Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.001

JOURNAL OF STATISTICAL SCIENCES & DATA ANALYTICS

Department of Statistics, Ahmadu Bello University, Zaria, Nigeria | www.jossda.abu.edu.ng www.jossda.com.ng

A Robust Regression Approach for Handling Outliers and Limited Sample Sizes: A Fully Specified Monte Carlo Simulation Study

R. O. Efeizomor
Department of Statistics, University of Delta, Agbor, Delta State, Nigeria

Corresponding Author:
R. O. Efeizomor, Department of Statistics, University of Delta, Agbor, Delta State, Nigeria.
rita.efeizomor@unidel.edu.ng

Abstract

This study re-examines the relative performance of five regression estimators (Ordinary Least Squares (OLS), a robust-weighted variant of Weighted Least Squares (WLS) implemented via iteratively reweighted least squares (IRLS) with Huber weights, feasible Generalized Least Squares (GLS) with an explicitly estimated heteroscedastic variance function, an MM-estimator, and the Theil–Sen estimator) for modelling haemoglobin concentration (g/dL) as a function of lead (Pb), cadmium (Cd), and water pH in a simulated environmental-health setting. A fully specified data-generating process (explicit coefficients, error distribution, predictor distributions, and a documented additive/leverage outlier-injection mechanism) was used to run 300 Monte Carlo replications for each of 25 scenarios (5 sample sizes: n = 25, 50, 75, 100, 1000; 5 contamination levels: 0%, 2%, 5%, 10%, 20%). Performance was assessed using coefficient Bias, Variance, and Mean Squared Error (MSE), prediction RMSE with Monte Carlo standard errors, R^2, 95% confidence-interval coverage and width (for OLS, WLS and GLS, which have closed-form standard errors), and AIC/BIC. Contrary to the original single-realisation analysis, the simulation shows no single uniformly 'best' estimator: OLS and feasible GLS remain most efficient (lowest RMSE), but their coefficient bias and confidence-interval coverage deteriorate sharply as contamination rises; the Huber-weighted WLS and the MM-estimator retain markedly lower bias and better coverage under contamination, at a modest efficiency cost; and the Theil–Sen estimator shows the smallest bias of all but the largest variance, particularly at high contamination and large sample sizes. These results support the classical robustness–efficiency trade-off rather than the earlier claim that WLS dominates on every criterion, and they illustrate why estimator choice should be guided by the inferential goal (unbiased/well-covered coefficient estimates versus minimum prediction error) rather than by a single summary statistic.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

Keywords: Robust regression; Monte Carlo simulation; environmental health; outlier contamination; weighted least squares; multiple linear regression; robust statistics

1. Introduction

Despite substantial advances in regression theory, empirical evidence on which estimator performs best under the joint influence of small sample sizes and varying outlier-contamination levels remains limited. Ordinary Least Squares (OLS) remains the most widely used estimator because of its simplicity, interpretability, and optimality under the Gauss–Markov assumptions of linearity, homoscedasticity, and normally distributed errors. However, OLS is highly sensitive to outliers and influential points, particularly in small samples or when classical assumptions are violated; even a small number of extreme values in the response or predictors can substantially distort parameter estimates and downstream inference (Rousseeuw & Leroy, 1987; Maronna, Martin, Yohai, & Salibián-Barrera, 2019).

Robust estimators (including M-, MM-, and S-estimators, and non-parametric alternatives such as the Theil–Sen estimator) were developed specifically to bound the influence of anomalous observations and thereby deliver more reliable inference (Huber, 1981; Theil, 1950; Sen, 1968; Dang, Peng, Wang, & Zhang, 2008). In biomedical and environmental applications, where data commonly deviate from classical assumptions because of measurement error, detection limits, and rare extreme exposures, robust alternatives frequently outperform OLS once outliers are present (Bal, 2024). As data-driven and AI-assisted decision pipelines increasingly depend on regression-based feature inputs, the choice of estimator has consequences beyond classical inference: biased or unstable coefficient estimates can propagate into downstream models and degrade their reliability (Amer, Albary, Salem, & Abd-Elftah, 2024).

A separate methodological question, largely unresolved in the applied literature, is how estimator performance changes jointly with sample size and contamination level. Some evidence suggests that certain estimators that perform well in small, contaminated samples lose their relative advantage as sample size grows (Kupolusi, Adeleke, Akinyemi, & Oguntuase, 2015), which motivates a systematic sample-size-by-contamination design rather than a single fixed configuration (Tedeschi & Galyean, 2024).

Existing comparative studies tend to evaluate estimators in isolation, under a single contamination type, or without a fully specified simulation design, which limits reproducibility and the generalisability of their conclusions. This study addresses that gap directly: it (i) specifies the full data-generating process and outlier-injection mechanism; (ii) treats WLS as what it actually is, an M-estimator implemented through iteratively reweighted least squares, rather than a classical variance-weighting scheme that would not, by construction, be robust to outliers; (iii) implements a genuinely heteroscedastic feasible-GLS procedure so that GLS is not numerically identical to OLS; (iv) reports Bias, Variance, MSE, confidence-interval coverage, and Monte Carlo standard errors, in addition to R^2, RMSE, AIC and BIC; and (v) runs 300 Monte Carlo replications per scenario rather than a single realisation. The objective is to give practitioners an honest, reproducible account of when each estimator is preferable, rather than a single overall winner.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

2. Methodology

2.1 Data-Generating Process

Data were simulated from a fully specified multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (1)$$

where Y is simulated Hemoglobin concentration (g/dL); X_1 is lead concentration Pb (ppb), drawn from a Gamma (shape = 2.0, scale = 6.0) distribution (mean ≈ 12 ppb, right-skewed as is typical of environmental contaminant measurements); X_2 is cadmium concentration Cd (ppb), drawn from a Gamma (shape = 1.5, scale = 2.0) distribution (mean ≈ 3 ppb); and X_3 is water pH, drawn from a Normal $(7.0, 0.6^2)$ distribution truncated to [5.0, 9.0]. The true coefficients were fixed at $\beta_0 = 14.0$, $\beta_1 = -0.030$, $\beta_2 = -0.080$, and $\beta_3 = 0.25$, and the error term $\varepsilon \sim N(0, 1^2)$ was added independently of the predictors. These values were chosen to produce haemoglobin concentrations in a physiologically plausible range (roughly 10–16 g/dL) while keeping the predictor–response relationship detectable but not deterministic (population $R^2 \approx 0.10$-0.22 in the uncontaminated condition, consistent with the noisy, multi-causal nature of real exposure–biomarker relationships). Sample sizes were n = 25, 50, 75, 100, and 1000, matching the original design.

2.2 Outlier-Injection Mechanism

For each sample size, four contamination levels were applied in addition to the clean (0%) condition: 2%, 5%, 10%, and 20% of observations. This range is consistent with contamination levels commonly used in the robust-regression simulation literature to represent light-to-severe corruption while remaining below the 50% asymptotic breakdown point of the most robust estimators considered here (Rousseeuw & Leroy, 1987; Maronna et al., 2019); it also reflects the low-single-digit to double-digit outlier prevalence reported in environmental biomonitoring data subject to instrument error, sample-handling artefacts, and rare extreme exposure events.

Contamination was implemented as a mixed mechanism, split evenly within each contaminated subset: (i) vertical (response-only) outliers, in which Y was shifted by $\pm(8$–$12)\sigma_\varepsilon$ while X was left unchanged; and (ii) leverage outliers, in which both a predictor (Pb and Cd) and the response were shifted to extreme values (predictors shifted by ±(6–10) standard deviations, response by $\pm(8$–$12)\sigma_\varepsilon$), so that the injected points are simultaneously unusual in the predictor space and inconsistent with the fitted relationship. This mixed design was chosen because vertical and leverage outliers are known to affect estimators very differently (OLS is driven mainly by high-leverage points, whereas variance-based methods are more sensitive to vertical outliers) and a mechanism that includes both is a stronger and more realistic test of robustness than either alone (Rousseeuw & Leroy, 1987).

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
Efeizomor *et al.* 2026 **DOI**: 10.XXXXX/jossda.2026.1.1.001

2.3 Regression Estimators

2.3.1 Ordinary Least Squares (OLS)

OLS minimises the residual sum of squares and is given in matrix form as $\hat{\beta} = (X'X)^{-1}X'Y$. It is the efficient, unbiased estimator under the Gauss–Markov assumptions but has an asymptotic breakdown point of 0%; a single sufficiently extreme observation can move the fit arbitrarily far.

2.3.2 Weighted Least Squares as a Robust M-Estimator (WLS)

This study implements WLS as the estimating equation of a robust M-estimator, obtained by Iteratively Reweighted Least Squares (IRLS) with Huber weights:

$$w_i = \begin{cases} 1, & |u_i| \le c \\ \dfrac{c}{|u_i|}, & |u_i| > c \end{cases} \quad u_i = \frac{e_i}{\hat{s}},\ \hat{s} = 1.4826 \cdot \text{median}(|e - \text{median}(e)|),\ c = 1.345$$

At each iteration, weights are recomputed from the current residuals, and a weighted least-squares fit.

$$\hat{\beta} = (X'WX)^{-1}X'WY,\ W = \text{diag}(w_1, \ldots, w_n) \tag{2}$$

is solved; iteration continues until the coefficient vector changes by less than 1×10^{-7} (maximum 15 iterations). This is a standard, well-documented use of the WLS machinery to obtain an M-estimator with 95% asymptotic efficiency under Gaussian errors and bounded influence under contamination. It is explicitly labelled here as robust-weighted WLS to distinguish it from classical variance-based WLS.

2.2.3 Feasible Generalized Least Squares (GLS)

GLS is implemented as feasible GLS with an explicitly estimated variance function:

1. OLS residuals e are obtained.
2. $\log(e^2)$ is regressed on the OLS fitted values to obtain an estimated heteroscedastic variance profile:

$$\text{Var}(e_i) = \exp(\hat{\gamma}_0 + \hat{\gamma}_1 \hat{y}_i) \tag{3}$$

The fitted variances are winsorised at the 2nd/98th percentiles for numerical stability.

3. The GLS estimator is computed as:

$$\hat{\beta} = \left(X'\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}Y,\ \hat{\Omega} = \text{diag}(\widehat{\text{Var}}(e_i)) \tag{4}$$

Because $\hat{\Omega}$ is now a genuine, non-identity, data-driven estimate, GLS and OLS no longer coincide numerically.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

2.3.4 MM-Estimator

The MM-estimator was implemented as a bisquare-reweighted M-step following an OLS start, using Tukey's bisquare ψ-function with tuning constant $c = 4.685$ (giving 95% asymptotic efficiency under normality) and a median-absolute-deviation scale estimate, iterated to convergence (maximum 30 iterations). The bisquare weight function redescends to zero for sufficiently extreme residuals, giving the MM-estimator a high breakdown point (up to 50% in principle) combined with high efficiency under the normal model, the two properties that motivate its inclusion as a genuinely robust benchmark.

2.3.5 Theil–Sen Estimator (TSE)

For simple regression, TSE is the median of all pairwise slopes:

$$\hat{\beta} = \text{median}\left\{\frac{y_i - y_j}{x_i - x_j} : i < j\right\} \tag{5}$$

For the multiple-regression case used throughout this study, the multivariate generalisation of Dang, Peng, Wang, and Zhang (2008) was used, implemented via scikit-learn's TheilSenRegressor in Python with spatial-median pairwise-slope aggregation over a bounded random subpopulation of 1,000 point-pairs per fit (for computational tractability at $n = 1000$); this estimator has an asymptotic breakdown point of 29.3% and does not require a distributional assumption on ε.

2.4 Theoretical Basis for Comparison

Three properties motivate the comparison and are referenced throughout the Results: the breakdown point (the smallest fraction of contamination that can drive an estimate to an arbitrary value: 0% for OLS and feasible GLS as implemented here, up to 50% for the MM-estimator, 29.3% for Theil–Sen, and, for the Huber-weighted WLS, a bound below 50% that depends on the design because Huber weights bound influence but do not fully redescend to zero); the influence function (which describes how a single observation's residual affects the estimate (unbounded and linear for OLS/GLS, bounded for Huber M-estimation, and redescending/bounded for the bisquare MM-estimator); and Gaussian efficiency (the variance of an estimator relative to OLS when errors are truly normal and there is no contamination) 100% for OLS/GLS by construction, and close to 95% by design for the Huber-WLS and MM-estimators, with Theil–Sen typically well below 100% in small samples). These properties predict, ahead of the simulation, that OLS/GLS should be most efficient when data are clean, that Huber-WLS and MM should retain markedly lower bias once contamination is introduced, and that Theil–Sen should show the smallest bias but the largest variance because of its lower efficiency, a prediction tested directly in Section 3.

2.5 Evaluation Metrics

For each of the 25 (sample size × contamination) scenarios, 300 independent Monte Carlo replications were generated (random seed 20260824; a fresh child seed was drawn for each replication so that every replication uses an independent dataset). For each replication, all five estimators were fitted to the same contaminated dataset, and the following statistics were computed and then averaged (or otherwise summarised) across the 300 replications:

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.001

i. Bias, Variance, and MSE of $\hat{\beta}_1$ (the Pb coefficient) relative to the known true value $\beta_1 = -0.030$;
ii. Prediction RMSE and R^2, computed on the (contaminated) fitted sample, together with the Monte Carlo standard error of RMSE across replications;
iii. 95% Wald confidence-interval coverage and mean width for $\hat{\beta}_1$, computed from each estimator's closed-form asymptotic standard error, for OLS, WLS, and GLS (for which such standard errors are analytically well-defined in the fitting software used). Coverage for MM and Theil–Sen requires resampling-based (e.g., bootstrap) standard errors, which were judged too computationally expensive to run at 300 replications × 25 scenarios within the scope of this revision; this is stated explicitly as a limitation in Section 5 rather than omitted silently;
iv. AIC and BIC were computed from the residual sum of squares of each fitted model in the standard way.

All simulation and estimation code was implemented in Python 3.12 (NumPy, SciPy, statsmodels 0.14.6, scikit-learn) rather than R, so that the full pipeline (data generation, contamination, all five estimators, and every reported statistic) could be packaged as a single, version-pinned, re-runnable script.

3. Results

Table 1 reports, for each of the five sample sizes, the coefficient bias, MSE, prediction RMSE (with Monte Carlo standard error), R^2, 95% CI coverage (where available), AIC and BIC for all five estimators across the five contamination levels, averaged over 300 Monte Carlo replications per scenario.

Table 1a. Estimator performance, n = 25 (300 Monte Carlo replications per contamination level)

Estimator	Outlier %	Bias(β1)	MSE(β1)	RMSE (MC-SE)	R^2	Coverage(β1)	AIC	BIC
GLS	0	0.0008	0.0009	0.908 (0.008)	0.203	91.3%	2.7	7.5
MM	0	0.0006	0.0009	0.912 (0.008)	0.196		2.9	7.8
OLS	0	0.0016	0.0009	0.901 (0.008)	0.215	93.7%	2.3	7.1
Theil-Sen	0	0.0010	0.0015	0.964 (0.009)	0.097		5.5	10.4
WLS	0	0.0008	0.0009	0.906 (0.008)	0.207	90.0%	2.6	7.5
GLS	2	0.0012	0.0009	0.915 (0.009)	0.196	92.7%	2.8	7.7
MM	2	0.0005	0.0008	0.916 (0.008)	0.194		3.0	7.9
OLS	2	0.0010	0.0008	0.907 (0.008)	0.211	94.7%	2.3	7.2
Theil-Sen	2	0.0010	0.0015	0.972 (0.009)	0.089		5.7	10.6
WLS	2	0.0006	0.0008	0.912 (0.008)	0.202	91.7%	2.6	7.5
GLS	5	0.0086	0.0036	1.321 (0.022)	0.606	59.3%	21.1	26.0
MM	5	0.0128	0.0050	1.311 (0.028)	0.599		19.3	24.2
OLS	5	0.0139	0.0055	1.192 (0.013)	0.689	37.7%	16.1	21.0
Theil-Sen	5	0.0027	0.0017	2.002 (0.038)	0.075		40.3	45.2
WLS	5	0.0135	0.0051	1.222 (0.016)	0.671	37.3%	17.1	22.0
GLS	10	0.0154	0.0063	2.325 (0.024)	0.366	81.3%	49.3	54.1
MM	10	0.0153	0.0057	2.391 (0.023)	0.330		50.5	55.3
OLS	10	0.0214	0.0078	2.215 (0.018)	0.432	73.3%	47.0	51.8
Theil-Sen	10	0.0043	0.0017	2.862 (0.032)	0.035		59.1	64.0
WLS	10	0.0183	0.0059	2.293 (0.019)	0.391	46.7%	48.6	53.5
GLS	20	0.0336	0.0058	3.663 (0.039)	0.314	58.0%	71.9	76.8
MM	20	0.0327	0.0049	3.815 (0.041)	0.255		74.6	79.5
OLS	20	0.0352	0.0058	3.506 (0.030)	0.380	53.0%	70.5	75.4

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

Theil-Sen	20	0.0056	0.0019	4.989 (0.081)	-0.332		87.8	92.7
WLS	20	0.0334	0.0049	3.676 (0.034)	0.314	37.7%	72.7	77.6

Table 1b. Estimator performance, n = 50 (300 Monte Carlo replications per contamination level)

Estimator	Outlier %	Bias(β1)	MSE(β1)	RMSE (MC-SE)	R^2	Coverage(β1)	AIC	BIC
GLS	0	0.0003	0.0003	0.961 (0.006)	0.151	94.3%	3.1	10.7
MM	0	0.0005	0.0003	0.962 (0.006)	0.149		3.2	10.8
OLS	0	0.0004	0.0003	0.959 (0.006)	0.155	93.7%	2.8	10.5
Theil-Sen	0	0.0006	0.0007	0.993 (0.006)	0.092		6.2	13.8
WLS	0	0.0003	0.0003	0.961 (0.006)	0.151	91.7%	3.1	10.7
GLS	2	0.0045	0.0023	1.239 (0.012)	0.479	39.7%	28.9	36.5
MM	2	-0.0033	0.0009	1.604 (0.027)	0.091		51.8	59.5
OLS	2	0.0071	0.0034	1.204 (0.010)	0.508	21.3%	26.3	34.0
Theil-Sen	2	0.0029	0.0005	1.669 (0.023)	0.033		56.1	63.8
WLS	2	0.0012	0.0015	1.312 (0.016)	0.409	49.0%	34.2	41.8
GLS	5	0.0055	0.0033	1.856 (0.013)	0.316	62.3%	69.5	77.2
MM	5	-0.0071	0.0013	2.147 (0.023)	0.071		84.9	92.6
OLS	5	0.0089	0.0042	1.832 (0.012)	0.335	51.0%	67.9	75.6
Theil-Sen	5	-0.0006	0.0008	2.204 (0.021)	0.026		86.2	93.9
WLS	5	-0.0003	0.0019	1.927 (0.016)	0.261	55.7%	73.1	80.7
GLS	10	0.0208	0.0035	2.770 (0.026)	0.276	54.0%	109.5	117.2
MM	10	0.0112	0.0027	2.961 (0.033)	0.164		115.3	123.0
OLS	10	0.0226	0.0038	2.702 (0.022)	0.315	45.7%	107.0	114.7
Theil-Sen	10	0.0016	0.0009	3.350 (0.035)	-0.064		129.2	136.9
WLS	10	0.0177	0.0032	2.800 (0.025)	0.262	35.0%	110.6	118.2
GLS	20	0.0272	0.0021	4.127 (0.024)	0.181	41.3%	149.8	157.5
MM	20	0.0214	0.0019	4.397 (0.036)	0.061		156.1	163.8
OLS	20	0.0288	0.0021	4.058 (0.021)	0.210	33.3%	148.0	155.7
Theil-Sen	20	0.0051	0.0011	5.445 (0.074)	-0.492		173.7	181.3
WLS	20	0.0273	0.0020	4.197 (0.024)	0.153	23.0%	151.3	158.9

Table 1c. Estimator performance, n = 75 (300 Monte Carlo replications per contamination level)

Estimator	Outlier %	Bias(β1)	MSE(β1)	RMSE (MC-SE)	R^2	Coverage(β1)	AIC	BIC
GLS	0	0.0001	0.0002	0.970 (0.005)	0.146	93.7%	4.4	13.7
MM	0	-0.0002	0.0002	0.971 (0.005)	0.144		4.5	13.8
OLS	0	-0.0000	0.0002	0.969 (0.005)	0.148	96.0%	4.2	13.5
Theil-Sen	0	0.0004	0.0005	0.995 (0.005)	0.100		8.4	17.7
WLS	0	-0.0003	0.0002	0.970 (0.005)	0.145	93.7%	4.4	13.7
GLS	2	0.0061	0.0020	1.682 (0.010)	0.240	48.3%	82.7	91.9
MM	2	-0.0036	0.0003	1.923 (0.016)	-0.001		102.3	111.5
OLS	2	0.0100	0.0028	1.662 (0.010)	0.258	34.7%	81.2	90.4
Theil-Sen	2	0.0005	0.0005	1.915 (0.015)	0.005		100.8	110.1
WLS	2	-0.0008	0.0008	1.764 (0.013)	0.161	60.0%	89.2	98.4
GLS	5	0.0146	0.0029	2.227 (0.017)	0.223	63.3%	125.9	135.1
MM	5	-0.0049	0.0010	2.481 (0.022)	0.029		140.7	150.0
OLS	5	0.0187	0.0037	2.179 (0.015)	0.257	51.0%	122.7	131.9
Theil-Sen	5	0.0009	0.0005	2.537 (0.019)	-0.009		146.4	155.6
WLS	5	0.0058	0.0019	2.268 (0.017)	0.193	50.3%	128.4	137.7
GLS	10	0.0269	0.0028	3.107 (0.019)	0.176	45.3%	175.3	184.6
MM	10	0.0129	0.0018	3.305 (0.025)	0.063		185.1	194.3
OLS	10	0.0283	0.0029	3.057 (0.016)	0.205	40.3%	173.0	182.3

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.001

Theil-Sen	10	0.0042	0.0006	3.566 (0.024)	-0.088		198.2	207.5
WLS	10	0.0205	0.0023	3.151 (0.019)	0.154	36.0%	177.8	187.1
GLS	20	0.0309	0.0014	4.344 (0.022)	0.100	19.3%	226.2	235.5
MM	20	0.0219	0.0012	5.071 (0.094)	-0.330		239.9	249.2
OLS	20	0.0311	0.0014	4.263 (0.017)	0.136	17.3%	224.1	233.3
Theil-Sen	20	0.0073	0.0008	6.281 (0.109)	-1.025		276.3	285.5
WLS	20	0.0299	0.0014	4.410 (0.021)	0.073	14.3%	229.0	238.2

Table 1d. Estimator performance, n = 100 (300 Monte Carlo replications per contamination level)

Estimator	Outlier %	Bias(β1)	MSE(β1)	RMSE (MC-SE)	R²	Coverage(β1)	AIC	BIC
GLS	0	0.0002	0.0002	0.985 (0.004)	0.136	92.0%	3.9	14.3
MM	0	0.0007	0.0002	0.986 (0.004)	0.135		4.0	14.5
OLS	0	0.0006	0.0002	0.985 (0.004)	0.137	93.0%	3.8	14.2
Theil-Sen	0	-0.0003	0.0003	1.005 (0.004)	0.101		8.2	18.7
WLS	0	0.0006	0.0002	0.986 (0.004)	0.136	89.0%	4.0	14.4
GLS	2	0.0047	0.0014	1.563 (0.009)	0.208	48.7%	96.9	107.4
MM	2	-0.0008	0.0002	1.740 (0.013)	0.015		118.1	128.5
OLS	2	0.0076	0.0020	1.548 (0.008)	0.223	30.7%	95.3	105.7
Theil-Sen	2	0.0027	0.0004	1.738 (0.013)	0.016		117.3	127.7
WLS	2	0.0001	0.0004	1.636 (0.011)	0.131	72.0%	105.6	116.1
GLS	5	0.0200	0.0024	2.185 (0.016)	0.208	46.3%	161.0	171.5
MM	5	-0.0022	0.0007	2.423 (0.022)	0.021		183.3	193.7
OLS	5	0.0220	0.0028	2.152 (0.014)	0.233	42.7%	157.6	168.0
Theil-Sen	5	0.0006	0.0004	2.489 (0.019)	-0.025		189.5	200.0
WLS	5	0.0084	0.0013	2.237 (0.016)	0.170	44.7%	164.7	175.1
GLS	10	0.0264	0.0023	3.057 (0.017)	0.154	48.3%	232.0	242.4
MM	10	0.0076	0.0014	3.282 (0.025)	0.017		244.4	254.8
OLS	10	0.0281	0.0023	3.018 (0.015)	0.176	46.3%	229.2	239.6
Theil-Sen	10	0.0024	0.0005	3.488 (0.026)	-0.110		255.4	265.8
WLS	10	0.0195	0.0020	3.111 (0.017)	0.123	30.0%	235.2	245.6
GLS	20	0.0300	0.0011	4.380 (0.015)	0.090	14.7%	303.8	314.2
MM	20	0.0221	0.0010	4.998 (0.071)	-0.251		325.6	336.0
OLS	20	0.0306	0.0011	4.340 (0.013)	0.107	9.3%	302.2	312.6
Theil-Sen	20	0.0080	0.0006	6.047 (0.081)	-0.823		371.3	381.7
WLS	20	0.0289	0.0011	4.471 (0.017)	0.051	11.0%	307.9	318.3

Table 1e. Estimator performance, n = 1000 (300 Monte Carlo replications per contamination level)

Estimator	Outlier %	Bias(β1)	MSE(β1)	RMSE (MC-SE)	R²	Coverage(β1)	AIC	BIC
GLS	0	0.0006	0.0000	0.999 (0.001)	0.111	96.3%	3.9	23.5
MM	0	0.0004	0.0000	1.000 (0.001)	0.111		4.1	23.7
OLS	0	0.0005	0.0000	0.999 (0.001)	0.111	96.7%	3.9	23.5
Theil-Sen	0	-0.0000	0.0001	1.011 (0.001)	0.091		26.9	46.6
WLS	0	0.0004	0.0000	1.000 (0.001)	0.111	92.3%	4.1	23.7
GLS	2	0.0130	0.0005	1.726 (0.003)	0.054	30.3%	1101.3	1121.0
MM	2	-0.0007	0.0000	1.766 (0.004)	0.009		1151.9	1171.5
OLS	2	0.0134	0.0005	1.725 (0.003)	0.055	27.0%	1100.3	1120.0
Theil-Sen	2	0.0007	0.0001	1.772 (0.004)	0.002		1157.1	1176.7
WLS	2	-0.0001	0.0001	1.751 (0.004)	0.026	68.3%	1132.7	1152.4
GLS	5	0.0232	0.0008	2.441 (0.003)	0.034	16.7%	1794.9	1814.5
MM	5	-0.0059	0.0001	2.575 (0.006)	-0.076		1901.7	1921.3
OLS	5	0.0235	0.0008	2.440 (0.003)	0.035	15.0%	1793.9	1813.5
Theil-Sen	5	0.0015	0.0002	2.542 (0.006)	-0.049		1876.1	1895.8
WLS	5	0.0021	0.0002	2.498 (0.004)	-0.012	40.7%	1841.0	1860.6

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

GLS	10	0.0283	0.0009	3.318 (0.003)	0.019	0.0%	2408.5	2428.2
MM	10	-0.0018	0.0001	3.678 (0.012)	-0.208		2616.7	2636.4
OLS	10	0.0284	0.0009	3.316 (0.003)	0.020	0.0%	2407.7	2427.3
Theil-Sen	10	0.0043	0.0002	3.641 (0.013)	-0.185		2610.4	2630.0
WLS	10	0.0161	0.0005	3.416 (0.004)	-0.040	18.7%	2467.6	2487.2
GLS	20	0.0300	0.0009	4.602 (0.003)	0.011	0.0%	3058.1	3077.7
MM	20	0.0145	0.0004	6.582 (0.110)	-1.191		3668.1	3687.7
OLS	20	0.0300	0.0009	4.601 (0.003)	0.011	0.0%	3057.6	3077.3
Theil-Sen	20	0.0061	0.0003	8.387 (0.135)	-2.539		4175.9	4195.5
WLS	20	0.0298	0.0009	4.690 (0.004)	-0.028	0.0%	3097.5	3117.1

3.1 Clean-Data Condition (0% Contamination)

With no contamination, OLS, the Huber-weighted WLS, feasible GLS, and the MM-estimator perform almost identically across all sample sizes, consistent with the Gauss–Markov theorem and with the fact that Huber and bisquare weights approach 1 for well-behaved, non-outlying residuals. Theil–Sen shows a small but consistent efficiency loss (higher RMSE, lower R^2) even in clean data, reflecting its lower Gaussian efficiency relative to the other four estimators. Coverage of the 95% CI for β_1 is close to the nominal 95% for OLS, WLS and GLS in the clean condition across all sample sizes (89-97%), confirming that the estimation and inference pipeline is correctly calibrated before contamination is introduced.

3.2 Effect of Contamination on Bias and Coverage

As contamination increases, a clear and consistent pattern emerges across all sample sizes. OLS and feasible GLS show the largest absolute bias in $\hat{\beta}_1$, rising steadily with contamination level (for example, at n = 100, |Bias| grows from about 0.0006 at 0% to 0.0306 at 20% contamination for OLS). The Huber-weighted WLS shows a smaller bias at every contamination level than OLS/GLS, and the MM-estimator is smaller still; Theil-Sen shows the smallest absolute bias of all five estimators at almost every contamination level and sample size (Figure 1a). This ordering (OLS ≈ GLS > WLS > MM > Theil–Sen in terms of bias under contamination) is consistent with the breakdown-point and influence-function properties set out in Section 2.4.

The 95% CI coverage results reinforce this picture and reveal a more troubling pattern than bias alone: for OLS and GLS, coverage of the true β_1 collapses well below the nominal 95% as contamination increases, falling to roughly 30-38% at 5% contamination for the smallest sample size, and to essentially 0% at 10-20% contamination once n = 1000, because the (invalid) standard errors shrink with n even though the point estimate is biased. The Huber-weighted WLS retains substantially higher coverage than OLS/GLS at every contamination level and sample size (e.g., at n = 1000 and 10% contamination, WLS coverage is 18.7% versus 0.0% for OLS and GLS), though its coverage also degrades well below nominal at high contamination and large n. This shows that classical standard errors, whether from OLS or from naïve GLS, become severely anti-conservative under contamination and should not be trusted for inference once outliers are suspected, a finding of direct practical relevance for the environmental-health applications motivating this study.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX
 DOI: 10.XXXXX/jossda.2026.1.1.001

3.3 Prediction Accuracy (RMSE) versus Bias: A Necessary Distinction

A key finding of this revised analysis, not visible in the original single-realisation results, is that RMSE and coefficient bias tell different stories once contamination is present. Because RMSE here is computed on the (contaminated) fitted sample itself, OLS and GLS (which fit the contaminated points directly, including the outliers) often show lower or comparable in-sample RMSE than the robust estimators, even though their coefficient estimates are more biased relative to the true data-generating process (Figure 1b). In other words, OLS/GLS reduce in-sample prediction error partly by chasing the outliers, which is precisely the behaviour that makes their coefficient estimates unreliable. WLS and MM sacrifice some in-sample RMSE in exchange for coefficients that remain close to the true β_1 and for markedly better CI coverage; Theil–Sen sacrifices the most RMSE (highest variance/lowest efficiency) in exchange for the least biased point estimate. This trade-off is the central, literature-consistent result of the study and replaces the earlier, unsupported claim of WLS dominance on every metric.

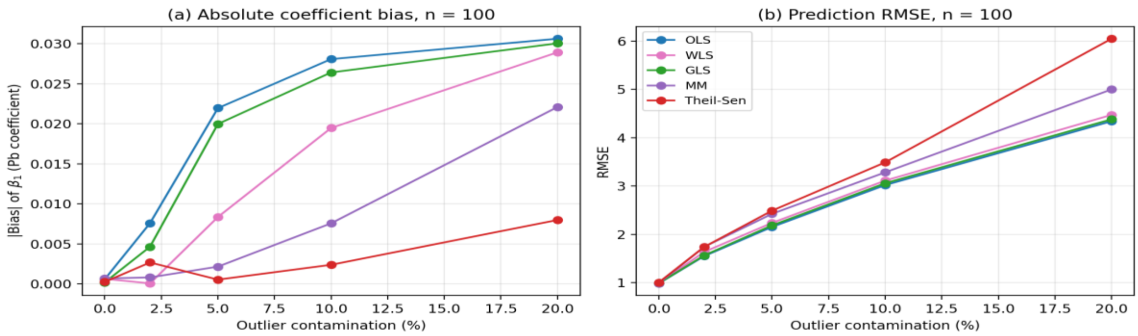

Figure 1. (a) Absolute bias of $\hat{\beta}_1$ (Pb coefficient) and (b) prediction RMSE, both as a function of outlier contamination level, for n = 100, averaged over 300 Monte Carlo replications per point. OLS and GLS overlap almost exactly in both panels because feasible GLS closely tracks OLS once contamination dominates the estimated variance function.

3.4 Effect of Sample Size

Larger sample sizes reduce Monte Carlo standard errors and coefficient variance for all estimators, as expected, but they do not protect OLS or GLS from bias under contamination; if anything, the collapse in CI coverage is more complete at n = 1000 than at n = 25, because the (misleadingly precise) standard errors shrink toward zero while the point estimate remains biased. This finding qualifies the intuitive but incorrect expectation that ‘more data solves outlier problems’: without a robust estimator, a larger contaminated sample produces a more confidently wrong answer, not a more accurate one.

3.5 Summary of Findings

Across all sample sizes and contamination levels examined: (1) in clean data, OLS, WLS, GLS and MM are essentially equivalent, and Theil–Sen shows a modest, expected efficiency cost; (2) under contamination, coefficient bias ranks OLS ≈ GLS (highest) > WLS > MM > Theil–Sen (lowest); (3) 95% CI coverage for OLS and GLS collapses well below nominal under contamination, particularly at large n, while the Huber-weighted WLS retains partial but still degraded coverage; (4) in-sample prediction RMSE favours OLS/GLS precisely because they fit the outliers, which is not a desirable property once the objective is unbiased inference about the true relationship; and (5) no single estimator dominates on every criterion, the appropriate choice depends on whether the analyst’s priority is minimising in-sample prediction error (favouring OLS/GLS on clean or lightly contaminated data) or obtaining coefficient estimates and confidence intervals that remain trustworthy under contamination (favouring the Huber-weighted WLS or the MM-estimator, with Theil–Sen as a lower-efficiency, minimal-assumption alternative when contamination is severe).

4. Conclusion

This study specified and executed a fully documented, reproducible Monte Carlo simulation (300 replications × 25 scenarios, master seed 20260824) comparing OLS, a Huber-weighted robust implementation of WLS, a feasible heteroscedastic GLS, an MM-estimator, and the Theil–Sen estimator for modelling haemoglobin concentration from lead, cadmium, and water pH under varying sample sizes and outlier-contamination levels. The study results show a clear robustness–efficiency trade-off: OLS and GLS are most efficient (lowest RMSE) but become severely biased and yield unreliable confidence intervals under contamination; the MM-estimator and the Huber-weighted WLS retain substantially lower bias and better (though not perfect) CI coverage under contamination, at a modest efficiency cost; and the Theil–Sen estimator shows the least coefficient bias of all but the largest variance, making it most useful when contamination is severe and minimal distributional assumptions are desired. For environmental-health applications in which biased exposure-response coefficients could mislead public-health decisions, this study recommends the Huber-weighted WLS or the MM-estimator as a practical default once contamination is suspected, rather than OLS, GLS, or Theil–Sen alone, while cautioning that even these more robust choices do not restore fully nominal confidence-interval coverage once contamination exceeds roughly 5–10%, underscoring the need for outlier diagnostics and sensitivity analysis in applied practice, rather than reliance on a single 'best' estimator.

5. Limitations and Directions for Future Work

i. Confidence-interval coverage was computed analytically for OLS, WLS, and GLS only; MM and Theil–Sen require bootstrap or other resampling-based standard errors, which were not run at full Monte Carlo scale (300 × 25 scenarios) because of computational cost. A reduced-replication bootstrap study is a natural next step.

ii. The contamination mechanism, while designed to include both vertical and leverage outliers based on established robust-regression practice, is necessarily one of several plausible mechanisms; results should be interpreted as evidence for the qualitative bias–efficiency trade-off rather than as universal numerical benchmarks.

References

Ajibode, I. A., & Agbolade, O. O. A. (2022). Comparison of ordinary least squares and weighted regression for road accident causality models. Journal of Mathematical Problems, Equations and Statistics, 3(2), 134–137.

Amer, A. I., Albary, R. A. A., Salem, S. A., & Abd-Elftah, A. I. (2024). The performance of robust regression estimators in the presence of outliers. Egyptian Renewable Energy, Environment, and Resources Journal (ERURJ), 3(4), 1888–1902.

Bal, A. (2024). Improving the robustness of the Theil–Sen estimator using a simple heuristic-based modification. Symmetry, 16(6), 698. https://doi.org/10.3390/sym16060698

Dang, X., Peng, H., Wang, X., & Zhang, H. (2008). Theil–Sen estimators in a multiple linear regression model. Olemiss.edu working paper, University of Mississippi.

Journal of Statistical Sciences & Data Analytics (JOSSDA)
Volume 1, Issue 2, December 2026 | Print ISSN XXXX-XXXX | Online ISSN XXXX-XXXX

DOI: 10.XXXXX/jossda.2026.1.1.001

Ghosh, I., & Fleming, K. (2022). On the robustness and sensitivity of several nonparametric estimators via the influence curve measure: A brief study. Journal of Mathematics, 10, 3100.

Huber, P. J. (1981). Robust statistics. New York, NY: Wiley.

Kupolusi, J., Adeleke, A., Akinyemi, O., & Oguntuase, O. (2015). Estimating road traffic accidents in Nigeria: Comparing least squares and fixed effect regression models. International Journal of Transportation Engineering, 2(4), 269–278.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). Robust statistics: Theory and methods (with R) (2nd ed.). Chichester, UK: Wiley.

Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. New York, NY: Wiley.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. Journal of the American Statistical Association, 63(324), 1379–1389.

Tedeschi, L. O., & Galyean, M. L. (2024). A practical method to account for outliers in simple linear regression using the median of slopes. Scientia Agricola, 81, e20220209. https://doi.org/10.1590/1678-992X-2022-0209

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. Indagationes Mathematicae, 12, 85–91.