Hidden in Plain Sight: Influential Sets in Linear Regression

Nikolas Kuschnig[1,*]

Gregor Zens[2]

Jesús Crespo Cuaresma[1,3,4,5,6]

[1] Vienna University of Economics and Business

[2] Bocconi University

[3] International Institute for Applied Systems Analysis

[4] Wittgenstein Centre

[5] Austrian Institute of Economic Research

[6] CESifo

August 20, 2022

Abstract

The robustness of inferential results drawn using econometric methods is central to their credibility. In this paper, we investigate the sensitivity of regression-based inference to the removal of influential sets of observations theoretically and empirically. We explore three approximate algorithms that address computational challenges and masking issues. We investigate sensitivity in previous development economics studies that assess the development impacts of slave trade, microcredit, and migration. The results of these studies are sensitive to relatively small influential sets. We show how sensitivity checks to influential sets can inform researchers about misspecification issues, the existence of heterogeneous effects, and problems with external validity.

Keywords: regression diagnostics, robustness, masking, influence

*Correspondence to nikolas.kuschnig@wu.ac.at, Welthandelsplatz 1, 1020 Vienna, Austria. The authors gratefully acknowledge helpful comments from Ryan Giordano, Daniel Peña, and Anthony Atkinson, as well as from participants in the 2021 Annual Congress of the Austrian Economic Association.

1 Introduction

Econometric methods are an important instrument of scientific discovery and vital for the design of evidence-based policy. By approximating real-world phenomena, they provide us with empirical insights, allow us to test theories, and facilitate prediction. The sensitivity of empirical inferences drawn to modeling assumptions is of particular importance in the field of applied econometrics, and is a long-standing and active subject of research. The existing literature tends to focus on sensitivity along the *horizontal* dimension of the data, i.e., related to the functional form of model specifications. Approaches that address these aspects include extreme bounds analysis (Leamer, 1983), shrinkage methods (Sims, 1980), model averaging (Steel, 2020), elaborate research designs that reduce model dependence (Angrist and Pischke, 2010), and randomization (Athey and Imbens, 2017). The sensitivity of inference to certain sets of observations, i.e., the *vertical* dimension of the data, however, has received less attention in the econometric literature.

There is a long of studies addressing the role of influential observations and outliers in regression models in the statistical literature(see Belsley et al., 1980, for instance). It is well known that single influential observations may hold considerable sway over regression results (Cook, 1979). Sensitivity checks based on the exclusion of such observations can provide important insights, and there are many approaches to identify them and account for their impacts (see e.g. Chatterjee and Hadi, 1986). This is not the case for *influential sets* of observations, both theoretically and empirically. Analyzing influential sets exactly is extremely expensive in terms of computation, while approximations must address *masking*, a phenomenon where certain observations obscure the influence of others (Chatterjee and Hadi, 1986). For these reasons, the literature has mostly sidestepped influential sets, for instance via robust estimators or resampling methods (see e.g. Hampel et al., 2005; Efron and Tibshirani, 1994).

In this paper, we investigate the sensitivity of regression-based inferential results to influential sets of observations. We evaluate three algorithmic approaches for assessing the sensitivity of regression statistics to influential sets. These algorithms are aimed at identifying maximally influential sets, and combine exact computation and approximate heuristics to avoid the costs of an exhaustive search. To illustrate our approach and guide applied researchers, we revisit several empirical studies in the field of economic development. We consider the number of observations that need to be excluded to overturn a particular result as a summary statistic, and find that some established results may be less robust than initially thought. In particular, the long-run impacts of bad geography in Africa may not be mediated by the slave trades (as proposed by Nunn and Puga, 2012), and their impacts on interpersonal trust are more heterogeneous than previously assumed (Nunn and Wantchekon, 2011). We find that masking issues consistently appear in practice, posing a challenge for credible sensitivity checks.

The remainder of the paper is structured as follows. In Section 2, we establish the theoretical framework and connect it to the relevant literature in statistics and econometrics. The algorithmic approaches are presented and illustrated in Section 3. In Section 4, we investigate the sensitivity to influential sets of previous studies on the development impacts of slave trade, microcredit, and migration, in order to illustrate our approaches. We discuss our findings and conclude in Section 5.

2 Influential sets in linear regression models

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is an $N \times 1$ vector containing observations of the dependent variable, $\mathbf{X}$ is an $N \times P$ matrix with the explanatory variables, β is a $P \times 1$ vector of coefficients to be estimated, and $\boldsymbol{\varepsilon}$ is an $N \times 1$ vector of independent error terms with zero mean and unknown variance σ^2. We denote the ith rows of $\mathbf{y}$ and $\mathbf{X}$ as y_i and x_i. The deletion of the rows indexed by a set I is indicated by a bracketed subscript (I), i.e. $\mathbf{y}_{(I)}$ denotes the vector of observations of the dependent variable without the elements identified by the set I. A set of observations, $\mathcal{S}$, is defined as a subset of the set of all observations $\tilde{\mathcal{S}} = \{s | s \in \mathbb{Z} \cap [1, N]\}$. We denote the cardinality of a set using $|\mathcal{S}_\alpha| = N_\alpha = \lceil N\alpha \rceil$, where $\alpha \in [0, 1]$. The empty set is denoted by $\varnothing$ and the set of all sets of cardinality N_α is referred to as $[\mathcal{S}]^\alpha$.

Our main focus lies on the sensitivity of some quantity of interest, λ, to removing influential sets from the sample. This quantity could be a coefficient value or standard error. We define influential sets following Belsley et al. (1980), as sets whose omission has a large impact on λ, when compared to the omission of most other sets of equal size. We measure the impacts of removing such sets with a generalized influence function (Hampel et al., 2005) that accommodates sets of observations. If we are interested in the sensitivity of the full sample, ordinary least squares (OLS) estimate of β in Equation 1 to dropping a set of observations $\mathcal{S}$, we compare $\lambda_{(\varnothing)}$ and $\lambda_{(\mathcal{S})}$. For this, we consider the following particular definition of λ as a function of the removed set

$$\lambda(\mathcal{S}) = \lambda_{(\mathcal{S})} = \left(\mathbf{X}'_{(\mathcal{S})}\mathbf{X}_{(\mathcal{S})}\right)^{-1} \mathbf{X}'_{(\mathcal{S})}\mathbf{y}_{(\mathcal{S})}, \tag{2}$$

where we suppress the dependence on the data in order to simplify notation. To assess the sensitivity of λ, we consider the *minimal influential set*, i.e. the smallest set whose omission achieves a target impact on λ. We formalize this set by defining the *maximally influential set* $\mathcal{S}^*_\alpha$, which achieves the maximal influence for a given number of omitted observations, as

$$\mathcal{S}^*_\alpha = \underset{\mathcal{S} \in [\mathcal{S}]^\alpha}{\arg\max}\, \Delta(\lambda, \mathcal{S}, \mathcal{T}), \tag{3}$$

where the influence function Δ measures the impact difference on λ, when removing a set $\mathcal{S}$ compared to a set $\mathcal{T}$. To ease notation, we will suppress the dependence on λ, and on $\mathcal{T}$, as long as $\mathcal{T} = \varnothing$. One example for Δ is the squared deviations from the full sample OLS estimator (see Equation 2), $\Delta(\mathcal{S}) = (\lambda_{(\varnothing)} - \lambda_{(\mathcal{S})})'(\lambda_{(\varnothing)} - \lambda_{(\mathcal{S})})$.

The *minimal influential set* $\mathcal{S}^{**}$ can then be defined as the set

$$\mathcal{S}^{**} = \mathcal{S}^*_{\arg\min_\alpha} \text{ s.t. } \Delta(\mathcal{S}^*_\alpha) \geq \Delta^*, \tag{4}$$

where Δ^* is a target value of choice. An example is the minimal influential set achieving a sign switch of a coefficient. This can be achieved by setting $\Delta^* = 0$ and using $\Delta(\mathcal{S}) = -\operatorname{sign}(\lambda_{(\varnothing)})\lambda_{(\mathcal{S})}$. After obtaining a minimal influential set, we may be interested in its size (both in absolute terms and relative to the full sample size) and potentially the characteristics of its members.

Identifying a minimal influential set using full enumeration is computationally prohibitive, and we must rely on approximations for all but the most trivial examples. In order to assess approximate approaches, we require some additional concepts related to the estimates and potential sources of error. First, observations that are truly members of a maximally influential set $\mathcal{S}^*_\alpha$ may not be identified in the estimated set $\hat{\mathcal{S}}^*_\alpha$, a phenomenon we call *masking*. These observations are *masked* by other, seemingly more influential, observations. Masked observations are given by the set difference $\mathcal{S}^*_\alpha - \hat{\mathcal{S}}^*_\alpha$. The severity of masking can be quantified using the number or share of masked observations, or by the difference in influence between the true and estimated sets. Second, the estimated impact of removing a set may not reflect the true value due to a poor approximation of $\Delta(\mathcal{S})$.

Both issues are tightly related to the concept of *jointly influential sets*. To analyze jointly influential sets, it helps to consider the partial influence of a set $\mathcal{S}$ on a scalar λ, given by the influence function $\Delta(\mathcal{S}) = \lambda_{(\varnothing)} - \lambda_{(\mathcal{S})}$. We use the shorthand $\delta_i = \delta_{i|\varnothing} = \Delta(\{i\}, \varnothing)$ for the effect of removing a single observation from the full sample. A *jointly influential set* is one that satisfies $|\delta_{i|j}| \gg |\delta_{i|\varnothing}|$ for all $i, j \in \mathcal{S}$, that is, the impact of removing i after j has been removed is much larger than removing i from the full sample. In other words, the impact of removing any member j from a jointly influential set greatly increases the impact of removing any other member i. A corollary of this definition is that the influence of a jointly influential set exceeds the sum of influences of its individual members, i.e. $\Delta(\mathcal{S}) > \sum_{i \in \mathcal{S}} \delta_i$.

2.1 Influential observations

The statistics literature is rich in methods for the identification of single influential observations, i.e. influential sets with cardinality $N_\alpha = 1$ (see e.g. Chatterjee and Hadi, 1986; Hampel et al., 2005; Maronna et al., 2019). Measures of influence are central to this pursuit, and there is a wide variety of interrelated statistics used for the purpose of measuring the impact of the removal of an observation (Chatterjee and Hadi, 1986). The residuals, $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$, and the leverage (the diagonal elements of the "hat matrix" $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$) are pivotal elements of such influence measures. In the linear regression setting, an observation is considered influential if it exhibits both a large residual and high leverage. A group of influence measures builds on the influence function of Hampel et al. (2005). These include Cook's distance (Cook, 1979), the Welsch-Kuh distance (Belsley et al., 1980), and modifications thereof. The notion of a confidence ellipsoid is used in other approaches for quantifying influence, including the likelihood distance (Cook and Weisberg, 1982), and robust estimation methods (Hampel et al., 2005; Huber, 1964; Lewis et al., 2021). In addition, a number of Bayesian approaches for assessing sensitivity to influential observations have been proposed (Box and Tiao, 1968; Kass et al., 1989; Pettit and Young, 1990; Verdinelli and Wasserman, 1991).

Most of these approaches consider holistic notions of sensitivity. However, in applied econometrics we tend to care about the sensitivity of a specific inferential quantity, such as the sign and significance of a coefficient. Assessing the sensitivity of these two to influential observations thus appears as a natural robustness check. The literature has produced a number of relevant results for assessing this type of sensitivity to influential observations. A prominent example considers the sensitivity of elements of the vector of OLS estimates. The influence of a single observation in this setting is given by

$$\delta_i = \lambda_{(\varnothing)} - \lambda_{(\{i\})} = \frac{\left(\mathbf{X}'\mathbf{X}\right)^{-1} x_i' e_i}{1 - h_i}, \tag{5}$$

where $h_i = \mathbf{H}_{ii}$ is the leverage of observation i. This statistic, termed DFBETA_i by Belsley et al. (1980), is a well known and widely available measure of influence for single observations on individual coefficient estimates. Similar results for σ^2 or the standard errors of coefficient estimates are also available. Belsley et al. (1980) use these quantities to build holistic measures of influence such as scaled versions of Equation 5 and the 'difference in fits' statistic, i.e. the change in predicted value of an observation if that observation is dropped. The partial effects of the removal of individual observations are directly available for many interesting quantities and settings, but may be expensive to compute in some cases. One example are the standard errors for parameter estimates, which rely on $(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}$. Updating formulas for matrix factorizations are helpful in such cases, but not always available or applicable — e.g. for the coefficients using instrumental variable estimation (as derived by Phillips, 1977).

2.2 Identifying influential sets

The identification of maximally influential sets is considerably more challenging than in the single observation case. Exactly determining a maximally influential set by enumeration requires a total of $\binom{N}{N_\alpha}$ calculations of the influence measure. This is computationally infeasible for all but the simplest problems. Despite the availability of closed-form results and fast approximations, solutions are generally intractable for $N_\alpha > 1$, even when influence measures can be calculated near instantly.[1] To make the analysis tractable, maximally influential sets need to be approximated, for instance using observations with large individual impacts (see e.g. Broderick et al., 2020). Importantly, there is a trade-off between the accuracy of the identified set and computational speed. Approximations may run into severe masking problems, especially in the presence of particularly influential observations or jointly influential sets.

The literature largely circumvents influential sets, inter alia focusing on robust statistics. Robust estimators, like M- and S-estimators, allow for inference that is resistant to outliers and influential observations (see, e.g., Hampel et al., 2005; Huber, 1964; Maronna et al., 2019). Overall, these robust methods are rarely used in applied work (Stigler, 2010). Riani et al. (2014) posit that this is due to (1) the number of decisions that need to be made (for instance regarding the breakdown point (the number of 'incorrect' observations an estimator can handle before producing an 'incorrect' result, see Maronna et al., 2019), (2) prohibitive computational cost, and (3) loss of information. Stigler (2010) notes that the concept of 'robustness' itself is elusive, and argues for using more investigative approaches.

Quantifying the sensitivity of inference to the particular (sub)sample employed can yield valuable insights into the data and issue under investigation. Examples include

[1] Consider a total number of observations of $N = 1,000$ and all potential sets of size $N_\alpha = 10$. Assume that every calculation of λ needs one microsecond — very roughly the time needed to compute the cross-product of a four-by-four matrix in R. Enumeration would still require about 8.35 billion years, or 1.8 times the age of the earth, which is safely out-of-scope for non-tenured researchers.

resampling methods such as the jackknife and the bootstrap, or model selection methods (Efron and Stein, 1981; Efron and Tibshirani, 1994; Hoeting et al., 1996). These yield holistic measures of sensitivity, which are generally approximate due to their daunting computational complexity.[2] Another important example are methods for the detection of outliers. Outliers have no precise definition, and detection is generally an unsupervised task, without a precisely defined objective function. For their detection, clustering methods are popular tools (Hautamaki et al., 2004; Kaufman and Rousseeuw, 2009; Kim and Krzanowski, 2007; Shotwell and Slate, 2011), but these approaches are also expensive computationally, motivating the use of sequential approximations (see e.g. Swallow and Kianifard, 1996).

An important strand of the literature focuses on the detection of influential sets or outliers based on the influence of single observations (Belsley et al., 1980; Broderick et al., 2020; Cook and Weisberg, 1982), on decomposition of the influence matrix (Peña and Yohai, 1995), or on deviance (see e.g. Atkinson and Riani, 2000; Atkinson et al., 2004). Simple heuristics and ad-hoc checks that are straightforward to compute and interpret are commonly used in practice (see e.g. Broderick et al., 2020; Nunn and Puga, 2012; Young, 2020), while more complicated procedures have received little attention. This is problematic, since there is a considerable risk of false negatives, particularly in settings where influential sets or jointly influential sets are present.

3 Algorithms to identify influential sets

In this section, we formalize three algorithms for identifying minimal influential sets. They differ in their trade-off between accuracy and speed, and variants or special cases of them have been used and theorized about in the literature. We start with an algorithm that is extremely cheap to compute, but prone to masking problems, and proceed with two algorithms that are able to address masking issues, at slightly increased computational cost. We discuss the use of approximate methods within these algorithms and illustrate their sensitivity to influential sets and their performance using a simple example.

3.1 Algorithm 0: Initial approximation

The first algorithm (Algorithm 0) builds on influence estimates for single observations, computed for the full sample. Maximally influential sets are built by ordering observations based on their initial, full-sample influence. The influence of sets of observations is then

[2]For instance, Young (2020) finds the delete-2 Jackknife to be infeasible to compute.

approximated by accumulating individual influence estimates. Algorithm 0 is extremely cheap to compute, but is prone to false negatives.

Algorithm 0: Initial approximation.

Result: $\hat{\mathcal{S}}^{**}$, an estimate of the minimal influential set.
choose the functions λ and Δ, target threshold Δ^*, and maximal size U;
compute $\delta_i = \Delta(\{i\})$ for all $i \in \tilde{\mathcal{S}}$;
let $\mathcal{S} \leftarrow \arg\max_i \delta_i$;
while $\Delta(\mathcal{S}) < \Delta^*$ **do**
  let $\mathcal{S} \leftarrow \mathcal{S} \cup \arg\max_j \delta_j$, for $j \notin \mathcal{S}$;
  let $\Delta(\mathcal{S}) \leftarrow \sum_{k \in \mathcal{S}} \delta_k$;
  if $|\mathcal{S}| \geq U$ **then return** *unsuccessful*;
end
return $\mathcal{S}$;

The algorithm works as follows. Given an influence function, target value, and a maximum size for the minimal influential set, we compute δ_i for each i. This computation is trivial in many interesting cases; otherwise, approximate methods could be used. The approach used by Broderick et al. (2020), for instance, is a special case of this algorithm, where δ_i is computed using a linear approximation.[3] The first iterated step is the proposal of a maximally influential set, based on the union of the observations with the largest partial influence. The influence of the set is estimated by summing the partial influence of observations in the set. These two steps are repeated until the specified threshold of the target function or the maximum size is reached. The result is an estimate of the minimal influential set and of its influence.

The method embodied in Algorithm 0 can yield useful insights, as demonstrated by the striking findings of Broderick et al. (2020). However, the low computational cost comes at the price of accuracy. First, proposing maximally influential sets based on initial, full-sample influence make the algorithm prone to problems related to masking of influential observations. The influence measure is not updated after the iterated removal of observations, and particularly influential observations may mask others. This means that the algorithm fails to account for jointly influential sets. Second, the influence approximation suffers from a downward bias (see the results in the Appendix) that is made considerably worse by the presence of jointly influential sets. Results obtained using Algorithm 0 are thus prone to convey a false sense of robustness to influential sets of observations.

[3] See the Appendix for a discussion of the performance of the approach in Broderick et al. (2020).

3.2 Algorithm 1: Initial binary search

The second algorithm considered, Algorithm 1, rectifies some critical issues of Algorithm 0, while retaining high computational efficiency. Like Algorithm 0, it builds maximally influential sets based on the ordering of the initial, full-sample influence. The influence of the proposed sets, however, is calculated exactly instead of being approximated by their cumulative partial influence. To minimize computational cost, the algorithm follows a binary search pattern to determine the minimal influential set. Algorithm 1 is thus designed to yield more precise results with minimal computational overhead.

Algorithm 1: Initial search.

Result: $\hat{\mathcal{S}}^{**}$, an estimate of the minimal influential set.
choose the functions λ and Δ, threshold Δ^*, and maximal size U, let $L \leftarrow 0$;
compute $\delta_i = \Delta(\{i\})$ for all $i \in \tilde{\mathcal{S}}$;
while $L \leq U$ **do**
  let $M \leftarrow \lfloor (L+U)/2 \rfloor$;
  let $\mathcal{S}$ be the union of the indices of the M largest δ_i;
  calculate $\Delta(\mathcal{S})$;
  if $\Delta(\mathcal{S}) < \Delta^*$ **then** let $L \leftarrow M+1$;
  if $\Delta(\mathcal{S}) \geq \Delta^*$ **then** let $U \leftarrow M-1$;
end
return $\mathcal{S}$;

Instead of sequentially increasing the size of proposed sets, Algorithm 1 sets the size of proposed sets, M, by iteratively halving the search interval $[L, U]$. In each step, the influence of the proposed set is computed exactly, and the bounds are updated depending on the result. If the target is reached, the upper bound is decreased to $M-1$, otherwise the lower bound is increased to $M+1$. If an approximate minimal influential set exists in the interval, it is found after $\mathcal{O}(\log U)$ steps. The initial upper bound can be set exogenously or using Algorithm 0 as a first approximation. The complexity of $\mathcal{O}(\log U)$ adds negligible overhead over Algorithm 0, making it this divide-and-conquer approach a practical choice for large scale problems.

Algorithm 1 performs considerably better than Algorithm 0, but its accuracy is not guaranteed. The identification of influential sets still relies on initial, full-sample influence estimates for individual observations, meaning that minimal influential sets may remain hidden due to masking. This is particularly problematic if jointly influential sets are present in the sample. Fortunately, masking does not directly affect the quality of influence estimates, only that of proposed sets. The algorithm relies on the (previously implicit) assumption that the influence of estimated maximally influential sets increases steadily. With the next algorithm, we present a simple approach that discards this assumption and is designed to address jointly influential sets more effectively.

3.3 Algorithm 2: Adaptive approximation

The third algorithm proposed, Algorithm 2, uses a simple adaptive procedure for identifying the minimal influential set. Maximally influential sets are identified iteratively, facilitating the discovery of potentially masked observations. Measures of influence are precise, since they are calculated exactly. In addition, the adaptive nature of the algorithm improves the accuracy of proposed maximally influential sets.

Algorithm 2: Adaptive approximation.

Result: $\hat{\mathcal{S}}^{**}$, an estimate of the minimal influential set.
choose the functions λ and Δ, target Δ^*, and maximal size U, let $\mathcal{S} \leftarrow \varnothing$;
while $\Delta(\mathcal{S}) < \Delta^*$ **do**
  compute $\Delta(\mathcal{S} \cup j)$, for all $j \in \tilde{\mathcal{S}} - \mathcal{S}$;
  let $\mathcal{S} \leftarrow \mathcal{S} \cup \arg\max_j \Delta(\mathcal{S} \cup j)$;
  if $|\mathcal{S}| \geq U$ **then return** *unsuccessful*;
end
return $\mathcal{S}$;

Algorithm 2 starts by computing the influence of all individual observations in the sample. Candidate maximally influential sets are built adaptively, by forming the union of the previous set (starting with the empty set) and the observation with the highest influence. In each step, the exact influence of the new set is computed. These steps are repeated until a minimal influential set is found, or the maximal size is reached. The complexity scales linearly with the cardinality of the set, so computational demand is rarely prohibitive and falls well short of, e.g., a jackknife approach. By adapting after every removal, this approach reduces the risk of masking problems, and allows us to investigate sensitivity to the presence of jointly influential sets in a more reliable manner.

3.4 Approximations and computational concerns

The computational complexity of determining minimal influential sets is daunting, motivating approximate procedures such as the ones presented above. Optimization steps aimed at improving the calculation of influence measure are crucial to improve the efficiency of these algorithms. An important and conceptually straightforward aspect is the efficient execution of the underlying linear algebra operations.[4] For many influence measures, updating formulas for inverse matrices, cross-products, and matrix factorizations (see e.g. Sherman and Morrison, 1950; Reichel and Gragg, 1990; Hammarling and Lucas, 2008) can improve computation considerably (for example in the case of standard errors or significance-based measures).

Further improvements are possible with approximate methods. When the number of covariates is large (e.g. in panel data regression models with cross-sectional fixed effects), it can be helpful to marginalize out nuisance covariates using the Frisch-Waugh-Lovell theorem before computing influence measures. The use of iterative solvers with suitable stopping criteria can also aid rapid computation (see e.g. Trefethen and Bau, 1997). Further speed gains are possible when accuracy is sacrificed. For instance, Broderick et al. (2020), propose a linear approximation to the influence measure that can be computed using automatic differentiation. In the context of Equation 5, however, this approach skips the effects of leverage, yielding $\delta_i \approx (\mathbf{X}'\mathbf{X})^{-1} x_i' e_i$. Due to the implicitly fixed leverage, the influence of all observations is biased downwards, with a particularly large bias for influential observations (see the Appendix). Similarly to the drawbacks of Algorithm 0, the linear approximation of Broderick et al. (2020) may yield appropriate results in the absence of influential observations, but performs worse in other scenarios.

3.5 An illustration

We illustrate the characteristics of the proposed algorithms using a simple example. Consider fitting a simple regression line to the data depicted in Figure 1. The three observations in the top-right of the scatter plot, marked (a), are influential on the estimated positive slope. This circumstance is reflected in standard diagnostics, and, unsurprisingly, all three proposed algorithms correctly identify this set of observations as influential. The identification of subsequent observations, however, is no longer consistent. The high influence of the three observations in the center-right of the scatter plot, marked (b), is masked by the first three observations. Neither Algorithm 0, nor Algorithm 1 unveil their influence. The adaptive nature of Algorithm 2 allows it to identify their influence, and correctly adds them to maximally influential sets of size four or above.

Figure 2 depicts the maximally influential sets identified at increasing size, as well as slopes after removing sets of size three and seven. Masking affects the identification of influential sets: Algorithms 0 and 1 (in the left panel) identify relatively inconsequential observations after the first three removals. This is reflected in the slope of the estimated

[4] Gains from optimized computation of linear algebra operations can be large in practice. When writing this paper, we were able to improve the speed of the influence functionality in the **ivreg** R package three-fold by optimizing its linear algebra operations.

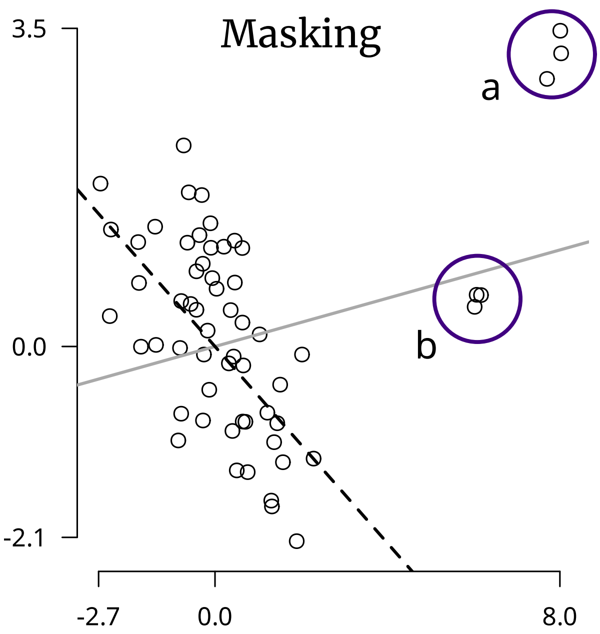


Figure 1: Regression line for a dataset with two influential sets of size three ($N = 50$). The set labeled (a) masks the influence of the set labeled (b). The solid gray line indicates the least-squares slope; the dashed line indicates the slope after removing both sets.

regression lines, which reaches a relative peak after three removals. Moreover, the approximation in Algorithm 0 even fails to account for the full influence of the first three observations, considerably underestimating the slope after the first removal. Algorithm 2 does not suffer from these problems; the estimated regression lines are accurate, and masking issues are avoided. For three removals, the implied regression line mirrors the one of Algorithm 1; there are no masking issues and associated downward bias for further removals. Speed-wise, all three algorithms can be effectively implemented instantaneously (under 0.1 seconds).

4 Empirical applications

In this section, we investigate the sensitivity to influential sets in different econometric models, which form the basis of recent empirical results in the context of development economics. We target influence on t values and consider the sign and significance of the main coefficient as thresholds of interest. For the applications considered, we report the number of observation removals needed to (1) lose significance of the parameter of interest (at a given level), (2) flip its sign, and (3) obtain a significant estimate of the opposite sign. For clarity, we focus on comparing the results obtained using Algorithms 0 and 2.[5]

[5]An application of all three algorithms to the empirical assessment of poverty convergence (following Ravallion, 2012; Crespo Cuaresma et al., 2022) is provided as an additional empirical illustration in the Appendix.

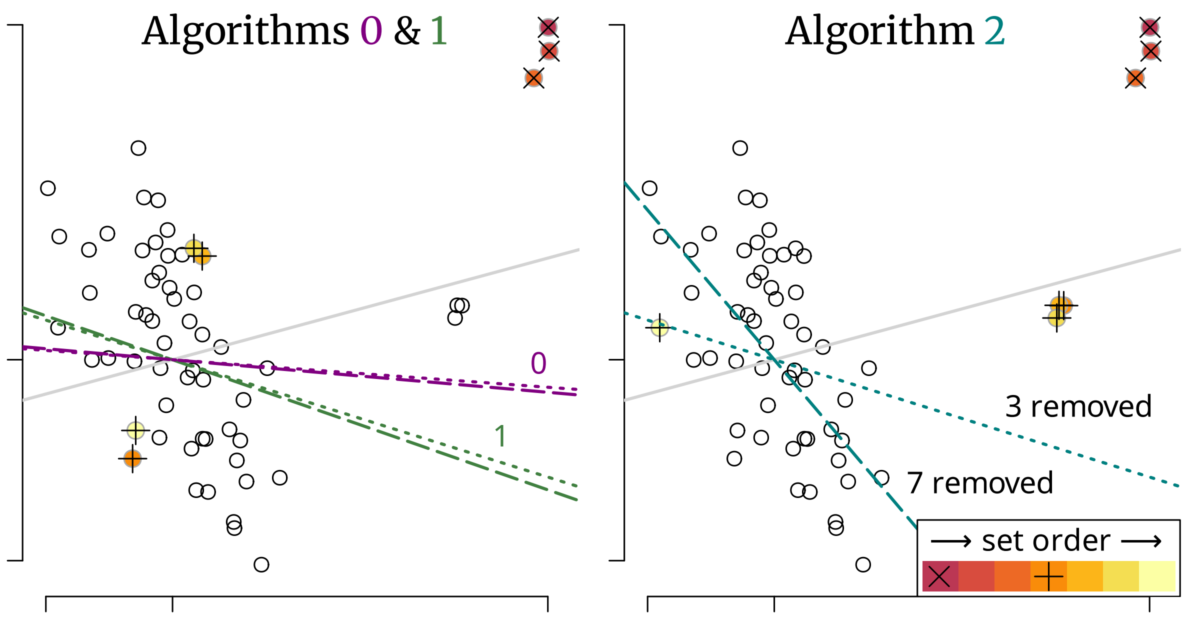


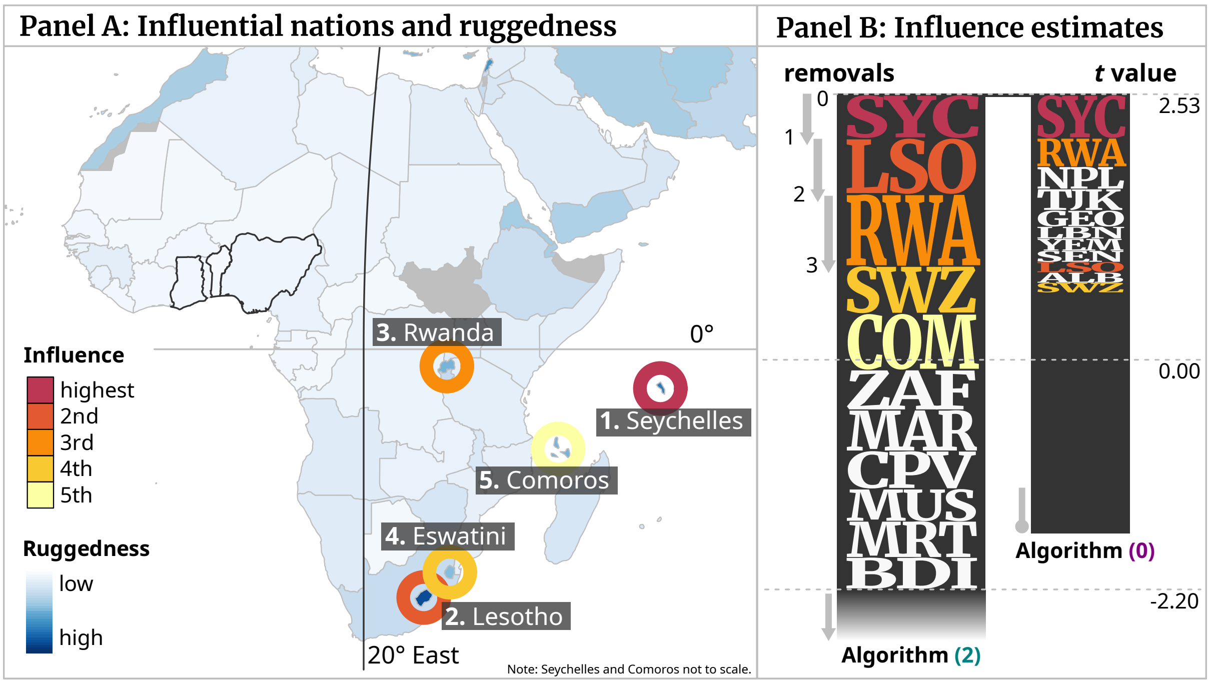
Figure 2: Influential set identification with Algorithms 0 and 1 (left panel, identical sets) and Algorithm 2 (right panel). Removed observations are color-coded and indicated by a cross (×, first three) and a crosshair (+, next four). Dotted and dashed lines are regression lines after removing three and seven observations, respectively; lines for Algorithm 0 are in purple, for Algorithm 1 in green, and for Algorithm 2 in teal.

4.1 Robustness of the effects of slave trades on development

The lackluster economic performance of many Sub-Saharan African nations over the last decades can partly be explained by historical factors, including colonial experiences (Acemoglu et al., 2001) and the slave trades (Nunn, 2008). The slave trades ravaged the African continent until (at least) the emergence of abolitionist movements in the 19th century and can be divided into the Trans-Saharan, the East African, and the Atlantic slave trade. Two recent studies investigate the causality, scale, and pathways of the long-term development impacts of the slave trades. Nunn and Puga (2012) presents evidence that more rugged terrain hinders development in the rest of the world, but not in Africa. They attribute this to the history of the slave trades. Nunn and Wantchekon (2011) argue that interpersonal trust is one of the mediators of the long-term development impacts of the Atlantic and East African slave trades.

In this section, we assess the robustness of these results to influential sets. We find that, despite their use of best-practice sensitivity checks, the results in Nunn and Puga (2012) are heavily reliant on a single influential set, and that the effect found in Nunn and Wantchekon (2011) is likely to be heterogeneous, and only applicable to the Atlantic slave trade.

Figure 3: Two nations drive the blessing of bad geography



Panel A shows ruggedness, the explanatory variable of interest in Nunn and Puga (2012), and highlights the five most influential observations. The three countries with darker borders in the top-right of Panel A are Ghana, Benin, and Nigeria, which, together with the 20° East meridian, are relevant to the results in Nunn and Wantchekon (2011). Panel B shows the reduction of t values (from 2.53 at the top) as observations (indicated with their ISO codes) are removed using Algorithms 2 (left) and 0 (right).

Geography, development, and unobservables

Nunn and Puga (2012) use a linear regression model with interaction terms to assess the differential effects of terrain ruggedness on income for African nations. They regress logged real GDP per capita in 2000 on a ruggedness measure and various controls. They find a significantly negative coefficient estimate for the ruggedness variable for the global sample, and a significantly positive one for ruggedness in Africa, implying that the effects of ruggedness in Africa are significantly different from those in the rest of the world.[6] The results are robust to the inclusion of a number of potential confounders, to alternative measures of income and ruggedness, and to influential observations, which the authors investigate explicitly. Specifically, they check the robustness to omitting observations that exceed a threshold of $|\text{DFBETA}_i| > 2/N$ (following Belsley et al., 1980), as well as the ten smallest (in terms of land area) and most rugged observations.

We focus on the robustness of the differential effect estimate found for Africa in their baseline specification with added controls. In Figure 3, Panel A, we show the five most

[6] Note that the effects of ruggedness in Africa are not significantly positive, as could be inferred from the title ('The blessing of bad geography in Africa'). Instead, there is no significant effect of ruggedness in Africa (also see Table 1).

influential observations identified by Algorithm 2. Using this algorithm, the influence of the five most influential observations increases after the first one is removed, indicating a jointly influential set. The removal of the two most influential observations overturns the differential effect of Africa; removing all five observations leads to a sign-flip, which becomes significant after eleven removals (see Panel B in Figure 3).

The results presented in Panel B of Figure 3, which shows the influence of removing observations on the t value of the interacted ruggedness, reveal that masking problems are present in the data. The observation for the Seychelles masks the identification of subsequent influential observations, like Lesotho, and the initial approximation only identifies two out of the top five and three out of the top ten most influential observations revealed by Algorithm 2. The influence estimates of Algorithm 0 (right bar) are therefore underestimated considerably due to masking. Since the influence approximation peaks before a significant sign-flip is attained, only two thresholds (for removing significance and an insignificant sign-flip) are found, at two and twenty removals. With Algorithm 2 (left bar), we can discern the pattern of a jointly influential set: while the Seychelles are the most influential observation individually, its removal increases the influence of other set members.

Algorithm 2 reveals that the effect of interest is sensitive to influential sets. To gain further insights, we investigate the characteristics of the elements of the influential set. First, the large influence of the Seychelles casts doubts on the interpretation of the results presented in Nunn and Puga (2012). The Seychelles were only settled in the 1770s and were arguably not impacted by the slave trades (Fauvel, 1909). Second, the influential set contains five small (in terms of land area and population) countries. This may indicate some survivorship bias, where the survival of countries (and thus their appearance in the data) is partially determined by characteristics such as land area, ruggedness, population size, and economic success.[7] As can be seen in Table 1, specifications that account for the past population size and land area do not yield significant results.

Slave trades and the origins of mistrust

Nunn and Wantchekon (2011) analyze the role of the Atlantic and East African slave trades as determinants of trust. They regress trust of relatives and trust of neighbors on a measure based on the number of slaves taken from a given ethnic group during the slave trade, controlling for a number of individual- and district-level variables. Additionally, they use country-fixed effects and cluster standard errors along ethnicities and districts. The design matrix contains over 20,000 observations across sub-Saharan Africa, and a

[7]Moreover, past population sizes are likely related to other, confounding, geographical features, and play an important role in mediating potential impacts of the slave trades.

Table 1: Sensitivity of the differential effect of ruggedness in Africa

	Baseline	Plain	Pop	Area	Sep
Ruggedness, Africa†	0.321	0.302	0.190	0.215	0.089
	(2.53)	(2.32)	(1.66)	(1.63)	(0.87)
Ruggedness	-0.231	-0.193	-0.231	-0.238	-0.231
	(-2.99)	(-2.38)	(-2.94)	(-3.08)	(0.08)
Coast distance	Yes	Yes	Yes	Yes	Yes
Population in 1400	–	–	Yes	–	–
Land area	–	–	–	Yes	–
Other controls	Yes	–	Yes	Yes	Yes
Thresholds†	2[5]{11}	2[7]{16}	–[3]{6}	–[4]{8}	–[1]{6}
Observations	170	170	168	170	49
R^2	0.537	0.421	0.571	0.554	0.317

Dependent variable is (log) GDP per capita in 2000. 'Thresholds' reports the number of observation necessary to remove significance (at the 5% level, if applicable), [flip the sign], and {significantly flip the sign} of the effect of ruggedness in Africa on GDP per capita. 'Baseline' and 'Plain' reproduce the results of Nunn and Puga (2012) (see columns 5 and 6 of Table 1 in the paper). 'Pop' presents the results from a specification that adds as a new covariate the population level in the year 1400. In 'Area', the land area of the country (in logs) is added as a control. 'Sep' reports the estimates of the model estimated exclusively using the sample of African economies. The reported t values are based on HC1 robust standard errors and presented in parentheses.

total of 78 regressors. Nunn and Wantchekon (2011) find statistically and economically significant effects of the slave trades on interpersonal trust. These results are robust to dropping Kenya and Mali, which were also affected by the Trans-Saharan slave trade, the use of different proxies for the slave trades, additional controls, and the use of instrumental variables.

We focus on the robustness of the impacts of slave trade on trust of relatives and neighbors, which we reproduce in Table 2. This setup can be considered computationally prohibitive for our purpose — in particular due to the use of two-way clustered standard errors, for which there is no updating formula. To facilitate computation, we only cluster standard errors after the removal of each observation, and not for proposed removals. Using this approach, we find that 105 removals (0.5% of the sample) lead to a loss of significance of the slave trade variable (exports/area in the table), and 380 removals (1.9% of the sample) lead to a sign-flip that becomes significant after 656 removals (3.3% of the sample) for the trust of relatives, with similar results for the trust of neighbors. We investigate the issue further by examining the members of the identified influential sets.

Table 2: The origins of mistrust

	Trust of relatives ∼		Trust of neighbors ∼	
	Pooled	West|East	Pooled	West|East
Exports/area†	-0.133	-0.145	-0.159	-0.168
	(-3.68)	(-3.84)	(-4.67)	(-4.48)
Exports/area, East		0.053		0.023
		(0.96)		(0.32)
Individual controls	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes
Thresholds†	105[380]{656}	78[301]{532}	161[425]{768}	133[323]{527}
Observations	20,062	7,549 | 12,513	20,027	7,523 | 12,504
Ethnicity clusters	185	62 | 123	185	62 | 123
District clusters	1,257	628 | 651	1,257	628 | 651
R^2	0.133	0.199 | 0.097	0.156	0.228 | 0.117

The row named 'Thresholds' gives the numbers of observation necessary to remove significance (at the 1% level), [flip the sign], and {significantly flip the sign} of the effects of the slave trade on trust. The columns labeled 'Pooled' reproduce the results of Nunn and Wantchekon (2011), which can be seen in columns (1) and (2) in Table 2 of their paper. The columns labeled with 'West|East' estimate separate models for observations West and East of the 20° meridian; thresholds refer to the coefficient for the Western subsets. Coefficient estimates are reported with t values based on two-way clustered standard errors in parentheses.

In Table 3, we see that 536 of the 600 most influential observations (89.3%) stem from three West African nations: Benin, Nigeria, and Ghana (marked with black borders in Figure 3). These countries were major centers of the Atlantic slave trade (Nunn, 2008), and their influence may suggest differences in impacts between the Atlantic and East-African slave trades. As can be seen in the columns labeled 'West|East' in Table 2, the trust of both relatives and neighbors is only significantly affected by the slave trade in Western regions, which were affected by the Atlantic slave trade.[8] When considering trust of relatives, overturning these effect estimates for this specification requires the removal of 78 (1.0%), 301 (4.0%), and 532 (7.0%) observations; these results are more robust than the full sample estimates. The finding that the Atlantic slave trade considerably drives trust impacts of the slave trades are in line with earlier results of Nunn (2008).

[8]We divided the dataset into two samples, to the West and to the East of the 20° Eastern meridian, which is marked in Figure 3.

Table 3: National split of the top 600 influential observations on trust in relatives

	Benin	Nigeria	Ghana	Other
Top 100	54	12	19	15
Top 101–200	85	6	5	4
Top 201–300	80	12	7	1
Top 301–400	52	27	14	7
Top 401-500	3	67	6	24
Top 501–600	3	69	15	13
Top 600	277	193	66	64

Summary of origin countries of the 600 most influential observations' on the effects of slave trade on the trust of relatives (see the first column of Table 2).

4.2 Microfinance, development, and influential observations

Microfinance is a tool for alleviating poverty and facilitating economic development in developing countries. Recently, many large-scale studies with experimental designs set out to quantify its impacts and evaluate its efficacy. In the context of an external validity assessment, Meager (2019) discusses and summarizes seven of these studies , which correspond to randomized control trials in Bosnia and Herzegovina (Augsburg et al., 2015), Mongolia (Attanasio et al., 2015), Ethiopia (Tarozzi et al., 2015), Mexico (Angelucci et al., 2015), Morocco (Crépon et al., 2015), the Philippines (Karlan and Zinman, 2011), and India (Banerjee et al., 2015). Broderick et al. (2020) assess the robustness of the average treatment effect to the removal of observations using an approximate variant of Algorithm 0. The model is a simple treatment effect model with a randomized treatment dummy; datasets are relatively large, ranging from 1,000 to 16,000 observations. Broderick et al. (2020) find that the effects of microcredit on household level business profits are not particularly robust, and that the size of estimates is usually driven by very few observations.

Table 4 presents the results of applying Algorithms 0 and 2 to the data of these seven studies. Few observations, both in relative and absolute terms, drive the full-sample estimates of the average treatment effect. In the studies corresponding to Ethiopia and Mexico, a single removal induces a sign-switch, which becomes significant after a few more removals. Results using Algorithm 0 are similar to the ones obtained by Broderick et al. (2020).[9] In fact, standard diagnostic indicators like DFBETA_i or assessments of the

[9]In principle, this is to be expected, since their approximation disregards leverage (see the Appendix), which is constrained in the context of a binary treatment.

Table 4: Sensitivity of the average treatment effect of microcredits

Study region	BIH		MON		ETH		MEX		MOR		PHI		IND	
Algorithm	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)
Sign-switch	14	13	16	15	1	1	1	1	11	11	9	9	6	6
Significance	49	39	43	37	117	13	20	12	35	33	74	54	41	35
Observations	1,195		961		3,113		16,560		5,498		1,113		6,863	

The reported values are the number of removals needed to induce a sign-switch of the average treatment effect, and have this sign-flipped coefficient become significant (at the 1% level) using Algorithm 0 and 2. Algorithm 2 outperforms consistently, but very few observations are needed to overturn results in all cases.

residuals, already indicate sensitivity, and — at a first glance — Algorithm 2 only seems to offer small improvements when more removals are necessary. However, Algorithm 2 avoids crucial issues with false negatives, which are troublesome for any sensitivity check. In the case of Ethiopia, Algorithm 2 suggests that the number of removals needed to induce a significant sign-switch is 13 (0.4%) instead of 117 (3.6%), since particularly influential observations mask the influence of subsequent removals.

The exceptional lack of robustness of the effects found using microcredit randomized control trials is striking. These results indicate a lack of power of the studies, implying that aggregation of many studies may be necessary for robust evidence (as done in Meager, 2019). Further, the results suggest that the effects of the interventions are zero for most of the population, and potentially large for few individuals. This is investigated further using aggregated distributional treatment effect analysis in Meager (2022), which concludes that the effects are usually concentrated at the top segment of the living standard distribution.

4.3 Migration, growth, and instruments

Migration reshapes global populations and, as a result, economic and cultural structures (see e.g. Tabellini, 2020). Droller (2018) investigates the long-term impacts of skilled, European migration to Argentina in the late 19th and early 20th century. The effect of migration is estimated using a shift-share instrument for identification, leveraging the arguable exogeneity of initial migrant shares and the conquest of new counties. The analysis is carried out using observations for 136 counties in the provinces of Buenos Aires, Santa Fe, Córdoba, and Entre Rios. Droller (2018) finds considerable impacts of migration on GDP per capita, education, and skilled labor. These results are robust to a number of checks, although the sensitivity of the results to sets or influential observations is not assessed in the original publication.

Table 5: Sensitivity of long-term migration impacts on development

	Baseline		Plain	
log GDP/capita ∼	1st stage	2SLS	1st stage	2SLS
European share†		5.778		6.35
		(3.40)		(4.31)
Instrument†	0.251		0.281	
	(6.01)		(4.84)	
Geographic controls	Yes	Yes	Yes	Yes
Socioeconomic controls	Yes	Yes	–	–
Province fixed effects	Yes	Yes	Yes	Yes
Thresholds†	10[18]{26}	5[11]{27}	10[18]{31}	6[11]{32}
Observations	136	136	136	136
R^2	0.816	0.622	0.688	0.591

The row named 'Thresholds' gives the numbers of observation necessary to remove significance (at the 5% level, if applicable), [flip the sign], and {significantly flip the sign} of either the shift-share instrument (in columns labelled '1st stage'), or the effects of the share of European migrants (as reflected by 2SLS estimates). The columns labelled 'Baseline' reproduce the 'Specification 2' in Table 4 of Droller (2018), the ones labelled 'Plain' reproduce 'Specification 1'. Coefficients are reported with t values based on HC0 robust standard errors in parentheses.

We reproduce the results of the study and present them in Table 5, computing sample size thresholds for influential sets for the first stage, and for the full two-stage least squares estimates (2SLS, implemented using the updating formulas by Phillips, 1977). Notably, leverage is exacerbated in 2SLS estimation, and influence approximations suffer particularly from the existence of high-leverage observations in this setting. We find that first stage results are relatively robust to influential sets. For the full results, we find that the removal of five observations overturns significance; eleven removals induce a sign-switch for the estimated effect of European migration on GDP per capita. The nature of influential sets and the covariate structure may indicate the existence of heterogeneous effects for Buenos Aires, where most late conquests occurred. While these estimates are not particularly sensitive, we cannot conclusively diagnose robustness.

In the wider context of assessing the sensitivity in inferential quantities for 2SLS estimation, a number of issues need to be considered. First, the numerical stability of estimates can be problematic, and is likely to deteriorate with removals, in particular in the presence of weak instruments.[10]. Second, the robust standard errors used by Droller (2018) are known to be considerably biased, and their finite sample behavior is primarily driven by the presence of high-leverage observations (Cribari-Neto et al., 2007).[11] Finally, the instrument is based on initial population shares (accounting for newly conquered counties), shifted with total migrant flows, a strategy that relies on the exogeneity of either the shares, or the shifts (Borusyak et al., 2022; Goldsmith-Pinkham et al., 2020).

5 Conclusions

In this paper, we investigated the sensitivity of inferential quantities to influential sets in linear regression models. We discussed three alternative algorithms aimed at the identification of influential sets, and quantifying the sensitivity to them. We showed how masking, where certain influential observations obscure the identification of others, complicates robustness checks by inducing false negatives, and discussed accuracy–speed trade-offs to address this issue. The practical relevance of sensitivity to influential sets was investigated by means of four empirical applications in the context of development economics. Our analysis suggested that sensitivity to influential sets plays an important role in applied studies, and its assessment can be very useful to address potential shortcomings of regression models. Masking issues are likely to be common in practice, and can be efficiently addressed using our approach.

Our notion of sensitivity builds on minimal influential sets, i.e. the smallest set of observations that changes a result of interest when it is removed. With this approach, sensitive results are not generally a conclusive indicator for a lack of validity; results that unveil sensitivity to influential sets must be interpreted with care. A great deal of interesting phenomena are exceedingly rare, and important insights may hinge on few observations. Examples include rare diseases, economic crises, and policy interventions that only affect a small part of the population. In settings where the phenomena assessed are not rare, however, pronounced sensitivity to influential sets may indicate a lack of internal validity. The analysis of influential sets can provide insights into possible shortcomings of the model, or potentially fruitful extensions of the analysis. The empirical studies analyzed in this piece provide good examples of the added value of assessing the composition of influential sets.

[10]In the study, collinearity is high due to the large number of geographical controls (which may address persistence issues noted by Kelly, 2020), and about 15 targeted removals can produce estimates that are numerically unstable.

[11]Results with more appropriate HC4 or HC5 standard errors are qualitatively similar, lowering t values of the baseline specification to 2.94 and 3.17, respectively.

In practice, the sensitivity to influential sets (or observations) is rarely assessed systematically in econometric applications. We argue that it should be understood as a standard part of regression diagnostics. For this purpose, summaries of minimal influential sets that overturn a result are useful diagnostic measures. Two important improvements upon related measures, such as Cook's distance, are their interpretability and salience, since summaries are directly tied to a result of interest, and influential sets can be analyzed in detail. Exact minimal influential sets are essentially unobtainable, but approximate algorithms can unveil a lack of robustness that would otherwise remain hidden in plain sight. However, for sensitivity checks, computational concerns must arguably take a backseat in the face of false negatives. Of the proposed algorithms, Algorithm 0 and Algorithm 1 sacrifice a great deal of accuracy for speed, and should be seen as interactive tools for quick sensitivity checks. Algorithm 2 offers a good baseline for a diagnostic tool, yielding more accurate results at reasonable computational costs.

There are several pathways for future work on the topic of influential sets in regression models. Further improvements in terms of computational efficiency and accuracy are conceivable, for instance via sampling-based approaches. In addition, a combination of the approach presented here with methods in the spirit of Peña and Yohai (1999) or Riani et al. (2014) may prove to be an interesting avenue of further research on the topic. The development of comprehensive measures that accessibly summarize sensitivity to influential sets can also be useful, in particular for practitioners. Lastly, the approach presented here opens the door for further replication studies and robustness checks of documented empirical phenomena, which may deliver valuable insights for researchers and policymakers.

References

Daron Acemoglu, Simon Johnson, and James A. Robinson. The colonial origins of comparative development: an empirical investigation. *American Economic Review*, 91(5):1369–1401, 2001. ISSN 0002-8282. doi:10.1257/aer.91.5.1369.

Manuela Angelucci, Dean Karlan, and Jonathan Zinman. Microcredit impacts: evidence from a randomized microcredit program placement experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, 7(1):151–82, 2015. doi:10.1257/app.20130537.

Joshua D. Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010. doi:10.1257/jep.24.2.3.

Susan Athey and Guido W. Imbens. The state of applied econometrics: causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32, 2017. doi:10.1257/jep.31.2.3.

Anthony Atkinson and Marco Riani. *Robust diagnostic regression analysis*. Springer, New York, NY, USA, 2000. ISBN 978-1-4612-1160-0. doi:10.1007/978-1-4612-1160-0.

Anthony C. Atkinson, Marco Riani, and Andrea Cerioli. *Exploring multivariate data with the Forward Search*. Springer, New York, United States, 2004. ISBN 978-0-387-21840-3. doi:10.1007/978-0-387-21840-3.

Orazio Attanasio, Britta Augsburg, Ralph De Haas, Emla Fitzsimons, and Heike Harmgart. The impacts of microfinance: evidence from joint-liability lending in Mongolia. *American Economic Journal: Applied Economics*, 7(1):90–122, 2015. doi:10.1257/app.20130489.

Britta Augsburg, Ralph De Haas, Heike Harmgart, and Costas Meghir. The impacts of microcredit: evidence from Bosnia and Herzegovina. *American Economic Journal: Applied Economics*, 7(1):183–203, 2015. doi:10.1257/app.20130272.

Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, 7(1):22–53, 2015. doi:10.1257/app.20130533.

Robert J. Barro and Xavier Sala-i Martin. Convergence. *Journal of Political Economy*, 100(2): 223–251, 1992. doi:10.1086/261816.

David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 1980. doi:10.1002/0471725153.

Kirill Borusyak, Peter Hull, and Xavier Jaravel. Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1):181–213, 2022. ISSN 0034-6527. doi:10.1093/restud/rdab030.

François Bourguignon. The growth elasticity of poverty reduction: explaining heterogeneity across countries and time periods, 2003.

George E. P. Box and George C. Tiao. A Bayesian approach to some outlier problems. *Biometrika*, 55(1):119–129, 1968. doi:10.1093/biomet/55.1.119.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: can dropping a little data change conclusions?, 2020.

Samprit Chatterjee and Ali S. Hadi. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393, 1986. doi:10.1214/ss/1177013622.

Ralph Dennis Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174, 1979. doi:10.2307/2286747.

Ralph Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*. New York: Chapman and Hall, 1982. ISBN 0-412-24280-0.

Bruno Crépon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1):123–50, 2015. doi:10.1257/app.20130535.

Jesús Crespo Cuaresma, Stephan Klasen, and Konstantin M. Wacker. There is poverty convergence. *SSRN Electronic Journal*, 2016. doi:10.2139/ssrn.2718720.

Jesús Crespo Cuaresma, Stephan Klasen, and Konstantin M. Wacker. When do we see poverty convergence? *Oxford Bulletin of Economics and Statistics*, 2022. ISSN 0305-9049. doi:10.1111/obes.12492.

Francisco Cribari-Neto, Tatiene C. Souza, and Klaus L. P. Vasconcellos. Inference under heteroskedasticity and leveraged data. *Communications in Statistics - Theory and Methods*, 36 (10):1877–1888, 2007. doi:10.1080/03610920601126589.

Federico Droller. Migration, population composition and long run economic development: evidence from settlements in the Pampas. *The Economic Journal*, 128(614):2321–2352, 2018. doi:10.1111/ecoj.12505.

Bradley Efron and Charles Stein. The Jackknife estimate of variance. *The Annals of Statistics*, 9(3), 1981. doi:10.1214/aos/1176345462.

Bradley Efron and Robert J. Tibshirani. *An introduction to the Bootstrap*. CRC Press, 1994. doi:10.1201/9780429246593.

A. A. Fauvel. *Unpublished documents on the history of the Seychelles islands anterior to 1810*. Government Printing Office Mahe, Seychelles, 1909. URL `www.loc.gov/item/unk83018617/`.

Ryan Giordano, Runjing Liu, Michael I. Jordan, and Tamara Broderick. Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Analysis*, -1(-1):1–34, 2022. ISSN 1936-0975. doi:10.1214/22-BA1309.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: what, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020. ISSN 0002-8282. doi:10.1257/aer.20181047.

Sven Hammarling and Craig Lucas. Updating the QR factorization and the least squares problem. Technical report, Manchester Institute for Mathematical Sciences, University of Manchester, 2008. URL `http://eprints.maths.manchester.ac.uk/id/eprint/1192`.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2005. doi:10.1002/9781118186435.

Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 430–433. IEEE, 2004. doi:10.1109/ICPR.2004.1334558.

Jennifer Hoeting, Adrian E. Raftery, and David Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis*, 22(3):251–270, 1996. doi:10.1016/0167-9473(95)00053-4.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi:10.1214/aoms/1177703732.

Paul Johnson and Chris Papageorgiou. What remains of cross-country convergence? *Journal of Economic Literature*, 58(1):129–75, 2020. doi:10.1257/jel.20181207.

Dean Karlan and Jonathan Zinman. Microcredit in theory and practice: using randomized credit scoring for impact evaluation. *Science*, 332(6035):1278–1284, 2011. doi:10.1126/science.1200138.

Robbert E. Kass, Luke Tierney, and Joseph B. Kadane. Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika*, 76(4):663–674, 1989. doi:10.1093/biomet/76.4.663.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: An introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. ISBN 978-0470316801. doi:10.1002/9780470316801.

Morgan Kelly. Understanding persistence, 2020. URL `https://ssrn.com/abstract=3688200`.

Sung-Soo Kim and Wojtek J. Krzanowski. Detecting multiple outliers in linear regression using a cluster method combined with graphical visualization. *Computational Statistics*, 22(1): 109–119, 2007. doi:10.1007/s00180-007-0026-3.

Stephan Klasen and Mark Misselhorn. Determinants of the growth semi-elasticity of poverty reduction, 2008.

Edward E. Leamer. Let's take the con out of econometrics. *American Economic Review*, 73(1): 31–43, 1983. URL `https://www.jstor.org/stable/1803924`.

John R. Lewis, Steven N. MacEachern, and Yoonkyung Lee. Bayesian restricted likelihood methods: conditioning on insufficient statistics in Bayesian regression. *Bayesian Analysis*, pages 1–38, 2021. doi:10.1214/21-BA1257.

Ricardo A Maronna, R Douglas Martin, Victor J Yohai, and Matías Salibián-Barrera. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019. doi:10.1002/9781119214656.

Rachael Meager. Understanding the average impact of microcredit expansions: a Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019. doi:10.1257/app.20170299.

Rachael Meager. Aggregating distributional treatment effects: a Bayesian hierarchical analysis of the microcredit literature. *American Economic Review*, 2022. ISSN 0002-8282. doi:10.1257/aer.20181811.

Nathan Nunn. The long-term effects of Africa's slave trades. *Quarterly Journal of Economics*, 123(1):139–176, 2008. ISSN 0033-5533. doi:10.1162/qjec.2008.123.1.139.

Nathan Nunn and Diego Puga. Ruggedness: the blessing of bad geography in Africa. *Review of Economics and Statistics*, 94(1):20–36, 2012. doi:10.1162/REST_a_00161.

Nathan Nunn and Leonard Wantchekon. The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101(7):3221–52, 2011. doi:10.1257/aer.101.7.3221.

Daniel Peña and Victor Yohai. A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association*, 94(446):434–445, 1999. doi:10.2307/2670164.

Daniel Peña and Victor J Yohai. The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):145–156, 1995. doi:10.1111/j.2517-6161.1995.tb02020.x.

Lawrenece I. Pettit and Karen D. S. Young. Measuring the effect of observations on Bayes factors. *Biometrika*, 77(3):455–466, 1990. doi:10.1093/biomet/77.3.455.

Garry D. A. Phillips. Recursions for the two-stage least-squares estimators. *Journal of Econometrics*, 6(1):65–77, 1977. doi:10.1016/0304-4076(77)90055-0.

Martin Ravallion. Why don't we see poverty convergence? *American Economic Review*, 102 (1):504–23, 2012. doi:10.1257/aer.102.1.504.

Lothar Reichel and William B. Gragg. Algorithm 686: FORTRAN subroutines for updating the QR decomposition. *ACM Transactions on Mathematical Software (TOMS)*, 16(4):369–377, 1990. doi:10.1145/98267.98291.

Marco Riani, Andrea Cerioli, Anthony C. Atkinson, and Domenico Perrotta. Monitoring robust regression. *Electronic Journal of Statistics*, 8(1):646–677, 2014. ISSN 1935-7524. doi:10.1214/14-EJS897.

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1): 124–127, 1950. doi:10.1214/aoms/1177729893.

Matthew S. Shotwell and Elizabeth H. Slate. Bayesian outlier detection with Dirichlet process mixtures. *Bayesian Analysis*, 6(4):665–690, 2011. doi:10.1214/11-BA625.

Christopher A. Sims. Macroeconomics and reality. *Econometrica*, pages 1–48, 1980. doi:10.2307/1912017.

Mark F. J. Steel. Model averaging and its use in economics. *Journal of Economic Literature*, 58(3):644–719, 2020. doi:10.1257/jel.20191385.

Stephen M. Stigler. The changing history of robustness. *American Statistician*, 64(4):277–281, 2010. ISSN 0003-1305. doi:10.1198/tast.2010.10159.

William H. Swallow and Farid Kianifard. Using robust scale estimates in detecting multiple outliers in linear regression. *Biometrics*, pages 545–556, 1996. doi:10.2307/2532894.

Marco Tabellini. Gifts of the immigrants, woes of the natives: lessons from the age of mass migration. *Review of Economic Studies*, 87(1):454–486, 2020. ISSN 0034-6527. doi:10.1093/restud/rdz027.

Alessandro Tarozzi, Jaikishan Desai, and Kristin Johnson. The impacts of microcredit: evidence from Ethiopia. *American Economic Journal: Applied Economics*, 7(1):54–89, 2015. doi:10.1257/app.20130475.

Lloyd N. Trefethen and David Bau. *Numerical Linear Algebra*. SIAM, 1997. doi:10.1137/1.9780898719574.

Isabella Verdinelli and Larry Wasserman. Bayesian analysis of outlier problems using the Gibbs sampler. *Statistics and Computing*, 1(2):105–117, 1991. doi:10.1007%2FBF01889985.

Alwyn Young. Consistency without inference: instrumental variables in practical application, 2020. URL `http://personal.lse.ac.uk/YoungA/CWOI.pdf`.

Appendix

Influential sets vs. single outliers: The case of the sample mean

Consider the model $\mathbf{y} = \mathbf{1}\theta + \boldsymbol{\varepsilon}$, where we are interested in the effect of potentially influential observations on the mean of the variable y, given by the estimate of θ. The $N > 3$ observations of y are contained in the vector $\mathbf{y}$. We show for this simple model and the influence function $\Delta(\mathcal{S}) = \hat{\theta}_{(\varnothing)} - \hat{\theta}_{(\mathcal{S})}$ that the influence of a influential set of two observations exceeds the sum of influences of its individual members.

Without loss of generality, assume that all observations are ordered by size and that the influential set is given by the largest observations, y_1 and y_2, so that we need to show that $\delta_1 + \delta_2 < \Delta(\{1, 2\})$. For the influence function chosen, this is equivalent to showing that

$$\hat{\theta} - \hat{\theta}_{(1)} + \hat{\theta} - \hat{\theta}_{(2)} < \hat{\theta} - \hat{\theta}_{(1,2)}, \tag{A1}$$

that is,

$$\frac{\sum_i y_i}{N} + \frac{\sum_{i \neq 1,2} y_i}{N-2} - \frac{\sum_{i \neq 1} y_i + \sum_{i \neq 2} y_i}{N-1} < 0. \tag{A2}$$

Noting that $\sum_{i \neq 1} y_i + \sum_{i \neq 2} y_i = \sum_i y_i + \sum_{i \neq 1,2} y_i$, Equation A2 can be written as

$$\left(\frac{1}{N} - \frac{1}{N-1}\right) \sum_i y_i + \left(\frac{1}{N-2} - \frac{1}{N-1}\right) \sum_{i \neq 1,2} y_i < 0. \tag{A3}$$

Since $(y_1 + y_2)/2 > \sum_{i \neq 1,2} y_i/(N-2)$, it follows that

$$\sum_i y_i = y_1 + y_2 + \sum_{i \neq 1,2} y_i > \sum_{i \neq 1,2} y_i + 2 \sum_{i \neq 1,2} y_i/(N-2), \tag{A4}$$

which implies that

$$\left(\frac{1}{N} - \frac{1}{N-1}\right) \sum_i y_i + \left(\frac{1}{N-2} - \frac{1}{N-1}\right) \sum_{i \neq 1,2} y_i <$$
$$\left(\frac{1}{N} - \frac{1}{N-1}\right) \left(\sum_{i \neq 1,2} y_i + \frac{2}{N-2} \sum_{i \neq 1,2} y_i\right) + \left(\frac{1}{N-2} - \frac{1}{N-1}\right) \sum_{i \neq 1,2} y_i = 0. \tag{A5}$$

This completes the proof. $\square$

Assessing linear approximations of influence

We use a simple Monte Carlo simulation exercise to study the precision of the linear approximation of the influence of sets of observations used by Broderick et al. (2020) and Giordano et al. (2022). We consider a simple setting based on regression with a single explanatory variable,

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}, \tag{A6}$$

Table A1: Summary statistics of the six indicators for approximation performance

	measure	mean	min	max	1^{st} dec	median	9^{th} dec
(1)	top error, absolute	0.002	0.000	0.091	0.000	0.001	0.005
(2)	top error (%)	0.054	0.003	0.685	0.017	0.042	0.105
(3)	most influential missed (%)	0.040	0.000	1.000	0.000	0.000	0.000
(4)	top five influential missed (%)	0.103	0.000	1.000	0.000	0.000	0.400
(5)	mean bias (%)	0.990	0.990	0.990	0.990	0.990	0.990
(6)	top bias (%)	0.888	0.315	0.967	0.830	0.901	0.932

where we simulate with $\beta = 1$, draw the errors from a standard normal distribution, and the observations of $\mathbf{x}$ from a $t(8)$ distribution, thus allowing for high leverage in the explanatory variable. We create samples of size $N = 100$ and compare influence estimates following Broderick et al. (2020) to the exact influence, repeating the exercise 100,000 times.

The results of our simulation exercise are presented in Table A1. We present six indicators to evaluate the performance of the approximation: (1) the absolute approximation error of the most influential observation, (2) the approximation error of the most influential observations relative to its influence, (3) the share of replications where the most influential observation is identified correctly, (4) the number of observations among the five most influential ones which are identified in the correct order, (5) the mean of the approximated influences relative to the true ones, and (6) as the maximum of the approximated influences relative to the true ones. Influence estimates of the most influential observation are smaller than the true value by 5.42% on average, which constitutes a relatively sizeable discrepancy. In the worst case, the largest influence was underestimated by 68.52%. The most influential observation was not identified in 3.97% of cases, and the top five most influential observations were not identified correctly in 10.28% of cases. At the 9th decile, three of the top five observations were identified in the correct order. The mean bias stems directly from the effect of leverage and can in principle be corrected. However, the worst case results, summarized in the maximum of the influence estimates relative to the true ones, shows that the approximation can be heavily biased in the presence of high leverage observations, with an average underestimation of 89%.

Figure A1 depicts the results of the simulation graphically for 10,000 replications. It shows the errors of the linear approximation compared to influence (absolute and scaled with respect to the mean influence and to the individual influence), leverage, and residuals. Errors appear large in comparison to the mean influence and biased downwards.

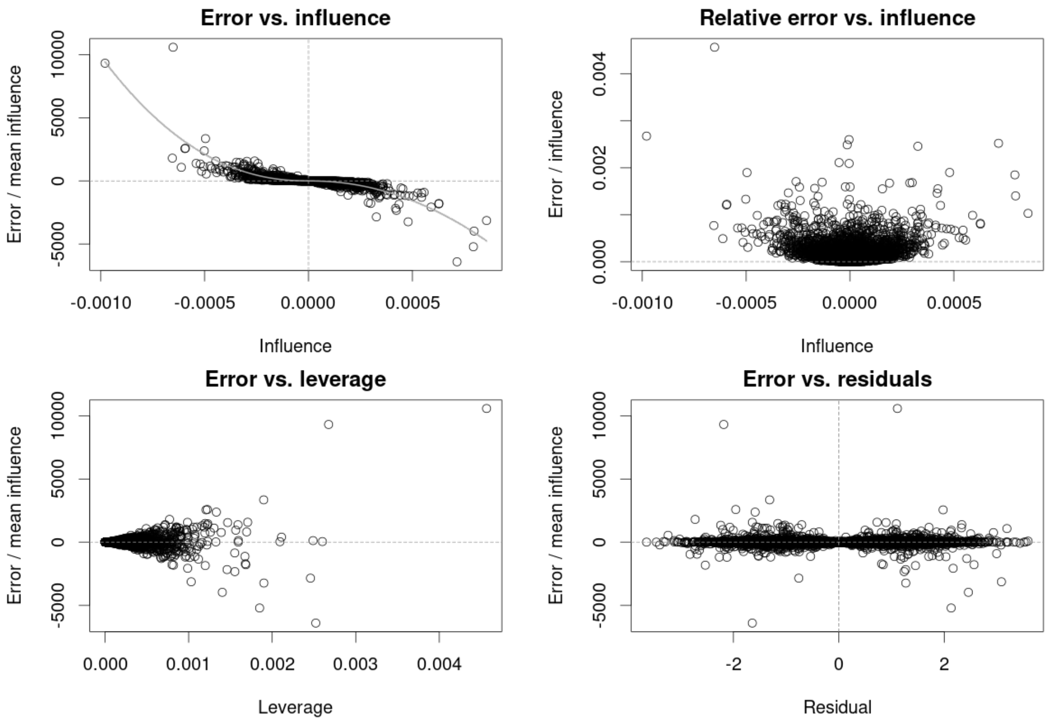


Figure A1: Errors of the linear approximation by Broderick et al. (2020) compared to influence (absolute and scaled with respect to the mean influence on the left, relative to the influence on the right), leverage (absolute and scaled), and residuals (absolute and scaled). Figures based on 10,000 Monte Carlo simulations.

The top right panel of Figure A1 shows that even relative errors display slight increases with the magnitude of the influence measure. This stems from the estimator's disregard for leverage, which is highlighted in the bottom left, where we see errors increasing with the leverage.

Influential sets in poverty convergence equations

Many empirical studies assess the patterns of cross-country convergence in living standards, as measured by GDP per capita (see e.g. Barro and Sala-i Martin, 1992; Johnson and Papageorgiou, 2020). Convergence in absolute poverty rates, however, has been examined less often. Ravallion (2012) addresses this question in a theoretical framework that is given by the combination of two stylized facts: higher average incomes tend to lead to lower poverty rates (Bourguignon, 2003), and mean incomes tend to converge across countries. Taken together, these concepts predict convergence in poverty rates. However, Ravallion (2012) does not detect poverty convergence in a sample of 89 countries, with the following model

$$T_i^{-1}(\ln H_{it} - \ln H_{it-1}) = \alpha + \beta \ln H_{it-1} + \varepsilon_{it}, \tag{A7}$$

where H_{it} denotes the poverty headcount ratio in country i and time period t, T_i is the country-specific observation period in years, and ε_{it} is an error term assumed to fulfill the standard assumptions of the linear regression model. This specification relates the annualized growth rate of the poverty headcount ratio to the log of the initial poverty headcount index. Ravallion (2012) obtains a positive, statistically insignificant estimate of β, i.e. effect of initial poverty on the subsequent growth of poverty rates.

Crespo Cuaresma et al. (2016, 2022) point out that the original, non-significant estimate of the convergence coefficient is likely due to a number of Eastern European countries exhibiting low initial poverty headcount ratios. The log-transformation in Equation A7 implies that small absolute changes translate into large growth rates in poverty headcount ratios for these economies. This makes the experience of these countries influential on the parameter estimates in Equation A7. When explicitly controlling for the poverty trajectories of Eastern European countries, there is indeed empirical evidence for cross-country convergence in poverty rates.

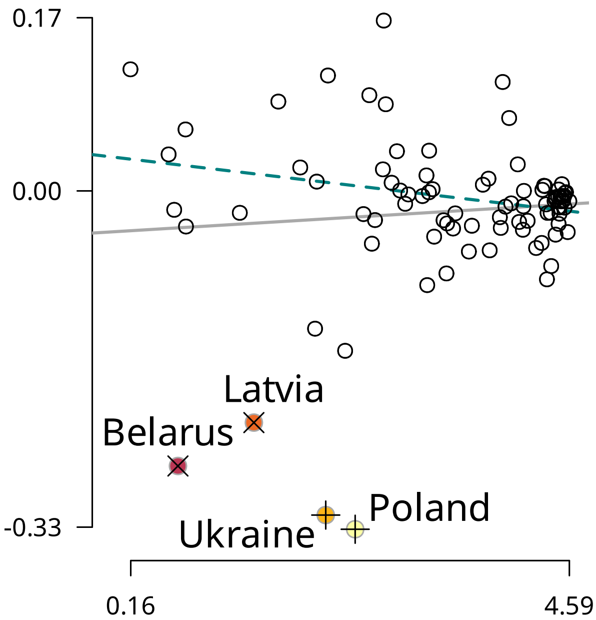


Figure A2: Data and regression line for Ravallion (2012) before (solid line) and after (dashed line) removing the influential set $\hat{\mathcal{S}}_4^*$ (highlighted via color and first crosses, then crosshairs). The horizontal axis represents the logarithm of the initial poverty headcount index; the vertical axis is annualized log differences of poverty headcount ratios.

We first revisit the problem using the same data set as used in both Ravallion (2012) and Crespo Cuaresma et al. (2016). Figure A2 presents the convergence scatter plot, where observations are colored according to their influence, as identified by Algorithm 2. In order to achieve statistically significant poverty convergence (a negative and significant estimate of β), we only need to remove a set of four countries from the dataset. Algorithm 0, however, only finds a threshold of sixteen removals in this relatively simple setting. The four influential countries are Belarus, Latvia, Ukraine, and Poland; subsequent removals would be the Russian Federation, Lithuania, Estonia, and Macedonia. This result stresses the need to take into account the different experience of Eastern European countries when analyzing cross-country poverty dynamics.

In an additional exercise, we investigate an alternative specification suggested in Crespo Cuaresma et al. (2016), which takes the form

$$T_i^{-1}(H_{it} - H_{it-1}) = \alpha + \beta H_{it-1} + \varepsilon_{it} \tag{A8}$$

where variable definitions are the same as in Equation A7. This regression specification is based on the concept of a semi-elastic relationship between poverty reduction and economic growth (see Klasen and Misselhorn, 2008). It relates changes in poverty headcount ratios to the initial level of poverty, instead of growth rates in poverty headcount ratios to the initial log level of poverty. Using this alternative specification, Crespo Cuaresma et al. (2016) find clear empirical evidence for poverty convergence in the original data by Ravallion (2012).

To assess how robust this finding is to influential sets of observations, we re-estimate the specification given by Equation A8 using an updated dataset sourced from PovCalNet.[12] Starting with the full sample of poverty headcount observations (using a poverty line of \$2 a day), we apply a number of data quality filters. First, observations that are not based on household surveys are excluded. Second, countries where the longest observation period is below ten years are excluded. Finally, when both income and consumption based poverty rates are available, consumption based data are preferred. This procedure leaves us with a sample of 124 countries. For each country, the longest time span available is used to compute annualized changes in poverty rates.

The full sample estimate of $\hat{\beta}$ is -0.019, with a standard error of 0.002. This implies significant poverty convergence. This result is relatively robust to the removal of influential sets. Algorithm 2 indicates thresholds at 26 (insignificance), 32 (sign-flip), and 42 (significant sign-flip) removals out of 124 observations. Algorithm 0 only reports a loss of significance after 79 removals. Interestingly, this issue is not only due to the influence approximation, but also due to masking, which leads to the identification of the encircled observations at the top left of the top panel in Figure A3. Algorithm 1 indicates a loss of significance only after 56 removals. As a result of this exercise, we conclude that poverty convergence appears to be a relatively robust empirical regularity.

[12] PovCalNet data can be obtained from http://iresearch.worldbank.org/PovcalNet/home.aspx.

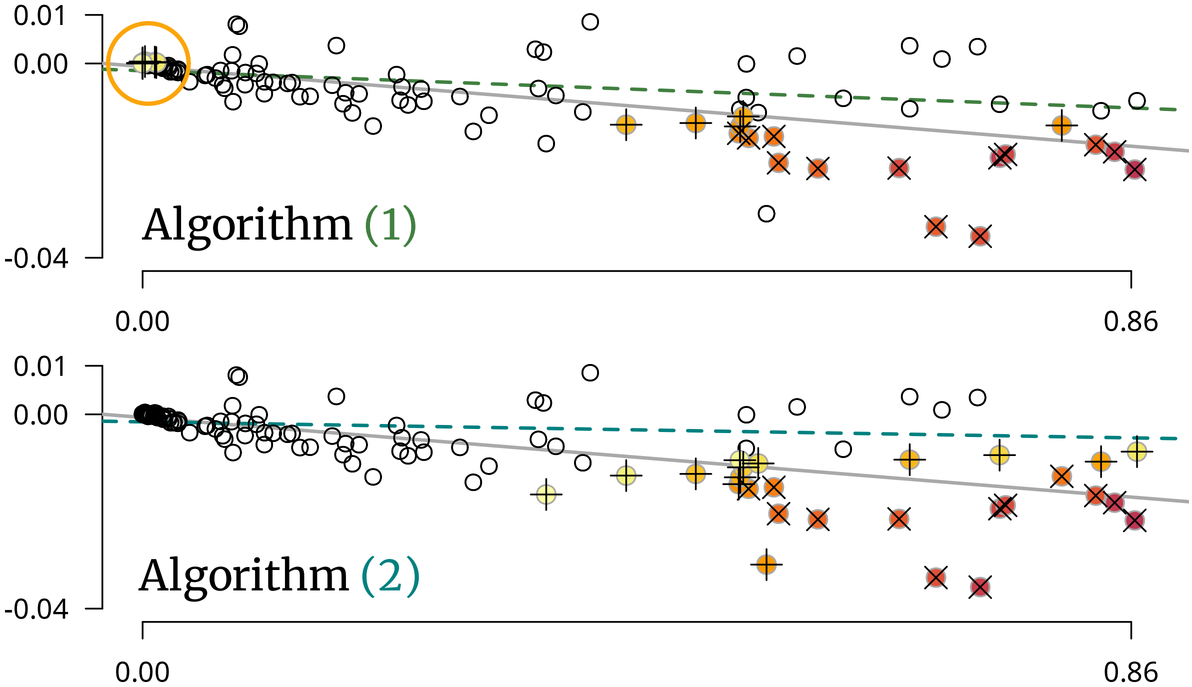


Figure A3: Data and regression line, following Crespo Cuaresma et al. (2016), before (solid line) and after (dashed line) removing the influential set $\hat{\mathcal{S}}^*_{26}$ (highlighted via color, and first crosses, then crosshairs). The top panel uses Algorithm 1, the bottom panel uses Algorithm 2.