NLLS Models Generally Outperform OLS Models for Quantifying Microbial Population Growth Data

Author: Shengge Tong (shengge.tong22@imperial.ac.uk)

Word Count:1323

Contents

Abstract

The abundance (density) of a single population may be critical to determining ecosystem dynamics and functional characteristics. And microbial (specifically, bacterial) growth rates is a hot research topic. This report focus on the model fitting and comparison of the population growth data. There are four models in the study, including OLS(Quadratic polynomial and Cubic polynomial) and NLLS (Logistic and Gompertz) models. The study tries to find the best model by comparing the AIC, AIC_c, R^2 (three measurements). The results showed that Gompertz model had best performance on fitting the data. Logistic model is sufficient for the simple growth case. OLS models are suitable for simple cases. In conclusion, Gompertz should be preferred if extreme performance is sought, as the model has more parameters and a relatively low probability of convergence, it may require more potential labor costs to adjust the initial values.

1 Introduction

Human society has a strong correlation with microbial growth. For example, Wine fermentation cannot be done without yeast, mold can make food spoiled, pneumococcus can make people sick, and some probiotics can keep people's intestinal health. There are four periods for bacterial growth:
(1) lag phase: After inoculation of the bacteria into the medium, there is a short adaptation process to the new environment (those that do not adapt can die due to the transfer). The curve is flat and stable during this period, as the bacteria multiply very little.
(2) logarithmic phase: The number of viable bacteria on the growth curve increases linearly during this period. The bacteria grow very rapidly in a steady geometric progression that can last from a few hours to a few days. The morphology, staining and biological activity of bacteria in this stage are typical and sensitive to the action of external environmental factors, so the study of bacterial traits is best in this stage.
(3) stationary phase: The total number of growing colonies in this period was in a flat phase, but the bacterial population viability was highly variable. The number of bacterial proliferations and deaths were gradually balanced in this period.
(4) decline phase: Bacteria multiply more and more slowly, and the number of dead bacteria increases significantly. The number of live bacteria is inversely related to the incubation time, the bacteria become long and swollen or deformed decay in this period, and the physiological metabolic activities tend to stagnate.

There are many mathematical models could be used to predict the growth curve. According to the past research and study, some NLLS models such as modifed Gompertz model (**(author?)** 5), Logistic model(**(author?)** 3) are widely used in this field. And our goal is to compare the OLS models (Cubic, Quadratic)(**(author?)** 4)(**(author?)** 1) with NLLS models by AIC, AIC_c, R^2 (**(author?)** 2). Thus, this report is mainly about which model has the best performance, and illustrate their pros and cons. Finally, a good strategy could be made for fitting the data in different cases.

2 Data and Methods

2.1 Data Preparation

The raw dataset for project is LogisticGrowthData.csv. And the field names are in LogisticGrowthMetaData.csv. They are both in the data directory. This project mainly focus on the relation between PopBio(abundance) with Time. Firstly, I loaded the data and combined Species, Medium, Temp and Citation columns to set up the ID coulmns. Secondly, I filtered the data for Time and PopBio larger than 0, removing meaningless data. Then I seperated them into several groups to get 256 independent experiments.

2.2 Models

There are 4 models to be used in the project. Polynomial models and Logistic model are fitted with non-logged data, and Gompertz model with logged data. Here are the formulas of each model. T is the time variable. $LogPopBio(T)$ is the ln(abundance) of T. N_0 is initial population size, r_{max} is the maximum growth rate , and N_{Max} is carrying capacity (commonly denoted by K in the ecological literature).

Quadratic polynomial model

$$PopBio(T) = A_0 + A_1 T + A_2 T^2 \tag{1}$$

Cubic polynomial model

$$PopBio(T) = A_0 + A_1 T + A_2 T^2 + A_3 T^3 \tag{2}$$

Logistic model

$$PopBio(T) = \frac{N_0 N_{Max} e^{r_{max} t}}{N_{Max} + N_0 (e^{r_{max} t} - 1)} \tag{3}$$

Gompertz model

$$LogPopBio(T) = N_O + (N_{Max} - N_0) e^{-e^{\frac{r_{max} e^1 (t_{lag} - T)}{(N_{Max} - N_0) log(10)} + 1}} \tag{4}$$

2.3 Computing Tools

Python 3.10.6 was used to do data preprocessing and model fitting. The packages *pandas* and *numpy* were used to preprocess the data. The *matplotlib* and*seaborn* were used to plot the data. The *scipy*, *sklearn* and *lmfit* were used to fit models. And *warnings* was used to ignore the warnings. *Bash* was used to run all the Python and LaTex scripts.

2.4 Model Fitting

For NLLS model, especially Gompertz, it is very important to determine the starting value. As the Gompertz has quite a large number of parameters: N_0, N_{max}, r_{max}, t_{lag}, the probability of model convergence is relatively low. For partial ID prediction, the initial value of the model needs to be continuously adjusted to converge. However, Linear (OLS) models do not care the starting value.

2.5 Model Selection

Three statistical methods(AIC, AIC_c, R^2) were used in this study to measure the performance of the models. The following show the equations of them:

$$AIC = 2k - 2ln(L) \tag{5}$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \tag{6}$$

$$R^2 = 1 - \frac{SSR}{SST} \tag{7}$$

k is the number of parameters and L is the likelihood function. A small k implies a concise model and a large L implies an accurate model, so the model is evaluated with a balance of conciseness and accuracy. In the case of small samples, AIC transforms to AIC_c. So AIC_c can be applied to any sample size. R^2 is generally used in regression to assess how good or bad the model is. The closer the value is to 1, the better the model performance is, and less than 0, it usually means that the model is very poor.

3 Results

After filtering the data (by $DataPreparation$), I got 256 sets in total for the model fitting. I got each models' fitting results and output the AIC, AIC_c, R^2 distributions in $results$ directory. Besides, I ploted each set of fitting results in the $sandbox$ directory.

3.1 Model Fitting Count

In Figure 1, it showed that the OLS models fitted most of the data, especially the Cubic Polynomial. Then the Quadratic Polynomial also fitted well. For the mechanistic models, Gompertz model fitted quite well but Logistic model only fitted half of the data successfully. The mainly reason of this is that there were low count of data points in some sets, so it caused under fitting problems.

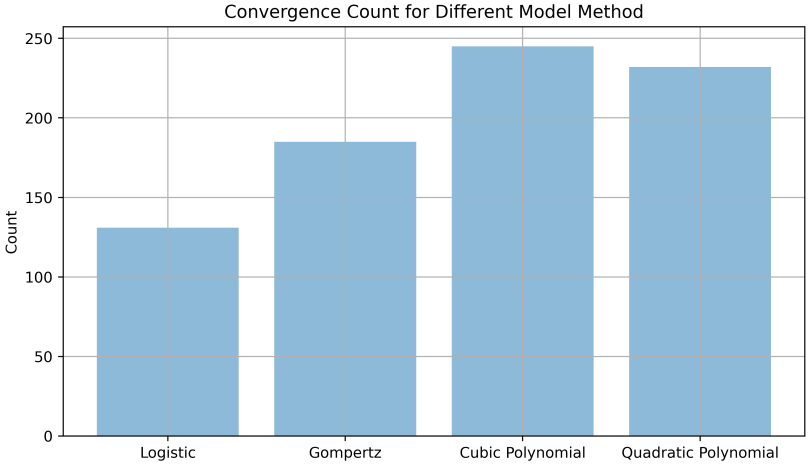


Figure 1: It showed the number of fitted models

3.2 Model Comparison

Figure 2 showed the R^2 distributions of models. Figure 3 shows the descriptive statistics of R^2. By comparing the R^2 of the models, it illustrated that the Cubic polynomial is the best.

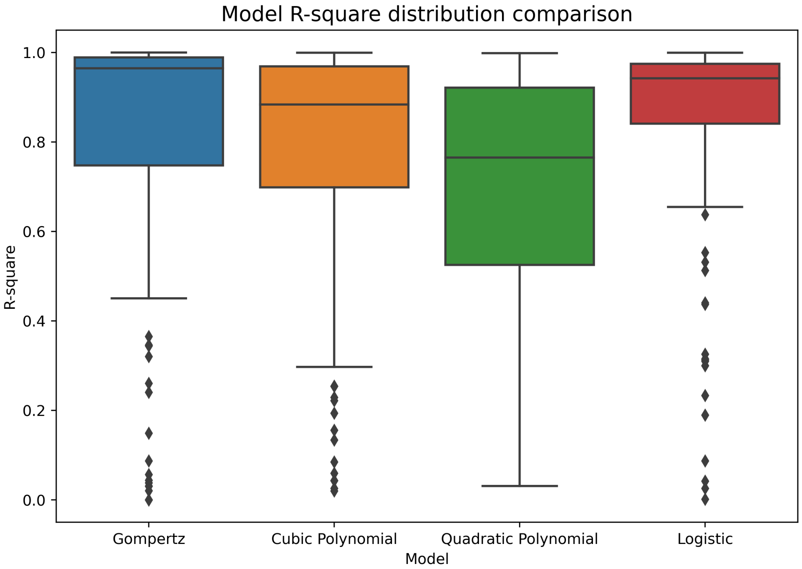
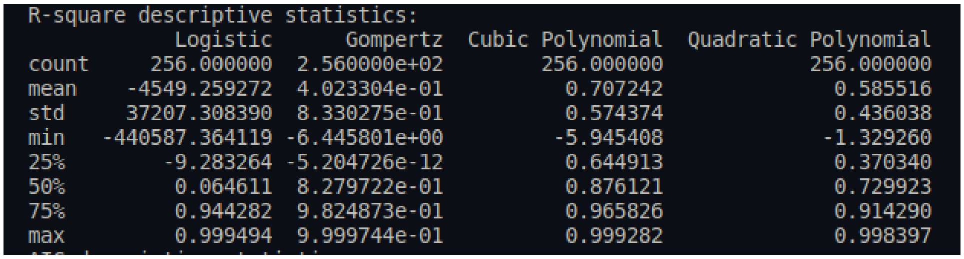


Figure 2: It showed the R^2 distribution



```
R-square descriptive statistics:
              Logistic      Gompertz  Cubic Polynomial  Quadratic Polynomial
count     256.000000  2.560000e+02        256.000000            256.000000
mean    -4549.259272  4.023304e-01          0.707242              0.585516
std     37207.308390  8.330275e-01          0.574374              0.436038
min   -440587.364119 -6.445801e+00         -5.945408             -1.329260
25%        -9.283264 -5.204726e-12          0.644913              0.370340
50%         0.064611  8.279722e-01          0.876121              0.729923
75%         0.944282  9.824873e-01          0.965826              0.914290
max         0.999494  9.999744e-01          0.999282              0.998397
```

Figure 3: It showed the R^2 descriptive statistics

Figure 4 showed the AIC distributions of models. Figure 5 shows the descriptive statistics of AIC. By comparing the $AICC$ of the models, it illustrated that the Gompertz is the best.

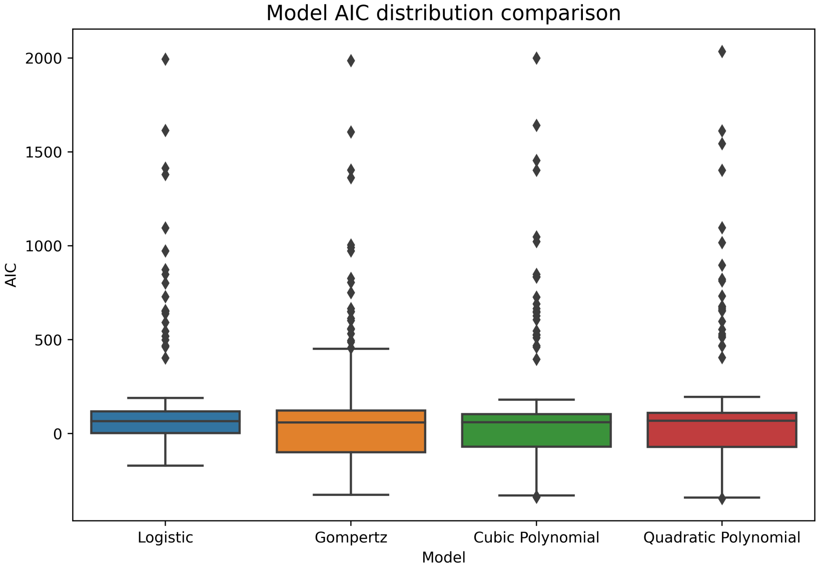
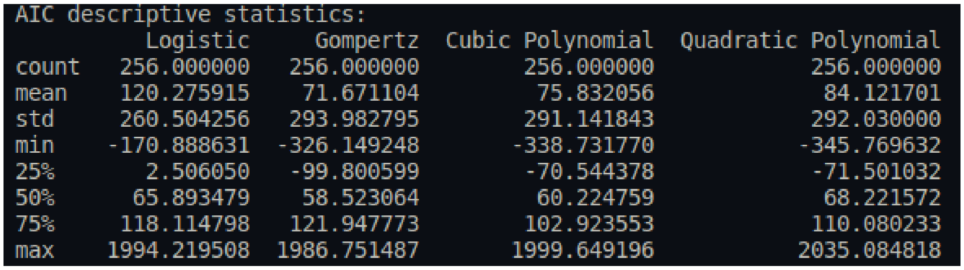


Figure 4: It showed the AIC distribution



```
AIC descriptive statistics:
          Logistic     Gompertz  Cubic Polynomial  Quadratic Polynomial
count   256.000000   256.000000        256.000000            256.000000
mean    120.275915    71.671104         75.832056             84.121701
std     260.504256   293.982795        291.141843            292.030000
min    -170.888631  -326.149248       -338.731770           -345.769632
25%       2.506050   -99.800599        -70.544378            -71.501032
50%      65.893479    58.523064         60.224759             68.221572
75%     118.114798   121.947773        102.923553            110.080233
max    1994.219508  1986.751487       1999.649196           2035.084818
```

Figure 5: It showed the AIC descriptive statistics

Figure 6 showed the AIC_c distributions of models. Figure 7 shows the descriptive statistics of AIC_c. By comparing the AIC_c of the models, it illustrated that the Gompertz is the best.

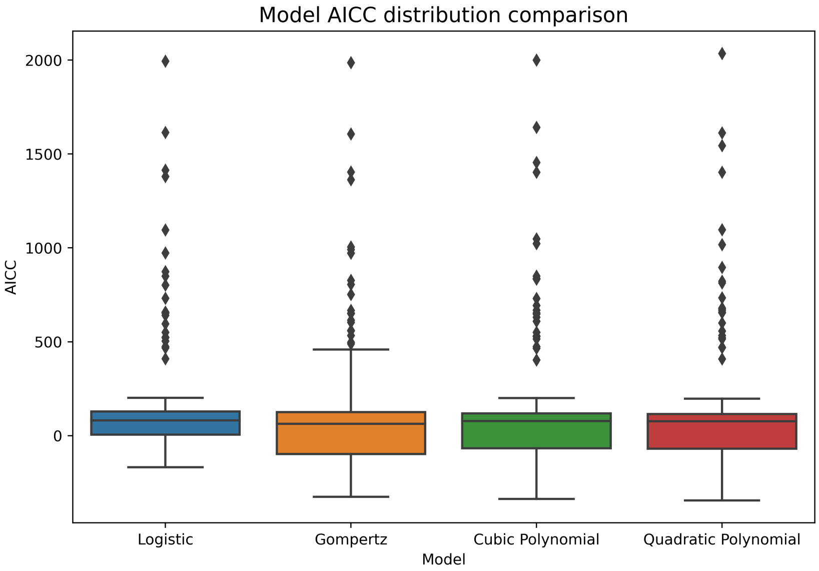
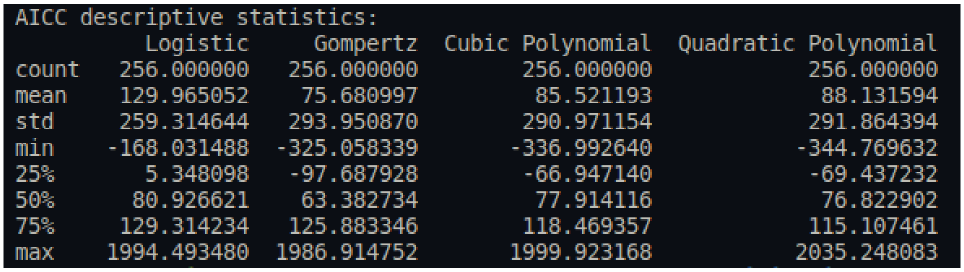


Figure 6: It showed the AIC_c distribution



```
AICC descriptive statistics:
          Logistic     Gompertz  Cubic Polynomial  Quadratic Polynomial
count   256.000000   256.000000        256.000000            256.000000
mean    129.965052    75.680997         85.521193             88.131594
std     259.314644   293.950870        290.971154            291.864394
min    -168.031488  -325.058339       -336.992640           -344.769632
25%       5.348098   -97.687928        -66.947140            -69.437232
50%      80.926621    63.382734         77.914116             76.822902
75%     129.314234   125.883346        118.469357            115.107461
max    1994.493480  1986.914752       1999.923168           2035.248083
```

Figure 7: It showed the AIC_c descriptive statistics

Figure 8 showed the AIC distributions of IDs. Figure 9 shows the AIC_c distributions of IDs. These two figures showed the performance of model fitting of different sets.

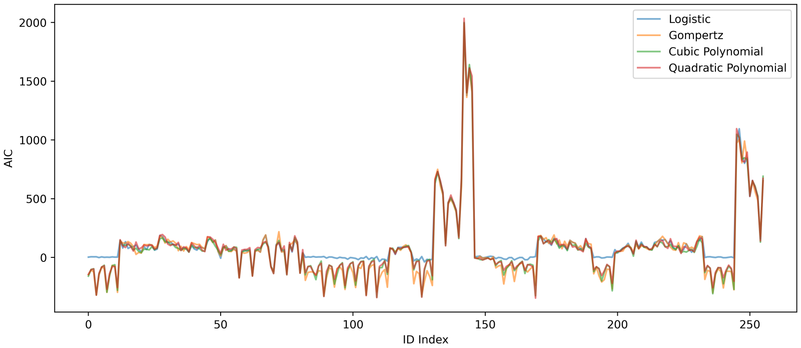


Figure 8: It showed the AIC distribution of IDs

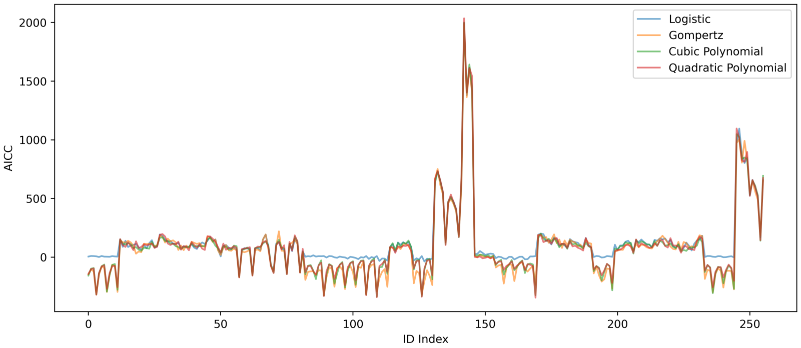


Figure 9: It showed the AIC_c distributions of IDs

4 Discussion

The main purpose of this project is to compare OLS with NLLS models and determine which models have the best performance on fitting the specific data (Microbial Population Growth Data). According to the study, the Cubic Polynomial and Quadratic Polynomial have few parameters with low model complexity, so they fit on most of the data sets. Meanwile, When using R^2, a basic statistical evaluation method, OLS stands out. Comparing the mean R^2 of the four models, $Cubic(0.702)$ is significantly larger than the other models and closest to 1. $Logistic(R^2 = -4549.25)$ performed terrible because of the huge amount of unfitted data. As those are machine learning models, if each set has more data points, it would greatly improve the accuracy and fitting ability of the models. However, apart from the non-fitted cases, the NLLS showed promising features. Whether comparing AIC or AIC_c, Gompertz model showed the best performance on fitting the data. In AIC, $Gompertz(71.67) < Cubic(75.83) < Quadratic(84.12) < Logistic(120.27)$. And in AIC_c, $Gompertz(75.68) < Cubic(85.52) < Quadratic(88.13) < Logistic(129.96)$ That is to say, Gompertz should be preferred if extreme performance is sought, as the model has more parameters and a relatively low probability of convergence, it may require more potential labor costs to adjust the initial values.

In conclusion, NLLS generally outperform OLS for quantifying microbial population growth data. Logistic model is sufficient for the simple growth case. Gompertz model is sufficient for the complex situation as it has the best performance. However,OLS models have advantages as well, they are suitable for simple cases. In order to get better performance for Gompertz, it takes more time to optimize it by adjusting the initial values.

References

[1] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.

[2] CLIFFORD M. HURVICH and CHIH-LING TSAI. Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika*, 78(3):499–509, 09 1991.

[3] Cesare Marchetti, Perrin S. Meyer, and Jesse H. Ausubel. Human population dynamics revisited with the logistic model: How much can be modeled and predicted? *Technological Forecasting and Social Change*, 52(1):1–30, 1996.

[4] Seiichi Sakanoue. Extended logistic model for growth of single-species populations. *Ecological Modelling*, 205(1):159–168, 2007.

[5] MH Zwietering, I Jongenburger, FM Rombouts, and K van 't Riet. Modeling of the bacterial growth curve. *Applied and environmental microbiology*, 56(6):1875—1881, June 1990.