ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 1, *pp.* 39–67

Making sense of sensitivity: extending omitted variable bias

Carlos Cinelli and Chad Hazlett

University of California, Los Angeles, USA

[Received August 2018. Final revision October 2019]

Summary. We extend the omitted variable bias framework with a suite of tools for sensitivity analysis in regression models that does not require assumptions on the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounders, naturally handles multiple confounders, possibly acting non-linearly, exploits expert knowledge to bound sensitivity parameters and can be easily computed by using only standard regression results. In particular, we introduce two novel sensitivity measures suited for routine reporting. The robustness value describes the minimum strength of association that unobserved confounding would need to have, both with the treatment and with the outcome, to change the research conclusions. The partial R^2 of the treatment with the outcome shows how strongly confounders explaining all the residual outcome variation would have to be associated with the treatment to eliminate the estimated effect. Next, we offer graphical tools for elaborating on problematic confounders, examining the sensitivity of point estimates and t-values, as well as 'extreme scenarios'. Finally, we describe problems with a common 'benchmarking' practice and introduce a novel procedure to bound the strength of confounders formally on the basis of a comparison with observed covariates. We apply these methods to a running example that estimates the effect of exposure to violence on attitudes toward peace.

Keywords: Causal inference; Confounding; Omitted variable bias; Regression; Robustness value; Sensitivity analysis

1. Introduction

Observational research often seeks to estimate causal effects under a 'no-unobserved-confounding' or 'ignorability' (conditional on observables) assumption (see for example Rosenbaum and Rubin (1983a), Pearl (2009) and Imbens and Rubin (2015)). When making causal claims from observational data, investigators marshal what evidence they can to argue that their result is not due to confounding. In 'natural' and 'quasi'-experiments, this often includes a qualitative account for why the treatment assignment is 'as if' random conditional on a set of key characteristics (see for example Angrist and Pischke (2008) and Dunning (2012)). Investigators seeking to make causal claims from observational data are also instructed to show 'balance tests' and 'placebo tests'. Although, in some cases, null findings on these tests may be consistent with the claim of no unobserved confounders, they are certainly not dispositive: it is *unobserved* variables that we worry may be both 'imbalanced' and related to the outcome in problematic ways. Fundamentally, causal inferences always require assumptions that are unverifiable from the data (Pearl, 2009).

Thus, in addition to balance and placebo tests, investigators are advised to conduct 'sensitivity

Address for correspondence: Chad Hazlett, Departments of Statistics and Political Science, University of California, Los Angeles, 8125 Math Sciences Building, Los Angeles, CA 90095, USA.
E-mail: chazlett@ucla.edu

 1369–7412/20/82039

analyses' examining how fragile a result is against the possibility of unobserved confounding. (Researchers may also wish to examine sensitivity to the choice of observed covariates; see Leamer (1983, 2016).) In general, such analyses entail two components:

(a) describing the type of unobserved confounders—parameterized by their relation to the treatment assignment, the outcome, or both—that would substantively change our conclusions about the estimated causal effect, and
(b) assisting the investigator in assessing the plausibility that such problematic confounding might exist, which necessarily depends on the research design and expert knowledge regarding the data-generating process.

A variety of sensitivity analyses have been proposed, dating back to Cornfield *et al.* (1959), with more recent contributions including Rosenbaum and Rubin (1983b), Robins (1999), Frank (2000), Rosenbaum (2002), Imbens (2003), Brumback *et al.* (2004), Frank *et al.* (2008, 2013), Hosman *et al.* (2010), Imai *et al.* (2010), Vanderweele and Arah (2011), Blackwell (2013), Carnegie *et al.* (2016b), Dorie *et al.* (2016), Middleton *et al.* (2016), Oster (2019) and Franks *et al.* (2019). Yet, such sensitivity analyses remain underutilized. For instance, in political science, out of 164 quantitative papers in the top three general interest publications (the *American Political Science Review*, *American Journal of Political Science* and *Journal of Politics*) for 2017, 64 papers clearly described a causal identification strategy other than a randomized experiment. Of these only four (6.25%) employed a formal sensitivity analysis beyond trying various specifications. In economics, Oster (2014) reported that most non-experimental empirical papers utilized only informal robustness tests based on coefficient stability in the face of adding or dropping covariates.

We argue that various factors contribute to this reluctant uptake. One is the complicated nature and strong assumptions that many of these methods impose, sometimes involving restrictions on or even a complete description of the nature of the confounder. A second reason is that, whereas training, convention and convenience dictate that users routinely report 'regression tables' (or perhaps coefficient plots) to convey the results of a regression, we lack readily available quantities that aid in understanding and communicating how sensitive our results are to potential unobserved confounding. Third, and most fundamentally, connecting the results of a formal sensitivity analysis to a cogent argument about what types of confounders may exist in one's research project is often difficult, particularly with research designs that do not hinge on a credible argument regarding the (conditionally) 'ignorable', 'exogenous' or 'as-if random' nature of the treatment assignment. To complicate things, some of the solutions that are offered by the literature can lead users to erroneous conclusions (see Section 6 for discussion).

In this paper we show how the familiar omitted variable bias (OVB) framework can be extended to address these challenges. We develop a suite of sensitivity analysis tools that do not require assumptions on the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounder and can be used to assess the sensitivity to multiple confounders, whether they influence the treatment and outcome linearly or not.

We first introduce two novel measures of the sensitivity of linear regression coefficients:

(a) the 'robustness value' RV, which provides a convenient reference point to assess the overall robustness of a coefficient to unobserved confounding. If the confounders' association to the treatment and to the outcome (measured in terms of partial R^2) are *both* assumed to be less than the robustness value, then such confounders cannot 'explain away' the observed effect. And,

(b) the proportion of variation in the outcome explained uniquely by the treatment, $R^2_{Y\sim D|\mathbf{X}}$, which reveals how strongly counfounders that explain 100% of the residual variance of the outcome would have to be associated with the treatment to eliminate the effect.

Both measures can be easily computed from standard regression output: one needs only the estimate's t-value and the degrees of freedom. To advance standard practice across a variety of disciplines, we propose routinely reporting RV and $R^2_{Y\sim D|\mathbf{X}}$ in regression tables.

Next, we offer graphical tools that investigators can use to refine their sensitivity analyses. The first is close in spirit to the proposal of Imbens (2003)—a bivariate sensitivity contour plot, parameterizing the confounder in terms of partial R^2 values. However, contrary to Imbens's maximum likelihood approach, the OVB-based approach makes the underlying analysis simpler to understand, easier to compute and more general. It side-steps assumptions on the functional form of the treatment assignment and on the distribution of the (possibly multiple, non-linear) confounders, and it easily extends contour plots to assess the sensitivity of t-values, p-values or confidence intervals. This enables users to examine the types of confounders that would alter their inferential conclusions, not just point estimates. The second is an 'extreme scenario' sensitivity plot, in which investigators make conservative assumptions about the portion of otherwise unexplainend variance in the outcome that is due to confounders. One can then see how strongly such confounders would need to be associated with the treatment to be problematic. In the 'worst case' of these scenarios, the investigator assumes that *all* unexplained variation in the outcome may be due to a confounder.

Finally, we introduce a novel bounding procedure that aids researchers in judging which confounders are plausible or could be ruled out, using the observed data in combination with expert knowledge. Whereas prior work (Frank, 2000; Imbens, 2003; Hosman *et al.*, 2010; Blackwell, 2013; Dorie *et al.*, 2016; Carnegie *et al.*, 2016a; Middleton *et al.*, 2016; Hong *et al.*, 2018) has suggested an informal practice of benchmarking the unobserved confounding by comparison with unadjusted statistics of observables, we show that this practice can generate misleading conclusions due to the effects of confounding itself, even if the confounder is assumed to be independent of the covariate(s) that are used for benchmarking. Instead, our approach formally bounds the strength of unobserved confounding with the same strength (or a multiple thereof) as a chosen observable or group of observables. These bounds are tight and may be especially useful when investigators can credibly argue to have measured the most important determinants (in terms of variance explained) of the treatment assignment or of the outcome.

In what follows, Section 2 describes the running example that will be used to illustrate the tools throughout the text—a study of the effect of violence on attitudes toward peace in Darfur, Sudan. Section 3 introduces the traditional OVB framework, how it can be used for a first approach to sensitivity analysis and some of its shortcomings. Next, Section 4 shows how to extend the traditional OVB with the partial R^2 parameterization and Section 5 demonstrates how these results lead to a rich set of tools for sensitivity analysis. We conclude by discussing how our proposal seeks to increase the use of sensitivity analyses in practice and how it compares with existing procedures, and highlighting important *caveats* when interpreting sensitivity results.

Open-source software for R (`sensemakr` (Cinelli and Hazlett, 2019) that implements the methods that are presented here is available from `https://cran.r-project.org/package=sensemakr`). A Shiny web application is also available from `https://carloscinelli.shinyapps.io/robustness_value/`. Finally, code to replicate all the analyses can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

2. Running example

In this section we briefly introduce the applied example that is used throughout the paper. (We describe only the most relevant details; further information is available in Hazlett (2019).) This serves as a background to illustrate how the tools that are developed here can be applied to address problems that commonly arise in observational research. We emphasize that the information that is produced by a sensitivity analysis is useful to the extent that researchers can wield domain knowledge about the data-generating process to rule out the types of confounders which are shown to be problematic. Thus, a real example helps to illustrate how such knowledge could be employed.

2.1. Exposure to violence in Darfur

In Sudan's western region of Darfur, a horrific campaign of violence against civilians began in 2003, sustaining high levels of violence through 2004 and killing an estimated 200000 people (Flint and de Waal, 2008). It was deemed genocide by then Secretary of State Colin Powell and has resulted in indictments of alleged genocide, war crimes and crimes against humanity in the International Criminal Court.

In the current case, we are interested in learning how being physically harmed during attacks on one's village changed individual attitudes towards peace. Clearly, we cannot randomize who is exposed to such violence. However, the means by which violence was distributed provide a tragic natural experiment. Violence against civilians during this time included both aerial bombardments by government aircraft, and attacks by a pro-government militia called the *Janjaweed*. Although some villages were singled out for more or less violence, within a given village violence was arguably indiscriminate. This argument is supported by reports such as

> 'The government came with Antonovs, and targeted everything that moved. They made no distinction between the civilians and rebel groups. If it moved, it was bombed. It is the same thing, whether there are rebel groups (present) or not .... The government bombs from the sky and the *Janjaweed* sweeps through and burns everything and loots the animals and spoils everything that they cannot take'

(transcript from an interview taken by the Darfurian Voices team; interview code 03072009_118_cf2009008).

One can further argue that attacks were indiscriminate within village on the basis that the violence that was promoted by the government was mainly used to drive people out rather than to target individuals. Within village, the bombing was crude and the attackers had almost no information about whom they would target, with one major exception: whereas both men and women were often injured or killed, women were targeted for widespread sexual assault and rape by the *Janjaweed*.

With this in mind, an investigator might claim that village and gender are sufficient for control of confounding and estimate the linear model

$$\text{PeaceIndex} = \hat{\tau}_{\text{res}}\,\text{DirectHarm} + \hat{\beta}_{f,\text{res}}\,\text{Female} + \text{Village}\,\hat{\boldsymbol{\beta}}_{v,\text{res}} + \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{res}} + \hat{\varepsilon}_{\text{res}} \qquad (1)$$

where PeaceIndex is an index measuring individual attitudes towards peace, DirectHarm is a dummy variable indicating whether an individual was reportedly injured or maimed during such an attack, Female is a fixed effect for being female and Village is a *matrix* of village fixed effects. Other pretreatment covariates are included through the matrix $\mathbf{X}$, such as age, whether they were a farmer, herder, merchant or trader, their household size and whether or not they voted in the past. The results of this regression show that, on average, exposure to violence (DirectHarm) is associated with more pro-peace attitudes on PeaceIndex.

Despite these arguments, not all investigators may agree with the assumption of no unobserved confounders. Consider, for example, a fellow researcher who argues that, although bombings were impossible to target finely, perhaps those in the centre of the village were more often harmed than those on the periphery. And might not those nearer the centre of each village also have different types of attitude towards peace, on average? This suggests that the author ought instead to have run the model

$$\text{PeaceIndex} = \hat{\tau}\,\text{DirectHarm} + \hat{\beta}_f\,\text{Female} + \text{Village}\,\hat{\boldsymbol{\beta}}_v + \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\gamma}\,\text{Center} + \hat{\varepsilon}_{\text{full}}, \tag{2}$$

i.e. our earlier estimate $\hat{\tau}_{\text{res}}$ would differ from our target quantity $\hat{\tau}$: but how badly? How 'strong' would a confounder like Center need to be to change our research conclusions? A simple violation of unconfoundedness such as this can be handled in a relatively straightforward manner by the traditional OVB framework, as we shall see in Section 3.

However, other sceptical researchers may question the claim that violence was conditionally indiscriminate with more elaborate stories, worrying that unobserved factors such as Wealth or PoliticalAttitudes remain as confounders, perhaps even acting through non-linear functions such as an interaction of these two. Additionally, we may also have domain knowledge about the determinants of the outcome or the treatment assignment that could be used to limit arguments about potential confounding. For example, considering the nature of the attacks and the special role that gender played, one may argue that, within village, confounders are not likely to be as strongly associated with the treatment as is the observed covariate Female.

How strong would these confounders need to be (acting as a group, possibly with non-linearities) to change our conclusions? And how could we *codify* and *leverage* our beliefs about the relative importance of Female to bound the plausible strength of unobserved confounders? In Sections 4 and 5, we show how extending the traditional OVB framework provides answers to such questions.

3. Sensitivity in an omitted variable bias framework

The OVB formula is an important part of the mechanics of linear regression models and describes how the inclusion of an omitted covariate changes a coefficient estimate of interest. In this section, we review the traditional OVB approach and illustrate its use as a simple tool for sensitivity analysis through bivariate contour plots showing how the effect estimate would vary depending on hypothetical strengths of the confounder. This serves not only as an introduction to the method, but also to highlight limitations that we shall address in the following sections.

3.1. The traditional omitted variable bias

Suppose that an investigator wishes to run a linear regression model of an outcome Y on a treatment D, controlling for a set of covariates given by $\mathbf{X}$ and Z, as in

$$Y = \hat{\tau} D + \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\gamma} Z + \hat{\varepsilon}_{\text{full}} \tag{3}$$

where Y is an $n \times 1$ vector containing the outcome of interest for each of the n observations and D is an $n \times 1$ treatment variable (which may be continuous or binary); $\mathbf{X}$ is an $n \times p$ matrix of *observed* (pretreatment) covariates including the constant; and Z is a single $n \times 1$ *unobserved* covariate (we allow a multivariate version of Z in Section 4.5). However, since Z is unobserved, the investigator is forced instead to estimate a restricted model:

$$Y = \hat{\tau}_{\text{res}} D + \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{res}} + \hat{\varepsilon}_{\text{res}} \tag{4}$$

where $\hat{\tau}_{\text{res}}$ and $\hat{\boldsymbol{\beta}}_{\text{res}}$ are the coefficient estimates of the restricted ordinary least squares with only D and $\mathbf{X}$, *omitting* Z, and $\hat{\varepsilon}_{\text{res}}$ its corresponding residual.

How does the observed estimate $\hat{\tau}_{\text{res}}$ compare with the desired estimate $\hat{\tau}$? Define as $\widehat{\text{bias}}$ the difference between these estimates: $\widehat{\text{bias}} := \hat{\tau}_{\text{res}} - \hat{\tau}$, where the circumflex clarifies that this quantity is a difference between sample estimates, not the difference between the expectation of a sample estimate and a population value. Using the Frisch–Waugh–Lovell theorem (Frisch and Waugh, 1933; Lovell, 1963, 2008) to 'partial out' the observed covariates $\mathbf{X}$, the classical OVB solution is

$$\begin{aligned}\hat{\tau}_{\text{res}} &= \frac{\text{cov}(D^{\perp\mathbf{X}}, Y^{\perp\mathbf{X}})}{\text{var}(D^{\perp\mathbf{X}})} \\ &= \frac{\text{cov}(D^{\perp\mathbf{X}}, \hat{\tau} D^{\perp\mathbf{X}} + \hat{\gamma} Z^{\perp\mathbf{X}})}{\text{var}(D^{\perp\mathbf{X}})} \\ &= \hat{\tau} + \hat{\gamma} \frac{\text{cov}(D^{\perp\mathbf{X}}, Z^{\perp\mathbf{X}})}{\text{var}(D^{\perp\mathbf{X}})} \\ &= \hat{\tau} + \hat{\gamma}\hat{\delta} \end{aligned} \tag{5}$$

where $\text{cov}(\cdot)$ and $\text{var}(\cdot)$ denote the *sample* covariance and variance; $Y^{\perp\mathbf{X}}$, $D^{\perp\mathbf{X}}$ and $Z^{\perp\mathbf{X}}$ are the variables Y, D and Z after removing the components linearly explained by $\mathbf{X}$ and we define $\hat{\delta} := \text{cov}(D^{\perp\mathbf{X}}, Z^{\perp\mathbf{X}})/\text{var}(D^{\perp\mathbf{X}})$. We then have

$$\widehat{\text{bias}} = \hat{\gamma}\hat{\delta}. \tag{6}$$

Although elementary, the OVB formula in equation (6) provides the key intuitions as well as a formulaic basis for a simple sensitivity analysis, enabling us to assess how the omission of covariates that we wished to have controlled for could affect our inferences. Note that it holds *whether or not equation (3) has a causal meaning*. In applied settings, however, we are typically interested in cases where the investigator has determined that the full regression, controlling for *both* $\mathbf{X}$ *and* the unobserved variable Z, would have identified the causal effect of D on Y; thus, hereafter we shall treat Z as an unobserved 'confounder' and continue the discussion as if the estimate $\hat{\tau}$, obtained with the inclusion of Z, is the desired target quantity. (We remind readers that conditions that endow regression estimates with causal meaning are extensively discussed in the literature: identification assumptions can be articulated in graphical terms, such as postulating a structural causal model in which $\{\mathbf{X}, Z\}$ satisfy the backdoor criterion for identifying the causal effect of D on Y (Pearl, 2009), or, equivalently, in counterfactual notation, stating that the treatment assignment D is conditionally ignorable given $\{\mathbf{X}, Z\}$, i.e. $Y_d \perp\!\!\!\perp D | \mathbf{X}, Z$, where Y_d denotes the potential outcome of Y when D is *set* to d (see Pearl (2009), Angrist and Pischke (2008) and Imbens and Rubin (2015)). We further note that the effect of D on Y may be non-linear, in which case a regression coefficient may be an incomplete summary of the causal effect (Angrist and Pischke, 2008). Finally, indiscriminate inclusion of covariates can induce or amplify bias (see Pearl (2011), Ding and Miratrix (2015), Middleton *et al.* (2016) and Steiner and Kim (2016) for related discussions). Here we assume that the researcher is interested in the estimates that one would obtain from running the regression in equation (3), controlling for $\mathbf{X}$ and Z.)

3.2. Making sense of the traditional omitted variable bias

One virtue of the OVB formula is its interpretability. The quantity $\hat{\gamma}$ describes the difference in the linear expectation of the outcome, when comparing individuals who differ by one unit on the confounder, but have the same treatment assignment status as well as the same value

for all remaining covariates. In broader terms, $\hat{\gamma}$ describes how looking at different subgroups of the unobserved confounder 'impacts' our best linear prediction of the outcome. Although a causal interpretation here is tempting, whether this difference in the distribution of the outcome within strata of the confounder can be attributed to a direct causal effect of the unobserved confounder on the outcome depends on structural assumptions. In many scenarios, however, this is unrealistic—since the researcher's goal is to estimate the causal effect of D on Y, usually Z is required only, along with $\mathbf{X}$, to block the back-door paths from D to Y (Pearl, 2009) or, equivalently, to make the treatment assignment conditionally ignorable. In this case, $\hat{\gamma}$ could reflect not only its causal effect on Y (if it has any) but also other spurious associations that are not eliminated by standard assumptions. Heuristically, however, referring to $\hat{\gamma}$ as the marginal 'impact' of the confounder on the outcome is useful, as long as the reader keeps in mind that it is an associational quantity with causal meaning only under certain circumstances.

By analogy, it would be tempting to think of $\hat{\delta}$ as the estimated marginal impact of the confounder on the *treatment*. However, causal interpretation aside, this is incorrect because it refers instead to the coefficient of the reverse regression, $Z = \hat{\delta}D + \mathbf{X}\hat{\psi} + \hat{\varepsilon}_Z$, and not the regression of the treatment D on Z, and $\mathbf{X}$, i.e. $\hat{\delta}$ gives the difference in the linear expectation of the confounder, when comparing individuals with the same values for the covariates, but differing by one unit on the treatment. This quantity will be familiar to empirical researchers who have used quasi-experiments in which the treatment is believed to be randomized only conditionally on certain covariates $\mathbf{X}$. In that case we may then check for 'balance' on other (pretreatment) observables once conditioning is complete. Hence, we can think of $\hat{\delta}$ as the (conditional) imbalance of the confounder with respect to the treatment—or simply 'imbalance'.

Thus, a useful mnemonic is that the omitted variable bias can be summarized as the unobserved confounder's 'impact times its imbalance'. Note that the imbalance component is quite general: whatever the true functional form dictating $\mathbb{E}[Z|D, \mathbf{X}]$ (or the treatment assignment mechanism), the only way in which Z's relationship to D enters the bias is captured by its 'linear imbalance', parameterized by $\hat{\delta}$. In other words, the linear regression of Z on D and $\mathbf{X}$ need not reflect the correct expected value of Z—rather it serves to capture the aspects of the relationship between Z and D that affects the bias.

3.3. Using the traditional omitted variable bias for sensitivity analysis

If we know the *signs* of the partial correlations between the confounder with the treatment and the outcome (the same as the signs of $\hat{\gamma}$ and $\hat{\delta}$) we can argue whether our estimate is likely to be underestimating or overestimating the quantity of interest. Arguments using correlational direction are common practice in econometrics work (for an example, see Angrist and Pischke (2017), pages 8–9). Often, though, discussing possible direction of the bias is not possible or not sufficient, and magnitude must be considered. How strong would the confounder(s) have to be to change the estimates in such a way to affect the main conclusions of a study?

3.3.1. Sensitivity contour plots

A first approach to investigate the sensitivity of our estimate can be summarized by a two-dimensional plot of bias contours parameterized by the two terms $\hat{\gamma}$ and $\hat{\delta}$. Each pair of hypothesized impact and imbalance parameters corresponds to a certain level of bias (their product) but, given an initial treatment effect estimate $\hat{\tau}_{\text{res}}$, we can also relabel the bias levels in terms of the 'adjusted' effect estimate, i.e $\hat{\tau} = \hat{\tau}_{\text{res}} - \hat{\gamma}\hat{\delta}$: the estimate from the ordinary least squares regression that we wish we had run, if we had included a confounder with the hypothesized level of impact and imbalance.

In our running example, a specific confounder that we wish we had controlled for is a binary indicator of whether the respondent lived in the centre or in the periphery of the village. How strong would this specific confounder have to be for its inclusion to affect our conclusions substantially? Fig. 1 shows the plot of adjusted estimates for several hypothetical values of impact and imbalance of the confounder Center.

Hypothetical values for the imbalance of the confounder lie on the horizontal axis. In this particular case, they indicate how those who were harmed are hypothesized to differ from those who were not harmed in terms of the proportion of people living in the centre of the village. Values for the hypothetical effect of the confounder on the outcome lie on the vertical axis, representing how attitudes towards peace differ on average for people living in the centre *versus* those in the periphery of the village, within strata of other covariates. The contour lines of the plot give the adjusted treatment effect at hypothesized values of the impact and imbalance parameters. They show the exact estimate that we would have obtained by running the full regression including a confounder with those hypothetical sensitivity parameters. No other information is required to know how such a confounder would influence the result. Note that here, and throughout the paper, we parameterize the bias in a way that it hurts our preferred hypothesis by reducing the absolute effect size. (Investigators may also argue that accounting for OVB would increase the effect size. Our tools apply to these cases as well; the arguments would just work in the opposite direction.)

This plot explicitly reveals the type of prior knowledge that we need to have to be able to rule out problematic confounders. As an example, imagine that the confounder Center has a conditional imbalance as high as 0.25—i.e., having controlled for the observed covariates, those who were physically injured were also 25 percentage points more likely to live in the centre of the village than those who were not. With such an imbalance, the plot reveals that the effect of

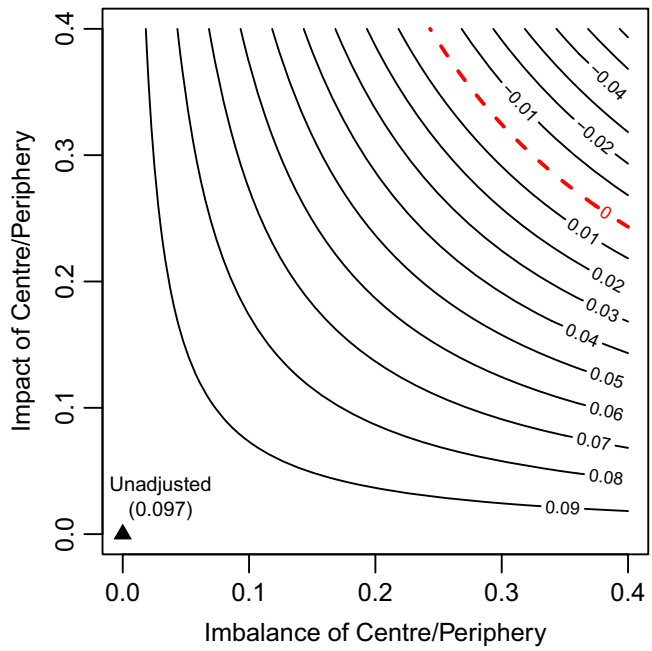


Fig. 1. Sensitivity contours of the point estimate—traditional OVB

living in the centre on the outcome (PeaceIndex) would have to be over 0.40 to bring down the estimated effect of DirectHarm to 0.

Determining whether this is good or bad news remains difficult and requires contextual knowledge about the process that generated the data. For instance, one could argue that, given the relatively homogeneous nature of these small villages and that their centres are generally not markedly different in composition from the peripheries, it is difficult to believe that being in the centre was associated with a 0.40 higher expected score on PeaceIndex (which varies only from 0 to 1). Regardless of whether the investigator can make a clear argument that rules out such confounders, the virtue of sensitivity analysis is that it moves the conversation from one where the investigator seeks to defend 'perfect identification' and the critic points out potential confounders, to one where details can be given and discussed about the degree of confounding that would be problematic.

3.3.2. Shortcomings of the traditional omitted variable bias

The traditional OVB has some benefits: as shown, with sound substantive knowledge about the problem, it is a straightforward exercise. But it also has shortcomings. In the previous example, Center was a convenient choice of confounder because it is a binary variable, and the units of measure attached to impact and imbalance are thus easy to understand as changes in proportions. This is not in general so. Imagine contemplating confounders such as PoliticalAttitudes: in what scale should we measure this? A doubling of that scale would halve the required impact and double the required imbalance. A possible solution is to standardize the coefficients, but this does not help if the goal is to assess the sensitivity of the causal parameter in its original scale.

Furthermore, the traditional OVB, be it standardized or not, does not generalize easily to multiple confounders: how should we assess the effect of confounders PoliticalAttitudes and Wealth, acting together, perhaps with complex non-linearities? Or, more generally, how should we consider all the other unnamed confounders acting together? Can we benchmark all these confounders against Female? Finally, how can we obtain the sensitivity of not only the point estimate, but also the standard errors, so that we could examine t-values, p-values or confidence intervals under hypothetical confounders?

4. Omitted variable bias with the partial R^2 parameterization

We now consider a reparameterization of the OVB formula in terms of partial R^2 values. Our goal is to replace the sensitivity parameters $\hat{\gamma}$ and $\hat{\delta}$ with a pair of parameters that uses an R^2-measure to assess the strength of association between the confounder and the treatment and between the confounder and the outcome, both assuming that the remaining covariates $\mathbf{X}$ have been accounted for. The partial R^2 parameterization is scale free and it further enables us to construct some useful analyses, including

(a) assessing the sensitivity of an estimate to any number or even *all* confounders acting together, possibly non-linearly,
(b) using the same framework to assess the sensitivity of point estimates as well as t-values and confidence intervals,
(c) assessing the sensitivity to extreme scenarios in which all or a big portion of the unexplained variance of the outcome is due to confounding,
(d) applying contextual information about the research design to bound the strength of the confounders and
(e) presenting these sensitivity results concisely for easy routine reporting, as well as providing visual tools for finer-grained analysis.

4.1. Reparameterizing the bias in terms of partial R^2

Let $R^2_{Z\sim D}$ denote the (sample) R^2 of regressing Z on D. Recall that for ordinary least squares the following result holds:

$$R^2_{Z\sim D} = \text{var}(\hat{Z})/\text{var}(Z) = 1 - \text{var}(Z^{\perp D})/\text{var}(Z) = \text{corr}(Z, \hat{Z})^2 = \text{corr}(Z, D)^2,$$

where $\hat{Z}$ are the fitted values given by regressing Z on D. Note that the R^2 is symmetric, i.e. it is invariant to whether we use the 'forward' or the 'reverse' regression since $R^2_{Z\sim D} = \text{corr}(Z, D)^2 = \text{corr}(D, Z)^2 = R^2_{D\sim Z}$. Extending this to the case with covariates $\mathbf{X}$, we denote the partial R^2 from regressing Z on D after controlling for $\mathbf{X}$ as $R^2_{Z\sim D|\mathbf{X}}$. This has the same useful symmetry, with $R^2_{Z\sim D|\mathbf{X}} = 1 - \text{var}(Z^{\perp \mathbf{X}, D})/\text{var}(Z^{\perp \mathbf{X}}) = \text{corr}(Z^{\perp \mathbf{X}}, D^{\perp \mathbf{X}})^2 = \text{corr}(D^{\perp \mathbf{X}}, Z^{\perp \mathbf{X}})^2 = R^2_{D\sim Z|\mathbf{X}}$.

We are now ready to express the bias in terms of partial R^2. First, by the Frisch–Waugh–Lovell theorem,

$$\begin{aligned}
\widehat{\text{bias}} &= \hat{\delta}\hat{\gamma} \\
&= \frac{\text{cov}(D^{\perp \mathbf{X}}, Z^{\perp \mathbf{X}})}{\text{var}(D^{\perp \mathbf{X}})} \frac{\text{cov}(Y^{\perp \mathbf{X}, D}, Z^{\perp \mathbf{X}, D})}{\text{var}(Z^{\perp \mathbf{X}, D})} \\
&= \frac{\text{corr}(D^{\perp \mathbf{X}}, Z^{\perp \mathbf{X}})\,\text{sd}(Z^{\perp \mathbf{X}})}{\text{sd}(D^{\perp \mathbf{X}})} \frac{\text{corr}(Y^{\perp \mathbf{X}, D}, Z^{\perp \mathbf{X}, D})\,\text{sd}(Y^{\perp \mathbf{X}, D})}{\text{sd}(Z^{\perp \mathbf{X}, D})} \\
&= \frac{\text{corr}(Y^{\perp \mathbf{X}, D}, Z^{\perp \mathbf{X}, D})\,\text{corr}(D^{\perp \mathbf{X}}, Z^{\perp \mathbf{X}})}{\text{sd}(Z^{\perp \mathbf{X}, D})/\text{sd}(Z^{\perp \mathbf{X}})} \frac{\text{sd}(Y^{\perp \mathbf{X}, D})}{\text{sd}(D^{\perp \mathbf{X}})}.
\end{aligned} \tag{7}$$

Noting that $\text{corr}(Y^{\perp \mathbf{X}, D}, Z^{\perp \mathbf{X}, D})^2 = R^2_{Y\sim Z|\mathbf{X}, D}$, that $\text{corr}(Z^{\perp \mathbf{X}}, D^{\perp \mathbf{X}})^2 = R^2_{D\sim Z|\mathbf{X}}$ and that $\text{var}(Z^{\perp \mathbf{X}, D})/\text{var}(Z^{\perp \mathbf{X}}) = 1 - R^2_{Z\sim D|\mathbf{X}} = 1 - R^2_{D\sim Z|\mathbf{X}}$, we can write equation (7) as

$$|\widehat{\text{bias}}| = \sqrt{\left(\frac{R^2_{Y\sim Z|D,\mathbf{X}} R^2_{D\sim Z|\mathbf{X}}}{1 - R^2_{D\sim Z|\mathbf{X}}}\right)} \frac{\text{sd}(Y^{\perp \mathbf{X}, D})}{\text{sd}(D^{\perp \mathbf{X}})}. \tag{8}$$

Equation (8) rewrites the OVB formula in terms that more conveniently rely on partial R^2 measures of association rather than raw regression coefficients. Investigators may be interested in how confounders alter inference as well, so we also examine the standard error. Let df denote the regression's degrees of freedom (for the restricted regression actually run). Noting that

$$\text{se}(\hat{\tau}_{\text{res}}) = \frac{\text{sd}(Y^{\perp \mathbf{X}, D})}{\text{sd}(D^{\perp \mathbf{X}})} \sqrt{\left(\frac{1}{\text{df}}\right)}, \tag{9}$$

$$\text{se}(\hat{\tau}) = \frac{\text{sd}(Y^{\perp \mathbf{X}, D, Z})}{\text{sd}(D^{\perp \mathbf{X}, Z})} \sqrt{\left(\frac{1}{\text{df} - 1}\right)}, \tag{10}$$

whose ratio is

$$\frac{\text{se}(\hat{\tau})}{\text{se}(\hat{\tau}_{\text{res}})} = \frac{\text{sd}(Y^{\perp \mathbf{X}, D, Z})}{\text{sd}(Y^{\perp \mathbf{X}, D})} \frac{\text{sd}(D^{\perp \mathbf{X}})}{\text{sd}(D^{\perp \mathbf{X}, Z})} \sqrt{\left(\frac{\text{df}}{\text{df} - 1}\right)}, \tag{11}$$

we obtain the expression for the standard error of $\hat{\tau}$:

$$\mathrm{se}(\hat{\tau}) = \mathrm{se}(\hat{\tau}_{\mathrm{res}})\sqrt{\left(\frac{1 - R^2_{Y\sim Z|D,\mathbf{X}}}{1 - R^2_{D\sim Z|\mathbf{X}}}\frac{\mathrm{df}}{\mathrm{df}-1}\right)}. \tag{12}$$

Moreover, with this we can further see the bias as

$$\widehat{|\mathrm{bias}|} = \mathrm{se}(\hat{\tau}_{\mathrm{res}})\sqrt{\left(\frac{R^2_{Y\sim Z|D,\mathbf{X}} R^2_{D\sim Z|\mathbf{X}}}{1 - R^2_{D\sim Z|\mathbf{X}}}\mathrm{df}\right)}. \tag{13}$$

4.2. Making sense of the partial R^2 parameterization

Equations (12) and (13) form the basis of the sensitivity exercises regarding both the point estimate and the standard error, with sensitivity parameters in terms of $R^2_{Y\sim Z|D,\mathbf{X}}$ and $R^2_{D\sim Z|\mathbf{X}}$. These formulae are computationally convenient—the only data-dependent parts are the standard error of $\hat{\tau}_{\mathrm{res}}$ and the regression's degrees of freedom, which are already reported by most regression software. In this section, we provide remarks that help to make sense of these results, revealing their simplicity in terms of regression anatomy. We also review some partial R^2 identities that may prove useful when reasoning about the sensitivity parameters.

4.2.1. Sensitivity of the point estimate

In the partial R^2 parameterization, the relative bias $|\widehat{\mathrm{bias}}/\hat{\tau}_{\mathrm{res}}|$ has a simple form (see the on-line supplement section A for details):

$$\text{relative bias} = \frac{\overbrace{|R_{Y\sim Z|D,\mathbf{X}} f_{D\sim Z|\mathbf{X}}|}^{\text{bias factor}}}{\underbrace{|f_{Y\sim D|\mathbf{X}}|}_{\text{partial } f \text{ of } D \text{ with } Y}} = \frac{\mathrm{BF}}{|f_{Y\sim D|\mathbf{X}}|}. \tag{14}$$

The numerator of the relative bias contains the partial Cohen's f of the confounder with the treatment, amortized by the partial correlation of that confounder with the outcome. (Cohen's f^2 can be written as $f^2 = R^2/(1-R^2)$, so, for example, $f^2_{D\sim Z|\mathbf{X}} = R^2_{D\sim Z|\mathbf{X}}/(1 - R^2_{D\sim Z|\mathbf{X}})$.) Collectively this numerator could be called the bias factor of the confounder: $\mathrm{BF} = |R_{Y\sim Z|D,\mathbf{X}} f_{D\sim Z|\mathbf{X}}|$, which is determined entirely by the two sensitivity parameters $R^2_{Y\sim Z|D,\mathbf{X}}$ and $R^2_{D\sim Z|\mathbf{X}}$. To determine the size of the relative bias, this is compared with how much variation of the outcome is uniquely explained by the treatment assignment, in the form of the partial Cohen's f of the treatment with the outcome. Computationally, $f_{Y\sim D|\mathbf{X}}$ can be obtained by dividing the t-value of the treatment coefficient by the square root of the regression's degrees of freedom—$f_{Y\sim D|\mathbf{X}} = t_{\hat{\tau}_{\mathrm{res}}}/\sqrt{\mathrm{df}}$. This enables us to assess easily the sensitivity to any confounder with a given pair of partial R^2 values; see Table 2 in the on-line supplement section D for an illustration procedure.

Equation (14) also reveals that, given a particular confounder (which will fix BF), the only property that is needed to determine the robustness of a regression estimate against that confounder is the partial R^2 of the treatment with the outcome (via $f_{Y\sim D|\mathbf{X}}$). This serves to reinforce the fact that robustness to confounding is an identification problem, impervious to sample size considerations. Whereas t-values and p-values might be informative with respect to the statistical uncertainty (in a correctly specified model), robustness to misspecification is determined by the share of variation of the outcome that the treatment uniquely explains.

A subtle but useful property of the partial R^2 parameterization is that it reveals an asymmetry in the role of the components of the bias factor. In the traditional OVB formulation, the

bias is simply a product of two terms with the same importance. The new formulation breaks this symmetry: the effect of the partial R^2 of the confounder with the outcome on the bias factor is bounded at 1. By contrast, the effect of the partial R^2 of the confounder with the treatment on the bias factor is unbounded (via $f_{D\sim Z|\mathbf{X}}$). This enables us to consider extreme scenarios, in which we suppose that the confounder explains *all* of the left-out variation of the outcome, and to see what happens as we vary the partial R^2 of the confounder with the treatment (Section 5.3).

4.2.2. *Sensitivity of the variance*

How the confounder affects the variance has a straightforward interpretation as well. The relative change in the variance, $\text{var}(\hat{\tau})/\text{var}(\hat{\tau}_{\text{res}})$, can be decomposed into three components:

$$\begin{aligned}\text{relative change in variance} &= \overbrace{(1 - R^2_{Y\sim Z|D,\mathbf{X}})}^{\text{VRF}} \underbrace{\frac{1}{1 - R^2_{D\sim Z|\mathbf{X}}}}_{\text{VIF}} \overbrace{\frac{\text{df}}{\text{df}-1}}^{\text{change in df}} \\ &= \text{VRF} \times \text{VIF} \times \text{change in df}, \end{aligned} \tag{15}$$

i.e. including the confounder in the regression reduces the variance of the coefficient of D by reducing the residual variance of Y (the variance reduction factor—VRF). In contrast, it raises the variance of the coefficient via its partial correlation with the treatment (the traditional variance inflation factor—VIF). Finally, the degrees of freedom must be adjusted to recover formally the answer that we would obtain from including the omitted variable. The overall relative change of the variance is simply the product of these three components.

4.2.3. *Reasoning about $R^2_{Y\sim Z|D,\mathbf{X}}$ and $R^2_{D\sim Z|\mathbf{X}}$*

For simplicity of exposition, throughout the paper we reason in terms of the sensitivity parameters $R^2_{Y\sim Z|D,\mathbf{X}}$ and $R^2_{D\sim Z|\mathbf{X}}$ directly. However, here we recall some identities of the partial R^2 scale that can aid interpretation depending on what can best be reasoned about in a given applied setting.

First, as noted in Section 4.1, researchers who are accustomed to thinking about or evaluating the strength of (partial) correlations can simply square those values to reason with the corresponding partial R^2s. Next, in some circumstances, researchers might prefer to reason about the relationship of the unobserved confounder Z and the outcome Y *without conditioning on the treatment assignment* D. (For instance, since D will usually be a *post-treatment* variable with respect to Z, this can make the association of Y and Z conditional on D more difficult to interpret, especially when we want to attach a causal meaning to the parameter (Rosenbaum, 1984). As argued in Section 3.2, however, recall that a causal interpretation of the association of Z with Y requires more assumptions than those usually invoked for the identification of the causal effect of D on Y.) This can be done by noting that, for a choice of $R_{Y\sim Z|\mathbf{X}}$ and $R_{D\sim Z|\mathbf{X}}$, we can reconstruct $R_{Y\sim Z|D,\mathbf{X}}$ by using the recursive definition of partial correlations:

$$R_{Y\sim Z|D,\mathbf{X}} = \frac{R_{Y\sim Z|\mathbf{X}} - R_{Y\sim D|\mathbf{X}} R_{D\sim Z|\mathbf{X}}}{\sqrt{(1 - R^2_{Y\sim D|\mathbf{X}})}\sqrt{(1 - R^2_{D\sim Z|\mathbf{X}})}}. \tag{16}$$

Therefore, if needed, we can reason directly about sensitivity parameters $R^2_{Y\sim Z|\mathbf{X}}$ and $R^2_{D\sim Z|\mathbf{X}}$.

Finally, it may be beneficial to reason in terms of how much explanatory power is added by including confounders. For this, recall that the partial R^2s are defined as

$$R^2_{Y\sim Z|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+Z} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}},$$
$$R^2_{D\sim Z|\mathbf{X}} = \frac{R^2_{D\sim \mathbf{X}+Z} - R^2_{D\sim \mathbf{X}}}{1 - R^2_{D\sim \mathbf{X}}}, \tag{17}$$

i.e. plausibility judgements about the partial R^2 boil down to plausibility judgements about the *total (or added) explanatory power* that we would have obtained in the treatment and the outcome regressions, if the unobserved confounder Z had been included. This may be particularly useful when contemplating multiple confounders acting in concert (as we shall discuss in Section 4.5), in which case other parameterizations (such as simple correlations or regression coefficients) become unwieldy.

4.3. Sensitivity statistics for routine reporting

Detailed sensitivity analyses can be conducted by using the previous results, as we shall show in the next section. However, widespread adoption of sensitivity analyses would benefit from simple measures that quickly describe the overall sensitivity of an estimate to unobserved confounding. These measures serve two main purposes:

(a) they can be routinely reported in standard regression tables, making the discussion of sensitivity to unobserved confounding more accessible and standardized;
(b) they can be easily computed from quantities found in a regression table, enabling readers and reviewers to initiate the discussion about unobserved confounders when reading papers that did not formally assess sensitivity.

4.3.1. The robustness value

The first quantity that we propose is the *robustness value* RV, which conveniently summarizes the types of confounders that would problematically change the research conclusions. Consider a confounder with equal association to the treatment and the outcome, i.e. $R^2_{Y\sim Z|\mathbf{X},D} = R^2_{D\sim Z|\mathbf{X}} = \mathrm{RV}_q$, where RV_q describes how strong that association must be to reduce the estimated effect by $100q\%$. By equation (14) (see the on-line supplement section A),

$$\mathrm{RV}_q = \tfrac{1}{2}\{\sqrt{(f_q^4 + 4f_q^2)} - f_q^2\} \tag{18}$$

where $f_q := q|f_{Y\sim D|\mathbf{X}}|$ is the partial Cohen's f of the treatment with the outcome multiplied by the proportion of reduction q on the treatment coefficient which would be deemed problematic. Confounders that explain $\mathrm{RV}_q\%$ both of the treatment and of the outcome are sufficiently strong to change the point estimate in problematic ways, whereas confounders with neither association greater than $\mathrm{RV}_q\%$ are not.

The robustness value thus offers an interpretable sensitivity measure that summarizes how robust the point estimate is to unobserved confounding. A robustness value that is close to 1 means that the treatment effect can handle strong confounders explaining almost all residual variation of the treatment and the outcome. In contrast, a robustness value that is close to 0 means that even very weak confounders could eliminate the results. Note that the robustness value can be easily computed from any regression table, recalling that $f_{Y\sim D|\mathbf{X}}$ can be obtained by simply dividing the treatment coefficient t-value by $\sqrt{\mathrm{df}}$.

With minor adjustment, robustness values can also be obtained for t-values, or lower and upper bounds of confidence intervals. Let $|t^*_{\alpha,\,\mathrm{df}-1}|$ denote the t-value threshold for a t-test with level of significance α and $\mathrm{df}-1$ degrees of freedom, and define $f^*_{\alpha,\,\mathrm{df}-1} := |t^*_{\alpha,\,\mathrm{df}-1}|/\sqrt{(\mathrm{df}-1)}$. Now construct an adjusted $f_{q,\alpha}$, accounting for both the proportion of reduction q of the point estimate and the boundary below which statistical significance is lost at the level of α:

$$f_{q,\alpha} := q|f_{Y\sim D|\mathbf{X}}| - f^*_{\alpha,\,\mathrm{df}-1}. \tag{19}$$

If $f_{q,\alpha} < 0$, then the robustness value is 0. If $f_{q,\alpha} > 0$, then a confounder with a partial R^2 of

$$\mathrm{RV}_{q,\alpha} = \tfrac{1}{2}\{\sqrt{(f^4_{q,\alpha} + 4f^2_{q,\alpha})} - f^2_{q,\alpha}\}, \tag{20}$$

both with the treatment and with the outcome, is sufficiently strong to make the adjusted t-test not reject the hypothesis $H_0 : \tau = (1-q)|\hat{\tau}_{\mathrm{res}}|$ at the α-level (or, equivalently, to make the adjusted $1-\alpha$ confidence interval include $(1-q)|\hat{\tau}_{\mathrm{res}}|$). When $\mathrm{RV}_{q,\,\alpha} > 1 - 1/f^2_q$ then, as with RV_q, we can conclude that no confounder with both associations lower than $\mathrm{RV}_{q,\,\alpha}$ can overturn the conclusion of such a test. In the rare cases when $\mathrm{RV}_{q,\,\alpha} \leqslant 1 - f^2_q$, setting $\mathrm{RV}_{q,\,\alpha} = (f^2_q - f^{*2}_{\alpha,\,\mathrm{df}-1})/(1 + f^2_q)$ restores the property that no confounder that is weaker on both associations would change the conclusion. Since we are considering sample uncertainty, $\mathrm{RV}_{q,\alpha}$ is a more conservative measure than RV_q. If we pick $|t^*_{\alpha,\,\mathrm{df}-1}| = 0$ then $\mathrm{RV}_{q,\alpha}$ reduces to RV_q. Also, for fixed $|t^*_{\alpha,\,\mathrm{df}-1}|$, $\mathrm{RV}_{q,\alpha}$ converges to RV_q when the sample size grows to ∞. See the on-line supplement section A for details.

4.3.2. *The $R^2_{Y\sim D|\mathbf{X}}$ as an extreme scenario analysis*

The second measure that we propose is the proportion of variation in the outcome that is uniquely explained by the treatment—$R^2_{Y\sim D|\mathbf{X}}$. Consider the following question: 'if an extreme confounder explained all the residual variance of the outcome, how strongly associated with the treatment would it need to be to eliminate the estimated effect?'. As it happens, the answer is precisely $R^2_{Y\sim D|\mathbf{X}}$.

Specifically, a confounder explaining *all* residual variance of the outcome implies that $R_{Y\sim Z|D,\mathbf{X}} = 1$. By equation (14), to bring the estimated effect down to 0 (relative bias 1), this means that $|f_{D\sim Z|\mathbf{X}}|$ needs to equal $|f_{Y\sim D|\mathbf{X}}|$, which implies that $R^2_{D\sim Z|\mathbf{X}} = R^2_{Y\sim D|\mathbf{X}}$. Thus, $R^2_{Y\sim D|\mathbf{X}}$ is not only the determinant of the robustness of the treatment effect coefficient but can also be interpreted as the result of an 'extreme scenario' sensitivity analysis.

4.4. Bounding the strength of the confounder by using observed covariates

Arguably, the most difficult part of a sensitivity analysis is taking the description of a confounder that would be problematic from the formal results, and reasoning about whether a confounder with such strength plausibly exists in one's study, given its design and the investigator's contextual knowledge. In this section, we introduce a novel bounding approach that can help to alleviate this difficulty. The rationale for the method is the realization that, although in some cases an investigator may not be able to make direct plausibility judgements about the strength of an unobserved confounder Z, she might still have grounds to make judgements about *its relative strength*, for instance, claiming that Z cannot possibly account for as much variation of the treatment assignment as some observed covariate X. How can we formally codify and leverage these claims regarding relative strength (or importance) of covariates for sensitivity analysis?

Clearly, there is not a unique way to measure the relative strength of variables (Kruskal and Majors, 1989). For the task at hand, however, any proposal must meet the minimal criterion of

solving the correct identification problem—essentially, this means that the chosen measure of relative strength must be sufficient to identify (or bound) the bias, and a new function (or bound) in terms of that measure must be derived (Cinelli *et al.*, 2019). Previous work has proposed informal benchmarking procedures that fail this minimal criterion and can generate misleading sensitivity analysis results, even if researchers had correct knowledge about the relative strength of Z (Frank, 2000; Imbens, 2003; Frank *et al.*, 2008; Blackwell, 2013; Dorie *et al.*, 2016; Carnegie *et al.*, 2016a; Middleton *et al.*, 2016). We elaborate on the pitfalls of this informal approach in Section 6.2 of the discussion.

Additionally, simply obtaining a formal identification result is not enough for it to be useful in applied settings—investigators must still be able to reason cogently about whether confounders are 'stronger' than observed covariates by using the chosen measure of relative strength. Since this depends on context, it is highly desirable to have a variety of measures for those relative comparisons (allowing researchers to choose those that are best suited for a given analysis) and that those measures have relevant interpretations (Kruskal and Majors, 1989). An example of the risks that are entailed by ignoring this requirement can be found in the coefficient of 'proportional selection on observables' that was advanced by Oster (2019), which will be discussed in Section 6.3.

With this in mind, here we offer three main alternatives to bound the strength of the unobserved confounder, by judging:

(a) how the *total* R^2 of the confounder compares with the total R^2 of a group of observed covariates;
(b) how the *partial* R^2 of the confounder compares with the partial R^2 of a group of observed covariates, having taken into account the explanatory power of remaining observed covariates, or
(c) how the partial R^2 of the confounder compares with the partial R^2 of a group of observed covariates, having taken into account the explanatory power of remaining observed covariates *and* the treatment assignment.

These are natural measures of relative importance for ordinary least squares and can be interpreted as comparisons of the consequences of dropping a (group of) variable(s) in variance reduction or prediction error (Kruskal and Majors, 1989).

The choice of bounding procedures that we should use depends on which of these quantities the investigator prefers and can most soundly reason about in their own research. In our running example, within a given village, one may argue that Female is the most important visible characteristic that could be used for exposure to violence, and it probably explains more of the residual variation in targeting than could any unobserved confounder. For this reason (as well as simplicity of exposition) in the main text we illustrate the use of the third type of bound, and we refer readers to the on-line supplement section B for further discussion and derivations of the other two variants. (Another reason that we employ this type of bound in the main text is that it is most closely related to approaches that are used by other sensitivity analyses with which we contrast our results. These include the informal benchmarks of Imbens (2003) as well as to the bounding proposal of Oster (2019), discussed in Section 6.)

Assume that $Z \perp \mathbf{X}$, or, equivalently, consider only the part of Z that is not linearly explained by $\mathbf{X}$. Now suppose that the researcher believes she has measured the key determinants of the outcome and treatment assignment process, in the sense that the omitted variable cannot explain as much residual variance (or cannot explain a large multiple of the variance) of D or Y in comparison with an observed covariate X_j. More formally, define k_D and k_Y as

$$k_D := \frac{R^2_{D\sim Z|\mathbf{X}_{-j}}}{R^2_{D\sim X_j|\mathbf{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|\mathbf{X}_{-j},D}}{R^2_{Y\sim X_j|\mathbf{X}_{-j},D}}, \tag{21}$$

where $\mathbf{X}_{-j}$ represents the vector of covariates $\mathbf{X}$ excluding X_j, i.e. k_D indexes how much variance of the treatment assignment the confounder explains relative to how much X_j explains (after controlling for the remaining covariates). To make things concrete, for example, if the researcher believes that the omission of X_j would result in a larger mean-squared error of the treatment assignment regression than would the omission of Z, this equals the claim that $k_D \leqslant 1$. The same reasoning applies to k_Y.

Given parameters k_D and k_Y, we can rewrite the strength of the confounders as

$$R^2_{D\sim Z|\mathbf{X}} = k_D f^2_{D\sim X_j|\mathbf{X}-j}, \qquad R^2_{Y\sim Z|D,\mathbf{X}} \leqslant \eta^2 f^2_{Y\sim X_j|\mathbf{X}_{-j},D}, \tag{22}$$

where η is a scalar which depends on k_Y, k_D and $R^2_{D\sim X_j|\mathbf{X}-j}$ (see the on-line supplement section B for details). These equations enable us to investigate the maximum effect that a confounder at most 'k times' as strong as a particular covariate X_j would have on the coefficient estimate. These results are also tight, in the sense that we can always find a confounder that makes the second inequality an equality. Further, certain values for k_D and k_Y may be ruled out by the data (for instance, if $R^2_{D\sim X_j|\mathbf{X}_{-j}} = 50\%$ then k_D must be less than 1).

Our bounding exercises can be extended to any subset of the covariates. For instance, the researcher can bound the effect of a confounder as strong as *all* covariates $\mathbf{X}$ or any subset thereof. The method can also be extended to allow different subgroups of covariates to bound $R^2_{D\sim Z|\mathbf{X}}$ and $R^2_{Y\sim Z|D,\mathbf{X}}$—thus, if a group of covariates $\mathbf{X}_1$ is known to be the most important driver of selection to treatment, and another group of covariates $\mathbf{X}_2$ is known to be the most important determinant of the outcome, the researcher can exploit this fact.

4.5. Sensitivity to multiple confounders

The previous results let us assess the bias that is caused by a single confounder. Fortunately, they also provide *upper bounds* in the case of *multiple* unobserved confounders. (See Hosman *et al.* (2010), section 4.1, for an alternative proof.) Allowing $\mathbf{Z}$ to be a set (matrix) of confounders and $\hat{\gamma}$ its coefficient vector, the full equation that we wished we had estimated becomes

$$Y = \hat{\tau} D + \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\gamma} + \hat{\varepsilon}_{\text{full}}. \tag{23}$$

Now consider the single variable $Z^* = \mathbf{Z}\hat{\gamma}$. The bias that is caused by omitting $\mathbf{Z}$ is the same as omitting the linear combination Z^*, and we can think about the effect of multiple confounders in terms of this single confounder. Estimating the regression with $\mathbf{X}$ and Z^* instead of $\mathbf{X}$ and $\mathbf{Z}$ gives the same results for $\hat{\tau}$:

$$Y = \hat{\tau} D + \mathbf{X}\hat{\boldsymbol{\beta}} + Z^* + \hat{\varepsilon}_{\text{full}}. \tag{24}$$

Accordingly, Z^* has the same partial R^2 with the outcome as the full set $\mathbf{Z}$. However, the partial R^2 of Z^* with the treatment must be less than or equal to the partial R^2 of $\mathbf{Z}$ with the treatment—this follows simply because the choice of the linear combination $\hat{\gamma}$ is the choice

that maximizes the R^2 with the outcome, and not with the treatment. Hence, the bias that is caused by a multivariate $\mathbf{Z}$ must be less than or equal to the bias that is computed by using equation (13).

A similar reasoning can be applied to the standard errors. Since the effective partial R^2 of the linear combination Z^* with the treatment is less than that of $\mathbf{Z}$, simply modifying sensitivity equation (12) to account for the correct degrees of freedom ($\mathrm{df} - k$ instead of $\mathrm{df} - 1$) will give conservative adjusted standard errors for a multivariate confounder. From a practical point of view, however, we note that further correction of the degrees of freedom might be an unnecessary formality—we are performing a hypothetical exercise, and we can always imagine to have measured Z^*.

Finally, note that the set of confounders $\mathbf{Z}$ is arbitrary; thus it accommodates non-linear confounders as well as misspecification of the functional form of the observed covariates $\mathbf{X}$. To illustrate the point, let $Y = \hat{\tau} D + \hat{\beta} X + \hat{\gamma}_1 Z + \hat{\gamma}_2 Z^2 + \hat{\gamma}_3 ZX + \hat{\gamma}_4 X^2 + \hat{\varepsilon}_{\mathrm{full}}$, and imagine that the researcher did not measure Z and did not consider that X could also enter the equation with a squared term. Now just call $\mathbf{Z} = (Z_1 = Z, Z_2 = Z^2, Z_3 = ZX, Z_4 = X^2)$ and all the previous arguments follow.

5. Using the partial R^2 parameterization for sensitivity analysis

Returning to our running example of violence in Darfur, we illustrate how these tools can be deployed in an effort to answer the following questions.

(a) How strong would a particular confounder (or group of confounders) have to be to change our conclusions?
(b) In a worst-case scenario, how vulnerable is our result to *many* or *all* unobserved confounders acting together, possibly non-linearly?
(c) Are the confounders that would alter our conclusions plausible, or at least how strong would they have to be relative to observed covariates?

5.1. Proposed minimal reporting: robustness value, $R^2_{Y\sim D|\mathbf{X}}$ and bounds

Table 1 illustrates the type of reporting that we propose should accompany linear regression models that are used for causal inference with observational data. Along with traditionally reported statistics, we propose that researchers present

(a) the partial R^2 of the treatment with the outcome and
(b) the robustness value RV, both for where the point estimate and the confidence interval would cross zero, or another meaningful reference value (for convenience, we refer to RV_q or $\mathrm{RV}_{q,\alpha}$ with $q = 1$ as simply RV or RV_α).

Table 1. Proposed minimal reporting on sensitivity to unobserved confounders†

Treatment	*Outcome, PeaceIndex:*					
	Estimate	*Standard error*	*t-value*	$R^2_{Y\sim D\|\mathbf{X}}$ *(%)*	*RV (%)*	$RV_{\alpha=0.05}$ *(%)*
DirectHarm	0.097	0.023	4.18	2.2	13.9	7.6

†df = 783; bound (Z as strong as Female), $R^2_{Y\sim Z|D\mathbf{X}} = 12\%$, $R^2_{D\sim Z|\mathbf{X}} = 1\%$.

Finally, to aid user judgement, we encourage researchers to provide plausible bounds on the strength of the confounder. These may be based on bounds employing meaningful covariates determined by the research context and design (Section 4.4), or in principle may be available from theory and previous literature.

For our running example of violence in Darfur, Table 1 shows an augmented regression table, including the robustness value RV of DirectHarm coefficient: 13.9%. This means that unobserved confounders explaining at least 13.9% of the residual variance of both the treatment and the outcome would explain away the estimated treatment effect. It also means that any confounder explaining less than 13.9% of the residual variance of both the treatment and the outcome would not be sufficiently my strong to bring down the estimated effect to 0. For cases where one association is over 13.9% and the other is below, we conduct additional analyses that are illustrated in the next subsection. Nevertheless, RV still fully characterizes the robustness of the regression coefficient to unobserved confounding—it provides a quick, meaningful reference point for understanding the minimal strength of bias necessary to overturn the research conclusions (i.e. any confounder with an equivalent bias factor of $\mathrm{BF} = \mathrm{RV}/\sqrt{(1-\mathrm{RV})}$.)

Adjusting for confounding may not bring the estimate to 0, but rather into a range where it is no longer 'statistically significant'. Therefore, the robustness value accounting for statistical significance, $\mathrm{RV}_{\alpha=0.05}$, is also shown in Table 1. For a level of significance of 5%, the robustness value goes down from 13.9% to 7.6%—i.e. confounders would need to be only about half as strong to make the estimate not statistically significant. Finally, the partial R^2 of the treatment with the outcome, $R^2_{Y\sim D|\mathbf{X}}$, in Table 1 gives a sensitivity analysis for an extreme scenario: if confounders explained 100% of the residual variance of the outcome, they would need to explain at least 2.2% of the residual variance of the treatment to bring down the estimated effect to 0.

Confronted with those results, we now need to judge whether confounders with the strengths that are revealed to be problematic are plausible. If we can claim to have measured the most important covariates in explaining treatment and outcome variation, it is possible to bound the strength of the confounder with the tools of Section 4.4 and to judge where it falls relative to these quantities. The footnote to Table 1 shows the strength of association that a confounder as strong as Female would have: $R^2_{Y\sim Z|D,\mathbf{X}} = 12\%$ and $R^2_{D\sim Z|\mathbf{X}} = 1\%$. As the robustness value is higher than either quantity, Table 1 readily reveals that such a confounder could not fully eliminate the point estimate. In addition, since the bound for $R^2_{D\sim Z|\mathbf{X}}$ is less than $R^2_{Y\sim D|\mathbf{X}} = 2.2\%$, a 'worst-case confounder' explaining *all* of the left-out variance of the outcome and as strongly associated with the treatment as Female would not eliminate the estimated effect either.

Domain knowledge about how the treatment was assigned or regarding the main determinants of the outcome is required to make any such comparisons meaningful. In our running example, a reasonable argument can be made that gender is one of the most visually apparent characteristics of an individual during the attacks, and that, within village, gender was potentially the most important factor to explain targeting due to the high level of sexual violence. Thus, if we can argue that total confounding as strongly associated with the treatment as Female is implausible, those bounding results show that it cannot completely account for the observed estimated effect.

These sensitivity exercises are exact when considering a single linear unobserved confounder and are conservative for multiple unobserved confounders, possibly acting non-linearly—this includes the explanatory power of *all left-out factors*, even misspecification of the functional form of observed covariates. It is worth pointing out that sensitivity to any arbitrary confounder with a given pair of partial R^2 values ($R^2_{Y\sim Z|D,\mathbf{X}}$, $R^2_{D\sim Z|\mathbf{X}}$) can also be easily computed with the information in Table 1; see the example in on-line supplement section D.

5.2. Sensitivity contour plots with partial R^2: estimates and t-values

The next step is to refine the analysis with tools that visually demonstrate how confounders of different types would affect point estimates and *t*-values, while showing where bounds on such confounders would fall under different assumptions on how unobserved confounders compare with observables. Note that, although we focus on the plots for point estimates and *t*-values, *p*-values can be obtained from the *t*-values, and the confidence interval end points by adjusting the estimate with the appropriate multiple of the standard errors.

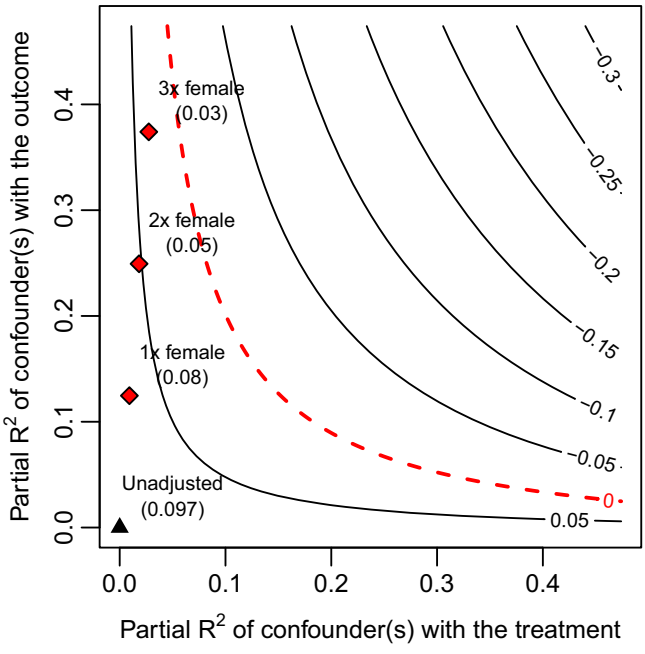
Perhaps the first plot that investigators would examine would be similar to Fig. 1, but now in the partial R^2 parameterization (Fig. 2(a)). The horizontal axis describes the fraction of the residual variation in the treatment (partial R^2) explained by the confounder; the vertical axis describes the fraction of the residual variation in the outcome explained by the confounder (as discussed in Section 4.2, axes could be transformed to show instead the total R^2 or the partial correlations among other options that may aid interpretation). The contours show the adjusted estimate that would be obtained for an unobserved confounder (in the full model) with the hypothesized values of the sensitivity parameters (assuming that the direction of the effects hurts our preferred hypothesis).

Whereas the contour plot that is used in illustrating the traditional OVB approach focused on a specific binary confounder—Center—the contour plot with the partial R^2 parameterization enables us to assess sensitivity to any confounder, irrespectively of its unit of measure. Additionally, since the sensitivity equations give an upper bound for the multivariate case, the same plot can be used to assess the sensitivity to any *group* of confounders, here including non-linear terms, such as the example of PoliticalAttitudes and Wealth acting together. If we choose a contour of interest (such as where the effect equals 0) and find the point with equal values on the horizontal and vertical axes (i.e. where it crosses a 45° line), this correspond to the robustness value, i.e. RV_q is a convenient, interpretable summary of a critical line of the contour plot.

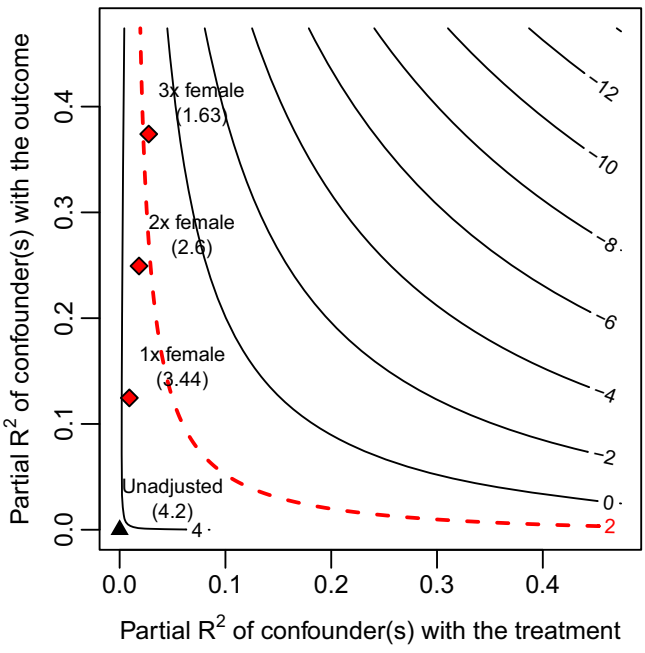
Further, the bounding exercise results in points on the plot showing the bounds on the partial R^2 of the unobserved confounder if it were k times as strong as the observed covariate Female. The first point shows the bounds for a confounder (or group of confounders) as strong as Female, as was also shown in Table 1. A second reference point shows the bounds for confounders *twice* as strong as Female, and finally the last point bounds the strength of confounders *three times* as strong as Female. The plot reveals that the *sign of the point estimate* is still relatively robust to confounding with such strengths, although the magnitude would be reduced to 77%, 55% and 32% of the original estimate.

Moving to inferential concerns, Fig. 2(b) now shows the sensitivity of the *t-value* of the treatment effect. As we move along the horizontal axis, not only the adjusted effect reduces, but we also obtain larger standard errors due to the variance inflation factor of the confounder. If we take the *t*-value of 2 as our reference (the usual approximate value for a 95% confidence interval), the plot reveals that the statistical significance of DirectHarm is robust to a confounder as strong as, or twice as strong, as Female. However, whereas confounders that are three times as strong as Female would not erode the point estimate to 0, we cannot guarantee that the estimate would remain statistically significant at the 5% level.

Altogether, these bounding exercises naturally lead to the questions: are such confounders plausible? Do we think it possible that confounders might exist that are three times as strong as Female? If so, what are they? Although we may not have complete confidence in answering such questions, we have moved the discussion from a qualitative argument about whether any confounding is possible to a more disciplined, quantitative argument that entices researchers to think about possible threats to their research design.



(a)



(b)

Fig. 2. Sensitivity contour plots in the partial R^2 scale with benchmark bounds: (a) sensitivity contour plot of the point estimate; (b) sensitivity contour plot of the t-value

5.3. Sensitivity plots of extreme scenarios

Even with a good understanding of the treatment assignment mechanism, investigators may not always be equipped to limit the association of the confounder convincingly with the outcome. In such cases, exploring sensitivity analysis to extreme scenarios is still an option. If we set $R^2_{Y\sim Z|D,\mathbf{X}}$ to 1 or some other conservative value, how strongly would such a confounder need to be associated with the treatment to change our estimate problematically? Although in some cases this exercise could reveal that confounders that are weakly related to the treatment would be sufficient to overturn the estimated effect, survival to extreme scenarios may help investigators to demonstrate the robustness of their results.

Applying this to our running example, results are shown in Fig. 3. The full curve represents the case where unobserved confounder(s) *explain all the left-out residual variance of the outcome*. On the vertical axis we have the adjusted treatment effect, starting from the case with no bias and going down as the bias increases, reducing the estimate; the horizontal axis shows the partial R^2 of the confounder with the treatment. In this *extreme scenario*, as we have seen, $R^2_{D\sim Z|\mathbf{X}}$ would need to be exactly the same as the partial R^2 of the treatment with the outcome to bring down the estimated effect to 0—i.e. it would need to be at least 2.2%: a value that is below the bound for a confounder once or twice as strong as Female (shown by the tick marks), which in this case is arguably one of the strongest predictors of the treatment assignment. In most circumstances, considering the worst-case scenario of $R^2_{Y\sim Z|D,\mathbf{X}} = 1$ might be needlessly conservative. Hence, we propose to plot other extreme scenarios, as shown in Fig. 3, where we consider different values of the partial R^2 of the unobserved confounder with the outcome, including 75% and 50%.

6. Discussion

6.1. Making formal sensitivity analysis standard practice

Given that ruling out unobserved confounders is often difficult or impossible in observational research, we might expect that sensitivity analyses would be a routine procedure in numerous

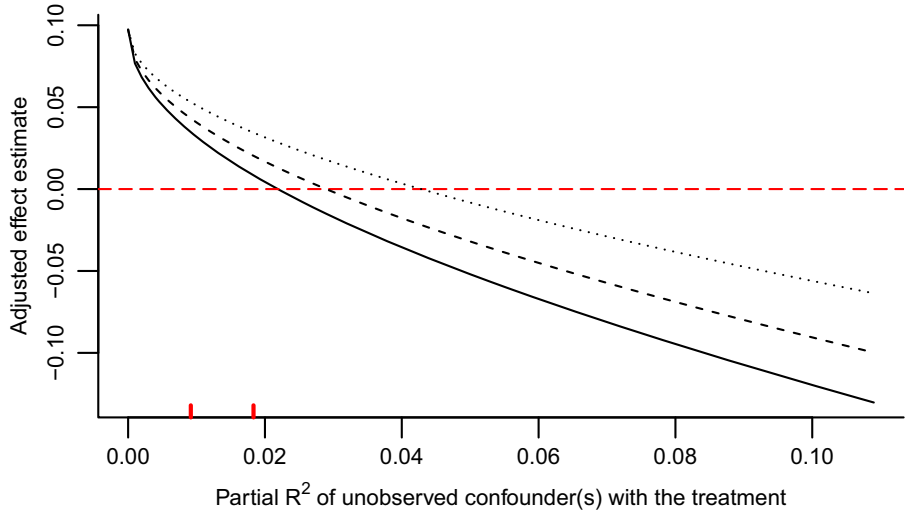


Fig. 3. Sensitivity analysis to extreme scenarios—partial R^2 of unobserved confounder(s) with the outcome: ———, 100%; – – –, 75%; · · · ·, 50%

disciplines. Why then are they not commonplace? We surmise that there are three main obstacles, which we directly address in this paper.

6.1.1. *Strong parametric assumptions*

First, the assumptions that many methods impose on the nature and distribution of unobserved confounders as well as on the treatment assignment mechanism may be difficult to sustain in some cases. For instance, Rosenbaum and Rubin (1983b), Imbens (2003), Carnegie *et al.* (2016a) and Dorie *et al.* (2016) required specifying the distribution of the confounder as well as modelling the treatment assignment mechanism; in another direction, the methods that were put forward in Robins (1999), Brumback *et al.* (2004) and Blackwell (2013) need to specify directly a confounding function parameterizing the difference in potential outcomes among treated and control units. Although assessing the sensitivity to some forms of confounding is an improvement over simply assuming no confounding (and users may be able to make suitable parameteric assumptions in some circumstances), widespread adoption of sensitivity analysis would benefit from methods that do not require users to make those restrictions *a priori*. Our derivations are rooted in the traditional OVB precisely to avoid those simplifying assumptions. As we have seen, the partial R^2 parameterization allows a flexible framework for assessing the sensitivity of the point estimate, as well as t-values and confidence intervals, allowing for multiple (possibly non-linear) confounders, even including misspecification of the functional form of the observed covariates.

6.1.2. *Lack of simple sensitivity measures for routine reporting*

A second obstacle to a wider adoption of sensitivity analysis is the lack of general, yet simple and interpretable sensitivity measures that users can report alongside other regression summary statistics. Our minimal reporting recommendation for regression tables (see Table 1) aims to fill this gap for regression models with

(a) the robustness value, which conveniently summarizes the minimal strength of association that a confounder needs to have to change the research conclusions, and
(b) the $R^2_{Y\sim D|\mathbf{X}}$, which works as an extreme scenario sensitivity analysis.

Regarding the robustness value in particular, we now discuss its relation to two other proposals that have been advocated in the literature: the *impact thresholds* of Frank (2000) and the E-value of VanderWeele and Ding (2017).

Frank (2000) proposed characterizing the strength of the unobserved confounder Z with what he denoted as its *impact*, defined as the product $R_{Y\sim Z|\mathbf{X}}R_{D\sim Z|\mathbf{X}}$ (not to confuse with $\hat{\gamma}$ of the impact times imbalance heuristic, as discussed in Section 3.2). This is then used to determine *impact thresholds*, which are defined as the minimum impact of the unobserved confounder that is necessary not to reject the null hypothesis of *zero effect*. However, as equation (14) reveals, the determinant of the bias is the bias factor $\text{BF} = R_{Y\sim Z|D,\mathbf{X}}f_{D\sim Z|\mathbf{X}}$, which does not have a one-to-one mapping with the confounder's impact. This can be made clear by rewriting the relative bias showing the product $R_{Y\sim Z|\mathbf{X}}R_{D\sim Z|\mathbf{X}}$ explicitly:

$$\text{relative bias} = \frac{|\overbrace{R_{Y\sim Z|\mathbf{X}}R_{D\sim Z|\mathbf{X}}}^{\text{Frank's impact}} - R_{Y\sim D|\mathbf{X}}R^2_{D\sim Z|\mathbf{X}}|}{|R_{Y\sim D|\mathbf{X}}(1 - R^2_{D\sim Z\mathbf{X}})|}. \tag{25}$$

Equation (25) reveals that

(a) an unobserved confounder with *zero impact* can still cause non-zero (downward) bias,
(b) an unobserved confounder with a *non-zero impact* can nevertheless induce zero bias (when impact $= R_{Y\sim D|\mathbf{X}} R^2_{D\sim Z|\mathbf{X}}$) and
(c) the two terms that compose the product $R_{Y\sim Z|\mathbf{X}} R_{D\sim Z|\mathbf{X}}$ do not enter symmetrically in the bias equation;

hence confounders with the *same impact* can cause *widely different biases*. This creates difficulties when trying to generalize the impact thresholds that were proposed in Frank (2000) to an arbitrary non-zero null hypothesis of regression coefficients. For instance, let q denote the relative bias and consider biases that move the effect towards (or through) zero. Solving equation (25) for *impact* gives us impact $= R_{Y\sim D|\mathbf{X}}\{q-(q-1)R^2_{D\sim Z|\mathbf{X}}\}$. Note that, given q and $R_{Y\sim D|\mathbf{X}}$, the impact that is necessary to bring about a relative bias of magnitude q still depends on the sensitivity parameter $R^2_{D\sim Z|\mathbf{X}}$—except when $q=1$ (for a numerical example, see the on-line supplement section A.5). Note that this is not a problem for the robustness value, since it acts as a convenient reference point uniquely characterizing any confounder with a bias factor of $\mathrm{BF} = \mathrm{RV}_q/\sqrt{(1-\mathrm{RV}_q)}$.

As to VanderWeele and Ding (2017), they have recently advanced the E-value: a sensitivity measure suited specifically for the *risk ratio*. For other effect measures, such as risk differences, the E-value is an approximation, whereas, if the researcher uses linear regression to obtain an estimate, the robustness value is exact. Also, whereas the robustness value parameterizes the association of the confounder with the treatment and the outcome in terms of percentage of variance explained (the partial R^2), the E-value parameterizes these in terms of risk ratios. Whether one scale is preferable over the other depends on context, and researchers should be aware of both options. Overall, we believe that the dissemination of measures such as the E-value and the robustness value is an important step towards the widespread adoption of sensitivity analysis to unobserved confounding. In current practice, robustness is often informally or implicitly linked to t-values or p-values, neither of which correctly characterizes how sensitive an estimate is to unobserved confounding. The extension of the robustness value to non-linear models is worth exploring in future research.

6.1.3. Difficulty in connecting sensitivity analysis to domain knowledge

Finally, the third and perhaps most fundamental obstacle to the use of sensitivity analysis is the difficulty in connecting the formal results to the researcher's substantive understanding about the object under study. This can be only partially overcome by statistical tools, as it relies on the nature of the domain knowledge that is used for plausibility judgements. In this paper we have showed how one can formally bound the strength of an unobserved confounder with the same strength (or a multiple thereof) as a chosen group of observed covariates, using three different types of comparison. This enables researchers to exploit knowledge regarding the relative importance of observed covariates: when researchers can credibly argue to have measured the most important determinants of the treatment assignment and of the outcome (in terms of variance explained), this bounding exercise can be a valuable tool. As we discuss next, previous attempts to make such comparisons have been problematic, either because of informal benchmarking practices that do not warrant the claims that they purport to make, or by relying on inappropriate choices of parameterization.

6.2. The risks of informal benchmarking

Although prior work has suggested informal benchmarking procedures using statistics of observed covariates $\mathbf{X}$ to help researchers to 'calibrate' their intuitions about the strength of the

unobserved confounder Z (Frank, 2000; Imbens, 2003; Frank and Min, 2007; Hosman *et al.*, 2010; Dorie *et al.*, 2016; Carnegie *et al.*, 2016a, b; Middleton *et al.*, 2016; Hong *et al.*, 2018), this practice has undesirable properties and can lead users to erroneous conclusions, even in the ideal case where they do have the correct knowledge about how Z compares with $\mathbf{X}$. This happens because the estimates of how the observed covariates are related to the outcome may be themselves affected by the omission of Z, regardless of whether we assume that Z is independent of $\mathbf{X}$. To illustrate this threat concretely, we first consider a simple simulation where there is no effect of D on Y, Z is orthogonal to X and, more importantly, Z *is exactly like* X. (We use structural equations, $Y = X + Z + \varepsilon_y$, $D = X + Z + \varepsilon_d$, $X = \varepsilon_x$ and $Z = \varepsilon_z$ where all disturbances are independent standard normal random variables. See also the on-line supplement section C.) The results are shown in Fig. 4.

Note that the informal benchmark point is still far from zero, leading the investigator to conclude incorrectly that a confounder 'not unlike X' would not be sufficient to bring down the estimated effect to 0—when in fact it would. This incorrect conclusion occurs although the investigator *correctly assumes* both that the unobserved confounder is 'no worse' than X (in terms of its strength of relationship to the treatment and outcome) and that $Z \perp X$. Fig. 4 also shows the formal bounds that are obtained with the procedures given in Section 4.4. Note that these would lead the researcher to the correct conclusion: an unobserved confounder with the same strength as X would be sufficiently powerful to bring down the estimated effect to 0.

Why exactly does this happen? Consider for a moment the difference between the coefficient on $\mathbf{X}$ in the full equation (3), $\hat{\beta}$, and its estimate in the restricted equation (4), $\hat{\beta}_{\text{res}}$. Using the same OVB approach of impact times imbalance, we arrive at $\hat{\beta}_{\text{res}} - \hat{\beta} = \hat{\gamma}\hat{\psi}$, where $\hat{\psi}$ is obtained

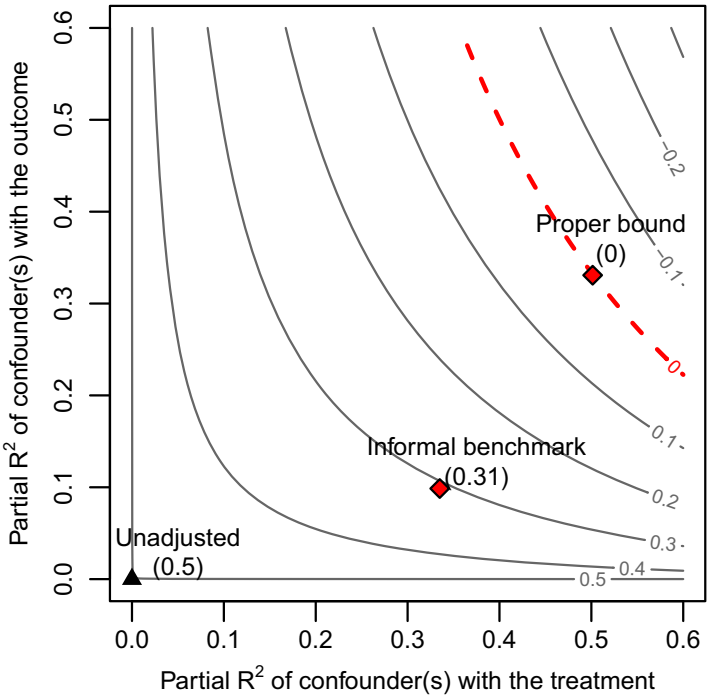


Fig. 4. Sensitivity contours of point estimate—informal benchmarking *versus* proper bounds

from the regression $Z = \hat{\delta} D + \mathbf{X}\hat{\psi} + \hat{\varepsilon}_Z$. Note that $\hat{\psi}$ can be non-zero even if $\mathbf{X} \perp Z$, because D is a collider (Pearl, 2009), and conditioning on D creates dependence between Z and $\mathbf{X}$. The reasoning holds whether we are using the regression coefficients themselves or other observed statistics, such as partial correlations, partial R^2 values or t-values. This renders claims of the type 'a confounder Z not unlike X could not change the research conclusions' unreliable when observed statistics without proper adjustment are used for benchmarking.

We can use the formal bounds that are derived in equation (22) to quantify how misleading claims using informal benchmarks would be. In the partial R^2 parameterization, this amounts to using as benchmarks $k_D R^2_{D\sim X_j|\mathbf{X}_{-j}}$ and $k_Y R^2_{Y\sim X_j|\mathbf{X}_{-j},D}$, instead of the proper bounds $k_D f^2_{D\sim X_j|\mathbf{X}_{-j}}$ and $\eta^2 f^2_{Y\sim X_j|\mathbf{X}_{-j},D}$. There are, thus, two discrepancies:

(a) an adjustment of baseline variance to be explained, when converting the partial R^2 to partial Cohen's $f^2 = R^2/(1-R^2)$, which affects both co-ordinates of the benchmark, and
(b) the collider bias due to the association of X_j with D, which affects only the bound on $R^2_{Y\sim Z|D,\mathbf{X}}$ via $\eta^2 \geqslant k_Y$.

The adjustment of baseline variance may affect informal benchmarks based on correlational (Frank, 2000), partial R^2 (Imbens, 2003) and t-value (Hosman *et al.*, 2010) measures. The collider bias may affect informal benchmarks that condition on D. Benchmarks that do not condition on D (such as in Frank (2000)) are not affected by collider bias.) Therefore, the stronger the association of X_j with the treatment, and the larger the multiples that are used for comparisons (k times as strong), the more misleading informal benchmarks will be. We thus advise against informal benchmarking procedures, and previous studies relying on these methods may warrant revisiting, especially those where benchmark points have strong association with the treatment assignment.

6.3. On the choice of parameterization

The approach of Hosman *et al.* (2010) is also rooted in the OVB framework, but it suffers from two main deficiencies. The first is the central role that informal benchmarking plays in their proposal, which can be seriously misleading as discussed in the previous section. The second issue is more subtle, but equally important: the choice of parameterization. Hosman *et al.* (2010) asked researchers to 'calibrate intuitions' about the strength of the confounder with the treatment by using a t-value. This is a problematic choice because the t-value incorporates information on both the strength of association and the sample size, the latter being irrelevant for identification concerns. What constitutes a large t-value for statistical significance does not map directly to what constitutes a large strength of a confounder, as this mapping varies significantly depending on sample size. (More precisely, the t-value in the expression of the bias is an artefact of both multiplying and dividing by the degrees of freedom, as in our equation (12). Although t-values can be useful for computational purposes (to utilize quantities that are routinely reported in regression tables), their dependence on sample size makes them inappropriate for contemplating how strongly related a confounder is to the treatment. Consider a t-value of 200. With 100 degrees of freedom, the confounder explains virtually all the residual variance of the treatment (a partial R^2 of 0.9975), whereas with 10 million degrees of freedom the confounder explains less than 0.5%. These are clearly confounders with very different strengths, and the partial R^2 clarifies this distinction.)

An alternative bounding argument has also been presented in Oster (2019) which, unlike the informal benchmarking practices that were previously discussed, provides a formal identification result. Nevertheless, the procedure proposed asks users to reason about a quantity that is very difficult to understand. More precisely, Oster (2019) asked researchers to make plausibility

judgements on two sensitivity parameters: $R_{\max}$ and δ_{Oster}. The $R_{\max}$-parameter is simply the maximum explanatory power that one could have with the full outcome regression, i.e. $R_{\max} = R^2_{Y\sim D+\mathbf{X}+\mathbf{Z}}$. As discussed in Section 4.2 (equation (17)) this has a one-to-one relationship with $R^2_{Y\sim \mathbf{Z}|\mathbf{X},D}$:

$$R^2_{Y\sim \mathbf{Z}|\mathbf{X},D} = \frac{R_{\max} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}}. \tag{26}$$

By contrast the second sensitivity parameter, δ_{Oster}, is not easily interpretable in substantive terms. Following Altonji *et al.* (2005), Oster (2019) defined 'indices' $W_1 := \mathbf{X}\hat{\boldsymbol{\beta}}$ and $W_2 := \mathbf{Z}\hat{\boldsymbol{\gamma}}$, where $\mathbf{X}$ is a matrix of observed covariates and $\mathbf{Z}$ a matrix of unobserved covariates. Critically, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are chosen such that $Y = \hat{\tau}D + W_1 + W_2 + \hat{\varepsilon}_{\text{full}}$ (Oster (2019) used population values. Here we use sample values to maintain consistency with the rest of the paper, but this has no consequence for the argument in question). The δ_{Oster}-parameter equals $\text{cov}(W_2, D)/\text{var}(W2) \times \text{var}(W_1)/\text{cov}(W_1, D)$ and is intended as a measure of 'proportional selection', i.e. how strongly the unobservables drive treatment assignment, relative to the observables. The problem here is that constructing indices W_1 and W_2 based on relationships to the outcome is not innocuous: δ_{Oster} captures not only the relative influence of $\mathbf{X}$ and $\mathbf{Z}$ over the treatment, but also their association with the outcome. To examine the simple case with only one covariate and one confounder and assuming $X \perp Z$, we have

$$\delta_{\text{Oster}} = \frac{\text{cov}(W_2, D)}{\text{var}(W_2)} \frac{\text{var}(W_1)}{\text{cov}(W_1, D)} = \frac{\text{cov}(\hat{\gamma}Z, D)}{\text{var}(\hat{\gamma}Z)} \frac{\text{var}(\hat{\beta}X)}{\text{cov}(\hat{\beta}X, D)} = \frac{\text{cov}(Z, D)}{\hat{\gamma}\text{var}(Z)} \frac{\hat{\beta}\text{var}(X)}{\text{cov}(X, D)} = \frac{\hat{\lambda}}{\hat{\gamma}} \frac{\hat{\beta}}{\hat{\theta}}, \tag{27}$$

where $\hat{\lambda}$ and $\hat{\theta}$ are the coefficients of the regression $D = \hat{\theta}X + \hat{\lambda}Z + \hat{\varepsilon}_D$. Consequently, claims that $\delta_{\text{Oster}} = 1$ implies that 'the unobservable and observables are equally related to the treatment' (Oster (2019), page 192) can lead researchers astray, as this quantity also depends on associations with the outcome. To see how, let the variables be standardized to mean 0 and unit variance, and pick $\hat{\beta} = \hat{\theta} = p$, $\hat{\gamma} = \hat{\lambda} = p/2$ and $\hat{\tau} = 0$. In this case, the confounder Z has either half or a quarter of the explanatory power of X (as measured by standardized coefficients or variance explained), yet $\delta_{\text{Oster}} = 1$.

Although researchers may be able to make arguments about relative explanatory power of observables and unobservables in the treatment assignment process, the δ_{Oster}-parameter does not correspond directly to such claims. Indeed, arguments made by researchers applying Oster (2019) suggest that they believe they are comparing the explanatory power of observables and unobservables over treatment assignment in terms such as correlation or variance explained (e.g. as in Jakiela and Ozier (2018), page 4, 'Following the approach suggested by Altonji, Elder, and Taber (2005) and Oster (2017), we estimate that unobservable country-level characteristics would need to be 1.44 times more correlated with treatment than observed covariates to fully explain the apparent impact of grammatical gender on the level of female labor force participation; unobserved factors would need to be 3.23 times more closely linked to treatment to explain the impact of grammatical gender on the gender gap in labor force participation'). By contrast, the parameter k_D that we introduced in our bounding procedure (Section 4.4) captures precisely this notion of the relative explanatory power of the unobservable and observable over treatment assignment, in terms of partial R^2 or total R^2, depending on the investigator's preference.

Such parameterization choices are more than notional when they drive a wedge between what investigators can argue about and the values of the parameters that these arguments imply. It is thus important that the sensitivity parameters that are used in these exercises be as transparent

as possible and match investigators' conception of what they mean. Hence, we employ R^2-based parameters, rather than t-values or quantities relating indices. The resulting sensitivity parameters not only correspond more directly to what investigators can articulate and reason about, but also lead to the rich set of sensitivity exercises that we have discussed. Of course, further improvements may be possible and future research should investigate whether such flexibility can be achieved with yet more meaningful parameterizations.

The tools that we propose here, like any other, have potential for abuse. We thus end with important *caveats*, in particular emphasizing that sensitivity analysis should not be used for automatic judgement, but as an instrument for disciplined arguments about confounding.

6.4. Sensitivity analysis as principled argument

Sensitivity analyses tell us what we would have to be willing to believe to accept the substantive claims that were initially made (Rosenbaum, 2005, 2010, 2017). The sensitivity exercises that were proposed here tell the researcher how strong unobserved confounding would have to be to change meaningfully the treatment effect estimate beyond some level we are interested in, and employ observed covariates to argue for bounds on unobserved confounding where possible. Whether we can rule out the confounders that are shown to be problematic depends on expert judgement. As a consequence, the research design and identification strategy as well as the story explaining the quality of the covariates that are used for benchmarking all play vital roles.

For this reason, we do not propose any arbitrary thresholds for deeming sensitivity statistics, such as the robustness value or the partial R^2 of the treatment with the outcome, sufficiently large to escape confounding concerns. In our view, no meaningful universal thresholds of the sort are possible to establish. In a poorly controlled regression on observational data, with no clear understanding of what (unobservables) might influence treatment uptake, it would be difficult to claim credibly that a robustness value of 15% is 'good news', since the investigator does not have the necessary domain knowledge to rule out the strength of unobserved confounders down to this level. In contrast, in a quasi-experiment where the researcher knows that the treatment was assigned in such a way that observed covariates account for almost any possible selection, a more credible case may be made that the types of confounders that would substantially alter the research conclusions are unlikely.

Similarly, we strongly warn against blindly employing covariates for bounding the strength of confounders, without the ability to argue that they are likely to be among the strongest predictors of the outcome or treatment assignment. A particular moral hazard is that weak covariates can make the apparent bounds look better. It is thus imperative for readers and reviewers to demand that researchers properly justify and interpret their sensitivity results, after which such claims can be properly debated. Sensitivity analysis is best suited as a tool for disciplined quantitative arguments about confounding, not for obviating scientific discussions by following automatic procedures.

This transition from a qualitative to a quantitative discussion about unobserved confounding can often be enlightening. As put by Rosenbaum (2017), page 171, it may 'provide grounds for caution that are not rooted in timidity, or grounds for boldness that are not rooted in arrogance'. A sensitivity analysis raises the bar for the sceptic of a causal estimate—not just any criticism can invalidate the research conclusions. The hypothesized unobserved confounder now must meet certain standards of strength; otherwise, it cannot logically account for all the observed association. Likewise, it also raises the bar for defending a causal interpretation of an estimate—proponents must articulate how confounders with certain strengths can be ruled out.

A final point of concern is the potential misuse of sensitivity analysis in the gate-keeping of publications. Sensitivity analysis should not be misappropriated as a tool for inhibiting 'imperfectly identified' research on relevant topics. Studies on important questions using state of the art research design, which turn out not to be robust to reasonable sources of confounding, should not be dismissed. On the contrary, with sensitivity analyses, we can conduct imperfect investigations, while transparently revealing how susceptible our results are to unobserved confounders. This gives future researchers a starting point and road map for improving on the robustness of these answers in their following inquiries.

Acknowledgements

We thank Neal Beck, Graeme Blair, Darin Christensen, Christopher M. Felton, Kenneth Frank, Adam Glynn, Erin Hartman, Paul Hünermund, Kosuke Imai, Ed Leamer, Ian Lundberg, Julian Schuessler, Brandon Stewart, Michael Tzen, Teppei Yamamoto and members of the 'Improving design in social science' workshop at the University of California, Los Angeles, for valuable comments and feedback. Thanks go to Aaron Rudkin for assistance developing the R package `sensemakr`. Thanks go to Fernando Mello for his examination of how many papers in political science journals have employed formal sensitivity analyses. We thank the reviewers and the Joint Editor and Associate Editor for valuable suggestions.

References

Altonji, J. G., Elder, T. E. and Taber, C. R. (2005) An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *J. Hum. Resour.*, **40**, 791–821.

Angrist, J. D. and Pischke, J.-S. (2008) *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton: Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2017) Undergraduate econometrics instruction: through our classes, darkly. *Technical Report*. National Bureau of Economic Research, Cambridge.

Blackwell, M. (2013) A selection bias approach to sensitivity analysis for causal effects. *Polit. Anal.*, **22**, 169–182.

Brumback, B. A., Hernán, M. A., Haneuse, S. J. and Robins, J. M. (2004) Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statist. Med.*, **23**, 749–767.

Carnegie, N. B., Harada, M. and Hill, J. L. (2016a) Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J. Res. Educ. Effect.*, **9**, 395–420.

Carnegie, N., Harada, M. and Hill, J. (2016b) treatsens: a package to assess sensitivity of causal analyses to unmeasured confounding.

Cinelli, C. and Hazlett, C. (2019) sensemakr: sensitivity analysis tools for OLS. *R Package Version 0.1.2*.

Cinelli, C., Kumor, D., Chen, B., Pearl, J. and Bareinboim, E. (2019) Sensitivity analysis of linear structural causal models. *Proc. Mach. Learn. Res.*, **97**, 1252–1261.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B. and Wynder, E. L. (1959) Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Natn. Cancer Inst.*, **22**, 173–203.

Ding, P. and Miratrix, L. W. (2015) To adjust or not to adjust?: Sensitivity analysis of M-bias and butterfly-bias. *J. Causl Inf.*, **3**, 41–57.

Dorie, V., Harada, M., Carnegie, N. B. and Hill, J. (2016) A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statist. Med.*, **35**, 3453–3470.

Dunning, T. (2012) *Natural Experiments in the Social Sciences: a Design-based Approach*. New York: Cambridge University Press.

Flint, J. and de Waal, A. (2008) *Darfur: a New History of a Long War*. London: Zed Books.

Frank, K. A. (2000) Impact of a confounding variable on a regression coefficient. *Sociol. Meth. Res.*, **29**, 147–194.

Frank, K. A., Maroulis, S. J., Duong, M. Q. and Kelcey, B. M. (2013) What would it take to change an inference?: Using Rubin's causal model to interpret the robustness of causal inferences. *Educ. Evaln Poly Anal.*, **35**, 437–460.

Frank, K. and Min, K.-S. (2007) Indices of robustness for sample representation. *Sociol. Methodol.*, **37**, 349–392.

Frank, K. A., Sykes, G., Anagnostopoulos, D., Cannata, M., Chard, L., Krause, A. and McCrory, R. (2008) Does NBPTS certification affect the number of colleagues a teacher helps with instructional matters? *Educ. Evaln Poly Anal.*, **30**, 3–30.

Franks, A., D'Amour, A. and Feller, A. (2019) Flexible sensitivity analysis for observational studies without observable implications. *J. Am. Statist. Ass.*, to be published.

Frisch, R. and Waugh, F. V. (1933) Partial time regressions as compared with individual trends. *Econometrica*, **1**, 387–401.
Hazlett, C. (2019) Angry or weary?: The effect of personal violence on attitudes towards peace in Darfur. *J. Conflct Resoln*, to be published.
Hong, G., Qin, X. and Yang, F. (2018) Weighting-based sensitivity analysis in causal mediation studies. *J. Educ. Behav. Statist.*, **43**, 32–56.
Hosman, C. A., Hansen, B. B. and Holland, P. W. (2010) The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Statist.*, **4**, 849–870.
Imai, K., Keele, L. and Yamamoto, T. (2010) Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci.*, **25**, 51–71.
Imbens, G. W. (2003) Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.*, **93**, 126–132.
Imbens, G. W. and Rubin, D. B. (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press.
Jakiela, P. and Ozier, O. (2018) Gendered language. *Policy Research Working Paper*. World Bank, Washington DC.
Kruskal, W. and Majors, R. (1989) Concepts of relative importance in recent scientific literature. *Am. Statistn*, **43**, 2–6.
Leamer, E. (1983) Let's take the con out of econometrics. *Am. Econ. Rev.*, **73**, 31–43.
Leamer, E. E. (2016) S-values: conventional context-minimal measures of the sturdiness of regression coefficients. *J. Econmetr.*, **193**, 147–161.
Lovell, M. C. (1963) Seasonal adjustment of economic time series and multiple regression analysis. *J. Am. Statist. Ass.*, **58**, 993–1010.
Lovell, M. C. (2008) A simple proof of the FWL theorem. *J. Econ. Educ.*, **39**, 88–91.
Middleton, J. A., Scott, M. A., Diakow, R. and Hill, J. L. (2016) Bias amplification and bias unmasking. *Polit. Anal.*, **24**, 307–323.
Oster, E. (2014) Unobservable selection and coefficient stability: theory and evidence. *Working Paper*. National Bureau of Economic Research, Cambridge.
Oster, E. (2019) Unobservable selection and coefficient stability: theory and evidence. *J. Bus. Econ. Statist.*, **37**, 187–204.
Pearl, J. (2009) *Causality*. New York: Cambridge University Press.
Pearl, J. (2011) Invited commentary: understanding bias amplification. *Am. J. Epidem.*, **174**, 1223–1227.
Robins, J. M. (1999) Association, causation, and marginal structural models. *Synthese*, **121**, 151–179.
Rosenbaum, P. R. (1984) The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Statist. Soc.* A, **147**, 656–666.
Rosenbaum, P. R. (2002) *Observational Studies*, pp. 1–17. New York: Springer.
Rosenbaum, P. R. (2005) Sensitivity analysis in observational studies. In *Encyclopedia of Statistics in Behavioral Science*, vol. 4, pp. 1809, 1814. New York: Wiley.
Rosenbaum, P. R. (2010) *Design of Observational Studies*. New York: Springer.
Rosenbaum, P. R. (2017) *Observation and Experiment: an Introduction to Causal Inference*. Cambridge: Harvard University Press.
Rosenbaum, P. R. and Rubin, D. B. (1983a) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
Rosenbaum, P. R. and Rubin, D. B. (1983b) Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B, **45**, 212–218.
Steiner, P. M. and Kim, Y. (2016) The mechanics of omitted variable bias: bias amplification and cancellation of offsetting biases. *J. Causl Inf.*, **4**, no. 2.
Vanderweele, T. J. and Arah, O. A. (2011) Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology*, **22**, 42–52.
VanderWeele, T. J. and Ding, P. (2017) Sensitivity analysis in observational research: introducing the E-value. *Ann. Intern. Med.*, **167**, 268–274.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Online supplementary material for "Making sense of sensitivity: extending omitted variable bias"'.