MSc Computational methods in Ecology and Evolution

Imperial College London

Department of Life Science

Non-linear least squares Gompertz models outperform linear models and non-linear Logistic growth models in fitting microbial population growth data

Author:
Chuxinyao Wang

*Word count:*2025
December 4, 2022

Non-linear least squares Gompertz models outperform linear models and non-linear Logistic growth models in fitting microbial population growth data

Chuxinyao Wang

November 2022

Abstract

AIC has been widely used in ecology, particularly in relation to model selection and model ranking. In this paper, 285 sets of microbial growth data were input and models were fitted to those data sets by using computational methods with R. The data were fitted by using a total of 6 models in both types of linear and non-linear least squared fits. By comparing the AIC, AIC_c and adjusted R^2 values of each model with the 285 data sets, the logistic growth model model was found to be stable for the majority of the 285 data sets and was the final optimal model option.

1 Introduction

Microbiology has long used growth rates to quantify phenotypic properties to determine bacterial growth rates, and growth rates are an important basis for all areas of microbiology(Hall et al. 2013). And the rate of change affects the trend of population growth. Estimating population growth has become a central focus of population biology. The use of modeling to fit population growth trends can more generally represent the population changes of a species. Much ecological work needs to drive the development of new predictive ecological methods that can reliably predict future changes in population size (Matthiopoulos et al. 2019). Therefore, modelling population growth came into being. The purpose of this article is to evaluate and select the model, and the standard used is to compare and compare AIC, AIC_c, and adjusted R^2 values based on procedures written by Johnson &

Omland (2004) .The Akaike Information Criterion (AIC) is an important reference index in the comparison of models in the field of ecology, where it is often necessary to compare and rank a number of potentially available models and to assess which of them is better able to deliver the expected results in the biological processes under study (Symonds & Moussalli 2010). And, Stephens et al. (2007) also mention the tendency of using IT techniques such as AIC to replace null hypothesis significance testings in ecology, although many ecologists are still opposed to such IT techniques as model selection under ecological applications, arguing that they do not facilitate the development of hypothesis sets (Eberhardt 2003, STEPHENS et al. 2005-02).

Regardless of whether the influence of IT technology on statistics is controversial or not, this paper mainly uses this feasible IT technology method to evaluate the proposed possible population growth model, and uses AIC as a criterion for assessing the fit of the model to compare the goodness of the models then select the best model among competing models.

2 Methods

2.1 Data collection

The data used to evaluate the model this time came from data on the growth of various microbial populations sampled in different journals(Data address). A total of 285 sets of data were obtained after classification by computer methods. The 285 sets of data include differences such as different species, different temperatures, and different culture media (see metadata in the Appendix. 1 for details), which makes model selections more convincing and allows them to be verified under different conditions.

2.2 Candidate Models

As mentioned by Johnson & Omland (2004), articulating a reasonable set of competing models should be chosen prior to data collection. The data was provided for model selection without my intervention, so the proposal of the candidate models can be regarded as the proposal before knowing the data. Considering the rationality and diversity of the model, the model is selected from the linear model and the nonlinear model respectively. The first is the linear function model:

$$N = r \cdot t + N_0 \tag{1}$$

Considering that population growth is affected by the growth rate and the initial population, the linear function of the growth rate as the slope and the initial population as the intercept is used as the first candidate model. The second model considers that the linear relationship of population growth may have a linear relationship between the logarithm of the population and time, so the second model is also a linear function model, but the dependent variable is changed to the logarithm of the population, and the intercept is also changed to Logarithm of the initial population size:

$$logN = r \cdot t + logN_0 \tag{2}$$

The third model considers that in addition to the growth rate, there are some unknown factors (a_1 and a_2 in $Equation_(3)$) that have an impact on the population change, so the quadratic polynomial linear model is used as the third alternative model:

$$N = a_1 \cdot t^2 + a_2 \cdot t + N_0 \tag{3}$$

Similarly, the logarithm of the population may have a linear relationship with the influencing factors(b_1 and b_2 in $Equation_(4)$), so the logarithm of the population is used as a quadratic polynomial model as the fourth alternative model:

$$logN = b_1 \cdot t^2 + b_2 \cdot t + logN_0 \tag{4}$$

The fifth model takes nonlinear model and logistic model as alternative models. It is an important equation of population biology, proposed by Verhulst in 1938 (Verhulst 1838). This model also takes into account the capacity of the environment, that is, the maximum number of people in a limited environment. The reason for including this model in the alternative model is also to determine the reliability of this model among many models. The differential equation form of the logistic growth model is expressed as:

$$\frac{dN}{dt} = r \cdot N \cdot (1 - \frac{N}{K}) \tag{5}$$

For the sixth model, I used the Gompertz model, a commonly used model for microbial population growth (Zwietering et al. 1990). According to Buchanan et al. (1997), the characteristic of this model is that the bacterial growth curve is divided into three phases: lag, stationary and exponential according to different growth rates. The model has four parameters:

$\log N_0$ (log of initial population), $\log N_{MAX}$ (log of maximum population) , t_{lag} (time when the lag phase ends), and t_{MAX} (time when the exponential phase ends). Gompertz model equation is:

$$\log_{N_t} = N_0 + (N_{MAX} - N_0) \cdot e^{-e \cdot r_{max} \cdot exp(1) \frac{t_{lag} - t}{(N_{MAX} - N_0) \cdot \log_{10} + 1)}} \tag{6}$$

These six models were all imported into the computer, tested one by one using the above 285 sets of data, and fitted the data into the model.

2.3 Model Fitting and Comparison

The process of model fitting requires the use of a computer. In R, after the six models are written, a preliminary test of the model is required. After grouping the original data, substituting a group of data to conduct preliminary verification on whether the model is reasonable, and the verification model can better reflect the changing trend of the data.

After completing the preliminary model and rationality verification, 285 sets of data need to be imported circularly, and the model comparison criteria, ie. AIC AIC_c, and adjusted R^2, are performed through the computer, and each set of calculated data is stored in turn for consequent comparison. For the calculation of AIC and AIC_c, I set the calculation scale on the exponential scale. Since all the models on the logarithmic scale, the dependent variables are often negative and unable to be calculated in the AIC calculation (involving the second order Logarithmic operation), so models under scale of logarithm were taken the natural logarithm index e of the dependent variable before calculating the sum of squared residuals, and then perform scale matching with other non-logarithmic scale models, thus AIC, $AICc$ comparison are more precise.

In addition, the reason why AIC and AIC_c are used as comparison criterion both is that in the 285 sets of data, the number of each set of data is different, while AIC_c is more accurate when the amount of data is small, and when the amount of data is large to a certain extent, the calculation result of AIC_c is basically AIC (Brewer et al. 2016). Therefore, including the two criteria at the same time for comparison can double judge the performance of the model under the same set of data.

For the final model comparison link, how to deal with the obtained AIC, AIC_c and adjusted R^2 becomes the focus of comparison. In Johnson &

Omland (2004) Model selecting approach part, they performed the likelihood calculation on AIC and compared the weight assignment of each model. This is a relatively complete and complicated process. I simplified it to make it easier to understand and accept. Specifically, the distribution of AIC and AIC_c is visualized and analyzed first. And calculate the average and variance of AIC and AIC_c of each model, and compare AIC_c first. In general, the smaller the average of AIC_c and AIC, the better the model. If the average of the models is not much different Next, compare its stability, that is, variance comparison. The adjusted R^2 is finally used as an auxiliary comparison index, and the closer to 1, the better model is.

2.4 Computing languages and tools

The computer language used in the whole process of model selection analysis and the use of each package will be presented in the following table.

Table 1: Computing languages and tools use

Computer languages	**Packages**	**Uses**
Python	Pandas	Data grouping and storing
R	ggplot2	Preliminary model fitting inspection work and model comparison criteria distribution plot drawing
	minpack.lm	Fit nonlinear Models using the Levenberg-Marquardt Algorithm within the package
	ggpubr	Typesetting for diagrams
LaTeX	natbib	Set the Citation and reference format to Harvard
	appendix	Create appendix
	hyperref	Create a link to direct to the data source address
	booktabs	Create tables
	graphicx	Input figures
Bash		Compile Miniproject report

3 Results

The calculation results of AIC, AIC_c, and adjusted R^2 descriptive statistics of each model are shown in Table 2, 3, and 4. From the mean and median of the AIC values, the AIC value of the Gompertz model is significantly smaller than that of the other 5 models' AIC. Also, AIC_c values have the same result, Gompertz also has the smallest AIC_c value.

Table 2: Descriptive statistics for AIC

model type	**mean**	**median**	**min**	**max**
linear model	133.02	96.24	-265.78	2423.012
log linear model	162.73	105.82	-262.19	3038.23
polynomial model	118.43	87.46	-267.74	2419.19
log polynomial model	133.65	94.42	-265.31	2467.60
logistic model	83.49	67.23	-253.09	1917.48
gompertz model	5.68	12.78	-319.09	16.24

Table 3: Descriptive statistics for AIC_c

model type	**mean**	**median**	**min**	**max**
linear model	137.86	105.00	-264.74	2423.18
log linear mdoel	156.73	99.82	-268.19	3032.23
polynomial model	127.68	108.16	-265.92	2419.47
log polynomial model	142.90	112.78	-263.49	2467.88
logistic model	90.74	81.59	-251.35	1917.86
gompertz model	13.01	17.23	-318.82	36.24

Table 4: Descriptive statistics for Adjusted R^2

model type	**mean**	**median**	**min**	**max**
linear model	-1768.3980	0.3870	-466006.9655	0.9826
log linear mdoel	-10.9571	0.6247	-1133.3364	0.9931
polynomial model	0.3590	0.8125	-23.2333	1.1204
log polynomial model	0.4334	0.8655	-23.1136	1.1157
logistic model	-3.5253e+28	0.9492	-7.6852e+30	0.99996
gompertz model	0.1168	0.9779	-23.8145	0.9988

However, in the adjusted R^2 table, large outliers appear, which is reflected in the large deviation of the adjusted R^2 values of the six models from the median. The minimum adjusted R^2 value of each group of models is far less than 1. Even in the logistic model, the adjusted R^2 value fitted by one data is too small to be regarded as an invalid value. Therefore, it is more efficient to compare medians here. From Table 3, the Gompertz model wins again among the six models, the median of adjusted R^2 is closest to 1.

The distribution box plots of AIC, AIC_c, and adjusted R^2 in the six models are shown in Figures 1, 2, and 3.

In the box plots of the first two AIC and AIC_c values, the Gompertz model is very stable, with fewer outliers than the other five groups of models. The worse performance is the logarithmic linear model.

The adjusted R^2 box plot excludes outliers that are too small, so as to better present the distribution of the adjusted R^2 value of the data sample. In the adjusted R^2 box plot, it can be seen that the fit of the Gompertz and Logistic models to the data is better than that of several other models, while the fit of the linear model is the lowest overall.

4 Discussion

The aim of this paper is to rank how well the candidate models fit the data. By comparing the AIC, AIC_c, and adjusted R^2 of different data, it can be concluded that the AIC and AIC_c median of the Gompertz model are the lowest, and the Gompertz model has the most stable fitting results for all data, and has the highest adjusted R2 value . Therefore, it can basically be determined that the optimal model choice among the candidate population growth models is the Gompertz model. Secondly, except for the huge outliers in the adjusted R^2 value of the logistic model, its AIC and AIC_c are only larger than the Gompertz model, so the logistic model ranks second only to the Gompertz model. After a similar comparison, it can be determined that the final ranking is $1.Gompertz > 2.Logistic > 3.Polynomial > 4.linear > 5.logpolynomial > 6.loglinear$.

It is important to note that the model with the logarithm is not as effective at fitting the data as the model without the logarithm. It appears that there is a weaker correlation between the population's logarithm and

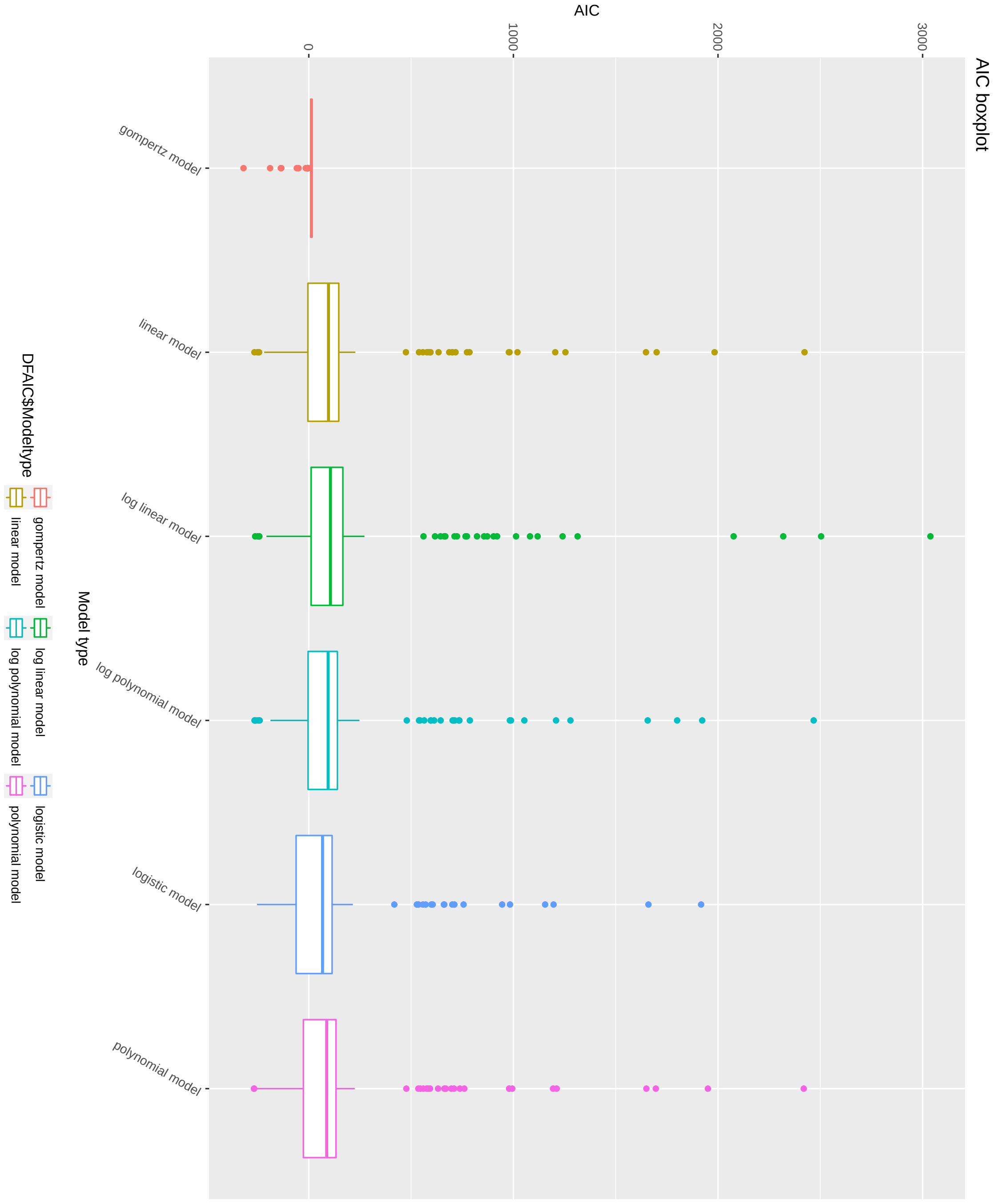

Figure 1: Box plot of AIC

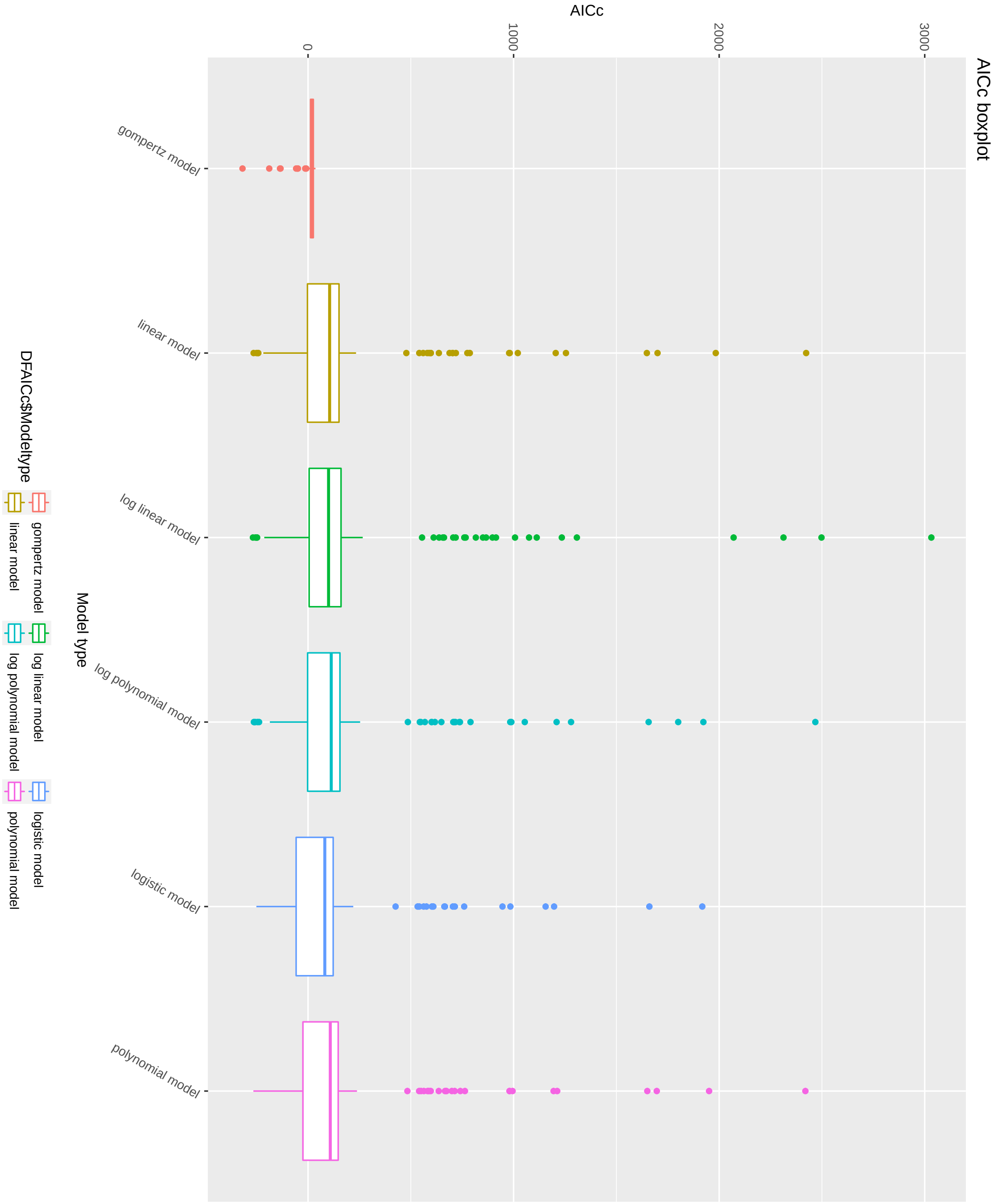

Figure 2: Box plot of AIC_c

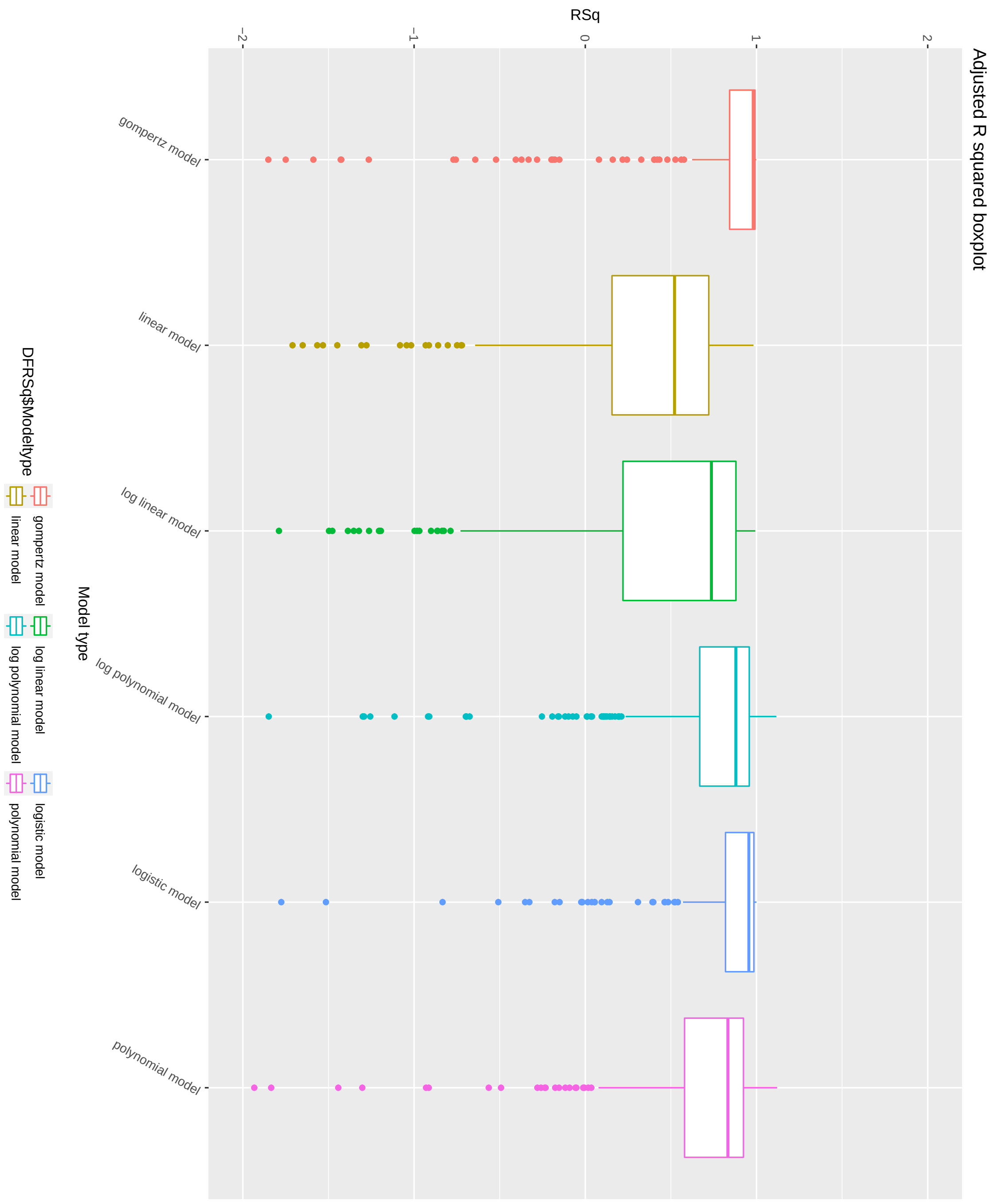

Figure 3: Box plot of adjusted R^2

the influencing variables than there is between the population as a whole and the influencing factors, but Gompertz model just becomes the exception, performs the best and merits additional study because it is fitted on a logarithmic scale.

Besides, according to the AIC value mentioned by Brewer et al. (2016), there is a penalty mechanism for the number of parameters, that is, the more parameters, the values of AIC and AIC_c will be reduced accordingly, and the same is true for the adjusted R2.(The number of parameters is $k_{linear} = 3, k_{loglinear} = 3, k_{polynomial} = 4, k_{logpolynomial} = 4, k_{logistic} = 4, k_{gompertz} = 4$) Therefore, AIC, AIC_c and adjusted R^2 are compared between models with the same number of parameters Finally, the conclusions that can be drawn are consistent with the overall comparison and the rankings have not changed.

This model comparison study did not carry out deeper calculations such as Johnson & Omland (2004) use of likelihood values and weighting scores to compare the fit between models, which was rougher and less accurate for model comparison and ranking.

Secondly, the data this time belong to the growth data of microbial population, lacking the comparison of the model fit to the change of population of other organisms. Perhaps, the best model in the growth of microbial population is Gompertz, but in a larger environment it may be The Gompertz model cannot better reflect the significant impact of the actual environment or other recessive factors on the biological population from a mathematical point of view. Therefore, considering the diversity of data is also a further determination of the performance of the model. Secondly, there are differences in data in different living environments (such as temperature, culture medium), which can also be used as one of the conditions to analyse the fitting status of the same set of candidate models under different environmental conditions, where research has been insufficient.

5 Appendix

Table 5: Meta data for 285 sets of data in model fitting

Time	Time at which measurement was taken.
PopBio	Population or biomass measurement.
Temp	Temperature at which the microbe was grown (degrees Celsius).
Time_units	Units time is measured in.
PopBio_units	Units population or biomass are measured in.
Species	Species or strain used.
Medium	Medium the microbe was grown in.
Rep	Replicate within the experiment.
Citation	Citation for the paper in which the study was recorded.

References

Brewer, M. J., Butler, A. & Cooksley, S. L. (2016), 'The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity', *Methods in Ecology and Evolution* **7**(6), 679–692.

Buchanan, R., Whiting, R. & Damert, W. (1997), 'When is simple good enough: a comparison of the gompertz, baranyi, and three-phase linear models for fitting bacterial growth curves', *Food Microbiology* **14**(4), 313–326.
URL: *https://www.sciencedirect.com/science/article/pii/S0740002097901258*

Eberhardt, L. L. (2003), 'What should we do about hypothesis testing?', *The Journal of Wildlife Management* **67**(2), 241–247.
URL: *http://www.jstor.org/stable/3802765*

Hall, B. G., Acar, H., Nandipati, A. & Barlow, M. (2013), 'Growth Rates Made Easy', *Molecular Biology and Evolution* **31**(1), 232–238.
URL: *https://doi.org/10.1093/molbev/mst187*

Johnson, J. B. & Omland, K. S. (2004), 'Model selection in ecology and evolution', *Trends in ecology & evolution* **19**(2), 101–108.

Matthiopoulos, J., Field, C. & MacLeod, R. (2019), 'Predicting population change from models based on habitat availability and utilization', *Proceedings of the Royal Society B* **286**(1901), 20182911.

Stephens, P. A., Buskirk, S. W. & del Rio, C. M. (2007), 'Inference in ecology and evolution', *Trends in Ecology Evolution* **22**(4), 192–197.
URL: *https://www.sciencedirect.com/science/article/pii/S0169534706004009*

STEPHENS, P. A., BUSKIRK, S. W., HAYWARD, G. D. & MARTINEZ DEL RIO, C. (2005-02), 'Information theory and hypothesis testing: a call for pluralism', *The journal of applied ecology.* **42**(1).

Symonds, M. R. E. & Moussalli, A. (2010), 'A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using akaike's information criterion', *Behavioral Ecology and Sociobiology* **65**, 13–21.

Verhulst, P.-F. (1838), 'Notice sur la loi que la population suit dans son accroissement', *Corresp. Math. Phys.* **10**, 113–126.

Zwietering, M. H., Jongenburger, I., Rombouts, F. M. & van 't Riet, K. (1990), 'Modeling of the bacterial growth curve', *Applied and Environmental Microbiology* **56**(6), 1875–1881.
URL: *https://journals.asm.org/doi/abs/10.1128/aem.56.6.1875-1881.1990*