Indonesian Journal of Science & Technology

Journal homepage: http://ejournal.upi.edu/index.php/ijost/

# Adaptive Robust Confidence Bands on Local Polynomial Regression Using Residual Bootstrap Percentiles

*Abil Mansyur*, Elmanani Simamora, Muliawan Firdaus, Tiur Malasari Siregar, Rizki Habibi*

Universitas Negeri Medan, Indonesia

*Correspondence: E-mail: abil@unimed.ac.id

## ABSTRACT

Ensuring reliable inference in local polynomial regression requires robust methods that can manage data irregularities, particularly outliers. This study introduces an adaptive robust approach for constructing confidence bands using residual bootstrap percentiles. Two robust weighting techniques (Huber and Tukey) were applied to address different levels of data contamination. The method was evaluated using both simulated datasets and real-world observations involving fluctuating patterns. Huber weighting produced more stable and narrower confidence bands under moderate anomalies, while Tukey weighting was more effective in handling extreme deviations. These differences arise because Huber downweights moderate residuals proportionally, whereas Tukey aggressively suppresses extreme outliers. Smoothing parameters were optimized through cross-validation to balance bias and variance effectively. This approach enhances the robustness of nonparametric regression because it maintains consistent confidence coverage despite data imperfections, offering a reliable tool for statistical inference in complex datasets.



## ARTICLE INFO

***Article History:***

*Submitted/Received 08 Feb 2025*

*First Revised 19 Mar 2025*

*Accepted 26 May 2025*

*First Available Online 26 May 2025*

*Publication Date 01 Sep 2025*

***Keyword:***

*Huber weights,*

*Local polynomial regression,*

*Outliers,*

*Residual bootstrap percentiles,*

*Robust confidence bands,*

*Tukey weights.*

## 1. INTRODUCTION

Local polynomial regression is a flexible non-parametric method that can capture non-linear patterns in the relationship between predictor and response variables. Unlike parametric models that assume a particular functional form, this method adjusts the estimation structure based on local data, making it suitable for analyzing datasets with irregular trends or containing noise. Combining moving averages and polynomial regression allows this technique to produce smooth and accurate estimates across data segments [1].

The flexibility of local polynomial regression has been applied in various fields. Some researchers used this method to detect extreme points in time series data, balancing between error and model complexity. In chemistry and biology, some researchers [2] demonstrated this method in predicting biological responses and toxicity, especially in complex datasets. The choice of optimal smoothing parameters can affect the performance of non-parametric regression and provide more accurate results in dealing with complex data [3]. The methods for robust regression are widely used to handle outliers and heteroscedasticity, as further discussed in the Methods section.

To improve the robustness to outliers, some researchers [4] introduced the Robust Local Weighted Regression approach, which uses adaptive weights based on the distance between the data points and the focus of estimation. Then, the robust local polynomial regression curve surface provides a smoother curve on data containing outliers. The principle of its work is an iterative approach, these weights are updated using residuals from the previous model, thereby increasing the robustness of the estimation to the influence of extreme data.

In the context of confidence bands, bootstrapping has proven to be an effective method, especially for non-parametric regression. Some researchers [5] introduced the percentile bootstrap method to construct confidence bands based on the distribution of resampling results. This approach requires no particular distributional assumptions, making it ideal for non-linear and complex data. Some researchers [6] noted that naïve bootstrap methods often result in inadequate coverage for confidence bands, particularly for data with high variability.

Some researchers [7,8] developed the bootstrap-t and percentile methods for local polynomial regression, which proved reliable in providing band estimates even at small sample sizes. Using bootstrap residuals, their study showed that the coverage of band probabilities can be significantly improved, especially through uncertainty simulation based on normally distributed true functions.

Several researchers have developed adaptive approaches in the development of local polynomial regression. The balance of bias and variance of adaptive modeling requires special attention in heteroscedastic data or irregular trends. Some researchers proposed penalized splines to regulate model complexity, providing flexibility in handling data fluctuations. Some researchers [9] developed a shape-constrained approach, while other researchers [10] advanced isotonic quantile regression, both improving probability coverage for nonparametric datasets.

We extend these advances by introducing an adaptive robust framework for constructing bootstrap residual percentile confidence bands in local polynomial regression. The proposed method integrates data-driven adjustment with Tukey and Huber weights, focusing on parameter tuning such as the weight cutoff and optimal smoothing parameters. These parameter tunings balance bias and variance and ensure consistent nominal coverage. The robustness and adaptability of the method are validated through simulations with controlled outlier proportions (5%, 10%, 15%) and real-world data from Kualanamu International

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

Airport. It contributes to the development of reliable statistical tools for real-world applications.

The remainder of this article will cover several main sections. The methodology section explains the concept and theory of robust local polynomial regression and the proposed approach for confidence band construction using percentile bootstrap residuals. This section also includes a hierarchical explanation of the algorithm, detailing the steps of adaptive confidence bands construction with Tukey and Huber weight integration. Next, simulation results are presented in two types of data: synthetic data generated through experimental simulations to control the proportion of outliers and noise, and real-world data from Kualanamu International Airport used to evaluate the method's applicability in practical situations. These results are analyzed to assess the ability of the bootstrap residual approach to ensure consistent coverage of confidence bands despite data fluctuations and extreme points during the COVID-19 pandemic. Finally, the concluding section summarizes the main findings, identifies the strengths of the proposed method, and provides recommendations for further development, both in the context of statistical methodology and real-world applications.

## 2. METHODS

The methodology proposed in this paper includes implementing robust weighting methods and a residual bootstrap algorithm specifically designed to handle data with complex characteristics, such as outliers and heteroscedasticity. This section consists of two main subsections. The first subsection discusses the basic concepts of robust local polynomial regression, including adjusting weighting parameters to improve robustness to extreme data. The second subsection describes the framework for robust residual bootstrap-based confidence band construction. This section systematically describes the implementation steps to achieve optimal confidence band construction consistent with the expected nominal probability coverage.

### 2.1. Robust Local Polynomial Regression

This section introduces two fundamental concepts: local polynomial regression and robust local polynomial regression. Local polynomial regression is a nonparametric model in which the regression function is approximated by locally fitting a polynomial to the data within the predictor space. In this approach, observations closer to the point of interest are given higher weights, which helps the model make more accurate estimates near the point of estimation [11]. There are two parameters for smoothing a scatterplot: the smoothing parameter and the polynomial degree. The smoothing parameter influences the fit of local regression, where its value ranges between zero and one. If $\alpha$ is too small, it approaches zero, the scatterplot smoothing becomes convoluted (noisy), which means that the number of nearest neighbors (k-Nearest Neighbors) included in the smoothing span is insufficient, so the variance becomes large. On the other hand, if $\alpha$ is too close to one, the scatterplot smoothing becomes smoother. However, the local polynomial regression may not fit the data within the smoothing span, so essential features of the averaging function may be distorted or lost altogether, which will have a significant bias. It is possible to choose the smoothing parameter $\alpha$ based on the researcher's subjectivity or search for optimal parameters based on a compromise of the bias-variance trade-off. Therefore, an $\alpha_{optimal}$ must be sought to balance bias and diversity to obtain a reasonable estimate.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

The choice of polynomial degree must also be considered. Researchers recommend low-degree polynomials such as degrees one and two. It is better if the selection of a high degree impacts greater diversity, even though the modeling bias is reduced. The choice of polynomial degree tends to be based on the researcher's wishes (subjectivity) by considering the characteristics of the data. We only focus on local polynomial regression of degree two and focus more on smoothing parameters. Several researchers [7,8] have shown that second-degree polynomials are more than sufficient to capture data patterns.

Next, we focus on the local polynomial regression model where bivariate data is available, which is expressed by mutually independent ordered pairs, $(x_1, y_1),\ (x_2, y_2), \cdots, (x_n, y_n)$. The function that relates $x_i$ to $y_i$ can be expressed in Equation (1),

$$y_i = m(x_i) + \varepsilon_i,\ i = 1, 2, \ldots, n \tag{1}$$

where $E(y_i|x_i) = m(x_i)$ and m is a smooth real-valued function that is unknown but will be estimated. The error term $\varepsilon_i$ is an independently distributed random variable with $E(\varepsilon_i|x_i) = 0$ and $Var(\varepsilon_i|x_i) = \sigma^2(x_i)$. If it is assumed that the function m is a continuous function and can be differentiated up to order $(p + 1)$, then the function $m(x_i)$ can be expressed as an approximation of the Taylor series expansion on $x_0$ in Equation (2),

$$m(x_i) \approx m(x_0) + m'(x_0)(x_i - x_0) + \frac{m''(x_0)}{2!}(x_i - x_0)^2 + \cdots \frac{m^p(x_0)}{p!}(x_i - x_0)^p \tag{2}$$

where $x_i \in \mathbb{N}(x_0)$ with $\mathbb{N}(x_0)$ defined as a set of data points in the vicinity, or often expressed as the nearest neighboring data points of $x_0$. Using the k-Nearest Neighbors (k-NN) principle, the regression function at a point $x_0$ is estimated by considering the weighted sum of squared errors, where the weights are determined by the distance between $x_0$ and the neighboring points in the predictor space. The k-NN principle plays a key role in identifying the local neighborhood for regression estimation. The local polynomial regression estimation problem is then reduced to estimating the polynomial regression in the neighborhood $\mathbb{N}(x_0)$ by minimizing the weighted sum of squared errors in Equation (3), as discussed by other reports [11].

$$\sum_{i=1}^{k} w_i(x_0) \left\{ y(x_i) - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right\}^2 \tag{3}$$

The weighting value $w_i(x_0)$ is obtained from the function $w_i(x_0) = W(|x_i - x_0|/\Delta(x_0))$ where $\Delta(x_0)$ is the maximum Euclidean distance $x_0$ to a point $x_i \in \mathbb{N}(x_0)$. By supposing $u = |x_i - x_0|/\Delta(x_0)$, the weight function W has the following properties [4],

(i) $W(u) > 0$, for $-1 < u < 1$;
(ii) $W(-u) = W(u)$;
(iii) $W(u)$ is a non-increasing function for $u \geq 0$;
(iv) $W(u) = 0$, for $u \leq -1$ or $u \geq 1$.

Several selections of weight functions can be found in [3]. This research does not concentrate on the weight function but rather on case studies that have been determined. We choose the tricube weight function as in Equation (4) [4].

$$W(u) = \begin{cases} (1 - |u|^3)^3, \text{for } |u| < 1 \\ 0 \quad , \text{ for } |u| \geq 1 \end{cases} \tag{4}$$

Other researchers derive the solution of Equation (3) in the form of a matrix equation in Equation (5),

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} = (\mathbf{X}_{\boldsymbol{\alpha}}^{\mathbf{T}} \mathbf{W}_{\boldsymbol{\alpha}} \mathbf{X}_{\boldsymbol{\alpha}})^{-1} \mathbf{X}_{\boldsymbol{\alpha}}^{\mathbf{T}} \mathbf{W}_{\boldsymbol{\alpha}} \mathbf{Y}_{\boldsymbol{\alpha}} \tag{5}$$

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

where, $\mathbf{X}_\alpha = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & \cdots & \vdots \\ 1 & (x_k - x_0) & \cdots & (x_k - x_0)^p \end{pmatrix}$, $\mathbf{Y_\alpha} = (y_1, \ldots, y_k)'$, $\hat{\boldsymbol{\beta}}_\alpha = (\hat{\beta}_0, \ldots \hat{\beta}_p)'$

and $\mathbf{W}_\alpha = \begin{pmatrix} w_1(x_0) & 0 & 0 & 0 \\ 0 & w_2(x_0) & 0 & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & \cdots & w_k(x_0) \end{pmatrix} = \mathrm{diag}\big(w_1(x_0), \ldots, w_k(x_0)\big).$

Based on other reports [7], the predicted point $x_0$ following Equation (6) is,

$$\hat{y}(x_0) = \sum_{j=0}^{p} \hat{\beta}_j \, x_0^j \tag{6}$$

The above procedure is repeated to predict specific points of interest, for example, the vector $\mathbf{X} = (x_1, \ldots x_n)'$. Next, we derive a robust local polynomial regression by first finding the local polynomial regression residual, $\hat{\varepsilon}_i = y_i - \hat{y}_i$. Then the robust weight is obtained from Equation (7),

$$r_i = W\left(\frac{\hat{\varepsilon}_i}{6s}\right) \tag{7}$$

where $s$ represents the median of $|\hat{\varepsilon}_i|$. The above procedure is repeated using Equation (3), but the weights are now, $r_i w_i$. By considering data points in $\mathbb{N}(x_0)$ to perform a curve fit. The iterative process conforms to the principle of iteration in optimization so that the regression curve converges or stops changing. The researcher's convergence tolerance determines whether to stop the iteration. The simulation only needs two or three iterations to get a reasonable model (see **Table 1**) [12].

**Table 1**. The algorithm below shows the steps for a robust local polynomial regression prediction.

| **Robust Local Polynomial Regression Algorithm** |
|---|
| 1. Perform data matching using a local polynomial regression procedure with the selection of the weight function W; |
| 2. Perform residual calculations, $\hat{\varepsilon}_i = y_i - \hat{y}_i$, for each data point; |
| 3. Determine *s* which is the median of $|\hat{\varepsilon}_i|$; |
| 4. Determine the robust weight $r_i$; |
| 5. Perform the local polynomial regression procedure again, but use the weights $r_i w_i$; |
| 6. Repeat steps 2 to 5 until the local polynomial regression predictions converge within the given tolerance. |

### 2.2. Proposed Robust Bootstrap Confidence Band Technique

We propose a robust bootstrap confidence band technique through several stages. The first stage searches for the optimal smoothing parameters ($\alpha_{optimal}$) based on the available data. The principle of searching for $\alpha_{optimal}$ considers the variance-bias trade-off. Too small a smoothing parameter results in high variance and low bias, while too large a parameter results in low variance and high bias. By determining the $\alpha_{optimal}$, the model can effectively capture the underlying data pattern while remaining robust to noise and outliers, thus achieving an ideal balance between bias and variance. A smoothing parameter search algorithm using Cross-Validation [7]. We use their provisions to obtain optimal smoothing parameters using the data, where the search for optimal smoothing parameters using Equation (8) is expressed by $\alpha_{optimal}$,

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

$$\alpha_{optimal} = \arg\min_{\alpha \in I}\left(CV(\alpha)\right), \quad I = (0,1) \tag{8}$$

where, $CV(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(y(x_i) - \hat{y}_{\alpha}^{-i}(x_i)\right)^2$ and $\hat{y}_{\alpha}^{-i}(x_i)$ is the model prediction at point $x_i$ by deleting one data point.

In the second stage, the robust local polynomial regression algorithm applies a data-matching process using the optimal smoothing parameters and low polynomial degrees. This process connects the predictor data $x_i$ with the predicted results $\hat{y}_i$. The result is a pair $(x_i, \hat{y}_i)$, which is used to calculate the residual $e_i$ and bootstrap. This step ensures that the regression model produces predictions that match the characteristics of the original data, just for information for readers that $e_i \neq \hat{\varepsilon}_i$. The notation $\hat{\varepsilon}_i$ indicates the residual of the local polynomial regression, while $e_i$ is the residual of the robust local polynomial regression.

The third stage performs a residual transformation with the Median Absolute Deviation (MAD) to be more robust to outliers than the traditional standard deviation. After the MAD is calculated, the residuals are transformed into standardized residuals using Equation (9),

$$\tilde{e}_i = |e_i|/MAD \tag{9}$$

where $MAD = 1.48261 \times \text{median}(|e - \text{median}(e)|)$ with e being the residual vector. The coefficient 1.48261 is used to adjust the MAD to be consistent with the standard deviation of the data following a normal distribution.

The fourth stage applies the robust approach (Huber and Tukey) to handle outliers. The Huber weights in Equation (10) are a robust method that integrates the advantages of two different approaches: 1) a linear approach where residuals with small values are treated as in conventional regression, thus giving them full weight, and 2) a constant approach where residuals with large values are given lower weights to reduce their impact on the model, making them more resistant to the influence of outliers. The Huber weight formulation is given in Equation (10),

$$w_i = \begin{cases} 1, & \text{if } |\tilde{e}_i| \leq c \\ c/|\tilde{e}_i|, & \text{if } |\tilde{e}_i| > c \end{cases} \tag{10}$$

where *c* is the cutoff that controls the transition between full weight (w = 1) and reduced weight; meanwhile, Tukey weights in Equation (11) are more aggressive in handling outliers, where residuals greater than the cutoff c are ignored or given zero weight (w = 0). The Tukey weight formulation is given in Equation (11).

$$w_i = \begin{cases} (1 - (\tilde{e}_i/c)^2)^2, & \text{if } |\tilde{e}_i| \leq c \\ 0, & \text{if } |\tilde{e}_i| > c \end{cases} \tag{11}$$

At this stage, the standardized residuals are adjusted by applying weights calculated using the Huber or Tukey method. This adjustment aims to produce weighted residuals, $\hat{e}_i = w_i \times \tilde{e}_i$, reducing the influence of outliers in the analysis and enhancing the accuracy of the primary data estimates.

The fifth stage involves the bootstrap sampling process using the weighted residuals to generate bootstrap-based confidence bands. **Table 2** presents an algorithm for constructing prediction confidence bands in local polynomial regression using the bootstrap residual percentile approach.

The theoretical basis of the Bootstrap Confidence Band in Equation (12) is summarized in the following theorem, which rigorously demonstrates the convergence property of the double bootstrap residuals. This result ensures the statistical validity of the confidence bands

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

generated by the algorithm, especially in achieving the specified coverage probability (CP) as the number of bootstrap iterations approaches infinity.

**Table 2**. Algorithm for constructing prediction confidence bands in local polynomial regression using the bootstrap residual percentile approach.

| **Bootstrap Confidence Band Algorithm** |
|---|
| 1. Determining the first bootstrap sample, $\hat{y}_i^* = \hat{y}_i + \hat{e}_i^*$, where $\hat{e}_i^*$ is a resampling with replacement from the sequence $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_n$; |
| 2. Matching the first bootstrap sample to Equation (6) to get the second bootstrap sample, $\hat{y}_i^{**}$, for $i = 1,2, \dots, n$; |
| 3. Performing double bootstrap calculations, $\hat{e}_i^{**} = \hat{y}_i^* - \hat{y}_i^{**}$ and normalizing, $\tilde{e}_i^* = \hat{e}_i^{**} - \frac{\sum_{j=1}^{n} \hat{e}_j^{**}}{n}$, for $i = 1,2, \dots, n$; |
| 4. Carry out independent random sampling with returns from $\tilde{e}_i^*$ to get $\mathbf{e}^{**} = (e_1^{**}, e_2^{**}, \dots, e_n^{**})$; |
| 5. Repeat the first step to the fourth step B times to obtain the sequence of bootstrap sample vectors, $\mathbf{e}^{**1}, \mathbf{e}^{**2}, \cdots \mathbf{e}^{**B}$; |
| 6. Determine the bootstrap estimate of the Prediction Confidence Band (PCB) for each data point, $x_i$ using Equation (12), |
| $\text{PCB}: (y_i): \hat{y}_i + \hat{H}^{-1}_{(\alpha/2)}(e_i) \le y_i \le \hat{y}_i + \hat{H}^{-1}_{(1-\alpha/2)}(e_i)$ (12) |
| where $\hat{H}^{-1}_{(1-\alpha/2)}(e_i) = e_i^{**B(1-\alpha/2)}$ is the $(1-\alpha/2)$-th bootstrap percentile of the estimated cumulative density distribution. |

**Theorem**. Suppose the double bootstrap estimate for the residual $e_i$ is a sequence of resampling statistics $e_i^{**1}, e_i^{**2}, \cdots, e_i^{**B}$. If the distribution of the resampling statistics is approximately a normal distribution with B tending to infinity, then the probability of:

$$P\left(\hat{y}_i + H^{-1}_{(\alpha/2)}(e_i) \le y_i \le \hat{y}_i + H^{-1}_{(1-\alpha/2)}(e_i)\right) \to 1 - \alpha,$$

where $H^{-1}_{(1-\alpha/2)}(e_i) = e_i^{**B(1-\alpha/2)}$ is the $(1-\alpha/2)$-th bootstrap percentile of the estimated cumulative density distribution.

**Proof of Theorem.** Let $H_{e_i}$ representing the cumulative distribution function (CDF) of the bootstrap sample sequence $e_i^{**1}, e_i^{**2}, \cdots, e_i^{**B}$. According to the assumption, as $B \to \infty$, the distribution of $e_i^{**j}$ converges to a normal distribution. Consequently, the empirical CDF $H_{e_i}$ asymptotically approximates the true normal CDF, denoted by $H_{\varepsilon_i}$. Since $H_{e_i}$ is a monotonic increasing function, the bootstrap quantile corresponding to $1-\alpha/2$, i.e., $H^{-1}_{(1-\alpha/2)}(\varepsilon_i)$, is the value that separates the top $1-\alpha/2$ proportion of the bootstrap sample. If $\varepsilon_i \sim N(\hat{\mu}, \hat{\sigma}^2)$, the quantile at $1-\alpha/2$ is approximately given by:

$$H^{-1}_{(1-\alpha/2)}(\varepsilon_i) \approx \hat{\mu} + z_{(1-\alpha/2)}\hat{\sigma},$$

where $\hat{\mu}$ is the sample mean, $\hat{\sigma}$ the sample standard deviation, and $z_{(1-\alpha/2)}$ is the standard normal quantile. Thus, the probability in the theorem can be written as:

$$P\left(\hat{y}_i - H^{-1}_{(\alpha/2)}(\varepsilon_i) \le y_i \le \hat{y}_i + H^{-1}_{(1-\alpha/2)}(\varepsilon_i)\right) = P\left(-H^{-1}_{(\alpha/2)}(\varepsilon_i) \le \varepsilon_i \le H^{-1}_{(1-\alpha/2)}(\varepsilon_i)\right)$$

Since $H_{e_i}$ approximates $H_{\varepsilon_i}$ for large B, it becomes:

$$P\left(\hat{\mu} - z_{(\alpha/2)}\hat{\sigma} \le \varepsilon_i \le \hat{\mu} + z_{(1-\alpha/2)}\hat{\sigma}\right) = P\left(-z_{(\alpha/2)} \le (\varepsilon_i - \hat{\mu})/\hat{\sigma} \le z_{(1-\alpha/2)}\right)$$

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

since $(\varepsilon_i - \hat{\mu})/\hat{\sigma} \sim N(0,1)$, the probability becomes: $P\left(-z_{(\alpha/2)} \leq Z \leq z_{(1-\alpha/2)}\right) = 1 - \alpha$, where Z is a standard normal random variable.

## 3. RESULTS AND DISCUSSION

### 3.1. Result

We use two data types to analyze the adjustment of boundary parameters on Huber and Tukey weights based on data characteristics to evaluate the application of the robust bootstrap confidence band technique. The first source is empirical data on the number of domestic flight passengers at Kualanamu Airport from January 2006 to March 2024. This local data was chosen because it captures fluctuations often driven by seasonal factors, long-term trends, and anomalies such as the COVID-19 pandemic. The second source is computer simulation data designed to simulate extreme conditions by adding outliers in a controlled manner. The addition of outliers aims to evaluate how boundary parameters adjust to the gradually increasing level of data deviation.

#### 3.1.1. Real Data

**Figure 1** presents two main visualizations illustrating the process and results of robust local polynomial regression analysis with a polynomial degree $p = 2$. **Figure 1a** shows the search for optimal smoothing parameters using the Cross-Validation method, while **Figure 1b** depicts the distribution of residuals derived from the applied regression model.

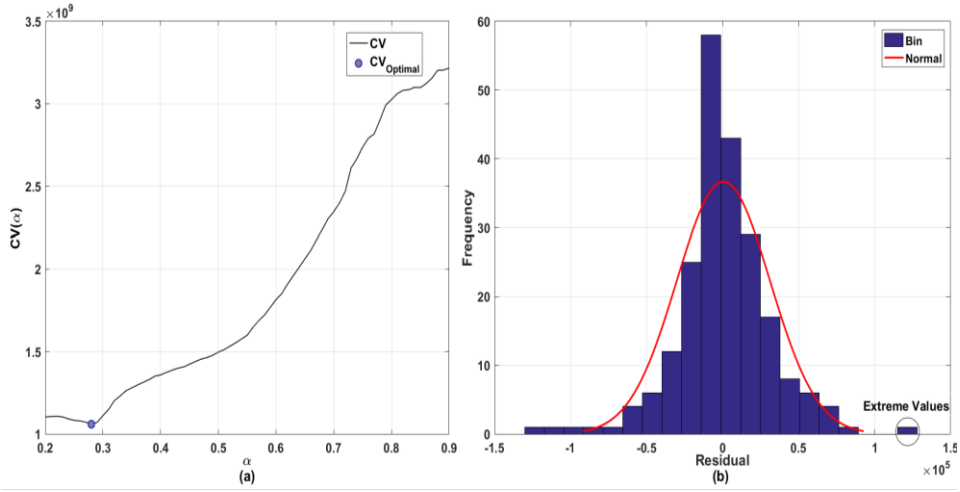


**Figure 1.** Smoothing parameter search and residual histogram using robust local polynomial regression.

**Figure 1a** depicts the relationship between the smoothing parameter ($\alpha$) and the Cross-Validation function CV($\alpha$). The blue circles on the curve mark the smoothing parameter values that achieve an ideal balance between bias and variance, $\alpha_{optimal}$ is 0.28. Values of $\alpha$ close to 0 increase variance and risk of overfitting as the model becomes too responsive to small data fluctuations. Values of $\alpha$ close to 1 reduce variance but risk over-smoothing, which can hide critical local variations in the data. The search for smoothing parameters uses the Cross-Validation function in Equation (8), which ensures that the selection process is based on the intrinsic characteristics of the dataset and remains unbiased. This approach minimizes subjective decisions and allows the model to adapt effectively to the underlying data structure.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

**Figure 1b** depicts the distribution of residuals through a histogram, which shows the difference between the actual values ($y_i$) and the predicted values from the robust local polynomial regression ($\hat{y}_i$) of the model. Most residuals are concentrated around 0, indicating that the model predictions are generally accurate. We include the red curve showing the normal distribution as a reference for evaluating the residuals. However, some extreme residuals (outliers) are visible at both ends of the histogram, highlighting the importance of adjusting the boundary parameters in the Huber or Tukey weighting function to reduce the influence of these anomalies.

We apply the confidence band technique proposed in the previous section using a sample size of $B = 100{,}000$ and an optimal smoothing parameter $\alpha_{optimal} = 0.28$. Consideration of the bootstrap sample size $B = 10{,}000$ is used to obtain an ideal bootstrap estimator [5]. In addition, a tolerance of 0.001 ensures that the optimal cutoff is precisely selected according to the nominal CP target of 0.95. This approach ensures the reliability and precision of the resulting confidence bands, even under difficult data conditions. **Figure 2** provides two main views explaining the relationship between cutoff and CP and PCB using the optimal cutoff of the Huber and Tukey methods on passenger data.

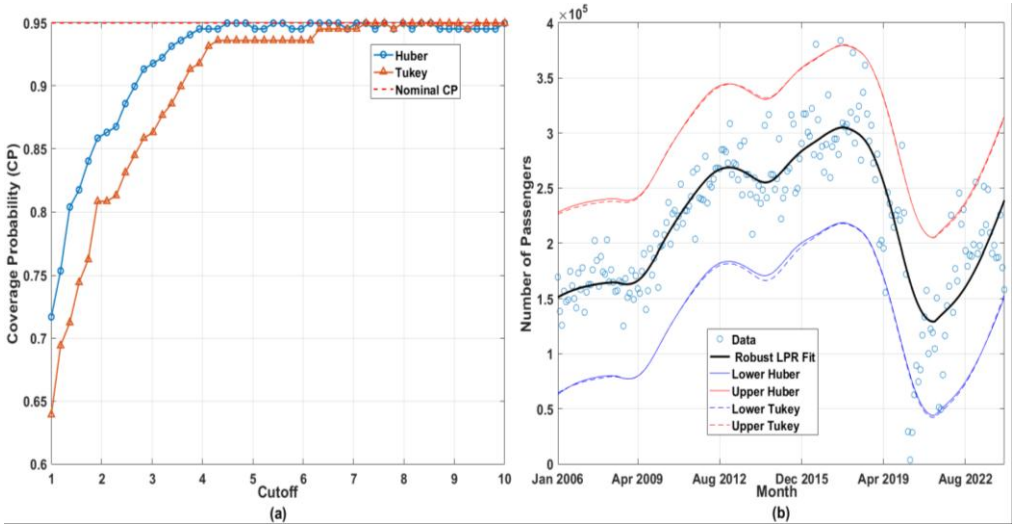


**Figure 2.** Coverage probability and bootstrap confidence bands using Huber and Tukey methods on passenger count data.

In **Figure 2a**, the relationship between the cutoff value and CP is depicted while both methods converge to the nominal CP. Huber's process can get closer to the nominal CP with a relatively low cutoff value due to its ability to deal with moderate outliers. Tukey's method needs a higher cutoff value to get to the same CP level as Huber's method, which shows how Tukey's method is more conservative in dealing with extreme outliers. These differences show that optimizing the cutoff value is crucial in enhancing the robustness and the precision of statistical modeling, mainly where there are differences in data quality. Huber's curve (marked by the blue circle) shows that CP increases with the increasing cutoff value until it reaches the nominal CP target (marked by the red dashed line). The optimal cutoff for Huber is 4.49, appears sharper at the beginning, and shows a better response to cutoff changes in the early stage. Tukey's curve (marked by the orange triangle) shows CP, which increases more slowly than Huber's. The optimal cutoff for Tukey is 7.25, which is higher than Huber's in achieving the CP target. Achieving Tukey's optimal limits indicates greater tolerance for extreme outliers.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

**Figure 2b** shows the bootstrap estimates for PCB from the two methods, applying the optimal cutoff. The bootstrap estimates for the confidence bands from Huber's method tend to be narrower, reflecting its sensitivity to data more profound in the principal distribution. In contrast, Tukey's method produces wider bands due to its conservative approach to excluding outliers. The difference between the two methods suggests a trade-off between precision and robustness in outlier management.

### 3.1.2. Artificial Data

In this section, we want to reinforce the results achieved on actual data through artificial data. The design for predictor X uses a set of values evenly distributed in the range $[-10,10]$ with a total of n = 100 data points. The generation of response data Y uses a sine function (X) added with a slight Gaussian noise with low variance, $\mathrm{Y}\sim\mathrm{N}(0,\ 0.1)$, to reflect the variance in real-world data. Then, the simulation controls outliers starting from 0% (no outliers), 5%, 10%, and 15%. These outliers are simulated by adding a large-scale Gaussian noise, $\mathrm{N}(0,25)$, to the response values $\mathrm{Y}$ at several randomly selected indices.

**Table 3** presents the optimal smoothing parameters with the polynomial degree $\mathrm{p} = 2$ based on artificial data. In data without outliers, the $\alpha_{optimal}$ value = 0.2 indicates minimal smoothing needs because the data is considered clean and homogeneous. The minimal $\mathrm{CV}(\alpha)$ value, which is 0.0113, reflects excellent prediction quality because there is no interference from outliers. When the percentage of outliers increases to 5%, the $\alpha_{optimal}$ value increases to 0.37. This increase indicates that the model requires more significant smoothing to suppress the impact of outliers. The minimum $\mathrm{CV}(\alpha)$ value also increases to 0.557, indicating that the prediction quality begins to decline due to outliers in the data. At an outlier percentage of 10%, the $\alpha_{optimal}$ value decreases again to 0.24, while the minimum $\mathrm{CV}(\alpha)$ value increases significantly to 1.719. The decrease in the $\alpha_{optimal}$ value reflects the model's efforts to adapt to more complex data due to outliers that begin to dominate. The minimum $\mathrm{CV}(\alpha)$ increase confirms that outliers significantly affect prediction quality. At the highest outlier level, 15%, the $\alpha_{optimal}$ value decreases further to 0.22, while the minimum $\mathrm{CV}(\alpha)$ value reaches 2.812. The decrease in $\alpha_{optimal}$ indicates that the model tries to be more sensitive to the primary data, although the disturbance from outliers still reduces the overall prediction quality. The simulation results reflect that outliers affect the optimal smoothing parameters and the prediction quality. The $\alpha_{optimal}$ value changes adaptively to balance smoothing and sensitivity to outliers.

**Table 3.** Optimal smoothing parameters using robust local polynomial regression

| | **Outlier Percentage** | | | |
|---|---|---|---|---|
| | **0%** | **5%** | **10%** | **15%** |
| $\alpha_{optimal}$ | 0.2000 | 0.370 | 0.240 | 0.220 |
| Minimum(CV(α)) | 0.0113 | 0.557 | 1.719 | 2.812 |

We use the results in **Table 3** to perform simulations to estimate robust bootstrap confidence bands. The simulations use B = 10,000 bootstrap samples to assess the distribution of the weighted residuals. With this distribution, bootstrap confidence bands are calculated based on the upper and lower quantiles of the residuals. These confidence bands are then used to calculate CP, the proportion of data within the band. Using this approach, we can evaluate the performance of the robustness method and determine the optimal threshold that ensures CP is close to the nominal target of 0.95 with an adjusted tolerance. The

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

simulation results are visualized to facilitate the interpretation of the relationship between the threshold, CP, and confidence band.

**Figure 3** provides two main perspectives on how the optimal threshold affects the performance of robust analysis in the face of outliers. **Figure 3a** depicts the estimation results and uncertainty on outliers-free data, while **Figure 3b** shows the interaction between the threshold and CP on data contaminated by 5% outliers. Thus, this figure emphasizes the importance of choosing the right optimal threshold and robust strategy to achieve accurate and reliable estimation.

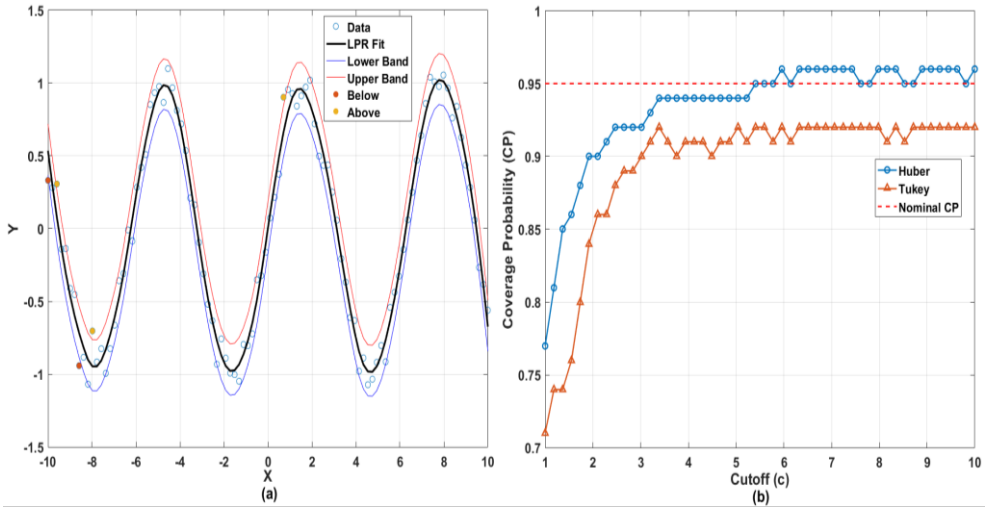


**Figure 3.** Bootstrap confidence bands without and 5% outliers.

**Figure 3a** applies a non-robust local polynomial regression model because the artificial data does not contain outliers, or the percentage of outliers is 0%. The black primary curve shows the model prediction results, which follow a sine function pattern considering the existing noise. The two blue and red curves represent the lower and upper limits of the confidence bands, respectively, which are calculated based on the bootstrap approach. Most of the data are within the confidence bands, reflecting that the model can capture the main characteristics of the data distribution. Some data points are outside the band boundaries; above and below represent noise (uncertainty). With a confidence level of 0.05, the bootstrap confidence bands provide a CP of 0.95, which matches the nominal coverage probability. For comparison, the naïve bootstrap approach also provides a CP of 0.95, independent of the cutoff value (c), as expected in scenarios without outliers.

**Figure 3b** visualizes the relationship between the cutoff value and CP when the data contains 5% outliers. The red horizontal line shows the nominal target CP of 0.95. The blue and orange lines depict the performance of the Huber and Tukey methods in achieving the target CP, respectively. The Huber method achieves the target CP with an optimal cutoff of 5.41, while the Tukey method fails to reach the nominal coverage in the given cutoff domain. We conclude that the Huber method shows its sensitivity to moderate outliers. That is, the Huber method does not directly ignore moderate outliers but only reduces the contribution of moderate outliers to the model proportionally. In contrast, the Tukey method requires a higher cutoff to achieve the target CP, reflecting its more conservative approach to dealing with extreme outliers. That is, the Tukey method has a strict policy against extreme outliers. Once a residual passes the cutoff, the outlier is considered “noise” or irrelevant error and removed from the analysis (weighted zero).

We increase the cutoff domain for cases where the data contains 10% and 15% outliers in the hope that the Tukey method can achieve the target CP. **Figure 4** presents an in-depth

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

analysis of the relationship between the cutoff value and CP with outliers in the data. **Figure 4a** shows where 10% of the data contains outliers, while **Figure 4b** illustrates the situation with 15% outliers. Both figures provide a comprehensive overview of the performance of the Huber and Tukey methods in achieving a nominal CP of 0.95.

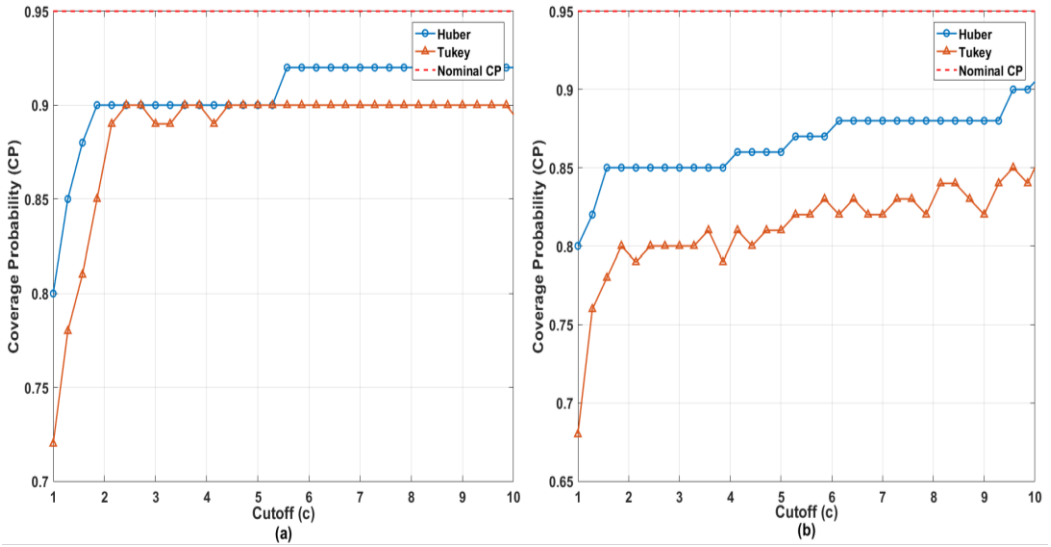


**Figure 4.** Coverage probability using the Huber and Tukey methods with 10% and 15% of data containing outliers.

In **Figure 4a**, Huber's method demonstrates its ability to achieve nominal CP quickly with a relatively low threshold. That is, Huber's method reflects its efficiency in balancing sensitivity to the underlying data while handling moderate outliers. In contrast, Tukey's method requires a higher threshold to achieve the target CP. Tukey's method takes a more conservative approach to handling extreme outliers, prioritizing protection against distortion even at the expense of a slower CP achievement rate.

**Figure 4b** shows how both methods adapt to an increase in the number of outliers to 15%. Huber's method maintains its relative efficiency despite requiring a slightly higher cutoff than in the 10% outlier condition. On the other hand, Tukey's method shows more significant fluctuations in CP before finally approaching the nominal value. The Tukey method is more sensitive to the high proportion of extreme outliers in the data.

### 3.2. Discussion

The proposed robust bootstrap confidence band technique contributes to developing robust statistical methods through Huber and Tukey approaches. We demonstrate the technique's ability to deal with data with outliers and noise, challenges often encountered in real-world data analysis. In comparison, Huber's approach is more sensitive to moderate outliers, while Tukey's method is conservative for dealing with extreme outliers. The simulation findings show that robust processes are adaptable to different data conditions and present the right strengths for various analysis needs for the following applications.

One of the key aspects of this proposal is the selection of optimal smoothing parameters, which significantly affect the bias and variance of the estimates. Artificial data simulations are used to evaluate the effect of the proportion of outliers. At the same time, additional validation is performed on real-world data, such as the number of airline passengers at Kualanamu International Airport. This approach confirms that the robust method is suitable for controlled data conditions and effective in handling complex data.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

The bootstrap confidence bands results indicate that the Huber method, with its narrower bands, effectively captures the essential structure of the data. This equilibrium between precision and robustness is crucial in applications that must balance high estimation accuracy with managing extreme data variations.

We also want to highlight some limitations, especially regarding the sensitivity of the choice of cutoff and the number of bootstraps used. Traditional approaches, such as naïve bootstrapping, tend not to perform well in preserving CP on data with outliers. Naïve bootstrapping often fails to maintain estimation accuracy because it does not integrate adaptive weights that can handle outliers. In this context, naïve bootstrap methods, described by [13], can produce unreliable estimates, particularly when the data contains outliers. This is because naïve bootstrap tends to sample more outliers than the original data, distorting the empirical distribution. In contrast, robust methods such as Huber and Tukey show superiority in handling complex data with disturbed distributions.

## 4. CONCLUSION

The search for optimal smoothing parameters using Cross-Validation shows that the $\alpha_{optimal}$ value varies depending on the level of outliers in the data. In a dataset without outliers, $\alpha_{optimal}$ approaches 0, while in a dataset with a high proportion of outliers, $\alpha_{optimal}$ increases to balance bias and variance. The combination of applying $\alpha_{optimal}$ with an adaptive robust approach to construct bootstrap residual confidence bands, using domestic passenger data from Kualanamu International Airport, shows that Huber weights provide narrower confidence bands with probability coverage close to the nominal target (95%), even under conditions with fluctuations caused by the COVID-19 pandemic. In contrast, Tukey weights produce wider bands due to their more conservative approach in handling outliers.

Simulations of artificial data demonstrate the adaptability of the method to outlier proportions up to 15%. Huber weights show better sensitivity to moderate outliers, while Tukey weights are more effective in handling extreme outliers. The difference in optimal threshold values between Huber and Tukey weights is evident in achieving the nominal coverage probability. With a lower threshold, Huber weights achieve nominal coverage faster than Tukey weights. However, Tukey weights offer higher tolerance to extreme outliers, although they require a higher threshold to achieve target coverage.

The construction of residual bootstrap confidence bands in local polynomial regression, through a robust adaptive approach integrating Tukey and Huber weights, demonstrates its effectiveness in handling data with varying degrees of outliers. This study shows that the robust residual bootstrap method, which separately examines the roles of Tukey and Huber weights, can handle complex data with noise and outliers. Applying this robust adaptive approach significantly contributes to statistics, especially in nonparametric statistics, for robust prediction confidence bands under various data conditions. Furthermore, the consideration of computational time efficiency is crucial, especially when dealing with large data sets or complex data structures. This approach enables accurate and reliable analysis of real-world data developments in various fields such as transportation, finance, healthcare, and others.

## 5. ACKNOWLEDGMENT

This research was funded by the PNBP Fund of Universitas Negeri Medan through the Applied Product Research Scheme, Contract No. 0032/UN33.8/PPKM/2023.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045

## 6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

## 7. REFERENCES

[1] Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, *83*(403), 596–610.

[2] Gajewicz-Skretna, A., Furuhama A., Yamamoto H., and Suzuki N. (2021). Generating accurate in silico predictions of acute aquatic toxicity for a range of organic chemicals: Towards similarity-based machine learning methods. *Chemosphere*, *280*, 130681.

[3] Alqasrawi, Y., Azzeh, M., and Elsheikh, Y. (2022). Locally weighted regression with different kernel smoothers for software effort estimation. *Science of Computer Programming*, *214*, 102744.

[4] Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–36.

[5] Efron, B. and Tibshirani, R. J. (1993). An introduction to the bootstrap. *Monographs on Statistics and Applied Probability*, *57*(1), 1-436.

[6] Politis, D. N. (1994). Bootstrap confidence intervals in nonparametric regression without an additive model. *Journal of Econometrics*, *63*(3), 125-145.

[7] Mansyur, A. and Simamora, E. (2022). Bootstrap-t confidence interval on local polynomial regression prediction. *Mathematics and Statistics*, *10*(6), 1178–1193.

[8] Mansyur, A., Simamora, E. and Ahmad, A. (2023). Percentile bootstrap interval on univariate local polynomial regression prediction. *Jurnal Teori dan Aplikasi Matematika*, *7*(1), 160–173.

[9] Chiang, H. D., Kato, K., Sasaki, Y., and Ura, T. (2021). Linear programming approach to nonparametric inference under shape restrictions: With an application to regression kink designs. *arXiv Preprint arXiv*, *2102*, 06586.

[10] Duembgen, L., and Luethi, L. (2022). Honest confidence bands for isotonic quantile curves. *arXiv Preprint arXiv*, *2206*, 13069.

[11] Cleveland, W. S. and Grosse, E. (1988). Regression by local fitting: methods, properties, and computational algorithms. *Journal of Econometrics*, *37*(1), 87–114.

[12] Cleveland, W. S. and McGill, R. (1984). The many faces of a scatterplot. *Journal of the American Statistical Association*, *79*(388), 807–822.

[13] de Andrade, L. R., Cirillo, M. A., and Beijo, L. A. (2014). Proposal of a bootstrap procedure using measures of influence in non-linear regression models with outliers. *Acta Scientiarum. Technology*, *36*(1), 93-99.

DOI: https://doi.org/10.17509/ijost.v10i2.84323
p- ISSN 2528-1410 e- ISSN 2527-8045