Making Sense of Sensitivity: Extending Omitted Variable Bias

Carlos Cinelli [*] Chad Hazlett[†]

Forthcoming in the Journal of the Royal Statistical Society, Series B (Statistical Methodology), 2020.
`https://doi.org/10.1111/rssb.12348`

Abstract

We extend the omitted variable bias framework with a suite of tools for sensitivity analysis in regression models that: (i) does not require assumptions on the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounders; (ii) naturally handles multiple confounders, possibly acting non-linearly; (iii) exploits expert knowledge to bound sensitivity parameters; and, (iv) can be easily computed 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 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, t-values, as well as "extreme scenarios". Finally, we describe problems with a common "benchmarking" practice and introduce a novel procedure to formally bound the strength of confounders based on comparison to observed covariates. We apply these methods to a running example that estimates the effect of exposure to violence on attitudes toward peace.

[*] PhD Candidate, Department of Statistics, University of California Los Angeles.
Email: carloscinelli@ucla.edu.

[†] Assistant Professor, Departments of Statistics and Political Science, University of California Los Angeles. Email: chazlett@ucla.edu URL: http://www.chadhazlett.com. 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 UCLA for valuable comments and feedback. Thanks to Aaron Rudkin for assistance developing the `R` package `sensemakr`. Thanks to Fernando Mello for his examination of how many papers in Political Science journals have employed formal sensitivity analyses. We thank anonymous reviewers and the editors for valuable suggestions.

1 Introduction

Observational research often seeks to estimate causal effects under a "no unobserved confounding" or "ignorability" (conditional on observables) assumption (see e.g. Rosenbaum and Rubin 1983b; Pearl 2009; 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 e.g. Angrist and Pischke 2008; Dunning 2012). Investigators seeking to make causal claims from observational data are also instructed to show "balance tests" and "placebo tests". While, 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 inference 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 analyses" examining how fragile a result is against the possibility of unobserved confounding.[1] In general, such analyses entail two components: (1) 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 (2) assisting the investigator in assessing the plausibility that such problematic confounding might exist, which necessarily depends upon 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 (1983a); Robins (1999); Frank (2000); Rosenbaum (2002); Imbens (2003); Brumback et al. (2004); Frank et al. (2008); Hosman et al. (2010); Imai et al. (2010); Vanderweele and Arah (2011); Blackwell (2013); Frank et al. (2013); Carnegie et al. (2016b); Dorie et al. (2016); Middleton et al. (2016); Oster (2017), and Franks et al. (2019).

Yet, such sensitivity analyses remain underutilized.[2] We argue that a number of factors contribute to this reluctant uptake. One is the complicated nature and strong assumptions 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, while 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", "exogeneous", or "as-if random" nature of the treatment assignment. To complicate things, some of the solutions 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

[1] Researchers may also wish to examine sensitivity to the choice of observed covariates, see Leamer (1983 2016).

[2] In political science, out of 164 quantitative papers in the top three general interest publications (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 4 (6.25%) employed a formal sensitivity analyses beyond trying various specifications. In economics, Oster (2014) reports that most of non-experimental empirical papers utilized only informal robustness tests based on coefficient stability in the face of adding or dropping covariates. See also Chen and Pearl (2015).

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: (i) 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, (ii) the proportion of variation in the outcome explained uniquely by the treatment, $R^2_{Y\sim D|\boldsymbol{X}}$, which reveals how strongly counfounders that explain 100% of the residual variance of the outcome would have to be associated with the treatment in order 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 the RV and $R^2_{Y\sim D|\boldsymbol{X}}$ 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' 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 *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. While prior work (Frank 2000; Imbens 2003; Hosman et al. 2010; Blackwell 2013; Dorie et al. 2016; Carnegie et al. 2016b; Middleton et al. 2016; Hong et al. 2018) has suggested an informal practice of benchmarking the unobserved confounding by comparison to 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) 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, how it compares to existing procedures, and highlighting important caveats when interpreting sensitivity results. Open-source software for

`R` implements the methods presented here.[3]

2 Running example

In this section we briefly introduce the applied example used throughout the paper.[4] This serves as a background to illustrate how the tools developed here can be applied to address problems that commonly arise in observational research. We emphasize that the information 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 shown to be problematic. Thus, a real world 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 200,000 (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*. While 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[5]

One can further argue that attacks were indiscriminate within village on the basis that the violence promoted by the government was mainly used to drive people out rather than target individuals. Within village, the bombing was crude and the attackers had almost no information about who they would target, with one major exception: while 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,

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

where *PeaceIndex* is an index measuring individual attitudes towards peace, *DirectHarm* a dummy variable indicating whether an individual was reportedly injured or maimed during such an attack,

[3] `R` package `sensemakr` (Cinelli and Hazlett 2019) available on CRAN: `https://cran.r-project.org/package=sensemakr`. Shiny App available on: `https://carloscinelli.shinyapps.io/robustness_value/`.

[4] We only describe the most relevant details, further information is available in Hazlett (2019).

[5] Transcript from interview taken by Darfurian Voices team. Interview code 03072009_118_cf2009008.

Female is a fixed effect for being female, and *Village* is a *matrix* of village fixed effects. Other pre-treatment covariates are included through the matrix $\boldsymbol{X}$, 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 center of the village were more often harmed than those on the periphery. And might not those nearer the center of each village also have different types of attitudes towards peace, on average? This suggests that the author ought to have instead run the model,

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

That is, 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 one can be handled in a relatively straightforward manner by the traditional OVB framework, as we will see in Section 3.

However, other skeptical researchers may question the claim that violence was conditionally indiscriminate with more elaborate stories, worrying that unobserved factors such as *Wealth* or *Political Attitudes* 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 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 "omitted variable bias" (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 upon hypothetical strengths of the confounder. This serves not only as an introduction to the method, but also to highlight limitations we will address in the following sections.

3.1 The traditional Omitted Variable Bias

Suppose 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 $\boldsymbol{X}$ and Z, as in

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

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); $\boldsymbol{X}$ is an $(n \times p)$ matrix of *observed* (pre-treatment) 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,

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

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

How does the observed estimate $\hat{\tau}_{\text{res}}$ compare to the desired estimate, $\hat{\tau}$? Let us define as $\widehat{\text{bias}}$ the difference between these estimates, $\widehat{\text{bias}} \; := \; \hat{\tau}_{\text{res}} \; - \; \hat{\tau}$, where the hat, $\widehat{(\cdot)}$, 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 (FWL) theorem (Frisch and Waugh 1933; Lovell 1963 2008) to "partial out" the observed covariates $\boldsymbol{X}$, the classic omitted variable bias solution is

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

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

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

While elementary, the OVB formula in Equation 6 provides the key intuitions as well as a formulaic basis for a simple sensitivity analysis, letting us assess how the omission of covariates 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, one is typically interested in cases where the investigator has determined that the full regression, controlling for *both* $\boldsymbol{X}$ *and* the unobserved variable Z, would have identified the causal effect of D on Y; thus, hereafter we will 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.[6]

[6] 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 $\{\boldsymbol{X}, Z\}$ 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 $\{\boldsymbol{X}, Z\}$, that is $Y_d \perp\!\!\!\perp D | \boldsymbol{X}, Z$, where Y_d denotes the potential outcome of Y when D is *set* to d (see Pearl 2009; Angrist and Pischke 2008; Imbens and Rubin 2015). We further note 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; Steiner and Kim 2016 for related discussions). Here we assume the researcher is interested in the estimates one would obtain from running the regression in Equation 3, controlling for $\boldsymbol{X}$ and Z.

3.2 Making sense of the traditional OVB

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 that 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.[7]

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, $Z = \hat{\delta} D + \boldsymbol{X}\hat{\boldsymbol{\psi}} + \hat{\varepsilon}_Z$, and not the regression of the treatment D on Z, and $\boldsymbol{X}$. That is, $\hat{\delta}$ 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 conditional on certain covariates $\boldsymbol{X}$. In that case we may then check for "balance" on other (pre-treatment) 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 $\mathbb{E}[Z|D, \boldsymbol{X}]$ (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 $\boldsymbol{X}$ 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 OVB 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 is common practice in econometrics work.[8] 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?

[7] While 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 former on the latter depends on structural assumptions. Suppose, for example, the "true" outcome model is assumed to be a linear structural equation where strict exogeneity holds, i.e., $Y = \tau D + \boldsymbol{X}\boldsymbol{\beta} + \gamma Z + \varepsilon$ and $\mathbb{E}[\varepsilon|D, \boldsymbol{X}, Z] = 0$. Then, $\hat{\gamma}$ could be interpreted as an estimate of the direct causal impact of a unit change of the confounder on the expected value of the outcome Y, holding the other covariates fixed. 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 to, along with $\boldsymbol{X}$, block the back-door paths from D to Y (Pearl 2009), or equivalently, 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 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.

[8] e.g. "Using a similar omitted-variables-type argument, we note that even if there are other confounders that we haven't controlled for, those that are positively correlated with private school attendance are likely to be positively correlated with earnings as well. Even if these variables remain omitted, their omission leads the estimates computed with the variables at hand to overestimate the private school premium." (Angrist and Pischke 2017, p.8-9)

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 OLS regression 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 we wish we had controlled for is a binary indicator of whether the respondent lived in the center or in the periphery of the village. How strong would this specific confounder have to be in order for its inclusion to substantially affect our conclusions? Figure 1 shows the plot of adjusted estimates for several hypothetical values of impact and imbalance of the confounder *Center*.

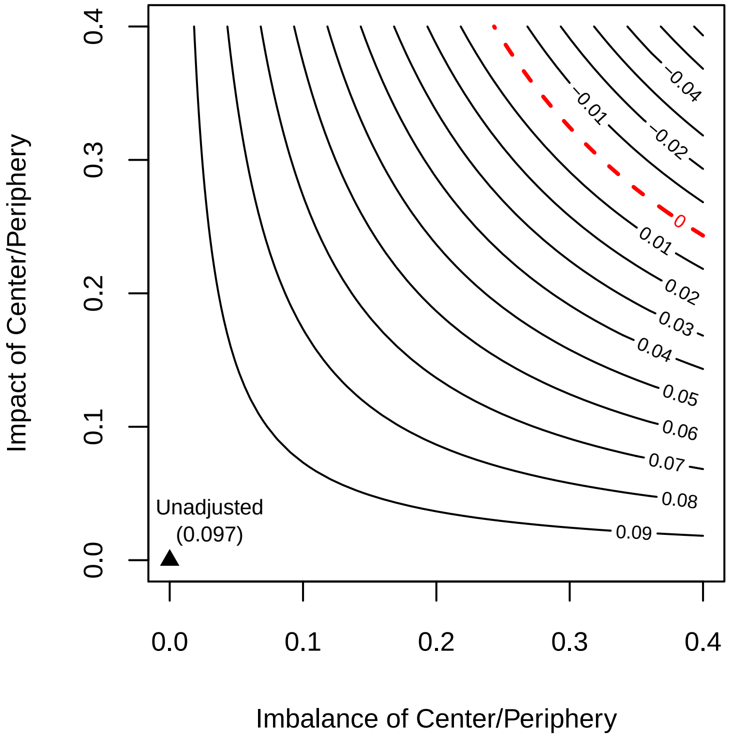


Figure 1: Sensitivity contours of the point estimate—traditional OVB

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 center of the village. Values for the hypothetical impact of the confounder on the outcome lie on the vertical axis, representing how attitudes towards peace differ on average for people living in the center 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 one 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. Notice 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.[9]

This plot explicitly reveals the type of prior knowledge one needs to have in order to be able to rule out problematic confounders. As an example, imagine the confounder *Center* has a conditional imbalance as high as 0.25—that is, having controlled for the observed covariates, those who were physically injured were also 25 percentage points more likely to live in the center of the village than those who were not. With such an imbalance, the plot reveals that the impact of living in the center on the outcome (Peace Index) would have to be over 0.40 in order to bring down the estimated effect of *DirectHarm* to zero.

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 centers are generally not markedly different in composition than the peripheries, it is hard to believe that being in the center was associated with a 0.40 higher expected score on Peace Index (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.

Shortcomings of the traditional OVB

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 the case. Imagine contemplating confounders such as *Political Attitudes*: 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 standardizing 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 *Political Attitudes* 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 OVB 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 the remaining covariates $\boldsymbol{X}$ have been accounted

[9] Investigators may also argue that accounting for omitted variable bias would increase the effect size, in the sense that the current estimate is conservative. Our tools apply to these cases as well, the arguments would just work in the opposite direction. For simplicity of exposition, in the paper we focus on the case where accounting for omitted variable bias reduces the effect size.

for. The partial R^2 parameterization is scale-free and it further enables us to construct a number of useful analyses, including: (i) assessing the sensitivity of an estimate to any number or even *all* confounders acting together, possibly non-linearly; (ii) using the same framework to assess the sensitivity of point estimates as well as t-values and confidence intervals; (iii) assessing the sensitivity to extreme-scenarios in which all or a big portion of the unexplained variance of the outcome is due to confounding; (iv) applying contextual information about the research design to bound the strength of the confounders; and (v) 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 OLS the following holds, $R^2_{Z\sim D} = \frac{\text{var}(\hat{Z})}{\text{var}(Z)} = 1 - \frac{\text{var}(Z^{\perp D})}{\text{var}(Z)} = \text{cor}(Z,\hat{Z})^2 = \text{cor}(Z,D)^2$, where $\hat{Z}$ are the fitted values given by regressing Z on D. Notice the R^2 is symmetric, that is, it is invariant to whether one uses the "forward" or the "reverse" regression since $R^2_{Z\sim D} = \text{cor}(Z,D)^2 = \text{cor}(D,Z)^2 = R^2_{D\sim Z}$. Extending this to the case with covariates $\boldsymbol{X}$, we denote the partial R^2 from regressing Z on D after controlling for $\boldsymbol{X}$ as $R^2_{Z\sim D|\boldsymbol{X}}$. This has the same useful symmetry, with $R^2_{Z\sim D|\boldsymbol{X}} = 1 - \frac{\text{var}(Z^{\perp \boldsymbol{X},D})}{\text{var}(Z^{\perp \boldsymbol{X}})} = \text{cor}(Z^{\perp \boldsymbol{X}}, D^{\perp \boldsymbol{X}})^2 = \text{cor}(D^{\perp \boldsymbol{X}}, Z^{\perp \boldsymbol{X}})^2 = R^2_{D\sim Z|\boldsymbol{X}}$.

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

$$\begin{aligned}
\widehat{\text{bias}} &= \hat{\delta}\hat{\gamma} \\
&= \left(\frac{\text{cov}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})}{\text{var}(D^{\perp \boldsymbol{X}})}\right)\left(\frac{\text{cov}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})}{\text{var}(Z^{\perp \boldsymbol{X},D})}\right) \\
&= \left(\frac{\text{cor}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})\text{sd}(Z^{\perp \boldsymbol{X}})}{\text{sd}(D^{\perp \boldsymbol{X}})}\right)\left(\frac{\text{cor}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(Z^{\perp \boldsymbol{X},D})}\right) \\
&= \left(\frac{\text{cor}(Y^{\perp \boldsymbol{X},D},\ Z^{\perp \boldsymbol{X},D})\text{cor}(D^{\perp \boldsymbol{X}},\ Z^{\perp \boldsymbol{X}})}{\frac{\text{sd}(Z^{\perp \boldsymbol{X},D})}{\text{sd}(Z^{\perp \boldsymbol{X}})}}\right)\left(\frac{\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(D^{\perp \boldsymbol{X}})}\right) \qquad (7)
\end{aligned}$$

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

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

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

$$\widehat{\text{se}}(\hat{\tau}_{\text{res}}) = \frac{\text{sd}(Y^{\perp \boldsymbol{X},D})}{\text{sd}(D^{\perp \boldsymbol{X}})}\sqrt{\frac{1}{\text{df}}} \qquad (9)$$

$$\widehat{\text{se}}(\hat{\tau}) = \frac{\text{sd}(Y^{\perp \boldsymbol{X},D,Z})}{\text{sd}(D^{\perp \boldsymbol{X},Z})}\sqrt{\frac{1}{\text{df}-1}}, \qquad (10)$$

whose ratio is

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

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

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

Moreover, with this we can further see the bias as

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

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 $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$. These formulae are computationally convenient—the only data dependent parts are the standard error of $\hat{\tau}_{\text{res}}$ and the regression's degrees of freedom, which are already reported by most regression software. In this section, we provide remarks that help making 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.

Sensitivity of the point estimate

In the partial R^2 parameterization, the relative bias, $\left|\frac{\widehat{\text{bias}}}{\hat{\tau}_{\text{res}}}\right|$, has a simple form:[10]

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

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.[11] Collectively this numerator could be called the "bias factor" of the confounder, $\text{BF} = |R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}|$, which is determined entirely by the two sensitivity parameters $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$. To determine the size of the relative bias, this is compared to 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, $f_{Y\sim D|\boldsymbol{X}}$ can be obtained by dividing the t-value of the treatment coefficient by the square-root of the regression's degrees of freedom—$f_{Y\sim D|\boldsymbol{X}} = t_{\hat{\tau}_{\text{res}}}/\sqrt{\text{df}}$. This allows one to easily assess sensitivity to any confounder with a given pair of partial R^2 values, see Table 2 in Supplement D for an illustrating procedure.

[10] See Supplement A for details.

[11] Cohen's f^2 can be written as $f^2 = R^2/(1-R^2)$, so, for example, $f^2_{D\sim Z|\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}}/(1-R^2_{D\sim Z|\boldsymbol{X}})$.

Equation 14 also reveals that, given a particular confounder (which will fix BF), the only property needed to determine the robustness of a regression estimate against that confounder is the partial R^2 of the treatment with the outcome (via $f_{Y\sim D|\boldsymbol{X}}$). This serves to reinforce the fact that robustness to confounding is an identification problem, impervious to sample size considerations. While 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 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 one. By contrast, the effect of the partial R^2 of the confounder with the treatment on the bias factor is unbounded (via $f_{D\sim Z|\boldsymbol{X}}$). This allows us to consider extreme scenarios, in which we suppose the confounder explains *all* of the left-out variation of the outcome, and see what happens as we vary the partial R^2 of the confounder with the treatment (Section 5.3).

Sensitivity of the variance

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

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

That is, including the confounder in the regression reduces the estimate of the variance of the coefficient of D by reducing the residual variance of Y (variance reduction factor—VRF). On the other hand, it raises the estimated 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 formally recover the answer one would obtain from including the omitted variable. The overall relative change of the estimated variance is simply the product of these three components.

Reasoning about $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$

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

First, as noted in Section 4.1, researchers 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.[12]

[12] For instance, since D will usually be a *post-treatment* variable with respect to Z, this can make the the association of Y and Z conditional on D harder to interpret, especially when one wants to attach a causal meaning to the parameter (Rosenbaum 1984). As argued in footnote 7, however, recall that a causal interpretation of the association of Z with Y requires more assumptions than the ones usually invoked for the identification of the causal effect of D on Y.

This can be done by noting that, for a choice of $R_{Y\sim Z|\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$, we can reconstruct $R_{Y\sim Z|D,\boldsymbol{X}}$ using the recursive definition of partial correlations,

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

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

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

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

That is, plausibility judgments about the partial R^2 boil down to plausibility judgments about the *total (or added) explanatory power* that one would have obtained in the treatment and the outcome regressions, had the unobserved confounder Z been included. This may be particularly useful when contemplating multiple confounders acting in concert (as we will 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 using the previous results, as we will 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: (i) they can be routinely reported in standard regression tables, making the discussion of sensitivity to unobserved confounding more accessible and standardized; and, (ii) they can be easily computed from quantities found on a regression table, allowing readers and reviewers to initiate the discussion about unobserved confounders when reading papers that did not formally assess sensitivity.

The robustness value

The first quantity 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. $R^2_{Y\sim Z|\boldsymbol{X},D} = R^2_{D\sim Z|\boldsymbol{X}} = RV_q$. The RV_q describes how strong that association must be in order to reduce the estimated effect by $(100 \times q)$%. By Equation 14 (see Supplement A),

$$RV_q = \frac{1}{2}\left(\sqrt{f_q^4 + 4f_q^2} - f_q^2\right) \tag{18}$$

where $f_q := q|f_{Y\sim D|\boldsymbol{X}}|$ 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 RV_q% of the residual variance both of the treatment and the outcome are sufficiently strong to change the point estimate in problematic ways, while confounders with neither association greater than RV_q% are not.

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

With minor adjustment, robustness values can also be obtained for t-values, or lower and upper bounds of confidence intervals. Let $|t^*_{\alpha,\text{df}-1}|$ denote the t-value threshold for a t-test with significance level of α and $\text{df}-1$ degrees of freedom, and define $f^*_{\alpha,\text{df}-1} := |t^*_{\alpha,\text{df}-1}|/\sqrt{\text{df}-1}$. 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,

$$f_{q,\alpha} := f_q - f^*_{\alpha,\text{df}-1} \tag{19}$$

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

$$RV_{q,\alpha} = \frac{1}{2}\left(\sqrt{f^4_{q,\alpha} + 4f^2_{q,\alpha}} - f^2_{q,\alpha}\right), \tag{20}$$

both with the treatment and with the outcome is sufficiently strong to make the adjusted t-test not reject the null hypothesis $H_0 : \tau = (1-q)|\hat{\tau}_{\text{res}}|$ at the α level (or, equivalently, to make the adjusted $1-\alpha$ confidence interval include $(1-q)|\hat{\tau}_{\text{res}}|$). When $RV_{q,\alpha} > 1 - 1/f^2_q$ then, as with the RV_q, we can conclude that no confounder with both associations lower than $RV_{q,\alpha}$ is able to overturn the conclusion of such a test. In the rare cases when $RV_{q,\alpha} \leq 1 - 1/f^2_q$, setting $RV_{q,\alpha} = (f^2_q - f^{*2}_{\alpha,\text{df}-1})/(1 + f^2_q)$ restores the property that no confounder weaker on both associations would change the conclusion.[13] Note that, since we are considering sample uncertainty, $RV_{q,\alpha}$ is a more conservative measure than RV_q. For a fixed $|t^*_{\alpha,\text{df}-1}|$, $RV_{q,\alpha}$ converges to RV_q when the sample size grows to infinity.

The $R^2_{Y\sim D|\boldsymbol{X}}$ as an extreme scenario analysis

The second measure we propose is the proportion of variation in the outcome uniquely explained by the treatment—$R^2_{Y\sim D|\boldsymbol{X}}$. 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 in order to eliminate the estimated effect?" As it happens, the answer is precisely the $R^2_{Y\sim D|\boldsymbol{X}}$.

Specifically, a confounder explaining *all* residual variance of the outcome implies $R_{Y\sim Z|D,\boldsymbol{X}} = 1$. By Equation 14, to bring the estimated effect down to zero (relative bias = 1), this means $|f_{D\sim Z|\boldsymbol{X}}|$ needs to equal $|f_{Y\sim D|\boldsymbol{X}}|$, which implies $R^2_{D\sim Z|\boldsymbol{X}} = R^2_{Y\sim D|\boldsymbol{X}}$. Thus, $R^2_{Y\sim D|\boldsymbol{X}}$ 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 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 alleviate this difficulty. The rationale for the method is the realization that, while in some cases an investigator may not be able make direct plausibility judgments 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.

[13] This occurs when the variance reduction due to an increase in $R^2_{Y\sim Z|D,\boldsymbol{X}}$ dominates its effect on the bias. Such cases are unlikely in practice, see Supplement A for details.

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 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. 2016b; 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 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 the ones that are best suited for a given analysis) and that those measures have relevant interpretations (Kruskal and Majors 1989). An example of the risks entailed by ignoring this requirement can be found in the coefficient of "proportional selection on observables" advanced by Oster (2017), 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: (i) how the *total* R^2 of the confounder compares with the *total* R^2 of a group of observed covariates; (ii) 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, (iii) 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 OLS, 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 one 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 likely 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, but we refer readers to Supplement B for further discussion and derivations of the other two variants.[14]

Assume $Z \perp \boldsymbol{X}$, or, equivalently, consider only the part of Z not linearly explained by $\boldsymbol{X}$. Now suppose 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 to a observed covariate X_j. More formally, define k_D and k_Y as,

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

[14] Another reason we employ this type of bound in the main text is that it is most closely related to approaches used by other sensitivity analyses to which we contrast our results. These include the informal benchmarks of Imbens (2003) as well as the bounding proposal of Oster (2017), discussed in Section 6.

Where $\boldsymbol{X}_{-j}$ represents the vector of covariates $\boldsymbol{X}$ excluding X_j. That is, 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 the omission of X_j would result in a larger mean squared error of the treatment assignment regression than the omission of Z, this equals the claim $k_D \leq 1$. The same reasoning applies to k_Y.

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

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

where η is a scalar which depends on k_Y, k_D and $R^2_{D\sim X_j|\boldsymbol{X}-j}$, (see Supplement B for details). These equations allow us to investigate the maximum effect 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 $R^2_{D\sim X_j|\boldsymbol{X}_{-j}} = 50\%$ 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 $\boldsymbol{X}$ or any subset thereof. The method can also be extended to allow different subgroups of covariates to bound $R^2_{D\sim Z|\boldsymbol{X}}$ and $R^2_{Y\sim Z|D,\boldsymbol{X}}$— thus, if a group of covariates $\boldsymbol{X}_1$ is known to be the most important driver of selection to treatment, and another group of covariates $\boldsymbol{X}_2$ 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 caused by a single confounder. Fortunately, they also provide *upper bounds* in the case of *multiple* unobserved confounders.[15] Allowing $\boldsymbol{Z}$ to be a set (matrix) of confounders and $\hat{\boldsymbol{\gamma}}$ its coefficient vector, the full equation we wished we had estimated becomes

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

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

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

Accordingly, Z^* has the same partial R^2 with the outcome as the full set $\boldsymbol{Z}$. However, the partial R^2 of Z^* with the treatment must be less than or equal to the partial R^2 of $\boldsymbol{Z}$ with the treatment—this follows simply because the choice of the linear combination $\hat{\boldsymbol{\gamma}}$ is the one the maximizes the R^2 with the outcome, and not with the treatment. Hence, the bias caused by a multivariate $\boldsymbol{Z}$ must be less than or equal the bias computed 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 $\boldsymbol{Z}$, simply modifying sensitivity Equation 12 to account for the correct degrees of freedom ($\text{df} - k$ instead of $\text{df} - 1$) will give conservative adjusted standard errors for a multivariate confounder. From a practical point of view, however, we

[15] See Hosman et al. (2010), Section 4.1, for an alternative proof.

note that further correction of the degrees of freedom might be an unnecessary formality—we are performing a hypothetical exercise, and one can always imagine to have measured Z^*.

Finally, note the set of confounders $\boldsymbol{Z}$ is arbitrary, thus it accommodates nonlinear confounders as well as misspecification of the functional form of the observed covariates $\boldsymbol{X}$. To illustrate the point, let $Y = \hat{\tau}D + \hat{\beta}X + \hat{\gamma}_1 Z + \hat{\gamma}_2 Z^2 + \hat{\gamma}_3(Z \times X) + \hat{\gamma}_4 X^2 + \hat{\varepsilon}_{\text{full}}$, and imagine the researcher did not measure Z and did not consider that X could also enter the equation with a squared term. Now just call $\boldsymbol{Z} = (Z_1 = Z, Z_2 = Z^2, Z_3 = Z \times X, Z_4 = X^2)$ 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: (i) How strong would a particular confounder (or group of confounders) have to be to change our conclusions? (ii) In a worst case scenario, how vulnerable is our result to *many* or *all* unobserved confounders acting together, possibly non-linearly? (iii) 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 we propose should accompany linear regression models used for causal inference with observational data. Along with traditionally reported statistics, we propose researchers present (i) the partial R^2 of the treatment with the outcome, and (ii) the robustness value, RV, both for where the point estimate and the confidence interval would cross zero (or another meaningful reference value).[16] Finally, in order to aid user judgment, we encourage researchers to provide plausible bounds on the strength of the confounder. These may be based upon 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.

Outcome: *Peace Index*

Treatment:	Est.	SE	t-value	$R^2_{Y\sim D\|\boldsymbol{X}}$	RV	$RV_{\alpha=0.05}$
Directly Harmed	0.097	0.023	4.18	2.2%	13.9%	7.6%
df = 783, Bound (Z as strong as *Female*): $R^2_{Y\sim Z\|D,\boldsymbol{X}} = 12\%$, $R^2_{D\sim Z\|\boldsymbol{X}} = 1\%$						

Table 1: Proposed minimal reporting on sensitivity to unobserved confounders

For our running example of violence in Darfur, Table 1 shows an augmented regression table, including the robustness value (RV) of the *Directly Harmed* 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 strong enough to bring down the estimated effect to zero. For cases where one association is over 13.9% and the other is below, we conduct additional analyses illustrated in the next subsection. Nevertheless, the 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.[17]

[16] For convenience, we refer to the RV_q or $RV_{q,\alpha}$ with $q = 1$ as simply the RV or RV_α

[17] That is, any confounder with an equivalent bias factor of BF $= RV/\sqrt{1 - RV}$.

Adjusting for confounding may not bring the estimate to zero, but rather into a range where it is no longer "statistically significant". Therefore, the robustness value accounting for statistical significance, $RV_{\alpha=0.05}$, is also shown in the table. For a significance level of 5%, the robustness value goes down from 13.9% to 7.6%—that is, 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, $R^2_{Y\sim D|\boldsymbol{X}}$, 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 zero.

Confronted with those results, we now need to judge whether confounders with the strengths revealed to be problematic are plausible. If one 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 judge where it falls relative to these quantities. The lower right corner of Table 1 shows the strength of association that a confounder as strong as *Female* would have: $R^2_{Y\sim Z|D,\boldsymbol{X}} = 12\%$ and $R^2_{D\sim Z|\boldsymbol{X}} = 1\%$. As the robustness value is higher than either quantity, the table readily reveals that such a confounder could not fully eliminate the point estimate. In addition, since the bound for $R^2_{D\sim Z|\boldsymbol{X}}$ is less than $R^2_{Y\sim D|\boldsymbol{X}} = 2.2\%$, 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 characteristic 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 one can argue that total confounding as strongly associated with the treatment as *Female* is implausible, those bounding results show 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 $(R^2_{Y\sim Z|D,\boldsymbol{X}}, R^2_{D\sim Z|\boldsymbol{X}})$ can also be easily computed with the information on the table, see example in Supplement D.

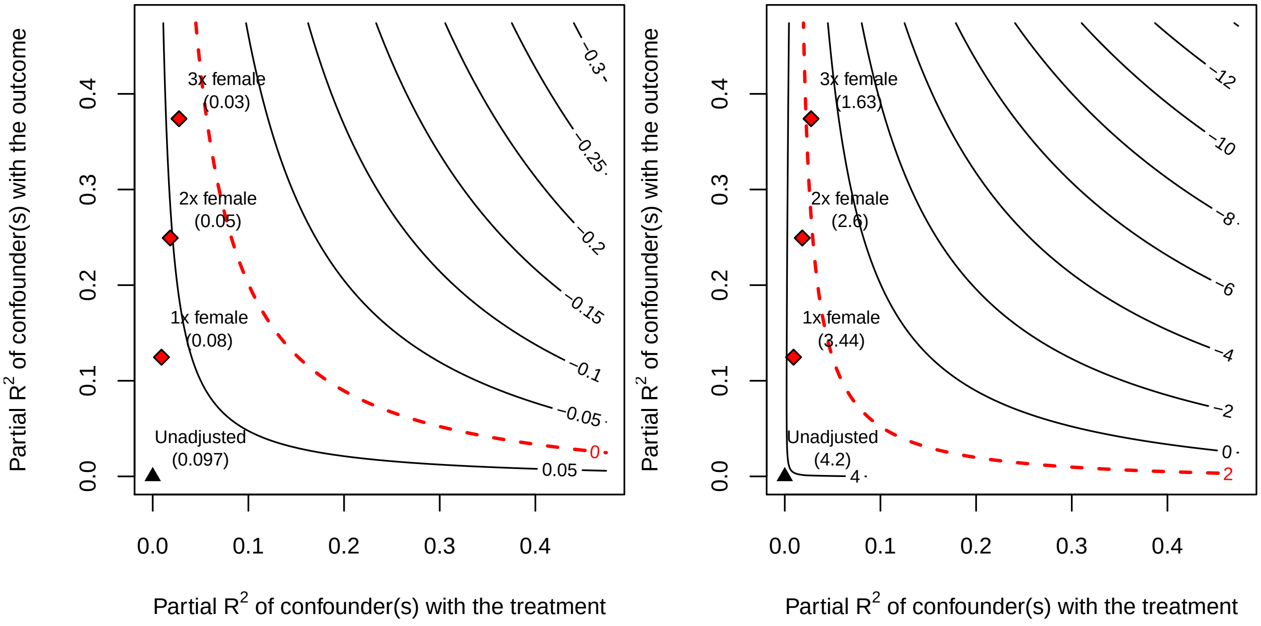
5.2 Sensitivity contour plots with partial R^2: estimates and t-values

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 to observables.[18]

Perhaps the first plot investigators would examine would be one similar to Figure 1, but now in the partial R^2 parameterization (Figure 2a). 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.[19] The contours show the adjusted estimate that would be obtained for an unobserved confounder (in the full model) with

[18] Here we focus on the plots for point estimates and t-values, but note 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.

[19] As discussed in Section 4.2, axes could be transformed to show instead (i) the total R^2 including the confounders $R^2_{Y\sim D+\boldsymbol{X}+Z}$ and $R^2_{D\sim\boldsymbol{X}+Z}$, (ii) the difference in the total R^2 including the confounders, i.e., $R^2_{Y\sim D+\boldsymbol{X}+Z} - R^2_{Y\sim D+\boldsymbol{X}}$ and $R^2_{D\sim\boldsymbol{X}+Z} - R^2_{D\sim\boldsymbol{X}}$, (iii) the partial correlations (by simply taking the square-root), (iv) the partial R^2 of the confounder with the outcome *not* conditioning on the treatment, among other options that may aid interpretation.



(a) Sensitivity contour plot of the point estimate

(b) Sensitivity contour plot of the t-value

Figure 2: Sensitivity contour plots in the partial R^2 scale with benchmark bounds

the hypothesized values of the sensitivity parameters (assuming the direction of the effects hurts our preferred hypothesis).

While the contour plot used in illustrating the traditional OVB approach focused on a specific binary confounder—*Center*—the contour plot with the partial R^2 parameterization allows us to assess sensitivity to any confounder, irrespective 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 *Political Attitudes* and *Wealth* acting together. Notice that if we choose a contour of interest (such as where the effect equals zero), and find the point with equal values on the horizontal and vertical axes (i.e. where it crosses a 45-degree line), this correspond to the robustness value. That is, the RV 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, respectively.

Moving to inferential concerns, Figure 2b 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 get 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 the statistical significance of *Directly Harmed* is robust to a confounder as strong as, or twice as strong as *Female*. However, whereas confounders *three times* as strong as *Female* would not erode the point estimate to zero, we cannot guarantee 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? While one 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.

5.3 Sensitivity plots of extreme scenarios

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

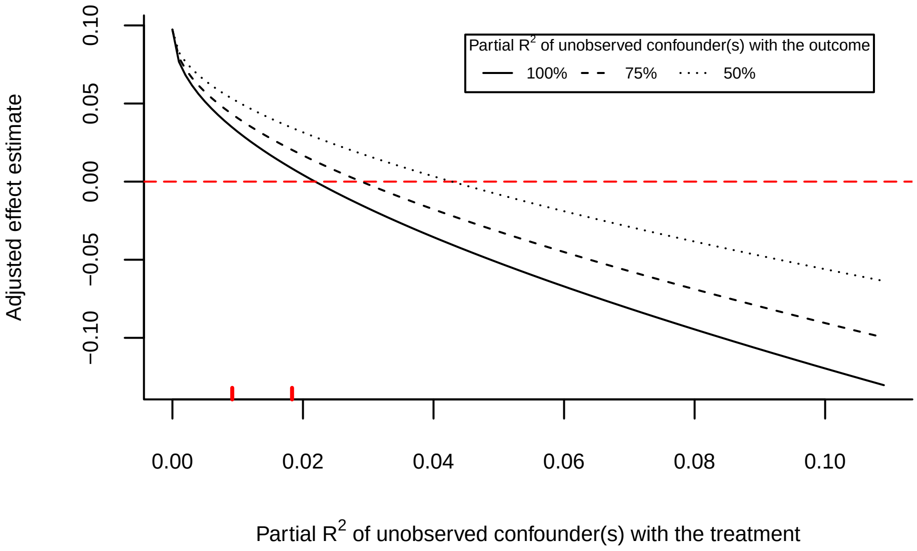


Figure 3: Sensitivity analysis to extreme scenarios

Applying this to our running example, results are shown in Figure 3. The solid 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, $R^2_{D\sim Z|\boldsymbol{X}}$ 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 zero—that is, it would need to be at least 2.2%, a value below the bound for a confounder once or twice as strong as *Female* (shown by red 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 $R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$ might be needlessly conservative. Hence, we propose plotting other extreme scenarios, as shown in Figure 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, one might expect that sensitivity analyses would be a routine procedure in numerous disciplines. Why then are they not commonplace? We surmise there are three main obstacles, which we directly address in this paper.

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 (1983a), Imbens (2003), Carnegie et al. (2016b) and Dorie et al. (2016) require specifying the distribution of the confounder as well as modeling the treatment assignment mechanism; in another direction, the methods put forward in Robins (1999), Brumback et al. (2004), Blackwell (2013) need to directly specify a confounding function parameterizing the difference in potential outcomes among treated and control units. While 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 nonlinear) confounders, even including mispecification of the functional form of the observed covariates.

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 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: (i) the robustness value, which conveniently summarizes the minimal strength of association a confounder needs to have to change the research conclusions, and (ii) the $R^2_{Y\sim D|\boldsymbol{X}}$, which works as an extreme-scenario sensitivity analysis. Regarding the robustness value in particular, we now discuss its relation to two other proposals advocated in the literature: the *impact thresholds* of Frank (2000) and the E-value of VanderWeele and Ding (2017).

Frank (2000) proposes characterizing the strength of the unobserved confounder Z with what he denotes as its *impact*, defined as the product $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$.[20] This is then used to determine *impact thresholds*, defined as the minimum impact of the unobserved confounder necessary to not

[20] Not to confuse with $\hat{\gamma}$ of the "impact times imbalance" heuristic, as discussed in Section 3.2.

reject the null hypothesis of *zero effect.* However, as Equation 14 reveals, the determinant of the bias is the bias factor BF $= R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}$, 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 $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$ explicitly,

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

Equation 25 reveals that: (i) an unobserved confounder with *zero impact* can still cause non-zero (downward) bias; (ii) an unobserved confounder with a *non-zero impact* can nevertheless induce zero bias (when impact $= R_{Y\sim D|\boldsymbol{X}} R^2_{D\sim Z|\boldsymbol{X}}$); and, (iii) the two terms that compose the product $R_{Y\sim Z|\boldsymbol{X}} \times R_{D\sim Z|\boldsymbol{X}}$ 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 proposed in Frank (2000) to arbitrary non-zero null hypothesis of regression coefficients.[21] Note 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 BF $= RV_q/\sqrt{1 - RV_q}$.

As to VanderWeele and Ding (2017), the authors 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, while 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 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.

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 upon the nature of the domain knowledge used for plausibility judgments. 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 comparisons. This allows 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 due to informal benchmarking practices that do not warrant the claims they purport to make, or by relying on inappropriate parameterization choices.

[21] Let q denote the relative bias and consider biases that move the effect toward (or through) zero. Solving Equation 25 for *impact* gives us impact $= R_{Y\sim D|\boldsymbol{X}}(q - (q-1)R^2_{D\sim Z|\boldsymbol{X}})$. Note that, given q and $R_{Y\sim D|\boldsymbol{X}}$, the *impact* necessary to bring about a relative bias of magnitude q still depends on the sensitivity parameter $R^2_{D\sim Z|\boldsymbol{X}}$—except when $q = 1$. For a numerical example, see Supplement A.5.

6.2 The risks of informal benchmarking

While prior work has suggested informal benchmarking procedures using statistics of observed covariates $\boldsymbol{X}$ to help researchers "calibrate" their intuitions about the strength of the unobserved confounder Z (Frank 2000; Imbens 2003; Hosman et al. 2010; Dorie et al. 2016; Carnegie et al. 2016b; 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 to $\boldsymbol{X}$. 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 one assumes Z to be independent of $\boldsymbol{X}$. To illustrate this threat concretely, let us 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*.[22] The results are shown in Figure 4.

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

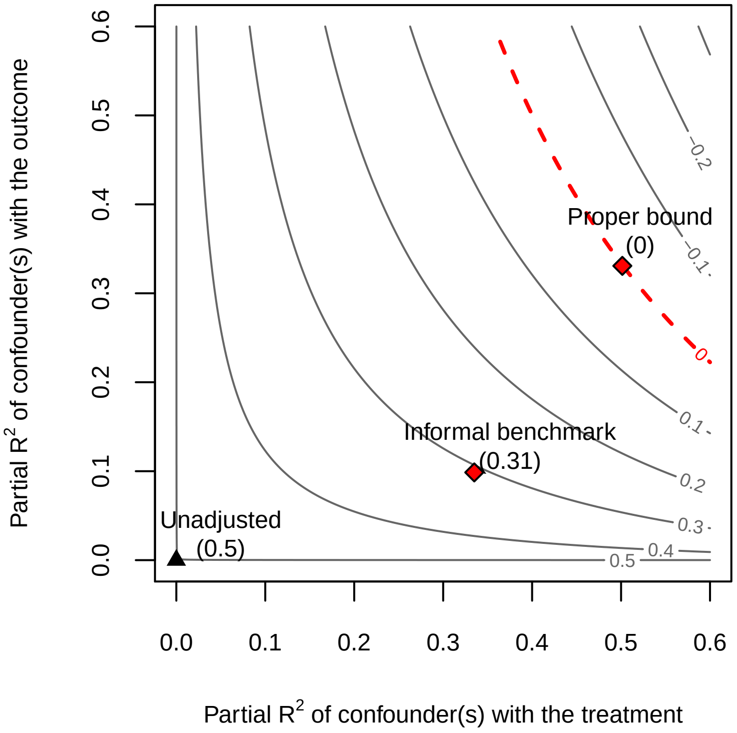


Figure 4: Sensitivity contours of point estimate — informal benchmarking *versus* proper bounds

[22] We use structural equations, $Y = X + Z + \varepsilon_y$, $D = X + Z + \varepsilon_d$, $X = \varepsilon_x$, $Z = \varepsilon_z$ where all disturbances, are independent standard normal random variables. See also Supplement C.

Why exactly does this happen? Consider for a moment the difference between the coefficient on $\boldsymbol{X}$ in the full Equation 3, $\hat{\boldsymbol{\beta}}$, and its estimate in the restricted Equation 4, $\hat{\boldsymbol{\beta}}_{\text{res}}$. Using the same OVB approach of "impact times imbalance", we arrive at $\hat{\boldsymbol{\beta}}_{\text{res}} - \hat{\boldsymbol{\beta}} = \hat{\gamma}\hat{\boldsymbol{\psi}}$, where $\hat{\boldsymbol{\psi}}$ is obtained from the regression $Z = \hat{\delta}D + \boldsymbol{X}\hat{\boldsymbol{\psi}} + \hat{\varepsilon}_Z$. Note that $\hat{\boldsymbol{\psi}}$ can be non-zero even if $\boldsymbol{X} \perp Z$, because D is a collider (Pearl 2009), and conditioning on D creates dependency between Z and $\boldsymbol{X}$. The reasoning holds whether one is 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 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 $k_D R^2_{D\sim X_j|\boldsymbol{X}_{-j}}$ and $k_Y R^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}$, instead of the proper bounds $k_D f^2_{D\sim X_j|\boldsymbol{X}_{-j}}$ and $\eta^2 f^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}$. There are, thus, two discrepancies: (i) 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 coordinates of the benchmark; and, (ii) the collider bias due to the association of X_j with D, which affects only the bound on $R^2_{Y\sim Z|D,\boldsymbol{X}}$ via $\eta^2 \geq k_Y$.[23] Therefore, the stronger the association of X_j with the treatment, and the larger the multiples used for comparisons ("k times as strong"), the more misleading informal benchmarks will be.[24] We thus advise against informal benchmarking procedures, and previous studies relying upon 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 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) ask researchers to "calibrate intuitions" about the strength of the confounder with the treatment 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.[25]

An alternative bounding argument has also been presented in Oster (2017) which, unlike the informal benchmarking practices previously discussed, provides a formal identification result. Nevertheless, the proposed procedure asks users to reason about a quantity that is very difficult to understand. More precisely, Oster (2017) asks researchers to make plausibility judgments on two sensitivity parameters, $R_{\max}$ and δ_{Oster}. The $R_{\max}$ parameter is simply the maximum explanatory

[23] 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.

[24] In our running example, since *female* explains less than 1% of the residual variance of the treatment, informal benchmarks would not be markedly different from the formal ones.

[25] The t-value in the expression of the bias is an artifact of both multiplying and dividing by the degrees of freedom, as in our Equation 12. While t-values can be useful for computational purposes (to utilize quantities 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 (partial R^2 of 0.9975), while 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.

power that one could have with the full outcome regression, i.e., $R_{\max} = R^2_{Y\sim D+\boldsymbol{X}+\boldsymbol{Z}}$. As discussed in Section 4.2 (Equation 17) this has a one to one relationship with $R^2_{Y\sim \boldsymbol{Z}|\boldsymbol{X},D}$,

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

By contrast the second sensitivity parameter, δ_{Oster}, is not easily interpretable in substantive terms. Following Altonji et al. (2005), Oster (2017) defines "indices" $W_1 := \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $W_2 := \boldsymbol{Z}\hat{\boldsymbol{\gamma}}$, where $\boldsymbol{X}$ is a matrix of observed covariates and $\boldsymbol{Z}$ a matrix of unobserved covariates. Critically, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are chosen such that $Y = \hat{\tau}D + W_1 + W_2 + \hat{\varepsilon}_{\text{full}}$.[26] 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 $\boldsymbol{X}$ and $\boldsymbol{Z}$ 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 "the unobservable and observables are equally related to the treatment" (Oster 2017, p.6) can lead researchers astray, as this quantity also depends upon associations with the outcome. To see how, let the variables be standardized to mean zero 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 one fourth of the explanatory power of X (as measured by standardized coefficients or variance explained), yet $\delta_{\text{Oster}} = 1$. While 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.[27] By contrast, the parameter k_D we introduce 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 these arguments imply. It is thus important that the sensitivity parameters used in these exercises be as transparent as possible and match investigators' conception of what the parameters imply. 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

[26] Oster (2017) uses 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.

[27] Indeed, arguments made by researchers applying Oster (2017) suggest 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. "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." (Jakiela and Ozier 2018, p.4)

lead to the rich set of sensitivity exercises 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 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 judgment, 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 prepared to believe in order to accept the substantive claims initially made (Rosenbaum 2005 2010 2017). The sensitivity exercises proposed here tell the researcher how strong unobserved confounding would have to be in order to meaningfully change 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 shown to be problematic depends on expert judgment. As a consequence, the research design, identification strategy as well as the story explaining the quality of the covariates 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 is 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 credibly claim 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. On the other hand, in a quasi-experiment where the researcher knows 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, p. 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 skeptic of a causal estimate—not just any criticism is able to invalidate the research conclusions. The hypothesized unobserved confounder now has to 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 gatekeeping 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 to not 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 roadmap for improving upon the robustness of these answers in their following inquiries.

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. *Journal of Human resources*, 40(4):791–821.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* 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.

Blackwell, M. (2013). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, 22(2):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. *Statistics in medicine*, 23(5):749–767.

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

Carnegie, N. B., Harada, M., and Hill, J. L. (2016b). Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420.

Chen, B. and Pearl, J. (2015). Exogeneity and robustness. Technical report, Tech. Rep.

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. *International Conference on Machine Learning.*

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. *journal of National Cancer Institute*, (23):173–203.

Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57.

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470.

Dunning, T. (2012). *Natural experiments in the social sciences: a design-based approach.* Cambridge University Press.

Flint, J. and de Waal, A. (2008). *Darfur: a new history of a long war.* Zed Books.

Frank, K. and Min, K.-S. (2007). Indices of robustness for sample representation. *Sociological Methodology*, 37(1):349–392.

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2):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. *Educational Evaluation and Policy Analysis*, 35(4):437–460.

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? *Educational Evaluation and Policy Analysis*, 30(1):3–30.

Franks, A., DAmour, A., and Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, (just-accepted):1–38.

Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401.

Hazlett, C. (2019+). Angry or weary? The effect of personal violence on attitudes towards peace in darfur. *Journal of Conflict Resolution*, Forthcoming.

Hong, G., Qin, X., and Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43(1):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. *The Annals of Applied Statistics*, pages 849–870.

Imai, K., Keele, L., Yamamoto, T., et al. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, 25(1):51–71.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *The American Economic Review*, 93(2):126–132.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Jakiela, P. and Ozier, O. (2018). Gendered language. *Policy Research Working Paper, World Bank*.

Kruskal, W. and Majors, R. (1989). Concepts of relative importance in recent scientific literature. *The American Statistician*, 43(1):2–6.

Leamer, E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Leamer, E. E. (2016). S-values: Conventional context-minimal measures of the sturdiness of regression coefficients. *Journal of Econometrics*, 193(1):147 – 161.

Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Lovell, M. C. (2008). A simple proof of the FWL theorem. *The Journal of Economic Education*, 39(1):88–91.

Middleton, J. A., Scott, M. A., Diakow, R., and Hill, J. L. (2016). Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323.

Oster, E. (2014). Unobservable selection and coefficient stability: Theory and evidence. *NBER working paper*.

Oster, E. (2017). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, pages 1–18.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121(1):151–179.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666.

Rosenbaum, P. R. (2002). Observational studies. In *Observational studies*, pages 1–17. Springer.

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. In *Encyclopedia of statistics in behavioral science*, volume 4, pages 1809, 1814. John Wiley & Sons Ltd.

Rosenbaum, P. R. (2010). *Design of observational studies*. Springer Series in Statistics.

Rosenbaum, P. R. (2017). *Observation and experiment: an introduction to causal inference*. Harvard University Press.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Steiner, P. M. and Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of causal inference*, 4(2).

Vanderweele, T. J. and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274.

Online supplementary material for “Making Sense of Sensitivity: Extending Omitted Variable Bias”

Carlos Cinelli & Chad Hazlett

A Simple measures for routine reporting

A.1 Preliminaries

For any univariate regression, recall $R^2 = t^2/(t^2+\text{df})$, $t^2 = \left(\frac{R^2}{1-R^2}\right)\text{df}$, and $f^2 = \frac{R^2}{1-R^2} = \frac{t^2}{\text{df}}$, where df is the regression's degrees of freedom. Repeating the partialling out procedure to allow for covariates, the partial R^2 of any covariate can be written in terms of its coefficient's t statistic and *vice-versa*.

For instance, the partial R^2 of the confounder with the treatment, conditional on $\boldsymbol{X}$, can be written as

$$R^2_{D\sim Z|\boldsymbol{X}} = \frac{t^2_{\hat{\delta}}}{t^2_{\hat{\delta}} + \text{df}}. \tag{28}$$

Analogously,

$$f^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}} = \frac{t^2_{\hat{\delta}}}{\text{df}}. \tag{29}$$

A.2 General strength of a confounder

Consider a confounder strong enough to change the estimated treatment effect by $(100 \times q)\%$. This means that $|\widehat{\text{bias}}| = q|\hat{\tau}_{\text{res}}|$. Hence, by equation 13 we have that

$$q|\hat{\tau}_{\text{res}}| = \sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\ R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}}}\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}. \tag{30}$$

Dividing both sides by $\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}$ and noting $\frac{|\hat{\tau}_{\text{res}}|}{\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\text{df}}} = \frac{|t_{\hat{\tau}_{\text{res}}}|}{\sqrt{\text{df}}} = f_{Y\sim D|\boldsymbol{X}}$, we obtain

$$\begin{aligned} q|f_{Y\sim D|\boldsymbol{X}}| &= \sqrt{\frac{R^2_{Y\sim Z|D,\boldsymbol{X}}\ R^2_{D\sim Z|\boldsymbol{X}}}{1 - R^2_{D\sim Z|\boldsymbol{X}}}} && (31)\\ &= |R_{Y\sim Z|D,\boldsymbol{X}} \times f_{D\sim Z|\boldsymbol{X}}| && (32)\\ &= \text{BF}. && (33) \end{aligned}$$

That is, to bring the estimated effect down by $(100 \times q)\%$, the bias factor (BF) of the confounder $\left(R_{Y\sim Z|D,\boldsymbol{X}} f_{D\sim Z|\boldsymbol{X}}\right)$ has to equal q times the partial f of the treatment with the outcome.

A.3 Extreme sensitivity scenarios

Considering the extreme case scenario where the confounders explain all the residual variance of the outcome, that is, $R^2_{Y\sim Z|D,\boldsymbol{X}} = 1$, a confounder strong enough to bring down the estimated effect to zero, that is $q = 1$, would need to satisfy $f^2_{Y\sim D|\boldsymbol{X}} = f^2_{D\sim Z|\boldsymbol{X}}$ which implies $R^2_{Y\sim D|\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}}$. This shows the partial R^2 of the treatment with the outcome is itself a measure of an extreme-scenario sensitivity analysis.

A.4 The Robustness Value (RV)

Now consider a confounder with $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = RV_q$. Rearranging terms and squaring Equation 31, one obtains

$$RV_q^2 + f_q^2 RV_q - f_q^2 = 0, \tag{34}$$

where $f_q := q|f_{Y\sim D|\boldsymbol{X}}|$. Solving the quadratic equation for RV_q,

$$RV_q = \frac{1}{2}\left(\sqrt{f_q^4 + 4f_q^2} - f_q^2\right) \tag{35}$$

gives us the equation for the robustness value for the point estimate. Note that, since the derivative of the bias with respect to both sensitivity parameters is positive, any confounder with both associations below RV_q is not strong enough to bring about a relative bias of q.

RV for t-values, or lower and upper bounds of confidence intervals

Imagine the researcher wants to know how strong a confounder would need to be for a $100(1-\alpha)\%$ confidence interval to include a change of $(100 \times q)\%$ of the treatment estimate. Consider again a confounder with equal association with the treatment and the outcome, $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = RV_{q,\alpha}$. By Equation 13,

$$|\hat{\tau}| = |\hat{\tau}_{\text{res}}| - \widehat{\text{se}}(\hat{\tau}_{\text{res}})\frac{RV_{q,\alpha}}{\sqrt{1 - RV_{q,\alpha}}}\sqrt{\text{df}}, \tag{36}$$

where we are assuming the bias reduces the absolute value of the estimated effect. For the opposite direction the subtraction would be changed to addition. Further, for any confounder with equal association with the treatment and the outcome, Equation 12 for the adjusted standard error simplifies to

$$\widehat{\text{se}}(\hat{\tau}) = \widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{\text{df}}{\text{df}-1}}. \tag{37}$$

Let $|t^*_{\alpha,\text{df}-1}|$ denote the t-value threshold for a t-test with significance level of α and $\text{df}-1$ degrees of freedom, and define $f^*_{\alpha,\text{df}-1} := |t^*_{\alpha,\text{df}-1}|/\sqrt{\text{df}-1}$. Now note that, for the adjusted t-test to not reject the hypothesis $H_0 : \tau = (1-q)\hat{\tau}_{\text{res}}$, we must have

$$
\begin{aligned}
|t^*_{\alpha,\text{df}-1}| &\geq \frac{|\hat{\tau}| - (1-q)|\hat{\tau}_{\text{res}}|}{\widehat{\text{se}}(\hat{\tau})} && (38)\\
&\geq \frac{q|\hat{\tau}_{\text{res}}| - \widehat{\text{se}}(\hat{\tau}_{\text{res}})\frac{RV_{q,\alpha}}{\sqrt{1-RV_{q,\alpha}}}\sqrt{\text{df}}}{\widehat{\text{se}}(\hat{\tau}_{\text{res}})\sqrt{\frac{\text{df}}{\text{df}-1}}} && (39)\\
&\geq \left(q|f_{Y\sim D|\boldsymbol{X}}| - \frac{RV_{q,\alpha}}{\sqrt{1-RV_{q,\alpha}}}\right)\sqrt{\text{df}-1}. && (40)
\end{aligned}
$$

Divide by $\sqrt{\text{df}-1}$ and rearrange terms to obtain,

$$\frac{RV_{q,\alpha}}{\sqrt{1-RV_{q,\alpha}}} \geq f_q - f^*_{\alpha,\text{df}-1} = f_{q,\alpha}, \tag{41}$$

where we define $f_{q,\alpha} := f_q - f^*_{\alpha,\text{df}-1}$. Our goal is to find the minimal strength of the confounder $RV_{q,\alpha}$ (which must be positive) such that this inequality holds. Thus, we have two cases. If $f_{q,\alpha} < 0$, then trivially $RV_{q,\alpha} = 0$. This happens when an inclusion of a covariate with zero predictive power would be enough not to reject the null hypothesis, either because the t-value is already low enough, or because it becomes low enough after adjusting for the loss in degrees of freedom.

Now consider the case where $f_{q,\alpha} > 0$, which means the minimum will happen in the equality. Rearrange terms and square to obtain,

$$RV^2_{q,\alpha} + f^2_{q,\alpha}RV_{q,\alpha} - f^2_{q,\alpha} = 0. \tag{42}$$

Solving the quadratic equation for $RV_{q,\alpha}$ gives us the robustness value for a reduction of $(100 \times q)\%$ to not be rejected at the significance level α,

$$RV_{q,\alpha} = \frac{1}{2}\left(\sqrt{f^4_{q,\alpha} + 4f^2_{q,\alpha}} - f^2_{q,\alpha}\right). \tag{43}$$

Note that, due to the variance reduction factor of Eq 15, it could be the case that increasing the sensitivity parameter $R^2_{Y\sim Z|D,\boldsymbol{X}}$ helps with statistical significance. When this happens, there can exist a set of confounders with lower $R^2_{Y\sim Z|D,\boldsymbol{X}}$ and $R^2_{D\sim Z|\boldsymbol{X}}$ than $RV_{q,\alpha}$ able to drive the t-statistic below significance. To check for such cases, we need to verify whether the derivative of the adjusted t-value with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is negative (the derivative with respect to $R^2_{D\sim Z|\boldsymbol{X}}$ is always negative). The t-value for $\hat{\tau}$ for testing the null hypothesis $H_0 : \tau = (1-q)\hat{\tau}_{\text{res}}$ can be written as,

$$t_{\hat{\tau},q} = \frac{\hat{\tau} - (1-q)\hat{\tau}_{\text{res}}}{\widehat{\text{se}}(\hat{\tau})} = \frac{f_q\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}} - \sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}\sqrt{R^2_{D\sim Z|\boldsymbol{X}}}}{1-\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}} \times \sqrt{\text{df}-1} \tag{44}$$

Dividing by $\sqrt{\text{df}-1}$ and taking the derivative with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ gives us,

$$\frac{\partial t_{\hat{\tau},q}}{\partial R^2_{Y\sim Z|D,\boldsymbol{X}}} = \frac{f_q\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}}}\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}} - \sqrt{R^2_{D\sim Z|\boldsymbol{X}}}}{2\sqrt{R^2_{Y\sim Z|D,\boldsymbol{X}}}(1-R^2_{Y\sim Z|D,\boldsymbol{X}})^{3/2}} \tag{45}$$

Equation 45 is negative when the numerator is less than zero, that is, when

$$\frac{R^2_{D\sim Z|\boldsymbol{X}}}{(1-R^2_{D\sim Z|\boldsymbol{X}})R^2_{Y\sim Z|D,\boldsymbol{X}}} > f_q^2 \tag{46}$$

For the point of equal association, $R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{D\sim Z|\boldsymbol{X}} = RV_{q,\alpha}$, the condition in Eq. 46 simplifies to $RV_{q,\alpha} > 1 - 1/f_q^2$. Note that, since $RV \geq 0$ this condition will often hold—for instance, for $q = 1$, whenever the partial R^2 of the treatment with the outcome is less or equal to 50%, the first order condition is guaranteed to hold.

When condition 46 does not hold, Eq 43 is still a useful and meaningful reference point of a specific contour line. However, one may want to alternatively define the $RV_{q,\alpha}$ as the maximum bound on both coordinates such that any confounder with (both) associations below that bound cannot bring the t-value below the chosen critical level. In that case, given a bound of $RV_{q,\alpha}$ on both coordinates, we can solve the following constrained minimization problem,

$$\min_{R^2_{Y\sim Z|D,\boldsymbol{X}},R^2_{D\sim Z|\boldsymbol{X}}} t_{\hat{\tau},q} \text{ s.t. } R^2_{Y\sim Z|D,\boldsymbol{X}} \leq RV_{q,\alpha} \text{ and } R^2_{D\sim Z|\boldsymbol{X}} \leq RV_{q,\alpha} \tag{47}$$

Since the derivative of the adjusted t-value with respect to $R^2_{D\sim Z|\boldsymbol{X}}$ is always negative, the optimum $R^2_{D\sim Z|\boldsymbol{X}}$ always reaches the bound. Next we have two cases: when (i) the derivative of the solution with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is negative, this means the optimum for both arguments reach the bound, and solving for a specific t-value threshold gives $RV_{q,\alpha}$ as before (Eq. 43); when (ii) the derivative of the solution with respect to $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is zero, then the optimal $R^2_{Y\sim Z|D,\boldsymbol{X}}$ is an interior point, which by Eq. 46 equals $R^2_{Y\sim Z|D,\boldsymbol{X}} = RV_{q,\alpha}/((1-RV_{q,\alpha})f_q^2)$. Solving for a specific t-value threshold gives us the bound,

$$RV_{q,\alpha} = \frac{f_q^2 - f^{*2}_{\alpha,\text{df}-1}}{1+f_q^2} \tag{48}$$

Finally, note that if one picks the threshold $|t^*_{\alpha,\text{df}-1}| = 0$ then $RV_{q,\alpha}$ trivially reduces to RV_q. Also note that, for fixed $|t^*_{\alpha,\text{df}-1}|$, when $df \to \infty$ we have that $RV_{q,\alpha} \to RV_q$, since standard errors become irrelevant when compared to the bias of the point estimate.

A.5 *Impact thresholds* (Frank 2000) for non-zero null hypothesis

In Section 6.1 we showed that a confounder's *impact*, as defined in Frank (2000), does not fully characterize the minimal strength of confounding necessary to bring about a certain amount of bias in the regression coefficient, except when the relative bias is unity (that is, when the null hypothesis of interest is *zero*). Thus, the *impact thresholds* obtained in Frank (2000) under the null of zero (in which case $R_{Y\sim Z|\boldsymbol{X}} = R_{D\sim Z|\boldsymbol{X}}$) cannot be immediately generalized to non-zero null hypotheses. Here we provide a simple illustrative numerical example. Consider the case with no observed covariates $\boldsymbol{X}$, a single unobserved confounder Z, all variables standardized to mean zero and unit variance and a sample of size $1,000$. Suppose $\hat{\tau}_{\text{res}} = 0.5$, $\widehat{\text{se}}(\hat{\tau}_{\text{res}}) = 0.0274$ and that we want to learn the minimal strength of Z necessary to bring this estimate to $\hat{\tau} = -0.5$ (a relative bias of 2). Solving the bias equation for the case where $R_{Y\sim Z|\boldsymbol{X}} = R_{D\sim Z|\boldsymbol{X}}$ one would obtain an impact threshold of $2/3$. However, this is not the minimal *impact* that would make $\hat{\tau} = -0.5$. As a counterexample, a confounder with an *impact* as low as 0.51 is sufficient to bring about a change of this magnitude, with $R_{Y\sim Z|\boldsymbol{X}} = 0.515$ and $R_{D\sim Z|\boldsymbol{X}} = 0.99$.

B Formal benchmark bounds

Suppose the researcher has substantive knowledge that certain covariates are “the most important predictors of the outcome” and other covariates “the most important predictors of the treatment assignment”. Imagine, also, that the researcher is willing to defend the claim that the unobserved confounder Z is not “as strong” as those covariates.

In order to use this information for bounding the strength of the confounder Z, we need to give it an operational meaning. We operationalize these types of claim as comparisons of the explanatory power of the confounder vis-a-vis the explanatory power of the observed covariates. Mathematically, we can quantify these comparisons using total or partial R^2 measures. Here we will assume that $Z \perp \boldsymbol{X}$ or, equivalently, that the following analysis applies to the part of Z not linearly explained by covariates $\boldsymbol{X}$.

B.1 Comparing total the R^2 of covariates with the total R^2 of the confounder

Although in the text we use the bounds by comparing partial R^2 measures, perhaps the simplest derivation is the comparison of the total R^2 of observed covariates with the total R^2 of the unobserved confounder Z. Consider an example in which the observed covariate X_j is assumed to be an important predictor of the treatment assignment D. If the researcher believes the correlation of X_j with D to be stronger than the correlation of Z with D, this implies,

$$R^2_{D\sim Z} < R^2_{D\sim X_j}. \tag{49}$$

We could use the same argument for comparing $R^2_{Y\sim Z}$ with $R^2_{Y\sim X_j}$. As it happens, such claims are sufficient to bound the sensitivity parameters. Let us generalize this notion by defining,

$$k_D := \frac{R^2_{D\sim Z}}{R^2_{D\sim X_j}}, \qquad k_Y := \frac{R^2_{Y\sim Z}}{R^2_{Y\sim X_j}}. \tag{50}$$

That is, k_D and k_Y measure how the correlation of Z, with D and Y, compares to the correlation of X_j with those same variables. Our goal here is to re-express both sensitivity parameters as a function of k_D and k_Y. Since $Z \perp \boldsymbol{X}$, we have that

$$R^2_{D\sim Z+\boldsymbol{X}} = R^2_{D\sim Z} + R^2_{D\sim \boldsymbol{X}} = k_D R^2_{D\sim X_j} + R^2_{D\sim \boldsymbol{X}} \tag{51}$$

$$R^2_{Y\sim Z+\boldsymbol{X}} = R^2_{Y\sim Z} + R^2_{Y\sim \boldsymbol{X}} = k_Y R^2_{Y\sim X_j} + R^2_{Y\sim \boldsymbol{X}}. \tag{52}$$

Now we can trivially re-express $R^2_{D\sim Z|\boldsymbol{X}}$ as function of k_D,

$$R^2_{D\sim Z|\boldsymbol{X}} = \frac{R^2_{D\sim Z+X} - R^2_{D\sim \boldsymbol{X}}}{1 - R^2_{D\sim \boldsymbol{X}}} \tag{53}$$

$$= k_D \left(\frac{R^2_{D\sim X_j}}{1 - R^2_{D\sim \boldsymbol{X}}} \right). \tag{54}$$

Analogous result holds for $R^2_{Y\sim Z|\boldsymbol{X}}$. What remains is to re-express $R^2_{Y\sim Z|D,\boldsymbol{X}}$. Using the standard recursive definition of partial correlations, we know that

$$\left|R_{Y\sim Z|\boldsymbol{X},D}\right| = \frac{\left|R_{Y\sim Z|\boldsymbol{X}} - R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}\right|}{\sqrt{1 - R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}}}}. \tag{55}$$

The only two terms of the RHS including the confounder, $R_{Y\sim Z|\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$, have been re-expressed as a function of k_D and k_Y above. We now show how to determine the sign of the correlations, by considering the direction of the strengths of the confounder act towards hurting our preferred hypothesis.

Let us assume the confounder acts towards reducing the absolute value of the effect size. If the effect size is positive ($R_{Y\sim D|\boldsymbol{X}} > 0$), this means $R_{Y\sim Z|D,\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$ must have the same signs. Consider, first, $R_{Y\sim Z|\boldsymbol{X},D} < 0$ and $R_{D\sim Z|\boldsymbol{X}} < 0$. This implies $R_{Y\sim Z|\boldsymbol{X}} < 0$, which means we are reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$. Now consider $R_{Y\sim Z|D,\boldsymbol{X}} > 0$ and $R_{D\sim Z|\boldsymbol{X}} > 0$. This implies $R_{Y\sim Z|\boldsymbol{X}} > 0$, which, again, means we are reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$. If the effect size is negative ($R_{Y\sim D|\boldsymbol{X}} < 0$), this now would mean that $R_{Y\sim Z|D,\boldsymbol{X}}$ and $R_{D\sim Z|\boldsymbol{X}}$ must have the opposite signs, and applying the previous arguments, we reach the same conclusion that we will be reducing the absolute value of $R_{Y\sim Z|\boldsymbol{X}}$.

Therefore, considering that the confounder acts towards *reducing* the magnitude of the estimate towards zero, we have that,

$$\left|R_{Y\sim Z|\boldsymbol{X},D}\right| = \frac{|R_{Y\sim Z|\boldsymbol{X}}| - |R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}|}{\sqrt{1 - R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}}}}. \tag{56}$$

Extending the previous arguments to multiple covariates is straightforward, since these results hold for any subset of $\boldsymbol{X}$.

B.2 Comparing the partial R^2 of covariates with the partial R^2 of the confounder

Now imagine the researcher is willing to make a more elaborate type of claim. For instance, the researcher believes that omitting X_j increases the mean squared error of the full treatment regression more than omitting Z. This means that, $R^2_{D\sim \boldsymbol{X}_{-j}+Z} < R^2_{D\sim \boldsymbol{X}}$, where $\boldsymbol{X}_{-j}$ represents all variables in $\boldsymbol{X}$ except X_j. If we now subtract of both sides $R^2_{D\sim \boldsymbol{X}_{-j}}$ and further divide them by $1 - R^2_{D\sim \boldsymbol{X}_{-j}}$, this gives us,

$$\frac{R^2_{D\sim \boldsymbol{X}_{-j}+Z} - R^2_{D\sim \boldsymbol{X}_{-j}}}{1 - R^2_{D\sim \boldsymbol{X}_{-j}}} < \frac{R^2_{D\sim \boldsymbol{X}} - R^2_{D\sim \boldsymbol{X}_{-j}}}{1 - R^2_{D\sim \boldsymbol{X}_{-j}}}. \tag{57}$$

Which means that

$$R^2_{D\sim Z|\boldsymbol{X}_{-j}} < R^2_{D\sim X_j|\boldsymbol{X}_{-j}}. \tag{58}$$

That is, we can compare the strength of Z to X_j by assessing their relative contribution to the partial R^2 of the treatment regression given the remaining covariates. Generalizing this notion define,

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}. \tag{59}$$

Our goal now is to re-express $R^2_{D\sim Z|\boldsymbol{X}}$ in terms of k_D.

Bounding $R^2_{D\sim Z|\boldsymbol{X}}$

From equation 59 we have that $|R_{D\sim Z|\boldsymbol{X}_{-j}}| = \sqrt{k_D}|R_{D\sim X_j|\boldsymbol{X}_{-j}}|$. Also, the assumption that $Z \perp \boldsymbol{X}$ implies $R_{Z\sim X_j|\boldsymbol{X}_{-j}} = 0$. Combining these two results, and using the standard recursive definition of partial correlations, gives us

$$\begin{aligned}
\left|R_{D\sim Z|\boldsymbol{X}}\right| &= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}} - R_{D\sim X_j|\boldsymbol{X}_{-j}} R_{Z\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (60)\\
&= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}}}{\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (61)\\
&= \frac{\sqrt{k_D}\left|R_{D\sim X_j|\boldsymbol{X}_{-j}}\right|}{\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}} && (62)\\
&= \sqrt{k_D}\left|f_{D\sim X_j|\boldsymbol{X}-j}\right|. && (63)
\end{aligned}$$

Hence,

$$R^2_{D\sim Z|\boldsymbol{X}} = k_D \times f^2_{D\sim X_j|\boldsymbol{X}-j}. \tag{64}$$

Also, notice that, since $R^2_{D\sim Z|\boldsymbol{X}} \leq 1$ this, means k_D cannot vary freely but rather is bounded by

$$k_D \leq \frac{1}{f^2_{D\sim X_j|\boldsymbol{X}-j}}. \tag{65}$$

As an example, if a researcher has a covariate that currently explains 50% of the residual variance of the treatment assignment (implying $f^2_{D\sim X_j|\boldsymbol{X}-j} = 1$), Equation 65 reveals it is *impossible* to have an *orthogonal* unobserved confounder Z stronger than that covariate.

Using multiple covariates. Now let us generalize the previous bound to multiple covariates. Let this set of covariates be $\mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})} = \{X_1, \ldots, X_j\}$. We will denote the complement of this set $\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}$. Thus, k_D now is defined as

$$k_D := \frac{R^2_{D\sim Z|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}{R^2_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}. \tag{66}$$

Applying the recursive definition of partial correlation to, $R_{D\sim Z|\boldsymbol{X}}$, $R_{D\sim Z|\boldsymbol{X}_{-(1)}}$, $R_{D\sim Z|\boldsymbol{X}_{-(1,2)}}$, up to $R_{D\sim Z|\boldsymbol{X}_{-(1,\ldots,j)}}$, and recalling the orthogonality of Z with $\boldsymbol{X}$, we have that,

$$R_{D\sim Z|\boldsymbol{X}} = \frac{R_{D\sim Z|\boldsymbol{X}_{-(1,\ldots,j)}}}{\sqrt{1-R^2_{D\sim X_1|\boldsymbol{X}_{-(1)}}}\sqrt{1-R^2_{D\sim X_2|\boldsymbol{X}_{-(1,2)}}}\cdots\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-(1,\ldots,j)}}}}. \tag{67}$$

Since, $R^2_{D\sim Z|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}} = k_D R^2_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}$, we obtain,

$$\left|R_{D\sim Z|\boldsymbol{X}}\right| = \frac{\sqrt{k_D}\left|R_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}\right|}{\sqrt{1-R^2_{D\sim X_1|\boldsymbol{X}_{-(1)}}}\sqrt{1-R^2_{D\sim X_2|\boldsymbol{X}_{-(1,2)}}}\cdots\sqrt{1-R^2_{D\sim X_j|\boldsymbol{X}_{-(1,\ldots,j)}}}}. \tag{68}$$

We can simplify this further by noticing the denominator is simply $\sqrt{1-R^2_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}$

$$\left|R_{D\sim Z|\boldsymbol{X}}\right| = \frac{\sqrt{k_D}\left|R_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}\right|}{\sqrt{1-R^2_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}} = \sqrt{k_D}\left|f_{D\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}\right|. \tag{69}$$

Bounding $R_{Y\sim Z|D,\boldsymbol{X}}$

We have two ways of bounding $R_{Y\sim Z|D,\boldsymbol{X}}$, making comparisons conditional or not conditional on D.

Comparisons not conditioning on D. As in the previous derivation, define,

$$k_Y := \frac{R^2_{Y\sim Z|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}{R^2_{Y\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}. \tag{70}$$

That is, we are asking the researcher to compare the explanatory power of the confounder against the explanatory power of $\mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}$ with respect to the outcome, conditioning on the remaining covariates $\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}$ but *not conditioning* on the treatment. Using the same recursive argument as before, we obtain

$$\left|R_{Y\sim Z|\boldsymbol{X}}\right| = \sqrt{k_Y}\left|f_{Y\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}\right|. \tag{71}$$

We can now bound $R^2_{Y\sim Z|D,\boldsymbol{X}}$ by noting again that

$$R_{Y\sim Z|D,\boldsymbol{X}} = \frac{R_{Y\sim Z|\boldsymbol{X}} - R_{Y\sim D|\boldsymbol{X}} R_{D\sim Z|\boldsymbol{X}}}{\sqrt{1 - R^2_{Y\sim D|\boldsymbol{X}}}\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}}}}, \tag{72}$$

then using the same argument as in B.2.

Comparisons conditioning on D. Here we have that k_D is defined as before, but k_Y compares the explanatory power of the confounder against the explanatory power of a covariate X_j with respect to the outcome, conditioning on both the remaining covariates $\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}$ and the treatment, that is,

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|\boldsymbol{X}_{-j},D}}{R^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}}. \tag{73}$$

To bound $R^2_{Y\sim Z|D,\boldsymbol{X}}$, we first need to investigate $R_{Z\sim X_j|\boldsymbol{X}_{-j},D}$. Expanding the partial correlation gives us

$$\begin{aligned}
\left|R_{Z\sim X_j|\boldsymbol{X}_{-j},D}\right| &= \left|\frac{R_{Z\sim X_j|\boldsymbol{X}_{-j}} - R_{D\sim Z|\boldsymbol{X}_{-j}} R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}_{-j}}}\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (74)\\
&= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-j}} R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1 - R^2_{D\sim Z|\boldsymbol{X}_{-j}}}\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (75)\\
&= \left|\frac{\sqrt{k_D} R_{D\sim X_j|\boldsymbol{X}_{-j}} R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1 - k_D R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}\sqrt{1 - R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}\right| && (76)\\
&= \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|. && (77)
\end{aligned}$$

where, f_{K_D} is defined to be,

$$f_{K_D} := \frac{\sqrt{k_D} R_{D\sim X_j|\boldsymbol{X}_{-j}}}{\sqrt{1 - k_D R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}}. \tag{78}$$

Combining theses results and Equation 73 we can proceed to bound $R_{Y\sim Z|D,\boldsymbol{X}}$:

$$
\begin{aligned}
|R_{Y\sim Z|D,\boldsymbol{X}}| &= \left| \frac{R_{Y\sim Z|\boldsymbol{X}_{-j},D} - R_{Y\sim X_j|\boldsymbol{X}_{-j}D} R_{Z\sim X_j|\boldsymbol{X}_{-j},D}}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j},D}}} \right| && (79) \\
&\leq \frac{\left|R_{Y\sim Z|\boldsymbol{X}_{-j},D}\right| + \left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right| \left|R_{Z\sim X_j|\boldsymbol{X}_{-j},D}\right|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-R^2_{Z\sim X_j|\boldsymbol{X}_{-j},D}}} && (80) \\
&= \frac{\sqrt{k_Y}\left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right| + \left|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}\right| \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}\sqrt{1-f^2_{K_D} \times f^2_{D\sim X_j|\boldsymbol{X}-j}}} && (81) \\
&= \left(\frac{\sqrt{k_Y} + \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-f^2_{K_D} \times f^2_{D\sim X_j|\boldsymbol{X}-j}}}\right)\left(\frac{|R_{Y\sim X_j|\boldsymbol{X}_{-j}D}|}{\sqrt{1-R^2_{Y\sim X_j|\boldsymbol{X}_{-j}D}}}\right) && (82) \\
&= \eta \left|f_{Y\sim X_j|\boldsymbol{X}_{-j},D}\right|. && (83)
\end{aligned}
$$

Hence, we have that,

$$
R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim X_j|\boldsymbol{X}_{-j},D}, \tag{84}
$$

where $\eta = \frac{\sqrt{k_Y} + \left|f_{K_D} \times f_{D\sim X_j|\boldsymbol{X}-j}\right|}{\sqrt{1-f^2_{K_D} \times f^2_{D\sim X_j|\boldsymbol{X}-j}}}$. Note the bound is tight. Without further assumptions, we can create an unobserved confounder Z that makes the inequality step in 80 an equality. One can extend this to multiple covariates by iteratively applying the recursive definition of partial correlation.

C Some numerical examples of informal benchmarking

Here we show how the informal benchmarking practices proposed in Frank (2000); Frank and Min (2007) and in Carnegie et al. (2016b) could lead users to erroneous conclusions. Starting with Carnegie et al. (2016b), consider the simulation in the `R` code presented in the left hand side of Figure 5. Note the unobserved confounder Z is exactly like X in terms of its association with the treatment D and the outcome Y; moreover, we also have that $Z \perp X$. Finally, note that, by construction, the unobserved confounder Z (which is as strong as X) is sufficient to bring the effect estimate down to zero. The right hand side of Figure 5 shows the output of Carnegie et al. (2016b) companion software, the `R` package `treatSens` (Carnegie et al. 2016a).[28] Note it incorrectly claims that the effect estimate would be robust to a confounder as strong as X (benchmark shown in the red "x" mark).

Now moving to Frank (2000) and Frank and Min (2007), one would first compute the "impact threshold" of a confounding variable and compare this to the "observed" *impact* of the covariate X. In the same simulation of Figure 5, these calculation are shown in the last part of the code (using the `R` package `konfound`). One then obtains an impact threshold of 0.469 (considering statistical significance of 5%), which, when contrasted with the "observed impact" of X, $R_{Y\sim X} \times R_{D\sim X} = 0.314$,

[28] As of 14 October 2019, the `R` package was removed from `CRAN` for lack of maitainance; archived versions can still be found in `https://cran.r-project.org/web/packages/treatSens/index.html`.

```
# cleans workspace
rm(list = ls())

# set seed for reproducibility
set.seed(10)

# loads packages
library(treatSens)
library(konfound)

# simulates data
n <- 500
x <- rnorm(n)
z <- rnorm(n)
d <- x + z + rnorm(n)
y <- x + z + rnorm(n)

# Carnegie et al method
sense <- treatSens(y ~ d + x)
sensPlot(sense)

# Frank's method
model <- lm(y ~ d + x)

## computes impact threshold
konfound(model, tested_variable = "d",
                alpha = 0.05)

## "observed impact" of X
cor(x, d)*cor(x, y)
```

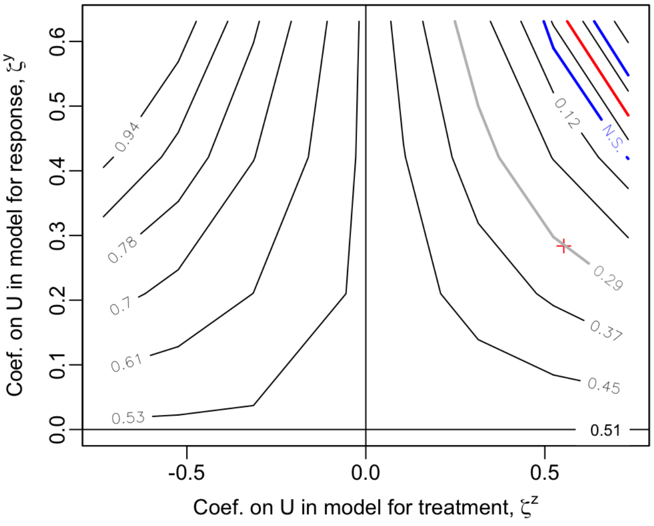


Figure 5: Examples of informal benchmarking.

Note: Code (left) and plot (right) for the incorrect informal benchmark bound produced from the methods of Carnegie, Harada and Hill (2006). Note the informal benchmark would lead one to *incorrectly* conclude that an unobserved confounder Z exactly like X would not be sufficient to explain away the estimated effect, when in fact it would (as shown in the red "x" mark). Code for Frank (2000) and Frank and Min (2007) is also shown in the left.

would lead an investigator to erroneously conclude that an unobserved confounder as strong as X would not be sufficient to explain away the effect estimate.

D Sensitivity tables

Bias Factor table

$R^2_{D\sim Z\mid\boldsymbol{X}}/R^2_{Y\sim Z\mid D,\boldsymbol{X}}$	**5%**	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**	**75%**	**80%**	**85%**	**90%**	**95%**	**99%**
5%	0.051	0.073	0.089	0.103	0.115	0.126	0.136	0.145	0.154	0.162	0.170	0.178	0.185	0.192	0.199	0.205	0.212	0.218	0.224	0.228
10%	0.075	0.105	0.129	0.149	0.167	0.183	0.197	0.211	0.224	0.236	0.247	0.258	0.269	0.279	0.289	0.298	0.307	0.316	0.325	0.332
15%	0.094	0.133	0.163	0.188	0.210	0.230	0.249	0.266	0.282	0.297	0.312	0.325	0.339	0.351	0.364	0.376	0.387	0.399	0.409	0.418
20%	0.112	0.158	0.194	0.224	0.250	0.274	0.296	0.316	0.335	0.354	0.371	0.387	0.403	0.418	0.433	0.447	0.461	0.474	0.487	0.497
25%	0.129	0.183	0.224	0.258	0.289	0.316	0.342	0.365	0.387	0.408	0.428	0.447	0.465	0.483	0.500	0.516	0.532	0.548	0.563	0.574
30%	0.146	0.207	0.254	0.293	0.327	0.359	0.387	0.414	0.439	0.463	0.486	0.507	0.528	0.548	0.567	0.586	0.604	0.621	0.638	0.651
35%	0.164	0.232	0.284	0.328	0.367	0.402	0.434	0.464	0.492	0.519	0.544	0.568	0.592	0.614	0.635	0.656	0.677	0.696	0.715	0.730
40%	0.183	0.258	0.316	0.365	0.408	0.447	0.483	0.516	0.548	0.577	0.606	0.632	0.658	0.683	0.707	0.730	0.753	0.775	0.796	0.812
45%	0.202	0.286	0.350	0.405	0.452	0.495	0.535	0.572	0.607	0.640	0.671	0.701	0.729	0.757	0.783	0.809	0.834	0.858	0.882	0.900
50%	0.224	0.316	0.387	0.447	0.500	0.548	0.592	0.632	0.671	0.707	0.742	0.775	0.806	0.837	0.866	0.894	0.922	0.949	0.975	0.995
55%	0.247	0.350	0.428	0.494	0.553	0.606	0.654	0.699	0.742	0.782	0.820	0.856	0.891	0.925	0.957	0.989	1.019	1.049	1.078	1.100
60%	0.274	0.387	0.474	0.548	0.612	0.671	0.725	0.775	0.822	0.866	0.908	0.949	0.987	1.025	1.061	1.095	1.129	1.162	1.194	1.219
65%	0.305	0.431	0.528	0.609	0.681	0.746	0.806	0.862	0.914	0.964	1.011	1.056	1.099	1.140	1.180	1.219	1.256	1.293	1.328	1.356
70%	0.342	0.483	0.592	0.683	0.764	0.837	0.904	0.966	1.025	1.080	1.133	1.183	1.232	1.278	1.323	1.366	1.408	1.449	1.489	1.520
75%	0.387	0.548	0.671	0.775	0.866	0.949	1.025	1.095	1.162	1.225	1.285	1.342	1.396	1.449	1.500	1.549	1.597	1.643	1.688	1.723
80%	0.447	0.632	0.775	0.894	1.000	1.095	1.183	1.265	1.342	1.414	1.483	1.549	1.612	1.673	1.732	1.789	1.844	1.897	1.949	1.990
85%	0.532	0.753	0.922	1.065	1.190	1.304	1.408	1.506	1.597	1.683	1.765	1.844	1.919	1.992	2.062	2.129	2.195	2.258	2.320	2.369
90%	0.671	0.949	1.162	1.342	1.500	1.643	1.775	1.897	2.012	2.121	2.225	2.324	2.419	2.510	2.598	2.683	2.766	2.846	2.924	2.985
95%	0.975	1.378	1.688	1.949	2.179	2.387	2.579	2.757	2.924	3.082	3.233	3.376	3.514	3.647	3.775	3.899	4.019	4.135	4.249	4.337
99%	2.225	3.146	3.854	4.450	4.975	5.450	5.886	6.293	6.675	7.036	7.379	7.707	8.022	8.325	8.617	8.899	9.173	9.439	9.698	9.900

Table 2: Bias factor table

Note: To use the bias factor table, first compute the absolute value of the partial (Cohen's) f of the treatment with the outcome. The partial f can be easily obtained in most regression tables by dividing the coefficient's t-value by the square-root of the degrees of freedom, that is, $f = \frac{t}{\sqrt{\text{df}}}$. To bring down the estimated effect to zero, the bias factor has to be greater than $|f|$. In our running example, $|f| \approx 0.15$. Looking at the table, we see that the coefficient of *Directly Harmed* would be robust to a confounder with, for instance, $R^2_{D\sim Z\mid\boldsymbol{X}} = 5\%$ and $R^2_{Y\sim Z\mid D,\boldsymbol{X}} = 40\%$, but would not be robust to a confounder with $R^2_{D\sim Z\mid\boldsymbol{X}} = 40\%$ and $R^2_{Y\sim Z\mid D,\boldsymbol{X}} = 5\%$. To assess the robustness of the coefficient to any change of $(100 \times q)\%$, just multiply $|f|$ by q. Any confounder with a bias factor less than $q|f|$ cannot cause a change of $(100 \times q)\%$ in the regression coefficient.