Monte Carlo Simulating Approach Based Fitting Model Evaluation

Yuji Sode

Preface

YujiSode
2024.08.18に更新

このチャプターの目次

Preface

A misleading case of numerical models was reported by means of experimental approach. This report describes a case that a discrete dataset causes the theoretical bias using an example of evaluating the curve fitting analyses from a paleontological perspective.

本書では数値モデルにおいて異なる解釈が生じる場合を実験的に示した. 離散データセットが理論上のバイアスを生じるケースを曲線フィッティングの評価を例に古生物学的観点より解析した.

Abstract

Regression or curve fitting based analyses have been employed in paleontological researches to estimate the interspecific relation, the growth, the body plan, or the evolution of extinct animals. Qualities and quantities of sample are different depending on each case of study because of the complicated preserving processes. These methodological and paleontological facts indicate a systematic difference in inputs between real samples and several theoretical analyses. This study reports a case of a sample-condition tolerant method, which was able to evaluate sample-model interpolation. The theoretical experiment was conducted based on the randomly extracted dataset from a quadratic function and the random simulation in order to assess fitting models. The least squares method was employed to calculate four fitting models against the generated dataset. Following three consequences were obtained from this experiment. (1) The randomly extracted samples remained some of original characteristics of the predetermined function. (2) The value of reduced chisquares could not recognize the best fitting model. (3) The frequency pattern of the generated sample corresponded to a single model-pattern.

Keywords: numerical model, curve fitting, random simulation, bias, paleontology

---

License

Change log

- Released: `[1.2.2] - 2024-08-17`
- Released: `[1.0.0] - 2024-08-10`
- Released: `[0.1.0 beta] - 2024-08-06`

NEXT
Introduction

Introduction

YujiSode
2024.08.06に更新

Introduction

Paleontological samples are qualitatively and quantitively different depending on studies, and this sample-heterogeneity requires methodological approach to evaluate sample-model interpolation. Mustoe and Smith (2023) reported detailed preservation of plants and animals based on clay-rich and fine-grained sediment-quality and opalization. Wani (2003) classified Cretaceous ammonoid assemblage into four facies by shell preservation and diversity. It was also reported that background probability distributions caused compositional differences in Cretaceous mollusk assemblages from Hokkaido, Japan (Sode and Okamoto, 2014; Sode and Okamoto, 2015). These investigations indicate that samples in paleontology appear discreteness in data; however, theoretical analyses are methodologically classified into discrete or continuous method. Hence, this study reports a new theoretical result on an evaluation of the sample-model interpolation based on a sample-condition tolerant method. Furthermore, studies between sample and theory can be abstracted as a data-fitting problem, hence it is important for evaluation of fitting results to undestand characteristics of both sample materials and models to apply. Nevertheless, it is also true that results of curve fittings or regression are the most likely outputs, which are determined under the given constraints and model (Figure 1).

The curve-fitting analysis is a interpolating method to estimate a relation between sample materials, such as Tanabe *et al.* (1979) and Tsujino *et al.* (2003). Tanabe *et al.* (1979) investigated the interspecific analysis regarding ontogenetic deveropments in the early phases of ammonites through observations in the median section of shells. They regressed the size of the first shell growth, which is from the protoconch to the nepionic constriction, called ammonitella against the diameter of the most primitive shell called protoconch by employing the least squares method. Tsujino *et al.* (2003) estimated allometric shell-growth of fragmented Japanese-ammonite called *Baculites tanakae* by means of three results: observation of shell ornament, regression of the height on the fragmentary-sample length based on the least squares method, and shell expansion that was on the basis of a linearly regressed result between height and length using the major axis method. Moreover, there are some studies that indicate sample-theory agreement, and Okamoto (1984), Okamoto (1988), and Okamoto and Shibata (1997) are good examples. Okamoto (1984) and Okamoto (1988) revealed theoretical shell growth of a heteromorphic-ammonite genus called *Nipponites*, and calculation for some physical properties (e.g. shell length, shell volume, and surface area) with complicated shell forms became available in some of the extinct cases. Okamoto and Shibata (1997) pointed ammonoids called *Polyptychoceras pseudogaultinum* had lived as nekton, plankton and/or benthos.

A random process based analysis is one of discrete approaches that can accept both discrete samples and continuous systems. This is because some analyses that have stochastic steps within themselves do not require arrangement of inputs. According to Motwani and Raghavan (1995), concept of algorithm using randomness can be traced back to the "Monte Carlo methods" used in numerical analysis and simulation. The bootstrap was introduced by Efron (1979) as a primitive interpretation of jackknife, in order to estimate sample distribution from an unknown probability distribution. Therefore, this study applied the random simulating approach through the bootstrap method to the evaluation of the sample-model interpolation.

PREV
Preface

NEXT
Materials and Methods

Materials and Methods

YujiSode
2024.08.07に更新

このチャプターの目次

Materials and Methods

The theoretical experiment was conducted based on Error Model, Sample Dataset Generation, Fitting Models, Value for Evaluation, and Random Process Based Analysis.

Error Model

An error model Q was introduced to simulate deviation in discrete data, based on the binary branching process with probability of 0.5 (Yotsuji, 2010). The error model Q was expressed using a random variable u and constant value p

$$Q(p) = 1 + p(2u - 1), \text{ where } u = [0, 1] \text{ and } p = [0, 1]. \tag{1}$$

A model with deviation F was simulated with using variable x as a product of predetermined function f and Q

$$F(x) = f(x)Q(p). \tag{2}$$

Sample Dataset Generation

A sample data model $F(x)$ was defined with variable x and deviation with twenty percent

$$F(x) = f(x)Q(0.2), \text{ where } f(x) = (3x - 2)^2. \tag{3}$$

A sample dataset was randomly generated as a hundred-set of discrete points through the formula 3.

Fitting Models

Four fitting models $f(x)^n$ against the generated sample data set were estimated using the least squares method with gnuplot version 5.2 (Williams, Kelley, *et al.*, 1986-1993, 1998, 2004, and 2007-2019)

$$f(x)^n = (Ax + B)^n, \text{ where } n = 1, 2, 3, \text{ and } -1. \tag{4}$$

Value for Evaluation

A dL/dx was used in the random-simulation as a variable that was able to distinguish the numerical-models (Formula 4) to fit the generated dataset, which was defined with the formula 3. This dL/dx was defined based on two values, and they are the approximated partial curve length dL with length L between two points and the horizontal change dx (Figure 2) as follows:

$$dL = \sum L - \max(L) \tag{5}$$

and

$$dx = \max(x) - \min(x). \tag{6}$$

Random Process Based Analysis

Bootstrap Method

Bootstrapping (Efron, 1979) is a practical method for estimating sample-distributional properties of an unknown or specified probability distribution. This method can be classified into two types based on random-resampling processes: nonparametric or parametric method. The nonparametric bootstrap method is available for estimation when the target distribution is unknown (e.g. discrete dataset), and the parametric one is available when the target can be specified by some parameters (e.g. numerical model). Consequently, this study applied a nonparametric bootstrap method to discrete generated-samples and a parametric bootstrap method to continuous systems in order to evaluate the sample-model interpolation. According to Efron (1979), the estimation is achieved based on a specified distribution that is called bootstrap distribution by random resampling, and each resampled specimen is called bootstrap sample. The value of dL/dx was selected as a bootstrap sample in order to estimate sample-distributional property for this experiment.

Monte Carlo Approximation

Bootstrap distribution was approximated through 1.0E2 sets of Monte Carlo simulations. Each simulation set had 1.0E4 sample sized ternary samples (Figure 2).

PREV
Introduction

NEXT
Results and Discussion

Results and Discussion

YujiSode
2024.08.07に更新

このチャプターの目次

Results and Discussion

Figure 3 shows results of the sample dataset generation and four different curve fitting models (Formula 4). The all plots represent a fixed dataset that is generated based on the formula 3. Four curves indicate different fitting models (Formula 4) that are calculated with the least squares method against this fixed dataset. Table 1 summarizes the parameter estimations by the least squares method. The details of the generated dataset is shown in Table 2. Every applied model indicated an increasing pattern (Figure 3). According to the values of reduced chisquares in the Table 1, a fitting model of $f(x)^{-1}$ had a distinctive trend, which was far from the fixed dataset. Three other models shared similarities in the graph patterns and the values of reduced chisquares (Figure 3 and the Table 1); however, according to gnuplot 5.2 (Williams, Kelley, *et al.*, 1986-1993, 1998, 2004, and 2007-2019), the reduced-chisquares, which exceed one, means outliers, systematic errors, errors that are non-normally distributed, or inappropriate error-estimation or fitting function. Nevertheless, it is true that a fitting model that has the quadratic form is an appropriate estimation due to the formula 3, which predetermined the functional form as $(3x-2)^2$. This contradiction between the value of reduced chisquares, which indicated an indeterminable result, and the sample data model, which had a determined solution, was due to a common methodological process and the predefined deviation. The least squares method minimizes a total distance between estimations and observations, and the value of reduced chisquares is also calculated based on each distance; additionally, these distances can be expanded by the deviation in the observed data.

Figure 4 shows a result that is expressed in frequencies of the respective measurement (Formula 5 and 6) of the sample dataset and four fitting models (Formula 4) through the random simulation. These frequencies were expressed by averages and standard deviations in this graph. Every frequency converged to a distinct distribution with unimodalty. The frequency that was derived from the fitting model of $f(x)^{-1}$ had a dissimilar shape of distribution to the other three. This distribution had a peak at the value between 1 and 2 whereas the others appeared the local maximum at the value between 3 and 5. The fitting model $f(x)^1$ showed the highest peak of the four local maximums; $f(x)^2$ and $f(x)^3$ had the second and third highest local maximums in their respective order. Figure 4 indicates that the measurements with the approximated partial curve length (Formula 5) and the horizontal change (Formula 6) are able to distinguish the four fitting models (Formula 4) based on standard deviations.

According to a frequency distribution of the sample-dataset (Figure 4), the observed peak height was between the first and third highest ones, and its distributional pattern had similarity to that of the two fitting models: $f(x)^2$ and $f(x)^3$. Moreover, the most appropriate model for the discrete dataset is determined as $f(x)^2$ since every peak derived from respective fitting models (Formula 4) is identified (Figure 4). This suggests that the discrete sample still remains some of nonlinear characteristics that are predetermined in the parent system. Consequently, the data fitting by $f(x)^3$ is evaluated as an overestimation (Figure 4).

PREV
Materials and Methods

NEXT
Conclusions

Conclusions

YujiSode
2024.08.12に更新

Conclusions

A case of the sample-condition tolerant approach that could evaluate sample-model interpolation was herein reported. This introduced method was based on the random simulation and a variable that could clearly outline the predetermined numerical-models. The experiment reached following three consequences. (1) The discrete dataset was randomly extracted from a model based on a quadratic function that was accompanied with twenty-percent deviation, and this dataset kept some of original characteristics of the predetermined system. (2) The value of reduced chisquares failed to recognize the best fitting model from four types for the reason that every estimation of fitting function is the most likely consequence based on the given conditions. (3) The frequency pattern of generated sample corresponded to a single model-pattern because every predetermined variable, which was calculated by the approximated partial curve length and the horizontal change, showed a convergence to the respective distinct-distributions depending on the fitting models. Finally, analyses of natural phenomena are usually established by some estimations and theories; however, it is a case that some key materials for investigation have a dissimilarity to the others in quality and/or quantity. The above results indicate the further understanding of a dataset can improve the fitting model to apply.

About This Book

Contents of this book were reorganized based on a report (Sode, 2024a), which was introduced on the 2024 Annual Meeting of the Palaeontological Society of Japan, and its supplementary materials (Sode, 2024b) with additions and corrections.

Acknowledgements

Prior to publish this book, the author appreciates the occasion and advice that were offered by Dr. Takashi Okamoto (Ehime University) in studying paleontlogy and statistics at Ehime University. The author would like to sincerely thank Dr. Yoshida Katsuhiko (National Institute for Environmental Studies, Japan) for his constructive discussions. The author would also like to thank Dr. Kentaro Izumi (Chiba University) for his encouraging discussion. The author is especially grateful to Dr. Rie S. Hori (Ehime University) for her patient mentoring. The author would also like to appreciate the kind encouragement, which pushed a decision to apply for the annual meeting of the Palaeontological Society of Japan, from the Intensive English Program Office of Global Engagement at North Carolina State University.

Zenn

Monte Carlo Simulating Approach Based Fitting Model…

Chapter 06 github.devで開く ↗

References Cited

YujiSode
2024.08.08に更新

このチャプターの目次

References Cited

Efron, B., 1979. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist*, vol. 7, no. 1, p. 1-26.

Motwani, R. and Raghavan, P., 1995. Randomized algorithms. Cambridge University Press. Printed in New York, United States of America. ISBN 0-521-47465-5.

Mustoe, G., E., and Smith, E., T., 2023. Timing of Opalization at Lightning Ridge, Australia: New Evidence from Opalized Fossils. *Minerals*, vol. 13, (12), 1471. https://doi.org/10.3390/min13121471

Okamoto, T., 1984: Theoretical morphology of *Nipponites* (a heteromorph ammonoid). *Fossils*, (36), p. 37-51. (in Japanese with English abstract).

Okamoto, T., 1988: Analysis of heteromorph ammonoids by differential geometry. *Palaeontology*, vol. 31, part 1, p. 35-52.

Okamoto, T. and Shibata, M., 1997: A cyclic mode of shell growth and its implications in a Late Cretaceous heteromorph ammonite *Polyptychoceras pseudogaultinum* (Yokoyama). *Paleont. Res.*, vol. 1, (1), p. 29-46.

Sode, Y. and Okamoto, T., 2014: Compositional Analysis of "*Baculites* facies" and its Biological Significance. *14th Meeting, Geol. Soc. Japan, Shikoku Branch, Abstract*, p. 9, P-1 (in Japanese).

Sode, Y. and Okamoto, T., 2015: Is "*Baculites* facies" biologically significant? *164th Regular Meeting, Palaeont. Soc. Japan, Abstract*, p.44 (in Japanese).

Sode, Y., 2024a: Applicability of Random Simulation to the Evaluation of Interpolation. *2024 Annual Meeting. Palaeont. Soc. Japan, Abstracts with Programs*, P72-C, p. 53.

Sode, Y., 2024b: Sode2024_AnnualMeeting_PalaeontSocJp-supplementaryData. The contents was derived on 2024-07-24 from: https://github.com/YujiSODE/Sode2024_AnnualMeeting_PalaeontSocJp-supplementaryData

Tanabe, K., Obata, I., Fukuda, Y. and Futakami, M., 1979: Early Shell Growth in Some Upper Cretaceous Ammonites and Its Implications to Major Taxonomy. *Bull. Natn. Sci. Mus.*, Ser. C (Geol.), vol. 5 (4), p. 153-176.

Tsujino, Y., Naruse, H., and Maeda, H., 2003: Estimation of allometric shell growth by fragmentary specimens of *Baculites tanakae* Matsumoto and Obata (a Late Cretaceous heteromorph ammonoid). *Paleont. Res.*, vol. 7, (3), p. 245-255.

Wani, R., 2003: Taphofacies models for Upper Cretaceous Ammonoids from the Kotanbetsu area, northwestern Hokkaido, Japan. *Palaeogeography*, *Palaeoclimatology*, *Palaeoecology*, vol. 199, p. 71-82.

Williams, T., Kelley, C., *et al.*, 1986-1993, 1998, 2004, and 2007-2019: Gnuplot 5.2: An Interactive Plotting Program. Organized by Merritt, E., A., et al. The contents was derived on 2023-12-09 from: http://www.gnuplot.info/

Yotsuji, T., 2010: Probability distribution random number generation method for computer simulation (in Japanese). Pleiades PUBLISHING Co.,Ltd. ISBN: 978-4903814353.

PREV
Conclusions

NEXT
Figure Captions 1

Figure Captions 1

YujiSode
2024.08.08に更新

Figure Captions

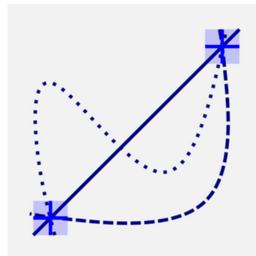

Figure 1. Schematic diagram showing three interpolations based on different models.
Every curve interpolates correctly the fixed two points under the respective given conditions.

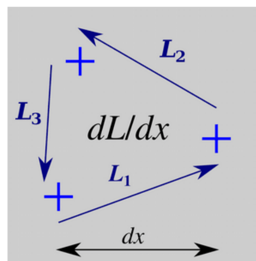

Figure 2. Schematic diagram of dL/dx.
This diagram shows vector relations of dL/dx between three points. A length L_i is expressed as a norm of a vector; $L_i = \|\boldsymbol{L}_i\|$, where $i = 1, 2, \text{ and } 3$.

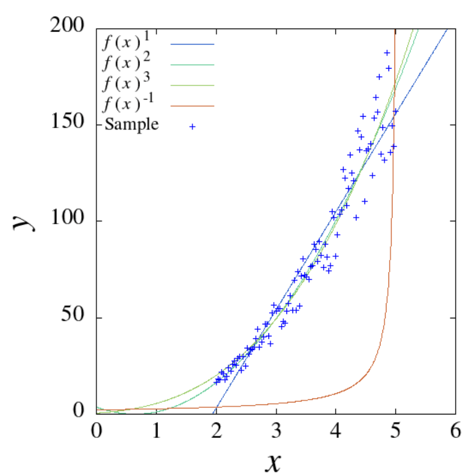

Figure 3. Graphs showing sample data and results of curve fittings.
Every point is plot of the generated data set from a model, which is $(3x - 2)^2$ with 20 percent of deviation. Every line is result of curve fitting using the least squares method.

PREV
References Cited

NEXT
Figure Captions 2

Figure Captions 2

YujiSode
2024.08.08に更新

このチャプターの目次

Figure Captions

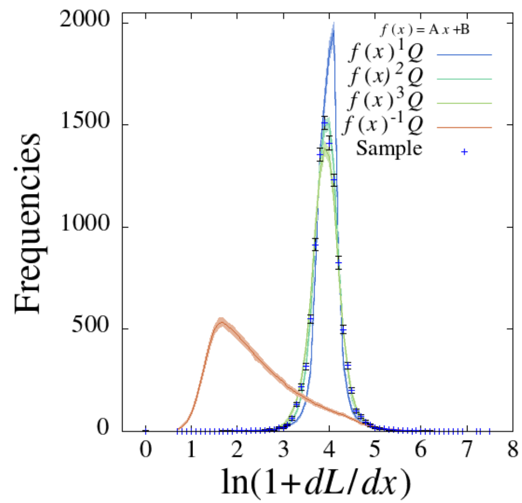

Figure 4. Graphs showing frequency distributions of $\ln(1+dL/dx)$.
This graph shows frequency distributions of $\ln(1+dL/dx)$, which are defined using 3 random points on a given function $f(x)$ or generated sample data. Every line is result of given function $f(x)$. Every point is result of generated sample data. Each graph represents average of a 100 sets of results from random simulation. Filled areas or error bars represent standard deviations.

Table 1. Results of the least squares method.
A and B are the estimated parameters for the fitting models against the generated sample data set.

Fitting models	A	B	Variance of residuals (reduced chisquares)
$(Ax+B)^{-1}$	-08.97E-02	45.23E-02	5612.88
$(Ax+B)^{1}$	50.97E+00	-98.81E+00	156.49
$(Ax+B)^{2}$	02.99E+00	-01.92E+00	119.89
$(Ax+B)^{3}$	95.01E-02	81.96E-02	125.26

PREV
Figure Captions 1

NEXT
Figure Captions 3

Figure Captions 3

YujiSode
2024.08.08に更新

このチャプターの目次

Figure Captions

Table 2. List of a generated sample dataset.
This dataset was randomly derived from a numerical model, which is $f(x)Q(0.2)$, where $f(x)=(3x-2)^2$, $Q(p)=1+p(2u-1)$, $u=[0,1]$, and $x=[2,5]$. The list of dataset was modified after Sode (2024b).

No.	x	y	No.	x	y	No.	x	y	No.	x	y
001	2	16.691	026	2.75758	37.3259	051	3.51515	71.3767	076	4.27273	125.321
002	2.0303	18.0962	027	2.78788	40.2827	052	3.54545	69.7152	077	4.30303	121.333
003	2.06061	17.8231	028	2.81818	46.9649	053	3.57576	76.439	078	4.33333	102.177
004	2.09091	21.8137	029	2.84848	46.7971	054	3.60606	77.235	079	4.36364	147
005	2.12121	20.9826	030	2.87879	40.4401	055	3.63636	88.1467	080	4.39394	137.257
006	2.15152	17.8059	031	2.90909	36.5564	056	3.66667	85.6039	081	4.42424	143.979
007	2.18182	19.6628	032	2.93939	52.4278	057	3.69697	79.4776	082	4.45455	154.503
008	2.21212	24.0979	033	2.9697	56.6774	058	3.72727	89.3861	083	4.48485	110.6
009	2.24242	22.0729	034	3	53.466	059	3.75758	82.6085	084	4.51515	136.836
010	2.27273	25.8463	035	3.0303	54.5675	060	3.78788	76.1514	085	4.54545	137.672
011	2.30303	27.7232	036	3.06061	54.844	061	3.81818	88.1289	086	4.57576	140.252
012	2.33333	25.2068	037	3.09091	45.3382	062	3.84848	81.4216	087	4.60606	124.032
013	2.36364	28.7694	038	3.12121	48.3018	063	3.87879	74.299	088	4.63636	153.459
014	2.39394	30.0033	039	3.15152	47.0636	064	3.90909	77.0904	089	4.66667	164.565
015	2.42424	22.8036	040	3.18182	53.3057	065	3.93939	105.073	090	4.69697	156.954
016	2.45455	29.7141	041	3.21212	57.6562	066	3.9697	102.012	091	4.72727	175.155
017	2.48485	25.0227	042	3.24242	61.5497	067	4	82.1583	092	4.75758	134.92
018	2.51515	34.2965	043	3.27273	53.9111	068	4.0303	93.048	093	4.78788	148.861
019	2.54545	31.2814	044	3.30303	69.6253	069	4.06061	103.806	094	4.81818	131.759
020	2.57576	33.3659	045	3.33333	54.0948	070	4.09091	106.162	095	4.84848	187.39
021	2.60606	33.6576	046	3.36364	73.881	071	4.12121	126.854	096	4.87879	179.724
022	2.63636	34.9513	047	3.39394	56.325	072	4.15152	122.542	097	4.90909	135.733
023	2.66667	39.0302	048	3.42424	71.704	073	4.18182	108.372	098	4.93939	149.813
024	2.69697	43.9933	049	3.45455	80.43	074	4.21212	117.325	099	4.9697	138.785
025	2.72727	34.8157	050	3.48485	71.9485	075	4.24242	134.415	100	5	157.063

YujiSode
興味のある分野はJavaScript, Tcl, 数学, 地質学, 生物学などです.

PREV
Figure Captions 2