



Wild Bootstrap Uncertainty for Outlier Detection Using MM-Estimation in Concrete Compressive Strength Data

Abil Mansyur*, Elmanani Simamora, Muliawan Firdaus, Rizki Habibi, Suvriadi Panggabean

Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Medan, Medan, Indonesia

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

ABSTRACT

The concrete compressive strength data may be heteroscedastic, contain influential observations, and exhibit atypical test results, such that a single robust fit is not adequate for the stability assessment of outlier decisions. In this paper, we develop a wild bootstrap-based uncertainty framework for observation-level outlier diagnostics under MM-estimation. First, a baseline MM-regression is fitted to the concrete compressive strength dataset. We apply the HC2 leverage correction to the centred residuals and perturb these residuals with three wild bootstrap multipliers: Mammen, Scaled Uniform Wild Bootstrap, and Symmetric Three-Point Bootstrap. Each pseudo-response is refitted using the same MM-estimation technique, and the outlier labels and robust weights are summarised by outlier probability, low-weight probability, mean bootstrap weight, and standard deviation of bootstrap weight. The results reveal that the three approaches detect partially overlapping outlier candidates but differ in the intensity of separation. Mammen has the sharpest separation, while SUWB has a compact core-candidate group, and STPB has a more conservative structure. The proposed methodology translates binary outlier labels to traceable diagnostic uncertainty measurements.

ARTICLE INFO

Article History:

Submitted/Received 29 Mar 2026

First Revised 25 Jun 2026

Accepted 11 Aug 2026

First Available Online 20 Oct 2026

Publication Date 01 Dec 2026

Keyword:

Concrete compressive strength,
MM-estimator,
Observation-level diagnostics,
Outlier uncertainty,
Robust regression,
Wild bootstrap.

1. INTRODUCTION

Concrete compressive strength data are not always well-behaved for regression analysis. The dataset analyzed in this study records concrete compressive strength as the response variable, with eight quantitative predictors describing cement, supplementary materials, water, aggregate composition, superplasticizer, and curing age. This type of data does not always conform to a regular regression setting. Error variance may differ across the covariate space, and a few observations can have enough leverage to alter the fitted model. Recent studies have mostly treated the dataset as a prediction problem, using interpretable models or tuned machine-learning algorithms. That direction is reasonable because the strength–mix relationship is nonlinear and depends on material interactions and curing conditions [1,2]. Predictive accuracy, however, does not address the diagnostic issue by itself. If several observations shift the fitted relationship, the engineering interpretation drawn from the model may also change.

Robust regression and outlier diagnostics provide a principled way to deal with contaminated observations without treating deletion as the first option [3–5]. In the framework, MM-estimators are attractive because they start with a high-breakdown estimate and regain efficiency via an M-estimation step under the ideal model. The computational side is no longer a major obstacle either. The Forward Search Data Analysis (FSDA) toolbox, for instance, includes MM-regression routines and diagnostic tools that can be used to follow changes in data structure [6–8]. Still, the usual output from a robust-regression analysis remains tied to a single fit: flagged observations, robust residuals, and final weights. These quantities are useful, but they do not indicate whether the same observations would remain flagged after controlled perturbations to the data. Recent studies have broadened robust inference beyond standard linear regression, including nonlinear regression models [9]. The focus of those developments, however, is mainly on model or parameter inference. The stability of outlier decisions at the observation level has received less direct treatment.

Resampling provides a natural mechanism for studying diagnostic stability. Bootstrap methods approximate the sampling behavior of complex estimators without relying entirely on a fixed parametric error distribution [10–13]. In a regression model, resampling is commonly used to quantify estimation uncertainty, including confidence intervals and confidence bands [14]. Robust bootstrap procedures have also been developed to reduce the effect of resampling atypical observations too frequently [15,16]. More recent work on stability-based outlier analysis has emphasized that outlier status should be assessed through repeated perturbations rather than only through a single fitted model [17,18]. However, many bootstrap applications in regression still focus on coefficient uncertainty or interval estimation, whereas the uncertainty of an observation being selected as an outlier is rarely treated as a primary object of analysis.

In heteroscedastic regression, the pairs bootstrap can be problematic because it resamples entire observations and, in doing so, may alter the design distribution. The wild bootstrap avoids this issue more directly: the covariates are kept fixed, while the residual part is perturbed by multipliers with zero mean and unit variance [19–22]. The idea has since been used in settings where the regression model may be mis specified and in linear regressions with many covariates [23,24]. Another related line of work uses skewness-calibrated two-point asymmetric Bernoulli weights to build adaptive wild bootstrap confidence intervals in linear regression [25]. In robust regression, the wild bootstrap has been combined with robust estimators to stabilize inference under heteroscedasticity and outliers [26]. Within this literature, the remaining gap is not the absence of robust estimators or bootstrap theory. The

gap lies in how these tools are combined to quantify the stability of observation-level outlier decisions in concrete compressive strength data.

To close this gap, the outlier decision is treated as an output that may vary across repeated robust fits, rather than as a fixed result from a single MM-regression run. For each wild bootstrap replication, the pseudo-response is formed from the baseline MM fit and the leverage-adjusted residual component. The same MM-regression settings are then applied again to the bootstrap data. For the i -th observation, the repeated fitting produces a sequence of outlier indicators; its average is reported as $\hat{p}_{out}(i)$, the empirical frequency with which the observation is flagged under bootstrap perturbation. A second summary, $\hat{p}_{wlow}(i)$, records how often the corresponding robust weight falls in the low-weight region. The mean and standard deviation of the bootstrap weights are also retained, since an observation may be repeatedly flagged or downweighted, or be unstable between the two criteria. The resulting output is an observation-level uncertainty map that describes the stability of each diagnostic decision, rather than a single list of outlier labels.

The study is organized around three objectives. The first is to develop an observation-level uncertainty framework for outlier detection by combining MM-estimation, leverage-adjusted residuals, and wild bootstrap refitting. The second is to compare the classical Mammen multiplier with SUWB and STPB, in order to assess whether different forms of bootstrap perturbation produce different stability patterns. The third is to identify priority observations in the concrete compressive strength dataset by reading several diagnostic summaries together: procedural outlier probability, low-weight probability, mean bootstrap weight, standard deviation of bootstrap weights, baseline fitted values, and baseline residuals.

This study has two main contributions. First, from a methodological perspective, this study expands the reporting of robust regression results. The analysis output no longer stops at a list of binary outliers. Each observation is summarized using an uncertainty measure. This summary is traceable, and the process is also auditable. Second, from a diagnostic contribution, outlier decisions are not read in isolation. They are read in conjunction with the behavior of the robust weights in each replication. From this reading, observations can be more clearly separated. Some observations consistently act as stable inliers. Some observations repeatedly emerge as outlier candidates. There are also borderline cases. The status of this last group changes depending on the resampling scheme used. The multiplier schemes are not used interchangeably. Mammen serves as the classical asymmetric benchmark [27], SUWB represents a bounded, continuous, symmetric multiplier, and STPB represents a bounded, discrete, symmetric multiplier. Comparing these three schemes allows the analysis to assess whether the resulting diagnostic conclusions are sensitive to the particular form of multiplier perturbation.

The remainder of this article is organized as follows. Section 2 describes the data source, regression model, MM-estimation procedure, and wild bootstrap refitting framework. Section 3 presents the outlier probability curves, robust weight stability clouds, and cross-scheme priority rankings. Section 4 concludes the article by summarizing the findings, methodological contribution, limitations, and directions for further work.

2. METHODS

Figure 1 summarizes the methodological workflow used to convert a single robust-regression outlier decision into an observation-level uncertainty map. The workflow is structured to follow the sequence outlined in the methodology section. The data source and variable structure are described in Section 2.1; the multiple linear regression model and baseline diagnostics are presented in Section 2.2; the baseline MM-estimation procedure and

robust weights are explained in Section 2.3; and the wild bootstrap refitting procedure is detailed in Section 2.4. Thus, the figure is intended not only as a graphical summary but also as a guide to the order of the computational procedure used in this study.

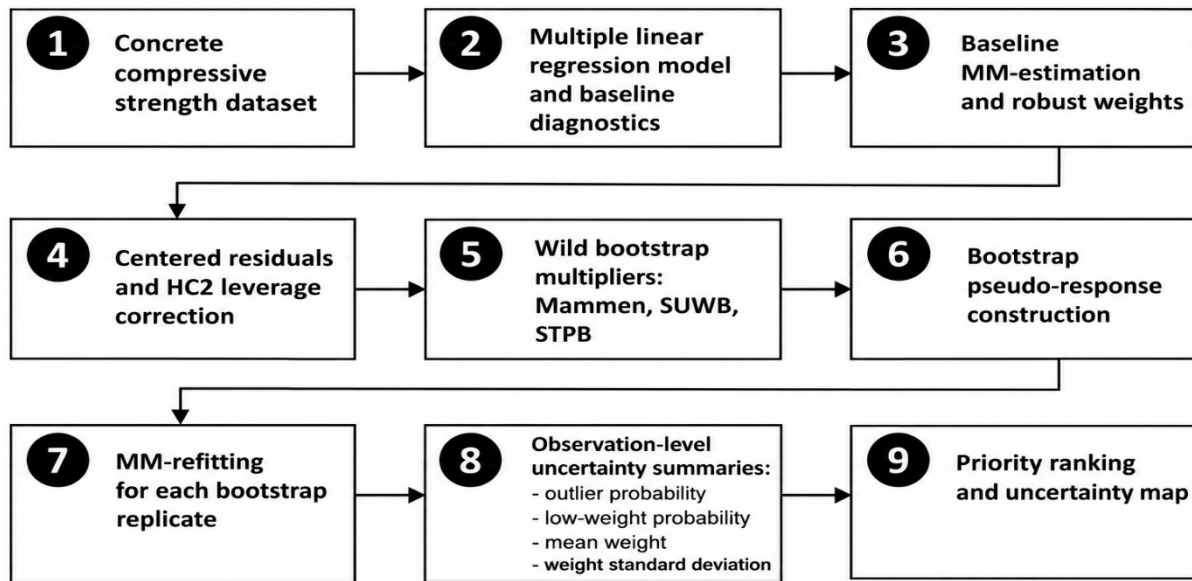


Figure 1. Workflow of the proposed wild bootstrap uncertainty framework for robust outlier diagnostics.

The procedure begins with the concrete compressive strength dataset and initial multiple linear regression diagnosis to examine the baseline residual and leverage structure. The baseline robust fit is then obtained using MM-estimation, which provides fitted values, residuals, and robust weights. The residuals from this baseline fit are centred and adjusted using the HC2 leverage correction before being perturbed with one of the selected wild-bootstrap multipliers: Mammen, SUWB, or STPB. These perturbed residual components are used to construct bootstrap pseudo-responses. The same MM-estimation technique is repeated for each bootstrap replicate, and the resulting outlier set and the robust weights are saved. The recurrent refitting process provides observation-level summaries, such as outlier probability, low-weight probability, mean bootstrap weight, and standard deviation of bootstrap weight. The summaries are then utilised to generate the priority ranking and the final uncertainty map for robust outlier diagnoses.

2.1 Data Sources

Yeh [28] introduced the concrete compressive strength dataset, which subsequently became a benchmark for evaluating predictive methods for concrete strength. The dataset has since supported a broad range of machine-learning studies. Feng et al. [29] used adaptive boosting to improve prediction accuracy, whereas Elhishi et al. [30] compared eight machine learning models for the same task. Elshaarawy et al. [31] further developed a machine-learning-based prediction tool with an interactive graphical user interface to facilitate practical use. Tak et al. [32] later employed the same dataset to develop a prediction model based on mix-design variables and curing age.

Rana et al. [26] examined the Concrete Compressive Strength Dataset, focusing on two factors: the presence of outliers and the possibility of non-constant variance. For outliers, they used residual-based Least Trimmed Squares (LTS). They found 6% outliers. They then demonstrated non-constant variance by plotting the residuals against the predicted values.

Meanwhile, Simamora *et al.* [27] used the same dataset from Kaggle and the UC Irvine Machine Learning Repository and presented Exploratory Data Analysis (EDA) to demonstrate the data characteristics. In the study, outliers were identified using the boxplot method, with 80 data points considered high-leverage. Indications of non-constant variance were further evaluated by plotting the residuals against the predicted values.

This article uses the concrete compressive strength dataset with 1,030 observations. **Table 1** describes the dataset's variable structure. All feature variables are numerical: seven mix-composition variables measured in kg/m³ and one variable representing concrete age in days. The target variable is concrete compressive strength measured in MPa. No missing values are reported so that the dataset can be analyzed without imputation. Blast Furnace Slag and Age are recorded as integer-valued variables, but they are treated as quantitative predictors in the regression model because their numerical magnitudes carry modeling information.

Table 1. Variables of the Concrete Compressive Strength Dataset

Variable Name	Role	Type	Units	Missing Values
Cement	Feature	Continuous	kg/m ³	No
Blast Furnace Slag	Feature	Integer	kg/m ³	No
Fly Ash	Feature	Continuous	kg/m ³	No
Water	Feature	Continuous	kg/m ³	No
Superplasticizer	Feature	Continuous	kg/m ³	No
Coarse Aggregate	Feature	Continuous	kg/m ³	No
Fine Aggregate	Feature	Continuous	kg/m ³	No
Age	Feature	Integer	day	No
Concrete Compressive Strength	Target	Continuous	MPa	No

2.2 Multiple Linear Regression Model

Specifically, the multiple linear regression model is expressed as shown in Equation (1).

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where y_i denotes the concrete compressive strength, and $p = 8$ predictors, namely Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, Fine Aggregate, and Age. The parameter β_0 is the intercept, $\beta = (\beta_1, \dots, \beta_p)^\top$ is the vector of regression coefficients associated with the covariates, and ε_i is the error term. Writing Equation (2) in matrix notation yields the multiple linear regression model in the following form

$$Y = X\beta + \varepsilon, \quad (2)$$

where $Y = (y_1, \dots, y_n)^\top$, X is the $n \times (p + 1)$ design matrix whose first column consists of ones, $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$, and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$. In classical linear regression, in order for the Ordinary Least Squares (OLS) method to be a Best Linear Unbiased Estimator, it must assume $\mathbb{E}(\varepsilon_i|X) = 0$ and $\text{Var}(\varepsilon_i|X) = \sigma^2$. However, concrete compressive strength data exhibit non-constant variation, and the presence of extreme or high-leverage observations is indicated as outliers. Violating these two classical assumptions can disrupt the stability of OLS predictions [26,27].

The multiple linear regression model was employed since the response variable, concrete compressive strength (MPa), is continuous and the predictors are quantitative mix-design variables. Logistic regression has not been used, as it is designed for binary or categorical responses and would require an artificial discretisation of compressive strength, which is not the intention of this work. Nonlinear or machine-learning methods may improve the accuracy

of concrete strength predictions. However, the present study focuses on observation-level diagnostic uncertainty, not the optimisation of predictions. The proposed MM-estimation and wild bootstrap refitting procedure depends on the availability of interpretable residuals, leverage values, robust weights, and HC2-adjusted residual components, which may be produced in a linear-in-parameters regression context. Thus, multiple regression provides an interpretable diagnostic baseline, while MM-estimation and repeated wild bootstrap refitting reduce the influence of unusual observations.

The OLS method determines the estimator β by minimizing the residual sum of squares, $\|Y - X\beta\|^2$. If the matrix $X^T X$ is invertible, then the OLS solution as shown in Equation (3).

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y. \quad (3)$$

Substituting $\hat{\beta}_{OLS}$ into Equation (2) yields the fitted values $\hat{Y} = X\hat{\beta}_{OLS}$ and the residuals $\hat{e} = Y - \hat{Y}$. Two diagnostic quantities that are relevant in the context of influential observations are the hat matrix as shown in Equation (4).

$$H = X(X^T X)^{-1} X^T, \quad (4)$$

and the leverage values h_{ii} , which are the diagonal elements of H and measure the extent to which the i observation lies in a relatively extreme region of the covariate space. When heteroscedasticity or contamination by outliers is present, OLS can become sensitive. Small changes in a single part of the data or in a few residual values may shift $\hat{\beta}_{OLS}$, thereby affecting the fitted values and residuals in meaningful ways. As a result, a set of outliers identified from a single model fit tends to be unstable. To address the issue, a linear model that remains linear in its parameters is introduced. The initial fitting step is carried out using OLS as a baseline, and outlier diagnostics are based on a robust MM-estimation procedure. The inferential uncertainty in the regression model is then assessed through wild bootstrap resampling. Accordingly, the results reported here not only provide an outlier or inlier classification but also summarize the stability of the classification decision and the robustness of the weights across repeated resampling.

2.3. Robust Estimation with MM-Estimator

Section 2.2 specifies the linear regression model in terms of its parameters. However, the specification does not eliminate difficulties often encountered with concrete compressive strength data. Extreme residuals, leverage points, and non-constant error variances can co-occur. When these features are present, OLS can be highly sensitive to changes in the scale of residuals or predicted values, and the inlier (or outlier) classification obtained from a single trial may be unstable. For the baseline analysis, the model is then estimated with the MM estimator, following the robust regression framework in [3-5] and the implementation principles used in robust diagnostic analysis [6-8]. This procedure combines high-breakdown initial estimation with efficient M-steps under the ideal model (See Equation 5).

$$e_i(\beta) = y_i - x_i^T \beta, \quad i = 1, \dots, n, \quad (5)$$

where x_i is a covariate vector containing the intercept. Robust estimators based on M-estimator are generally written as solutions to the system of Equation (6).

$$\sum_{i=1}^n \psi\left(\frac{e_i(\beta)}{s}\right) x_i = 0, \quad (6)$$

where $\psi(\cdot)$ is the influence function, which is usually chosen as a redescending one (e.g., Tukey bisquare weight) to prevent the dominance of extreme observations, and s is the robust scale for standardizing the residuals. In the Iteratively Reweighted Least Squares (IRLS)

implementation, the solution to (2.6) is obtained through iteratively running weighted least squares regression. For each iteration, the weights $w_i = \frac{\psi(u_i)}{u_i}$ are updated based on the standardized residuals $u_i = \frac{e_i(\beta)}{s}$, resulting as shown in Equation (7).

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n w_i [e_i(\beta)]^2, \quad (7)$$

with the conversion $w_i = \lim_{u \rightarrow 0} \psi(u)/u$.

The weights $w_i \in [0,1]$ capture the robustness mechanism of the estimator. Values of w_i close to 1 indicate that the corresponding observations are treated as in OLS. A small value of w_i (close to 0) means its contribution is significantly trimmed. The MM-estimator is formulated in two stages to combine robustness with efficiency. The first stage yields an initial estimator $\hat{\beta}_S$ together with a robust scale estimator $\hat{s}_S$ having a high breakdown point, such as that obtained from S-estimation. The second stage refines the M-estimation with a tuned ψ function to achieve a specific asymptotic efficiency (eff), for example, $\text{eff} = 0.95$ under normality (see Equation 8).

$$(\hat{\beta}_{MM}, \hat{s}, \hat{w}) \leftarrow \text{M-step}(\hat{\beta}_S, \hat{s}_S, \psi, \text{eff}). \quad (8)$$

Equation (8) is the output of the MM-estimator method for linear regression, which uses a two-stage strategy. The first stage produces a robust initial estimate that limits the influence of outlying observations. Specifically, estimators $\hat{\beta}_S$ and $\hat{s}_S$ are constructed to prevent the fitted model from being determined by a small set of extreme data points. The next stage applies the M-step to increase efficiency while preserving the robustness obtained at the outset. Computationally, the method is carried out using IRLS, where each iteration updates the weights w_i based on the standardized residuals and then updates $\hat{\beta}$ through weighted regression. The Algorithm 1 translates the weighting scheme and efficiency criterion into a clear and reproducible computational procedure.

Algorithm 1. Robust MM-estimator for linear regression through the S-step and M-step based on IRLS

Input:

- Response vector $Y \in \mathbb{R}^n$, design matrix $X \in \mathbb{R}^{n \times (p+1)}$.
- Loss function $\rho_0(\cdot)$ for the S-step and $\rho_1(\cdot)$ for the M-step. Their influence functions are defined as $\psi_k(u) = \rho'_k(u)$, for $k = 0, 1$.
- The tuning constant c_0 is selected to obtain a high breakdown point in the S-step, whereas c_1 is chosen to achieve the desired asymptotic efficiency in the M-step, for example around 95% under normality.
- The scale calibration constant $b \in (0,1)$ is chosen so that the scale estimator is consistent under the normality assumption.
- The convergence criterion is specified through a tolerance $\delta > 0$ and a maximum number of iterations $T_{\max}$.

Output:

- $\hat{\beta}_{MM}$, the robust estimator of the regression coefficients.
 - $\hat{s}$, the robust scale obtained from the S-step.
 - $\hat{w} = (w_1^{(t)}, \dots, w_n^{(t)})$, the final weight vector.
-

Stage 0 — Robust initialization

- (i) Ensure that X has full column rank. If the problem is ill-conditioned, use a numerically stable least-squares solver, such as QR factorization, or Cholesky factorization when the weighted matrix is positive definite.
- (ii) Construct an initial value $\beta^{(0)}$ that is resistant to outliers through an S-initial procedure. Other methods may also be used to obtain the initial value $\beta^{(0)}$, such as Least Trimmed Squares (LTS) or Least Median of Squares (LMS).
- (iii) Set $t \leftarrow 0$.

Stage I — S-estimation to obtain a stable and outlier-resistant robust scale.

- (i) Define the residuals for a candidate β according to equation (2.5).
- (ii) For each candidate β , define the robust scale $s(\beta)$ as the solution $s > 0$ of the M-scale Equation (9).

$$\frac{1}{n} \sum_{i=1}^n \rho_0 \left(\frac{e_i(\beta)}{s} \right) = b. \quad (9)$$

- (iii) In implementation, $s(\beta)$ is usually obtained by a scale-iteration procedure until convergence.
- (iv) Choose the S-estimator as the scale minimizer (see Equation 10)

$$\hat{\beta}_S = \arg \min_{\beta} s(\beta), \quad \hat{s}_S = s(\hat{\beta}_S). \quad (10)$$

Note: Stage I provides high breakdown protection against high-leverage points or outliers, and the scale $\hat{s}_S$ is then used as a stable reference in Stage II.

Stage II — IRLS-based M-estimation with fixed scale to improve efficiency

- (i) Set $\beta^{(0)} \leftarrow \hat{\beta}_S$ and use $\hat{s}_S$ as a fixed standardization scale, although limited updating may also be allowed.
- (ii) Repeat until convergence or until $t = T_{\max}$:
 - 2.1. Compute the residuals $e_i^{(t)} = y_i - x_i^T \beta^{(t)}$, $i = 1, \dots, n$.
 - 2.2. Standardize the residuals as $u_i^{(t)} = \frac{e_i^{(t)}}{\hat{s}_S}$.
 - 2.3. Compute the IRLS weights

$$w_i^{(t)} = \begin{cases} \frac{\psi_1(u_i^{(t)})}{u_i^{(t)}}, & u_i^{(t)} \neq 0, \\ \lim_{u \rightarrow 0} \frac{\psi_1(u)}{u}, & u_i^{(t)} = 0. \end{cases}$$

- 2.4. Form the weight matrix $W^{(t)} = \text{diag}(w_1^{(t)}, \dots, w_n^{(t)})$.
- 2.5. Update the parameter by Weighted Least Squares (WLS)

$$\beta^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} Y.$$

- 2.6. Stop if

$$\frac{\|\beta^{(t+1)} - \beta^{(t)}\|}{(1 + \|\beta^{(t)}\|)} < \delta.$$

- 2.7. Set $t \leftarrow t + 1$ and repeat.

- (iii) Final output of the procedure:

$$\hat{\beta}_{MM} \leftarrow \beta^{(t)}, \quad \hat{w} \leftarrow (w_1^{(t)}, \dots, w_n^{(t)})^T, \quad \hat{s} \leftarrow \hat{s}_S.$$

The main results of robust modelling include the estimated coefficients $\hat{\beta}_{MM}$, the estimated values $\hat{y}_i = x_i^T \hat{\beta}_{MM}$, the residuals $\hat{e}_i = y_i - \hat{y}_i$, and the robust weights $\hat{w}_i$. To provide an operational and auditable definition of outliers, two criteria are explicitly distinguished:

standardized residual-based flagging and weight-based downweighting. In the first criterion, the residuals are standardized with a robust scale and weighted leverage correction (see Equation 11).

$$r_i = \frac{\hat{e}_i}{\hat{s}\sqrt{1-h_i}}, \quad (11)$$

with $h_i = (H_Z)_{ii}$, $H_Z = Z(Z^\top Z)^{-1}Z^\top$ and $Z = W_0^{1/2}X_1$, where $w_0 = (w_{01}, \dots, w_{0n})^\top$ and $W_0 = \text{diag}(w_0)$. Unlike classical leverage, which is based on equation (2.4), the standardized residuals in the study use the weighted leverage $h_i = (H_Z)_{ii}$ to align with the baseline MM-regression.

An observation is declared flagged as an outlier if $|r_i| > c_r$, where c_r is a threshold calibrated at a certain level of confidence. The set of outliers from the fit baseline is denoted by $\mathcal{O}$. In the second criterion, observations are declared downweighted if $\hat{w}_i \leq \tau_w$. Unlike flagging, the criterion is not intended as a detection label but rather as a measure of how much the influence of an observation is reduced during robust estimation.

Separating these two criteria is important because the main aim of the study is to measure the uncertainty in outlier detection. The baseline decision $1\{i \in \hat{\mathcal{O}}\}$ is a discontinuous function of the data, so small changes in the fit results or in the scale can change the status of an observation. Therefore, Section 2.4 concerns building a wild bootstrap resampling and refitting the MM-regression iteratively so that, for each observation, we obtain the sequence $\{w_i^{(b)}, 1\{i \in \hat{\mathcal{O}}\}\}_{b=1}^B$. From here, we define stability measures such as: $\hat{p}_{out}(i)$ is the selection frequency, $\hat{\mu}(w_i)$ is the weighted average, $\hat{\sigma}(w_i)$ is the standard deviation of the weights, $\hat{p}_{low}(i)$ is the low weight frequency for the i -th observation across bootstrap replications, and the normalized switching index. The wild bootstrap refitting procedure to construct the uncertainty map is summarized in Algorithm 2 in Section 2.4.

2.4. Wild Bootstrap Refitting

Section 2.4 presents the wild bootstrap refitting procedure used to construct the outlier uncertainty map. The procedure follows the fixed-design logic of the wild bootstrap in heteroscedastic regression [19-22]. It adapts it to robust diagnostic refitting as motivated by robust bootstrap procedures for contaminated regression data [15,16,26]. In each scheme, the covariate design is treated as fixed, whereas the residual component is perturbed by a multiplier with zero mean and unit variance.

The initial step is to fit the baseline MM-regression as in Section 2.3 to obtain the residual vector. The residual vector is centered as $\tilde{e} = \hat{e} - \bar{\hat{e}}1_n$ with $1_n = (1, 1, \dots, 1)^\top \in \mathbb{R}^n$ and $\bar{\hat{e}} = \frac{1}{n} \sum_{i=1}^n \hat{e}_i$. To generate a pseudo-response, the residuals can be adjusted using a Heteroskedasticity-Consistent (HC) leverage correction. The adjustment is optional and is used when the inference needs to be stabilized on high-leverage observations. In this article, the HC2 correction is used (see Equation 12).

$$e_i^{adj} = \frac{\tilde{e}}{\sqrt{\max(1-h_i, \partial)}}, \quad (12)$$

where h_i represents the weighted leverage of the baseline MM-regression as defined in (11). Meanwhile, $\partial > 0$ is a small constant to prevent division by zero when $1 - h_i$ is very small. With the construction, highly leveraged observations do not automatically dominate the bootstrap variation.

In applying the wild bootstrap, the classical Mammen multiplier is used as the benchmark scheme because it is a standard two-point distribution for heteroscedastic regression resampling [21]. The Mammen multiplier follows a two-point distribution

$$v_i = \begin{cases} (1 - \sqrt{5})/2, & \text{with probability } (\sqrt{5} + 1)/(2\sqrt{5}) \\ (1 + \sqrt{5})/2, & \text{with probability } (\sqrt{5} - 1)/(2\sqrt{5}) \end{cases}$$

In addition, two alternative multiplier specifications are introduced in this article for diagnostic comparison: the Scaled Uniform Wild Bootstrap (SUWB) and the Symmetric Three-Point Bootstrap (STPB). SUWB is used to represent a bounded continuous symmetric perturbation, whereas STPB is used to represent a bounded discrete symmetric perturbation. The comparison of Mammen, SUWB, and STPB is therefore not only computational. It is designed to examine whether the stability of outlier decisions is affected by multiplier asymmetry, discreteness, and the concentration of bootstrap perturbations. Observations that remain prominent under all three schemes are interpreted as more stable diagnostic priorities than observations that become prominent only under one scheme.

$$v_i \sim \begin{cases} -\sqrt{3}, & \text{with probability } 1/6 \\ 0, & \text{with probability } 2/3 \\ \sqrt{3}, & \text{with probability } 1/6 \end{cases}$$

Suppose the simulation performs bootstrap sampling B times, generating each bootstrap sample vector $\mathbf{v}^{(b)} = (v_1^{(b)}, \dots, v_n^{(b)})$ independently and identically distributed. Each component of $v_i^{(b)}$ satisfies $\mathbb{E}(v_i^{(b)}) = 0$ and $\text{Var}(v_i^{(b)}) = 1$. A pseudo-response is generated for each $i = 1, \dots, n$ (see Equation 13).

$$y_i^{*(b)} = \hat{y}_i + e_i^{adj} v_i^{(b)}, i = 1, \dots, n. \quad (13)$$

for the b -th replicate, the pseudo-response vector is defined by $\mathbf{Y}^{*(b)} = (y_1^{*(b)}, y_2^{*(b)}, \dots, y_n^{*(b)})^\top$, so that ordered data pairs are obtained $(\mathbf{X}, \mathbf{Y}^{*(b)})$.

Next, we refit the MM-regression on $(\mathbf{X}, \mathbf{Y}^{*(b)})$, utilizing the initial estimates from the baseline to speed up the computation (thus maintaining the same robust procedure, but reducing the iteration overhead). Each refit produces two main outputs: (i) the set of outliers $\hat{\mathcal{O}}^{(b)}$ returned by the MM-regression at the same confidence level, and (ii) the robust weights $w_i^{(b)}$ for all observations. The uncertainty measure for outlier detection is operationally defined in terms of the resampling frequency. First, the outlier selection frequency for the i -th observation is defined as shown in Equation (14).

$$\hat{p}_{\text{out}}(i) = \frac{1}{B} \sum_{b=1}^B 1\{i \in \hat{\mathcal{O}}^{(b)}\}, \quad (14)$$

which interprets the procedural probability of observation being flagged as an outlier under wild bootstrap noise. Second, the procedure separates the stability of the downweighting mechanism through low-weight events with a threshold w_{thr} (see Equation 15).

$$\hat{p}_{\text{low}}(i) = \frac{1}{B} \sum_{b=1}^B 1\{w_i^{(b)} < w_{\text{thr}}\}. \quad (15)$$

These two measures are not necessarily identical: an observation can frequently be flagged as an outlier without always having a very low weight, or vice versa, and the separation is necessary to interpret the borderline versus core outlier structure in the output.

In addition to probabilistic measures, weight stability is also summarized at the observation level. Two summaries are used: the mean weight across replicates and the standard deviation of the weights across replicates. For each observation, these two summaries are defined in Equation (16). These summaries are used to map the pairs $(\hat{\mu}(w_i), \hat{\sigma}(w_i))$ and to assess the degree of weight switching between replicates. The value of $\hat{\mu}(w_i)$ represents the long-term trend of the weights, while $\hat{\sigma}(w_i)$ measures the volatility of the weights and reflects the sensitivity of the downweighting mechanism to resampling variations

$$\hat{\mu}(w_i) = \frac{1}{B} \sum_{b=1}^B w_i^{(b)}, \quad \hat{\sigma}(w_i) = \left\{ \frac{1}{B-1} \sum_{b=1}^B \left(w_i^{(b)} - \hat{\mu}_w(i) \right)^2 \right\}^{1/2}. \quad (16)$$

Next, a summary of the resampling results is compiled at the observation level using $\hat{p}_{out}^{(b)}$, $\hat{p}_{wlow,i}^{(b)}$, $\hat{\mu}_{w,i}^{(b)}$, and $\hat{\sigma}_{w,i}^{(b)}$. These summaries are then presented in a stability table per method, along with baseline diagnostics. Furthermore, supporting visualizations are used to evaluate the consistency of outlier classification and the stability of robust weighting. The computational steps for obtaining these measures are presented in Algorithm 2.

Algorithm 2. Wild bootstrap uncertainty map for outliers in MM-regression

- Step 1. Fit MM-regression to the data pair (X, Y) to obtain $\hat{\beta}$, fitted values $\hat{y}$, centered residuals $\tilde{e}_i$, and robust weights $\hat{w}_i$.
 - Step 2. Compute the leverage-adjusted residuals using the HC-2 correction in (12).
 - Step 3. For $b = 1, \dots, B$, generate wild bootstrap multipliers using one of the selected distributions (Mammen, SUWB, or STPB).
 - Step 4. Construct the pseudo-response according to (13).
 - Step 5. Fit MM-regression to $(X, Y^{*(b)})$, then store the outlier set $\mathcal{O}^{(b)}$ and the robust weights $w_i^{(b)}$.
 - Step 6. Estimate the probability that the i -th observation is classified as an outlier across the B resampling results using (14).
 - Step 7. Estimate the probability that the robust weight associated with the i -th observation falls in the low-weight region across the bootstrap replicates using (15).
 - Step 8. Form the overall summary using (14), (15), and (16), then construct the ranking plot and the weight-stability cloud.
-

3. RESULTS AND DISCUSSION

3.1. Simulation Results

Section 3 presents simulation results to assess the uncertainty of outlier detection in three wild bootstrap schemes, namely Mammen, SUWB, and STPB, based on the concrete compressive strength data in Section 2.1. The presentation of the results is divided into two subsections. Subsection 3.1 discusses the general patterns of outlier probability, weight stability, and the positions of problematic observations, using visualization of the simulation results. Subsection 3.2 presents the ranking of priority observations. The ranking allows the consistency of outlier candidates across multiplier schemes to be evaluated. It also identifies observations whose classification is more sensitive to the resampling specification.

3.1.1 Outlier Uncertainty Pattern and Weight Stability

Based on the concrete compressive strength dataset in Section 2.1, **Figure 2** presents three rank–probability curves for p_{out} in three wild bootstrap schemes: Mammen, SUWB, and STPB.

For each the i -th observation, $p_{\text{out}}(i)$ is defined as the proportion of $B = 10,000$ bootstrap replications when the observation is classified as an outlier by the MM-regression procedure. Operationally, $p_{\text{out}}(i)$ is calculated by summing the outlier indicators for the observation across all replications, then dividing by B . Thus, $p_{\text{out}}(i)$ represents the relative frequency of the observation being classified as an outlier throughout the bootstrap process. Next, all $p_{\text{out}}(i)$ values are sorted in descending order and plotted against their ranks. Therefore, the x-axis represents the rank (Rank, descending order) and the y-axis represents $p_{\text{out}}(i)$. By construction, the point at the first rank (leftmost) always indicates the observation with the largest $p_{\text{out}}(i)$ value, that is, the observation most frequently identified as an outlier throughout the bootstrap replicates.

In **Figure 2(a)**, under the Mammen scheme, the curve starts very high and then drops sharply at the initial ranks. The pattern indicates a small number of observations with very large p_{out} , so their outlier status is relatively consistent across replicates. After the initial drop, the curve tends to flatten into a moderate-probability range for a considerable number of ranks. The segment indicates a set of observations with unstable outlier status, as their probability is at an intermediate level and easily changes when the data is perturbed. At further ranks, the curve approaches zero and persists for a long time. The segment indicates that most observations are rarely classified as outliers, so their inlier status is relatively stable.

In **Figure 2(b)**, under the SUWB scheme, the curve does not reach a value of 1, and the decline is more gradual. It means that no observations are always flagged as outliers in almost all replications. The p_{out} values are more evenly distributed in the small to medium range, then gradually decrease towards zero. Interpretatively, SUWB produces a more continuous risk sequence. Consequently, the number of outlier candidates that need to be reviewed is more dependent on the chosen p_{out} threshold, because the separation between high and low probabilities is not as strong as in Mammen.

In **Figure 2(c)**, under the STPB scheme, the maximum value of p_{out} is relatively low (around 0.34–0.35) and the initial part of the curve tends to be flat before then dropping rapidly towards zero. The pattern indicates that no observation dominates as an outlier in most of the replications. Instead, membership in the outlier set tends to vary across several candidate observations. It leads to smaller and more homogeneous outlier probabilities at the observation level. Accordingly, STPB produces more cautious labeling decisions because the highest probability remains limited even though some candidates continue to reappear.

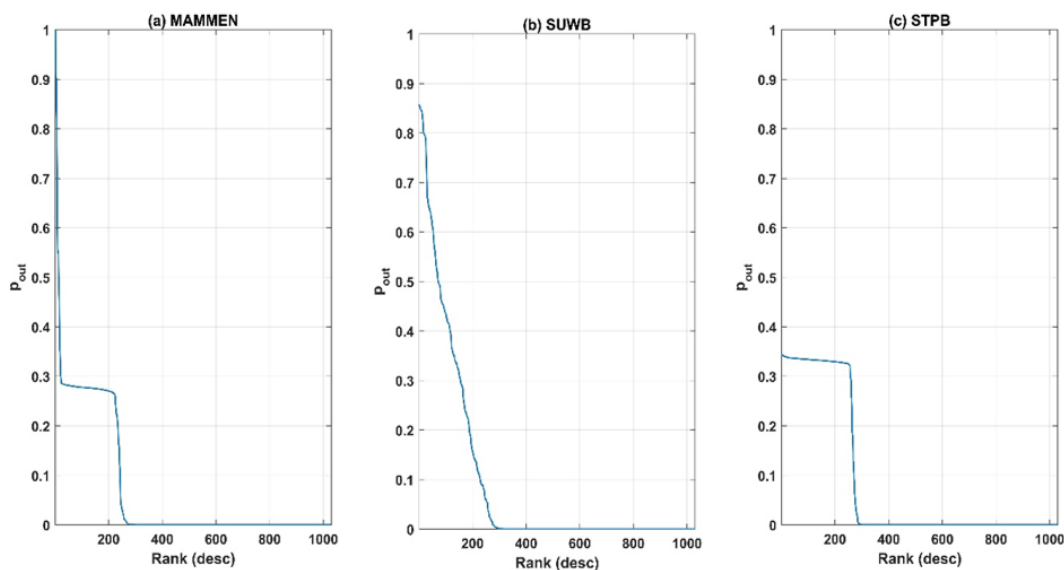


Figure 2. Outlier Probability Rank Curves in Wild Bootstrap: Mammen, SUWB, and STPB.

The three rank-probability curves show different patterns across the multiplier schemes. Mammen gives the clearest separation between observations with high outlier probability and those that remain in the inlier group. For most inliers, the probabilities stay close to zero. The Mammen scheme is appropriate when the main goal is to identify the most consistent outlier candidates for further analysis. SUWB provides a more detailed ranking based on probability. For that reason, it is more useful when the emphasis is on ordering observations by risk rather than on obtaining a sharp separation. STPB produces lower maximum probabilities and a more even distribution among the leading candidates. As a result, it is more suitable when a more cautious decision is preferred at the observation level.

For the purpose of presenting auditable results, the interpretation of the curves can be formulated as a concise procedure. First, define priority candidates as the observations with the largest p_{out} and check their consistency among schemas. Second, identify instability ranges as relatively flat curve segments at intermediate probabilities, as these segments indicate candidates with states that are prone to change between replicates. Third, treat long segments close to zero as stable inlier clusters. With these steps, the analysis output no longer stops at binary labels but becomes a summary of decision uncertainty that can be traced back to the curve shapes at each multiplier.

As a complement to the outlier probabilities, **Figure 3** presents a weight stability cloud for each observation. The summary is used to assess the stability of the downweighting mechanism in MM-regression. Observations with high $\mu(w_i)$ and small $\sigma(w_i)$ tend to maintain their influence consistently throughout the estimation process, while observations with lower average weights and larger standard deviations indicate that their influence changes between replications.

Figure 3(a) shows the widest distribution of visible points. Most observations cluster in the region of high $\mu(w_i)$ and low $\sigma(w_i)$, indicating stability. However, some observations fall outside the cluster, displaying medium-weight averages and large standard deviations. This suggests that, in the Mammen scheme, weight changes across replications are not uniform. Even adjacent weight averages can differ significantly in standard deviation, indicating Mammen is more sensitive to variation in the stability of influences across observations.

Figure 3(b) displays a more regular distribution pattern. The relationship between $\mu(w_i)$ and $\sigma(w_i)$ is systematic: as the weighted average changes, the standard deviation follows. The highest $\sigma(w_i)$ values occur at medium $\mu(w_i)$, while at higher $\mu(w_i)$, $\sigma(w_i)$ decreases gradually. This suggests that SUWB creates a continuous stability structure, structuring the influence of observations as a spectrum from most stable to relatively unstable, rather than sharply dividing them.

Figure 3(c) shows a more concentrated pattern than the other two methods. Most of the points fall along a narrow downward path between the weighted mean and the standard deviation. It indicates that the weight variation in STPB is more structured, so that the standard deviation of the weights is closely related to their mean value. Under these conditions, the differences between observations are not as big as those seen in Mammen. Notably, among observations near the dividing line between inliers and downweighted candidates, the pattern of weight changes is consistently regular, suggesting a predictable shift as observations transition from inliers to downweighted candidates.

Comparing the three schemes, each produces a different stability character. For instance, Mammen exhibits the highest weighting heterogeneity, making it more informative for identifying observations that are truly sensitive to bootstrap perturbations. In contrast, SUWB provides a more gradual pattern and is easier to interpret as a hierarchical stability level. Meanwhile, STPB produces a more uniform structure, making the weighting uncertainty

appear more consolidated within a specific group of observations. Taken together, these differences indicate that the choice of multiplier not only affects the frequency with which observations are detected as outliers but also the form of stability of the observational influence in the robust procedure.

In the context of concrete compressive strength data, simulation results provide information that is not fully captured by the outlier probability curve. Observations with a high outlier probability and a low weighted mean are considered relatively consistent outlier candidates. Observations with a medium outlier probability but a high weighted standard deviation are more appropriately treated as sensitive to changes in the bootstrap data. Meanwhile, observations with a weighted mean close to one and a small standard deviation can be considered stable in the estimation process. Therefore, a joint reading of outlier probability and weight stability provides a more complete basis for interpretation than using either indicator alone.

Overall, the simulation results confirm that the three wild bootstrap schemes do not yield identical stable patterns. Mammen shows greater variation between observations, SUWB shows a more gradual pattern, and STPB shows a more regular structure. The finding supports the use of uncertainty maps as a reporting tool, as the results obtained go beyond marking outliers and also indicate the stability of the status and influence of each observation across repeated bootstrapping.

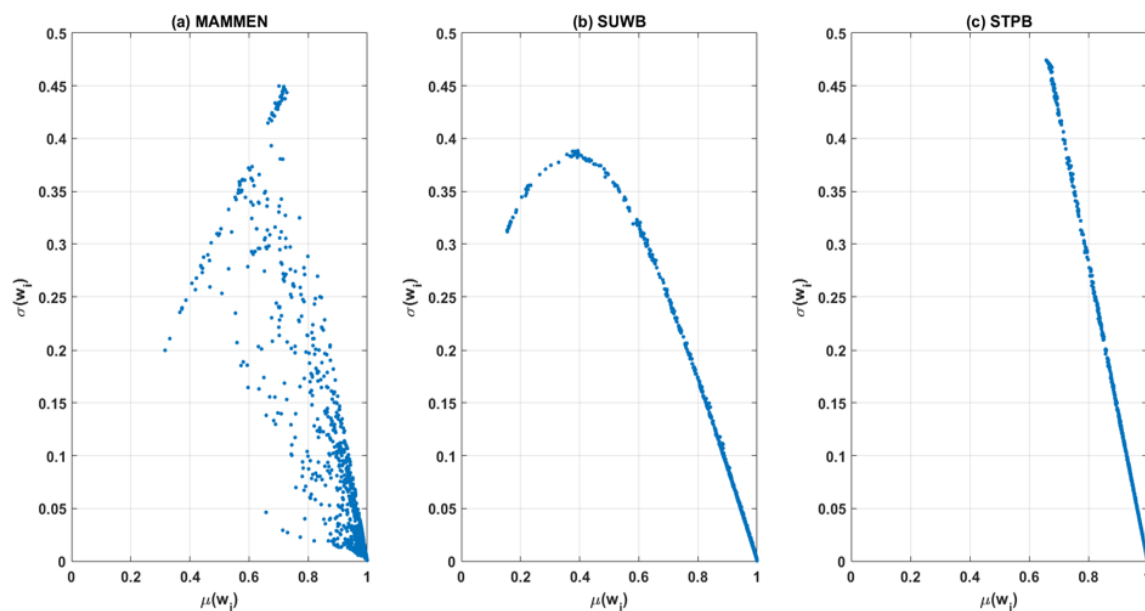


Figure 3. Robust weight stability cloud based on the mean and standard deviation of bootstrap robust weights in the Mammen, SUWB, and STPB schemes.

Figure 4 plots baseline residuals against the baseline fitted values, with points colored by the procedural outlier probability, $\hat{p}_{out}(i)$, for the Mammen, SUWB, and STPB wild bootstrap schemes. Most observations cluster in low-to-medium fitted values, with residuals near zero and low outlier probabilities. In the high-fitted range, larger negative residuals emerge, revealing scheme-specific differences. These deviations indicate that instability sources are concentrated in a few observations rather than being uniformly distributed across the data.

Figure 4(a) shows the clearest separation between stable observations and those frequently detected as outliers. Points with high fit and large negative residuals are much more likely to be outliers than the main group. In **Figure 4(b)**, problematic observation locations stay the same, but changes in outlier probability are more gradual, making the

distinction between stable, borderline, and frequently flagged points less sharp. **Figure 4(c)** displays more widely distributed and lower outlier probabilities, even for the most geometrically extreme observations. The simulation results indicate all three wild bootstrap schemes identify similar problem areas but differ in how strictly they assign outlier status. For concrete compressive strength data, it is important because it shows that outlier interpretation depends not only on residual size but also on the stability of decisions across repeated bootstrapping.

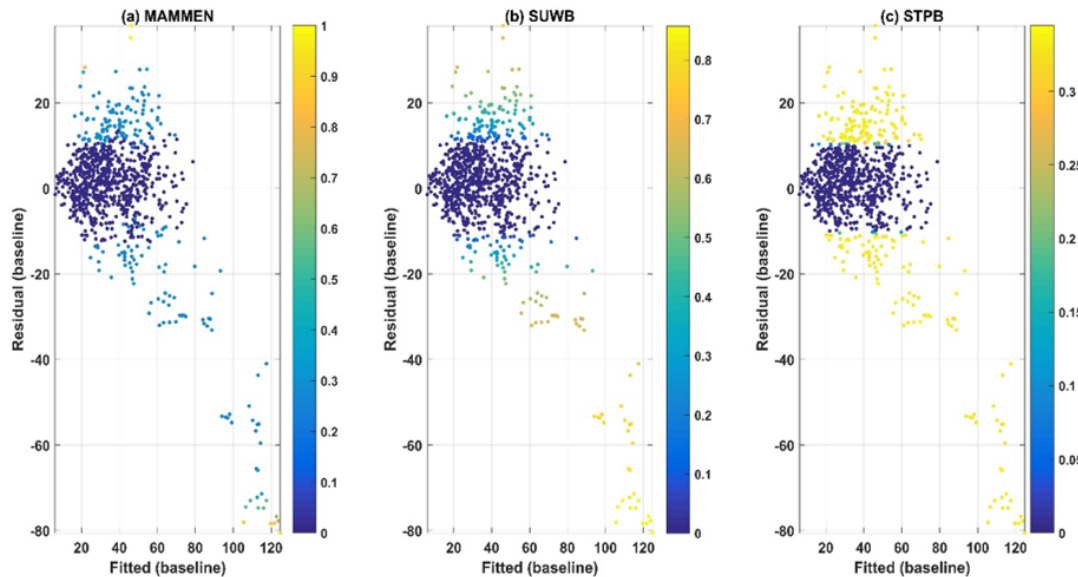


Figure 4. Baseline residuals versus baseline fitted values colored by the procedural outlier probability, $\hat{p}_{out}(i)$, under the Mammen, SUWB, and STPB schemes.

3.1.2 Cross-Scheme Priority Observation Ranking

As a complement to **Figure 4**, observation-level analysis uses a ranking table to identify points most often detected as outliers for each multiplier scheme. The approach highlights both the probability of procedural outliers and their relationship to robust weights, initial residuals, and initial fitted values for each observation.

In the Mammen scheme, the separation between stable observations and observations frequently detected as outliers is most evident. According to **Table 2**, observations 382 and 384 achieve $p_{out} = 1$, with p_{wlow} of 0.9702 and 0.9384, respectively, and very small initial weights w_0 . However, their residual profiles differ from those of the other leading candidates. Observations 382 and 384 have relatively large positive initial residuals of 38.012 and 35.131, respectively, with initial fitted values of approximately 46. In contrast, observations 57, 43, 611, and 4 also have very high p_{out} , but at larger initial fitted values of approximately 105 to 125, with very negative initial residuals of approximately -78 to -81, and $w_0 = 0$. These results indicate that in Mammen, there are at least two priority observation groups, namely the group with large positive residuals in the medium fitting and the group with very large negative residuals in the high fitting.

The Scaled Uniform Wild Bootstrap (SUWB) method shows that the dominant observations still have high fitted values with large negative residuals, but their rank order is more uniform. **Table 2** shows that observations 57, 4, 611, 42, 793, 621, 617, 43, and 7 rank highest, with probability of being an outlier values ranging from 0.846 to 0.857. The probability of being strongly downweighted values for the group are also nearly identical, ranging from 0.828 to 0.841. The average observation weights (i.e., a measure of the influence each observation has

on the analysis) are very low, ranging from 0.155 to 0.167, and all have initial weights. The initial residuals (i.e., differences between observed and fitted values) for the group are also consistent, being very negative in the high initial fitted set. The finding suggests that in SUWB, observations frequently detected as outliers also tend to experience consistent strong downweighting. Therefore, SUWB forms a more compact group of core candidate outliers.

The STPB shows a different pattern. As shown in **Table 2**, p_{out} values for top observations are narrowly clustered (0.33–0.34), and mean weights remain high (0.66–0.67) for observations 57, 611, 43, 382, 4, and 384. It shows that frequently detected outliers in the STPB do not always get low weights. The link between outliers and low weights in the STPB is weaker than in SUWB. Outlier membership in STPB varies more across replicates, making single-observation separation less clear.

Table 2. Twenty-Five Observations with the Highest Probability of Procedural Outliers in the Wild Bootstrap Scheme.

Skema	i	$\hat{p}_{out}(i)$	$\hat{p}_{wlow}(i)$	$\hat{\mu}(w_i)$	$\hat{\sigma}(w_i)$	$w_0(i)$	$\hat{e}_i$	$\hat{y}_i$
Mammen	382	1	0.9702	0.31601	0.19986	0.0088526	38.012	46.061
	384	1	0.9384	0.33206	0.21102	0.054973	35.131	45.991
	57	0.9253	0.5738	0.36609	0.23595	0	-80.59	124.85
	43	0.9054	0.5429	0.37338	0.23955	0	-78.38	119.7
	611	0.9017	0.5506	0.3727	0.23898	0	-78.149	105.55
	4	0.8564	0.4771	0.3867	0.24672	0	-78.382	121.76
	15	0.8198	0.2768	0.41878	0.25665	0.27003	28.311	21.825
	42	0.7187	0.4011	0.40672	0.2627	0	-77.797	123.82
	7	0.6472	0.3647	0.4219	0.26764	0	-76.767	122.79
	623	0.558	0.3413	0.43675	0.27919	0	-74.732	113.2
	621	0.5534	0.3362	0.44031	0.27758	0	-74.531	106.64
	605	0.5498	0.3409	0.43713	0.27957	0	-74.793	116.01
	793	0.547	0.3267	0.44296	0.27694	0	-73.037	117.49
	617	0.5007	0.3275	0.44642	0.2875	0	-73.048	109.07
	9	0.4843	0.2749	0.4417	0.27295	0.31439	27.175	21.001
	815	0.4531	0.3092	0.46064	0.28965	0	-72.295	112.73
	770	0.4022	0.3014	0.4718	0.30112	0	-71.457	115.02
	115	0.3553	0.2819	0.4678	0.25923	0.32273	-22.322	47.541
	1	0.3489	0.269	0.46621	0.28434	0.28925	27.814	54.494
	225	0.3319	0.2789	0.55926	0.20724	0.44862	-19.226	29.297
	67	0.3029	0.279	0.53099	0.33277	0	-65.934	112.95
	77	0.2995	0.2796	0.58051	0.18885	0.48005	-18.46	60.978
	814	0.2913	0.2913	0.7013	0.44969	0.23827	-24.506	64.156
	75	0.2856	0.2798	0.59656	0.16466	0.50797	-17.776	49.098
	5	0.2855	0.2761	0.55318	0.3443	0	-65.542	112.16
Scaled Uniform Wild Bootstrap	57	0.8574	0.8406	0.15543	0.31162	0	-80.59	124.85
	4	0.8545	0.8403	0.15477	0.31306	0	-78.382	121.76
	611	0.854	0.8355	0.15898	0.31467	0	-78.149	105.55
	42	0.8525	0.8348	0.15964	0.31666	0	-77.797	123.82
	793	0.8507	0.8333	0.16185	0.31652	0	-73.037	117.49
	621	0.8474	0.8302	0.16456	0.31891	0	-74.531	106.64
	617	0.8471	0.8286	0.16742	0.32082	0	-73.048	109.07
	43	0.8469	0.8285	0.16444	0.31923	0	-78.38	119.7
	7	0.8463	0.8283	0.16517	0.3206	0	-76.767	122.79
	623	0.8453	0.8295	0.16374	0.31833	0	-74.732	113.2
	605	0.8439	0.8283	0.16676	0.32142	0	-74.793	116.01
	815	0.8397	0.8216	0.17419	0.32503	0	-72.295	112.73
	770	0.8384	0.8214	0.17371	0.32521	0	-71.457	115.02

Table 2 (continue). Twenty-Five Observations with the Highest Probability of Procedural Outliers in the Wild Bootstrap Scheme.

Skema	i	$\hat{p}_{out}(i)$	$\hat{p}_{wlow}(i)$	$\hat{\mu}(w_i)$	$\hat{\sigma}(w_i)$	$w_0(i)$	$\hat{e}_i$	$\hat{y}_i$
Scaled Uniform Wild Bootstrap	5	0.8286	0.8098	0.18432	0.32998	0	-65.542	112.16
	67	0.8261	0.8071	0.18583	0.3322	0	-65.934	112.95
	25	0.8115	0.7902	0.20402	0.34447	0	-59.577	114.42
	36	0.803	0.7779	0.21721	0.34581	0	-53.33	94.06
	27	0.7998	0.7736	0.22063	0.34819	0	-53.677	97.15
	32	0.7984	0.7742	0.21807	0.35193	0	-56.713	111.94
	34	0.7978	0.7749	0.21993	0.35127	0	-54.757	99.21
	18	0.7954	0.7702	0.22315	0.35382	0	-55.128	113.59
	3	0.7948	0.7676	0.22534	0.35174	0	-53.529	96.12
	31	0.7947	0.7718	0.22262	0.35226	0	-55.185	112.77
	35	0.7912	0.7688	0.22424	0.35501	0	-54.281	110.3
	13	0.7891	0.7654	0.22959	0.35551	0	-52.845	98.18
	326	0.3448	0.3448	0.788135	0.291917	0.835413	13.31797	31.29651
	400	0.344	0.0313	0.833464	0.229923	0.883851	11.49688	30.12528
	80	0.3424	0.3106	0.811537	0.260564	0.713909	-12.4442	56.06582
Symmetric Three-Point Bootstrap	448	0.3423	0.3022	0.821899	0.246768	0.871076	12.0053	38.90409
	115	0.3417	0.3417	0.658136	0.474185	0.322726	-22.3216	47.54109
	449	0.3417	0.3417	0.774276	0.31319	0.815999	13.98099	40.06552
	670	0.3416	0.3337	0.817249	0.25359	0.722308	-12.2063	30.63796
	447	0.3413	0.3413	0.728726	0.376713	0.751734	15.99588	40.05012
	131	0.3412	0.3412	0.658515	0.473931	0.455741	23.69631	53.32354
	4	0.3408	0.3408	0.65894	0.473815	0	-78.3822	121.757
	30	0.3406	0.3376	0.80581	0.269821	0.852977	12.68886	28.23692
	877	0.3401	0.0433	0.837496	0.226105	0.886008	11.40858	23.97381
	40	0.34	0.34	0.65978	0.473575	0.040347	-31.199	69.77298
	105	0.3398	0.06	0.836594	0.227455	0.748426	-11.4497	47.16992
	622	0.3397	0.3397	0.660173	0.47354	0.168479	-26.4694	63.27902
	667	0.3394	0.3394	0.789358	0.293622	0.678814	-13.4136	28.52406
	443	0.3392	0.3392	0.778255	0.309348	0.818762	13.88849	33.07021
	854	0.3392	0.3118	0.820981	0.249678	0.868648	12.09936	30.28122
	308	0.339	0.0001	0.846933	0.213692	0.895164	11.02517	42.33186
	770	0.3387	0.3387	0.661038	0.473104	0	-71.457	115.0235
	383	0.3385	0.3385	0.696507	0.423692	0.689704	17.76078	48.57909
	403	0.3384	0.3201	0.81912	0.252726	0.865901	12.20483	43.69637
	69	0.338	0.338	0.661769	0.472887	0.450414	23.82606	19.35293
	687	0.338	0.338	0.756252	0.340889	0.623948	-14.8656	31.33156
	946	0.338	0.338	0.731687	0.375367	0.752305	15.97892	22.78674

The cross-method comparison in **Table 3** shows that several observations remain as prime candidates despite changing the multiplier scheme. Based on the average p_{out} across methods, observations 57, 611, 43, 382, 4, and 384 rank among the top candidates with $\bar{p}_{out}$ ranging from 0.677 to 0.705. Across these six observations, Mammen and SUWB both yield high outlier probabilities, whereas STPB yields a more compressed distribution around 0.33. These results suggest that these six observations are worthy of prioritization for further investigation because they remain prominent across the various multiplier schemes. In contrast, observations such as 621, 623, and 793 rise primarily due to the stronger contribution of SUWB compared to the other two methods. Thus, **Table 3** not only demonstrates the candidate consensus across methods but also reveals observations that are sensitive to the choice of resampling scheme.

Table 3. Twenty-Five Observations with the Highest Mean Probability of Procedural Outliers in the Three Wild Bootstrap Schemes.

i	$\hat{p}_{out}(i)$			$\hat{\mu}(w_i)$			$\bar{p}_{out}(i)$
	Mammen	SUWB	STPB	Mammen	SUWB	STPB	
57	0.9253	0.8574	0.3308	0.366085	0.155425	0.66884	0.7045
611	0.9017	0.854	0.3325	0.3727	0.158975	0.66715	0.696067
43	0.9054	0.8469	0.3305	0.373383	0.164438	0.66918	0.694267
382	1	0.7266	0.3328	0.316011	0.302594	0.666913	0.686467
4	0.8564	0.8545	0.3408	0.386698	0.154771	0.65894	0.6839
384	1	0.7024	0.3291	0.332055	0.328605	0.670713	0.677167
42	0.7187	0.8525	0.3324	0.406717	0.159642	0.6673	0.634533
7	0.6472	0.8463	0.3334	0.421897	0.16517	0.666329	0.608967
15	0.8198	0.6261	0.3259	0.418778	0.412671	0.673856	0.5906
621	0.5534	0.8474	0.3347	0.440314	0.16456	0.664989	0.5785
623	0.558	0.8453	0.332	0.436753	0.163742	0.667689	0.578433
793	0.547	0.8507	0.3317	0.442963	0.161851	0.667976	0.576467
605	0.5498	0.8439	0.3296	0.437126	0.166758	0.670135	0.574433
617	0.5007	0.8471	0.3349	0.446417	0.167422	0.664788	0.5609
815	0.4531	0.8397	0.3362	0.460639	0.174194	0.663518	0.543
770	0.4022	0.8384	0.3387	0.471804	0.173708	0.661038	0.526433
67	0.3029	0.8261	0.3365	0.530989	0.185828	0.662901	0.4885
5	0.2855	0.8286	0.3355	0.553176	0.184321	0.664164	0.4832
9	0.4843	0.6224	0.3271	0.441703	0.422634	0.672674	0.477933
25	0.2793	0.8115	0.332	0.66379	0.204021	0.667701	0.474267
36	0.2853	0.803	0.3291	0.567211	0.217212	0.670651	0.472467
3	0.2798	0.7948	0.3331	0.577738	0.225341	0.666699	0.469233
34	0.2798	0.7978	0.3291	0.576225	0.219929	0.670592	0.4689
35	0.2836	0.7912	0.3317	0.688169	0.224237	0.668002	0.468833

The ranking tables in the appendix show that changing the multiplier scheme does not remove the main outlier candidates. However, it does affect how tightly the ranking values are grouped and how sharply the observations are separated. **Tables 2** and **3** provide an important numerical complement to **Figure 4** and help assess the stability of outlier candidates across different resampling schemes.

3.2 Discussion

The results in **Figures 2–4** and **Tables 2–3** show that the three wild bootstrap schemes identify partially overlapping groups of problematic observations, but they differ in how strongly they separate stable inliers, borderline cases, and recurrent outlier candidates. This finding confirms that $\hat{p}_{out}(i)$ should not be read alone. A more reliable interpretation requires a joint reading of $\hat{p}_{out}(i)$, $\hat{p}_{wlow}(i)$, the mean bootstrap weight, the standard deviation of bootstrap weights, and the baseline residual-fitted pattern.

These results are in line with the usual caution in robust regression: atypical observations should be treated as influential cases to be examined, not as points to be removed automatically [3–5]. Robust diagnostics and forward-search procedures can already describe where the leverage is high, how large the residuals are, and whether the fitted structure changes when particular observations receive less influence [6–8]. The analysis in this study uses that diagnostic information as a starting point. Its additional focus is the repeatability of the outlier decision after the residual component is perturbed through the wild bootstrap.

Thus, an observation is not assessed only by its status in the baseline robust fit, but also by how often it returns as an outlier candidate across heteroscedastic bootstrap replications.

The comparison with concrete strength prediction studies is also important. Recent studies on concrete compressive strength have mainly emphasized prediction accuracy, explainability, or model comparison in machine-learning settings [1,2,29-32]. Such studies are useful for estimating strength from mix-design variables, but they often treat problematic observations only as preprocessing issues or residual anomalies. The present results show that outlier analysis can be reported as a separate diagnostic component. This is useful because observations with high fitted values and large negative residuals, as well as observations with moderate fitted values and large positive residuals, may have different engineering interpretations and may require different forms of data checking.

Under the Mammen scheme, the separation between observations is most pronounced. Several observations reach very high $\hat{p}_{out}(i)$ and are also associated with a high probability of being underweighted. However, the leading candidates are not homogeneous. Some have large positive residuals at moderate fitted values, whereas others have very large negative residuals at high fitted values. This result indicates that Mammen is sensitive to more than one deviation pattern. Its benefit is clear separation, but its drawback is that borderline observations may be pushed into sharper diagnostic categories because of the stronger asymmetric multiplier perturbation.

SUWB produces a more compact diagnostic structure. The top observations tend to have high $\hat{p}_{out}(i)$, high $\hat{p}_{wlow}(i)$, low mean bootstrap weight, and large negative residuals in the high-fitted region. The alignment among these indicators makes SUWB easier to interpret as a core-candidate detector. This pattern is useful when the practical goal is to prioritize a smaller set of observations for data-quality review. The drawback is that observations outside the dominant high-fitted negative-residual pattern may receive less diagnostic emphasis than in Mammen.

STPB produces the most restrained decision pattern. The maximum $\hat{p}_{out}(i)$ values are more compressed, and some leading candidates retain relatively high mean weights. This means that repeated inclusion in the outlier set does not always coincide with persistent downweighting. From a practical standpoint, STPB is useful as a cautious diagnostic scheme because it avoids overly strong individual labeling. Its drawback is reduced separation, which can make it harder to distinguish stable outlier candidates from method-sensitive borderline cases.

The relationship between these results and bootstrap theory is direct. Wild bootstrap methods are designed to preserve the fixed covariate structure while perturbing the residual component under heteroscedasticity [19-22]. Robust bootstrap procedures further aim to avoid instability caused by repeatedly resampling influential observations [15,16,26]. By combining these ideas, the proposed framework uses bootstrap refitting not primarily to estimate coefficient uncertainty, but to evaluate the stability of diagnostic decisions. This is the main methodological contribution relative to standard robust regression and standard bootstrap inference.

The benefit of the proposed framework is its auditability. Each observation receives interpretable diagnostic summaries rather than a single binary label. Observations such as 57, 611, 43, 382, 4, and 384 remain prominent across multiplier schemes and therefore deserve priority in further inspection. In contrast, observations whose rankings depend strongly on one multiplier scheme should be interpreted as method-sensitive rather than as definitive outliers. This distinction is valuable for concrete data because potential causes of atypicality

may include measurement error, special mix designs, curing-age effects, batch differences, or unusual material combinations.

The framework also has limitations. First, the analysis is applied to one benchmark dataset, so broader validation on other construction-material datasets is still required. Second, the baseline model is a linear robust regression model, whereas the true relationship between mix composition, curing age, and strength may be nonlinear. Third, the ranking depends on choices such as the outlier threshold, low-weight threshold, bootstrap size B , and robust tuning constants. Fourth, the framework identifies diagnostic priorities but does not by itself determine whether an observation is a measurement error, a valid rare case, or a structurally meaningful extreme mix. These limitations should be addressed through sensitivity analysis and domain-level verification.

4. CONCLUSION

We developed an uncertainty-based framework for outlier detection in robust regression by combining MM-estimation with wild bootstrap refitting on concrete compressive strength data. Instead of reducing robust diagnostics to a single binary outlier label, the framework summarizes each observation through the following quantities: $\hat{p}_{out}(i)$, $\hat{p}_{wlow}(i)$, the mean bootstrap weight, and the standard deviation of bootstrap weights. This formulation makes it possible to examine not only whether an observation is repeatedly identified as an outlier, but also whether its influence remains stable across bootstrap replications.

The empirical results show that Mammen, SUWB, and STPB produce overlapping but non-identical diagnostic patterns. Mammen gives the clearest separation between recurrent outlier candidates and stable inliers. SUWB produces a compact core group in which high outlier probability, high low-weight probability, and low mean bootstrap weight tend to occur together. STPB yields a more restrained pattern, with compressed outlier probabilities and weaker agreement between repeated outlier detection and persistently low robust weights. These differences confirm that outlier status should be interpreted as a stability measure that can depend on the multiplier scheme used for resampling.

The cross-method comparison identifies observations that remain prominent under several multiplier specifications and separates them from observations whose diagnostic status is more method-sensitive. In the concrete compressive strength dataset, this distinction is useful for prioritizing observations for further data-quality inspection and for avoiding overly rigid deletion decisions. The framework therefore extends conventional robust diagnostics by providing an auditable observation-level summary of diagnostic stability.

Several limitations should be noted. The framework was evaluated on a single benchmark dataset and used a linear MM-regression baseline. The conclusions may change under nonlinear robust regression, local regression, or machine-learning models with different residual structures. The results may also depend on the selected outlier threshold, low-weight threshold, bootstrap replication size, and robust tuning constants. Therefore, future research should examine sensitivity to these choices, apply the framework to additional engineering datasets, compare it with alternative robust and resampling procedures, and extend the approach to nonlinear or nonparametric robust models. Such extensions would strengthen the use of wild bootstrap uncertainty maps as transparent diagnostic tools for robust data analysis.

5. ACKNOWLEDGMENT

This research was funded by the PNPB Fund of Universitas Negeri Medan through the University Policy Research Scheme, Contract No. 0calab03/UN33.8/PPKM/PK/2025.

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] Liu, G., and Sun, B. (2023). Concrete compressive strength prediction using an explainable boosting machine model. *Case Studies in Construction Materials*, 18, e01845.
- [2] Chen, X., Zhang, X., and Chen, W. Z. (2024). Advanced predictive modeling of concrete compressive strength and slump characteristics: A comparative evaluation of BPNN, SVM, and RF models optimized via PSO. *Materials*, 17, 4791.
- [3] Wiley.Rabbi, F., Khalil, A., Khan, I., Almuqrin, M. A., Khalil, U., and Andualem, M. (2022). Robust model selection using the out-of-bag bootstrap in linear regression. *Scientific Reports*, 12(1), 10992.
- [4] Riani, M., Atkinson, A. C., Corbellini, A., Farcomeni, A., and Laurini, F. (2024). Information criteria for outlier detection avoiding arbitrary significance levels. *Econometrics and Statistics*, 29, 189–205.
- [5] Zhang, C., Zhang, T., Yin, F., and Zoubir, A. M. (2023). Data-adaptive M-estimators for robust regression via bi-level optimization. *Signal Processing*, 210, 109063.
- [6] Giudici, P., Raffinetti, E., and Riani, M. (2025). Robust machine learning models: linear and nonlinear. *International Journal of Data Science and Analytics*, 20, 1043–1050.
- [7] Atkinson, A. C., Riani, M., and Cerioli, A. (2010). The forward search: Theory and data analysis. *Journal of the Korean Statistical Society*, 39(2), 117–134.
- [8] Riani, M., Perrotta, D., and Torti, F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, 116, 17–32.
- [9] Jana, S., and Ghosh, A. (2026). Robust inference for non-linear regression models with applications in enzyme kinetics. *arXiv:2409.15995*.
- [10] Pötscher, B. M., and Preinerstorfer, D. (2023). How reliable are bootstrap-based heteroskedasticity robust tests?. *Econometric Theory*, 39, 789–847.
- [11] MacKinnon, J. G. (2023). Fast cluster bootstrap methods for linear regression models. *Econometrics and Statistics*, 26, 52–71.
- [12] Wang, Z., Peng, L., and Kim, J. K. (2022). Bootstrap inference for the finite population mean under complex sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4), 1150–1174.
- [13] Webb, M. D. (2023). Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics/Revue canadienne d'économique*, 56(3), 839–858.
- [14] Mansyur, A., Simamora, E., Firdaus, M., Siregar, T. M., and Habibi, R. (2025). Adaptive robust confidence bands on local polynomial regression using residual bootstrap percentiles. *Indonesian Journal of Science and Technology*, 10(2), 355–368.
- [15] Salibian-Barrera, M., and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30(2), 556–582.

- [16] Salibian-Barrera, M. (2006). Bootstrapping MM-estimators for linear regression with fixed designs. *Statistics and Probability Letters*, 76(12), 1287–1297.
- [17] Heng, Q., Shen, H., and Lange, K. (2024). High-dimensional outlier detection via stability. *arXiv:2401.14359*.
- [18] Heng, Q., and Lange, K. (2025). Bootstrap estimation of the proportion of outliers in robust regression. *Statistics and Computing*, 35(1), 3.
- [19] Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, 14(4), 1261–1295.
- [20] Liu, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics*, 16(4), 1696–1708.
- [21] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1), 255–285.
- [22] Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: Wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis*, 49(2), 361–376.
- [23] Kline, P., and Santos, A. (2012). Higher order properties of the wild bootstrap under misspecification. *Journal of Econometrics*, 171(1), 54–70.
- [24] Li, W. (2025). Wild bootstrap inference for linear regressions with many covariates. *arXiv:2506.20972*.
- [25] Simamora, E., Mansyur, A., Mukhtar, Firdaus, M., and Habibi, R. (2026). Adaptive wild bootstrap confidence intervals using skewness-calibrated two-point asymmetric Bernoulli weights in linear regression. *Mathematics and Statistics*, 14(1), 147–166.
- [26] Rana, S., Midi, H., and Imon, A. H. M. R. (2012). Robust wild bootstrap for stabilizing the variance of parameter estimates in heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering*, 2012, 730328.
- [27] Simamora, E., Mansyur, A., Firdaus, M., and Habibi, R. (2025). Residual bootstrap analysis on multiple linear regression: Case study of concrete compressive strength. *AIP Conference Proceedings*, 3316(1), 040003.
- [28] Yeh, I. C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12), 1797–1808.
- [29] Feng, D. C., Liu, Z. T., Wang, X. D., Chen, Y., Chang, J. Q., Wei, D. F., and Jiang, Z. M. (2020). Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. *Construction and Building Materials*, 230, 117000.
- [30] Elhishi, S., Elashry, A. M., and El-Metwally, S. (2023). Unboxing machine learning models for concrete strength prediction using XAI. *Scientific Reports*, 13(1), 19892.
- [31] Elshaarawy, M. K., Alsaadawi, M. M., and Hamed, A. K. (2024). Machine learning and interactive GUI for concrete compressive strength prediction. *Scientific Reports*, 14(1), 16694.
- [32] Tak, M. S. N., Feng, Y., and Mahgoub, M. (2025). Advanced machine learning techniques for predicting concrete compressive strength. *Infrastructures*, 10(2), 26.