assumed to have zero mean and σ standard deviation independently of t and y. This is a well-analysed one-variable example for unweighted linear regression (Nau, 2016).

Bacterial concentrations measured by the traditional plate-count method follow a lognormal distribution. This is because high concentrations are diluted and plated so that the number of colonies counted falls within the same order of magnitude (approximately 150–300, considering practical constraints; see Corry et al., 2007). As a result, the relative error of the concentration estimate is independent of both time and bacterial concentration. Therefore, when fitting a primary model to the y_i values (i.e. the natural logarithm of the cell concentrations at the times t_i), the residuals are expected to be independent, identically, and normally distributed, just as we need it for data fitting.

Below we summarize some properties of the above linear regression (Nau R. 2016). We will need these when comparing the regression output of BRM- and GF-fitting to that of LM.

When fitting the linear model, $\mu_{max} \cdot t + y_0$, to a (t_i, y_i) set of independent observations ($i = 1 \ldots N$), using the Least Squares criterion, the parameter estimators $\widehat{y_0}$ and $\widehat{\mu_{max}}$ are explicit expressions of the t_i, y_i values. The standard error of fit provides an estimate for the σ standard deviation of the stochastic measurement error ε. The respective ratios between the two parameter estimates, $\widehat{y_0}/\mathrm{se}(\widehat{y_0})$ and $\widehat{\mu_{max}}/\mathrm{se}(\widehat{\mu_{max}})$ (the reciprocal of the relative errors of the parameter estimates) follow Student's t-distribution. This allows calculating confidence intervals for the estimates.

The standard error of the slope-estimate will be proportional to the standard error of the fit:

$\mathrm{se}(\widehat{\mu_{max}}) = \mathrm{se}_{\mathrm{fit}} / C$ (for a detailed proof see Nau, 2016),

where C is a constant, depending solely on the sampling times.

As the relative error of the slope estimate is constant, it is the *logarithm* link-function that should stabilize the variance of the $\widehat{\mu_{max}}$ estimator (Akkermans et al., 2018). This implies that, if the main source of variability is the measurement error in the logcount data, then the logarithm of the μ_{max} estimates should play the role of observed data to be modelled as a function of environmental factors (secondary model).

Fitting a reparametrized version of the linear model may result in a non-linear regression. In the example above, the linear model is presented in its canonical form, $y(t) = \mu_{max} \cdot t + y_0$ where the response variable is linear also in the parameters $\mu_{\max}$ and y_0. This linearity of the model in the parameters (and not that the model is linear in x) is the very reason why the Least Squares method yields explicit formulae for $\widehat{\mu_{max}}$ and $\widehat{y_0}$.

Consider a rearrangement of the linear model: $y = \mu_{max}(t - t_0)$, where $t_0 = -y_0/\mu_{max}$ is the intercept with the horizontal t axis. The model itself does not change, only its parameters are recombined. Fitting this

reparametrized linear model using the Least Squares method results in non-linear regression, which typically requires an iterative procedure. In this context, Student's t-distribution is only approximately valid for $\widehat{t_0}$, the estimator for t_0. The "more non-linear" the regression, the less applicable the t-test becomes to error analysis. Nevertheless, the transformation can be useful if t_0 offers biological interpretation, which also makes it easier to provide initial parameter estimates needed for non-linear fitting.

Linear rescaling of variables. Typically, also non-linear regression is carried out by minimizing the sum of squares of the residuals, using an iterative algorithm. This has straightforward consequences to the case when the variables are rescaled by a proportionality constant. We demonstrate this below by changing the units of the axes of primary models:

1. Rescale the time unit from hours to, say, days. The sampling points then become: $t_i' = t_i/24$. The standard error of fit, se_{fit}, remains unchanged since it is measured on the vertical axis. However, the estimate for the maximum slope and its standard error are both altered by the factor 24. Therefore, the *relative* error of the slope estimate (the reciprocal of the t-score) does not change.
2. Rescale the logcount observations from decimal to, say, natural logarithm. This will change both the standard error of fit and the standard error of the slope by the factor Ln(10)≈2.3. The *relative error* of the slope, again, does not change.

It can be readily seen that, if the response variable depends on the parameter in a non-linear manner, then the relative error of the slope estimate does change.

Monte-Carlo simulation. The above described three models were used for regression in our simulation study: the Linear Model (LM, Equation (3)); the model of Baranyi and Roberts (1994) (BRM, Equation (7)) and the reparametrized Gompertz function (GF, Equation (10)).

Seven sampling times, t_i (i = 1…7) were considered (0h, 12h, 24h, 36h, 48h, 60h, 72h) and 1000 Monte Carlo simulations were performed for each model, with the following parameters: $\mu_{max} = 0.5(h^{-1})$, $y_0 = 5$ (Ln cell/ml) and (for BRM and GF only) $\lambda = 8$(h), $y_{max} = 23$(Ln cell/ml). The y_i response values obtained for each model were perturbed by using normally distributed random errors (Gaussian noise) with zero mean and standard deviation $\sigma = 1$, for each y_i obtained at t_i. The perturbed y_i values were then fitted (i) by the same equation that generated them and (ii) by the other sigmoid curve, to estimate the growth parameters and their associated standard errors.

Our focus was not only the histograms obtained for the parameter estimates but also that of the relative errors of the parameter estimates. We also examined the correlation between the estimates.

Computation. We used Microsoft Excel, its statistical functions and Add-Ins, for simulation and linear regression. An in-house VBA (Visual Basic for Applications) package was used for non-linear regression that implemented the Levenberg – Marquardt algorithm (Press et al., 2007).

3. Results

Sigmoid curve fitting. It can be readily seen that the geometry of a $y(t)$ sigmoid curve is, by and large, defined by the $[y_0, y_{max}]$ band and the tangent of the $y(t)$ curve drawn to its t_f inflexion point. Therefore, if the slope of that tangent is a parameter of the sigmoid function used for fitting (as in the case of the GF; see Equation (10)), then a linear rescaling of the variables proportionally changes that parameter estimate as well as its standard error. Consequently, it is indifferent whether (i) GF is fitted to the natural logarithm of the cell concentrations or (ii) to their decimal logarithm, and the obtained rate estimate is multiplied by Ln(10)≈2.3 to get the slope on the natural log scale.

This equivalence does not hold for the other sigmoid curve, BRM, whose μ_{max} parameter is not exactly the maximum slope of the sigmoid function. Instead, it represents a "potential" maximum specific growth rate that the organism would achieve if it was inoculated without any change in the environment (i.e. no lag phase) and at sufficiently low initial concentration, so the proximity of the y_{max} carrying capacity does not reduce the exponential rate at the onset. As can be seen from the model construction (see also Fig. 5), the slope of the curve at the t_f inflexion is

$$\mu(t_f) = \alpha(t_f) \cdot \mu_{max} \cdot \left(1 - exp\left(y(t_f) - y_{max}\right)\right) \quad (11)$$

So, μ_{max} is the *potential* and not the *observed* maximum slope (which is $\mu(t_f)$) of the $y(t)$ curve. Table 1 demonstrates that these two algorithms do not result in the same estimate for μ_{max}:

(i) Formula (7) is used to fit the *natural logarithm* of the cell concentration
(ii) Formula (7) is used to fit the *decimal logarithm* of the concentrations, then the obtained rate is multiplied by the Ln(10)≈2.3 factor.

In the second case, the result can be noticeably different from that of the correct first one. The example curve in Table 1 is from Buchanan and Klawitter (1991), which is recorded under ID = ELC0374 in the ComBase database (www.combase.cc).

The reason for the discrepancy goes back to the *u(x)* inhibition function in formula (4). If the decimal logarithm of x is put in the place of y, as in the method (ii) above, then 10^y, not exp(y) should be used in Equation (7a). The GF-regression does not have this problem as there the dimension of the argument of the exponential function is time, not logcount.

In our experience, the discrepancy is generally not too big (<5%), if there are many observations in the exponential phase. However, if only a few points have been measured there, and the inoculum level is not very low, then the second method can overestimate the maximum specific growth rate by as much as 20–50%.

Error distribution of the parameter estimates. When fitting the $y_1 \ldots y_N$ data by an $f(t;\mathbf{p})$ growth function, the regression is close to linear, if the f function is close to linear in the components of the $\mathbf{p}$ vector of parameters. Because of its origin, regression by BRM should be closer to linear regression. That indeed, this is the case, we will illustrate by the simulation data described in Materials and Method.

Fig. 2 shows that, based on properties like the symmetry of the histograms produced by the $\widehat{\mu_{max}}$ and $re(\widehat{\mu_{max}}) = se(\widehat{\mu_{max}})/\widehat{\mu_{max}}$ estimates, the gold standard linear regression is much closer to the BRM-than to the GF-regression. Note that the horizontal axes of the respective histograms in Fig. 2 are intentionally not on the same scale. The purpose of the figure is to compare the shape of the distributions of the estimated parameters, rather than simply evaluating the goodness of fit. Specifically, we aim to investigate how closely these distributions align with those characterizing the linear regression. This comparison provides insights into the reliability of the standard error and confidence interval

Table 1
Fitting the model of Baranyi and Roberts (1994), BRM, to the growth curve with ID = ELC0374 in ComBase. First, $y_i = Ln(x_i)$ then the $\log_{10}(x_i)$ values were fitted and, in the latter case, the results were put back to the natural log scale (see Fig. 1).

Method to fit BRM	$\widehat{\mu_{max}}$ (h^{-1})	se ($\widehat{\mu_{max}}$)	$\widehat{\lambda}$ (h)	se ($\widehat{\lambda}$)	$\widehat{y_{max}}$	se ($\widehat{y_{max}}$)
Eq. (7) fitted to $y_i = Ln(x_i)$	0.442	0.0703	21.9	4.70	4.79	22.9
Eq. (7) fitted to $y_i = \log_{10}(x_i)$, then the appropriate parameter estimates are multiplied by Ln (10)≈2.3	0.542	0.0942	26.5	4.16	4.98	22.9

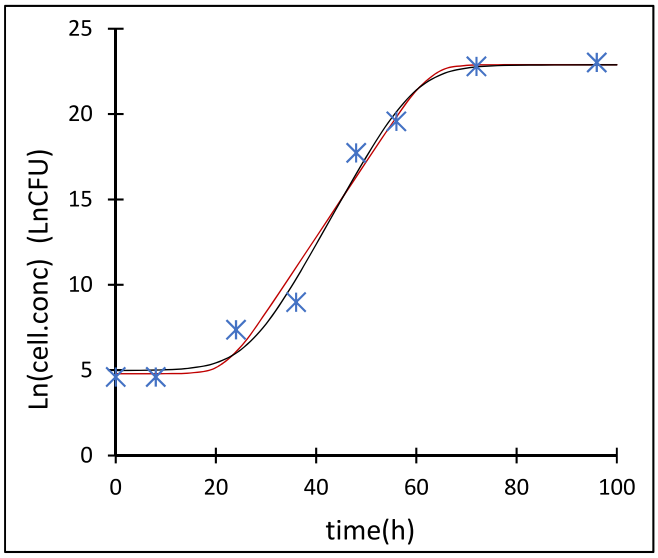


Fig. 1. Results of the regression described in Table 1. Blue stars: logcounts measurements. Red continuous line: fitting BRM to the natural logarithm of the cell concentrations. Black continuous line: fitting BRM to the decimal logarithm of the cell concentrations and the estimated rate parameter is converted back to natural logarithm scale via multiplying the rate by Ln(10) ≈ 2.3. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

estimations for the parameters, which are routinely carried out using the t-distribution assumption. Such analysis helps decide, which of the models offers more adequate quantification of uncertainty, which is crucial for making decisions based on these estimates. Our focus is on the precision of the parameter estimation, rather than to see which model fits the data better. In other words, our study compares two (BRM- and GF-) *regression algorithms*, regardless of which model describes the biological growth better. To show this, we also BRM-fitted the datasets simulated by the GF function and vice versa. The result was the same: the error estimation produced by BRM-regression is more reliable than the one produced by GF. This means that the conventional confidence interval estimation is more accurate in the BRM than in the GF case.

Another known property of the linear regression is that the estimate of the intercept with the vertical axis linearly correlates with the slope estimate (Nau, 2016). Therefore, we reparametrized the two sigmoid functions, BRM and GF (Equations (7) and (10)), by $h_0 = \mu_{max}\lambda$, to get the intercepts for the sigmoid models, too, and to see the correlation between the $\widehat{\mu_{max}}$ and $\widehat{h_0}$ estimates for both cases. Note that, apart from such numerical purposes, h_0 also has physiological significance, as it quantifies the "work to be done" by the cells during the lag phase (Rolfe et al., 2012). This parameter has been shown to have no correlation with the temperature (therefore neither with the maximum specific growth rate) in the normal physiological growth region. The parameter h_0 reflects certain aspects of inoculation procedure (for example inoculating from early or late stationary phase?), just as the inoculum level does. Taking it as a temperature-independent constant represents the assumption that, at higher and lower temperatures, the same physiological processes happen (the same work is to be done) during the lag phase, just at higher and lower rates.

Fig. 3 shows that the "dry" correlation between the two parameter estimates, $\widehat{\mu_{max}}$ and $\widehat{h_0}$, for both BRM and GF, are strongly positive (unlike their wet correlation, which is close to zero as said above). The resultant R^2 indicator was >0.7 for BRM, but was 0.61 for GF, confirming that the statistical properties of BRM-regression are closer to those of the linear regression.

4. Discussion

Predictive microbiology is based on mathematical models that describe the key characteristics of microbial responses to their environment. Estimating the parameters of the model should always be accompanied by quantifying their uncertainty (e.g., confidence intervals, standard errors), while taking into account that these are also estimations. The more robust and reliable the error estimation, the more it can improve decision-making.

This study compared the statistical properties of error estimates obtained by fitting two sigmoid functions. We wanted to know how closely they align with those obtainable by linear regression. We demonstrated that the latter one is closer to the BRM-fit than to the GF-fit, regardless of which model describes bacterial growth better. This is because BRM is a refinement of LM, the Linear Model. We also pointed out that "dry variability" and "dry" correlation can be improved by rescaling and reparametrization.

A point in the case overlooked in the past, is that BRM should be used for the natural, not the decimal logarithm of the bacterial concentration, otherwise it becomes an empirical curve fitting, like GF, and the parameter estimation can be distorted. We saw an example for this in Fig. 1.

When estimating parameters from observations, it is essential to verify whether the estimates align with existing knowledge. A quantified assessment of confidence in the estimates can be achieved via standard errors or confidence intervals. While a small standard error of a parameter estimate may seem reassuring, it can be misleading if not properly founded, even if the parameter estimate itself is accurate. Reparametrization and rescaling can bring the regression closer to linear (Figs. 2–3), making the confidence interval estimation more reliable. However, these transformations do not change the model itself but can change the estimates of the model parameters, their standard errors, and the correlations between them.

It is worth noting that the technique of rescaling played an important role during the development of BRM already. Specifically, the mechanistic assumption behind the model was that, during the lag phase, cells need to produce a key metabolic product, $P(t)$, such as a new enzyme, required for the new environment. This product influences the specific growth rate according to Michaelis-Menten inhibition, a concept well-known in enzyme kinetics (Srinivasan, 2022):

$$\frac{dy}{dt} = \frac{P(t)}{K_P + P(t)}\mu_{max} \tag{12}$$

It is evident that the effect of this inhibition depends only on the ratio $q(t) = P(t)/K_P$, which naturally suggests rescaling the variable $P(t)$ via dividing it by K_P. This transformation led also to the reparametrization $\alpha_0 = q_0/(1+q_0)$ where $q_0 = P(0)/K_P$ represents a dimensionless initial value that characterizes the initial physiological state: how much suited the cells are to the new environment (see the Glossary table). This formulation reduced the number of parameters by one. Furthermore, Baranyi and Roberts (1994) showed that the parameters α_0 and $h_0 = -\ln(\alpha_0)$ have biological interpretations, too.

Another assumption of BRM was that $P(t)$ is produced at an exponential rate, at the same μ_{max} specific rate, at which the cells grow in their exponential growth phase. This assumption further reduces the model's complexity while maintaining a biologically reasonable basis for the used simplification.

One advantage of linear regression is that it consistently provides unique and robust parameter estimates. In this study, we demonstrated that fitting GF is less robust than fitting BRM regardless of which model describes the biological process better. In case of GF-fits, the maximum specific growth rate was way overestimated in 10 out of 1000 cases (for these fits the real relative error of $\widehat{\mu_{max}}$ was more than 100%; i.e. $\widehat{\mu_{max}}$ >1.0 h^{-1} while the data were simulated with μ_{max} = 0.5 h^{-1}). In our simulation, the logcount data had (intentionally) relatively big error,

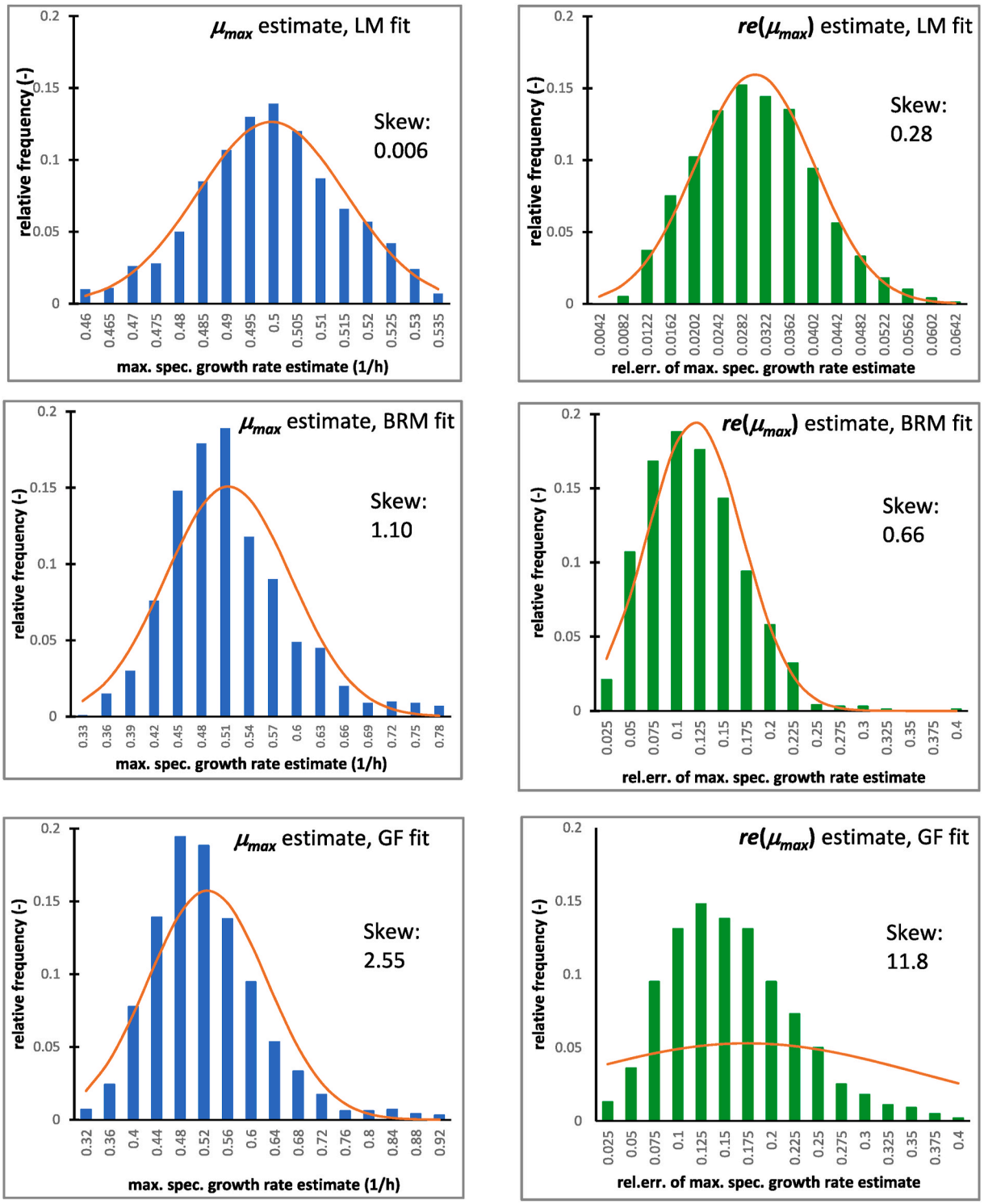


Fig. 2. Histograms for the estimates of the maximum spec. growth rates (blue columns) and their relative errors (green columns) in case of LM, BRM and GF functions. The red continuous lines represent the normal distribution that has the same mean and standard deviation that the respective histogram does. The fitted raw data were 1000 growth curves, one of which is shown in Fig. 4. Each set consisted of N = 7 simulated logcounts generated by the same model that was used for fitting. The simulation parameters were identical for the three models including those of the Gaussian noise. The closer the skewness of a histogram to zero, the more symmetric the distribution. The distribution of the LM-fit turned out to be closer to BRM-fit than to that of the GF-fit. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

this is why outlier bad estimates were obtained in about 1% of the GF-fits. No such "relative error >100%" situation was recorded with the BRM-fit.

The poor quality was caused by the limited number of data points during the highly influential exponential phase. If just one point had an unusually large error, that distorted the GF-fit more than the BRM-fit. Fig. 4 provides an example for this. When BRM was used for fitting, no such bad estimation occurred (even if the raw data was generated by GF), and a unique set of parameter estimates was consistently obtained, for any reasonable initial values for the regression. On the other hand, it

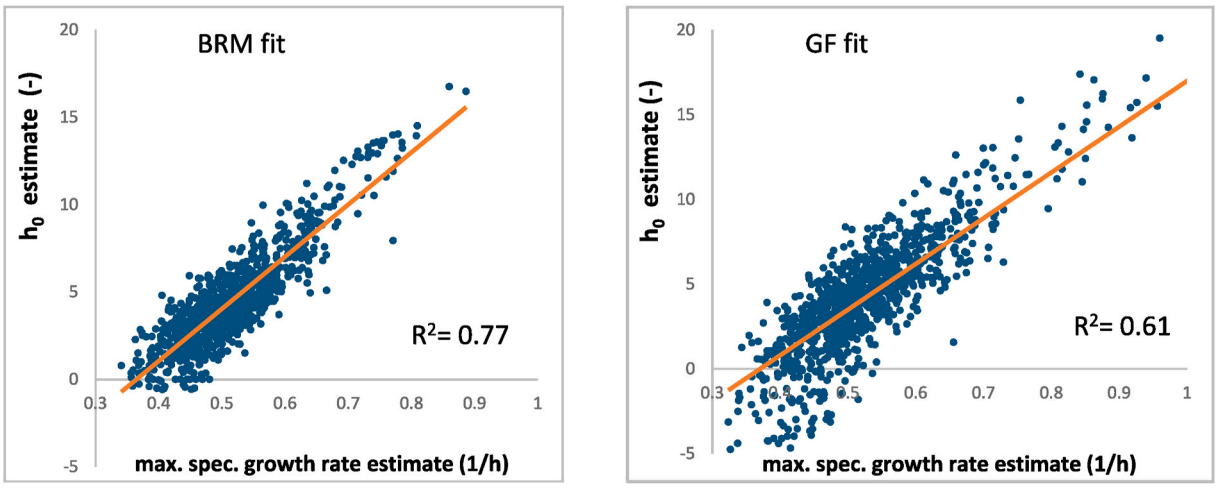


Fig. 3. "Dry" correlation between the maximum specific growth rate and the "work to be done" parameter, i.e. the $\widehat{\mu_{max}}$ and $\widehat{h_0}$ estimates, when fitting the simulated data, used for Fig. 2, by BRM and GF. The relationship is closer to linear for the BRM fit.

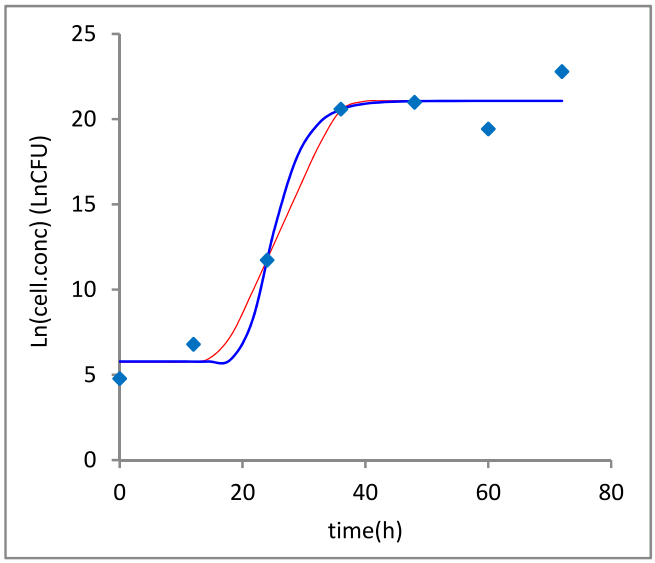


Fig. 4. An example for a GF-simulated data set (blue diamonds) fitted by GF (blue thick continuous line) and BRM (red thin continuous line). The GF-estimate of the maximum specific growth rate is more than double of that produced by the BRM regression, and more than three times greater than the true value used to generate the data. The asymmetry seen in Fig. 2 with the GF regression is caused by similar outliers, which constitute approximately 1% of the 1000 simulated raw datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

can be tested that, if accurately measured bacterial concentrations are all in the exponential phase, except the inoculum and one stationary phase point, then GF would way overestimate the maximum specific growth rate unlike BRM would.

It is noteworthy that, just as μ_{max} and h_0 (see Fig. 3), the estimates of the specific growth rate and lag time also showed a strong positive correlation. This is unrelated to their biological ("wet") relationship. Biologically, we would expect the opposite trend: as temperature increases (but remains sub-optimal), the lag phase would shorten, and the maximum specific growth rate would increase, resulting in a negative correlation. The first correlation refers to the estimates, while the second refers to the real biological parameters at different temperatures.

Finally, a property of the lag parameter estimates provide even stronger support for using the BRM for regression. Namely, one important aspect of the lag estimate is that it can be negative, too. While this may sound controversial, in fact it does make sense for both GF and BRM as shown by Fig. 5. Remember that, for BRM, this means that the initial value of the product $P(0)$ is greater than K_P, so the time point when $P(t)$ reached K_P (or equivalently, when $q(t)$ reaches 1, or $\alpha(t)$ reaches 0.5) was in the past compared to the time of inoculation. GF does not have a similar mechanistic derivation, as the lag is defined via the geometric properties of the sigmoid function (Zwietering, 1990). Additionally, in GF, the y_0 parameter is not the actual $y(0)$ value but the lower asymptote of the fitted sigmoid function. Typically, this should be close to the $y(0)$ level, but not necessarily so, especially when the lag parameter is negative as shown in Fig. 5.

5. Conclusions and recommendations

Rescaling and reparametrization are important to (i) to reduce the complexity of the model (i.e. reducing the number of parameters or/and simplifying its formulae); (ii) to make the new parameters easier to interpret; and (iii) to make the estimation procedure closer to linear regression.

When rescaling a variable or reparametrizing a model, the equations must undergo invertible transformations; otherwise, it is no longer the same model. For instance, when the Gompertz equation, designed by its inventor to describe population dynamics, is applied to the logarithm of that population, the resulting model becomes purely empirical, essentially a black-box model with no connection to its original biological interpretation.

To develop equations for predictive models, one should prefer starting first with a well-established basic model rooted in fundamental science, containing well-analysed mathematical relationships at its core; then refine it with extra assumptions. This way, the simpler model can be nested in the more complex one, like LM is a special case of BRM. This is a useful property of the model, as it can be decided, for example by an F-test, whether the refinement is needed at all when fitting a particular dataset.

Predictive microbiology still needs more mathematical rigour. Precise use of terms and concepts (like defining the lag phase, see Fig. 5), is one of the indicators of progress. Clarifying the purpose and methods behind rescaling and reparametrization techniques is part of the effort needed to advance this still relatively young scientific discipline.

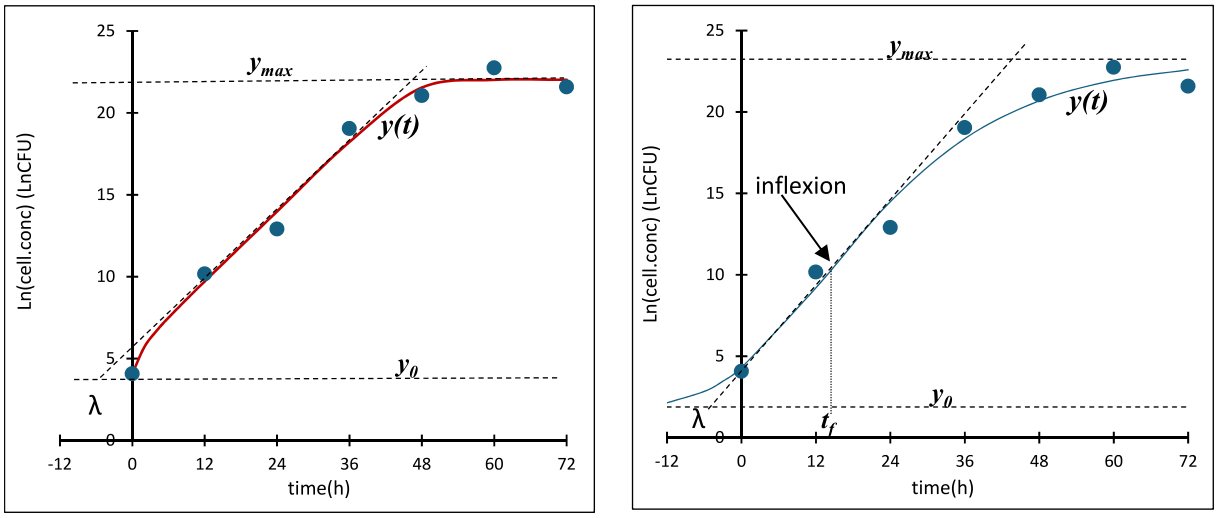


Fig. 5. Case of negative lag. Blue circles: GF-simulated data set. Red and blue continuous lines: fitted by BRM and GF. For the lag estimate, $\widehat{\lambda} < 0$ in both cases, though with different interpretations. The maximum specific growth rate is $\mu_{max} = y(t_f)/(t_f - \lambda)$, where t_f is the time of inflexion point of $y(t)$. For BRM, the $y(t)$ fit is not even a sigmoid function, if the lag parameter is negative. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Glossary of notations

P	Parameter
$\widehat{p}$	parameter estimate
se($\widehat{p}$)	standard error of parameter estimate
t	Time
t_f	time of inflexion of a sigmoid curve
$x(t)$, $y(t)$	bacterial concentration and its natural logarithm: $y(t) = \text{Ln } x(t)$, at the time t
x_0, y_0	Initial bacterial concentration and its natural logarithm: $y_0 = y(0) = \text{Ln } x(0) = \text{Ln } x_0$
$\mu(t)$	specific growth rate of a bacterial population at the time t: $\mu(t) = \text{d}x/\text{d}t/x$
μ_{max}	maximum specific growth rate of a cell population in a constant environment
λ	lag parameter of a sigmoid curve
$P(t)$	(temporal) concentration of a key metabolic product needed for the new environment
K_P	Michaelis-Menten constant for $P(t)$
$q(t)$	$q(t) = P(t)/K_P$
q_0	$q_0 = q(0)$
α_0	$0 \leq \alpha_0 = \alpha(0) \leq 1$ "suitability" parameter, or initial physiological state
h_0	$h_0 = \mu_{max}\lambda$. In the model of Baranyi and Roberts (1994): $h_0 = -\text{Ln}(\alpha_0)$
se_{fit}	standard error of fit
t_0	intercepts of the linear growth model with the time axis: $y(t) = \mu_{max}\cdot(t - t_0)$, where $t_0 = -y_0/\mu_{max}$

CRediT authorship contribution statement

Maha Rockaya: Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation. **József Baranyi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

MR has been supported by the Stipendium Hungaricum Scholarship Programme of the Ministry of Foreign Affairs and Trade of Hungary, via the Tempus Public Foundation.

The authors thank Mariem Ellouze for consultations and many ComBase users for pointing out anomalies in the implementation of BRM in regression tools.

References

Akkermans, S., Nimmegeers, P., Van Impe, J.F., 2018. A tutorial on uncertainty propagation techniques for predictive microbiology models: a critical analysis of state-of-the-art techniques. Int. J. Food Microbiol. 282, 1–8.

Baranyi, J., 1992. Notes on reparametrization of bacterial growth curves. Food Microbiol. 9, 3.

Baranyi, J., Roberts, T.A., 1994. A dynamic approach to predicting bacterial growth in food. Int. J. Food Microbiol. 23 (3–4), 277–294.

Buchanan, R.L., Klawitter, L.A., 1991. Effectiveness of *Carnobacterium piscicola* LK5 for controlling the growth of *Listeria monocytogenes* Scott A in refrigerated foods. J. Food Saf. 12, 219–236.

Corry, J.E., Jarvis, B., Passmore, S., Hedges, A., 2007. A critical review of measurement uncertainty in the enumeration of food micro-organisms. Food Microbiol. 24 (3), 230–253.

Nau, R., 2016. Statistical forecasting: notes on regression and time series analysis. Stepwise and All Possible Regressions. Available online: https://people.duke.edu/~rnau/regstep.htm. (Accessed 2 May 2019).

Nauta, M.J., 2000. Separation of uncertainty and variability in quantitative microbial risk assessment models. Int. J. Food Microbiol. 57 (1–2), 9–18.

Pérez-Rodríguez, F., Valero, A., 2013. Predictive Microbiology in Foods. Springer, New York, pp. 1–10.

Press, W.H., 2007. Numerical recipes. In: The Art of Scientific Computing, third ed. Cambridge University Press.

Rolfe, M.D., et al., 2012. Lag phase is a distinct growth phase that prepares bacteria for exponential growth and involves transient metal accumulation. J. Bacteriol. 194 (3), 686–701.

Srinivasan, B., 2022. A guide to the Michaelis–Menten equation: steady state and beyond. FEBS J. 289, 6086–6098.

Turner Jr, M.E., Bradley, Jr E.L., Kirk, K.A., Pruitt, K.M., 1976. A theory of growth. Math. Biosci. 29 (3–4), 367–373.

Zwietering, M.H., Jongenburger, I., Rombouts, F.M., Van't Riet, K.J.A.E.M., 1990. Modeling of the bacterial growth curve. Appl. Environ. Microbiol. 56 (6), 1875–1881.