Testing Most Influential Sets

Lucas D. Konrad [1] Nikolas Kuschnig [2]

[1]Vienna University of Business and Economics
[2]Monash University

April 27, 2026
ICLR 2026

Motivation

1. **Scientific discovery:** Rugged terrain generally hinders economic development, but not in Africa. What if this is driven by just two small island nations?
2. **Fairness auditing:** An algorithmic decision-making system produces different outcomes for a protected group. What if the disparity can be explained by only a handful of data points?
3. **Data cleaning:** A single influential point among a thousand samples flips a strong correlation to a null result. Should we trust the original finding or the one without the outlier?
4. **Data preprocessing:** A microcredit experiment shows negligible outcome variations overall, except for a few outliers. How should we prepare and analyze the sample?

Introduction

Inferences may be sensitive to small **most influential sets**.[1] These sets are:

- intuitive to interpret
- directly tied to the quantity of interest, and
- highlight support of estimates in the data.

[1] **cook1982**; **belsley1980**; **chatterjee1986**; **hadi1993**; BGM23; KZC21; FH23; RH25; Hu+24.

Introduction

Inferences may be sensitive to small **most influential sets**.[1] These sets are:

- intuitive to interpret
- directly tied to the quantity of interest, and
- highlight support of estimates in the data.

Research Gap

It's *unclear how to interpret and deal with* most influential sets — there is no statistical framework for judging influence.

Research Question

How influential should we expect a given set to be?

[1] **cook1982**; **belsley1980**; **chatterjee1986**; **hadi1993**; BGM23; KZC21; FH23; RH25; Hu+24.

Illustration — influential sets

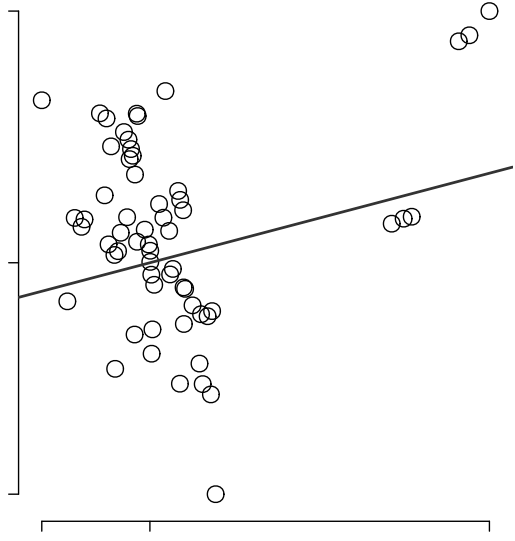


Figure: Two influential sets in a univariate regression.

Illustration — influential sets

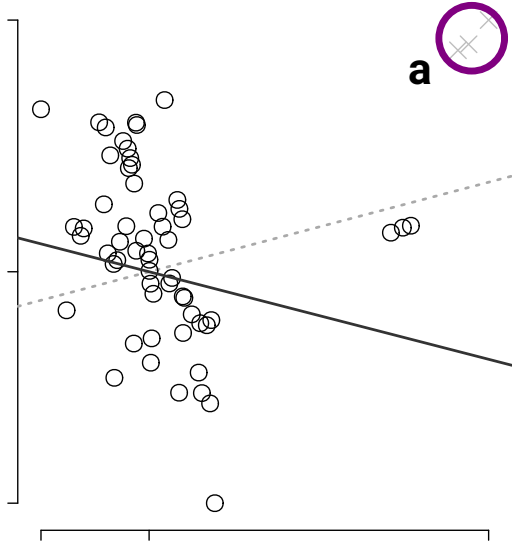


Figure: Two influential sets in a univariate regression.

Illustration — influential sets

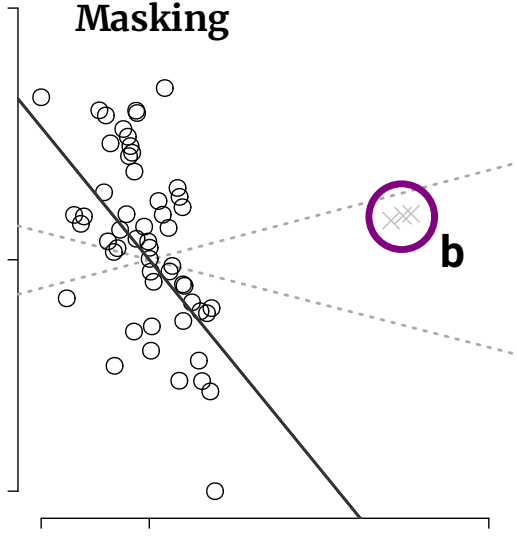


Figure: Two influential sets in a univariate regression.

Overview

We develop a **statistical framework** to evaluate influential sets, focusing on *linear regression coefficients*:

- Distribution of maximum influence
- Efficient implementation for formal tests
- Empirical validation across domains

We (1) provide the first rigorous theoretical results to interpret influence, and (2) demonstrate their practical utility by resolving contested findings.

Setup — Influence

- Let $\mathcal{S} \subset \{1, \ldots, N\}$ be an index-set of size $|\mathcal{S}| = k$
- Use a subscript $\hat{\theta}_{-\mathcal{S}}$ to denote a quantity θ without $\mathcal{S}$

Definition (Influence)

$$\Delta\left(\mathcal{S}; \phi\right) = \phi(\hat{\theta}) - \phi(\hat{\theta}_{-\mathcal{S}})$$

is the *influence* of subset $\mathcal{S}$ on target function $\phi : \mathbb{R}^Q \mapsto \mathbb{R}$.

Definition (Most Influential Set)

For a positive integer $k \ll N$, the k-*most influential set* is

$$\mathcal{S}_k^{\max} := \underset{\mathcal{S} \subset [N], |\mathcal{S}| \leqslant k}{\arg\max}\, \Delta\left(\mathcal{S}; \phi\right),$$

We denote the maximum influence as $\Delta^{\max} = \Delta\left(\mathcal{S}_k^{\max}; \phi\right)$.

Setting — Linear Least Squares

Influence Functions

We do not work with generally applicable *influence functions*, as they systematically underestimate the impact of (a) *sets* of data points and (b) highly *influential* data points.[2]

Instead, we use **exact influence** in a tractable setting: LS.

$$\hat{\theta} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda\right)^{-1} \mathbf{X}'\mathbf{y},$$

yielding predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ and residuals $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$.

[2] BYF20; Hu+24; Hua+25; KL17.

Influence Distribution — Illustration

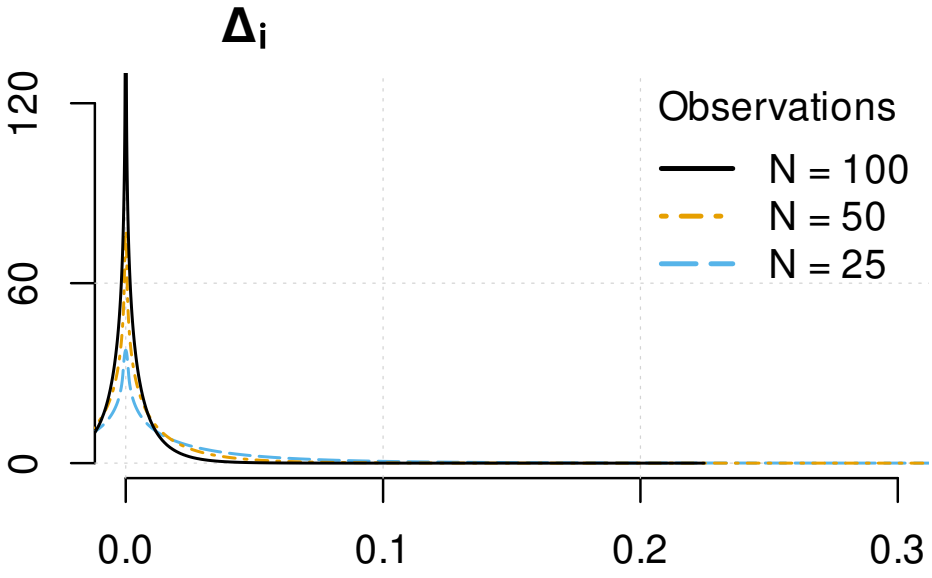


Figure: Probability density of single-observation Δ_i.

Influence Distribution for Sets — Illustration

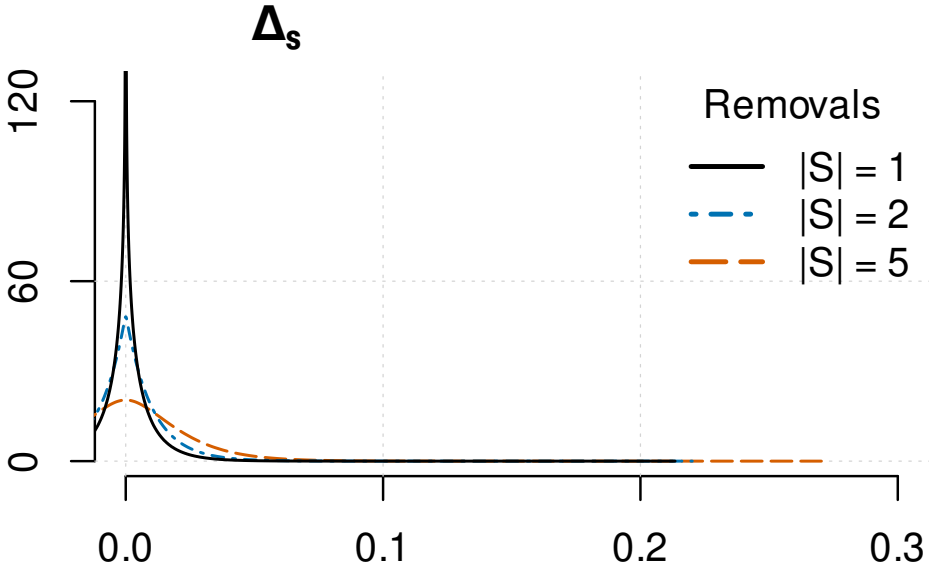


Figure: Probability density of Δ_S for increasing set sizes.

Influence Distribution

For OLS, the **influence of observation** i is well-known [belsley1980]:

$$\Delta_i(\hat{\beta}) = \frac{X_i \cdot R_i}{\sum_n^N X_n^2} (1 - H_i)^2 ,$$

where R_i is the residual and H_i the leverage.

Proposition

The influence of some set $\mathcal{S}$ on the least-squares estimator $\hat{\theta}$ is

$$\Delta(\mathcal{S}) = \left(\mathbf{X}'_{-\mathcal{S}}\mathbf{X}_{-\mathcal{S}} + \lambda \mathbf{I}_P\right)^{-1} \mathbf{X}'_{\mathcal{S}} r_{\mathcal{S}}, \tag{1}$$

where $\lambda \geqslant 0$ is an optional penalization parameter.

- We can work out distributions, e.g., for $X, R \sim_{iid} \mathcal{N}(0, 1)$:

$$\Delta_i \xrightarrow{d} \frac{\chi_1^2 - \chi_1^2}{2\,\chi_{N-1}^2},$$

Addressing Set Selection

- Conveniently, as set size k increases $\Delta_{\mathcal{S}} \to \mathcal{N}(0, \sigma_k)$
- But the set of interest $\mathcal{S}$ is *unlikely to be a random sample*
- Consider the most influential observation in seven microcredit RCTs [Mea19]

Addressing Set Selection

- Conveniently, as set size k increases $\Delta_{\mathcal{S}} \to \mathcal{N}(0, \sigma_k)$
- But the set of interest $\mathcal{S}$ is *unlikely to be a random sample*
- Consider the most influential observation in seven microcredit RCTs [Mea19]

	BIH	MON	ETH	MEX	MOR	PHI	IND
N	1,195	961	3,113	16,560	5,498	1,113	6,863
p_0	.000260	.001700	.000104	.000004	.000036	.002310	.000046
A	192	29	480	12,499	1388	21	1086

Table: Sample sizes, influence p-values (under normality), and the hypothetical number of hypothesis tests A for which naive adjustment would still yield $p \leqslant .05$ under H_0.

Addressing Set Selection

- Conveniently, as set size k increases $\Delta_{\mathcal{S}} \to \mathcal{N}(0, \sigma_k)$
- But the set of interest $\mathcal{S}$ is *unlikely to be a random sample*
- Consider the most influential observation in seven microcredit RCTs [Mea19]

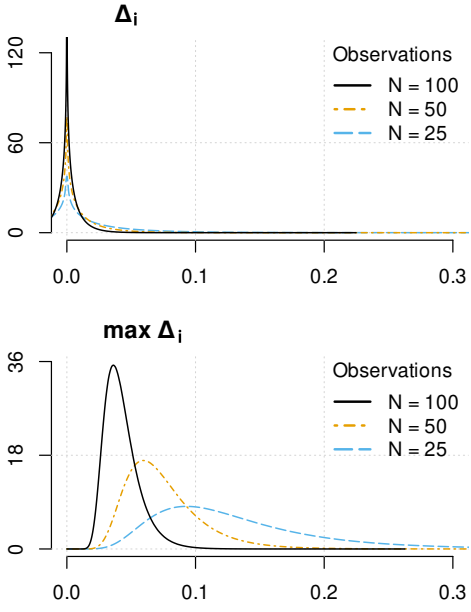
	BIH	MON	ETH	MEX	MOR	PHI	IND
N	1,195	961	3,113	16,560	5,498	1,113	6,863
p_0	.000260	.001700	.000104	.000004	.000036	.002310	.000046
A	192	29	480	12,499	1388	21	1086

Table: Sample sizes, influence p-values (under normality), and the hypothetical number of hypothesis tests A for which naive adjustment would still yield $p \leqslant .05$ under H_0.

Extreme Value Theory

Our goal is to characterize $\Delta^{\max}$ — defined by a maximal operation its distribution is governed by extreme value theory rather than classical asymptotics.

Influence Extreme Value Distribution — Illustration



Influence Extreme Value Distribution

Target

We seek the limiting EVD H such that $\Delta^{\max} \in \mathrm{MDA}(H)$, i.e., $\Delta^{\max}$ lies in the maximum domain of attraction of H.

It is not clear that the Generalized Extreme Value Distribution applies!

- Standard EVT requires $i.i.d$ observations
- **Sets create dependence between observations!**

We distinguish two regimes [BGM23; KZC21] based on the subset size k:

1. **Constant-size sets:** k remains fixed as $N \to \infty$.
2. **Relative-size sets:** k grows proportionally with N, i.e., $k = pN$ for some $p \in (0, 1)$.

Result 1

Theorem (EVD for constant-size sets)

Suppose $\mathbb{E}[X^2] < \infty$, *and that the thicker upper tail of* X_i, R_i *is polynomial with coefficients* $\xi_x, \xi_r < \infty$. *If* $|\mathcal{S}_k^{\max}|$ *remains constant as* $N \to \infty$, *then*

$$\lim_{N\to\infty} \Delta^{\max} \sim \textit{Fréchet}(a, b, \xi),$$

with location parameter a, *scale parameter* b, *and shape parameter* $\xi = \min\{\xi_x, \xi_r\}$.

Corollary

If the tail coefficients of both X_i *and* R_i *are infinite, then*

$$\lim_{N\to\infty} \Delta^{\max} \sim \textit{Gumbel}(a, b).$$

Result 2

Theorem (EVD for relative-size sets)

If $\{X_n R_n\}_{n=1}^{N}$ *satisfies the conditions of a CLT and* $|\mathcal{S}_k^{\max}|$ *grows proportionally with* N*, then*

$$\lim_{N\to\infty} \Delta^{\max} \sim \textit{Gumbel}(a, b).$$

- Constant-size sets are dominated by the heaviest tail, and especially small sets can exert extreme influence.
- For growing sets, $\Delta^{\max}$ converges to a well-behaved Gumbel distribution —

Practical Implementation

With a most influential set of interest, we can:

1. **Determine the EVD family** based on the tail behavior of X, R
2. **Estimate EVD parameters** using bias-corrected block maxima MLE
3. **Perform hypothesis test** that the observed influence reflects natural sampling variation, against the alternative H_1 of excessive influence.

This procedure is useful in practice:

- Straightforward in terms of computation
- Simulation results show rapid convergence to our theoretical predictions — results are applicable with small samples
- EVD parameter estimation works reasonably well

Practical Illustration

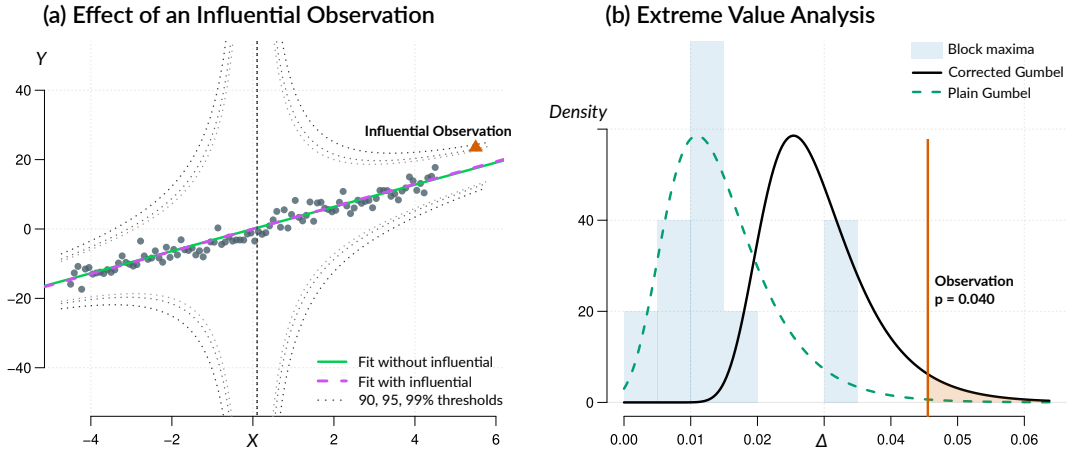


Figure: Panel A depicts observations, estimated regression lines, and conditional significance regions at the 10, 5, and 1% levels (dotted lines). Panel B illustrates the extreme value analysis: a histogram of block maxima in the background, and fitted Gumbel distributions with (solid) and without (dashed) bias correction.

Application to Economic Development and Geography

- We resolve the controversial finding that **rugged terrain benefits African economies** compared to the rest of the world [NP12].
- The Seychelles are **excessively influential**, both individually and in combination with other outliers [flagged by KZC21].

Table: Influence of Ruggedness on log(GDP per capita in 2000)

Influential Set	$\Delta(\mathcal{S})$	$\hat{\tilde{a}}$	$\hat{b}$	p-value
Seychelles	0.077	0.020	0.004	$< 1e^{-16}$
Seychelles + Rwanda	0.070	0.028	0.006	0.001
Seychelles + Eswatini	0.077	0.020	0.004	$< 1e^{-16}$
Seychelles + Comoros	0.061	0.028	0.006	0.004

More applications (big-headed sparrows etc.) are in the paper.

Conclusion

- We provide a rigorous framework to test **whether most influential sets are excessively influential** or just *natural sampling variation*.
- This transforms *ad-hoc sensitivity checks* into **formal hypothesis tests**
- We can determine when most influential sets should **genuinely overturn results** and when they naturally reflect information.
- This enables more **robust and transparent decision-making** where reliability matters, from medical trials to policy evaluation.

Future work remains on identifying most influential sets, characterizing finite sample behavior, and EV parameter estimation.

References I

[BGM23] T. Broderick, R. Giordano, and R. Meager. *An Automatic Finite-Sample Robustness Metric: When Can Dropping a Little Data Make a Big Difference?* arXiv:2011.14999 [stat.ME]. 2023 (cit. on pp. 3, 4, 18).

[BYF20] S. Basu, X. You, and S. Feizi. "On Second-Order Group Influence Functions for Black-Box Predictions". In: *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, pp. 1139–1149 (cit. on p. 10).

[FH23] D. Freund and S. B. Hopkins. "Towards Practical Robustness Auditing for Linear Regression". In: *arXiv preprint* (2023). DOI: 10.48550/arXiv.2307.16315 (cit. on pp. 3, 4).

[Hu+24] Y. Hu et al. "Most Influential Subset Selection: Challenges, Promises, and Beyond". In: *Advances in Neural Information Processing Systems*. NeurIPS. 2024. arXiv: 2409.18153 (cit. on pp. 3, 4, 10).

[Hua+25] J. Y. Huang et al. *Approximations to Worst-Case Data Dropping: Unmasking Failure Modes*. arXiv:2408.09008 [stat.ME]. 2025 (cit. on p. 10).

References II

[KL17] P. W. Koh and P. Liang. "Understanding Black-Box Predictions via Influence Functions". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pp. 1885–1894 (cit. on p. 10).

[KZC21] N. Kuschnig, G. Zens, and J. Crespo Cuaresma. *Hidden in Plain Sight: Influential Sets in Linear Regression*. Tech. rep. 8981. CESifo, 2021 (cit. on pp. 3, 4, 18, 23).

[Mea19] R. Meager. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments". In: *American Economic Journal: Applied Economics* 11.1 (2019), pp. 57–91. DOI: 10.1257/app.20170299 (cit. on pp. 14–16).

[NP12] N. Nunn and D. Puga. "Ruggedness: The Blessing of Bad Geography in Africa". In: *Review of Economics and Statistics* 94.1 (2012), pp. 20–36. DOI: 10.1162/REST_a_00161 (cit. on p. 23).

[RH25] I. Rubinstein and S. Hopkins. "Robustness Auditing for Linear Regression: To Singularity and Beyond". In: *International Conference on Learning Representations*. 2025 (cit. on pp. 3, 4).