Earthline Journal of Mathematical Sciences
ISSN (Online): 2581-8147
Volume 2, Number 2, 2019, Pages 491-503
https://doi.org/10.34198/ejms.2219.491503

Evaluation of Methods of Interpolation (Existing and Proposed) using Monte Carlo Simulation Technique and Delete-D Jackknife Method

K. A. Awopeju[1,*], B. F. Ajibade[2], R. A. Efeizormor[3] and B. E. Omokaro[4]

[1]Department of Statistics, Nnamdi Azikiwe University, Awka, Anambra State, Nigeria
e-mail: ak.awopeju@unizik.edu.ng

[2]Petroleum Training Institute, Effurun, Warri, Delta State, Nigeria

[3]College of Education, Agbor, Delta State, Nigeria

[4]Department of Statistics, Delta State Polytechnic, Otefe, Delta State, Nigeria

[*]Corresponding author

Abstract

It is possible to encounter missing value in a research. Missing value may be as a result of none response in primary data collection or unavailability of data in the case of secondary data. It may occur within a set of observations or at the tail end of the observations. The paper addresses missing value within a set of observations (interpolation). Existing methods considered are linear, log-linear, Catmull-Rom spline and cardinal spline and a method of estimating missing value is presented. For evaluation of the methods, random data are simulated using Monte Carlo simulation approach and analytical approach is used to determine the most effective method. Among the findings, linear, log-linear and the proposed method give high precision estimate compare to the CS and CSR methods. With the use of tension parameter, CS is better than CSR method.

1. Introduction

Missing values can be estimated using either interpolation or extrapolation procedure

Received: May 31, 2019; Accepted: August 3, 2019

2010 Mathematics Subject Classification: 62-XX.

Keywords and phrases: simulation, precision, interpolation, extrapolation, linearity, curvature.

depending on the position of the missing value(s). It is referred to as extrapolation if the observations are to be estimated beyond data collected but interpolation if the missing value is within a set of observations. Therefore, interpolation procedure for a series fills missing values within a series by interpolating from values that are non-missing.

A number of factors can lead to missing value in a research. This implies the reason for missing data may not be correlated with the idiosyncratic. In survival analysis, missing record of a patient may be as a result of death or recovery from the ailment. This was taken care of in censoring. Censoring may be to the right or left as the case may be. In primary data collection, non-response may be classified as missing value as the respondent has the sole right to supply information to be used for the research (Gelman and Hill [2]). Often, respondents fail to provide answers to certain questions, which lead to missing data for the dependent or independent variables. Some researchers recommended total drop of the variable because such variable or response may lead to biasness of the estimators. Missing values in data can hamper a research and prevent meaningful conclusion and can also make panel data to be unbalanced (Yen and Jones [8]).

Natural hazard can also be a cause for missing value is record keeping or data collection. A country terribly hit by natural hazard may not have accurate financial record for the period of the hazard. Countries at war may find it very difficult to keep accurate record in terms of activities in the country. Considering Nigeria, in the year 1967-1969, the country was hit by civil war which paralyses economic activities through out the period. Nevertheless, the financial institutions in the country have records for financial activities for the period (see CBN Bulletin). Such information may not be accurate and difficult to verified. A noble way of recording such information is to report the observations as missing value(s). Therefore, in analytical study, there may be need to complete the record for the use of mathematical expressions.

A number of researchers have worked to fill the gap in data collection which led to some existing methods of estimating extrapolation and interpolation, see Fung [5], Cheema [3], Iwueze et al. [6]. Some schools of taught agreed that extrapolation should be done using regression approach (time series regression), (Iwueze et al. [6]) or moving average method while some depend on ordinary averages (Mahir and Al-Khazaleh [1]). State-space model provides a flexible approach to time series analysis, especially for simplifying maximum likelihood estimation and handling missing values (Tsay [7]).

Also, in the estimation of interpolation, some of the commonly used methods are linear method, log-linear method, Catmull-Rom spline method, cubic spline method and cardinal spline method which were formulated by researchers in the past. The methods have received acceptability by researchers especially in econometrics and statistics but there is always room for improvement due to advancement in the activities of man.

This paper reviews the methods of estimating missing value within set of observations (interpolation), compares for efficiency, that is, accuracy and then proposes an alternative method of interpolation.

2. Methodology

The data are assumed to be measured at equidistance and discrete in nature. Let y_i be the measures such that $i = 1, 2, 3, ..., n,$ where n is the total data points. Given bivariables x_i and y_i, the value of any point y can be estimated at a given point x in-between the two points.

Linear interpolation technique: The method computes a linear approximation value based on the previous and next non-missing value in the series of interest. Mathematically, the interpolation can be done using the expression

$$IV_{lin} = (1 - \gamma) P_{i-1} + \gamma P_{i+1},$$

where P_{i-1} is the previous value before the missing value, P_{i+1} is the next value after the missing value, γ is the position of the missing value divided by the number of missing values in a row.

If only one missing value exists, the interpolation will be done considering the missing value half way between the previous and the next value. If two missing values exist, the first will be computed as one-third of the distance between the previous value and the next value, the second value will be computed as two-third of the distance.

Log-linear method: The method is similar to linear method but the series are logged before the computation is made and the interpolated value is exponentiated to get the natural number. In the case of negativity in the series, then, absolute values are considered which is one of the shortcomings of the method as logarithm of negative number cannot be estimated. Mathematically, the expression is

$$IV_{pos} = Exp[(1-\gamma)\log(P_{i-1}) + \gamma\log(P_{i+1})]$$

and

$$IV_{neg} = -Exp[(1-\gamma)\log(-P_{i-1}) + \gamma\log(-P_{i+1})].$$

The first expression is used for positive series and the second is used for negative series. For both linear and log-linear methods, there must be previous and next non-missing value(s).

Cardinal spline method: The method is based on the previous two non-missing values and the next non-missing values. This implies for the usage of cardinal spline method, the missing value must be located among non-missing values (at least four, two before and two after the missing value). The method fits the missing data to a non-linear or curved pattern. Denoting P_{i-1} and P_{i-2} as the previous two non-missing values and P_{i+1} and P_{i+2} as the numbers proceeding the missing value. The interpolated value is calculated as:

$$IV_{cs} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (1-t)(\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2})$$
$$-(2\gamma^3 - 3\gamma^2)P_{i+1} + (1-t)(\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}),$$

where t is called the tension parameter and affects the curvature of the spline. The shortcoming of the method is that the missing value must have two values before and after, otherwise, the method cannot be applied.

Catmull-Rom spline: The method is a special case of the cardinal spline with the tension parameter, t, set as zero. Therefore, the expression becomes

$$IV_{cr} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (1-0)(\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2}) - (2\gamma^3 - 3\gamma^2)P_{i+1}$$
$$+ (1-0)(\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}),$$

$$IV_{cr} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2}) - (2\gamma^3 - 3\gamma^2)P_{i+1}$$
$$+ (\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}).$$

For cardinal spline and Cardinal-Rom Spline, the methods will return missing value if P_{i-1} or P_{i+1} is missing. Note that if P_{i-2} is missing, it is set equal to P_{i-1} and if P_{i+2} is missing, it is set to equal to P_{i+1}.

Proposed Method

It is possible to have missing value in the middle of series or rightly or leftly shift. It is also possible to have missing value at the beginning of a series. If the missing value is at the tail end of the series, then it can be calculated using extrapolation methods. Since the paper centralizes on extrapolation, therefore, missing value(s) at the end of a series is not considered.

Case 1

Given a set of observations $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ and x_9. Let us assume the missing value is within the set of observations, say x_6, then the estimate of $\hat{x}_6$ is

$$\hat{x}_6 = \frac{x_4 + x_5 + x_7 + x_8}{4} = \frac{1}{4}[x_4 + x_5 + x_7 + x_8].$$

The difference between x_6 and $\hat{x}_6$ is that the first is actual value and the second is an estimated of the first. The closeness of the values implies accuracy of technique used. The closer the actual value to the estimate, the better the technique.

In the above expression, two values before the missing value and two values after the missing values are considered.

If the missing value is located at point 4, data points 2, 3, 5 and 6 will be used for the estimate. If the missing value is at the point 2, it becomes a different approach but similar. See Case 2.

Case 2

Given a set of observations $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$ and x_9. Let us assume the missing value is within the set of observations, say x_2, then the estimate of $\hat{x}_2$ is

$$\hat{x}_2 = \frac{0 + x_1 + x_3 + x_4}{4} = \frac{1}{4}[0 + x_1 + x_3 + x_4].$$

If the missing value is at position 1, the computation or usage of the series may start from point 2 and if the missing value is at the position 6, the usage of the data may stop at point 5.

Case 2 is appropriate when it is not possible to have two data points before the missing value.

Generally, for a missing value, the estimate can be computed using the expression

$$x_i = \frac{1}{4}[x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}].$$

It is possible for x_{i-2} to be zero, a case where the missing value is data point 2. Then, the expression becomes

$$x_i = \frac{1}{4}[x_{i-1} + x_{i+1} + x_{i+2}].$$

This is because x_{i-2} is replaced with zero.

3. Evaluation of the Five Methods of Interpolation

It should be noted that resulting value from the use of the expressions for interpolation is an estimate. For the evaluation of a method, the closeness of the estimate to the actual value shows the precision of the estimate. This implies for proper evaluation, the missing value must be known and then estimated. The most appropriate method for such evaluation is delete-d jackknife method where d represents number of observations to be elected from the series. In this case, d equals 1 to have just one missing value. From the literature, the position of d can be randomly selected. Monte Carlo simulation method can also be used to have repeated observations.

A. Single data series set of observations with missing value centralized

Given the data set below, the data point five is removed (deleted), provide the estimate using the methods of interpolation stated

(i) 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10

(ii) 2, 4, 6, 8, 10, 12, 14, 16, and 18

(iii) 1, 4, 6, 10, 12, 14, 18 and 20.

In the above set of observations, the first series has constant difference of 1 the second set of observations has constant difference of 2 while the third series has varies difference. See Robert and Casella [4].

As instructed, the series become

(1) 1, 2, 3, 4, *x*, 6, 7, 8, 9, and 10

(2) 2, 4, 6, 8, x, 12, 14 , 16 and 18

(3) 1, 4, 6, 10, x, 14, 18 and 20.

Note that the missing values are 5, 10 and 12 respectively. Best method among the five methods will provide the closest estimate. The methods are linear (L), log-linear (LL), cardinal spline (CS), Catmull-Rom spline (CRS) and the proposed method (PM).

The expressions are

L Method $IV_{lin} = (1-\gamma)P_{i-1} + \gamma P_{i+1}$

LL method $IV_{pos} = Exp[(1-\gamma)\log(P_{i-1}) + \gamma \log(P_{i+1})]$

CS Method $IV_{cs} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (1-t)(\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2})$

$$- (2\gamma^3 - 3\gamma^2)P_{i+1} + (1-t)(\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}),$$

CSR Method $IV_{cr} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2})$

$$- (2\gamma^3 - 3\gamma^2)P_{i+1} + (\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}),$$

PM Method $x_i = \frac{1}{4}[x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}].$

Using L Method

$$IV_{lin} = (1-\gamma)P_{i-1} + \gamma P_{i+1}$$

Since only one value is missing, the value of γ is 0.5. Therefore, $(1-\gamma)$ is also 0.5.

For the first series, $P_{i-1} = 4$ and $P_{i+1} = 6$.

Therefore, the estimate is $IV_{lin} = (1-0.5)4 + 0.5(6) = 5$.

For the second series, $P_{i-1} = 8$ and $P_{i+1} = 12$.

Therefore, the estimate is $IV_{lin} = (1-0.5)8 + 0.5(12) = 10$.

For the third series, $P_{i-1} = 10$ and $P_{i+1} = 14$.

Therefore, the estimate is $IV_{lin} = (1-0.5)10 + 0.5(14) = 12$.

Using LL Method

$$IV_{pos} = Exp[(1-\gamma)\log(P_{i-1}) + \gamma\log(P_{i+1})].$$

This is similar to L method where γ is 0.5.

For the first series, $P_{i-1} = 4$ and $P_{i+1} = 6$.

Therefore, the estimate is $IV_{pos} = Exp[(1-0.5)\log 4 + 0.5\log 6] = 4.89898 \cong 5$.

For the second series, $P_{i-1} = 8$ and $P_{i+1} = 12$

Therefore, the estimate is $IV_{pos} = Exp[(1-0.5)\log 8 + 0.5\log 12] = 9.79796 \cong 10$.

For the third series, $P_{i-1} = 10$ and $P_{i+1} = 14$.

Therefore, the estimate is $IV_{pos} = Exp[(1-0.5)\log 10 + 0.5\log 14] = 11.8322 \cong 12$.

Using CS Method

$$IV_{cs} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (1-t)(\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2})$$
$$-(2\gamma^3 - 3\gamma^2)P_{i+1} + (1-t)(\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}).$$

Set the tension parameter at 0.5, then $1 - t = 0.5$.

For the first series, the values of P_{i-1}, P_{i-2}, P_{i+1} and P_{i+2} are 4, 3, 6 and 7 respectively.

By substitution,

$$IV_{cs} = (2(0.5)^3 - 3(0.5)^2 + 1)4 + 0.5[(0.5)^3 - 2(0.5)^2 + 0.5](6-3)$$
$$-[2(0.5)^3 - 3(0.5)^2]6 + 0.5[(0.5)^3 - (0.5)^2](7-6)$$
$$= 2 + 0.1875 - (-3) - 0.0625$$
$$= 5.125 \approx 5.$$

For the second series, the values of P_{i-1}, P_{i-2}, P_{i+1} and P_{i+2} are 6, 8, 12 and 14 respectively.

By substitution,

$$IV_{cs} = (2(0.5)^3 - 3(0.5)^2 + 1)8 + 0.5[(0.5)^3 - 2(0.5)^2 + 0.5](12 - 6)$$
$$- (2(0.5)^3 - 3(0.5)^2)12 + 0.5[(0.5)^3 - (0.5)^2](14 - 12)$$
$$= 4 + 0.375 - (-6) + (-0.125) = 10.25 \approx 10.$$

For the third series, the values of P_{i-1}, P_{i-2}, P_{i+1} and P_{i+2} are 4, 6, 12 and 14 respectively.

By substitution,

$$IV_{csr} = (2(0.5)^3 - 3(0.5)^2 + 1)6 + 0.5((0.5)^3 - 2(0.5)^2 + 0.5)(12 - 4)$$
$$- (2(0.5)^3 - 3(0.5)^2)12 + 0.5((0.5)^3 - (0.5)^2)(14 - 12)$$
$$= 3 + 0.5 - (-6) - 0.125 = 9.375 \approx 9.$$

Using CSR Method

$$IV_{csr} = (2\gamma^3 - 3\gamma^2 + 1)P_{i-1} + (\gamma^3 - 2\gamma^2 + \gamma)(P_{i+1} - P_{i-2}) - (2\gamma^3 - 3\gamma^2)P_{i+1}$$
$$+ (\gamma^3 - \gamma^2)(P_{i+2} - P_{i+1}),$$

where γ is 0.5, and for the first series, $P_{i-1} = 4$, $P_{i-2} = 3$, $P_{i+1} = 6$ and $P_{i+2} = 7$.

Therefore, the estimate is

$$IV_{csr} = (2(0.5)^3 - 3(0.5)^2 + 1)4 + ((0.5)^3 - 2(0.5)^2 + 0.5)(6 - 3)$$
$$- (2(0.5)^3 - 3(0.5)^2)6 + ((0.5)^3 - (0.5)^2)(7 - 6)$$
$$= 2 + 0.375 - (-3) + (-0.125) = 5.25 \cong 5.$$

For the second series, $P_{i-1} = 8$, $P_{i-2} = 6$, $P_{i+1} = 12$ and $P_{i+2} = 14$.

Therefore, the estimate is

$$IV_{csr} = (2(0.5)^3 - 3(0.5)^2 + 1)8 + ((0.5)^3 - 2(0.5)^2 + 0.5)(12 - 6)$$
$$- (2(0.5)^3 - 3(0.5)^2)12 + ((0.5)^3 - (0.5)^2)(14 - 12)$$
$$= 4 + 0.75 - (-6) + (-0.25) = 10.5 \cong 11.$$

For the third series, $P_{i-1} = 6$, $P_{i-2} = 4$, $P_{i+1} = 12$ and $P_{i+2} = 14$.

Therefore, the estimate is

$$IV_{csr} = (2(0.5)^3 - 3(0.5)^2 + 1)6 + ((0.5)^3 - 2(0.5)^2 + 0.5)(12 - 4)$$
$$- (2(0.5)^3 - 3(0.5)^2)12 + ((0.5)^3 - (0.5)^2)(14 - 12)$$
$$= 3 + 1 - (-6) - 0.25 = 9.75 \cong 10.$$

Using PM Method

$$x_i = \frac{1}{4}[x_{i-2} + x_{i-1} + x_{i+1} + x_{i+2}].$$

For the first series, the values are 3, 4, 6 and 7.

Therefore, the estimate is

$$x_i = \frac{1}{4}[3 + 4 + 6 + 7] = 5.$$

For the second series, the values are 6, 8, 12 and 14.

Therefore, the estimate is

$$x_i = \frac{1}{4}[6 + 8 + 12 + 14] = 10.$$

For the third series, the values are 6, 10, 14 and 18.

Therefore, the estimate is

$$x_i = \frac{1}{4}[6 + 10 + 14 + 18] = 12.$$

Table 1. Summary of the results using the five methods.

S/N	Missing Value	L Method	LL Method	CS Method	CSR Method	PM Method
1	5	5	4.89898	5.125	5.25	5
2	10	10	9.79796	10.25	10.5**	10
3	12	12	11.8322	9.375	9.75**	12

Linear method gave the same value as the delete value, as well as log-linear method but the log-linear method gave approximately the values, not exact. This shows the superiority of linear method over log-linear method. CS method also resulted to approximate values with the third estimate significantly differ from the missing value. CSR method gave approximated values with the second result slightly higher than the missing value and the third result lower than the missing value. This implies for the series under consideration, CSR has low interpolation strength.

From the above table, the proposed method (PM) gives exact values as estimate which makes its precision to be 100% accurate. This implies the proposed method can be used for estimate of missing value (interpolation).

B. Validity of Case 2 of the proposed method

Considering a situation where the number of observations preceding the missing value is just one, Case 1 of the proposed method may not be accurate but Case 2 was presented to handle such problem. Using the same set of random numbers generated for Case 1 with different position of missing value.

Given the data set below, the data point two is removed (deleted), provide the estimate using the methods of interpolation stated

(i) 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10

(ii) 2, 4, 6, 8, 10, 12, 14, 16, and 18

(iii) 1, 4, 6, 10, 12, 14, 18 and 20.

As instructed, the series become

(1) 1, *y*, 3, 4, 5, 6, 7, 8, 9, and 10

(2) 2, *y*, 6, 8, 10, 12, 14 , 16 and 18

(3) 1, *y*, 6, 10, 12, 14, 18 and 20.

Note that the missing values are 2, 4 and 4 respectively.

The expression for Case 2 is

$$x_i = \frac{1}{4}[x_{i-1} + x_{i+1} + x_{i+2}]$$

This is because x_{i-2} is replaced with zero.

For series 1, the interpolated value is

$$x_1 = \frac{1}{4}[1 + 3 + 4] = 2.$$

For series 2, the interpolated value (estimate) is

$$x_2 = \frac{1}{4}[2 + 6 + 8] = 4.$$

For series 3, the interpolated value is

$$x_3 = \frac{1}{4}[1 + 6 + 10] = 4.25 \approx 4.$$

From the above results, the estimates are adequate for interpolation when the position of the missing value is at data point two. For missing data point one, different approach can be taken such as considering the starting point of the data as point two.

4. Summary

Four methods of interpolation were considered in this paper. Among the methods, two methods attempts to fix the missing link in the series using linearity approach and the other two methods use curvature. The missing link between a series can be bridged using a straight line or curve. Based on the data used for review, the methods that attempted the link with a straight line seem to be better than the methods that attempted the link with curve irrespective of the location of the missing value.

A method was presented as a new approach of finding or estimating the missing value. By estimate, the method is capable of producing a value very close to the actual missing value irrespective of the location of the missing link. The new method (proposed method) is capable of interpolation, estimating missing value within a series.

Monte Carlo simulation approach was used to generate set of random numbers varying the conditions and the common difference among the numbers to produce different conditions. Jackknife delete-d method was used to create missing link among the simulated observations, thereby producing missing value that are known ahead of the computation. The idea is to check level of precision of the methods for possible evaluation. The principle used in the paper is the closeness of the estimate to the actual missing value. A better method produces closest value as an estimate among the methods.

Considering linear, log-linear, cardinal spline, Catmull-Rom spline and proposed method (PM), linear, log-linear and proposed method were able to produce the closest estimate which shows the methods are better than cardinal spline and Catmull-Rom spline.

Considering the rigour in the computation of linear and log-linear, since the proposed method is as good as the two methods, it is better to adopt the proposed method to prevent human error in the computation. Therefore, the proposed method is highly recommended for interpolation.

Further test can be carried out on the strength of the method using a more rigour computation such as the use of replicate to determine the power-of-test.

References

[1] R. Ahmad Mahir, and A. M. H. Al-Khazaleh, Estimation of missing data by using the filtering process in a time series modeling, 2008. arXiv:0811.0659 [stat.ME]

[2] Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, 2006.

[3] J. R. Cheema, Some general guidelines for choosing missing data handling methods in educational research, *Journal of Modern Applied Statistical Methods* 13(2) (2014), Article 3. https://doi.org/10.22237/jmasm/1414814520

[4] Christian P. Robert and George Casella, *Monte Carlo Statistical Methods*, 2nd ed., Springer Text in Statistics, New York: Springer-Verlag, 2004.

[5] David S. Fung, Methods for the estimation of missing values in time series, Ph.D. Thesis, Edith Cowan University, Perth, Australia, 2006.

[6] I. S. Iwueze, E. C. Nwogu, V. U. Nlebedim, U. I. Nwosu and U. E. Chinyem, Comparison of methods of estimating missing values in time series, *Open Journal of Statistics* 8 (2018), 390-399. https://doi.org/10.4236/ojs.2018.82025

[7] Ruey S. Tsay, *Analysis of Financial Time Series*, 2nd ed., John Wiley and Sons, Inc., 2005.

[8] Steven T. Yen and Andrew M. Jones, Individual cigarette consumption and addiction: a flexible limited dependent variable approach, *Health Econ.* 5 (1996), 105-117.