Pham T., Talavera O. & Tsapin A. (2021). Shock contagion, asset quality, and lending behavior: The case of war in Eastern Ukraine. *Kyklos, 2*(74), 243–269. https://doi.org/10.1111/kykl.12261
Phan D. H. B., Tran V. T. & Iyke B. N. (2022). Geopolitical risk and bank stability. *Finance Research Letters, Part B*(46), 102453. https://doi.org/10.1016/j.frl.2021.102453
Shabir M., Jiang P., Wang W. & Işık Ö. (2023). COVID-19 pandemic impact on banking sector: A cross-country analysis. *Journal of Multinational Financial Management, 67*, 100784. https://doi.org/10.1016/j.mulfin.2023.100784
Shanti R., Siregar H., Zulbainarni N. & Tony (2024). Revolutionizing banking: digital transformation of neobanks for improved performance. *Journal of Risk and Financial Management, 5*(17), 188. https://doi.org/10.3390/jrfm17050188
Sharma R. & Chauhan R. (2023). Impact of Basel III liquidity and capital regulations on bank lending and financial stability: Evidence from emerging countries. *Journal of Corporate Accounting & Finance, 4*(34), 28–45. https://doi.org/10.1002/jcaf.22630
Xiazi X. & Shabir M. (2022). Coronavirus pandemic impact on bank performance. *Frontiers in Psychology, 13*, 1014009. https://doi.org/10.3389/fpsyg.2022.1014009
Zhao J., Wang K., Ibrahim H. & Chen Y. (2024). The impact of digital financial inclusion on bank performance: An investigation of mechanisms and heterogeneity. *PLOS ONE, 8*(19), e0309099. https://doi.org/10.1371/journal.pone.0309099

Дослідження не отримувало зовнішнього фінансування.
Автори заявляють про відсутність конфлікту інтересів.
Штучний інтелект для підготовки рукопису не використовувався.

Стаття надійшла до редакції / Received: 08.06.2026
Статтю прийнято до публікації / Accepted: 21.06.2026
Оприлюднено / Published: 30.07.2026

UDC 330.43:519.237.5
JEL: C15; C21; C52
DOI: https://doi.org/10.32983/2222-4459-2026-6-233-244

IMPACT OF STRUCTURED OUTLIERS ON REGRESSION INFERENCE

©2026 **MANZHOS T. V., MELNYK O. O.**

UDC 330.43:519.237.5
JEL: C15; C21; C52

Manzhos T. V., Melnyk O. O. Impact of Structured Outliers on Regression Inference

The article explores the problem of the reliability of regression analysis results in the presence of atypical observations. In econometric studies, such observations can occur in various types of data: macroeconomic, financial, regional, sectoral, market, or broader socioeconomic data. Their appearance is not always related to random errors or technical inaccuracies. Often, they reflect crisis periods, structural shifts, the heterogeneity of the studied objects, the concentration of certain indicators, or the presence of local patterns in the data. Under these conditions, it is important not only to detect outliers but also to understand how their nature affects coefficient estimation, statistical conclusions, and the economic interpretation of the model. The article aims to compare the consequences of different types of outliers for estimating the parameters of a linear regression. The article considers five scenarios: random vertical outliers, non-random vertical outliers, high-leverage observations that align with the basic regression relationship, high-leverage observations that distort this relationship, and clustered high-leverage observations that might form a separate local regime. The study is based on Monte Carlo simulations with 1000 repetitions. The base model includes three regressors with a given correlation structure; one of the coefficients is zero. This allows for assessing not only the bias of the main coefficient but also the appearance of a false effect for a variable that, according to the model, does not affect the outcome. The article compares three evaluation approaches: the usual least squares method, Huber regression, and the least trimmed squares method. The quality of evaluation is analyzed based on coefficient bias, root mean square error, sign stability, the size of false effects, and the frequency of large false effects. Simulation results show that the impact of outliers on a regression model is determined not only by their presence but also by how they actually occur. Vertical outliers primarily increase the variability of residuals and are relatively well handled by Huber regression. On the other hand, high-leverage observations require separate interpretation: if they follow the underlying regression relationship, they should not be automatically considered problematic. But if such observations change the slope of that relationship, they can significantly bias coefficient estimates and create false effects for correlated regressors. In such cases, the least trimmed squares method turns out to be more effective than Huber regression. At the same time, clustered observations with high leverage remain difficult to correct even after trimming part of the data, which may indicate structural heterogeneity in the sample. The scientific novelty of the article lies in the fact that different types of atypical observations are compared not only by their presence but also by the nature of their impact on econometric estimation. The practical value of the results lies in forming guidelines for choosing a robust method. If the problem is mainly related to large residuals, Huber regression is appropriate. If the atypical observations have high leverage and alter the shape of the regression relationship, it is more reasonable to use trimming methods or conduct additional subgroup and local regime analyses.

Keywords: *regression inference; structured outliers; high-leverage observations; least squares method; Huber regression; least trimmed squares method; Monte Carlo modeling; econometric estimation.*
Fig.: *4.* ***Tabl.:*** *3.* ***Formulae:*** *17.* ***Bibl.:*** *16.*

Manzhos Tetiana V. *– PhD (Physics and Mathematics), Associate Professor, Associate Professor of the Department of Higher Mathematics, Kyiv National Economic University named after Vadym Hetman (54/1 Beresteiskyi Ave., Kyiv, 03057, Ukraine)*
E-mail: *tmanzhos@kneu.edu.ua*
ORCID: *https://orcid.org/0000-0003-1393-9116*

Melnyk Olha O. *– PhD (Physics and Mathematics), Associate Professor, Associate Professor of the Department of Higher Mathematics, Kyiv National Economic University named after Vadym Hetman (54/1 Beresteiskyi Ave., Kyiv, 03057, Ukraine)*
E-mail: *melnyk.olha@kneu.edu.ua*
ORCID: *https://orcid.org/0000-0002-4399-176X*

УДК 330.43:519.237.5
JEL: C15; C21; C52

Манжос Т. В., Мельник О. О. Вплив структурованих викидів на регресійний висновок

У статті досліджено проблему надійності результатів регресійного аналізу за наявності нетипових спостережень. В економетричних дослідженнях такі спостереження можуть траплятися в різних типах даних: макроекономічних, фінансових, регіональних, галузевих, ринкових або ширших соціально-економічних. Їх поява не завжди пов'язана з випадковими помилками чи технічними неточностями. Часто вони відображають кризові періоди, структурні зрушення, неоднорідність досліджуваних об'єктів, концентрацію окремих показників або наявність локальних режимів у даних. За таких умов важливо не лише виявити викиди, а й зрозуміти, як їхня природа впливає на оцінювання коефіцієнтів, статистичні висновки та економічну інтерпретацію моделі. Метою статті є порівняння наслідків різних типів викидів для оцінювання параметрів лінійної регресії. У роботі розглянуто п'ять сценаріїв: випадкові вертикальні викиди, невипадкові вертикальні викиди, спостереження з високим важелем, що узгоджуються з базовою регресійною залежністю, спостереження з високим важелем, що спотворюють цю залежність, а також кластеризовані спостереження з високим важелем, які можуть формувати окремий локальний режим. Методологічною основою дослідження є моделювання методом Монте-Карло з 1000 повторень. Базова модель містить три регресори із заданою кореляційною структурою; один із коефіцієнтів дорівнює нулю. Це дає змогу оцінити не тільки зміщення основного коефіцієнта, а й появу хибного ефекту для змінної, яка за умовами моделі не впливає на результативну ознаку. У статті порівнюються три підходи до оцінювання: звичайний метод найменших квадратів, регресія Губера та метод найменших усічених квадратів. Якість оцінювання аналізується за зміщенням коефіцієнта, середньоквадратичною похибкою, стабільністю знаку, величиною хибного ефекту та частотою появи великих хибних ефектів. Результати моделювання свідчать, що вплив викидів на регресійну модель визначається не лише фактом їх наявності, а й тим, як саме вони виникають. Вертикальні викиди насамперед збільшують варіативність залишків і порівняно добре коригуються регресією Губера. Натомість спостереження з високим важелем потребують окремого тлумачення: якщо вони відповідають базовій регресійній залежності, їх не варто автоматично вважати проблемними. Якщо ж такі спостереження змінюють нахил цієї залежності, вони можуть істотно зміщувати оцінки коефіцієнтів і створювати хибні ефекти для корельованих регресорів. У таких випадках метод найменших усічених квадратів виявляється ефективнішим за регресію Губера. Водночас кластеризовані спостереження з високим важелем залишаються складними для корекції навіть після усікання частини даних, що може вказувати на структурну неоднорідність вибірки. Наукова новизна статті полягає в тому, що різні типи нетипових спостережень порівнюються не лише за їхньою наявністю, а й за характером впливу на економетричне оцінювання. Практична цінність результатів полягає у формуванні орієнтирів для вибору робастного методу. Якщо проблема пов'язана переважно з великими залишками, доцільною є регресія Губера. Якщо ж нетипові спостереження мають високий важіль і змінюють форму регресійної залежності, більш обґрунтованим є застосування методів з усіканням спостережень або проведення додаткового аналізу підгруп і локальних режимів.

Ключові слова: *регресійний висновок; структуровані викиди; спостереження з високим важелем; метод найменших квадратів; регресія Губера; метод найменших усічених квадратів; моделювання методом Монте-Карло; економетричне оцінювання.*
Рис.: *4.* ***Табл.:*** *3.* ***Формул:*** *17.* ***Бібл.:*** *16.*

Манжос Тетяна Василівна *– кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики, Київський національний економічний університет імені Вадима Гетьмана (просп. Берестейський, 54/1, Київ, 03057, Україна)*
E-mail: *tmanzhos@kneu.edu.ua*
ORCID: *https://orcid.org/0000-0003-1393-9116*

Мельник Ольга Олександрівна *– кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики, Київський національний економічний університет імені Вадима Гетьмана (просп. Берестейський, 54/1, Київ, 03057, Україна)*
E-mail: *melnyk.olha@kneu.edu.ua*
ORCID: *https://orcid.org/0000-0002-4399-176X*

Linear regression remains one of the most widely used tools in applied economics. Researchers rely on it not only to summarize statistical associations, but also to estimate economically interpretable relationships and to support substantive conclusions about signs, magnitudes, and comparative effects. This makes the problem of atypical observations especially important. When outliers enter the data, they can affect estimation, weaken interpretation, and in some cases create effects that are not present in the underlying economic mechanism.

At the same time, it is misleading to speak about "outliers" as though they constituted a single and uniform problem. Some observations are unusual only because the dependent variable takes an extreme value. Others are unusual in the regressor space. Still others form small but structured clusters that may reflect a subgroup, a distinct regime, or a different data-generating process. In practice, these cases are often treated too mechanically. Researchers may delete observations, winsorize variables, or switch to a robust estimator without asking what type of contamination is actually present.

In economic research, atypical observations may appear in many different types of data, including macroeconomic, financial, regional, sectoral, market, and socioeconomic datasets. Such observations are not always random errors or purely technical anomalies.

They may reflect crisis episodes, abrupt structural changes, heterogeneity among observed units, concentration of economic indicators, or local regimes within the data. For this reason, the relevant question is not only whether outliers are present, but how their mechanism affects coefficient estimates, statistical inference, and the economic interpretation of a regression model.

The present study is motivated by a simple but practically important idea: the inferential consequences of outliers depend on their structure. A vertical outlier is atypical mainly in the dependent variable. A leverage point is atypical in the regressor space. A good leverage point remains broadly consistent with the main regression relationship, whereas a bad leverage point combines unusual regressor values with a response pattern that contradicts that relationship. A clustered form of bad leverage may indicate not just noise, but a small local regime embedded in the data.

This study examines how different mechanisms through which outliers enter economic data affect coefficient estimation and interpretation in linear regression models. The study focuses on the relationship between the type of atypical observation and the reliability of regression inference. In particular, it compares vertical outliers, non-random vertical outliers, good leverage points, bad leverage points, and clustered bad leverage contamination. Using a Monte Carlo simulation design, the study evaluates how OLS, Huber regression, and a trimming-based least-squares procedure respond to these mechanisms in terms of coefficient bias, RMSE, sign stability, and false-effect detection.

The central research question is: how do different mechanisms of structured outlier contamination affect the reliability of coefficient estimates and their economic interpretation in linear regression models?

LITERATURE REVIEW

Robust regression developed as a response to the well-known sensitivity of least squares estimation. P. Huber's robust-regression framework introduced estimators that reduce the influence of large residuals while retaining least-squares-like behavior near the center of the error distribution [1]. This makes Huber regression particularly relevant for vertical outliers and heavy-tailed disturbances.

A second major line of work focuses on high-breakdown and trimming-based procedures. P. Rousseeuw introduced the least median of squares estimator, replacing the ordinary least-squares objective with a more resistant criterion [2]. The related least trimmed squares idea minimizes the sum of the smallest squared residuals and can resist a substantial fraction of contamination. P. Rousseeuw and A. Leroy systematized robust regression and outlier detection [3], while P. Rousseeuw and K. Van Driessen developed the FAST-LTS algorithm for practical computation [4].

The distinction between outliers and leverage points is fundamental for regression analysis. P. Rousseeuw and B. van Zomeren emphasized that leverage points are observations whose explanatory-variable values are far from the main cloud of regressors [5]. Such observations may be either useful or harmful depending on whether they are consistent with the response relationship. This distinction is important for applied economics because atypical observations may represent meaningful economic situations rather than errors, for example crisis episodes, concentrated market outcomes, exceptional regional units, unusual financial observations, or other structurally important cases.

Later developments in robust regression broadened the available methodological toolkit. MM-estimators combine a high breakdown point with high efficiency under standard conditions [6]. The broader robust-statistics literature frames robustness through influence functions, breakdown points, and efficiency trade-offs [7; 8]. Quantile-based and absolute-deviation methods offer additional alternatives [9], while diagnostic approaches stress the importance of identifying influential observations rather than treating robustness as a purely automatic correction [10–12].

Recent work is especially relevant for the question addressed here. Z. Gao and H. Moon study regression models with potentially endogenous outliers and show that common robust methods such as Huber and LAD may still be biased when outliers are structurally related to the regressors [13]. Applied empirical research also shows that influential observations can materially affect inference and that mechanical procedures such as winsorization or truncation may not fully solve the problem [14]. In applied econometrics, high-breakdown robust regression has also been proposed as a useful but underused tool for empirical work [15]. Recent surveys of robust anomaly detection further emphasize that atypical observations may be either harmful or informative depending on their structure [16].

Against this background, the present study compares a classical benchmark, a residual-robust method, and a trimming-based method across a compact set of structured contamination scenarios. The goal is to identify which contamination mechanisms alter which parts of inference and to translate that mapping into clear recommendations for applied researchers.

RESEARCH AIM AND QUESTIONS

The aim of the study is to evaluate how different mechanisms through which outliers enter the data affect the reliability of coefficient estimates and their economic interpretation in linear regression models.

The study addresses the following research questions:

1. How do vertical and leverage-based outliers differ in their effect on coefficient bias and RMSE?
2. Are good leverage points harmful when they remain consistent with the true regression relationship?
3. Is Huber regression sufficient under structured contamination, or is a trimming-based procedure required?
4. Which contamination mechanisms generate false effects for an irrelevant regressor?

RESEARCH METHODS

Baseline model

The simulation is based on the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i.$$

The true parameter values are fixed as

$$\beta_0 = 1, \beta_1 = 1, \beta_2 = 0.5, \beta_3 = 0.$$

The choice $\beta_3 = 0$ is deliberate. It creates a clean benchmark for studying false-effect detection, that is, cases in which the estimator assigns an economically meaningful coefficient to a regressor that is in truth irrelevant.

The regressor vector is

$$x_i = (x_{1i}, x_{2i}, x_{3i})^{\mathsf{T}},$$

and it is generated from a multivariate normal distribution,

$$x_i \sim N(0, \sum).$$

Instead of assuming a single common pairwise correlation, the covariance matrix is specified as

$$\sum = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix},$$

with baseline values

$$\rho_{12} = 0{,}30, \rho_{13} = 0{,}20, \rho_{23} = 0{,}40.$$

The disturbance term is generated as

$$\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right), \sigma_\varepsilon = 1.$$

Thus, π denotes the contamination rate, ρ_{jk} denotes a pairwise regressor correlation, and σ_ε denotes the error standard deviation.

Simulation parameters

The main reported specification uses $n = 250$ observations and $R = 1000$ Monte Carlo replications. The contamination rates are given by

$$\pi \in \{0, 0.01, 0.05, 0.10\}.$$

For each contamination rate π, the number of contaminated observations is set equal to the integer part of πn. In the main design with $n = 250$, this gives 0, 2, 12, and 25 contaminated observations for $\pi = 0$, 0.01, 0.05, and 0.10, respectively.

Estimation methods

Three estimation methods are compared. First, ordinary least squares serves as the classical benchmark:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{\mathsf{T}}\beta)^2.$$

Second, Huber regression is used as a residual-robust method. Its loss function is denoted by $l_c(u)$ and defined as

$$l_c(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le c, \\ c|u| - \frac{1}{2}c^2, & |u| > c, \end{cases}$$

with tuning constant $c = 1.345$. In words, the Huber loss behaves quadratically for moderate residuals but grows only linearly for large residuals. This preserves efficiency near the center while limiting the influence of extreme residuals.

Third, a trimming-based least-squares procedure is used. It iteratively fits least squares on the h observations with the smallest squared residuals:

$$\hat{\beta}_{trim} = \arg\min_{\beta} \sum_{i=1}^{h} r_{(i)}^2(\beta),$$

where the ordered squared residuals satisfy

$$r_{(1)}^2(\beta) \le r_{(2)}^2(\beta) \le \ldots \le r_{(n)}^2(\beta).$$

In the present computational implementation, this estimator should be interpreted as an iterative trimmed least-squares approximation rather than as a full FAST-LTS routine. The algorithm starts from the OLS estimate, computes squared residuals, keeps $h = 187$ observations with the smallest squared residuals in the main $n = 250$ design, and refits least squares on this retained subset. The procedure is repeated for at most 10 iterations. It stops earlier if both the trimmed residual objective and the coefficient vector change by less than 10^{-6} between two consecutive iterations. The simulations were implemented in Python using NumPy and pandas; figures were generated with Matplotlib.

Contamination scenarios

The simulation considers one clean benchmark and five contamination scenarios. In the formulas in this subsection, a tilde denotes the value assigned to a selected observation after the contamination mechanism has been applied. Thus, $\tilde{x}_i$ and $\tilde{y}_i$ are modified values for contaminated observations only; observations not selected for contamination retain their baseline values.

C0: No contamination. All observations follow the baseline data-generating process.

C1: Random vertical outliers. A random subset of observations receives a shock in the dependent variable:

$$\tilde{y}_i = y_i + a s_y z_i,$$

where s_y is the standard deviation of the clean dependent variable, $z_i \in \{-1,1\}$, and $a = 6$. The regressors remain unchanged.

C2: Non-random vertical outliers. The same vertical shock is applied, but selection into contamination depends on x_{1i}. The probability of selection is proportional to

$$p_i = \frac{\exp(\alpha_1 x_{1i})}{\sum_{j=1}^{n} \exp(\alpha_1 x_{1i})}, \ \alpha_1 = 1.5.$$

This design captures cases in which contamination is structurally linked to the regressor space rather than assigned randomly.

C3: Good leverage points. Selected observations receive shifted regressor values,

$$\tilde{x}_i \sim N(\mu_L, \sum), \ \mu_L = (\lambda, \lambda, \lambda)^{\mathsf{T}}, \ \lambda = 4,$$

while the dependent variable continues to follow the true regression relation:

$$\tilde{y}_i = \beta_0 + \beta_1 \tilde{x}_{1i} + \beta_2 \tilde{x}_{2i} + \beta_3 \tilde{x}_{3i} + \varepsilon_i.$$

C4: Bad leverage points. Selected observations again receive shifted regressor values, but now follow an alternative response rule:

$$\tilde{y}_i = \beta_0 - \beta_1 \tilde{x}_{1i} + \beta_2 \tilde{x}_{2i} + \beta_3 \tilde{x}_{3i} + \varepsilon_i.$$

This reverses the sign of the x_1 effect among contaminated observations.

C5: Clustered bad leverage contamination. A subgroup is generated from a more concentrated local regime using the same location shift $\mu_L = (\lambda, \lambda, \lambda)^{\mathsf{T}}$ as in C3 and C4, but with a smaller covariance matrix:

$$\tilde{x}_i \sim N(\mu_L, 0.25\sum),$$

$$\tilde{y}_i = \beta_0 - \beta_1 \tilde{x}_{1i} + \beta_2 \tilde{x}_{2i} + \beta_3 \tilde{x}_{3i} + \delta + \varepsilon_i.$$

Here, $\delta = 2$. This scenario is intended to mimic a distinct subgroup or local regime rather than isolated noise.

Evaluation criteria and Monte Carlo uncertainty

The first criterion is coefficient bias:

$$Bias(\hat{\beta}_1) = E(\hat{\beta}_1 - \beta_1).$$

The second criterion is root mean squared error:

$$RMSE(\hat{\beta}_1) = \sqrt{E\left[(\hat{\beta}_1 - \beta_1)^2\right]}.$$

The third criterion is sign error:

$$P\left(sign(\hat{\beta}_1) \neq sign(\beta_1)\right).$$

The fourth criterion is false-effect magnitude, defined as the mean absolute value of the estimated coefficient $\hat{\beta}_3$. Since $\beta_3 = 0$, a large false-effect value indicates that an estimator attributes an effect to a regressor that is in truth irrelevant. The fifth criterion is the frequency with which the absolute value of $\hat{\beta}_3$ exceeds 0.10.

Because the simulation uses a finite number of replications, all reported metrics are Monte Carlo estimates. For a generic metric *g*, the Monte Carlo standard error can be estimated as

$$MCSE(\hat{g}) = \frac{s(g_r)}{\sqrt{R}},$$

where g_r is the value of the metric in replication *r*. For proportions, such as sign error or the frequency of large false effects, the approximation is

$$MCSE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{R}}.$$

With $R = 1000$, the maximum Monte Carlo standard error for a proportion is approximately 0.016. In the *Tbl. 1*, RMSE estimates are reported together with their Monte Carlo standard errors in parentheses. Accordingly, the interpretation below focuses on economically and statistically substantial differences rather than on very small numeric gaps.

RESULTS

The main results are reported for $n = 250$, $\pi = 10\%$, and $R = 1000$. The *Tbl. 1* reports the accuracy of estimating the main coefficient β_1: bias, RMSE with Monte Carlo standard errors, and the frequency of sign errors. The *Tbl. 2* reports false-effect diagnostics for β_3, whose true value is zero. All values in the tables and all figures in this section are generated from the same simulation run, which makes the comparison across scenarios internally consistent.

The clean benchmark confirms the expected efficiency ranking in uncontaminated data. OLS and Huber regression have very similar RMSE values, 0.069 and 0.071, respectively, while trimmed LS has a larger RMSE of 0.109. This is not a failure of trimming; rather, it reflects the usual robustness-efficiency trade-off. Trimming becomes useful when contamination is present, but it can be less efficient when the model is clean.

Sign error is essentially zero across all scenarios. This does not mean that the estimators are unaffected by contamination. Instead, it shows that in the present design the main inferential distortion is not a reversal of the coefficient sign, but a distortion of coefficient magnitude and an increase in false-effect detection for the irrelevant regressor x_3.

Table 1

Accuracy of estimating the main coefficient β_1 (n = 250, π = 10%, R = 1000)

Scenario	Method	Bias of β_1	RMSE of β_1 (MCSE)	Sign error
Clean	OLS	–0.002	0.069 (0.002)	0.000
Clean	Huber	–0.002	0.071 (0.002)	0.000
Clean	Trimmed LS	–0.005	0.109 (0.003)	0.000
Random vertical	OLS	–0.007	0.207 (0.005)	0.000
Random vertical	Huber	–0.000	0.081 (0.002)	0.000
Random vertical	Trimmed LS	0.004	0.103 (0.002)	0.000
Non-random vertical	OLS	–0.009	0.214 (0.005)	0.000
Non-random vertical	Huber	–0.006	0.084 (0.002)	0.000
Non-random vertical	Trimmed LS	–0.007	0.109 (0.003)	0.000
Good leverage	OLS	–0.003	0.058 (0.001)	0.000
Good leverage	Huber	–0.003	0.059 (0.001)	0.000
Good leverage	Trimmed LS	–0.005	0.096 (0.002)	0.000
Bad leverage	OLS	–0.698	0.706 (0.003)	0.001
Bad leverage	Huber	–0.622	0.633 (0.004)	0.000
Bad leverage	Trimmed LS	–0.107	0.252 (0.009)	0.003
Clustered bad leverage	OLS	–0.465	0.473 (0.003)	0.000
Clustered bad leverage	Huber	–0.449	0.458 (0.003)	0.000
Clustered bad leverage	Trimmed LS	–0.289	0.369 (0.006)	0.002

Source: compiled by the author based on Monte Carlo simulation results.

Table 2

False-effect diagnostics for the irrelevant coefficient β_3 (n = 250, π = 10%, R = 1000)

Scenario	Method	False-effect magnitude	Large false-effect rate
Clean	OLS	0.058	0.160
Clean	Huber	0.059	0.173
Clean	Trimmed LS	0.091	0.385
Random vertical	OLS	0.170	0.633
Random vertical	Huber	0.068	0.246
Random vertical	Trimmed LS	0.087	0.356
Non-random vertical	OLS	0.175	0.647
Non-random vertical	Huber	0.069	0.249
Non-random vertical	Trimmed LS	0.087	0.379
Good leverage	OLS	0.053	0.135
Good leverage	Huber	0.054	0.148
Good leverage	Trimmed LS	0.086	0.356
Bad leverage	OLS	0.456	1.000
Bad leverage	Huber	0.453	1.000
Bad leverage	Trimmed LS	0.151	0.464
Clustered bad leverage	OLS	0.369	0.999
Clustered bad leverage	Huber	0.367	0.998
Clustered bad leverage	Trimmed LS	0.278	0.787

Source: compiled by the author based on Monte Carlo simulation results.

Note: RMSE values in the Tbl. 1 are reported together with their Monte Carlo standard errors in parentheses. The largest RMSE MCSE in the Tbl. 1 is 0.009. For proportions, the maximum possible MCSE with R = 1000 is approximately 0.016. In the Tbl. 2, false-effect magnitude is the mean absolute value of the estimated β_3 coefficient, and the large false-effect rate is the share of replications in which its absolute value exceeds 0.10.

Vertical contamination

A first clear pattern emerges in the two vertical-contamination scenarios. Under random vertical contamination, OLS RMSE rises sharply to 0.207, whereas Huber regression reduces it to 0.081. Under non-random vertical contamination, the same ranking remains visible: OLS RMSE equals 0.214, while Huber regression yields 0.084. Trimmed LS also improves on OLS in both cases, but it is not as efficient as Huber.

These numbers are substantively coherent. Vertical outliers primarily inflate residual dispersion, and Huber regression is designed exactly for that setting because it directly downweights unusually large residuals. The comparison between random and non-random vertical contamination is also instructive. Once contamination becomes related to the regressor space, the problem becomes more structured; nevertheless, residual robustness still performs well in the present design. The RMSE comparison in *Fig. 1* places these findings next to the leverage scenarios and shows that residual robustness is most efficient when the anomaly is mainly vertical, whereas leverage-based contamination requires a different response.

Bad leverage contamination

The most damaging scenario in the simulation is bad leverage contamination. Here OLS bias in $\hat{\beta}_1$ reaches –0.698, and RMSE increases to 0.706. Huber regression helps only slightly, lowering RMSE to 0.633. By contrast, trimmed LS reduces RMSE to 0.252, which is still worse than in the clean benchmark but clearly preferable to the other two methods. The bias comparison in *Fig. 2* shows why this case is qualitatively different from vertical contamination: the problem is not only higher noise, but a systematic pull on the estimated slope.

The rate-based comparison in *Fig.* 3 reinforces this interpretation. As the share of bad leverage observations increases, OLS and Huber deteriorate rapidly, while trimmed LS remains substantially more stable.

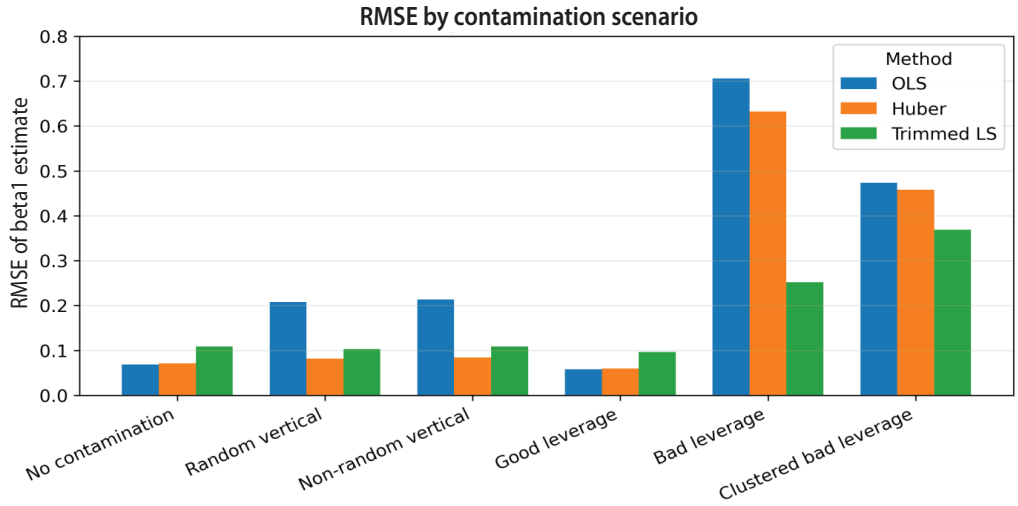

Fig. 1. RMSE of $\hat{\beta}_1$ across contamination scenarios (n = 250, π = 10%)

Source: compiled by the authors based on Monte Carlo simulation results.

Good leverage points

The results for good leverage contamination are equally important, even though they are less dramatic. Under good leverage contamination, OLS RMSE is 0.058, which is slightly below the clean-data benchmark, while Huber regression performs almost identically at 0.059. Trimmed LS is noticeably less efficient, with RMSE 0.096.

The main lesson is conceptual rather than numerical: leverage alone is not necessarily harmful. Observations can be unusual in the regressor space and still provide valid information about the underlying economic relationship. In applied work, such observations may correspond to extreme but substantively meaningful economic cases rather than mistakes in the data. Removing them mechanically simply because they are far away in the regressor space can therefore be a mistake.

The same mechanism appears in the false-effect metric. Since $\beta_3 = 0$, a large absolute value of indicates that the estimator is generating spurious evidence for an irrelevant regressor. Under bad leverage contamination, false-effect magnitude is 0.456 for OLS and 0.453 for Huber, while trimmed LS reduces it to 0.151. This false effect does not arise from nothing. It is partly transmitted through the correlation structure of the regressors. Given $\rho_{13} = 0.20$ and $\rho_{23} = 0.40$ in the baseline covariance matrix, distortions in and can propagate to through regressor correlation. This pattern is summarized visually in *Fig. 4.*

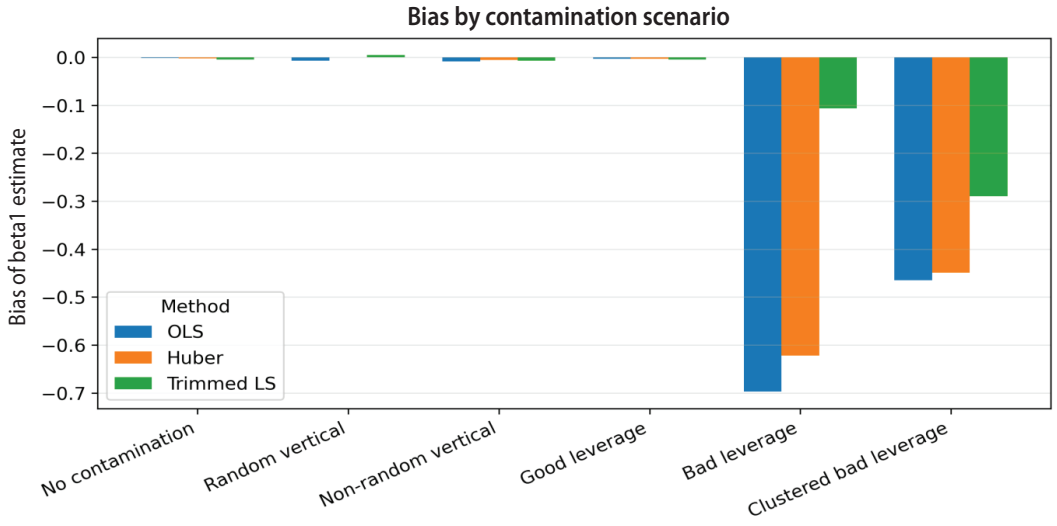

Fig. 2. Bias of $\hat{\beta}_1$ across contamination scenarios (n = 250, π = 10%)

Source: compiled by the authors based on Monte Carlo simulation results.

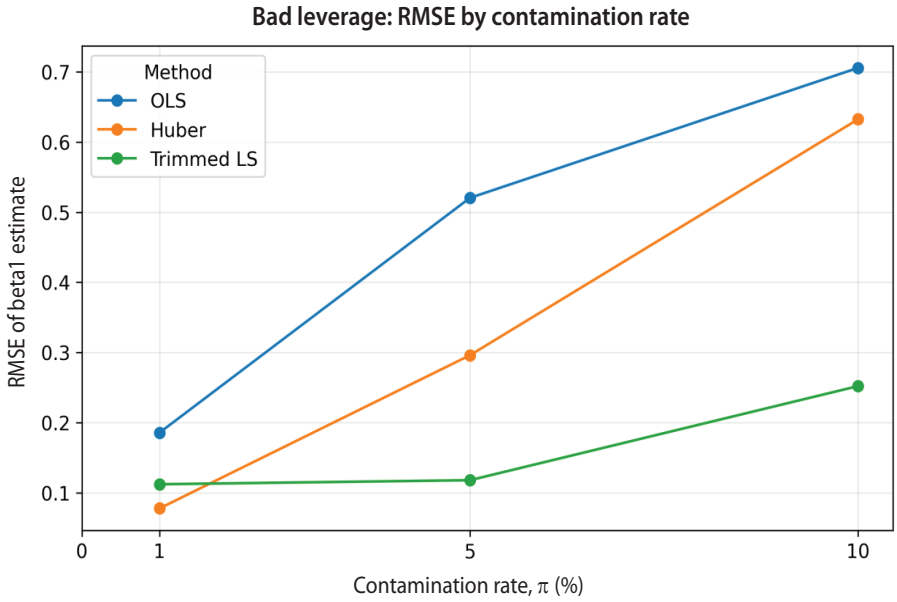

Fig. 3. RMSE of $\hat{\beta}_1$ under bad leverage contamination as the contamination rate increases

Source: compiled by the authors based on Monte Carlo simulation results.

Clustered bad leverage contamination

Clustered bad leverage contamination is especially revealing because it combines contamination with local structure. OLS and Huber again perform poorly, with RMSE values of 0.473 and 0.458, respectively. Trimmed LS improves the result, but only partially: its RMSE remains 0.369, the false-effect magnitude is 0.278, and the large false-effect rate is still 0.787.

This pattern suggests that clustered contamination should not be interpreted merely as a collection of isolated outliers. A coherent cluster may represent a subgroup, a crisis episode, a local institutional regime, or some other structurally distinct mechanism. Robust estimation helps, but it does not fully solve the interpretive problem. Once contamination has this grouped structure, the researcher may need to think not only about estimator choice, but also about specification choice.

DISCUSSION

The simulation results support a mechanism-based view of robustness. The same contamination rate can produce very different inferential consequences depending on how contamination enters the data-generating process. For that reason, the blanket recommendation to "use a robust estimator" is often too vague to be genuinely helpful in applied work.

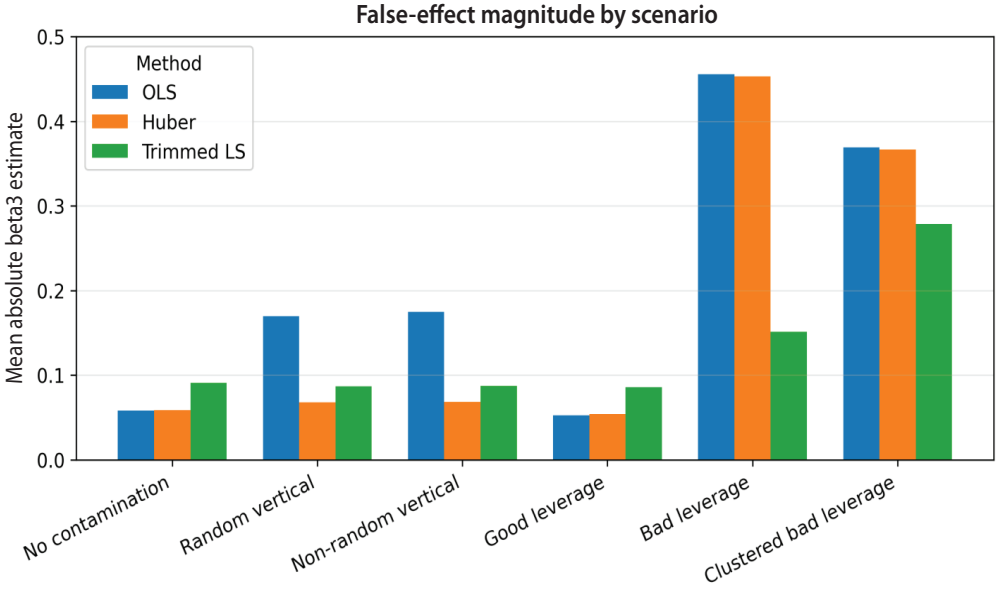

Fig. 4. False-effect magnitude for $\hat{\beta}_3$ across contamination scenarios (n = 250, π = 10%)

Source: compiled by the authors based on Monte Carlo simulation results.

When the problem is primarily vertical, Huber regression performs well because it attacks the relevant source of distortion directly. Large residuals receive less weight, while the central part of the sample continues to contribute almost as in least squares. In practice, this means that if diagnostics suggest isolated or moderately structured vertical contamination, a residual-robust procedure may be sufficient without more aggressive trimming.

The interpretation is different for good leverage points. Here the simulation shows that unusual observations in the regressor space are not automatically a problem. Indeed, they may enrich the support of the data and improve identification of the regression slope. This is especially relevant in economics, where some observations may be extreme because they capture real economic heterogeneity, structural change, crisis conditions, market concentration, or local regimes. The practical implication is straightforward: distance in the regressor space should not by itself justify deletion.

Bad leverage contamination is much more dangerous because the problem is no longer the size of the residual alone. Instead, the unusual observations actively pull the fitted relationship away from the dominant pattern in the data. In that setting, Huber regression only partially helps, since it remains centered on residual size rather than on the structural position of the observation. Trimming-based robustness performs better because it is capable of anchoring the fit on the majority structure of the sample.

The clustered bad leverage case is perhaps the most important applied warning. Once contamination forms a coherent subgroup, it may reflect something economically meaningful: a crisis regime, a sector-specific mechanism, a measurement regime, or an institutional split in the population. Robust regression can reduce the distortion, but it cannot by itself explain why the subgroup exists. That is why a robust estimator should sometimes be seen not as the final solution, but as a diagnostic signal that the model specification may need to be reconsidered.

Taken together, the results suggest the following practical logic. *First,* identify whether the anomaly is mainly vertical or leverage-based. *Second,* distinguish between leverage that is compatible with the dominant pattern and leverage that contradicts it. *Third,* if the contamination appears clustered, move beyond purely mechanical robustness and consider subgroup analysis, interaction terms, or alternative regime-specific specifications.

In applied work, the choice of robust method should be informed by diagnostics rather than by a mechanical rule. Vertical contamination can be examined through standardized residuals and residual Q-Q plots after a preliminary OLS fit. Leverage points can be identified using hat values, for example values exceeding $2(k+1)/n$, or through robust Mahalanobis distances computed on the regressor vector. Good and bad leverage observations can then be distinguished by combining leverage diagnostics with residual diagnostics: high leverage with small residuals is closer to good leverage, whereas high leverage with large residuals is more consistent with bad leverage. Clustered contamination requires additional checks, such

as scatterplots of robust residuals against fitted values, inspection of sectoral or regional groups, or formal subgroup tests. The practical classification that follows from this diagnostic logic is summarized in the *Tbl. 3*.

the issue may reflect structural heterogeneity rather than mere noise. In that setting, robust estimation should be combined with subgroup analysis or alternative specifications.

Table 3

Practical classification of contamination mechanisms

Contamination type	Main problem	Practical recommendation
Random vertical	Large residual shocks in y	Huber regression is often sufficient
Non-random vertical	Residual contamination linked to regressors	Use robust regression and diagnose the selection mechanism
Good leverage	Unusual x-values but correct relationship	Do not remove automatically
Bad leverage	Coefficient distortion and false effects	Prefer trimming-based or high-breakdown robust methods
Clustered bad leverage	Possible subgroup or alternative regime	Combine robust regression with subgroup or regime analysis

Source: compiled by the authors based on Monte Carlo simulation results.

CONCLUSIONS

This study examined how structured outlier contamination impacts regression inference in linear models. The central conclusion is that outliers should not be treated as a homogeneous problem. Different contamination mechanisms distort different parts of inference and therefore call for different modeling responses.

The *main finding*s can be summarized as follows:

1. **Vertical contamination is mainly a residual problem.** In both random and non-random vertical scenarios, Huber regression substantially reduces RMSE relative to OLS and keeps false-effect magnitude relatively low.
2. **Good leverage is not inherently harmful.** Observations can be extreme in the regressor space and still remain informative. In the simulation, OLS performs very well under good leverage contamination, which cautions against mechanical deletion of unusual observations.
3. **Bad leverage contamination is much more destructive than vertical contamination.** Once unusual observations also contradict the underlying relationship, OLS becomes strongly biased and can create spurious effects in irrelevant regressors.
4. **Trimming-based robustness is preferable under leverage-based contamination.** The trimmed LS procedure clearly outperforms both OLS and Huber under bad leverage and improves results under clustered bad leverage as well, although the improvement is incomplete in the clustered case.
5. **Clustered contamination should often be interpreted substantively.** When problematic observations form a coherent subgroup,

The study also has several limitations. *First,* the main reported design focuses on $n = 250$, although the broader simulation framework can be extended to additional sample sizes. *Second,* the trimming-based estimator used here is an iterative approximation rather than a full FAST-LTS implementation. *Third,* the analysis is based on simulated cross-sectional linear models and therefore does not address panel structure, time dependence, endogenous regressors, or empirical datasets with real institutional complexity. *Fourth,* the contamination parameters $a = 6$, $\lambda = 4$, $\alpha_1 = 1.5$, and $\delta = 2$ are chosen to make the mechanisms visible in a controlled design; sensitivity to these parameters is left for future work. Fifth, the bad leverage scenario is implemented through a sign reversal in the effect of x_1. Other forms of bad leverage, such as intercept shifts or simultaneous changes in several coefficients, may generate different numerical patterns.

These limitations point directly to future research. A natural extension would be to compare the current trimming approach with standard FAST-LTS and MM-estimators in the same contamination map. Another useful step would be to study panel-data settings or regime-switching environments, where clustered contamination may be even more relevant. Finally, the simulation evidence could be connected to applied case studies in economics, where the distinction between good leverage, bad leverage, and subgroup structure often has immediate substantive meaning.

Even with these limitations, the paper offers a practically useful conclusion: robustness should be chosen with respect to the mechanism of contamination, not treated as a one-size-fits-all adjustment. The main contribution is therefore a clearer classification of when residual robustness is enough, when trimming-

based robustness is preferable, and when the data may be signaling deeper structural heterogeneity. ■

BIBLIOGRAPHY

1. Huber P. J. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*. 1973. Vol. 1. Iss. 5. P. 799–821. DOI: https://doi.org/10.1214/aos/1176342503
2. Rousseeuw P. J. Least median of squares regression. *Journal of the American Statistical Association*. 1984. Vol. 79. Iss. 388. P. 871–880. DOI: https://doi.org/10.1080/01621459.1984.10477105
3. Rousseeuw P. J., Leroy A. M. Robust Regression and Outlier Detection. New York : Wiley, 1987. DOI: https://doi.org/10.1002/0471725382
4. Rousseeuw P. J., Van Driessen K. Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*. 2006. Vol. 12. P. 29–45. DOI: https://doi.org/10.1007/s10618-005-0024-4
5. Rousseeuw P. J., van Zomeren B. C. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*. 1990. Vol. 85. Iss. 411. P. 633–639. DOI: https://doi.org/10.1080/01621459.1990.10474920
6. Yohai V. J. High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*. 1987. Vol. 15. Iss. 2. P. 642–656. DOI: https://doi.org/10.1214/aos/1176350366
7. Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A. Robust Statistics: The Approach Based on Influence Functions. New York : Wiley, 1986. DOI: https://doi.org/10.1002/9781118186435
8. Maronna R. A., Martin R. D., Yohai V. J. Robust Statistics: Theory and Methods. Chichester : Wiley, 2006. DOI: https://doi.org/10.1002/0470010940
9. Koenker R., Bassett G. Regression quantiles. *Econometrica*. 1978. Vol. 46. Iss. 1. P. 33–50. DOI: https://doi.org/10.2307/1913643
10. Cook R. D. Detection of influential observation in linear regression. *Technometrics*. 1977. Vol. 19. Iss. 1. P. 15–18. DOI: https://doi.org/10.1080/00401706.1977.10489493
11. Belsley D. A., Kuh E., Welsch R. E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. New York : Wiley, 1980. DOI: https://doi.org/10.1002/0471725153
12. Atkinson A. C., Riani M. Robust Diagnostic Regression Analysis. New York : Springer, 2000. DOI: https://doi.org/10.1007/978-1-4612-1160-0
13. Gao Z., Moon H. R. Robust estimation of regression models with potentially endogenous outliers via a modern optimization lens. *arXiv preprint*. 2024. DOI: https://doi.org/10.48550/arXiv.2408.03930
14. Leone A. J., Minutti-Meza M., Wasley C. E. Influential observations and inference in accounting research. *The Accounting Review*. 2019. Vol. 94. Iss. 6. P. 337–364. DOI: https://doi.org/10.2308/accr-52396
15. Zaman A., Rousseeuw P. J., Orhan M. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*. 2001. Vol. 71. Iss. 1. P. 1–8. DOI: https://doi.org/10.1016/S0165-1765(00)00404-3
16. Rousseeuw P. J., Hubert M. Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*. 2018. Vol. 8. Iss. 2. Art. e1236. DOI: https://doi.org/10.1002/widm.1236

REFERENCES

Atkinson A. C. & Riani M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer. https://doi.org/10.1007/978-1-4612-1160-0

Belsley D. A., Kuh E. & Welsch R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley. https://doi.org/10.1002/0471725153

Cook R. D. (1977). Detection of influential observation in linear regression. *Technometrics, 1*(19), 15–18. https://doi.org/10.1080/00401706.1977.10489493

Gao Z. & Moon H. R. (2024). Robust estimation of regression models with potentially endogenous outliers via a modern optimization lens. *arXiv preprint*. https://doi.org/10.48550/arXiv.2408.03930

Hampel F. R., Ronchetti E. M., Rousseeuw P. J. & Stahel W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley. https://doi.org/10.1002/9781118186435

Huber P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics, 5*(1), 799–821. https://doi.org/10.1214/aos/1176342503

Koenker R. & Bassett G. (1978). Regression quantiles. *Econometrica, 1*(46), 33–50. https://doi.org/10.2307/1913643

Leone A. J., Minutti-Meza M. & Wasley C. E. (2019). Influential observations and inference in accounting research. *The Accounting Review, 6*(94), 337–364. https://doi.org/10.2308/accr-52396

Maronna R. A., Martin R. D. & Yohai V. J. (2006). *Robust Statistics: Theory and Methods*. Chichester: Wiley. https://doi.org/10.1002/0470010940

Rousseeuw P. J. & Hubert M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery, 2*(8), e1236. https://doi.org/10.1002/widm.1236

Rousseeuw P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 388*(79), 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw P. J. & Leroy A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley. https://doi.org/10.1002/0471725382

Rousseeuw P. J. & van Zomeren B. C. (1990). Unmasking multivariate outliers and leverage points. *Jour-*

nal of the American Statistical Association, 411(85), 633–639. https://doi.org/10.1080/01621459.1990.10474920
Rousseeuw P. J. & Van Driessen K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery, 12*, 29–45. https://doi.org/10.1007/s10618-005-0024-4
Yohai V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics, 2*(15), 642–656. https://doi.org/10.1214/aos/1176350366
Zaman A., Rousseeuw P. J. & Orhan M. (2001). Econometric applications of high-breakdown robust regression techniques. *Economics Letters, 1*(71), 1–8. https://doi.org/10.1016/S0165-1765(00)00404-3

Дослідження не отримувало зовнішнього фінансування.
Автори заявляють про відсутність конфлікту інтересів.
Інструменти генеративного ШІ використовувалися для редагування мови, допомоги у форматуванні та технічній підтримці під час підготовки рукопису. Автори несуть повну відповідальність за науковий зміст, дизайн моделювання, інтерпретацію результатів та фінальний текст статті.

Стаття надійшла до редакції / Received: 08.06.2026
Статтю прийнято до публікації / Accepted: 21.06.2026
Оприлюднено / Published: 30.07.2026

УДК 330.16:336.76:004.9
JEL: D81; G11; G41; O33
DOI: https://doi.org/10.32983/2222-4459-2026-6-244-253

МЕТОДОЛОГІЧНІ ЗАСАДИ ФОРМУВАННЯ ПОВЕДІНКОВОГО ПРОФІЛЮ ІНВЕСТОРА В УМОВАХ ЦИФРОВІЗАЦІЇ ТА КРИЗОВОЇ НЕВИЗНАЧЕНОСТІ

©2026 **ПРОКОПЕНКО М. В.**

УДК 330.16:336.76:004.9
JEL: D81; G11; G41; O33

Прокопенко М. В. Методологічні засади формування поведінкового профілю інвестора в умовах цифровізації та кризової невизначеності

У статті розв'язується наукова проблема недостатньої інтеграції поведінкових, цифрових і кризових чинників у традиційні підходи до профілювання інвестора. Метою дослідження є обґрунтування концептуально-методологічного підходу до формування поведінкового профілю інвестора шляхом систематизації ключових поведінкових упереджень, визначення їх проявів у цифровому фінансовому середовищі та врахування кризової невизначеності як контекстного чинника інвестиційної поведінки. Методологічну основу становлять положення поведінкових фінансів, теорії перспектив, концепції обмеженої раціональності та підходи до оцінювання ризик-толерантності. Інформаційною базою є наукові праці з портфельної теорії, поведінкових фінансів, соціального трейдингу, цифровізації інвестиційного процесу та інформаційного перевантаження. У роботі використано методи кабінетного дослідження, теоретичного узагальнення, порівняльного аналізу, класифікації, систематизації, логіко-структурного моделювання та концептуального синтезу. У результаті систематизовано п'ять груп поведінкових параметрів: упередження, пов'язані з оцінюванням ризику та втрат; інформаційно-когнітивні упередження; соціальні упередження; упередження самосприйняття; фінансово-поведінкові ефекти. Запропоновано концептуальну матрицю, яка поєднує вид упередження, його групову належність, прояв у інвестиційній поведінці та можливий індикатор вимірювання. Обґрунтовано, що цифрове середовище, соціальний трейдинг, інформаційне перевантаження та кризова невизначеність мають розглядатися як активні чинники трансформації поведінкового профілю, а не лише як зовнішні умови прийняття рішень. Наукова новизна полягає у формуванні узагальненої моделі поведінкового профілю інвестора як динамічної системи взаємопов'язаних індивідуальних, соціальних, фінансово-поведінкових і контекстних компонентів. Практична цінність результатів полягає в можливості використання запропонованого підходу для розроблення опитувальників, інструментів інвесторської діагностики та подальших економіко-математичних моделей інвестиційної поведінки.

Ключові слова: *поведінкові фінанси; поведінковий профіль інвестора; інвестиційні рішення; поведінкові упередження; цифровізація; кризова невизначеність; ризик; соціальний трейдинг.*
Табл.: *1.* ***Формул:*** *1.* ***Бібл.:*** *28.*

Прокопенко Матвій Вадимович *– аспірант, Київський національний університет імені Тараса Шевченка (вул. Володимирська, 60, Київ, 01033, Україна)*
E-mail: *matvii.prokopenko@gmail.com*
ORCID: *https://orcid.org/0000-0003-1457-8762*

UDC 330.16:336.76:004.9
JEL: D81; G11; G41; O33

Prokopenko M. V. The Methodological Principles of Forming an Investor Behavioral Profile under Digitalization and Crisis Uncertainty

The article addresses the research gap related to the insufficient integration of behavioral, digital, and crisis-related factors into traditional approaches to investor profiling. The aim of the study is to substantiate a conceptual and methodological approach to forming an investor behavioral profile by systematizing key behavioral biases, identifying their manifestations in the digital financial environment, and considering crisis uncertainty as a contextual factor of investment behavior. The methodological basis of the research combines behavioral finance, prospect theory, bounded rationality, and approaches to financial risk tolerance assessment. The information base consists of scientific publications on portfolio theory, behavioral finance, social trading, digitalization of investment processes, and information overload. The study applies desk research, theoretical generalization, comparative analysis, classification, systematization, logical/structural modeling, and conceptual synthesis. The results systematize five groups of behavioral parameters: biases related to the assessment of risk and losses;