Growth curves: Regression and ranking of growth curves

1 Introduction

Growth curves are widely used to model animal growth. In general experimental data of a population is gathered and growth curves are fitted to the samples and one is chosen to represent the growth of the population. These curves are generally sigmoidal or attenuating, and common models include the logistic, Gompertz, and von Bertalanffy curves, as shown in Table 1.

Typically there are many different parametrisations of these curves, for example see E. Tjørve and K. M. Tjørve (2010) which compares the different different parametrisations of the Richards curve. These different parametrisations result in mathematically equivalent forms, yet due to the nature of the numerical algorithms different parametrisations may work better for different data sets Myhrvold (2013).

2 Literature

2.1 Regression of growth curves

Model name	Number of results
Logistic	3908
Gompertz	1566
Von Bertalanffy	1477
Richards	593
Weibull	591
Michaelis-Menten	179
Monomolecular	129
He Legendre	88
Korf	39
Morgan-Mercer-Flodin	16
Levakovic	6

Table 1: Number of results according to web of science search across all databases, including the model name and the phrase "growth curve"

Various mathematical models can be used to explain and interpret growth. Asymptotic models (often sigmoidal) are commonly used to describe deter-

Paper	Number of Models
Myhrvold (2013)	77
Do and Miar (2019)	10
García-Muñiz et al. (2019)	10
Ghavi Hossein-Zadeh (2015)	5
Kheirabadi and Rashidi (2019)	5
Cooper et al. (2008)	5
Gbangboche et al. (2008)	4
Araújo et al. (2012)	4
Keskin et al. (2009)	4
Perotto, Cue, and A. J. Lee (1992)	4
Xie et al. (2020)	3
Erickson et al. (2015)	3
DeNise and Brinks (1985)	2
Waheed et al. (2011)	2
Yakupoglu and Atil (2001)	2
Ersoy, Mehmet, and Aktan (2006)	1
Bathaei and Leroy (1998)	1
Raji, Alade, and Duwa (2014)	1
Ghiasi, Lupi, and Mokhtari (2018)	1
Li and Zhao (2019)	1
Lehman and Woodward (2008)	1
Chinsamy (1990)	1

Table 2: The number of growth curves compared in various papers

minate growth, while a wide range of models can be used for indeterminate growth. Models are commonly written in algebraic form although differential models do exist, for examples see Brunner et al. (2019) and Myhrvold (2013). Models in differential form can sometimes be more flexible however they are more difficult to regress to data sets, in this paper only models that can be written in closed algebraic forms will be considered.

Despite the number of available models one commonly finds works in which very few models are compared to each other, typically less than three: for example

2.1.1 Uncertainty in one variable

The determination of growth curves parameters is an example of explicit regression, i.,e.

$$y \approx f(x : \beta) \tag{1}$$

Where y is the growth variable (size, length, etc.), x is the time variable, and β is the vector of fitted parameters. y is only approximately equal to $f(x)$ due to uncertainty (unknown errors) in x and or y. For a sample i from the population, the exact equation is written as:

$$y_i = f(x_i + \delta_i : \beta) + \epsilon_i \tag{2}$$

With δ_i is the error in x, ϵ is the error in y and the subscript i refers to the specific sample. These errors are assumed to be normally distributed and independent.

If there are significant errors in only the y variable, i.e. $\delta_i = 0$ for any i, then ordinary least squares (OLS) can be used to perform the regression. This is because a key assumption of OLS is that the measurements of the independent variable (x, age) are without errors. If the x variable has errors then it is fundamentally flawed to use OLS.

Depending on how the samples are taken there may be errors in the ages of the animal — this is typically the case when feral animals are being sampled and the age is estimated by features of the animal for example see Mayberry et al. (2018). Commonly the growth variable (y) can be accurately measured (e.g. in the case of weight), which means that there would still be error in only one of the variables. The growth function (f) can be inverted ($f^-1 = g$) and OLS can be used with y_i as the independent variable, i.e.

$$x_i = g(y_i : \beta) + \epsilon_i \tag{3}$$

This method of inversion to use OLS is referred to as inverted OLS (OLSI) in this work. This method measures the residuals in a fundamentally different way, shown graphically in Figure 1. The variable which has no measurement error is assumed to be exact and the residual is measured as occurring entirely in the other variable. Thus the use of OLSI can give different regressed parameters to OLS.

The use of OLSI has been recommended numerous times over the years (Kaufmann 1981; Myhrvold 2013), yet it is common find that it is not specified if inverted functions are used, or to examine the methodology and see the use of non-inverted functions despite errors in the age variables, for example Erickson et al. (2015).

Although the inversion of the growth curve functions is typically simple, using inverted functions does result in different restrictions in the variables, i.e. the function g must accept any of the population sample y values, while f must accept any of the x values. To give an example, the logistic curve is given by:

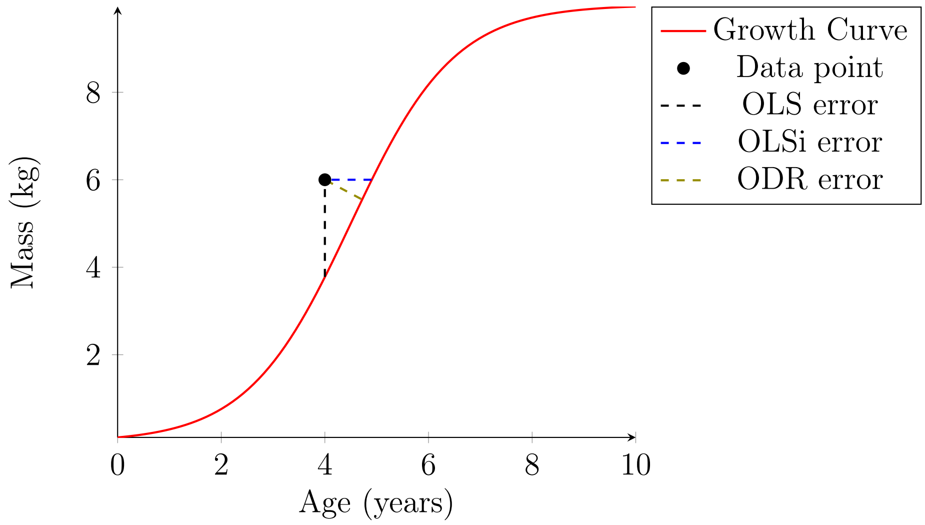


Figure 1: Example of the different measurements of fit according to OLS, OLSi and ODR. The growth curve is a logistic curve, with $a = 10$, $b = -1$ and $c = -4.5$

$$y = f(x) = \frac{a}{1 + e^{-b(-c+x)})} \qquad a > 0, \quad b > 0 \tag{4}$$

$$x = g(y) = \frac{bc - \ln(-1 + \frac{a}{y})}{b} \qquad a > y, \quad b > 0 \tag{5}$$

One cannot take the natural log of 0 and hence the inversion of f introduces a constraint that was not present in the un-inverted form.

2.1.2 Uncertainty in two variables

If there is significant uncertainty in both the x and y values of the sample then an error-in-variable regression model must be used. A commonly used model is orthogonal distance regression (ODR).

Graphically, in ODR the error of the fit by a line orthogonal to the curve, shown in Figure 1. A robust, and widely used implementation of ODR, was introduced by Boggs et al. (1989), updated in Zwolak, Boggs, and Watson (2007). This algorithm is available in various programming languages including Python and R. One should weight the errors in the different variables, typically by the estimated variance in the parameters.

This method along with OLS and OLSI is used in this work, when appropriate.

2.2 Criteria for model selection

Once the growth curve parameters have been fitted, one must select a model(s) that best describes the data. Although this has been been done by "author judgement" there are statistical parameters that give indications how well the models fit the data.

2.2.1 The coefficient of determination

The most common widely used parameter is the coefficient of determination (R^2). This is the proportion of variance in the dependent variable predicted from the independent variable. The best possible value of R^2 is 1, which indicates the model exactly fits the data, with worse models having a lower value.

$$R^2 = 1 - \frac{RSS}{TSS} \tag{6}$$

where RSS is the residual sum of squares and TSS is the total sum of squares.

It is important to note that R^2 gives no measure of over-fitting, and that increasing the number of parameters in a model will always increase R^2 or keep it constant.

2.2.2 Information Criterions

The Akaike information criterion (AIC) is an estimator of the quality of a model and is commonly used to give a statistical basis of model selection (Akaike 1974). An alternative criterion that is widely used is the Bayesian information criterion (BIC), also known as the Schwarz information criterion (Schwarz 1978). With both criterions the "preferred" model has the lowest value. Both of the criterions are similar measures the goodness of model fit and penalise it according to the number of model parameters. This can be interpreted as an implementation of Occam's Razor, i.e. for models that are similar in accuracy, the simpler model (i.e. has less adjustable parameters) is the better choice. The criterions are defined as:

$$AIC = 2k - 2\ln(L) \tag{7}$$

$$BIC = k\ln(n) - 2\ln(L) \tag{8}$$

$$-2\ln(L) = n\ln\frac{RSS}{n} \tag{9}$$

where RSS is the residual sum of squares, n is the number of samples, k is the number of parameters, L is the maximum value of the likelihood functions.

Note that errors are assumed to be independent and identically distributed with zero mean, which specifies the variance of the error which contributes to the count of parameters, k. To give an example: a linear fit,

$$y \approx \beta_o + \beta_1 x \tag{10}$$

has $k = 3$ if there are errors in one variable, and $k = 4$ if there are errors in both variables.

As the the information criterions are used to compare models it is common practice to calculate them as Δ values, i.e.

$$\Delta IC_k = IC_k - \min(IC) \tag{11}$$

where IC is some information criterion, $\min(IC)$ is the minimum criterion found, and the subscript k refers to a particular model. For example if one was comparing models, M_1 and M_2 with AIC's of 10 and 12 respectively then the ΔAIC's are 0 and 2.

Both criterions may may select models incorrect models when small data sets are used: AIC tends to select models that over-fit while BIC selects those that under-fit (Hurvich and Tsai 1989; McQuarrie 1999). Corrections can be defined to correct for this behaviour, giving the corrected AIC (AICc) and BIC (BICc):

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \tag{12}$$

$$BICc = k\ln(n)\frac{n}{n-k-1} - 2\ln(L) \tag{13}$$

Both of the correction converge to the uncorrected criterion when the number of samples is large, and their use is commonly advised over the uncorrected samples (K. Burnham and D. Anderson 2002).

2.2.3 Interpreting the criterions

If there is a "true model" is in the set of available models then BIC will select the true model with 100% probability, as $n \to \infty$. A "true model" is the model that generates the data. The existence of such a "true model" when using real data is questionable. It was famously noted "All models are wrong" (Box 1976), or to be more explicit, it is unlikely that any real system can be described exactly by some simple model.

There is no guaranteed that AIC will select the "true model" as it may select a more complex, but better fitting, wrong model (K. Burnham and D.

Anderson 2002). As growth is unlikely to be governed exactly by any growth model, the use of BIC doesn't hold any advantage over AIC. Some authors strongly suggest that AIC should be used over BIC, while others recommend reporting both (K. P. Burnham, D. R. Anderson, and Huyvaert 2011; A. H. Lee et al. 2013). Both AIC, BIC and their corrected forms are calculated in this paper.

Plausible models are generally taken as having a $\Delta AICc \leq 4$, with implausible models having $\Delta AICc > 14$ (K. P. Burnham, D. R. Anderson, and Huyvaert 2011). The exact cut-off value to not consider a model varies, with some recommending that models up to $\Delta = 7$ should be considered.

3 Case studies

3.1 How the choice of regression method changes parameters

Data for the ceratopsian dinosaur *Psittacosaurus lujiatunensis* is given in Erickson et al. (2009, 2015). Femoral lengths and growth lines were used to estimate specimen mass and age respectively, with data shown in Table 3. Mass estimates were used to compare the growth curve to other dinosaurs and living vertebrates Erickson et al. (2015). Three growth curves were computed and a logistic curve was chosen based on its R^2 value, and because the Gompertz curve was "biologically unreasonable". Issues with the methodology have been raised by Myhrvold, (Myhrvold 2013, 2015), leading to a later response by the original authors in Erickson et al. (2015). These issues will not be delved into here, instead the focus is on the choice of regression method, and what impact this has on the regression. Table 3 clearly shows that there is (large) age estimation, as age is approximated from growth lines, some of which are destroyed. The uncertainty in the femoral length measurements is marginal in comparison. Thus it would be justifiable to perform inverse OLS (OLSi) as discussed above and in Kaufmann (1981).

Erickson et al. (2015) chose the three parameter logistic function:

$$y = \frac{a}{1 + \exp -b(x - c)} \tag{14}$$

with parameter values of $a = 37.48$, $b = 0.55$ and $c = 9.03$. The uncertainties in these parameters were not given so the developed code was used to perform OLS followed by OLSi regression and these results are presented in Table 4.

Based on the results, it is clear that Erickson et al. (2015) used the OLS method as the fitted parameters (Table 4) are exactly as reported. OLSi

Femoral length (mm)	Growth lines	Age estimate (years)	Mass estimation
49-55	0	0.5	0.45
49-55	0	0.5	0.45
34	1	1	0.12
38-45	1	1	0.29
52	2	2	0.50
80-85	3	3	1.61
80-83	3-4	3	1.97
103	2	3	3.50
101-105	2-3	4	3.30
95-98	3-4	4	3.01
108-112	3-4	5	4.02
110-111	4	5	4.34
135-144	5	7	8.40
149-150	4	7	7.87
135-137	5-6	7	8.05
135	5	7	10.80
164-165	6-7	8	10.36
153-156	5	8	11.97
148	5-7	8	14.36
189-190	7-8	9	21.94
199-201	8	10	25.96
200-202	6	11	25.96

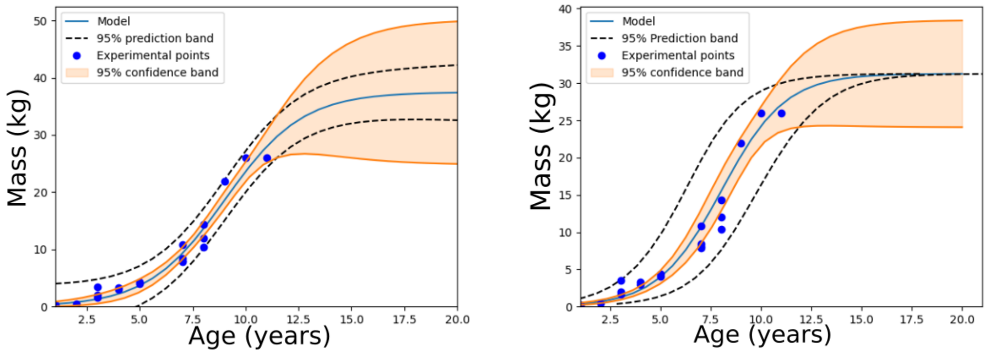
Table 3: Growth data of *Psittacosaurus lujiatunensis* from Erickson et al. (2015)

Parameter	OLS	OLSi
a	37.48 ± 6.48	31.26 ± 3.67
b	0.55 ± 0.08	0.63 ± 0.05
c	9.03 ± 0.72	8.04 ± 0.56

Table 4: Fitted parameters for the *Psittacosaurus lujiatunensis* data, using inverted and normal OLS with Equation 14

and OLSi and OLS give different nominal parameters (i.e. central parameters in the confidence range), with OLSi giving smaller ranges for the parameters. This is most noticeable in the a parameter in which the parameter decreases by nearly half.

Thus the use of the OLSi and OLS on this data will give different curves, as shown in Figure 2, due to the differences in nominal values and uncertainties. Although the shape of the curve is similar there are striking differences in the confidence band. This is expected, as the methods differ based on to which variable the “fitting error” is associated with. The figure clearly shows the influence that the choice of regression method can have, as interpretation of the growth curves would differ, with regression with mass as the independent variable being statistically advised (Kaufmann 1981; Myhrvold 2013). Note that the large uncertainty near the asymptote is simplyu due to the lack of data points above 13 years.



(a) Age as independent variable (OLS) (b) Mass as independent variable (iOLS)

Figure 2: Comparison of choice of independent variable for the *Psittacosaurus lujiatunensis* data