An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?

Tamara Broderick, Ryan Giordano, and Rachael Meager *

August 14, 2020

Abstract

This paper investigates the sensitivity of common econometrics data analyses to the removal of a small fraction of the data. We demonstrate that empirical results can often depend strongly on only a small number of observations and that this sensitivity is not captured by ordinary standard errors and does not disappear asymptotically. Theory and simulations show that the main driver of the sensitivity we measure is a low signal to noise ratio in the inference problem. We provide a finite-sample method to approximately compute the number of observations or fraction of the data which has the greatest influence on a given result when removed from the sample. Our approximation is automatically computable within an established statistical framework, and we provide finite-sample error bounds for linear and instrumental variables regression. Examining several empirical applications from top economics journals we find that even 2-parameter linear regression analyses of experimental data can be highly non-robust by our metric; while some applications are robust, in others the sign of a treatment effect can be changed by dropping less than 0.01% of the sample even when standard errors are small.

*We thank Avi Feller, Jesse Shapiro, Michael Kremer, Heather Sarsons and Kirill Borusyak for helpful discussions. All mistakes are our own. Corresponding Author: Rachael Meager, reachable at r.meager@lse.ac.uk.
1 Introduction

Robustness checks in empirical economics research are motivated by the understanding that our conclusions may be sensitive to certain aspects of the analysis problem in non-obvious and undesirable ways. These checks are often focused on perturbations to the model or analytical choices, as practitioners often assume that the sensitivity of conclusions to the data itself is already captured by the standard errors. In this paper we show that empirical conclusions can depend on the realization of the finite sample data in ways which neither frequentist standard errors nor Bayesian posterior intervals capture. We propose a generally applicable metric to quantify one aspect of this sensitivity: the extent to which the conclusions of the analysis can be reversed by dropping a small fraction of the sample. We show that sensitivity of empirical results in economics to less than 0.1% of the sample, or less than 10 data points in some cases, can be substantial even when the conventional standard errors are relatively small.

The question of how our empirical conclusions would have changed had we received different data lies at the core of statistical inference. The goal of empirical research is to learn about the population of interest, not the sample at hand. The conventional frequentist standard error captures a type of data sensitivity that corresponds to a particular experiment in which the data is repeatedly randomly sampled from the population of direct interest. We might call this variability “sampling uncertainty” and quantify it with standard error estimates, which capture the variability in the estimate we can expect to see across these realizations. Yet this does not capture the only way in which the finite-sample data may have been different to what it is; in practice data may be garbled, or missing, or drawn from a population defined more broadly than exactly corresponds to the initial sample.

In practice, the goals of applied social science are quite different from simply characterising how a specific parameter estimate might change in a rarefied resampling experiment. Researchers and policymakers must extrapolate at least from data collected in the past to policy environments in the future, in a world we know to be changing; thus we may be concerned with multiple similar populations at once. Moreover, we rarely believe that our statistical model is a true description of the data-generating process, so our interest in the parameter estimates is rarely confined to the world of the model. Rather, the parameters are often proxies or guiding quantities used to capture information about a deeper or perhaps more general mechanism. We tend to more readily interpret the sign and magnitude of effects rather than their exact numerical value, because we recognize that our analyses are informing contexts that are likely to be very different than those that produced the data at hand. Particularly when studying problems that affect many communities, social scientists seek to build a robust, transportable narrative understanding of an intervention or variable that has relevance beyond the specific model
we can build or population we can access in any given moment. In such cases, the
data-dependency of our conclusions may not be fully captured by the rarefied full-
resampling experiment, and there may well be other ways in which the data might
have realized differently that would concern us if we knew our conclusions were not
robust to these perturbations.

Consider the following two examples which exemplify this dichotomy. The ran-
donized agricultural trials that motivated many early developments in 20th century
statistics are well-served by standard errors and sampling uncertainty. The quantity
measured—say, average crop yield under well-controlled conditions—is precisely the
quantity that matters for decision making. There is good reason to believe that the
differences between one field and another in an experimental setup, and between
the experimental setup and future versions of the same setup, are well-modeled by
independent sampling variation. On the other hand, consider randomized controlled
trials investigating the efficacy of microcredit in different countries (Meager, 2020).
The average change in, say, household profit, is a useful but imperfect summary of
what we care about both because the distribution of the gains across individuals
matters and because household profit or income itself often is a proxy for a broader
notion of economic wellbeing.\footnote{If the average profit were to increase slightly through one individual becoming wealthy while leaving
all others destitute, one could consider this a failed intervention. In contrast, were a single plant to
produce an entire harvest’s worth of corn, the outcome still would be positive, if strange.}
Furthermore, the difference between one country and
another, between one method of microcredit intervention and another, or between
past interventions and future interventions, cannot be plausibly fully captured by
random sampling variation. Despite these shortcomings, it is still useful to summa-
rize complex outcomes with a small number of statistics, and sampling uncertainty
still has a clear and informative meaning. Yet, being cognizant of how imperfectly
the conceptual statistical model fits with the reality of the data and generality of
our desired conclusions, practitioners might reasonably wish to know whether their
conclusions are robust to a broader variety of perturbations of the data than simply
random resampling.

In this paper we propose an approach to measure the extent to which a small
fraction of the sample has influenced the central claims or conclusions of the in-
fERENCE in practice. We examine a particular data perturbation, that of dropping
a small proportion of data points, which is both finite-sample and model-free. We
quantify the effects of such perturbations with the following three key metrics. For
a particular fraction, $\alpha$ (e.g., $\alpha = 0.001$), we define the Most Influential Set (MIS)
as the set of no more than $100\alpha\%$ of all the observations which has the greatest
influence on a given estimator when removed from the sample. We define the change
induced by dropping the MIS as the Maximum Influence Perturbation (MIP). Fi-
nally, we define the minimum proportion of points, $\alpha$, required to produce a change
of a particular size to be the Perturbation-Inducing Proportion (PIP). We argue
that, in many cases, conclusions should not be considered robust if the fraction
of the data able to reverse the claim when dropped—i.e., the PIP—is small. For
example, if the estimated sign and significance of the average increase in household
profit after increasing access to microcredit can be reversed by dropping 0.1% of
the sample (this is often fewer than 10 data points), then it becomes challenging to
argue that one should be confident that the effect will be positive even in somewhat
different populations or outside of the specific model used to estimate it. The exact
value of the MIP below which a conclusion should be considered non-robust is, of
course, subjective, and will vary from one application to another. Nevertheless, we
find that sign reversal from removing 0.01% of the data—and even from removing
1 observation from a sample of more than 15,000 data points—can and does occur
in practice in applied economics.

Although the change generated by removing any particular set of data points is
exactly computable in the finite sample for any empirical claim by cycling through
all the data removal combinations and re-running the analysis for each, this is
computationally infeasible in most cases due to the combinatorially large number
of potentially left-out sets. Fortunately, the problem of finding the MIS admits
an approximation that is easily and automatically computable for many estimators
commonly used in practice. We refer to our “Approximate MIS,” “Approximate
MIP,” and “Approximate PIP” as the AMIS, AMIP, and APIP, respectively. We
show that the approximation will tend to be most accurate when considering the
smallest fractions of the data, which is desirable as this is the most important use
case. We further show that when re-computation of the original estimator is possible
then the AMIS provides a finite-sample lower bound on the exact MIP, since in the
worst case the approximation chose a sub-optimal set of points to leave out. In every
empirical application we examined for which the perturbation is small, the AMIP
successfully delivers the change required, although to instill appropriate caution we
also show some larger proposed removal fractions for which the AMIP diverges from
the true metric actual observed change at the MIS.

We investigate what features of the model and data might cause a large MIP by
studying the AMIP using a combination of theory and simulations. In particular,
our approximation is derived from the empirical influence function, which allows us
to draw theoretical connections to robust standard errors, gross outlier robustness,
and other classical measures of robustness. We find that, for fixed \( \alpha \), the AMIP does
not disappear as the amount of data goes to infinity, in contrast with the robust
standard errors of large-\( N \) asymptotics. We also find that AMIP is typically driven
by the signal to noise ratio rather than the shape of the tail of the influence function,
in contrast to the “gross error sensitivity” of Mosteller and Tukey (1977); Huber
(1981). A further distinction between our work and that of much of classical robust
statistics is our emphasis on quantifying robustness for a particular dataset and es-

\footnote{Note that, if one can compute the MIS, then one can easily compute the MIP, and if one can compute
the MIP for all \( \alpha \), then one can easily compute the PIP.}
timator, rather than designing estimators that satisfy certain robustness properties asymptotically. Though both perspectives are certainly valuable, the distinction has some practical implications. For example, in the framework of Huber (1981); Hampel (1986), conditional means are never robust to gross outliers. Yet we show via simulation that conditional means (and thus that linear regression and other conditional mean methods) can have a numerically large (and so robust) APIP if the signal to noise ratio is sufficiently high.

Computing the AMIP for three empirical economics applications shows that the distinction between standard errors and AMIP can be considerable in practice. Even linear regression of outcomes on an intercept and a binary treatment indicator from randomized controlled trials in which the standard errors are reasonably tight can be highly non-robust. We examine an example from the cash transfers literature in development economics (Angelucci and De Giorgi, 2009) in which we find that the direct effects are quite robust while the spillover effects are highly non-robust, and demonstrate that trimming or Winsorizing the data will not eliminate this sensitivity. We examine the Oregon Medicaid Study from Finkelstein et al. (2012) and find a range of robustness results, including the finding that the instrumental variables analysis is no less robust than the intention to treat analysis. Finally, we examine seven randomized trials of microcredit aggregated in Meager (2020), and show that neither linear regressions nor Bayesian hierarchical analysis can provide robust estimates of treatment effects but that the latter does provide a robust finding on the variation in effects across studies.

Throughout the paper we work within a general local sensitivity framework which has a long history in econometrics and statistics, and encompasses papers such as Hampel (1986); Huber (1981); Horowitz and Manski (1995); Gustafson (2000); Mahajan (2006); Andrews et al. (2017). Metrics of local robustness are a complement to global sensitivity analyses such as those proposed by Leamer (1984, 1985); Sobol (2001), and Saltelli (2004) among others, and the breakdown frontiers approach of He et al. (1990); Masten and Poirier (2017) rather than a substitute. The APIP is also no substitute for tailored robustness checks designed by the researchers themselves to investigate specific concerns about sensitivity of the results to certain structures or assumptions. Passing the robustness check that we propose in this paper does not forestall the need for global sensitivity analysis, other robustness checks, or attempts to robustify analyses such as those proposed in Mosteller and Tukey (1977); Hansen and Sargent (2008), or Chen et al. (2011). We do show however in empirical applications that the AMIP provides a good approximation, even to the impact of relatively large changes to the data set.

The remainder of this paper is organized as follows. In Section 2 we define the MIS, MIP, and the PIP and their approximations, and then briefly explain implementation. We then show, using a published example from development economics, that in an actual linear regression problem from the microeconomics literature the
sign of the regression coefficient may be controlled by a single data point even in a
sample of over 15,000. Section 3 explores why such sensitivities can arise in practice,
by analyzing the nature of the approximate metric using both theory and simula-
tions. Here we show that the sensitivity captured by our metric is formally distinct
from standard errors, and distinct from yet related to the leverage in a regression
problem. Section 4 explores the sensitivity of three applications from applied mi-
croeconomics, and Section 5 discusses some implications of our findings for future
work.

2 A proposed metric

In this section we lay out the formal definition of the metric we propose, and discuss
the computation of a tractable approximation. The framework we lay out below
applies to researchers using ordinary least squares regression (OLS), instrumental
variables regression (IV), maximum likelihood estimators, generalized method of
moments estimators, Bayesian maximum a-posteriori estimators, and variational
Bayes estimators. We consider the estimation of some parameter $\theta$, which could be
the target of any of the above methods, from a single average treatment effect to a
vector of Bayesian posterior credible intervals.

Formally, we consider estimators defined as the root of an empirical estimating
equation. This encompasses most common estimators in econometrics, including
the full list above. Suppose we have $N$ data points, $x_1, \ldots, x_N$. We will let $\vec{x} =
(x_1, \ldots, x_N)$ denote the complete dataset. Let $G(\theta, x_n)$ be a function that is twice
continuously differentiable in a $D$-dimensional parameter, $\theta$. Finally, since we wish
to consider alterations to the data in a general way, we must make explicit that
data always enters the analysis problem under a weighting scheme – even using
the “unweighted” data corresponds to assigning a weight of 1 to each point. We
define scalar-valued data weights, $w_1, \ldots, w_N$, with $\vec{w} = (w_1, \ldots, w_N)$, and denote
the vector of unit weights as $\vec{1} = (1, \ldots, 1)$.

**Definition 1.** Define $\hat{\theta}(\vec{w})$ as the solution to the weighted estimating equation:

\[ \sum_{n=1}^{N} w_n G(\hat{\theta}(\vec{w}), x_n) = 0. \] (2.1)

Estimators of the form Eq. 2.1 are called “Z-estimators.” Typically, Eq. 2.1
is used with weights that are identically 1, i.e., $\vec{w} = \vec{1}$. When we write $\hat{\theta}$ without
an argument, we will take this implicitly to mean the solution $\hat{\theta} := \hat{\theta}(\vec{1})$ of the
“unweighted” analysis. The advantage of introducing weights to Eq. 2.1 is that

---

3Sometimes Eq. 2.1 is associated with “M-estimators” which optimize a smooth objective function,
since such M-estimators typically take the form of a Z-estimator. However, some Z-estimators, such
as instrumental variables regression, do not optimize any particular empirical objective function, so the
notion of Z-estimator is in fact more general.
dropping a set of datapoints can be represented by setting the weights $w_n = 0$ for $n$ in the dropped set, a fact which we will use shortly to derive our approximate robustness metrics. 

In empirical economics papers, the central claims of a research paper are focused more on important attributes of $\hat{\theta}$, such as its sign and significance, than on the specific numerical value of the parameter. In particular, frequentist analysts might be concerned if there was a small fraction of the data which:

1. Changes the sign of the parameter estimate when removed,
2. Changes the significance status of the result when removed, or
3. Generates a significant result of the opposite sign when removed.

The metric we propose encompasses these and many more changes. Bayesian analysts may not be interested in the statistical significance of their results, but may be uncomfortable if large changes in the posterior distribution can be generated by removal of a small amount of data. If the number of data points needed to change the posterior probability that a parameter is positive from 90% to 10% is small, this suggests a sensitivity to the data not captured by the posterior itself. Similarly, one might be concerned if the order of magnitude of a parameter can be altered by removing 0.1% of the sample. In the context of hierarchical models, one might be particularly concerned with the order of magnitude of hypervariances in hierarchical models, as these influence the shrinkage in the model and conclusions relating to generalizability of results. Another set of claims one could investigate focus on the cost-effectiveness of certain policy programs, for which the magnitude of effects is central.

Thus, in general, the quantity of interest is a vector of properties of the estimated quantity $\hat{\theta}$, and removing data points may have a different effect on different components of the vector: for example, one might be interested in both bounds of a confidence interval on $\hat{\theta}$, but these two bounds might well be affected differently by the removal of the same data. To formally describe what aspect of $\hat{\theta}$ is important to the analysis, we specify a more general quantity of interest that may be a function of $\hat{\theta}$, as well how much this quantity of interest would have to change in order to overturn some qualitative conclusion.

**Definition 2.** Define a quantity of interest by a continuously differentiable real-valued function $\phi(\hat{\theta}, \vec{w}) : \mathbb{R}^D \times \mathbb{R}^N \mapsto \mathbb{R}$. In a slight abuse of notation, let $\hat{\phi}(\vec{w}) := \phi(\hat{\theta}(\vec{w}), \vec{w})$. Define the “signal” $\Delta$, with $\Delta > 0$, such that if the value of $\phi$ were to increase by more than $\Delta$, some key qualitative conclusion of our analysis would be overturned. Define the “noise” $\hat{\sigma}_\psi$ to be the estimate of the standard deviation of the limiting distribution of $\sqrt{N}(\hat{\phi}(\vec{I}) - \text{plim}_{N \to \infty} \phi(\vec{I}))$ as given by the sandwich covariance estimate for $\hat{\theta}$ and the delta method. (We defer a formal definition of $\hat{\sigma}_\psi$ until Definition 5 in Section 3.1.)
It will turn out that the “signal to noise ratio,” $\Delta / \hat{\sigma}_\psi$, will be the principle determinant of robustness according to the AMIP. Note that the standard errors for $\phi(\vec{1})$ are given by $\hat{\sigma}_\psi / \sqrt{N}$, so that, if $\Delta / \hat{\sigma}_\psi$ is small but $\sqrt{N} \Delta / \hat{\sigma}_\psi$ is large, then a result can be both highly statistically significant but non-robust according to the AMIP. However, for the moment we will defer these and related results to Section 3 below.

In the simplest case we will take $\phi(\theta, \vec{w})$ to be some component of $\theta$, i.e. $\phi(\theta, \vec{w}) = \theta_d$ for some $1 \leq d \leq D$. If, for example, $\hat{\theta}_d < 0$ and our qualitative conclusions are based on the sign of $\theta_d$ being negative, then one would take $\Delta = |\hat{\theta}_d|$. In order to consider questions of statistical significance, say of $\theta_d$, we suppose that we have an estimate of the standard error, $\sigma_d(\theta, \vec{w})$, which depends smoothly on $\theta$ and $\vec{w}$, and define quantities like $\phi(\theta, \vec{w}) = \theta_d + 1.96 \sigma_d(\theta, \vec{w})$. (It is precisely for the standard error that we allow explicit weight dependence in $\phi$.) Without loss of generality, we focus only on perturbations that increase $\phi$ (that is, positive $\Delta$), since one can apply the same reasoning to $-\phi$ to investigate decreases.

### 2.1 A Tractable Approximation

This Most Influential Set (MIS) is the set of data observations, no greater than a certain size, having the greatest influence on a given inference result when removed from the given sample. In theory, the MIS is always exactly computable in practice, because one can cycle through all combinations of candidate removal points and re-run the analysis for each. But this can be both computationally demanding and time consuming, especially if the data set is large and the claims are numerous. The number of candidate sets that remove no more than $M$ out of $N$ points is $\sum_{M=1}^{M} \binom{N}{M}$, which is extremely large even for moderate $M$. To make matters worse, if one is interested in the effect of removing a fixed fraction $\alpha$ of the data, then $M$ grows as $N$ grows large. Exactly finding the MIS is thus computationally prohibitive in general.

We now derive an approximation to the influence of any given data point on a given estimate, and thus develop a practical approximate robustness metric that is computable with negligible cost.

We can represent dropping a set of datapoints by setting the weights $w_n = 0$ for $n$ in the dropped set. Specifically, if we let

$$W_\alpha := \{ \vec{w} : \text{No more than } \lfloor \alpha N \rfloor \text{ elements of } \vec{w} \text{ are } 0 \text{ and the rest are } 1 \},$$

denote the set of weights corresponding to the set of weights that drop no more than $M = \lfloor \alpha N \rfloor$ points, then the problem of finding the MIS is equivalent to solving

$$\vec{w}^{**} := \arg \max_{\vec{w} \in W_\alpha} \left( \phi(\vec{w}) - \phi(\vec{1}) \right). \tag{2.2}$$
So far, nothing has been gained, since \( \hat{\theta}(\vec{w}) \) is, in general, a non-linear function of \( \vec{w} \) and the number of elements in \( W_\alpha \) is combinatorially large. However, we can approximate the solution to Eq. 2.2 by forming an approximation to \( \phi(\vec{w}) \) that renders Eq. 2.2 easy to solve. Observe that \( \hat{\theta}(\vec{w}) \) is well-defined even for weights that are not integers. In fact, under regularity conditions given by the implicit function theorem (which we will discuss shortly), \( \hat{\theta}(\vec{w}) \) is a continuously differentiable function of \( \vec{w} \) with a closed-form expression for \( \frac{d\hat{\theta}(\vec{w})}{d\vec{w}} \). A local change to the unit weight vector \( \vec{1} \) in the direction of some \( \vec{w} \) can be conceptualized by a small movement along the line that connects \( \vec{1} \) to \( \vec{w} \), and we can approximate the behavior of \( \hat{\theta}(\vec{w}) \) along this line using its slope at \( \vec{1} \). In this sense, the change in \( \hat{\theta}(\vec{w}) \) produced by small changes to \( \vec{w} \) is captured by the derivative \( \frac{d\hat{\theta}(\vec{w})}{d\vec{w}} \) evaluated at \( \vec{w} = \vec{1} \).

Concretely, we can use the chain rule to define the following Taylor series approximation to \( \phi(\vec{w}) \).

**Definition 3.** Fix a differentiable quantity of interest, \( \phi : \mathbb{R}^D \mapsto \mathbb{R} \). The influence score of data point \( n \) for \( \phi \) at \( \vec{w} \) is

\[
\psi_n := \frac{\partial \phi(\vec{w})}{\partial w_n} \bigg|_{\vec{w}} = \frac{\partial \phi(\hat{\theta}, \vec{w})}{\partial \theta} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} \frac{\partial \hat{\theta}(\vec{w})}{\partial w_n} \bigg|_{\vec{w}} + \frac{\partial \phi(\theta, \vec{w})}{\partial w_n} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}}.
\] 

Let \( \vec{\psi} := (\psi_1, \ldots, \psi_N) \). We define the corresponding linear approximation as

\[
\phi^{\text{lin}}(\vec{w}) := \phi(\vec{1}) + \sum_{n=1}^N (w_n - 1)\psi_n \approx \phi(\vec{w}).
\]

Unless stated otherwise, we will consider influence scores evaluated at \( \vec{w} = \vec{1} \), the base case where all the data is used for the analysis.

As the name suggests, the “influence scores” are closely related to the influence function of classical robust statistics. In fact, the influence scores are the values of the empirical influence function evaluated at the observed data, as we discuss in Section 3.1.

Using Definition 3, we can define our approximate robustness metrics as follows.

**Definition 4.** For a proportion of left out points \( \alpha \), let

\[
\vec{w}^* := \arg \max_{\vec{w} \in W_M} \left( \phi^{\text{lin}}(\vec{w}) - \phi(\vec{1}) \right).
\]

Then define

\[
\hat{S}_\alpha := \{ n : w^*_n = 0 \} \quad \text{(AMIS)}
\]
\[
\hat{\Psi}_\alpha := \phi^{\text{lin}}(\vec{w}^*) - \phi(\vec{1}) \quad \text{(AMIP)}
\]
\[
\hat{\alpha}^*_\Delta := \inf \alpha : \hat{\Psi}_\alpha > \Delta, \text{ or } NA \text{ if no such } \alpha \text{ exists.} \quad \text{(APIP)}
\]

The reason to use \( \phi^{\text{lin}}(\vec{w}) \) in Definition 4 is that the optimization problem Eq. 2.5
can be written using in terms of the influence scores as

$$
\vec{w}^\ast = \arg \max_{\vec{w} \in W} \sum_{n=1}^{N} (w_n - 1)\psi_n = \arg \max_{\vec{w} \in W \atop n: w_n = 0} (-\psi_n).
$$

The solution, $\vec{w}^\ast$, is thus simply the weight vector that zeros out the $\lceil \alpha N \rceil$ most negative values of the influence scores $\psi_n$. In other words, to effect the largest positive change, one removes the observations with the most negative influence.

Similarly, we see that $\hat{\Psi}_\alpha = \sum_{n \in \hat{S}_\alpha} (-\psi_n)$ is simply minus the sum of the most negative influence scores. In contrast to the intractable MIS, computing $\hat{S}_\alpha$ is no more difficult than computing and sorting the influence scores, $\vec{\psi}$. Additionally, once the $\vec{\psi}$ have been computed and sorted, one can also easily compute $\hat{S}_\alpha$ and $\hat{\Psi}_\alpha$ for any $\alpha$, and so compute the APIP, $\hat{\theta}_\Delta^\ast$.

Of course, the computationally tractable approximation of Definition 4 is only useful if the Taylor series approximation in Eq. 2.4 is a good one. As with any linear approximation, we might expect that the approximation in Eq. 2.4 will perform well for small perturbations of the data. Specifically, when the distance from $\vec{w}$ to $\vec{1}$ is small, i.e., when relatively few entries of $\vec{w}$ are different from 1, we might expect that the error $\| \phi^{\text{lin}}(\vec{w}) - \phi(\vec{w}) \|$ is also small. Correspondingly, for large perturbations of the data, i.e. when $\| \vec{w} - \vec{1} \|$ is large, we would expect the accuracy of the approximation Eq. 2.4 to diminish. In general, the accuracy of the approximation will depend on the smoothness of the map $\vec{w} \mapsto \phi(\vec{w})$, which in turn depends on features of the optimization problem and of the data itself. We discuss the features of the approximation in more detail in Section 3 and check the quality of the approximation in our simulations and applications in Section 4.

We primarily consider the removal unit to be the individual data points themselves, since this is the level at which an assumption of independence and identical distribution is typically made. The metrics we discuss can be expanded to consider groupings or pairings of data points if appropriate. For example, one might insist on removing an equal number of control and treatment observations. When the data points are paired or grouped in ways that depend on the data itself, then the intuition we develop about the sources of non-robustness in Section 3 may no longer apply, though the linear approximation of Definition 4 can still be used.

2.1.1 Derivative of $\hat{\theta}(\vec{w})$

For Z-estimators, the derivative $d\hat{\theta}(\vec{w})/d\vec{w}^T$ is tractable to compute even though $\hat{\theta}(\vec{w})$ is only implicitly defined. While this result is a corollary of the implicit function theorem and is well-known in both statistics and mathematics (e.g., Krantz and Parks (2012); Hampel (1974)), we provide an intuitive argument for its construction here. Consider the more general problem of an estimate $\hat{\theta}(\vec{w})$ produced by solving an estimating equation that itself depends on the weights $\vec{w}$ and the parameter $\theta$.
(the data dependence is left implicit for the moment):

\[ \gamma(\hat{\theta}(\vec{w}), \vec{w}) = 0. \]  \hspace{1cm} (2.6)

As long as the map \( \theta \mapsto \gamma(\theta, \vec{w}) \) is continuously differentiable with a full-rank Jacobian matrix at \( \vec{w} \), the derivative \( d\hat{\theta}(\vec{w})/d\vec{w}^T \) exists by the implicit function theorem (Krantz and Parks, 2012, Theorem 3.3.1). Moreover, as \( \hat{\theta}(\vec{w}) \) satisfies Eq. 2.6 for all \( \vec{w} \), we can use the chain rule to compute the total derivative of Eq. 2.6 with respect to \( \vec{w} \) to derive an expression for \( d\hat{\theta}(\vec{w})/d\vec{w}^T \).

\[ \frac{d\gamma(\hat{\theta}(\vec{w}), \vec{w})}{d\vec{w}^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} = 0 \Rightarrow \]

\[ \frac{\partial \gamma(\theta, \vec{w})}{\partial \theta^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} \frac{d\hat{\theta}(\vec{w})}{d\vec{w}} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} + \frac{\partial \gamma(\theta, \vec{w})}{\partial \vec{w}^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} = 0 \Rightarrow \]

\[ \frac{d\hat{\theta}(\vec{w})}{d\vec{w}} \bigg|_{\vec{w}} = - \left( \frac{\partial \gamma(\theta, \vec{w})}{\partial \theta^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} \right)^{-1} \frac{\partial \gamma(\theta, \vec{w})}{\partial \vec{w}^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}}, \]

where the last line follows from the fact that we have already assumed for the implicit function theorem that the matrix of \( \theta \)-partial derivatives is full-rank.

In our case, we are applying Eq. 2.6 with \( \gamma(\theta, \vec{w}) = \sum_{n=1}^{N} w_n G(\theta, x_n) \), giving

\[ \frac{d\hat{\theta}(\vec{w})}{d\vec{w}^T} \bigg|_{\vec{w}} = - \left( \sum_{n=1}^{N} w_n \frac{\partial G(\theta, \vec{w})}{\partial \theta^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} \right)^{-1} \left( G(\hat{\theta}(\vec{w}), x_1), \ldots, G(\hat{\theta}(\vec{w}), x_N) \right). \]  \hspace{1cm} (2.7)

Using the chain rule thus gives a closed form expression for the influence scores:

\[ \vec{\psi} = - \left( G(\hat{\theta}, x_1)^T \vspace{1cm} \right) \left( \sum_{n=1}^{N} \frac{\partial G(\theta, \vec{w})}{\partial \theta^T} \bigg|_{\hat{\theta}(\vec{w}), \vec{w}} \right)^{-T} \frac{\partial \phi(\theta, \vec{w})}{\partial \theta} \bigg|_{\hat{\theta}, \vec{w}} + \frac{\partial \phi(\theta, \vec{w})}{\partial \vec{w}} \bigg|_{\hat{\theta}, \vec{w}}. \]  \hspace{1cm} (2.8)

The computationally intensive part of Eq. 2.8 is typically the evaluation of \( d\hat{\theta}(\vec{w})/d\vec{w}^T \), though this computation can be re-used for any number of functions \( \phi \).

The expression Eq. 2.8 is model-free in the sense that it does not rely on a comparison to a specific alternative econometric approach nor on the idea that, say, some fraction of the data is drawn from some alternative statistical model. Given the data and the estimating equation chosen by the analyst, one requires no further structure on the problem in order to compute and interpret \( d\hat{\theta}(\vec{w})/d\vec{w}^T \): it is an exact finite sample expression.

Using automatic differentiation software such as Python’s autograd library Maclaurin et al. (2015), equation Eq. 2.8 and \( \vec{\psi} \) can be computed automatically as long as the estimating equation \( \theta \mapsto G(\theta, x_n) \) is accessible to the user. We provide an open source software package for automatic implementation in R using Python under the hood, discussed in more detail in Section 2.3.
2.2 On OLS regression

Econometric analyses of causal relationships—the focus of applied microeconomics—are often conducted using linear regressions estimated via Ordinary Least Squares. The research is often done in view of decisions that must be made by policymakers in an environment which is somewhat different to the environment studied, if only because the research occurs before the decision is made and the world can change over time. It is not possible to eliminate these changes, so the typical approach is to argue that such changes are small in practice. In that case, it becomes especially important to understand whether small perturbations to the analysis at hand could yield major changes in the conclusions and policy recommendations. Even if economics researchers seek only to uncover truths about the world, a result determined by a small fraction of the sample is at greater risk of being a local phenomenon rather than a deeper truth that characterizes fundamental economic patterns across many contexts; it would be valuable to quantify this risk.

Linear regression of economic outcomes on explanatory variables is often used in economics not because the researcher believes the conditional mean dependence is truly linear but because this analysis allows one to estimate an average treatment effect or local average treatment effect in a transparent and straightforward manner. The rationale for this approach usually involves the invocation of the law of large numbers to justify the focus on the sample mean and the central limit theorem to justify the use of Gaussian confidence intervals even in the absence of a finite-sample Gaussianity assumption on the regression errors. In the remainder of this section we show that the influence of the data on the estimates of regression parameters in the finite sample can be complex even when the sample is large.

In view of this central application, consider linear mean regression of some outcome \( y = (y_1, y_2, ... y_N) \) on some explanatory variable \( x = (x_1, x_2, ... x_N) \) estimated via OLS. Suppose for simplicity that these variables have been demeaned and that relevant other factors have been partialed out of \( y \). Then, estimating the model \( y = x^T \theta + \epsilon \), the OLS estimate is

\[
\hat{\theta}(w) = \left( \sum_{n=1}^{N} w_n x_n x_n^T \right)^{-1} \sum_{n=1}^{N} w_n y_n x_n.
\]  

(2.9)

Defining \( \epsilon_n = y_n - \hat{\theta}_n x_n \) and applying Eq. 2.8 yields:

\[
\frac{d\hat{\theta}(\vec{w})}{d\vec{w}} \bigg|_{\vec{1}} = -\left( \sum_{n'=1}^{N} w_{n'} x_{n'} x_{n'}^T \right)^{-1} \sum_{n=1}^{N} w_n x_n \left( y_n - \hat{\theta}_n^T x_n \right)_{w=(1,1,\ldots,1)}
\]

(2.10)

Given the rationale above, one might expect that in large samples there ought
not to be a small number of data points that wholly determine the results of a linear
regression conducted via ordinary least squares. We now show that this intuition is
misplaced: \(\psi_n\) can be very large in practice. Consider as an example the set of seven
randomized controlled trials of expanding access to microcredit discussed by Meager
(2019). For illustrative purposes we single out the study with the largest sample
size, (Angelucci et al., 2015). This study has approximately 16,500 households. A
full treatment of all seven studies is in Section 4.3 along with tables and figures of
the results discussed below.

We consider the headline results on household business profit regressed on an
intercept and a binary variable indicating whether a household was allocated to the
treatment group or to the control group. Thus, allowing \(Y_{ik}\) to denote the profit
measured for household \(i\) in site \(k\), and \(T_{ik}\) to denote their treatment status, we first
estimate the following model via ordinary least squares:

\[
Y_{ik} = \beta_0 + \beta_1 T_{ik} + \epsilon_{ik}. \tag{2.11}
\]

We confirm the main findings of the study in estimating an average treatment
effect (ATE) of -4.55 USD PPP per 2 weeks, with a standard error of 5.88. Here,
our parameter is \(\theta = (\beta_0, \beta_1)\), and we are interested in whether we can change the
sign of \(\beta_1\) from negative to positive, so we take \(\phi(\theta) = \beta_1\). We then compute \(\psi_n\)
for each data point in the sample, which takes less than 2 seconds in R via Python
using our implementation below.

Examining \(\vec{\psi}\), we find that one household has \(\psi_n = 4.95\); removing that single
household should flip the sign if the approximation is accurate. In this case we can
manually remove the data point and re-run the regression. We indeed find that the
ATE is now 0.4 with a standard error of 3.19. Moreover, by removing 15 households
we can generate an ATE of 7.03 with a standard error of 2.55: a significant result
of the opposite sign. These results and comparable analyses for other microcredit
RCTs are presented in Section 4.3.

How is it possible for the absence of a single household to flip the sign of an
estimate that was ostensibly based on all the information from a sample of 16,500?
The fact that the original estimate was imprecise plays a role here, but it is not
decisive, and we find examples of statistically significant results that can be over-
turned by removing less than 1% of the sample in Section 4.1 and Section 4.2. One
might also suspect that this sensitivity arises because sample means are non-robust
in the Huber sense, but this is not decisive either: we find applications in which
it is necessary to remove more than 10% of the sample to change the sign, and
can simulate cases in which no amount of removal will change the sign. We also
investigate the results of fitting a Bayesian hierarchical model with a more realistic
data-generating process to the set of seven experiments and find that this does not
resolve the sensitivity either (Section 4.3). Instead, both theory and simulations
suggest that the major determinant of this sensitivity is the signal to noise ratio in
2.3 Automated implementation

The AMIS, AMIP, and APIP can be computed automatically on a range of problems, and we provide an open-source software implementation in R using Python’s auto-differentiation capacity under the hood. Since \( \vec{\psi} \) is given in closed form as a function of the estimating equation \( G(\theta, x_n) \) and its partial derivative \( \partial G(\theta, x_n)/\partial \theta \), one need only provide a Python implementation of \( G(\theta, x_n) \) in order to use Python’s autograd library (Maclaurin et al. (2015)) to automatically compute \( \vec{\psi} \) and, in turn, the AMIS and AMIP. The most computationally expensive part of computing \( \vec{\psi} \) is usually computing the partial derivative \( \sum_{n=1}^{N} \partial G(\theta, x_n)/\partial \theta \) and its inverse, but this computation is common to all functions of interest \( \phi \), and so need only be computed once to investigate a wide range of quantities of interest.

Our package is available on Github in the repository rgiordan/zaminfluence. Currently, we handle OLS and IV regression fully automatically, including weighted versions and robust or clustered standard errors. The package can handle general Z-estimators if the user provides a Python implementation of the estimating equation and functions of interest. We also provide functions to automatically rerun regressions removing the AMIS, thus allowing the user to easily check the quality of the approximation in their application.

To illustrate the ease of use of the package, consider the microcredit example from the previous section. Suppose we have a linear regression for which we care about the estimated coefficient on a variable called “treatment.” Once the researcher has run the regression using R’s \texttt{lm()} function and defined the resulting object as e.g. \texttt{reg_fit}, the user need only run the following to compute and rank the \( \vec{\psi} \), where the function of interest is the “treatment” regressor.

```r
reg_influence <- ComputeModelInfluence(reg_fit)
grad_df <- GetTargetRegressorGrads(reg_influence, "treatment")
influence_dfs <- SortAndAccumulate(grad_df)
```

The object \texttt{influence_dfs} produces a dataframe of the influence scores, the associated change in the coefficient from both individual removal and cumulative removal in rank order, the row at which one can locate the data point in the original dataset, and other metrics. If the researcher is specifically interested in how many data points she needs to remove to change the sign of the result, the significance of the result, or to generate a result of the opposite sign, she then need only run the following to produce a table of results similar to those we show in our applications.

```r
GetRegressionTargetChange(influence_dfs, "num_removed")
```
Further details and options for these functions can be found at the online repository.

3 Underlying theory and interpretation

In this section we aim to understand how and why an empirical analysis can exhibit a lack of robustness according to the AMIP metric. Particularly in light of the microcredit example above, and given the tendency in economics to rely on large-sample approximations to understand the relationship of the sample statistics to the population, it is worth understanding how and why in sample of over 10,000 data points it is possible that one single point could control the sign of the difference in means.

To understand this, we must address the nature of the automatic metric we provide, which is an approximation to the true metric. We use the approximation only because the exact combinatorial problem of computing MIP is difficult to solve in practice. It is therefore desirable to investigate the broader theoretical foundations of the approximate metric, and characterise the error of the approximation, so we turn to this first. We show that our AMIP metric is distinct from, yet fundamentally connected to, both the sampling variance of the estimator and the sensitivity usually studied in "robust" statistics. In Section 3.1 we show that AMIP is a norm on the classical influence function, which connects this metric to both sampling variance and classical robustness results (Hampel, 1986; Huber, 1983).

We then derive bounds on the performance of our approximate metric in capturing the true influence metric. In Section 3.2, we show that when the analysis can be re-run under the proposed removal combination AMIS, we provide an exact finite-sample lower bound on the MIP. In Section 3.2.2 we provide a more refined analysis we show that for OLS and IV the linear approximation’s error will be much smaller than the effect we are trying to quantify when \( \alpha \), the fraction of data removed, is small. We are able to demonstrate this scaling by providing explicit, computable finite-sample error bounds for IV and OLS regression using the framework of Giordano et al. (2019).\footnote{It is more difficult to provide explicit finite-sample error bounds for general Z-estimators, but Giordano et al. (2019) suggests that the qualitative results for IV and OLS will hold more generally.}

Having established the theoretical underpinnings and practical performance of the approximation, we then turn to understanding its behavior. We seek to understand which features of the data and inference problem determine when and how a small fraction of the sample can have a large influence on empirical conclusions according to our AMIP metric. In Section 3.3, we provide intuition for the determinants of AMIP, decomposing \( \hat{\Psi}_{\alpha} \) into contributions from the scale and shape of the distribution of influence scores. Again, our analysis is not asymptotic, although we explore the regime where \( N \) is large and \( \alpha \) stays fixed in order to contrast the
AMIP with the behavior of classical standard errors. We find that the signal to noise ratio of the inference problem is the key determinant of the sensitivity we study, and that other potentially 'suspicious' features of the problem such as having an unduly small \( N \), or fat tails, are of secondary importance. We then show that our metric is related to, yet distinct from, leverage in a linear regression and that merely having high leverage is not enough to create a lack of robustness. Finally, we discuss that non-significant results have a particular lack of robustness according to the AMIP metric as \( N \) becomes large, although in the finite sample there is no fixed relationship between significance and AMIP -robustness.

Approximations are of interest only to the extent that we can show that they actually do perform well in practice. Throughout the section, where we use asymptotic approximation in order to simplify the analysis, we do so purely for tractability and not out of any interest in large-sample behaviour. In Section 4 we show experimentally that the AMIS is always able to deliver the change it claims to deliver to the inference, even when \( \alpha \) is on the order of 10% of the sample. Although the computed AMIS provides only a lower bound on the MIP in these samples, at least this offers reassurance that the approximation identifies genuine sensitivities in the data.

### 3.1 The influence function

In this section we relate the AMIP to some selected aspects of classical robust statistics. The point of connection is the influence function of Hampel (1974, 1986) and the closely related concept of von Mises expansions of statistical functionals (Mises, 1947; Reeds, 1976). Heuristically, the influence function measures the effect on a statistic \( T \) of adding an infinitesimal amount of data with value, \( x \), to some base or reference distribution \( F \); we will denote the influence function by \( IF(x; T, F) \).

For the definition and properties of the influence function that we will use in this section, see Hampel (1986) or, for a more rigorous treatment, see Reeds (1976).

Standard results give that the asymptotic variance is measured by the squared \( L_2 \) norm, \( \int IF(x; T, F)^2dF(x) \). Additionally, classical robustness, being principally concerned with a small number of gross outliers, aims to control the \( L_\infty \) norm, \( \sup_x |IF(x; T, F)| \). As we will show, our AMIP metric can also be expressed as different norm on the influence function. As discussed in Hampel (1986), many classical results in robust statistics can be understood in terms of influence functions, and our present work is no exception.

Despite this similarity, two key differences set our work apart from much of classical robust statistics. First, we are not concerned with gross errors. We are willing to tentatively believe that all our data came from the same data generating distribution, and even accept that our model is an adequate fit for the data we saw. Given this, we wish only to find out what proportion of the data we saw informs our substantive conclusions.
Second, we are concerned with the data we saw and estimator we chose, not with the asymptotic behavior of our estimator. In contrast to much of classical robustness, we do not attempt to design estimators that will work for all \( n \) and all realizations of the data. In this sense, we are addressing robust statistics’ aim (iii) of Hampel (1986, Chapter 1.1c), “to identify and give a warning about highly influential data points.” Our emphasis on the data at hand allows us to study only the empirical influence function (evaluated at the observed empirical distribution), rather than the population influence function (evaluated at the unknown population distribution), which in turn allows us to provide the global bounds results of Section 3.2.

We will now make the connection between the AMIP and the influence function concrete. By analogy with our Eq. 2.1, define \( \hat{\theta}(F) \) to be the solution to

\[
\int G(\hat{\theta}(F), x) dF(x) = 0,
\]

for a data distribution function \( F \) (Hampel, 1986, Section 4.2c, Def. 5). For simplicity, in this section we restrict our attention to \( \phi \) that do not depend explicitly on \( \vec{w} \), and let \( \phi(F) := \phi(\hat{\theta}(F), \emptyset) \).\(^5\) Letting \( \hat{F} \) denote the empirical distribution function placing weight \( N^{-1} \) on each datapoint \( x_1, \ldots, x_N \), we thus recover that \( \hat{\theta}(\hat{F}) \) is precisely the solution to Eq. 2.1 with \( \vec{w} = \frac{1}{N} \vec{1} \), which is equivalent to the solution with \( \vec{w} = \vec{1} \), since the solution of Eq. 2.1 is invariant to the scale of the estimating equation. In the notation of Eq. 2.1, the influence function for Z-estimators is given by (Hampel, 1986, Eq. 4.2.9) as

\[
IF(x; \phi, F) = -\frac{\partial \phi(\theta, \emptyset)}{\partial \theta^T} \bigg|_{\hat{\theta}(F)} \left( \int \frac{\partial G(\theta, x)}{\partial \theta} \bigg|_{\hat{\theta}(F)} dF(x) \right)^{-1} G(\hat{\theta}(F), x).
\]

Comparing with our Eq. 2.8, we see that

\[
N \psi_n = IF(x_n; \phi, \hat{F}). \tag{3.1}
\]

That is, the linear approximation of Definition 3 is employing what (Hampel, 1986, Section 2.1e) refers to as the “empirical influence function”, which is a differential approximation to the behavior of \( \phi(\hat{\theta}(F)) \) in a neighborhood of \( \hat{F} \).

### 3.1.1 The influence function and variance

The “gross error sensitivity” of a statistical function is defined as \( \sup_x |IF(x; \phi, F)| \), or, equivalently, \( \|IF(x; \phi, F)\|_\infty \) (Hampel, 1986, Eq. 2.1.13). A central goal of classical robust statistics is to design estimators for which the gross error sensitivity is

---

\(^5\)As in ordinary calculus in Euclidian space, allowing for explicit \( F \) dependence in \( \phi \) requires only adding an extra influence function describing the dependence of \( \phi(\theta, F) \) on \( F \) with \( \theta \) held fixed. This is slightly notationally burdensome and not typical in the analysis of the influence functions M-estimators, so we omit this dependence for simplicity.
bounded. Note that the gross error sensitivity is usually defined at the population distribution, $F$, rather than the empirical distribution $\hat{F}$. In this sense, classical statistics uses the influence function at the population distribution as a guide to designing estimators with desirable properties. In contrast, we are using the influence function evaluated at the empirical distribution to approximate the effect of small changes to our dataset. We are using the same tool, but for different purposes.

The AMIP can remain finite or even quite small even if $\|IF(x; \phi, \hat{F})\|_\infty$ diverges, as long as. It turns out that the AMIP is more closely related to a different norm on the influence function. Recall that the limiting standard deviation of a statistic $T$ is given by the norm $\|IF(x; \phi, F)\|_2$ (Hampel, 1986, Eq. 2.1.8), in the sense that

$$\sqrt{N}(T(\hat{F}) - T(F)) \rightsquigarrow N(0, \|IF(x; T, F)\|_2^2).$$

Defining $\phi_0 := \operatorname{plim}_{N \to \infty} \phi(\bar{1})$, and using Eq. 3.1, we would expect to estimate the variance of the limiting distribution of $\sqrt{N}(\phi(\bar{1}) - \phi_0)$ by

$$\|IF(x; \phi, \hat{F})\|^2_2 = \int IF(x; \phi, \hat{F})^2 d\hat{F}(x)$$

$$= \frac{1}{N} \sum_{n=1}^{N} IF(x_n; \phi, \hat{F})^2$$

$$= N \sum_{n=1}^{N} \psi_n^2$$

$$= N \|\bar{\psi}\|_2^2. \quad (3.2)$$

Recall that we have already defined $\hat{\sigma}_\psi$ in Definition 2 to be the estimate of the standard deviation as given by the sandwich covariance estimate for $\hat{\theta}$ and the delta method.

**Definition 5.** Define the “sandwich covariance matrix” for $\hat{\theta}(\bar{w})$ to be \(^6\)

$$H(\bar{w}) := \sum_{n=1}^{N} w_n \left[ \frac{\partial G(\hat{\theta}(\bar{w}), x_n)}{\partial \theta} \right]_{\hat{\theta}(\bar{w})}$$

$$S(\bar{w}) := \sum_{n=1}^{N} w_n G(\hat{\theta}(\bar{w}), x_n) G(\hat{\theta}(\bar{w}), x_n)^T$$

$$\hat{\Sigma}_\theta(\bar{w}) := N H(\bar{w})^{-1} S(\bar{w}) H(\bar{w})^{-1}.$$
As with $\hat{\theta}$, take $\hat{\Sigma}_\theta := \hat{\Sigma}_\theta(\bar{\theta})$ and $\hat{\sigma}_\psi^2 := \hat{\sigma}_\psi^2(\bar{\theta})$.

In the case of linear models, $\hat{\Sigma}_\theta$ is also known as the “robust” standard error covariance, but it also applies to more general Z-estimators (see Van der Vaart (2000, Theorem 5.23, Example 5.25), or our informal discussion in Appendix B). Specifically, if we define $\theta_0 := \lim_{N \to \infty} \hat{\theta}$, then standard regularity conditions give that $\hat{\Sigma}_\theta$ consistently estimates the covariance of the limiting distribution of $\sqrt{N}(\hat{\theta} - \theta_0)$. An application of the delta method (Van der Vaart, 2000, Section 3) gives that $\hat{\sigma}_\psi^2$ consistently estimates the variance of the limiting distribution of $\sqrt{N}(\phi(\bar{\theta}) - \phi_0)$.

By plugging the quantities of Definition 5 into Eq. 2.7, we see that

$$\sum_{n=1}^N \frac{d\hat{\theta}(\bar{w})}{dw_n} \bigg|_{\bar{\theta}} \frac{d\hat{\theta}(\bar{w})^T}{dw_n} \bigg|_{\bar{\theta}} = H(\bar{\theta})^{-1}S(\bar{\theta})H(\bar{\theta})^{-1} = \frac{1}{N} \hat{\Sigma}_\theta. \tag{3.3}$$

$$= \frac{1}{N} \hat{\Sigma}_\theta. \tag{3.4}$$

By combining Eqs. 2.8, 3.2 and 3.3 (and recalling that, for the present section, we are assuming that $\phi$ does not depend explicitly on $\bar{w}$), we get that

$$\| IF(x; \phi, \hat{F}) \|_2^2 = N \sum_{n=1}^N \psi_n^2 = N \frac{\partial \phi(\theta, \emptyset)}{\partial \theta} \bigg|_{\hat{\theta}} \sum_{n=1}^N \frac{d\hat{\theta}(\bar{w})}{dw_n} \bigg|_{\bar{\theta}} \frac{d\hat{\theta}(\bar{w})^T}{dw_n} \bigg|_{\bar{\theta}} \frac{\partial \phi(\theta, \emptyset)}{\partial \theta} \bigg|_{\hat{\theta}} = \hat{\sigma}_\psi^2. \tag{3.3}$$

Thus we see that the standard delta method sandwich covariance estimator, influence function estimator, and influence score norm $N \| \psi \|_2$ coincide. The relationship between $\hat{\sigma}_\psi$ and AMIP will be explored in Section 3.3.

### 3.1.2 AMIP as a norm on the influence function

Since both the gross error sensitivity and asymptotic standard deviation can be expressed as norms on the influence function, it is interesting to ask whether the AMIP also corresponds to a norm. In this section we answer in the affirmative. Given a measure $F$ and a proportion $\alpha$, one could re-write the AMIP as the following norm applied to the empirical influence function. For an $F$-measurable function, $f(x)$, define $\bar{f}_F = \int f(x) dF(x)$, and

$$\| f \|_{F,\alpha} := \sup_{S: F(S) < \alpha} \left( - \int_S (f(x) - \bar{f}_F) dF(x) \right).$$

Because $f(x) - \bar{f}$ must have both negative and positive regions, $\| f \|_{F,\alpha} \geq 0$ and $\| f \|_{F,\alpha} = 0$ implies that $f(x) = 0$ $F$-almost surely. By linearity of the integral,
\|f\|_{F,\alpha} = a \|f\|_{F,\alpha}$. And by the definition of the supremum,

\[
\|f + g\|_{F,\alpha} = \sup_{S: F(S) < \alpha} \left( -\int_S \left( f(x) - \bar{f}_F + g(x) - \bar{g}_F \right) dF(x) \right) \\
\leq \sup_{S: F(S) < \alpha} \left( -\int_S (f(x) - \bar{f}_F) dF(x) \right) + \sup_{S: F(S) < \alpha} \left( -\int_S (g(x) - \bar{g}_F) dF(x) \right) \\
= \|f\|_{F,\alpha} + \|g\|_{F,\alpha},
\]

so \(\|\cdot\|_{F,\alpha}\) satisfies the triangle inequality, and is a well-defined norm. Using Eq. 3.1, and comparing with Definition 4, we can see that

\[
\hat{\Psi}_\alpha = \|IF(x; \phi, \tilde{F})\|_{F,\alpha}.
\]

Consequently, like the asymptotic standard deviation and gross-error sensitivity, our AMIP can be understood as a norm on the influence function.

### 3.2 Bounds on approximation error

Our ability to understand the quality of the approximation \(\hat{\Psi}_\alpha\) for \(\Psi_\alpha\) depends on the specific estimation problem. We first discuss the fact that for any problem for which estimation is relatively cheap, one can always check the quality of the approximation directly by removing the AMIS and re-computing \(\hat{\theta}(\hat{S}_\alpha)\), providing a finite sample lower bound on the true maximal achievable change. Next, we use the framework of Giordano et al. (2019) to show that show that for the case of IV and OLS regression, we can write closed-form, computable, finite-sample bounds for the approximation error and analyze its asymptotic behavior under certain regularity assumptions. Finally we discuss certain cases in which the approximation is likely to perform poorly.

#### 3.2.1 Lower bound via re-computation

Although it is typically too expensive to run the combinatorially large number of estimation problems required to compute the MIS, one can often easily re-run the analysis for a small handful of candidate sets of left-out points. This re-computation is possible to do in most cases for linear regressions or linear GMM systems. In such cases, one can simply check the approximation by manually removing the data points in the AMIS and re-running the analysis to see if the desired change is achieved in practice. That is, for desired change \(\Delta\), one can first compute \(\hat{\alpha}^*_\Delta\) and the corresponding weight \(\vec{w}_\Delta\), re-run the estimation procedure to exactly compute \(\hat{\theta}(\vec{w}_\Delta)\), and see whether \(\phi(\vec{w}_\Delta) = \phi(\hat{\theta}(\vec{w}_\Delta), \vec{w}_\Delta) > \phi(\vec{1}) + \Delta\). If it is, then non-robustness has been shown irrespective of the approximation.

In fact, recalling the definitions of the exact MIS weight \(\vec{w}^{**}\) in Eq. 2.2 and the approximate AMIS weight \(\vec{w}^*\) in Eq. 2.5, we have that \(\phi(\vec{w}^*) - \phi(\vec{1})\) is a lower bound
for \( \phi(\tilde{w}^*) - \phi(\tilde{\theta}) \) for any \( \alpha \) since, by definition,
\[
\phi(\tilde{w}^*) - \phi(\tilde{\theta}) = \max_{\tilde{w} \in W_\alpha} \left( \phi(\tilde{w}) - \phi(\tilde{\theta}) \right) \geq \phi(\tilde{w}^*) - \phi(\tilde{\theta}).
\]
Consequently, the approximation \( \hat{S}_\alpha \) can conclusively demonstrate non-robustness.

Without more fine-grained control on the error in the approximation of \( \phi_{\text{lin}}(\tilde{w}) \) for \( \phi(\tilde{\theta}(\tilde{w})) \), \( \hat{S}_\alpha \) cannot prove robustness. We now turn to discussion of such fine-grained control for OLS and IV regression.

### 3.2.2 Finite sample bounds on the approximation error

We now derive explicit error bounds for the influence function approximation to IV and OLS regression. Though it may not be easy to produce explicit error bounds for general Z-estimators, the scaling as \( \alpha \to 0 \) is expected to be similar given the results in Giordano et al. (2019).

Suppose we observe regressors, \( x_n \in \mathbb{R}^D \), instruments \( z_n \in \mathbb{R}^D \), and responses, \( y_n \in \mathbb{R} \), for \( n = 1, \ldots, N \), and use the estimating equation
\[
G(\theta, \tilde{w}) := \frac{1}{N} \sum_{n=1}^{N} w_n (y_n - \beta^T x_n) z_n.
\]
Eq. 3.5 encompasses both IV regression and OLS, the latter by taking \( z_n = x_n \).

Using this observation, we will prove bounds for IV regression, from which bounds for OLS will follow automatically.

Let \( \hat{\beta}(\tilde{w}) \) be the linear regression estimator for the weight vector \( \tilde{w} \), with \( \hat{\beta} = \hat{\beta}(\tilde{w}) \). Define the linear approximation to \( \hat{\beta}(\tilde{w}) \) as
\[
\hat{\beta}_{\text{lin}}(\tilde{w}) := \hat{\beta} + \frac{d\hat{\beta}(\tilde{w})}{d\tilde{w}^T} \bigg|_{\tilde{w}} (\tilde{w} - \tilde{\theta}).
\]
We will directly bound the error \( \|\hat{\beta}(\tilde{w}) - \hat{\beta}_{\text{lin}}(\tilde{w})\|_2 \). Analogous bounds will then hold for summary functions \( \phi \) that are sufficiently smooth in \( \tilde{w} \) and Lipschitz with constant \( L_\phi \) in \( \beta \) on some convex, compact set containing \( \hat{\beta}_{\text{lin}}(\tilde{w}) \) and \( \hat{\beta}(\tilde{w}) \), since
\[
|\phi(\hat{\beta}_{\text{lin}}(\tilde{w}), \tilde{w}) - \phi(\hat{\beta}(\tilde{w}), \tilde{w})| \leq L_\phi \left( \|\hat{\beta}(\tilde{w}) - \hat{\beta}_{\text{lin}}(\tilde{w})\|_2 \right)
\]
For matrices, let \( \|\cdot\|_{\text{op}} \) denote the operator norm (i.e., the largest eigenvalue).
For a particular \( \tilde{w} \) whose entries are either 1 or 0, we will express the error of \( \|\hat{\beta}(\tilde{w}) - \hat{\beta}_{\text{lin}}(\tilde{w})\|_2 \) in terms of the following quantities.

**Definition 6.** Given a \( \tilde{w} \) whose entries are either 1 or 0 and the linear regression
estimate $\hat{\beta} = \hat{\beta}(\bar{1})$, define the following quantities.

$$
\mathcal{N} := \{n : w_n = 0\}
$$

$$
\hat{\epsilon}_n := y_n - \hat{\beta}^T x_n
$$

$$
C_{op} := \left\| \left( \frac{1}{N} \sum_{n=1}^{N} z_n x_n^T \right)^{-1} \right\|_{op}
$$

$$
\xi_1 := \left\| \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} z_n x_n^T \right\|_2
$$

$$
\xi_2 := \left\| \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} z_n \hat{\epsilon}_n \right\|_2
$$

$$
\Delta_{lin} := \left\| \hat{\beta}^{lin}(\bar{w}) - \hat{\beta} \right\|_2.
$$

Note that all the quantities in Definition 6 can be easily computed from $\hat{\beta}$ and $\bar{w}$ without without running any additional regressions (i.e., without performing any additional matrix inverses).

Using Definition 6, the analysis of Giordano et al. (2019, Section 6) can control finite-sample error of $\hat{\beta}^{lin}(\bar{w})$ as an approximation to $\hat{\beta}(\bar{w})$. The result is most conveniently expressed in terms of the following quantities which are all functions of quantities in Definition 6.

**Definition 7.**

$$
\alpha := \frac{|\mathcal{N}|}{N}
$$

$$
\hat{C}_{op} := \frac{3}{2} C_{op}
$$

$$
\hat{\beta} := \frac{\Delta_{lin} + 2\alpha^2 \hat{C}_{op}^2 \xi_1 \xi_2}{1 - 2\alpha^2 \hat{C}_{op}^2 \xi_1^2}.
$$

Our key regularity assumption is as follows.

**Assumption 1.** Assume that $\alpha C_{op} \xi_1 \leq \frac{1}{3}$.

Importantly, Assumption 1 can be easily checked for any particular dataset and weight vector using the quantities in Definition 6. Assumption 1 essentially requires that the regressors in the left-out set are not too large relative to the average regressors, particularly as $\alpha \to 0$. For example, when the regressors and instruments are bounded (i.e., $\max_{n \in |\mathcal{N}|} \|x_n\|_2$ and $\max_{n \in |\mathcal{N}|} \|z_n\|_2$ are bounded for all $N$), then $\alpha C_{op} \xi_1 \leq \frac{1}{3}$ will always hold for sufficiently small $\alpha$.

In Appendix D, we prove Lemma 1, which is a technical results essentially stating that, under Assumption 1, Assumptions 1-5 of Giordano et al. (2019) are satisfied with particular constants. The precise statement of Lemma 1 is slightly notationally
burdensome, and so we do not state it here. However, from Lemma 1, the following theorem follows immediately.

**Theorem 1.** For a particular dataset and any weight vector \( \vec{w} \) such that Assumption 1 is satisfied,

\[
\left\| \hat{\beta}(\vec{w}) - \hat{\beta}^{\text{lin}}(\vec{w}) \right\|_2 \leq \alpha^2 \hat{C}_{op}^2 \xi_1 (\xi_2 + \mathcal{B} \xi_1) \quad \text{and} \quad \left\| \hat{\beta}(\vec{w}) - \hat{\beta} \right\|_2 \leq \alpha C_{op} (\xi_2 + \mathcal{B} \xi_1).
\]

**Proof.** Both results depend on Lemma 1 of Appendix D, which asserts that all of the assumptions for applying the bounds of Giordano et al. (2019) are satisfied. The first result then follows from Giordano et al. (2019, Corollary 3) and Giordano et al. (2019, Lemma 10). The second result follows from Giordano et al. (2019, Corollary 2). \( \square \)

Note that Theorem 1 is not an asymptotic result. On the contrary, Theorem 1 holds deterministically for any finite \( N \) and \( \alpha \), in terms of constants and conditions that can be easily computed from \( \hat{\beta} \) using the quantities in Definition 6 and Definition 7. Further, Theorem 1 applies to any \( \vec{w} \), including \( \vec{w} \) that are chosen adversarially (as in the present paper), as long as \( \vec{w} \) satisfies the regularity condition Assumption 1.

When \( \alpha \to 0 \) as \( N \to \infty \) and Assumption 1 holds uniformly, we have the following consistency result.

**Corollary 1.** For each \( N \), choose any sequence of weight vectors \( \vec{w}(N) \) with corresponding \( \alpha(N) \) such that \( \lim_{N \to \infty} \alpha(N) = 0 \). Each of the constants in Definition 6 thus depend on \( N \). Assume that each the \( C_{op}(N) \), \( \xi_1(N) \), and \( \xi_2(N) \) are eventually uniformly bounded as \( N \to \infty \). Assume that there exists an \( N_0 \) such that, for all \( N > N_0 \), Assumption 1 holds. Then, as \( N \to \infty \),

\[
\left\| \hat{\beta}(\vec{w}) - \hat{\beta} \right\|_2 = O(\alpha) \to 0 \\
\left\| \hat{\beta}(\vec{w}) - \hat{\beta}^{\text{lin}}(\vec{w}) \right\|_2 = O(\alpha^2) \to 0.
\]

In particular, Corollary 1 states that the error \( \left\| \hat{\beta}(\vec{w}) - \hat{\beta}^{\text{lin}}(\vec{w}) \right\|_2 \) goes to zero at a faster rate than the actual change, \( \left\| \hat{\beta}(\vec{w}) - \hat{\beta} \right\|_2 \). For sequences of \( \alpha \) that do not go to zero, the error \( \left\| \hat{\beta}^{\text{lin}}(\vec{w}) - \hat{\beta}(\vec{w}) \right\|_2 \) does not go to zero in general. However, the bounds of Theorem 1 still apply, and the error will be expected to be small for small \( \alpha \).

### 3.2.3 Cases to Approach with Caution

In virtually all cases we examine in our applications in section Section 4, we manually re-run the analysis without the data points in the removal set \( \hat{S}_\alpha \), and find that the
change suggested by the approximation is achieved in practice. This is true even when the required removal is up to 10%. However, there are some notable cases in which it is advisable to approach the AMIP with caution which we discuss in the next section.

First and foremost, the impact of large changes to the data set on the order of removing 30% of the data or more are unlikely to be well approximated by $\hat{\Psi}_\alpha$. As with any approximation resting on a Taylor series expansion, the quality of the approximation rests on the similarity between the true function and its linearization. We do not view this as a major drawback for the method, since analysts are typically concerned with robustness to small changes; it is hardly concerning that the AMIP and MIP differ if one removes one third of the sample.

A notable case in which we often know that the AMIP is incorrect is when it reports that there is no feasible way to effect a particular change in the sample, i.e., when $\hat{\alpha}_\Delta = NA$ in Definition 4. Consider, for example, a sign change on the treatment effect estimated in a randomized controlled trial. Unless there is complete separation in the sample across the treatment and control groups (i.e. every outcome for every individual in the treatment group lies either above or below those of every individual in the control group) this is not true. When the linear approximation cannot find any set of points whose removal could reverse the desired result, this is usually reflective of the fact that $\hat{\Psi}_\alpha$ is not a good approximation for $\Psi_\alpha$ for the large $\alpha$ that would be required to produce such a large change in $\Psi_\alpha$. However, since, for small $\alpha$, we expect that $\hat{\Psi}_\alpha \approx \Psi_\alpha$, so we may confidently assert that there is no small $\alpha$ that could reverse our result.

Another case in which the $\hat{\Psi}_\alpha$ can fail to approximate the truth well is that of bounded parameters whose true value lies near the boundary. Because the linearized function of the parameter $\theta^*_d$ is not bounded, the approximation can diverge from the truth when the true $\theta$ and original $\hat{\theta}$ are bounded, and will tend to do so near the boundary itself. For linear regression on a continuous outcome this issue is not a concern for the regression coefficients, but it is a concern with the estimation of associated variance parameters. This is also a concern in many Bayesian models which use bounds on parameters to improve estimation, and in hierarchical models in which variances at different levels are used for shrinkage as well as inference and for which the hypervariances could be quite small in practice. It can help to linearize the problem using unconstrained parameters (e.g. linearly approximating the log variance rather than variance). However, as we discuss in Section 4.3, simply transforming to an unconstrained space is still not necessarily guaranteed to produce accurate approximations near the boundary.

### 3.3 What determines AMIP robustness?

In this section we explore some potential sources of the sensitivity captured by the AMIP. We use a combination of simulations and theory to investigate the behavior
of the AMIP in response to different features of the sample or underlying inference problem. Our discussion is general for Z-estimators, but our simulations will focus on the case of OLS regression.

We find that the sensitivity captured by the AMIP is primarily driven by the signal-to-noise ratio (SNR) of the inference problem, by which we mean the ratio $\Delta / \hat{\sigma}_\psi$, where $\Delta$ is the size of a perturbation that will overturn some substantive conclusion, as in Definition 2. While the AMIP is also partly determined by shape of the underlying distribution of the data, the shape itself primarily matters in terms of its influence on the SNR; for example, fat-tailed data is an issue but only because it produces a large noise scale, not because it produces large individual outliers.

We further show that the problem is not due to having insufficiently large $N$ and not due to conditional means being Huber non-robust in general. Finally we offer a corollary relating to the interpretation of statistical significance—that, for fixed $\alpha$ as $N$ grows large, insignificance is always non-robust according to the AMIP.

### 3.3.1 Monte Carlo Set-Up

We set up the following Monte Carlo experiment which we will use to explore the range of behavior of the $\hat{\Psi}_\alpha$ in the case of simple linear regression. Consider regressing an outcome variable $y_n$ on a single explanatory variable $x_n$ where each of the variables has been de-meaned. Typically one estimates the scalar $\beta$ parameter that governs the relationship between $x_n$ and the conditional mean of $y_n$ using the following linear regression model estimated using the OLS estimator:

$$ y_n = \beta x_n + \epsilon_n $$

(3.6)

In each case we create $N$ draws of $x_n$ from some chosen distribution which has mean 0 and some chosen scale $\sigma_X = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x_n^2}$. We also take $N$ draws of $\epsilon_n$ from some chosen distribution which has mean 0 and some chosen scale $\sigma_\epsilon = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \epsilon_n^2}$. We produce $y_n$ according to the linear model itself with some chosen $\beta$.

Consider for illustration drawing $x_n$ and $\epsilon_n$ from independent Gaussians where the variance of $x_n$ is much larger than $\epsilon_n$; we expect the results of the regression to be robust in this case. Specifically, just as an example, consider the case where $\sigma_X = 12.3$, $\sigma_\epsilon = 1.2$, and $\beta = -1$. When $N = 10,000$, producing the $y_n$ draws from this model and estimating $\hat{\beta}$ produces a statistically significant and indeed highly accurate estimate (we got -1.00071 with a standard error of 0.00195). But is it robust according to the $\hat{\Psi}_\alpha$? Yes: Even adversarially removing 1% of the sample $(\alpha = 0.01)$ produces a change in $\hat{\beta}$ of only about 1/3rd of 1% (a 0.0038 unit change).

In this case, according to the linear approximation, there is no proportion of the data that one can remove in order to change the sign at all, nor the significance (nor, of course, both). This is not literally true—unless there is separation in the data, there will always be some way to change the significance and/or sign by removing
data points—but in this case the local linear approximation cannot find a way. This makes sense because the map \( \alpha \mapsto \hat{\Psi}_\alpha \) is concave (\( \hat{\Psi}_\alpha \) is the cumulative sum of sorted values), so \( \hat{\Psi}_1 \) is at most 100 times \( \hat{\Psi}_{0.01} \). But even 100 \( \cdot \hat{\Psi}_{0.01} \) in this case would only be a change of 0.38 units, not sufficient to change the sign nor even the significance in this case. The graph in Figure 1 illustrates the situation: as \( \alpha \) increases along the horizontal axis, it is possible to change the value of \( \hat{\beta} \) in both the positive and negative directions as shown by the two red lines (the blue interval is the 95% confidence interval). But in this case, \( \hat{\beta} \) changes so little that there is no local change to the sample which can produce even a change in the statistical significance of the estimate.

One might imagine that the classical Gaussian regression experiment could not be made sensitive since it corresponds to the idealized set-up in which the asymptotics of the CLT hold literally in the finite sample. However, we will show in Section 3.3.4 that even in this case a low signal to noise ratio in the data can lead to sufficiently large \( \hat{\Psi}_\alpha \) as to overturn the sign and significance of the estimates in practice.

### 3.3.2 Theoretical Decomposition

To understand the properties of the AMIP and how we should expect different features of the data set to influence it more generally, we decompose \( \hat{\Psi}_\alpha \) into contributions from the scale, \( \hat{\sigma}_\psi \), and from the shape of the distribution of the \( \psi_n \). Our emphasis will be on the expected behavior as \( N \) grows and \( \alpha \) remains fixed but small, particularly when contrasting with the behavior of standard error estimates, though our analysis is not otherwise asymptotic.

Recall from Section 3.1.1 the standard error estimate \( \hat{\sigma}_\psi^2 = N \| \tilde{\psi} \|_2^2 \). Note also that

\[
\sum_{n=1}^{N} \psi_n = 0,
\]

since, by definition of \( \hat{\theta} \) in Eq. 2.1, \( \sum_{n=1}^{N} G(\hat{\theta}, x_n) = 0 \), and so \( \sum_{n=1}^{N} \left. \frac{\partial \hat{\theta}(\tilde{w})}{\partial w_n} \right|_{\tilde{w}} = 0 \). \(^7\)

To understand the asymptotic behavior of \( \hat{\Psi}_\alpha \), we define the vector \( \tilde{\gamma} = (\gamma_1, \ldots, \gamma_N) \),

\(^7\)The influence scores in fact sum to zero for any statistical functional that does not depend explicitly on \( N \).
where

\[ \gamma_n := \frac{N \psi_n}{\hat{\sigma}_\psi} \Rightarrow \]

\[ \frac{1}{N} \sum_{n=1}^{N} \gamma_n = 0 \]

and

\[ \frac{1}{N} \sum_{n=1}^{N} \gamma_n^2 = 1 \]

so that the \( \gamma_n \) can be thought of as mean zero, unit variance observations. Then, for any subset \( S \subseteq [1, \ldots, N] \) of indices,

\[ \sum_{n \in S} \psi_n = \hat{\sigma}_\psi \frac{1}{N} \sum_{n \in S} \gamma_n. \]

Since \( \hat{\sigma}_\psi \) does not depend on \( S \), and the sorting of \( \gamma_n \) is the same as the sorting for \( \psi_n \), it follows that the MIP is given by

\[ \Gamma_{\alpha} := -\alpha \frac{1}{|S_\alpha|} \sum_{n \in S_\alpha} \gamma_n \]

\[ \hat{\Psi}_\alpha = \hat{\sigma}_\psi \Gamma_{\alpha}. \]

In this form, we see that AMIP depends on the behavior of two quantities. First, we have \( \hat{\sigma}_\psi \), which converges to \( \sigma_\psi \), the same quantity that determines the standard errors. The scale \( \hat{\sigma}_\psi \) does not, of course, depend on the shape of the distribution of the \( \psi_n \), only on their \( L_2 \) norm. All such shape dependence is captured in \( \Gamma_{\alpha} \). We now use this decomposition to analyze the role of the data shape, scale, and sample size.

Note that, for small fixed \( \alpha \) as \( N \to \infty \), we expect both \( \Gamma_{\alpha} \) and \( \hat{\sigma}_\psi \) to converge to non-zero quantities. We have already discussed how \( \hat{\sigma}_\psi \to \sigma_\psi > 0 \). Conditional on \( \hat{\sigma}_\psi \), which converges to a constant, the term \( -\frac{1}{|S_\alpha|} \sum_{n \in S_\alpha} \gamma_n \) is a sample average of \( |\hat{\mathcal{S}}_{\alpha}| \) observations, all of which typically with the same sign since \( \alpha \ll 1/2 \). Consequently, we expect that \( \Gamma_{\alpha} \) converges to a non-zero constant as well.

### 3.3.3 The role of data shape

Data shape, and particularly the presence of fat tails, seems potentially central to the presence of highly-influential data points for inference using Z-estimators. However, we now use the decomposition of Eq. 3.7 to show that the values that the shape parameter \( \Gamma_{\alpha} \) can take are strictly bounded at a moderate value. The shape can thus only have a limited influence on the robustness that we measure. In contrast, the scale \( \hat{\sigma}_\psi \) can be arbitrarily large, and does not disappear as \( N \) goes to infinity.

It turns out that \( \Gamma_{\alpha} \) is bounded by construction. Recall from the previous section
that \( \frac{1}{N} \sum_{n=1}^{N} \gamma_n = 0 \) and \( \frac{1}{N} \sum_{n=1}^{N} \gamma_n^2 = 1 \). From these constraints it follows by a Lagrange multiplier argument (see Appendix C) that

\[
|\Gamma_\alpha| \leq \sqrt{\alpha(1-\alpha)},
\]

(3.8)

the worst-case value of \( \sqrt{\alpha(1-\alpha)} \) being obtained when all of the \( \gamma_n \) for \( n \in S_\alpha \) are equal to one another. Note that Eq. 3.8 holds exactly in finite sample; it is not an asymptotic result.

An illustrative example of two different \( \gamma \) distributions is shown in Figure 3. The left plot shows a heavy-tailed distribution, and the right plot shows a light-tailed distribution, each scaled to have unit sample variance. We are leaving out \( M = 4 \) datapoints, which are shown in red. The heavy tailed distribution has one large negative entry, but the other entries are correspondingly smaller due to the constraint \( \frac{1}{N} \sum_{n=1}^{N} \gamma_n^2 = 1 \). In contrast, the light-tailed distribution has a moderate number of fairly large entries. The result is that, summed over the four left-out points, the total \( \Gamma_\alpha \) is larger for the light tailed distribution than for the heavy-tailed distribution.

For \( \alpha = 0.01 \), corresponding to removing 1% of datapoints, Eq. 3.8 gives to \( \Gamma_\alpha \leq 0.0995 \). However, the worst-case bound may not reflect a very realistic distribution of influence scores. To investigate the range of \( \Gamma_\alpha \) that might be encountered in practice, we generate Monte Carlo estimates of \( \Gamma_\alpha \) using 1,000,000 samples of \( \gamma_n \) with tail behavior given by several common distributions. The results are shown in Table 3.3.3. As we can see, the archetypal heavy-tailed Cauchy distribution, which has no moments due to its extreme kurtosis, in fact has quite a small shape parameter \( \Gamma_\alpha \). This is because the worst-case of \( \Gamma_\alpha \) occurs when most of the left-out values all take the same value, and the Cauchy tends to be dominated by one or two extreme values after rescaling. Of course, the estimate \( \hat{\sigma}_\psi \) is extremely large for the Cauchy draws (\( \hat{\sigma}_\psi \) diverges), and it is for this reason that Cauchy influence scores result in a few points with very high impact. In contrast, the left tails of the re-centered exponential distribution, which is bounded below, have the largest value of \( \Gamma_\alpha \), due to the fact that most of the left-out set has a similar value. When the exponential is flipped so that the left tail is unbounded, \( \Gamma_\alpha \) assumes a much smaller value. In total, the range of \( \Gamma_\alpha \) exhibited in Table 3.3.3 is relatively small, and, for this reason, we argue that the adversarial sensitivity is largely determined by the scale \( \hat{\sigma}_\psi \) rather than the shape \( \Gamma_\alpha \).

This demonstrates the perhaps counterintuitive result that conditional on a given distributional scale, a Cauchy with no moments can have a smaller \( \Gamma_\alpha \) than a Gaussian or even a proper uniform distribution. Intuitively, the reason a uniform distribution of the of the same scale as the Cauchy produces more extreme MIP values is because there are more data points to remove far from the median. Fat tails seem pathological because they produce distributions without moments, but in fact they are pathological because they tend to produce large scale in the data,
Table 1: Monte Carlo estimates of $\Gamma_\alpha$ with $\alpha = 0.01$ and 1,000,000 draws.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\Gamma_\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Worst case</td>
<td>0.0995</td>
</tr>
<tr>
<td>2 Normal</td>
<td>0.0266</td>
</tr>
<tr>
<td>3 Exponential</td>
<td>0.0460</td>
</tr>
<tr>
<td>4 Flipped exp</td>
<td>0.0099</td>
</tr>
<tr>
<td>5 T(10)</td>
<td>0.0300</td>
</tr>
<tr>
<td>6 T(3)</td>
<td>0.0408</td>
</tr>
<tr>
<td>7 T(2)</td>
<td>0.0361</td>
</tr>
<tr>
<td>8 Cauchy</td>
<td>0.0022</td>
</tr>
<tr>
<td>9 Uniform</td>
<td>0.0172</td>
</tr>
<tr>
<td>10 Binary(0.01)</td>
<td>0.0299</td>
</tr>
<tr>
<td>11 Binary(0.1)</td>
<td>0.0299</td>
</tr>
<tr>
<td>12 Binary(0.5)</td>
<td>0.0301</td>
</tr>
</tbody>
</table>

not because they have heavy tails.

3.3.4 The role of the SNR

Recall that we will consider our estimate to be non-robust if there exists a sufficiently small $\alpha$ capable of producing a change of size at least $\Delta$, as in Definition 2. In terms of our decomposition, our estimate is non robust if there exists a sufficiently small $\alpha$ such that

$$\Delta \leq \hat{\sigma}_\psi \Gamma_\alpha \iff \frac{\Delta}{\hat{\sigma}_\psi} \leq \Gamma_\alpha.$$ 

The ratio $\Delta/\hat{\sigma}_\psi$, which we are calling the SNR, does not depend on $\alpha$. As we have just seen, the curve $\alpha \mapsto \Gamma_\alpha$ is bounded above by $\sqrt{\alpha(1-\alpha)}$, and does not vary widely across typical probability distributions. Consequently, the principal determinant of whether the AMIP will detect non-robustness is whether the SNR is sufficiently small.

As a demonstration of the key role of the SNR, we now examine the sensitivity of linear regression results for a wide range of $\sigma_X$ and $\sigma_\epsilon$ for the Gaussian case. As applied economists often face situations in which the variation in $X$ is smaller than the variation in $\epsilon$ – especially in randomized trials which have binary $x_n$ or quasi-experiments with small changes in $x_n$ – we consider a range of $\sigma_X \in (0, 4]$ and a range of $\sigma_\epsilon \in (0, 12.5]$. We keep $\beta$ fixed at $-1$.

The results of these simulations are shown as a heatmap grid in figure Figure 2, where a darker colour indicates a highly sensitive analysis, in which only a small proportion of the sample needs to be removed to effect these three major changes: generating the opposite sign, changing the significance, and generating a significant result of the opposite sign. The grey areas indicate those in which the linear approximation returns $\hat{\alpha}_\Delta^* = N.A$, areas for which it is not possible to generate these
changes with the linear approximation.

We see results ranging from cases where removing less than 5% of the sample can generate a significant conclusion of the opposite sign to cases where even changing the significance alone is impossible within our approximate framework. Overall the results show that the larger is $\sigma_e$ relative to $\sigma_X$, conditional on $\beta$, the more sensitive are the regression results to minor perturbations of the data set.

3.3.5 The role of sample size

One might imagine that as more data is collected it would eventually become impossible to find a small fraction of the data that strongly influences the results. However, we now show that the uncertainty measured by $\hat{\Psi}_\alpha$ does not disappear as $N \to \infty$ if $\alpha$ is a fixed proportion of the sample. To see this, first recall that the usual Gaussian asymptotic approach to quantifying the uncertainty around an estimate from a sufficiently regular estimator delivers the (approximate) two-sided 95% interval

$$\hat{\theta} - \theta_0 \in \left\{ -1.96 \frac{\hat{\sigma}_\psi}{\sqrt{N}}, 1.96 \frac{\hat{\sigma}_\psi}{\sqrt{N}} \right\}. \quad (3.9)$$

This interval shrinks to zero as $N$ gets very large.

Let us compare the sampling uncertainty with the range suggested by the AMIP with $\alpha = 0.01$. Let $\Gamma_{0.01}^+$ denote the shape parameter for a target function $\phi(\theta, \bar{w}) = \theta$, and $\Gamma_{0.01}^-$ the shape parameter for $\phi(\theta, \bar{w}) = -\theta$. Applying the decomposition from Eq. 3.7, we find that the uncertainty implied by the AMIP is

$$\hat{\theta}(\bar{w}) - \hat{\theta} \in \{ -\Gamma_{0.01}^- \hat{\sigma}_\psi, \Gamma_{0.01}^+ \hat{\sigma}_\psi \}. \quad (3.10)$$

This width of the interval implied by AMIP does not shrink to zero as $N \to \infty$. The width of both the standard error and adversarial sensitivity intervals is determined by the scale $\hat{\sigma}_\psi$, but the former is scaled by $1.96/\sqrt{N}$, and the latter by $\Gamma_{0.01}^+$ and $\Gamma_{0.01}^-$. As a result, there may be genuine sensitivities to small perturbations of the data set even in very large samples, and moreover these sensitivities would be masked by simply examining the standard errors or applying the usual approximate Gaussian uncertainty interval for the estimator of choice. Taken together these results demonstrate again that the uncertainty measured by $\hat{\Psi}_\alpha$ is not captured by standard errors and which, unlike standard errors, one should not expect to disappear as $N \to \infty$, as long as $\alpha$ is a fixed proportion of the sample.

3.3.6 Relationship to Huber Robustness

From the perspective of classical robust statistics, one might think that the sensitivity exhibited in the microcredit example from Section 2.2 arises because averages are not robust statistics in the Huber sense of “gross error sensitivity” (Huber, 1983; Kim and White, 2004). But this is not the case. As our simulation results show, it is
quite possible for a mean to be highly robust to the small perturbations considered by the $\hat{\Psi}_\alpha$. In the heatmap grid in figure Figure 2 we see many cases in which one must remove 30% or more of the sample in order to simply change the significance of the estimate, and many more cases in which there is no local change that could effect changes to sign or significance detected by our approximation. Gross error sensitivity concerns the influence that a huge change in the value of a small fraction of the data set would have on the inference; MIP sensitivity concerns the influence that dropping a small fraction of the data set would have on the inference.

### 3.3.7 A Corollary on Statistical Significance

The fact that the sensitivity captured by the $\hat{\Psi}_\alpha$ does not disappear asymptotically, unlike the standard errors, prompts a corollary insight concerning the interpretation of statistical significance in large samples. As $N$ grows large, the standard errors shrink towards zero, but the effect that the largest influential fraction has on the estimate according to the AMIP does not. Thus, when a result is insignificant, for a given $\hat{\Psi}_\alpha$ there will generally be some $N$ for which we are able to move the parameter estimate far enough away from zero that the result becomes significant. In this sense, non-significance itself exhibits an inherent non-robustness according to the $\hat{\Psi}_\alpha$. Significant results do not exhibit this inherent lack of robustness except in the case that the point estimates are small relative to the size of $\hat{\Psi}_\alpha$ so that the SNR is small. This observation is not particular to the AMIP, and would apply equally to any robustness measure that did not disappear asymptotically.

### 3.4 AMIP and leverage

We now show that the $\psi$ metric and thus the influence curve is fundamentally linked to the concept of leverage in linear regression. Leverage is often loosely defined as the extent to which results hinge on a single data point, but is formally defined as the influence of the observed outcome data $Y$ on the predicted values $\hat{Y}$. For linear regression, leverage is given by the relevant diagonal entry of the “hat matrix” $H = X(X'X)^{-1}X'$. This sometimes called the projection matrix, because $\hat{Y} = HY$, and so this is the projection of the $Y$ variable into the space spanned by the columns of the $X$ matrix. The leverage of a single observation $y_n$ on its own fitted value $\hat{y}_n$ is the $n$th diagonal entry of this hat matrix, denoted

$$h_{nn} = x_n'(X'X)^{-1}x_n.$$  \hfill (3.11)

In the context of multivariate regression, we have

$$\frac{d\hat{\beta}}{dw_n} = (X'X)^{-1}x_n\epsilon_n.$$  \hfill (3.12)
Taking $\phi(\beta, \tilde{w}) = \hat{y}_n$, the influence score is

$$
\psi_n = \frac{d\hat{y}}{dw_n} = \left( \frac{d\hat{\beta}}{dw_n} \right)' x_n = \epsilon_n x_n' (X'X)^{-1} x_n. 
$$

(3.13)

This expression formalizes the conceptual link made in Chatterjee et al. (1986) between influence, leverage and large values of $\epsilon_n$, which are there considered a kind of outlier. As they note, observing that a particular data point is an outlier in the error space is not enough to guarantee that this point has an outsize influence on the results. Observing that a data point has a high leverage score relative to other points is also not enough to guarantee that a given data point has an outsize influence on the results. As the re-expression of the influence function above shows, what matters is the confluence of these two events.

4 Applied experiments

4.1 Cash transfers

We now apply our techniques to examine the robustness of the main analysis from Angelucci and De Giorgi (2009), one of the flagship studies showing the impact of cash transfers on ineligible households, also known as “spillover effects”. The authors employ a randomized controlled trial to study the impact of Progresa, a social program giving cash gifts to eligible poor households in Mexico. The randomization occurs at the village level, and the authors can therefore study the impact on the non-eligible “non-poor” households in the villages randomly selected to receive Progresa, as well as studying the main effect on the poor households.

The main results of the paper show that there are strong positive impacts of

---

8It is even easier to see this in the univariate regression case. In this case leverage is

$$
h_{nn} = \frac{x_n^2}{\sum_{n=1}^{N} x_n^2}.
$$

(3.14)

It turns out that the influence function $\psi$ is proportional to the leverage function, but the proportionality is governed by an additional aspect of the data: the relative scale of the $\epsilon$ and $X$ variables. To see this, continuing within a univariate regression context, we make the following augmentation and substitution:

$$
\left. \frac{d\hat{\beta}}{dw_n} \right|_1 = \frac{-x_n \epsilon_n}{\sum_{n=1}^{N} x_n^2} = \frac{-x_n^2 \epsilon_n}{x_n \sum_{n=1}^{N} x_n^2} = \frac{\epsilon_n}{x_n} h_{nn}.
$$

---
Progresa both on eligible poor households and on the non-eligible households on total household consumption measured as an index (see Table 1 of Angelucci and De Giorgi (2009)). The authors study three different time periods separately to detect any change in the impact between the short and long term. They further condition on a large set of variables (household poverty index, land size, head of household gender, age, whether speak indigenous language, literacy; at the locality level, poverty index and number of households) to ensure a comparable comparison between households in the treatment and control villages. In this case these controls are important as the effects on the “nonpoor” households are significant at the 5% level when the controls are included, while they are only significant at the 10% level in a simple regression on a dummy for treatment status.

The full data for the paper is available on the website of the *American Economic Review* due to the open-data policies of the journal and the authors. We are able to successfully replicate the results of their analysis with the controls and without, and we proceed with the controls in this exercise as their preferred specification. Therefore, for the time periods indexed as t=8,9,10 in the dataset provided and employing K control variables $X_{it}$ for household $i$ in period $t$, we run the following regression:

$$C_{indit} = \beta_0 + \beta_1 \text{treat}_{poor} + \beta_2 \text{treat}_{nonpoor} + \sum_{k=1}^{K} \beta_{2+k} X_{ijk} + \epsilon_{it}$$

We are able to exactly replicate the results of Table 1 in Angelucci and De Giorgi (2009). We then perform a sensitivity analysis to assess how many data points one need remove to change the sign, the significance, or to generate a significant result of the opposite sign to that found in the full sample. The results are shown in Table 3. We find that the inferences on the direct effects are robust, but the inferences on the indirect effects are quite sensitive. For the analysis of the poor (“treatp”), one typically needs to remove around 5% and even up to 10% of the sample to effect these changes. For the analysis of the nonpoor (“treatnp”), one need remove less than 0.5% of the data to make these large changes, and in one case removing only 5 data points in a sample of approximately 10,000 households can change the sign. These differential robustness results likely reflect the differential signal to noise ratio that one might expect comparing direct and indirect effects, and suggest the merits of a cautious approach to the spillovers literature more broadly.

We can check the quality of the approximation by examining the extent to which these results, generated using the linear approximation to the combinatorial problem, reflect the reality of the analysis by re-running with the implicated households removed. The results are shown in Table 3 and indicate that even when 10% of the sample is suggested for removal, the linear approximation correctly identifies a combination of data points that can make these changes to the conclusions of the study.
The authors of this study were aware of the potential for highly influential points in the tails of the consumption data. To deal with this, they removed what they considered outliers before running the analysis. The authors truncated the sample by deleting households with consumption indices greater than 10,000 units from the analysis. While there are very few such households, we have verified that they do make a major difference to the results of Table 1, and the resulting inferences on the nonpoor households are even more sensitive without the truncation. However, as our results show, even the truncated data set remains non-robust in the case of the analysis of the nonpoor households. This shows that truncation (or, more formally, Winsorization) does not necessarily produce robust inference.

4.2 The Oregon medicaid experiment

In early 2008, the state of Oregon opened a waiting list for new enrollments in its Medicaid program for low-income adults. It then drew names by lottery from the 90,000 people who signed up, and those who won the lottery could sign up for Medicaid along with any of their household members, creating a randomization into treatment and control groups at the household level. The Finkelstein et al 2012 study measures outcomes 1 year after the treatment group received Medicaid. About 25% of the treatment group did indeed have Medicaid coverage by the end of the trial (a somewhat surprisingly low compliance rate). The main analysis both investigates treatment assignment as treatment itself (“intent to treat” or ITT analysis) and uses treatment assignment as an instrumental variable for takeup of insurance coverage (“local average treatment effect” or LATE analysis).

The outcomes of interest are grouped into health care utilization indicators, compliance with recommended preventative care, financial strain related to medical expenditures, and health outcomes themselves both physical and mental. The data sample we have access to consists of survey responders (N = 23,741), some of which are from the same households. All regressions include household size fixed effects, survey wave fixed effects, and the interaction between the two. All standard errors are clustered on the household and all regressions are weighted using survey weights defined by the variable “weight_12m”. At present we consider Table 9 of Finkelstein et al 2012, which shows the impact on health outcomes themselves. We have access to the following variables, presented in Table 9 in the following order: a binary indicator of a self-reported measure of health being good/very good/excellent (not fair or poor), a binary indicator on self-reported health not being poor, a binary indicator on health being about the same or improving over the last six months, the number of days of good physical health in the past 30 days, the number of days on which poor physical or mental health did not impair usual activities, the number of days mental health was good in the past 30 days, and indicator on not being depressed in last two weeks. We successfully replicate Table 9 of Finkelstein et al 2012 exactly.
Consider first the ITT analysis conducted with a variety of control variables. We now perform a sensitivity analysis to assess how many data points one need remove to change the sign, the significance, and produce a significant results of the opposite sign in this analysis. The results, with all fixed effects and controls included and clustering at the household level, are shown in Table 4. The graphics showing the proportional removal results are shown in figure Figure 5. The table shows that there are variables for which the sign can be changed by removing 0.05% of the data or less, or around 100 data points in a sample of approximately 22,000. It typically requires the removal of around 1% of the data to produce a significant result of the opposite sign, although some of the results are more robust, and require almost 5% removal to be reversed.

We again check the quality of the approximation and find that it always delivers the change it claims to deliver. The inference on the impact of Medicaid on these seven outcomes produced by re-running the regressions manually removing the data points in $\hat{S}_{\Delta}^*$ is shown in Table 6. Here the linear approximation still reliably uncovers combinations of data points that can deliver the proposed changes even when 5% of the sample is removed. Again we note this is a lower bound on the sensitivity as it may have been possible to discover a smaller set of points achieving the same change in results had one solved the full combinatorial optimization problem of finding the MIS.

Consider now the LATE analysis which is performed using the two-stage-least-squares estimator. The results, with all fixed effects and controls included and clustering at the household level, are shown in Table 5. The results of re-running the analysis with the data points in $\hat{S}_{\Delta}^*$ removed are shown in Table 7. The robustness of the IV results is very similar to the ITT case. This is perhaps surprising given recent work suggesting that instrumental variables might often be somewhat unreliable relative to other methods (Young (2019)). However, in this case the instrument was very strong and the effects were large, which may be why the IV’s robustness is comparable to the OLS in this case.

### 4.3 Seven RCTs of microcredit

We now consider the data from seven randomized controlled trials of expanding access to microcredit selected by Meager (2019) for evidence aggregation. In this section we will examine the sensitivity of the comparison of the means in the treatment and control groups from each of the studies separately, as well as the sensitivity of a Bayesian hierarchical model fit to the entire data set for evidence aggregation purposes. We specifically focus on the more complex tailored mixture model from Meager (2020) as one might hope that this model’s somewhat more realistic data-generating process alleviates the sensitivity to some extent; unfortunately we will show that this is not the case for the average effects, although we do find that the estimated variance in effects across studies is robust.
4.3.1 Linear Regression Analysis

Each of the microcredit studies, conducted across seven different countries, selected certain communities to randomly receive greater access to microcredit either by building a branch, or a combination of a branch and some active outreach, or by randomly selecting borrowers among those who applied. The selected studies are: Angelucci et al. (2015), Attanasio et al. (2015), Augsburg et al. (2015), Banerjee et al. (2015), Crépon et al. (2015), Karlan and Zinman (2011), and Tarozzi et al. (2015), six of which were published in a special issue of the American Economics Journal: Applied Economics on microcredit, and are commonly considered to represent the most solid evidence-base for understanding the impact of microcredit.

We first follow the original studies and Meager (2019) in analyzing the impact of this access itself as the treatment of interest. The studies range in their sample sizes from around 1,000 households in Mongolia Attanasio et al. (2015) to around 16,500 households in Mexico Angelucci et al. (2015). We consider the headline results on household business profit regressed on an intercept and a binary variable indicating whether a household was allocated to the treatment group or to the control group. Thus, allowing $Y_{ik}$ to denote the profit measured for household $i$ in site $k$, and $T_{ik}$ to denote their treatment status, we estimate the following model via ordinary least squares:

$$Y_{ik} = \beta_0 + \beta T_{ik} + \epsilon_{ik}$$ (4.1)

It is hard to imagine a more straightforward analysis. This regression model compares the means in the treatment and control groups and estimates the difference as $\hat{\beta}$. With a sample size of 1,000 or 16,500, the intuition one is often prompted to develop in one’s econometric education is that these two means - and thus their difference - are likely estimated with a high degree of accuracy. We follow Meager (2019) in omitting the control variables or fixed effects from the regressions in order to examine the robustness of this fundamental procedure, but in principle this should make no difference to the estimate $\hat{\beta}$ and indeed it does not (Meager, 2019).9

We now analyze the robustness of the microcredit results for household profit to the removal of a small fraction of the data. The results in table Table 8 show that in this set of studies removing a small number of data points can change the sign, the significance, and generate a result of the opposite sign which would be deemed significant at the 5% level. As discussed in Section 2.2, the largest study is among the most sensitive – a single data point in the 16,561 households in Mexico determines the sign. To change both the sign and significance – that is, to turn Mexico’s noisy negative result into a "strong" positive result – one need remove only

---

9While it may in principle make a difference to the inference on $\beta$ by affecting the standard errors, it turns out that in these studies the additional covariates make very little difference to the standard errors. We also do not cluster the standard errors at the community level for the same reason; the results are not substantially changed. Running the regression above in each of the 7 studies delivers almost identical results to the preferred specification, as it should if intrachannel correlations are weak and covariates are not strongly predictive of household profit.
15 data points, less than 0.001% of the sample. Mongolia, the smallest study in terms of sample size, is among the most robust in terms of sign changes, as it takes 2% of the sample to change the sign. The Philippines has the largest standard error yet is the most robust in terms of our ability to generate a significant result of the opposite sign, which would require the removal of more than 5% of the sample. Illustrative graphs of these removal results are presented in Figure 6.

It may seem unsurprising that some of these results are highly non-robust, as they are all non-significant; yet some of these non-significant results are more robust than some of the significant results in the Cash Transfers and Oregon Medicaid examples (consider the Philippines study, for example). It might also seem natural to implicate the fat tails of the household profit variable, a phenomenon well-documented in Meager (2020) and known to reduce the efficiency of the mean as an estimator of location. We have shown in Section 3 that tail shape generally matters most in its role in generating outcome data with very large scale, so in the remainder of the section we focus on scale issues. We discuss the issue of the shape of the profit tail and address related concerns about disparate removal from treatment and control groups further in Appendix E.

To explore this issue of data scale further, we analyze another outcome reported in five of the seven studies: household consumption spending on temptation goods such as alcohol, chocolate and cigarettes. This variable had a much smaller scale than household profits, and was estimated with greatest precision of all six variables investigated by Meager (2019). The results are shown in Table 10. While these analyses are somewhat more robust than the profit analyses, the difference is not large. Mongolia and India are now faring quite well, but it is still possible to change the sign of many of the estimates by removing around 1% of each sample or less, and to generate a "strong" result of the opposite sign by removing 2% of the data or less. Understanding that the relative scale is what matters rather than the absolute scale, it is not surprising that the temptation analysis is only marginally more robust than the profit analysis: the temptation analysis also had by far the smallest estimated $\hat{\beta}$ of all variables.

We again test how well the linear approximation on which the AMIP is based is performing in this case. The results of manually re-running the profit analysis with the largest influence points removed are shown in table Table 9 and for temptation goods in table Table 11. While there may still be some other combinations of fewer points that lead to greater changes that one might find if one ran the full combinatorial analysis, the AMIS at least offers a finite sample lower bound on the sensitivity. In this case, the lower bound on the sensitivity is large (because the percentage of points removed is very small), indicating that the true sensitivity is high.
4.3.2 Bayesian Hierarchical Tailored Mixture Model

We now turn to the analytical approach from Meager (2020) in the hopes that the results may be more robust. One might imagine that regularization from the priors as well as from the shrinkage performed by the hierarchical component of the model ought to lessen the sensitivity of the results with respect to the data. Moreover, the approach was specifically motivated by the desire to capture important features of the data-generating process such as fatter tails.

The model is fit to all the data from the seven RCTs, and each of the seven distributions is modeled using a spike at zero and two lognormal tail distributions, one for the positive realizations of profit and one for the negative realizations. Within the model, microcredit can affect the proportion of data assigned to each of these three components as well as affecting the location and scale of the lognormal tails. There is a hierarchical shrinkage element to the model for each parameter, and particular interest is given to the hypervariances of the treatment effects, because these capture the variation in effects across studies which provides information about the transportability of results across settings. In this section we will focus on the treatment effect of microcredit on the location parameters of the tail distributions.

The models in the original paper were fit via Hamiltonian Monte Carlo (HMC) in the software package Stan (Carpenter et al., 2017). It is possible to compute the AMIP for HMC, or for any Markov Chain Monte Carlo method using the tools of Bayesian local robustness (Gustafson, 2000), but the sensitivity of simulation-based estimators is beyond the scope of this paper. However, there are ways to estimate Bayesian posteriors via optimization; perhaps the most notable among these are Variational Bayes (VB) techniques (Blei et al., 2016). To proceed in this section therefore we fit the model using a variant of Automatic Differentiation Variational Inference (ADVI) (Kucukelbir et al., 2017), and apply the sensitivity analysis to the ADVI estimates. Specifically, we apply the version of ADVI described in Giordano et al. (2018, Section 5.2). Since the posterior uncertainty estimates of vanilla ADVI are notoriously inaccurate, we estimated posterior uncertainty using linear response covariances, again from Giordano et al. (2018, Section 5.2). We verified that the fit and overall results of the ADVI and MCMC estimation are similar.

Within the VB implementation of the model, we compute the influence score $\psi_n$ for each data point. Consider the effect of microcredit on the location parameter of the positive tail of profit, where the majority of the data in the samples is located. For each of the seven countries, Figure 7 shows the changes that one can make by removing data points successively from the most influential point downwards, considering the removal of up to 2% of the total sample for each of the individual treatment effects on the means of the positive tails (the results are similar for the negative tails). The bars in this case show the central 95% posterior interval. While a Bayesian analyst is not generally concerned with statistical significance, our analyst might be concerned that these marginal posteriors show a large sensitivity to
the removal of a small fraction of the sample. Parameter value regions which had a posterior mass of less than 2% in the original analysis can end up with a posterior mass of 50% when we remove only 1% of the sample.

The hypermeans and hypervariances for these location parameters are somewhat more robust than the country-specific parameters, and the hypervariances even more than the hypermeans. Figure 8 shows the change in the posterior marginals for the these parameters for both the negative and positive tails. The hypervariances are roughly twice as robust in the sense that their magnitude can be altered only half as much as the hypermeans can. This is quite interesting since from a conventional robust statistics viewpoint, variances tend to be highly non-robust statistics. Our finding of a relatively robust hypervariance suggests that while it is possible to change the individual treatment effects by changing certain points, there must be different points removed for the different effects. It is easy to move each of the treatment effects around individually, but it is harder to move all of them in different directions away from each other by removing only a small part of the sample. Thus, the aggregate results are more stable, and the aggregate conclusions are more robust than the individual conclusions of the papers on which the analysis is based.

Checking the accuracy of the AMIP in this case uncovers an instance in which the approximation fails to represent the underlying situation well in practice. We take the function of interest to be the log of the hypervariance and attempted make it as negative as possible. That is, we tried to find a set of points which, once removed, would make the hypervariance close to zero. Though the AMIP suggests that such a change is possible, Figure 9 shows that the approximation actual estimator diverge as the hypervariance approaches the boundary at zero.

5 Conclusion

There are different ways of quantifying the dependence between the sample and the conclusions of statistical inference. While the idea of data dependence has become synonymous with standard errors in frequentist statistics, the notions are only equivalent under a certain paradigm that considers only a classical resampling exercise for the purpose of evaluating a specific parameter within a given model. Yet there are other ways of conceiving of and quantifying the dependence of empirical results on the sample data, and researchers might find a different story under a different paradigm both in theory and in practice.

The goal of inference is not to learn about the sample but to learn about the population, and moreover in practice researchers often wish to conceive of that population in a relatively broad manner. If minor alterations to the sample can generate major changes in the inference, and we know the environment in which we do economics is changing all the time, we ought to be less confident that we have

---

10 And indeed, she is.
learned something fundamental about the population we seek to understand, for whom we ultimately seek to make policy.

Working within an established general framework for local sensitivity, we have provided a computable metric, the AMIP, to quantify robustness to small perturbations of data in finite sample. The AMIP can be computed automatically for many common analysis methods and the quality of the approximation can usually be checked in practice. We find that common methods for data analysis in economics can and often do display important sensitivities to small fractions of the data. These sensitivities are present in all three of the empirical applications we studied here, even when the inferential procedures are straightforward, although to differing extents. In many cases, removing less than 1% of the sample data can generate a strong, statistically significant result of the opposite sign to that claimed in the study; in some cases, removing 1 data point can change the sign and removing 15 data points will generate significance on this new sign.

However, we do find more robust cases in certain applications, suggesting that applied economic analyses do differ in their sensitivities according to the AMIP. Those analyses which are more robust are not obviously identifiable in terms of having smaller standard errors or fatter tails or inferior significance status, which confirms our theoretical findings that the sensitivity measured by AMIP is not captured by conventional metrics of sampling uncertainty. Instead, the sensitivity we analyze appears to be largely driven by the relative scale of the outcome and regressors as well as the size of their covariation, a set of factors succinctly captured in the signal to noise ratio for the given statistical problem. This intuition is also distinct from that which underlies general Huber or robust statistics: indeed, we have shown that conditional means can perform well if the signal to noise ratio is high.

This paper has proposed a new robustness metric which is both relevant to economic analysis in theory and which uncovers important sensitivities in practice. It now seems desirable to develop new statistical methods to address the presence of this kind of sensitivity, and moreover, to develop these methods in view of the actual goals and uses of economics research rather than relying on a classical resampling paradigm that bears little resemblance to the practice of applied social science.
## 6 Figures

Table 2: Number of Data Points Affecting Conclusions of Cash Transfers Analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>N</th>
<th>Change Sign</th>
<th>Change Significance</th>
<th>Change Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatnp, t=10</td>
<td>4,266</td>
<td>30</td>
<td>3</td>
<td>101</td>
</tr>
<tr>
<td>treatnp, t=9</td>
<td>3,838</td>
<td>21</td>
<td>3</td>
<td>53</td>
</tr>
<tr>
<td>treatnp, t=8</td>
<td>4,624</td>
<td>5</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>treatp, t=10</td>
<td>10,518</td>
<td>697</td>
<td>435</td>
<td>1,049</td>
</tr>
<tr>
<td>treatp, t=9</td>
<td>9,630</td>
<td>345</td>
<td>146</td>
<td>653</td>
</tr>
<tr>
<td>treatp, t=8</td>
<td>10,936</td>
<td>252</td>
<td>83</td>
<td>520</td>
</tr>
</tbody>
</table>

Table 3: Manual Re-Runs Of The Cash Transfers Analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Beta (SE)</th>
<th>Re-run for sign</th>
<th>Re-run for significance</th>
<th>Re-run for both</th>
</tr>
</thead>
<tbody>
<tr>
<td>treatnp, t=10</td>
<td>21.49 (9.41)</td>
<td>-0.57 (6.75)</td>
<td>16.26 (8.93)</td>
<td>-11.96 (6.84)</td>
</tr>
<tr>
<td>treatnp, t=9</td>
<td>22.85 (10)</td>
<td>-0.37 (7.54)</td>
<td>16.51 (9.11)</td>
<td>-11.83 (7.11)</td>
</tr>
<tr>
<td>treatnp, t=8</td>
<td>-5.44 (7.13)</td>
<td>0.26 (6.41)</td>
<td>-12.97 (6.78)</td>
<td>10.44 (5.86)</td>
</tr>
<tr>
<td>treatp, t=10</td>
<td>33.86 (4.47)</td>
<td>-2.56 (3.54)</td>
<td>4.81 (3.68)</td>
<td>-11.34 (3.99)</td>
</tr>
<tr>
<td>treatp, t=9</td>
<td>27.92 (5.77)</td>
<td>-1.38 (4.41)</td>
<td>7.08 (4.55)</td>
<td>-11.26 (4.78)</td>
</tr>
<tr>
<td>treatp, t=8</td>
<td>17.31 (4.59)</td>
<td>-0.66 (3.75)</td>
<td>7.28 (4.1)</td>
<td>-9.11 (3.67)</td>
</tr>
</tbody>
</table>
Table 4: Number of Data Points Affecting Conclusions of Oregon Medicaid Table 9 ITT Results

<table>
<thead>
<tr>
<th>Outcome</th>
<th>N</th>
<th>Change Sign</th>
<th>Change Significance</th>
<th>Change Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>health_genflip_bin_12m</td>
<td>23,361</td>
<td>286</td>
<td>163</td>
<td>422</td>
</tr>
<tr>
<td>health_notpoor_12m</td>
<td>23,361</td>
<td>156</td>
<td>101</td>
<td>224</td>
</tr>
<tr>
<td>health_chgflip_bin_12m</td>
<td>23,407</td>
<td>198</td>
<td>106</td>
<td>292</td>
</tr>
<tr>
<td>notbaddays_tot_12m</td>
<td>21,881</td>
<td>74</td>
<td>11</td>
<td>147</td>
</tr>
<tr>
<td>notbaddays_phys_12m</td>
<td>21,384</td>
<td>88</td>
<td>20</td>
<td>166</td>
</tr>
<tr>
<td>notbaddays_ment_12m</td>
<td>21,601</td>
<td>124</td>
<td>42</td>
<td>216</td>
</tr>
<tr>
<td>nodep_screen_12m</td>
<td>23,147</td>
<td>123</td>
<td>43</td>
<td>225</td>
</tr>
</tbody>
</table>

Table 5: Number of Data Points Affecting Conclusions of Oregon Medicaid Table 9 IV Results

<table>
<thead>
<tr>
<th>Outcome</th>
<th>N</th>
<th>Change Sign</th>
<th>Change Significance</th>
<th>Change Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>health_genflip_bin_12m</td>
<td>23,361</td>
<td>275</td>
<td>162</td>
<td>383</td>
</tr>
<tr>
<td>health_notpoor_12m</td>
<td>23,361</td>
<td>155</td>
<td>100</td>
<td>219</td>
</tr>
<tr>
<td>health_chgflip_bin_12m</td>
<td>23,407</td>
<td>197</td>
<td>106</td>
<td>292</td>
</tr>
<tr>
<td>notbaddays_tot_12m</td>
<td>21,881</td>
<td>73</td>
<td>10</td>
<td>145</td>
</tr>
<tr>
<td>notbaddays_phys_12m</td>
<td>21,384</td>
<td>87</td>
<td>20</td>
<td>165</td>
</tr>
<tr>
<td>notbaddays_ment_12m</td>
<td>21,601</td>
<td>123</td>
<td>42</td>
<td>215</td>
</tr>
<tr>
<td>nodep_screen_12m</td>
<td>23,147</td>
<td>123</td>
<td>42</td>
<td>220</td>
</tr>
</tbody>
</table>
### Table 6: Manual Re-Runs Of The Oregon Medicaid ITT Analysis

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Beta (SE)</th>
<th>Re-run for sign</th>
<th>Re-run for significance</th>
<th>Re-run for both</th>
</tr>
</thead>
<tbody>
<tr>
<td>health_genflip_bin_12m</td>
<td>0.039 (0.008)</td>
<td>-0.004 (0.008)</td>
<td>0.013 (0.008)</td>
<td>-0.021 (0.008)</td>
</tr>
<tr>
<td>health_notpoor_12m</td>
<td>0.029 (0.005)</td>
<td>-0.001 (0.005)</td>
<td>0.008 (0.005)</td>
<td>-0.009 (0.004)</td>
</tr>
<tr>
<td>health_chgflip_bin_12m</td>
<td>0.033 (0.007)</td>
<td>-0.002 (0.006)</td>
<td>0.011 (0.007)</td>
<td>-0.015 (0.006)</td>
</tr>
<tr>
<td>notbaddays_tot_12m</td>
<td>0.381 (0.162)</td>
<td>-0.013 (0.157)</td>
<td>0.306 (0.161)</td>
<td>-0.309 (0.153)</td>
</tr>
<tr>
<td>notbaddays phys_12m</td>
<td>0.459 (0.175)</td>
<td>-0.017 (0.169)</td>
<td>0.328 (0.172)</td>
<td>-0.344 (0.165)</td>
</tr>
<tr>
<td>notbaddays ment_12m</td>
<td>0.603 (0.184)</td>
<td>-0.027 (0.178)</td>
<td>0.34 (0.181)</td>
<td>-0.381 (0.175)</td>
</tr>
<tr>
<td>nodep_screen_12m</td>
<td>0.023 (0.007)</td>
<td>-0.001 (0.007)</td>
<td>0.013 (0.007)</td>
<td>-0.015 (0.007)</td>
</tr>
</tbody>
</table>

### Table 7: Manual Re-Runs Of The Oregon Medicaid IV Analysis

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Beta (SE)</th>
<th>Re-run for sign</th>
<th>Re-run for significance</th>
<th>Re-run for both</th>
</tr>
</thead>
<tbody>
<tr>
<td>health_genflip_bin_12m</td>
<td>0.133 (0.026)</td>
<td>-0.006 (0.025)</td>
<td>0.044 (0.026)</td>
<td>-0.047 (0.024)</td>
</tr>
<tr>
<td>health_notpoor_12m</td>
<td>0.099 (0.018)</td>
<td>-0.003 (0.015)</td>
<td>0.027 (0.016)</td>
<td>-0.03 (0.015)</td>
</tr>
<tr>
<td>health_chgflip_bin_12m</td>
<td>0.113 (0.023)</td>
<td>-0.006 (0.022)</td>
<td>0.039 (0.022)</td>
<td>-0.051 (0.022)</td>
</tr>
<tr>
<td>notbaddays_tot_12m</td>
<td>1.317 (0.563)</td>
<td>-0.023 (0.535)</td>
<td>1.078 (0.558)</td>
<td>-1.035 (0.524)</td>
</tr>
<tr>
<td>notbaddays phys_12m</td>
<td>1.585 (0.606)</td>
<td>-0.04 (0.577)</td>
<td>1.131 (0.597)</td>
<td>-1.169 (0.568)</td>
</tr>
<tr>
<td>notbaddays ment_12m</td>
<td>2.082 (0.64)</td>
<td>-0.062 (0.607)</td>
<td>1.171 (0.625)</td>
<td>-1.293 (0.603)</td>
</tr>
<tr>
<td>nodep_screen_12m</td>
<td>0.078 (0.025)</td>
<td>-0.005 (0.024)</td>
<td>0.046 (0.024)</td>
<td>-0.05 (0.023)</td>
</tr>
</tbody>
</table>
Table 8: Number of Data Points Affecting Conclusions of Microcredit Profit Analysis

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>Change Sign</th>
<th>Change Significance</th>
<th>Change Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>16,560</td>
<td>1</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>Mongolia</td>
<td>961</td>
<td>16</td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>Bosnia</td>
<td>1,195</td>
<td>14</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>India</td>
<td>6,863</td>
<td>6</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td>Morocco</td>
<td>5,498</td>
<td>11</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Philippines</td>
<td>1,113</td>
<td>9</td>
<td>10</td>
<td>63</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>3,113</td>
<td>1</td>
<td>45</td>
<td>66</td>
</tr>
</tbody>
</table>

Table 9: Manual Re-Runs Of The Microcredit Profit Analysis

<table>
<thead>
<tr>
<th>Country</th>
<th>Beta (SE)</th>
<th>Re-run for sign</th>
<th>Re-run for significance</th>
<th>Re-run for both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>-4.55 (5.88)</td>
<td>0.4 (3.19)</td>
<td>-10.96 (5.57)</td>
<td>7.03 (2.55)</td>
</tr>
<tr>
<td>Mongolia</td>
<td>-0.34 (0.22)</td>
<td>0.02 (0.18)</td>
<td>-0.44 (0.22)</td>
<td>0.36 (0.15)</td>
</tr>
<tr>
<td>Bosnia</td>
<td>37.53 (19.78)</td>
<td>-2.23 (15.63)</td>
<td>43.73 (18.89)</td>
<td>-34.93 (14.32)</td>
</tr>
<tr>
<td>India</td>
<td>16.72 (11.83)</td>
<td>-0.5 (8.22)</td>
<td>22.89 (10.27)</td>
<td>-16.64 (7.54)</td>
</tr>
<tr>
<td>Morocco</td>
<td>17.54 (11.4)</td>
<td>-0.57 (9.92)</td>
<td>21.72 (11)</td>
<td>-18.85 (9.01)</td>
</tr>
<tr>
<td>Philippines</td>
<td>66.56 (78.13)</td>
<td>-4.01 (57.2)</td>
<td>155.89 (77.37)</td>
<td>-135.41 (53.51)</td>
</tr>
<tr>
<td>Ethiopia</td>
<td>7.29 (7.89)</td>
<td>-0.05 (2.51)</td>
<td>15.36 (7.76)</td>
<td>-8.75 (1.85)</td>
</tr>
</tbody>
</table>

Table 10: Number of Data Points Affecting Conclusions of Microcredit Temptation Analysis

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>Change Sign</th>
<th>Change Significance</th>
<th>Change Both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>16,435</td>
<td>12</td>
<td>14</td>
<td>55</td>
</tr>
<tr>
<td>Mongolia</td>
<td>961</td>
<td>3</td>
<td>12</td>
<td>162</td>
</tr>
<tr>
<td>Bosnia</td>
<td>996</td>
<td>10</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>India</td>
<td>6,827</td>
<td>41</td>
<td>8</td>
<td>85</td>
</tr>
<tr>
<td>Morocco</td>
<td>5,487</td>
<td>3</td>
<td>14</td>
<td>23</td>
</tr>
</tbody>
</table>

Table 11: Manual Re-Runs Of The Microcredit Temptation Analysis

<table>
<thead>
<tr>
<th>Country</th>
<th>Beta (SE)</th>
<th>Re-run for sign</th>
<th>Re-run for significance</th>
<th>Re-run for both</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexico</td>
<td>-0.08 (0.09)</td>
<td>0 (0.09)</td>
<td>-0.18 (0.09)</td>
<td>0.18 (0.09)</td>
</tr>
<tr>
<td>Mongolia</td>
<td>1.52 (2.1)</td>
<td>-0.03 (0.97)</td>
<td>4.21 (2.08)</td>
<td>-7.37 (2.41)</td>
</tr>
<tr>
<td>Bosnia</td>
<td>-5.8 (2.82)</td>
<td>0.39 (2.13)</td>
<td>-4.87 (2.69)</td>
<td>5.13 (1.98)</td>
</tr>
<tr>
<td>India</td>
<td>-1.64 (0.58)</td>
<td>0.04 (0.51)</td>
<td>-1.05 (0.54)</td>
<td>1.06 (0.49)</td>
</tr>
<tr>
<td>Morocco</td>
<td>-0.42 (0.72)</td>
<td>0.05 (0.67)</td>
<td>-1.35 (0.67)</td>
<td>1.25 (0.6)</td>
</tr>
</tbody>
</table>
Figure 1: Simulation results for linear regression. Values of $\hat{\beta}$ on the vertical axis, values of $\alpha$ (proportion of the data removed) on the horizontal axis. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions, the blue shaded area is the 95% confidence interval.

Figure 2: Simulation results for linear regression at differing scales of $\sigma_x$ and $\sigma_\epsilon$. A darker red colour indicates a highly sensitive analysis, in which only a small proportion of the sample needs to be removed to effect these three major changes: generating the opposite sign, changing the significance, and generating a significant result of the opposite sign. A lighter red colour indicates greater robustness. The grey areas indicate $\Psi_{\alpha} = N.A$, a failure of the linear approximation to locate any way to effect these changes.
Figure 3: Simulations. Illustrative distributions of $\tilde{\gamma}$ with traditionally-defined heavy and light tails. Here, $N = 20$, $\alpha = 0.2$, and $M = 4$. For both plots, $\sum_{n=1}^{N} \gamma_n^2 = 1$. 
Figure 4: Cash Transfers Analysis. Values of $\beta$ on the vertical axis, values of $\alpha$ (proportion of the data removed) on the horizontal axis. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions, the blue shaded area is the 95% confidence interval.
Figure 5: Oregon Medicaid Analysis. Values of $\hat{\beta}$ on the vertical axis, values of $\alpha$ (proportion of the data removed) on the horizontal axis. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions, the blue shaded area is the 95% confidence interval.
Figure 6: Microcredit Analysis. Values of $\hat{\beta}$ on the vertical axis, values of $\alpha$ (proportion of the data removed) on the horizontal axis. The red lines show how $\hat{\beta}$ can be altered by adversarial removal in both directions, the blue shaded area is the 95% confidence interval.
Figure 7: Change in marginal posterior distributions of the 7 study-specific treatment effects on the location parameter of the positive tail of Profit (within the Lognormal specification) achieved by removing the proportion of the data indicated on the horizontal axis.
Figure 8: Change in marginal posterior distribution of the hypermean and hypervariance of the treatment effect on the location parameter of the positive tail of profit (first row) and the negative tail of profit (second row), achieved by removing the proportion of the data indicated on the horizontal axis.

Figure 9: Comparison of the change in the posterior indicated by the approximation with the actual change in the hypervariance (here of the negative tail) achieved by re-running the analysis with the data points removed.
References


Appendix A  Motivating Examples

Example 1 (Linear regression). Let $y_n$ denote the response and $z_n$ denote the regresseors for linear regression. In this case, $x_n = (y_n, z_n)$ and the regression coefficient solves Eq. 2.1 with $G(\theta, x_n) = z_n(y_n - z^T_n \theta)$. This can be thought of as a method of moments estimator imposing the condition that the residuals $y_n - z^T_n \theta$ be orthogonal to the regressors.

Example 2 (Smooth optimization problems). Suppose that $f(\theta, x_n) \in \mathbb{R}$ is a three-times continuously differentiable objective function, and we wish to find $\theta$ solving

$$\hat{\theta} := \arg\min_{\theta} \frac{1}{N} \sum_{n=1}^{N} f(\theta, x_n).$$

Under appropriate regularity conditions, the optimization problem is equivalent to satisfying the first order condition $\frac{1}{N} \sum_{n=1}^{N} \frac{\partial f(\theta, x_n)}{\partial \theta} |_{\hat{\theta}} = 0$. So the optimization problem is equivalent to Eq. 2.1 with $G(\theta, x_n) = \frac{\partial f(\theta, x_n)}{\partial \theta} |_{\hat{\theta}}$.

Regression is, of course, a special case of a smooth optimization problem, as are generalized method of moments estimators. Many common econometrics methods can be cast as optimization problems, but not all, as the following example shows.

Example 3 (Instrumental variables). Suppose we have responses $y_n$, regressors $r_n$, and instruments $z_n$. Then $x_n = (y_n, r_n, z_n)$ and the instrumental variables (IV) estimator $\theta$ solves Eq. 2.1 with $G(\theta, x_n) = z_n(y_n - \theta^T x_n)$. Note that the IV estimator does not, in general solve an optimization problem. It if did, then $\frac{1}{N} \sum_{n=1}^{N} \partial G(\theta, x_n) / \partial \theta = \frac{1}{N} \sum_{n=1}^{N} z_n x_n^T$ would be the Hessian matrix of the objective function, but this is impossible, because $\frac{1}{N} \sum_{n=1}^{N} z_n x_n^T$ is not symmetric in general.

Example 4 (Sequences of optimization problems). Suppose that $\theta = (\theta_1, \theta_2)$, and we estimate $\theta_1$ by first solving an optimization problem

$$\hat{\theta}_1 := \arg\min_{\theta_1} \frac{1}{N} \sum_{n=1}^{N} f_1(\theta_1, x_n),$$

and then use $\hat{\theta}_1$ as a hyperparameter for a subsequent optimization problem:

$$\hat{\theta}_2 := \arg\min_{\theta_2} \frac{1}{N} \sum_{n=1}^{N} f_2(\hat{\theta}_1, \theta_2, x_n).$$

Jointly, this is equivalent to solving Eq. 2.1 with $G(\theta, x_n) = \left( \frac{1}{N} \sum_{n=1}^{N} \frac{\partial f_1(\hat{\theta}_1, x_n)}{\partial \theta_1} |_{\hat{\theta}_1}, \frac{1}{N} \sum_{n=1}^{N} \frac{\partial f_2(\hat{\theta}_1, \theta_2, x_n)}{\partial \theta_2} |_{\hat{\theta}_1, \theta_2} \right)$.

Example 5 (Sample mean). A simple sample mean is a case where $S_\alpha$ and $\Psi_\alpha$ are analytically tractable. Let $\bar{x}$ be a vector of $N$ scalar observations, and let $G(\theta, x_n) = \theta - x_n$, so that $\hat{\theta} = \frac{1}{N} \sum_{n=1}^{N} x_n$. The re-weighted estimate is given by $\hat{\theta}(\bar{w}) = \sum_{n=1}^{N} w_n x_n / \sum_{n=1}^{N} w_n$. We will take $\phi(\theta, \bar{w}) = \theta$. Without loss of generality,
Figure 10: A graph of Example 5 for $1 \leq M \leq 100$ using 1000 standard normal datapoints. The horizontal blue line shows the original sample mean, and the red curve shows $\hat{\theta}(\vec{w})$ as more and more points are left out. The vertical black line shows that the sign of $\hat{\theta}(\vec{w})$ changes when 10 points are left out.

1261 let $\vec{x}$ be sorted so that $x_1 \leq x_2 \leq \ldots \leq x_N$. When there are $N'$ points remaining in the sample, the effect on $\hat{\theta}$ of removing datapoint $n$ is $-x_n/N'$. Consequently, the most influential datapoint to remove is always the most negative observation remaining, so $S_\alpha = \{1 \ldots M\}$, $\hat{\theta}(\vec{w}^*) = \frac{1}{N-M} \sum_{n=M+1}^{N} x_n$, and $\Psi_\alpha = \hat{\theta}(\vec{w}^*) - \hat{\theta}$.

Figure 10 shows this analysis for 1000 standard normal datapoints which happened to have a sample mean of $-0.026$. We took $M \leq 100$ ($\alpha < 10\%$). The difference between the red curve, which shows $\hat{\theta}(\vec{w})$, and the horizontal blue line, which shows $\hat{\theta}$, is $\Psi_\alpha$ for increasing $M$. The re-weighted $\hat{\theta}(\vec{w})$ crosses zero at $M = 10$ ($\alpha = 1\%$). If the conclusions of an analysis rested on the fact that the sign of $\hat{\theta}$ were negative, one would take $\Delta = 0.026$ and find that $M = 10$ sufficed to produce a change of size $\Delta$ and overturn the analysis. If 1% of the sample were not considered too large, then the analysis would not be considered robust.

**Example 6** (Linear regression). Consider linear regression as defined in Example 1. In this case,

$$\hat{\theta}(\vec{w}) = \left( \sum_{n=1}^{N} w_n z_n z_n^T \right)^{-1} \sum_{n=1}^{N} w_n y_n z_n.$$

The fact that the weights occur in the $\left( \sum_{n=1}^{N} w_n z_n z_n^T \right)^{-1}$ term mean that the ordering of the datapoints’ influence depends on which datapoints have already been removed. For example, suppose that the regressors would be nearly colinear were it not for two similar datapoints, $x_1$ and $x_2$. The effect of removing only one of $x_1$ or $x_2$ alone might be small, but the effect of removing both of them could be very
large. For this reason, exactly computing $S_\alpha$ and $\Psi_\alpha$ for linear regression is hard in general.

Note that if one were to “fix” the regressors in the re-weighting, defining the weighted estimating equation

$$\frac{1}{N} \sum_{n=1}^{N} z_n \left( w_n y_n - z_n^T \hat{\theta}(\bar{w}) \right) = 0,$$

then $S_\alpha$ and $\Psi_\alpha$ would have closed forms as in Example 5. For the remainder of the paper we take the view that removing a datapoint involves removing the regressor as well, though which is most appropriate depends on the context.

**Example 7** (Optimization.). In the setting of Example 2, the Jacobian of $G(\theta, x_n)$ is the Hessian of the objective function $f(\theta, x_n)$, and Eq. 2.8 takes the form

$$\frac{d\hat{\theta}(\bar{w})}{d\bar{w}^T} \bigg| _{\bar{w}} = - \left( \sum_{n=1}^{N} w_n \frac{\partial^2 f(\theta, \bar{w})}{\partial \theta \partial \theta^T} \bigg| _{\theta(\bar{w}), \bar{w}} \right)^{-1} \left( \frac{\partial f(\theta, x_1)}{\partial \theta} \bigg| _{\theta(\bar{w})}, \ldots, \frac{\partial f(\theta, x_N)}{\partial \theta} \bigg| _{\theta(\bar{w})} \right).$$

Note that, in order for Eq. 2.8 to apply, the Hessian matrix must be non-degenerate at $\hat{\theta}(\bar{w}), \bar{w}$.

**Example 8** (A component of $\hat{\theta}$). When $\phi$ simply picks out one entry of the vector $\theta$, i.e. $\phi(\theta) = \theta_d$, then $\psi_n$ is simply the $(d, n)$-th entry of the matrix $d\hat{\theta}(\bar{w})/d\bar{w}^T$.

**Appendix B**  Asymptotic Properties of the Influence Function

**B.1 Covariance of M-estimators: standard and robust versions.**

Suppose we have an objective function, $g$, that decomposes as a sum over datapoints, $x_n, n = 1, \ldots, N$. Let $x = (x_1, \ldots, x_N)$. Let an estimate of the parameter $\theta \in \mathbb{R}^D$ be defined as a root of the summed objective function, i.e.,

$$\hat{\theta} := \theta \text{ such that } \sum_{n=1}^{N} g(x_n, \theta) =: G(x, \theta) = 0. \quad \text{(B.1)}$$

By definition, $G\left(\hat{\theta}, x\right) = 0$. The MLE of smooth likelihoods is such an estimator, where $g$ is the gradient of the log likelihood. Let $\theta_0$ denote the “true” value (that is, the root of Eq. B.1). Assume all the smoothness and regularity you need, Taylor
expand a single term around $\theta_0$, and evaluate at $\hat{\theta}$ to get

$$0 = G(x, \hat{\theta}) = G(x, \theta_0) + \frac{dG}{d\theta} \bigg|_{\theta_0} (\hat{\theta} - \theta_0) + O\left(\left\|\hat{\theta} - \theta_0\right\|^2\right) \Rightarrow$$

$$\hat{\theta} - \theta_0 = -\left(\frac{dG}{d\theta} \bigg|_{\theta_0}\right)^{-1} G(x, \theta_0) + O\left(\left\|\hat{\theta} - \theta_0\right\|^2\right). \quad (B.2)$$

This is sort of a “master formula”, of which different regression standard errors arise as special cases.

### B.2 Correctly specified likelihoods

First, suppose that $g$ is the gradient of a correctly specified log likelihood, $\ell(x_n, \theta)$. Then

$$g(x_n, \theta) = \nabla \ell(x_n, \theta)$$

$$G(x, \theta_0) = \sum_{n=1}^{N} \nabla \ell(x_n, \theta_0)$$

$$\frac{dG}{d\theta} \bigg|_{\theta_0} = \sum_{n=1}^{N} \nabla^2 \ell(x_n, \theta_0)$$

By standard properties of correctly-specified likelihoods,

$$\mathbb{E} [\nabla \ell(x_n, \theta)] = 0, \quad (B.3)$$

and

$$\text{Cov} (\nabla \ell(x_n, \theta_0)) = \mathcal{I}, \quad (B.4)$$

where $\mathcal{I}$ is the Fisher information. By the law of large numbers, and again a property of correctly-specified likelihoods,

$$\frac{1}{N} \frac{dG}{d\theta} \bigg|_{\theta_0} = \frac{1}{N} \sum_{n=1}^{N} \nabla^2 \ell(x_n, \theta_0) = \mathbb{E} \left[\nabla^2 \ell(x_n, \theta_0)\right] \rightarrow -\mathcal{I}. \quad (B.5)$$

By the Central limit theorem,

$$\frac{1}{\sqrt{N}} G(x, \theta_0) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} \nabla \ell(x_n, \theta_0) \xrightarrow{\text{dist}} \mathcal{N}(0, \text{Cov} (\nabla \ell(x_n, \theta_0))) = \mathcal{N}(0, \mathcal{I}). \quad (B.6)$$
By smoothness assumptions, \( O \left( \sqrt{N} \| \hat{\theta} - \theta_0 \|^2 \right) \to 0. \) Putting this all together in Eq. B.2,

\[
\sqrt{N} \left( \hat{\theta} - \theta_0 \right) = - \left( \frac{1}{N} \frac{dG}{d\theta} \right)_{\theta_0}^{-1} \frac{1}{\sqrt{N}} G \left( x, \theta_0 \right) + O \left( \sqrt{N} \| \hat{\theta} - \theta_0 \|^2 \right) \\
\rightarrow_{\text{dist}} \mathcal{I}^{-1} N \left( 0, \mathcal{I} \right) \\
= \mathcal{N} \left( 0, \mathcal{I}^{-1} \right).
\]

Typically, \( \mathcal{I}^{-1} \) is estimated with the negative inverse Hessian of the log likelihood, i.e., the “observed Fisher information”:

\[
\hat{\mathcal{I}} := - \frac{1}{N} \sum_{n=1}^{N} \nabla^2 \ell \left( x_n, \hat{\theta} \right) = - \frac{1}{N} \frac{dG}{d\theta} \bigg|_{\hat{\theta}}.
\]

I will briefly observe that this could be motivated by considering the non-asymptotic sensitivity estimator at \( \hat{\theta} \) rather than \( \theta_0 \). The formal difference is merely forming the Taylor expansion around \( \hat{\theta} \) and not \( \theta_0 \). The conceptual difference is interesting, but beyond the scope.

### B.3 Robustness to misspecification.

Let us consider how results change if the likelihood is not correct. In this case, we are simply calculating the asymptotic behavior of a smooth optimization problem. The results of Section B.2 used the fact that the model was correctly specified in two places. First, it was used in Eq. B.3. This is not a particularly material assumption; we simply define the “true” \( \theta_0 \) as the one for which Eq. B.3 is true. Less benign is the assumption that the quantity \( \mathcal{I} \) in Eq. B.4 and Eq. B.5 is the same. This is not, in general, true. Let us define

\[
\mathbb{E} \left[ \nabla^2 \ell \left( x_n, \theta_0 \right) \right] := -H.
\]

Then, everything else follows as before, except we get

\[
V = \mathbb{E} \left[ \text{Cov} \left( \nabla \ell \left( x_n, \theta_0 \right) \right) \right],
\]

where the \( x_n \) is drawn according to whatever distribution generates your data – they must be independent, but might not be identically distributed. And then,

\[
\sqrt{N} \left( \hat{\theta} - \theta_0 \right) \rightarrow_{\text{dist}} H^{-1} N \left( 0, V \right) \\
= \mathcal{N} \left( 0, H^{-1} V H^{-1} \right).
\] (B.7)
Now we need to estimate two different quantities.

\[
\hat{H} := -\frac{1}{N} \sum_{n=1}^{N} \nabla^2 \ell (x_n, \hat{\theta})
\]

\[
\hat{V} := \frac{1}{N} \sum_{n=1}^{N} \nabla \ell (x_n, \hat{\theta}) \nabla \ell (x_n, \hat{\theta})^T.
\]

(B.8)

Now, \( \hat{V} \) is the sample variance of the observed scores (using the fact that \( \frac{1}{N} \sum_{n=1}^{N} \nabla \ell (x_n, \hat{\theta}) = 0 \) by definition). In Section B.2 we effectively assumed that \( \hat{H} = \hat{V} \). In finite sample, even this is not necessarily true; it was motivated in Section B.2 by purely asymptotic assumptions.

Note that, in regression problems, \( \nabla \ell (x_n, \hat{\theta}) \) is given by

\[
\ell (x_n, \theta) = \frac{1}{2} \left( y_n - x_n^T \theta \right)^2
\]

\[
\nabla \ell (x_n, \hat{\theta}) = x_n \left( y_n - x_n^T \hat{\theta} \right).
\]

This shows that Eq. B.7 is in fact the standard heteroskedasticity-consistent robust standard error for regressions.

### B.4 Robustness to within-group covariances.

Finally, let us consider grouping data together. Both Section B.2 and Section B.3 required that \( x_1, ..., x_N \) be independent of one another. Otherwise, the expected score covariance as measured by Eq. B.5 or by Eq. B.8 is not the same variance to be used in the Central limit theorem, Eq. B.6.

As a first step, note that, even when the \( x_1, ..., x_N \) are believed to be independent, there is in fact a choice to be made. For if \( x_1, ..., x_N \) are independent, then the pairs \( (x_1, x_2), (x_3, x_4), ..., (x_{N-1}, x_N) \) are also independent. So it is perfectly well-motivated to write

\[
z_m = (x_{2m+1}, x_{2m+2}), \text{ for } m = 1, ..., \frac{N}{2}
\]

\[
\ell (z, \theta) = \sum_{m=1}^{M} \ell (z_m, \theta)
\]

\[
= \sum_{m=1}^{M} \left( \ell (x_{2m+1}, \theta) + \ell (x_{2m+2}, \theta) \right)
\]

\[
= \sum_{n=1}^{N} \ell (x_n, \theta)
\]

\[
= \ell (x, \theta)
\]

(B.9)
and so replace Eq. B.8 with

$$
\hat{V}_{\text{paired}} := \frac{1}{M} \sum_{m=1}^{M} \nabla \ell (z_m, \hat{\theta}) \nabla \ell (z_m, \hat{\theta})^T
$$

(B.10)

The existence of cross-terms in Eq. B.10 shows that \( \hat{V}_{\text{paired}} \) is not equal to \( \hat{V} \) in Eq. B.8, though if the \( x_n \) are truly independent then the difference disappears asymptotically. The same argument could be made for any grouping, or, indeed, groupings of different sizes where you imagine that the group size is independent and random.

We see no reason to use \( \hat{V}_{\text{paired}} \) if you believe that \( x_1, \ldots, x_N \) are independent. However, if you believe that some of the \( x_n \) are dependent within a certain grouping, but that the groups are independent of one another, then you can simply re-write the problem using these groups:

$$
z_m = \left( x_{gm(1)}, \ldots, x_{gm(N_m)} \right), \text{ for } m = 1, \ldots, M,
$$

and use the reasoning of Section B.3 applied to the grouped random variables as in Eq. B.10. A simple example is the problem we’re considering, where the groups are villages, and the \( x_n \) are observations for people. The objective function is unchanged, as seen in Eq. B.9. The score covariance estimator will be noisier in general, since it contains the cross-terms that would otherwise be absent in Eq. B.8.

But this may be a small price to pay for a correct model specification.

### Appendix C  Bounds on the shape parameter

As this is \( \alpha \) times the truncated mean of a non-generate distribution, \( \Gamma_\alpha^+ \) will not tend to 0 as \( N \) increases. Indeed, it will remain large to the extent that this truncated mean is large, which occurs when the scale of the \( \gamma_n \) variable is large. In fact, it is possible to derive an upper bound on this object, to understand the worst-case scenario of maximal sensitivity for a fixed \( N \) and \( \alpha \) given some observed standard errors \( \sigma \).

Consider the problem of choosing \( \{\gamma_n\}_{n=1}^N \) to maximize \( \Gamma_\alpha^+ \) subject to the two constraints that \( N^{-1} \sum_{n=1}^N \gamma_n^2 = 1 \) and \( N^{-1} \sum_{n=1}^N \gamma_n = 0 \). For a given \( \alpha \) and \( N \), what matters is the value of the influence scores of the data points we are going to discard: this is the set for which \( n \in \hat{S}_\alpha \), and we denote members of this set by \( \gamma_m \) for \( m = 1, 2, \ldots M \) where applicable. Thus the Lagrangian form of the problem that defines the worst case scenario is

$$
\mathcal{L} = \inf_{\lambda_\mu, \lambda_\sigma} \max_{\{\gamma_n\}_{n=1}^N} \left( \sum_{m=1}^{M} \gamma_m + \lambda_\mu \sum_{n=1}^{N} \gamma_n + \frac{1}{2} \left( \sum_{n=1}^{N} \gamma_n^2 - N \right) \right)
$$

(C.1)
Taking first order conditions with respect to both some candidate \( \gamma_m \) and some retained \( \gamma_n \) (with slight abuse of notation),

\[
1 + \lambda_{\mu} + \lambda_{\sigma} \gamma_m = 0 \\
\lambda_{\mu} + \lambda_{\sigma} \gamma_n = 0
\]

Summing over the indices in the first equation and plugging in the constraint \( \sum_{n=1}^{N} \gamma_n = 0 \) gives \( \lambda_{\mu} = -M/N = -\alpha \). Squaring the second equation and summing over the indices and using the constraint \( \sum_{n=1}^{N} \gamma_n^2 = N \) delivers \( \lambda_{\sigma}^2 = \alpha(1-\alpha) \).

Putting these together, and employing the negative root of \( \lambda_{\sigma} \), we have

\[
\gamma_m = \frac{-(1 + \lambda_{\mu})}{\lambda_{\sigma}} = \frac{1 - \alpha}{\sqrt{\alpha(1-\alpha)}} = \sqrt{\frac{1 - \alpha}{\alpha}}
\]

This is the worst possible value of some candidate \( \gamma_m \). Thus, plugging that into the definition of \( \Gamma_{\alpha}^+ \), in general it turns out that

\[
\Gamma_{\alpha}^+ \leq \sqrt{\alpha(1-\alpha)} \quad \text{(C.2)}
\]

and that this upper bound is attained when exactly M values are omitted and when \( \gamma_m \) from these omitted points all take the same value.

### Appendix D  IV and OLS Proofs

In this section we prove the key step in Theorem 1 using the results from Giordano et al. (2019), which we will abbreviate as HOIJ. For example, we will refer to Giordano et al. (2019, Assumption 1) as HOIJ Assumption 1. Additionally, for this section only, we will exclusively use the notation of Giordano et al. (2019). Our Section 3.2 draws the connection between this section’s notation and the notation of the rest of the present paper.

Suppose we observe regressors, \( x_n \in \mathbb{R}^D \), instruments \( z_n \in \mathbb{R}^D \), and responses, \( y_n \in \mathbb{R} \), for \( n = 1, \ldots, N \). We assume that the observations are exchangeable. It will be convenient to define the residual \( \varepsilon_n(\theta) := \theta^T x_n - y_n \). Our estimating equation
partial derivatives are then
\[
G(\theta, \vec{w}) := \frac{1}{N} \sum_{n=1}^{N} w_n \epsilon_n(\theta) z_n
\]
\[
d_\theta^1 G(\theta, \vec{w}) = H(\theta, w) := \frac{1}{N} \sum_{n=1}^{N} w_n z_n x_n^T
\]
\[
d_\theta^k G(\theta, \vec{w}) = 0 \text{ for } k > 1.
\]

We now consider each of the assumptions and conditions of HOIJ Section 4.1 in this context, proving the following lemma:

**Lemma 1.** Under our Assumption 1 of Section 3.2, HOIJ Assumptions 1-5 are satisfied with \( \rho = 1/3 \) and

\[
\Omega_\theta(\mathcal{B}) := \left\{ \theta : \| \theta - \hat{\theta} \|_2 < \mathcal{B} \right\} \quad \text{with} \quad \mathcal{B} = \frac{\| \hat{\theta}^1(w) - \hat{\theta} \|_2 + 2\alpha^2 C_{op}^2 \xi_1 \xi_2}{1 - 2\alpha^2 C_{op}^2 \xi_1^2} > 0.
\]

**Proof.** For HOIJ Assumption 1, we will use HOIJ Lemma 8 to choose a suitable value of \( \mathcal{B} \) once the other constants are established.

HOIJ Assumption 2 follows immediately by inspection, as the estimating equation is linear in \( \theta \).

In this case, \( H(\theta, 1_N) = \hat{H} = \frac{1}{N} \sum_{n=1}^{N} z_n x_n^T \) does not depend on \( \theta \). So HOIJ Assumption 3 is satisfied with

HOIJ Assumption 3: \( C_{op} := \| \hat{H}^{-1} \|_{op} \).

HOIJ Assumption 4 is satisfied for \( k \geq 2 \) with \( M_k = 0 \). For \( k = 1 \), we observe again that \( d_\theta^1 G(\theta, 1_N) \) does not depend on \( \theta \), and so we can take

HOIJ Assumption 4: \( M_1 := \| \hat{H} \|_2 \)

\[
M_2 := 0, \ldots.
\]

Note that HOIJ Assumption 4 does not require control for \( k = 0 \).

HOIJ Assumption 5 is more involved and depends on the types of weights we are considering. Let us adopt the notation of Section 3.2 that \( w \) is zero in entries indexed by a set \( \mathcal{M} \subseteq [N] \) and one otherwise. It will turn out to be convenient to define the following quantities (which match the definitions of Definition 6 but in
the present section’s notation).

\[\begin{align*}
\alpha &:= \frac{|\mathcal{N}|}{N} \\
\xi_1 &:= \left\| \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} z_n x_n^T \right\|_2 \\
\xi_2 &:= \left\| \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} \varepsilon_n(\hat{\theta}) z_n \right\|_2.
\end{align*}\]

The stochastic and asymptotic behavior of \(\xi_1\) and \(\xi_2\) will obviously depend on how the weights \(w\) are chosen and on the tail behavior of \(x_n\) and \(y_n\).

For \(k \geq 2\), we can take \(\lambda_k = 0\). For \(k = 1\), \(d^1_0 G(\theta, w)\) is again independent of \(\theta\), so we have

\[
\sup_{\theta \in \Omega(\theta)} \| d^1_0 G(\theta, w) - d^1_0 G(\theta, 1_N) \|_2 = \alpha \xi_1.
\]

For \(k = 0\), we must rely on the definition of \(\Omega(\theta)\).

\[
\sup_{\theta \in \Omega(\theta)} \| G(\theta, w) - G(\theta, 1_N) \|_2 = \alpha \sup_{\theta \in \Omega(\theta)} \left\| \frac{1}{|\mathcal{N}|} \sum_{n = 1}^N \varepsilon_n(\theta) z_n \right\|_2 \\
\leq \alpha \left( \sup_{\theta \in \Omega(\theta)} \left\| (\theta - \hat{\theta})^T \frac{1}{|\mathcal{N}|} \sum_{n \in \mathcal{N}} z_n x_n^T \right\|_2 + \xi_2 \right) \\
\leq \alpha (\mathcal{B} \xi_1 + \xi_2).
\]

We can then make use of HOIJ Lemma 1 to set

\[\begin{align*}
\text{HOIJ Assumption 5:} & \quad \lambda_0 := \alpha (\mathcal{B} \xi_1 + \xi_2) \\
\lambda_1 &:= \alpha \xi_1 \\
\lambda_2 &:= 0, \ldots
\end{align*}\]

Next, we turn to HOIJ Condition 1, for which we will require that

\[\text{HOIJ Condition 1:} \quad \rho := C_{op} \lambda_1 + C_{op}^2 M_2 \lambda_0 = \alpha C_{op} \xi_1 < 1.\]

Should HOIJ Condition 1 fail to be satisfied, then our bounds cannot be applied, essentially because \(d^1_0 G(\theta, w)\) is not smooth enough to guarantee the strong convexity of \(H(\hat{\theta})\) using HOIJ Lemma 2. Supposing that HOIJ Condition 1 is satisfied with \(\rho < 1\), we define \(\tilde{C}_{op} := \frac{1}{1 - \rho} C_{op}\) as in HOIJ Lemma 2. In Theorem 1, we take...
\( \rho = 1/3 \), for reasons to be described below.

Finally, we can calculate the approximation and error bounds (though we cannot use the bounds to control the error until we have chosen \( B \) to satisfy HOIJ Lemma 8). Let us consider \( k_{IJ} = 1 \) for simplicity. We have

\[
\hat{\theta}_{1}^{IJ}(w) = \hat{\theta} + \delta_{w}^{1} \hat{\theta}(1_{N})
\]

\[
= \hat{\theta} + \alpha \left( \frac{1}{N} \sum_{n=1}^{N} x_{n} x_{n}^{T} \right)^{-1} \frac{1}{|\mathcal{V}|} \sum_{n \in \mathcal{V}} x_{n} \varepsilon_{n}(\hat{\theta}) \quad \text{(HOIJ Equation 3)}
\]

\[
\left\| \hat{\theta}_{1}^{IJ}(w) - \hat{\theta}(w) \right\|_{2} \leq \tilde{C}_{\alpha} \op \lambda_{0}^{2} + 2 \tilde{C}_{\alpha} \lambda_{1} + \tilde{C}_{\alpha} \lambda_{2}^{2}
\]

\[
= 2\alpha^{2} \tilde{C}_{\alpha} \xi_{1}(\mathcal{B} \xi_{1} + \xi_{2}).
\]

In order to apply HOIJ Lemma 8 to show that \( \Omega_{\theta}(\mathcal{B}) \) satisfies HOIJ Assumption 1, it will suffice to choose \( B \) so that

\[
\mathcal{B} > \left\| \hat{\theta}_{1}^{IJ}(w) - \hat{\theta} \right\|_{2} + 2\alpha^{2} \tilde{C}_{\alpha} \xi_{1}(\mathcal{B} \xi_{1} + \xi_{2}) \iff
\]

\[
\mathcal{B} > \frac{\left\| \hat{\theta}_{1}^{IJ}(w) - \hat{\theta} \right\|_{2} + 2\alpha^{2} \tilde{C}_{\alpha} \xi_{1} \xi_{2}}{1 - 2\alpha^{2} \tilde{C}_{\alpha} \xi_{1}^{2}}.
\]  

(D.1)

As long as \( 2\alpha^{2} \tilde{C}_{\alpha} \xi_{1}^{2} < 1 \), a positive solution to Eq. D.1 exists and can be readily computed from the data at hand.

Recall that, to satisfy HOIJ Condition 1 above, we took \( \rho = \alpha C_{\op} \xi_{1} < 1 \). Noting this fact allows the interpretation of Eq. D.1 as an additional, stricter condition on \( \rho \), since

\[
2\alpha^{2} \tilde{C}_{\alpha} \xi_{1}^{2} < 1 \iff
\]

\[
\left( \frac{\rho}{1 - \rho} \right)^{2} < \frac{1}{2} \iff
\]

\[
\rho \left( 1 + \frac{1}{\sqrt{2}} \right) < \frac{1}{\sqrt{2}} \iff
\]

\[
\rho < \frac{1}{1 + \sqrt{2}} < 1.
\]

Consequently, to satisfy both HOIJ Condition 1 and HOIJ Lemma 8, it suffices to have

\[
\alpha C_{\op} \xi_{1} < \frac{1}{1 + \sqrt{2}}.
\]  

(D.2)

which can be satisfied by requiring \( \alpha C_{\op} \xi_{1} < \rho := 1/3 \), the value assumed in Assumption 1.

\[ \square \]
Appendix E  Paired Removal Procedures

In this appendix we discuss the option of enforcing a removal protocol to ensure balanced removal of data points across groups within the study. This concern may arise for practitioners who are accustomed to analyzing randomized controlled trials and who believe that symmetric operations must be performed on treatment and control groups not just in expectation but in practice, in each sample realization. In this section we will show that it is difficult to define a coherent notion of balanced removal and that the development of a paired removal metric would be a challenging problem in itself. However, to demonstrate that even a rather punitive approach to the balance problem will not generally eliminate the sensitivity we measure here, we then attempt such an exercise on the microcredit profit data.

Before we examine the issue of pairing data points, we wish to point out that randomization does not and is not intended to provide balance across treatment and control groups in every finite sample realization. It only produces balance in expectation and in asymptopia. Moreover, if researchers are worried about "doing something different" to the treatment and control groups then they should not be engaged in common practices such as Winsorizing or removing outliers based on extreme values in the outcome or in any of the covariates, since this will generically affect the two groups differently in any given sample.

Warnings aside, suppose that a researcher analyzing a randomized trial wishes to use our metric but to enforce "balanced" removal across treatment and control groups. In general, it is difficult to define a notion of balance at the individual data point level. Consider assigning each outcome data point in the treatment group a rank and assigning each outcome data point in the control group a rank. One might imagine that enforcing removal paired by ranks would be desirable to ensure parity or balance across groups, but unless there are exactly the same number of households in the treatment and control groups, there will be a rank mismatch at one end of the data or the other and some extreme values will have no rank-match.

One can avoid having an unmatched data point by binning the outcome data up into discrete quantiles and enforcing removal of the entire quantile across both groups if any data point in the quantile is implicated, but the number of data points in each of the quantiles will generally differ across quantiles both within each group and across the groups. This would ensure that data points of similar magnitude are removed from both groups – and thus that removal is "balanced" along that notion – but cannot ensure parity of number of data points removed across groups. Even greater trouble is caused by the frequent presence of ties in the data we analyze in economics, although it is possible to use jittering to handle that to some extent when the ties are few.