JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 7, No. 1, January 2023, pp. 160-173

Percentile Bootstrap Interval on Univariate Local Polynomial Regression Prediction

Abil Mansyur[1], Elmanani Simamora[2], Ahmad[3]
[1,2]Departement of Mathematics, Universitas Negeri Medan, Indonesia
[3]Departement of Mathematics Education, Universitas Muhammadiyah Purwokerto, Indonesia
abil@unimed.ac.id[1], elmanani_simamora@unimed.ac.id[2], ahmad@ump.ac.id[3]

Article History:
Received : 03-11-2022
Revised : 03-12-2022
Accepted : 16-12-2022
Online : 12-01-2023

Keywords:
Simulation;
Percentile Bootstrap;
Local Polynomial;
Regression;
Paired Bootstrap;
Residual Bootstrap.

ABSTRACT

This study offers a new technique for constructing percentile bootstrap intervals to predict the regression of univariate local polynomials. Bootstrap regression uses resampling derived from paired and residual bootstrap methods. The main objective of this study is to perform a comparative analysis between the two resampling methods by considering the nominal coverage probability. Resampling uses a nonparametric bootstrap technique with the return method, where each sample point has an equal chance of being selected. The principle of nonparametric bootstrapping uses the original sample data as a source of diversity in contrast to parametric bootstrapping, where the variety comes from generating a particular distribution. The simulation results show that the paired and residual bootstrap interval coverage probabilities are close to nominal coverage. The results showed no significant difference between paired bootstrap interval and percentile residual. Increasing the bootstrap sample size sufficiently large gives the scatterplot smoothness of the confidence interval. Applying the smoothing parameter by choice gives a second-order polynomial regression with a smoother distribution than the first-order polynomial regression. The scatterplot shows that the second-degree polynomial regression can capture the data curvature feature compared to the first-degree polynomial. The bands made from second-degree polynomials give a narrower width than first-degree polynomials. In contrast, applying optimal smoothing parameters to the model provides different conclusions by using smoothing parameters based on choice. In addition to the differences based on the scatterplot, the bootstrap estimates of the coverage probability are also other. Selecting smoothing parameters based on a particular value provides probability coverage with the paired bootstrap method for the first-degree local polynomial regression is 0.93, while the second-degree local polynomial is 0.96. The probability of coverage based on the residual bootstrap method for the first-degree local polynomial regression is 0.95, while the second-degree local polynomial is 0.96. The probability coverage based on the optimal parameters of the paired bootstrap method for the first-degree local polynomial regression is 0.945, while the second-degree local polynomial is 0.93. The residual bootstrap method gives the first-degree local polynomial regression of 0.95, while the second-degree local polynomial is 0.93. In general, both bootstrap methods work well for estimating prediction confidence intervals.

https://doi.org/10.31764/jtam.v7i1.11752

——————————— ◆ ———————————

A. INTRODUCTION

Efron & Tibshirani (1994) were pioneers in introducing the bootstrap method as a resampling technique that is very useful for estimating statistics without fulfilling certain assumptions. In the bootstrap method, there is bootstrapping terminology which is a resampling procedure from the original data to produce many simulated samples (bootstrap samples). Simamora et al. (2015) suggested using a computer with a high ability level to perform bootstrapping in the simulation. Expensive simulations are options if the statistics are not in closed form or more complex statistics that do not require certain assumptions. Bootstrapping is a resampling technique that is useful for analysing difficult statistics without strict rules or the parametric assumptions of the applied model are not met (Solci et al., 2022). The working principle of bootstrapping relies on resampling the empirical distribution, which can be done by weakening parametric assumptions. One of the complex and sensitive statistics is constructing a confidence interval for nonparametric regression prediction. In practice, constructing standard confidence intervals based on asymptotic distribution theory can be wildly inaccurate (Diciccio & Efron, 1996). The curve features of the Local Polynomial Regression and the lower and upper limits of the interval are far from the truth.

The failure of the normality approach to provide a valid confidence interval has prompted some alternative methods. Eubank & Speckman (1993) proposed a bias-corrected confidence band in a nonparametric kernel regression model. The consideration is only on the bands generated from the kernel estimator of the regression curve and the selection of rounds based on the behaviour of the data. The consideration is only on the bands generated from the kernel estimator of the regression curve and the selection of rounds based on the behaviour of the data. The results of the Monte Carlo simulation show that the mean response confidence interval has asymptotically correct coverage and behaves well in small sample studies. They concluded that Bonferroni-type bands have conservative asymptotic coverage behaviour for large samples without bias correction. Then, Xia (1998) answered the open-ended question on page 1298 of (Eubank & Speckman, 1993). They used local polynomial regression model matching to construct confidence intervals for the mean of response using cross-validation and plug-in methods to select bandwidth.

Härdle & Bowman (1988) discuss the performance of bootstrapping and direct methods on a nonparametric regression model. They use the principle of good local adaptive choice of local smoothing parameters. This principle is applied to bootstrap sampling to estimate mean squared errors and percentile intervals from nonparametric estimates at test points. These two applications compare bootstrap performance with a simple "plug-in" method based on direct estimation (asymptotic expansion). In general, the performance of these two methods is generally very similar. However, bootstrap has the slight advantage of not being as sensitive to second derivatives. Moreover, in the confidence interval construct, the bootstrap can reflect features such as skewness but slightly less than the target confidence interval due to inaccuracies in centring. Ringle et al. (2012) warned that the correct setting could provide a reasonable bootstrap confidence interval estimate. Poor choice of options can lead to a significantly biased estimate of the standard error and cause the bootstrap estimate to become unstable.

Özdemir (2013) showed a better pencil bootstrap interval on the probability of error in Type I and more efficient computation time. Aguirre-Urreta & Rönkkö (2017) revealed that the confidence interval of the pencil bootstrap is the most straightforward approach, but it is necessary to consider the exact statistical distribution. This approach will work best if the statistical distribution is symmetrical and centred on the original estimate. Then, (Jung et al.,

2019) show that the coverage probability of the bootstrap percentile confidence interval is closer to the nominal coverage. They study general structured component analysis without the need for distributional assumptions. Gultom et al. (2022) applied the Gompertz Growth Model with Levenberg–Marquardt iteration on the soybean growth process. They conclude that the bootstrap resampling process in the growth model does not change the characteristics of the data (information from the data), and aims to fulfill the assumption of residual normality.

This approach will work best if the statistical distribution is symmetrical and centred on the original estimate. Then, (Jung et al., 2019) show that the coverage probability of the bootstrap percentile confidence interval is closer to the nominal coverage. The application of the paired and residual nonparametric bootstrap method aims to construct predictive confidence intervals for local polynomial regression. Then perform a comparative study between the two nonparametric bootstrap methods by considering the bootstrap estimate for the standard error of the pivot quantity and the proximity of the empirical probability coverage of the bootstrap to the nominal range. The basic idea uses the results of (Mansyur & Simamora, 2022).

The organisation of this paper is as follows. The first section presents an introduction covering the background and proposals of this research. The second part describes the research method and summarises the concepts and theories of the local polynomial regression model and bootstrap. This section also provides a new algorithm (novelty) related to the bootstrap estimation of confidence intervals for local polynomial regression predictions with nested bootstrap using paired and residual bootstrap methods. The third section deals with the results and discussion of the simulation of the new algorithm. The last section includes research conclusions and suggestions for further development.

B. METHODS

This research method is a combination of literature review and simulation. The literature review aims to provide a framework for deriving a new algorithm. At the same time, the sample data follows the data generated from the literature of (Eubank & Speckman, 1993). The simulation sample follows the resampling of the generation sample data using paired and residual bootstraps. Figure 1 presents the flow of thinking after conducting a literature review. This flow only displays the main stages in the simulation and analyses the output of the simulation. The following section will explain some parts of these stages. Section 1 summarises the literature review on the concepts and theories of local polynomial regression, paired bootstrap and residuals. Section 2 presents the proposed new algorithm of bootstrap percentile confidence intervals for predicting local polynomial regression in detail, as shown in Figure 1.

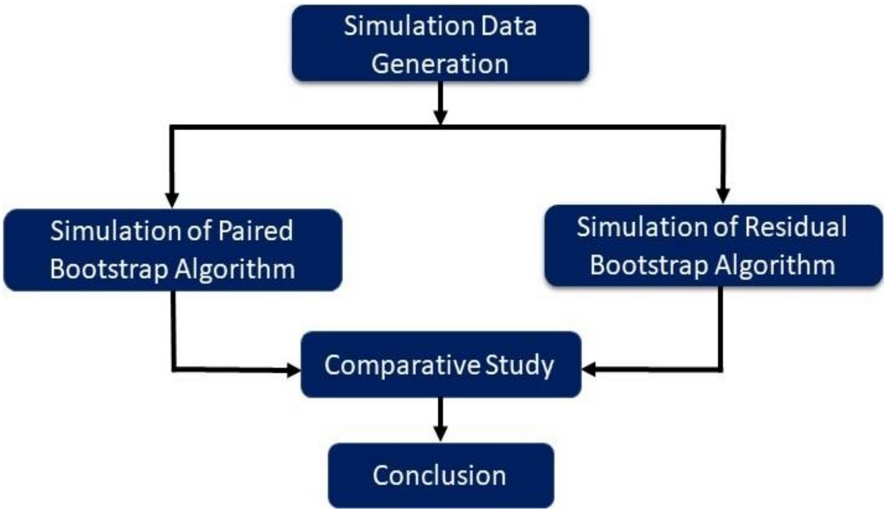
Simulation Data Generation
Simulation of Paired Bootstrap Algorithm
Simulation of Residual Bootstrap Algorithm
Comparative Study
Conclusion

Figure 1. Simulation Flow and Simulation Output Analysis

1. Summary of Literature Review

This section summarises the concepts and theories of local polynomial regression. To make writing easier, LPR-1 is an acronym for first-degree Local Polynomial Regression (LPR), and LPR-2 is an acronym for second-degree Local Polynomial Regression. Then in this section also summarises the paired and residual bootstrap method in general. The combination of the LPR concept and theory and the bootstrap method proposes two new algorithms for bootstrap percentile interval estimation based on the empirical distribution of pivot quantities. We mention the terminology of the two algorithms with the paired bootstrap percentile interval and residual with the acronyms CI-Paired and CI-Residual, respectively.

a. Univariate Local Polynomial Regression

Nonparametric regression is an extension of the parametric regression model. The average response does not have a specific trend but is constructed according to a data-based factual information set. Local polynomial regression (LPR) is a nonparametric regression model that estimates the relationship between the independent and dependent variables without assuming any functional form. Cleveland (1979) presented a univariate LPR model in the following form,

$$y(x_i) = g(x_i) + \varepsilon_i \text{ for } i = 1,2,\cdots,n, \tag{1}$$

where $g(x_i)$ is an unknown smoothing function and ε_i is a random variable independently and identically distributed. The function g is the expectation of the response that needs to be estimated. Meanwhile, the random variable ε has the expectation $E(\varepsilon) = 0$ and the constant variance, $Var(\varepsilon) = \sigma^2$. According to (de Brabanter et al., 2013), as long as the $(p+1)^{th}$ derivative of g at the point of interest x_0 exists, the function $g(x_i)$ can be approximated locally with a polynomial degree p,

$$g(x_i) \approx \sum_{j=0}^{p} \frac{g^j(x_0)}{j!}(x_i - x_0)^j \equiv \sum_{j=0}^{p} \beta_j(x_i - x_0)^j. \tag{2}$$

Polynomial matching locally or, in other words, looks for the estimated parameter β_j on the right-hand side of equation (2) using the weighted least squares method with minimizing the problem,

$$\hat{\beta}_j = \min_{\beta} \sum_{i=1}^{k} \left[y(x_i) - \sum_{j=0}^{p} \beta_j(x_i - x_0)^j\right]^2 W(|x_0 - x_i|/\Delta(x_0) \tag{3}$$

where $k = \lfloor \gamma n \rfloor$ represents the number of points $x_i \in \mathbb{N}(x_0)$. The γ parameter is the curve smoothing parameter, and n is the sample size. The W function is the selectable weight function and $\Delta(x_0) = \underset{x_i \in \mathbb{N}(x_0)}{\text{maximum}} |x_i - x_0|$. Cleveland (1979) characterizes the W function as follows:

1) $W(x) > 0$, for $|x| < 1$;
2) $W(-x) = W(x)$;
3) $W(x)$ is a non-increasing function for $x \geq 0$;
4) $W(x) = 0$, for $|x| \geq 1$.

Researchers usually choose the γ value of between zero and one. It is necessary to consider the magnitude of the γ value, where the γ value close to zero will predict overfitting or a wavy curve surface. The γ value close to one will provide a smooth surface curve prediction or underfitting but omit the original data features. Mansyur & Simamora (2022) offer a search algorithm for optimal smoothing parameters using cross-validation. The choice of the polynomial degree also determines the curve's smoothness. Usually, researchers use the low-degree polynomial, where the first degree is a linear polynomial and the second degree is a quadratic polynomial.

Fan & Gijbels (1960) provide a solution to equation (3) using the weighted least squares method in the form of a matrix equation,

$$\widehat{\boldsymbol{\beta}}_\gamma = \left(\boldsymbol{X}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{X}_\gamma\right)^{-1} \boldsymbol{X}_\gamma^T \boldsymbol{W}_\gamma \boldsymbol{Y}, \quad (4)$$

where

$$\boldsymbol{X}_\gamma = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^p \\ 1 & (x_2 - x_0) & \cdots & (x_2 - x_0)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_k - x_0) & \cdots & (x_k - x_0)^p \end{pmatrix}; \; \boldsymbol{Y} = \begin{pmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_k) \end{pmatrix}; \; \widehat{\boldsymbol{\beta}}_\gamma = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix};$$

and $\boldsymbol{W}\gamma$ is a diagonal matrix of size k×k whose diagonal element contains the sequence $W(|x_0 - x_1|/\Delta(x_0), W(|x_0 - x_2|/\Delta(x_0), \cdots, W(|x_0 - x_k|/\Delta(x_0)$.

Following the similarity as in the case of the linear regression model, the prediction of LPR at a point x_0 using the weighted least squares method yields,

$$\hat{y}(x_0) = \hat{g}(x_0) = \sum_{j=0}^{p} \hat{\beta}_j x_0^j. \quad (5)$$

Readers interested in studying more about the weighted least squares method can read the literature of (Draper & Smith, 1998).

b. Paired and Residual Bootstrapping

Efron & Tibshirani (1994) present two resampling methods, paired and residuals bootstrap processes, in a linear regression model. They give an open problem on page 113, which is the best between paired and residual Bootstrapping? The answer is left to us to what extent we trust the linear regression model. We conclude that there are four exciting provisions from the results of the analysis of (Efron & Tibshirani, 1994) regarding the results of the percentile regression simulation of cholestyramine data, namely:

1) Paired bootstrapping is slightly more sensitive than residual bootstrapping;
2) The error and mean of the response for paired bootstrapping do not depend on the covariates of the original data. The reason is that the covariates are random, unlike the residuals, in which the covariate structure is unchanged;
3) The residual bootstrap has the same suitability as the original data;
4) The bootstrap method does not provide a unique conclusion for the particular concept.

Efron & Tibshirani (1994) claim that the two methods are equivalent when the sample size reaches infinity (asymptotic). The difference will appear if the sample size is relatively small. Chernick & LaBudde (2014) also review these two methods. Unfortunately, this literature does not contain exciting statements in the bootstrapping process, only focusing on algorithms and coding in the R programming language.

Based on the two types of literature provides information that the difference lies only in the resampling scheme, and we will summarize it further. Suppose the linear regression model is $y_j = x_j^T \beta + \varepsilon_j$, where (x_j, y_j) is an ordered pair of responses and a covariate vector of size $p \times 1$. The difference between the resampling schemes of the two methods is as follows.

1) Paired bootstrap takes a simulated sample (bootstrap sample) from the original data $(x_1, y_1), \cdots, (x_n, y_n)$ independently with returns. Each original data point has an equal chance of being taken as a sample point, $1/n$. The resampling process allows a bootstrap sample to have two or more of the same sample points or an original data point to be taken twice or more as members of the bootstrap sample.
2) Residual Bootstrap performs the first procedure by matching the original data $(x_1, y_1), \cdots, (x_n, y_n)$ into the model to get $\hat{y}_j = x_j^T \hat{\beta}$. Then calculate each residual $\hat{\varepsilon}_j = y_j - x_j^T \hat{\beta}$ which gives the residual vector $\hat{\boldsymbol{\varepsilon}}^T = (\hat{\varepsilon}_1, \cdots \hat{\varepsilon}_n)$. Determines $\hat{y}_j^* = x_j^T \hat{\beta} + \hat{\varepsilon}_j^*$ where taking $\hat{\varepsilon}_j^*$ from a point on the vector $\hat{\boldsymbol{\varepsilon}}^T$ independently with the return. The probability that each $\hat{\varepsilon}_j \in \hat{\boldsymbol{\varepsilon}}^T$ is drawn as a bootstrap sample point $\hat{\varepsilon}_j^*$ is the same, i.e. $1/n$, and it is possible to be drawn twice or more as a member of a bootstrap sample, $\hat{\boldsymbol{\varepsilon}}^{*T} = (\hat{\varepsilon}_1^*, \cdots, \hat{\varepsilon}_n^*)$. We repeat this process for another observation to get a bootstrap sample $(x_1, \hat{y}_1^*), \cdots, (x_n, \hat{y}_n^*)$.

2. Paired and Residual Bootstrap Percentile Interval

Wasserman (2004) and (Wasserman, 2006) give the pivot quantity $Z_n = \hat{\theta}_n - \theta$ for the exact interval,

$$\hat{\theta}_n \pm H^{-1}_{(1-\alpha/2)}, (6)$$

where $H^{-1}_{(1-\alpha/2)} = z_n^{(1-\alpha/2)}$ is the (1–α/2)-th quantile of the H distribution. The H distribution is a Cumulative Density Function (CDF) of the unknown Z_n pivot. Suppose $\widehat{H}$ is a bootstrap estimate for the H distribution considering the pivot $Z_n^{*b} = \hat{\theta}_n^{*b} - \hat{\theta}_n$. The bootstrap percentile interval for the theta parameter is

$$\hat{\theta}_n \pm \widehat{H}^{-1}_{(1-\alpha/2)} \ (7)$$

where $\widehat{H}^{-1}_{(1-\alpha/2)} = z_n^{*(1-\alpha/2)}$ is the bootstrap percentile (1−α/2)-th of the $\widehat{H}$ distribution. Hall & Horowitz (2013) performed a double bootstrap resampling to get the desired bootstrap percentile interval. It informs that multiple pivots in bootstrap can be done by considering the conditions in the model. We can derive two bootstrap percentile interval algorithms based on the literature review.

a. Paired Bootstrap Percentile Interval Algorithm

Based on the explanation of paired bootstrap in the previous section, we derive the steps of the paired bootstrap percentile interval algorithm.

1) Specifying a bootstrap sample $(x_1^*, y_1^*), \cdots, (x_n^*, y_n^*)$ with conditions as described in Paired Bootstrap.
2) Fitting a model using the bootstrap sample in step (1) uses equation (5) to get $(x_1^*, \hat{y}_1^*), \cdots, (x_n^*, \hat{y}_n^*)$.
3) Calculating the residual from each bootstrap sample point $\hat{\varepsilon}_i^* = y_i^* - \hat{y}_i^*$.
4) Normalizing for each bootstrap residual from step (3), $\tilde{\varepsilon}_i^* = \hat{\varepsilon}_i^* - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^*}{n}$. Suppose the bootstrap residual normalization vector $\tilde{\boldsymbol{\varepsilon}}^{*T} = (\tilde{\varepsilon}_1^*, \cdots \tilde{\varepsilon}_n^*)$.
5) Determining the residual vector with bootstrap double $\tilde{\boldsymbol{\varepsilon}}^{**T} = (\tilde{\varepsilon}_1^{**}, \cdots \tilde{\varepsilon}_n^{**})$ using independent and random retrieval with returns from the bootstrap residual normalization vector $\tilde{\boldsymbol{\varepsilon}}^{*T} = (\tilde{\varepsilon}_1^*, \cdots \tilde{\varepsilon}_n^*)$.
6) Repeating steps (1) to (5) B times to get the distribution $\widehat{H}(\tilde{\varepsilon}_i^{**})$. The distribution of $\widehat{H}(\tilde{\varepsilon}_i^{**})$ is CDF of the pivot $\tilde{\varepsilon}_i^{**}$.
7) Determining the percentile interval of the paired bootstrap (CI-Paired) at point x_0 using equation (5),

   $\hat{y}_n(x_0) \pm \widehat{H}^{-1}_{(1-\alpha/2)}(\tilde{\varepsilon}_0^{**})$, (8)

   where $\hat{y}_n(x_0)$ is prediction of LPR at point x_0 with using original data and $\widehat{H}^{-1}_{(1-\alpha/2)}(\tilde{\varepsilon}_0^{**}) = \tilde{\varepsilon}_0^{**B(1-\alpha/2)}$ is the bootstrap percentile (1-α/2)-th of the $\widehat{H}(\tilde{\varepsilon}_0^{**})$ distribution.

b. Residual Bootstrap Percentile Interval Algorithm

Following the sampling of the residual bootstrap, we derive the steps of the residual bootstrap percentile interval algorithm.

1) Fitting the original data $(x_1,y_1),\cdots, (x_n,y_n)$ into the model using equation (5) to get $(x_1, \hat{y}_1), \cdots, (x_n, \hat{y}_n)$.
2) Calculating the residuals from each point of the original data $\hat{\varepsilon}_i = y_i - \hat{y}_i$.
3) Normalizing for each residual data point in step (2) to get $\hat{\varepsilon}_i - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i}{n}$. Suppose the residual normalization vector is $\tilde{\boldsymbol{\varepsilon}}^{T} = (\tilde{\varepsilon}_1, \cdots \tilde{\varepsilon}_n)$.
4) Determining a bootstrap sample point $\hat{y}_i^* = \hat{y}_i + \tilde{\varepsilon}_i^*$ where $\tilde{\varepsilon}_i^*$ is an independent random sampling with the return of the residual normalized vector $\tilde{\boldsymbol{\varepsilon}}^{T} = (\tilde{\varepsilon}_1, \cdots \tilde{\varepsilon}_n)$. Suppose the bootstrap sample point set is $(x_1, \hat{y}_1^*), \cdots, (x_n, \hat{y}_n^*)$.
5) Fitting the bootstrap sample $(x_1, \hat{y}_1^*), \cdots, (x_n, \hat{y}_n^*)$ into the model using equation (5) to get a double bootstrap sample $(x_1, \hat{y}_1^{**}), \cdots, (x_n, \hat{y}_n^{**})$.

6) Calculating the residual from point each of the double bootstrap sample $\hat{\varepsilon}_i^{**} = \hat{y}_i^{*} - \hat{y}_i^{**}$.
7) Normalizing for each residual of double bootstrap sample, $\tilde{\varepsilon}_i^{**} = \hat{\varepsilon}_i^{**} - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{**}}{n}$. Suppose the normalization vector of double bootstrap residual is $\tilde{\boldsymbol{\varepsilon}}^{**T} = (\tilde{\varepsilon}_1^{**}, \cdots \tilde{\varepsilon}_n^{**})$.
8) Determining the residual vector of the bootstrap-trio sample $\tilde{\boldsymbol{\varepsilon}}^{***T} = (\tilde{\varepsilon}_1^{***}, \cdots, \tilde{\varepsilon}_n^{***})$ where $\tilde{\varepsilon}_i^{***}$ is the take independently and randomly with returns from vektor $\tilde{\boldsymbol{\varepsilon}}^{**T} = (\tilde{\varepsilon}_1^{**}, \cdots \tilde{\varepsilon}_n^{**})$.
9) Repeating steps (1) to (8) B times to get the distribution of $\widehat{H}(\tilde{\varepsilon}_i^{***})$. The distribution of $\widehat{H}(\tilde{\varepsilon}_i^{***})$ is the CDF of the pivot $\tilde{\varepsilon}_i^{***}$.
10) Determining the percentile interval of the residual bootstrap (CI-Residual) at point x_0 using equation (5),

$$\hat{y}_n(x_0) \pm \widehat{H}_{(1-\alpha/2)}^{-1}(\tilde{\varepsilon}_0^{***}), \tag{9}$$

where $\hat{y}_n(x_0)$ is prediction of LPR at point x_0 with using original data and $\widehat{H}_{(1-\alpha/2)}^{-1}(\tilde{\varepsilon}_0^{***}) = \tilde{\varepsilon}_0^{***B(1-\alpha/2)}$ is the bootstrap percentile (1-α/2)-th of the $\widehat{H}(\tilde{\varepsilon}_0^{***})$ distribution.

C. RESULT AND DISCUSSION

The design of the independent variable x and the dependent variable y in the simulation for the two algorithms follows the following conditions. The independent variable (covariate) x is an equidistant points design with $x_{\min} = 0$ and $x_{\max} = 2\pi$. The sample size will affect the distance from one point to another in the observation domain. The dependent variable (response) y comes from the trigonometric function $g(x) = \sin 2x$ with the addition of an error normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 0.2$. In addition, we need to choose a weight function where researchers generally, such as (Cleveland, 1979), (Cleveland & Grosse, 1991), and (Cleveland et al., 1988) use the tricube weight function,

$$W(u) = \begin{cases} (1 - |u|^3)^3 & \text{, untuk } |u| < 1 \\ 0 & \text{, untuk } |u| \geq 1 \,. \end{cases} \tag{9}$$

The smoothing of the LPR curve uses two parameters, γ as the smoothing parameter and p as the degree of LPR. Researchers may choose the magnitude of the γ parameter provided that it is not too close to zero and not too close to one or use the optimal search using cross-validation. The simulation uses two alternatives to determine smoothing parameters. The first is to select the value γ = 0.5, and the second is to use the optimal γ search algorithm in (Mansyur & Simamora, 2022). The goal is to analyze whether there is an optimal γ influence. The simulation considers low-degree polynomials, namely $p = 1$ and $p = 2$. The interval construction uses a 95% confidence interval or a significance level of $\alpha = 5\%$.

Figure 2 is the simulation result for the first algorithm with a sample size of n =100 with the number of bootstrap samples B = 1000 and a smoothing parameter γ = 0.5. Figure 2(a) is a scatterplot of CI-Paired and LPR-1 where the CI-Paired for the upper boundary (green curve) and lower boundary (blue curve) have jagged or wavy surfaces. The surface of the LPR-1 curve (curve in black) is very far from the curvature feature of the trigonometric function $g(x) = \sin 2x$. The CI-Paired coverage probability of LPR-1 is 0.93, which is close to nominal coverage. On the other hand, Figure 2(b), derived from CI-Paired based on LPR-2, shows a smoother surface and its curvature features follow the trigonometric function $g(x) = \sin 2x$. The area formed by LPR-1 is wider than LPR-2. As a result, the CI-Paired band of LPR-1 is broader than that of LPR-2. The CI-Paired coverage probability of LPR-2 is the same as the nominal coverage, as shown in Figure 2.

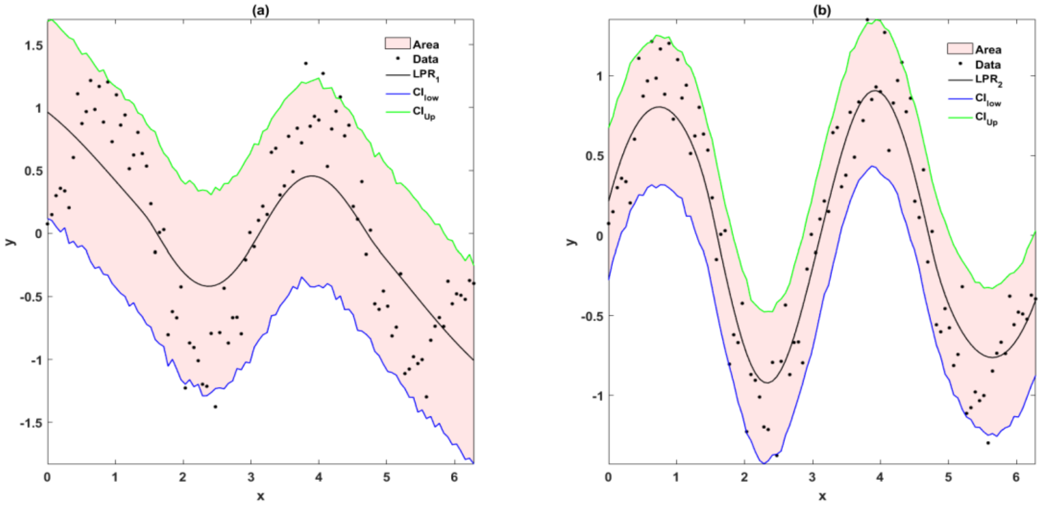

Figure 2. Scatterplot of Paired Bootstrap Percentile Interval and Local Polynomial Regression Curve with Bootstrap Number of Samples *B* = 1000 and γ = 0.5

The literature of (Efron & Tibshirani, 1994) on page 47, reveals that to get an ideal estimator, it is necessary to increase the number of bootstrap samples. To achieve that, we need to increase the number of bootstrap samples, say *B* = 10000. Considering that the larger *B* size will result in an expensive simulation is necessary. Figure 3 is a simulation with the same conditions as Figure 2 but only differs in the number of bootstrap samples. The simulation results show that Figure 3 gives a smoother curve scatterplot than Figure 2. The curve feature does not change, but the probability of coverage of CI-Paired from LPR-2 becomes 0.96, as shown in Figure 3.

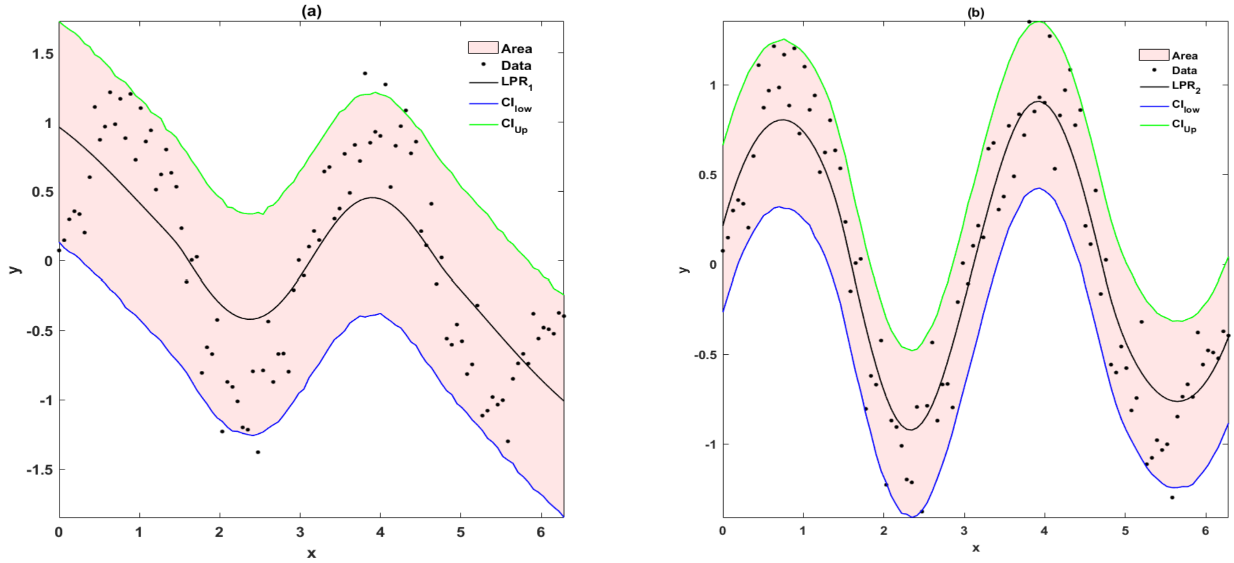

Figure 3. Scatterplot of Paired Bootstrap Percentile Interval and Local Polynomial Regression Curve with Bootstrap Number of Samples *B* = 10000 and γ = 0.5

Using the same sample design and conditions above, we apply the second algorithm to the simulation with a Bootstrap B = 10000 sample lot. Figure 4 shows that there is no significant difference from the conclusions of the first algorithm.

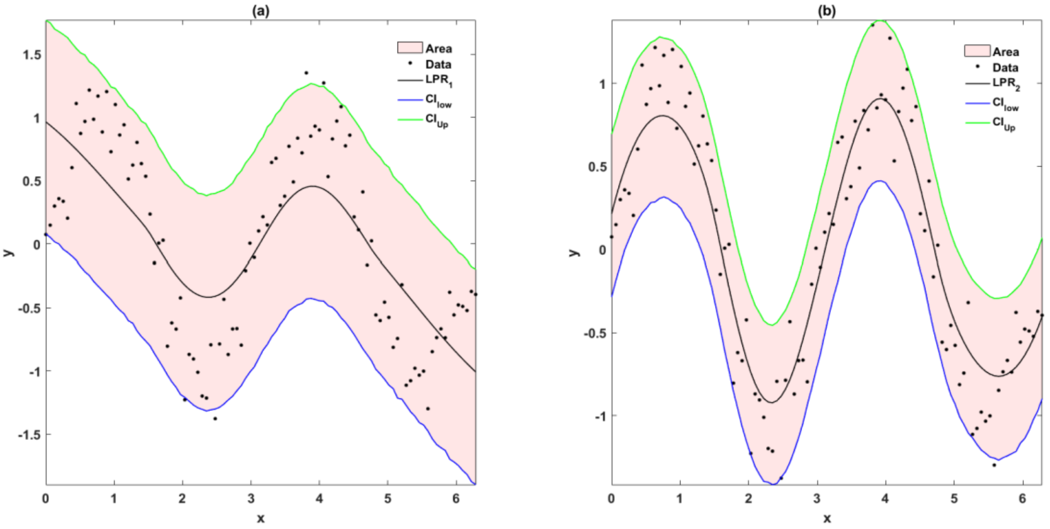

Figure 4. Scatterplot of Residual Bootstrap Percentile Interval and Local Polynomial Regression Curve with Bootstrap Number of Samples B = 10000 and γ = 0.5

The coverage probability of the CI-Residual of LPR-1 is the same as the nominal coverage, while the CI-Residual of LPR-2 is 0.96. The simulation applies the search for optimal smoothing parameters from the (Mansyur & Simamora, 2022) algorithm, where the sample design conditions are the same as above. Mansyur & Simamora (2022) used the cross-validation function to get the optimal γ value. The cross-validation function uses the formula,

$$CV(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\left\{y(x_i) - \hat{y}_{\gamma}^{-i}(x_i)\right\}^2, \tag{10}$$

where $\hat{y}_{\alpha}^{-i}(x_i)$ is the prediction of the LPR at the point x_i for which the value of $y(x_i)$ is removed from original data. The simulation gives γ_{Optimal} = 0.09 for LPR-1 with $CV(\gamma_{\text{Optimal}})$ = 0.0499 and γ_{Optimal} = 0.25 for LPR-2 with $CV(\gamma_{\text{Optimal}})$ = 0.049. Because LPR-2 gives a smaller CV value, we use γ_{Optimal} = 0.25, as shown in Figure 5.

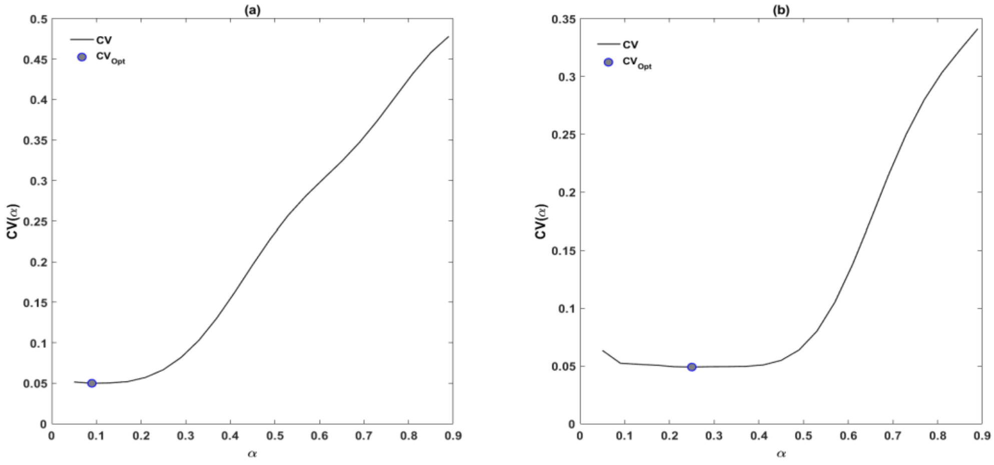

Figure 5. Scatterplot of Optimal Gamma Search with Sample Size n = 100

Figure 6 is the simulation result of the paired percentile bootstrap interval algorithm, which applies γ_{Optimal} = 0.25 for LPR-1 and LPR-2, and the number of bootstrap samples is B = 10000, as shown in Figure 6.

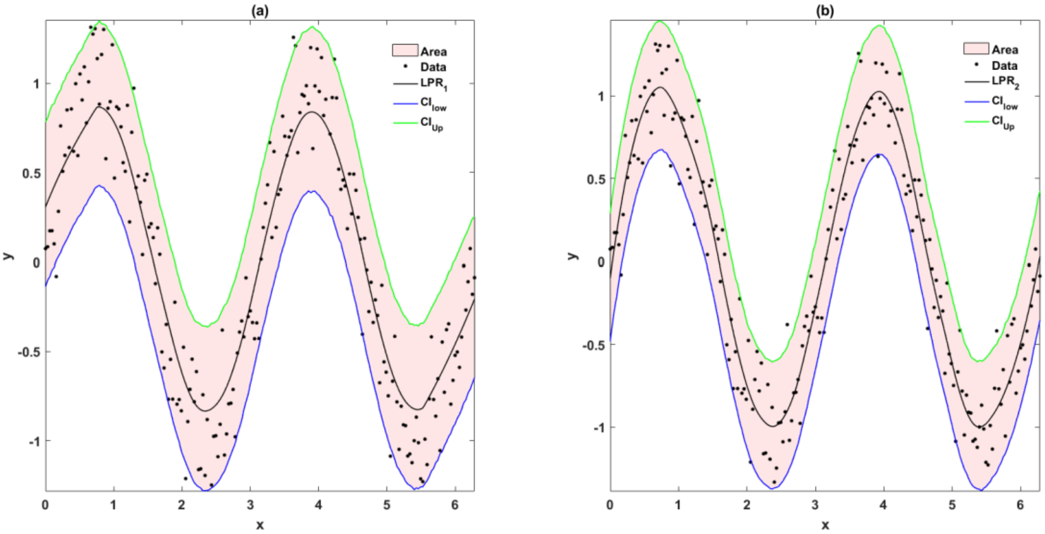

Figure 6. Scatterplot of Paired Bootstrap Percentile Interval and Local Polynomial Regression Curve with Bootstrap Number of Samples B = 10000 and γ_{Optimal} = 0.25

The scatterplot shows the smooth surface of the LPR-1 and LPR-2 curves following the curvature feature of the trigonometric function $g(x) = \sin 2x$. However, the bandwidth of the CI-Paired from LPR-2 is narrower than that of the LPR-1. The coverage probability of CI-Paired from LPR-1 is 0.945, while LPR-2 is 0.93. Figure 7 is a simulation result of the bootstrap residual, which shows the same conclusion as Figure 6 but has a different coverage probability. The probability of coverage of the CI-Residual from LPR-1 is 0.95, while the LPR-2 is 0.93, as shown in Figure 7.

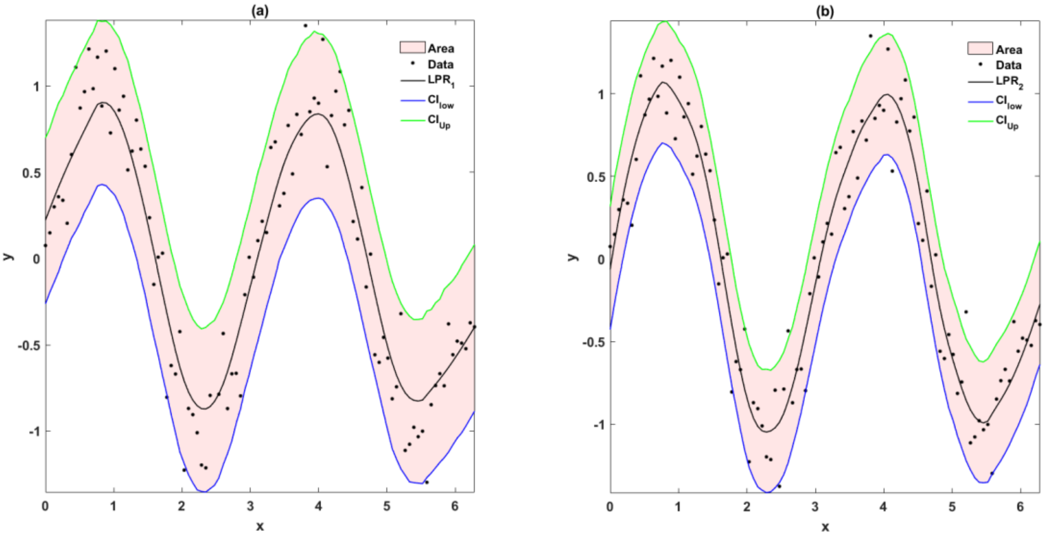

Figure 7. Scatterplot of Residual Bootstrap Percentile Interval and Local Polynomial Regression Curve with Bootstrap Number of Samples B = 10000 and γ_{Optimal} = 0.25

D. CONCLUSION AND SUGGESTIONS

Nonparametric regression models generally require a large enough sample size to capture the curve features. The two new algorithms can work well at relatively small sample sizes. However, for the polynomial regression of degree one with the selection of the smoothing parameter α = 0.5, it cannot characterize the sample from a particular function. The bandwidth resulting from the regression of the first-degree polynomial is wider than the second-degree polynomial regression. Still, there is no guarantee that the coverage probability will be the same as the nominal coverage. On the other hand, second-degree polynomial regression can characterize the behaviour of the data derived from the generation of a particular function, and the probability coverage is close to the nominal probability coverage.

The smoothness of the curve is also affected by the number of bootstrap samples. If the number of bootstrap samples is relatively small, the surface of the curve is more jagged and wavy, especially for first-degree polynomial regression. At the same time, the second-degree polynomial has a smoother curvilinear surface even though the number of bootstrap samples is relatively small. The purpose of increasing the number of bootstrap only to smooth the surface of the curve does not change the behavior of the curve curve, which is analogous to the conclusion of (Gultom et al., 2022).

The scatterplot shows that applying the optimal smoothing parameter to the local polynomial regression model improves performance. Both local polynomial regressions can capture curve features based on the behaviour of the sample derived from the generation of a particular function. The band thickness of the first-degree polynomial is more proportional than that of the second-degree polynomial. The second-degree polynomial regression band trend is narrower than the first-degree polynomial regression for both algorithms. The probability coverage of the two algorithms is not significantly different. However, the coverage probability of the first-degree polynomial is better than that of the second-degree polynomial.

The simulation results conclude that the bootstrap method can improve the performance of complex and sensitive statistics where certain assumptions are not met. Applying the optimal smoothing parameter concludes that the two algorithms do not have a significant difference, and both local polynomial regressions do not show much difference. It counters the conclusion of open-ended questions by (Efron & Tibshirani, 1994), which conclude that paired bootstrapping has few disadvantages compared to residual bootstrapping.

We provide some suggestions for readers who wish to continue this article. Perhaps, the reader is interested in determining the BCA bootstrap confidence interval in a local polynomial regression model by adapting an existing procedure. Readers may also be interested in studying comparative studies, for example, between the bootstrap-t interval method and the bootstrap percentile, to determine the best interval between both. Another study that may be more interesting is the application of the wild bootstrap method to local polynomial regression prediction intervals where heteroscedasticity is present. In addition, readers can examine other topics related to the combination of the local polynomial regression concept with the bootstrap concept, which is the impact of this article. The priority for further research on bootstrap methods is no longer about polynomial degrees or smoothing.

ACKNOWLEDGEMENT

This research was funded by the PNBP Fund of the State University of Medan, through the Basic Research Scheme, Contract No. 0057 /UN33.8/PPKM/2022.

REFERENCES

Aguirre-Urreta, M., & Rönkkö, M. (2017). *Statistical Inference with PLSc Using Bootstrap Confidence Intervals*. https://www.researchgate.net/publication/315690307

Chernick, M. R., & LaBudde, R. A. (2014). *An Introduction to Bootstrap Methods with Applications to R*.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836. https://doi.org/10.1080/01621459.1979.10481038

Cleveland, W. S., Devlin, S. J., & Grosse, E. (1988). REGRESSION BY LOCAL FITIING Methods, Properties, and Computational Algorithms. In *Journal of Econometrics* (Vol. 37).

Cleveland, W. S., & Grosse, E. (1991). Computational methods for local regression. In *Statistics and Computing* (Vol. 1).

de Brabanter, K., de Brabanter, J., & de Moor, B. (2013). Derivative Estimation with Local Polynomial Fitting Irène Gijbels. In *Journal of Machine Learning Research* (Vol. 14).

Diciccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, *11*(3), 189–228.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis, Third Edition*.

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Eubank, R. L., & Speckman, P. L. (1993). Confidence Bands in Nonparametric Regression. In *Source: Journal of the American Statistical Association* (Vol. 88, Issue 424).

Fan, J., & Gijbels, I. (1960). Local Polynomial Modelling and Its Applications. In *Tensor Methods in Statistics P. McCullagh* (Vol. 21, Issue 2).

Gultom, F. R. P., Solimun, S., & Nurjannah, N. (2022). Bootstrap Resampling in Gompertz Growth Model with Levenberg–Marquardt Iteration. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *6*(4), 810. https://doi.org/10.31764/jtam.v6i4.8617

Hall, P., & Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *Annals of Statistics*, *41*(4), 1892–1921. https://doi.org/10.1214/13-AOS1137

Härdle, W., & Bowman, A. W. (1988). Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bands. *Journal of the American Statistical Association*, *83*(401), 102–110. https://doi.org/10.1080/01621459.1988.10478572

Jung, K., Lee, J., Gupta, V., & Cho, G. (2019). Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02215

Mansyur, A., & Simamora, E. (2022). Bootstrap-t Confidence Interval on Local Polynomial Regression Prediction. *Mathematics and Statistics*, 10(6), 1178–1193. https://doi.org/10.13189/ms.2022.100604

Özdemir, A. F. (2013). Comparing two independent groups: A test based on a one-step M-estimator and bootstrap-t. *British Journal of Mathematical and Statistical Psychology*, *66*(2), 322–337. https://doi.org/10.1111/j.2044-8317.2012.02053.x

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's Comments: A Critical Look at the Use of PLS-SEM in "MIS Quarterly" Author(s). In *Source: MIS Quarterly* (Vol. 36, Issue 1).

Simamora, E., Subanar, & Kartiko, S. H. (2015). Asymptotic property of semiparametric bootstrapping kriging variance in deterministic simulation. *Applied Mathematical Sciences*, *9*(49–52). https://doi.org/10.12988/ams.2015.52104

Solci, C. C., Reisen, V. A., Rodrigues, P. C., Solci, C. C., & Reisen, V. A. (2022). *Robust Local Bootstrap for WeaklyStationary Time Series in the Presence ofAdditive Outliers*. https://doi.org/10.21203/rs.3.rs-2054445/v1

Wasserman, L. (2004). *All of Nonparametric Statistics* (G. Casella, S. Fienberg, & I. Olkin, Eds.; 1st ed.). Springer New York. https://doi.org/10.1007/978-0-387-21736-9

Wasserman, L. (2006). *All of Statistics: A Concise Course in Statistical Inference* (Vol. 26). Springer New York. https://doi.org/10.1007/978-0-387-21736-9

Xia, Y. (1998). Bias-Corrected Confidence Bands in Nonparametric Regression. In *Source: Journal of the Royal Statistical Society. Series B (Statistical Methodology)* (Vol. 60, Issue 4). https://www.jstor.org/stable/2985963