sensemakr: Sensitivity Analysis Tools for OLS in R and Stata

Carlos Cinelli **cinelli@uw.edu**
Department of Statistics
University of Washington
Seattle, WA, USA

Jeremy Ferwerda **jeremy.a.ferwerda@dartmouth.edu**
Department of Government
Dartmouth College
Hanover, NH, USA

Chad Hazlett **chazlett@ucla.edu**
Department of Statistics
University of California, Los Angeles
Los Angeles, CA, USA

Abstract

This tutorial introduces the package sensemakr for R and Stata, which implements a suite of sensitivity analysis tools for regression models developed in Cinelli and Hazlett (2020, 2022). Given a regression model, sensemakr can compute sensitivity statistics for routine reporting, such as the *robustness value*, which describes the minimum strength that unobserved confounders need to have to overturn a research conclusion. The package also provides plotting tools that visually demonstrate the sensitivity of point estimates and t-values to hypothetical confounders. Finally, sensemakr implements formal bounds on sensitivity parameters by means of comparison with the explanatory power of observed variables. All these tools are based on the familiar "omitted variable bias" framework, do not require assumptions regarding the functional form of the treatment assignment mechanism nor the distribution of the unobserved confounders, and naturally handle multiple, non-linear confounders. With sensemakr, users can transparently report the sensitivity of their causal inferences to unobserved confounding, thereby enabling a more precise, quantitative debate as to what can be concluded from imperfect observational studies.

Keywords: causal inference, sensitivity analysis, omitted variable bias, robustness value, R, Stata, bounds.

1. Introduction

Across disciplines, investigators face the perennial challenge of making and defending causal claims using observational data. The most common identification strategy in these circumstances is to adjust for a set of *observed* covariates deemed sufficient to control for confounding, with linear regression remaining among the most popular statistical method for making such adjustments. Researchers who argue that a regression coefficient unbiasedly reflects a causal relationship must also be able to argue that there are no *unobserved* confounders—a difficult or impossible assumption to defend in most applied settings.[1] What value can we draw from these studies, knowing that this ideal condition is likely to fail? Fortunately, the assumption of *zero* unobserved confounding need not hold precisely for an observational study to remain substantively informative. In these cases, sensitivity analyses play a useful role by allowing researchers to quantify how strong unobserved confounding needs to be in order to substantially change a research conclusion, and by aiding in determining whether confounding of such strength is plausible.

Although numerous methods for sensitivity analyses have been proposed, these tools are still under-utilized.[2] As argued in Cinelli and Hazlett (2020), several reasons may contribute to the low adoption of these methods. First, many of these methods impose complicated and strong assumptions regarding the nature of the confounder, which many users cannot or are not willing to defend. Second, while users routinely report regression tables or coefficient plots, until recently investigators have lacked "standard" quantities that can easily and correctly summarize the robustness of a regression coefficient to unobserved confounding. Finally, connecting the results of a formal sensitivity analysis to a cogent argument about what types of confounders may exist in one's research project remains difficult, particularly when there are no compelling arguments as to why the treatment assignment should be approximately "ignorable," "exogeneous," or "as-if random."

This tutorial introduces the `R` and `Stata` package `sensemakr` (Cinelli et al., 2020a,b), which implements a suite of sensitivity analysis tools proposed in Cinelli and Hazlett (2020) to address these challenges, in the context of regression adjustment using ordinary least squares (OLS). Within the familiar regression framework and without the need for additional assumptions (beyond those that justified using OLS in the first place), `sensemakr` enables analysts to easily answer a variety of common sensitivity questions, such as:

- How strong would an unobserved confounder (or a group of confounders) have to be to change a research conclusion?
- In a worst-case scenario, how robust are the results to all unobserved confounders acting together, possibly non-linearly?
- How strong would confounding need to be, relative to the strength of observed covariates, to change the answer by a certain amount?

Specifically, given a full regression model, or simply standard statistics found in conventional regression tables, `sensemakr` is able to: (i) compute sensitivity statistics for routine reporting, such as the *robustness value* describing the minimum strength that unobserved confounders would need to have to overturn the research conclusions; (ii) provide graphical tools that enable users to visually explore the implications of unobserved confounding,

[1] This condition is also known as "selection on observables," "conditional igorability," "conditional exogeneity," "conditional exchangeability," or "backdoor admissibility" (Angrist and Pischke, 2008; Pearl, 2009; Imbens and Rubin, 2015; Hernán and Robins, 2020).

[2] Dating back to at least Cornfield et al. (1959), a partial list of sensitivity analysis proposals includes Rosenbaum and Rubin (1983); 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. (2016); Dorie et al. (2016); Middleton et al. (2016); Oster (2017); Cinelli et al. (2019); AlexanderM. Franks and Feller (2020).

such as contour plots showing adjusted point estimates and t-values under confounding of various strengths, as well as plots showing adjusted estimates under extreme (pessimistic) scenarios; and (iii) place formal bounds on the *maximum* strength of confounding, based on plausibility judgments regarding how unobserved confounders compare with observed variables. These tools do not require additional assumptions regarding the functional form of the treatment assignment mechanism nor on the distribution of the unobserved confounders, and naturally handle multiple confounders, possibly acting non-linearly.

In what follows, Section 2 briefly reviews the omitted variable bias framework for sensitivity analysis developed in Cinelli and Hazlett (2020), which provides the theoretical foundations for the tools in `sensemakr`. Next, Section 3 describes the basic functionality and provides a practical introduction to sensitivity analysis using `sensemakr` for `R`. Section 4 describes advanced usage of the `R` package, and shows how to leverage individual functions for customized sensitivity analyses. Finally, Section 5 describes `sensemakr` for `Stata`, and Section 6 concludes with a brief discussion of what sensitivity analysis can and cannot do in practice.

2. Sensitivity analysis in an omitted variable bias framework

In this section, we briefly review the omitted variable bias (OVB) framework for sensitivity analysis presented in Cinelli and Hazlett (2020). This method builds on a scale-free reparameterization of the OVB formula in terms of partial R^2 values, which allows us to: (i) assess the sensitivity of point estimates, t-values, and confidence intervals under the same conceptual framework; (ii) easily assess the sensitivity of multiple confounders acting together, possibly non-linearly; (iii) exploit knowledge of the relative strength of variables to posit plausible bounds on unobserved confounding; and (iv) construct a set of summary sensitivity statistics suitable for routine reporting.

2.1 The OVB framework

The starting point of our analysis is a "full" linear regression model of an outcome Y on a treatment D, controlling for a set of covariates given by *both* $\mathbf{X}$ and Z,

$$Y = \hat{\tau} D + \mathbf{X}\hat{\beta} + \hat{\gamma} Z + \hat{\varepsilon}_{\text{full}}, \tag{1}$$

where Y is an $(n \times 1)$ vector containing the outcome of interest for each of the n observations and D is an $(n \times 1)$ treatment variable (which may be continuous or binary); $\mathbf{X}$ is an $(n \times p)$ matrix of *observed* covariates including the constant; and Z is a single $(n \times 1)$ *unobserved* covariate (we discuss how to extend results for a multivariate Z below).

Equation 1 is the regression model that the investigator *wished* she had run to obtain a valid causal estimate of the effect of D on Y. Nevertheless, Z is unobserved. Therefore, the feasible regression the investigator is able to estimate is the "restricted“ model *omitting* Z, that is,

$$Y = \hat{\tau}_{\text{res}} D + \mathbf{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{\text{res}}. \tag{2}$$

Given the discrepancy of what we wish to know and what we actually have, the main question we would like to answer is: how do the observed point estimate and standard error of the restricted regression, $\hat{\tau}_{\text{res}}$ and $\widehat{\text{se}}(\hat{\tau}_{\text{res}})$, compare to the desired point estimate and standard error of the full regression, $\hat{\tau}$ and $\widehat{\text{se}}(\hat{\tau})$?

2.1.1 OVB with the partial R^2 parameterization

Define as $\widehat{\text{bias}}$ the difference between the full and restricted estimates, $\widehat{\text{bias}} := \hat{\tau}_{\text{res}} - \hat{\tau}$. Now let (i) $R^2_{D\sim Z|\mathbf{X}}$ denote the (sample) share of residual variance of the *treatment* D explained by the omitted variable Z, after accounting for the remaining covariates $\mathbf{X}$; and, (ii) $R^2_{Y\sim Z|D,\mathbf{X}}$ denote the share of residual variance of the *outcome* Y explained by the omitted variable Z, after accounting for $\mathbf{X}$ and D. Cinelli and Hazlett (2020) have shown that these quantities are sufficient for determining the bias, adjusted estimate, and adjusted standard errors of the full regression of Equation 1.

More precisely, the bias can be written as,

$$|\widehat{\text{bias}}| = \sqrt{\frac{R^2_{Y\sim Z|D,\mathbf{X}}\ R^2_{D\sim Z|\mathbf{X}}}{1-R^2_{D\sim Z|\mathbf{X}}}}\left(\frac{\widehat{\sigma}_{y.dx}}{\widehat{\sigma}_{d.x}}\right) = \sqrt{\frac{R^2_{Y\sim Z|D,\mathbf{X}}\ R^2_{D\sim Z|\mathbf{X}}}{1-R^2_{D\sim Z|\mathbf{X}}}} \times \widehat{\text{se}}(\hat{\tau}_{\text{res}}) \times \sqrt{\text{df}}. \tag{3}$$

Here df stands for the degrees of freedom of the restricted regression actually run, and $\widehat{\sigma}_{y.dx}$, $\widehat{\sigma}_{d.x}$ denote the residual standard deviations of the outcome and treatment regressions actually run. For *computational convenience*, it is possible to rewrite the bias formula in terms of the classical standard error estimate, which is usually reported in regression tables.

Moreover, the classical estimated standard error of $\hat{\tau}$ can be recovered with,

$$\widehat{\text{se}}(\hat{\tau}) = \sqrt{\frac{1-R^2_{Y\sim Z|D,\mathbf{X}}}{1-R^2_{D\sim Z|\mathbf{X}}}} \times \widehat{\text{se}}(\hat{\tau}_{\text{res}}) \times \sqrt{\left(\frac{\text{df}}{\text{df}-1}\right)}. \tag{4}$$

Given hypothetical values of $R^2_{D\sim Z|\mathbf{X}}$ and $R^2_{Y\sim Z|D,\mathbf{X}}$, Equations 3 and 4 allow investigators to examine the sensitivity of point estimates and standard-errors (and consequently t-values, confidence intervals or p-values) to the inclusion of any omitted variable Z with such strengths. Conversely, given a critical threshold deemed to be problematic, one can find the strength of confounders capable of bringing about a bias reducing the adjusted effect to that threshold. Another useful property of the OVB formula with the partial R^2 parameterization is that the effect of $R^2_{Y\sim Z|D,\mathbf{X}}$ on the bias is bounded. This allows investigators to contemplate extreme sensitivity scenarios, in which the parameter $R^2_{Y\sim Z|D,\mathbf{X}}$ is set to 1 (or another conservative value), and see what happens as $R^2_{D\sim Z|\mathbf{X}}$ varies.

2.2 Sensitivity statistics for routine reporting

The previous formulas can be used to assess the sensitivity of an estimate to confounders with any hypothesized strength. However, making sensitivity analyses standard practice benefits from simple and interpretable sensitivity statistics that can quickly summarize the robustness of a study result to unobserved confounding. With this in mind, Cinelli and Hazlett (2020) propose two main sensitivity statistics for routine reporting: (i) the (observed) partial R^2 of the treatment with the outcome, $R^2_{Y\sim D|\mathbf{X}}$; and, (ii) the *robustness value*, $RV_{q,\alpha}$. These statistics serve two main purposes:

1. They can be easily displayed alongside other summary statistics in regression tables, making sensitivity analysis to unobserved confounding simple, accessible, and standardized;

2. They can be easily computed from quantities found in a regression table, thereby enabling readers and reviewers to assess the sensitivity of results they see in print, even if the original authors did not perform sensitivity analyses.

2.2.1 The partial $R^2$2 of the treatment with the outcome

In addition to quantifying how much variation of the outcome is explained by the treatment, the partial R^2 of the treatment with the outcome also conveys how robust the point estimate is to unobserved confounding in an "extreme scenario." Specifically, suppose the unobserved confounder Z explains *all* residual variance of the outcome, that is, $R_{Y\sim Z|D,\mathbf{X}} = 1$. For this confounder to bring the point estimate to zero, it must explain *at least* as much residual variation of the treatment as the residual variation of the outcome that the treatment currently explains. Put differently, if $R_{Y\sim Z|D,\mathbf{X}} = 1$, then we must have that $R^2_{D\sim Z|\mathbf{X}} \geq R^2_{Y\sim D|\mathbf{X}}$, otherwise this confounder cannot logically account for all the observed association between the treatment and the outcome (Cinelli and Hazlett, 2020).

2.2.2 The Robustness Value

The second sensitivity statistic proposed in Cinelli and Hazlett (2020) is the *robustness value*. The robustness value $RV_{q,\alpha}$ quantifies the *minimal* strength of association that the confounder needs to have, *both* with the treatment and with the outcome, so that a confidence interval of level $1-\alpha$ includes a change of q% of the current estimated value.

Let $f_q := q|f_{Y\sim D|\mathbf{X}}|$, where $|f_{Y\sim D|\mathbf{X}}|$ is the partial *Cohen's f* of the treatment with the outcome multiplied by the percentage reduction q deemed to be problematic.[3] Also, let $|t^*_{\alpha,\text{df}-1}|$ denote the t-value threshold for a t-test with significance level of α and df-1 degrees of freedom, and define $f^*_{\alpha,\text{df}-1} := |t^*_{\alpha,\text{df}-1}|/\sqrt{\text{df}-1}$. Finally, construct $f_{q,\alpha}$, which "deducts" from $f_{Y\sim D|\mathbf{X}}$ both the proportion of reduction q of the point estimate and the boundary below which statistical significance is lost at the level of α. That is, $f_{q,\alpha} := f_q - f^*_{\alpha,\text{df}-1}$. We then have that $RV_{q,\alpha}$ is given by (Cinelli and Hazlett, 2020, 2022),

$$RV_{q,\alpha} = \begin{cases} 0, & \text{if } \ f_{q,\alpha} < 0 \\ \frac{1}{2}\left(\sqrt{f^4_{q,\alpha} + 4f^2_{q,\alpha}} - f^2_{q,\alpha}\right), & \text{if } \ f^*_{\alpha,\text{df}-1} \leq f_q < 1/f^*_{\alpha,\text{df}-1} \\ \dfrac{f_q^2 - f^{*2}_{\alpha,\text{df}-1}}{1+f_q^2}, & \text{otherwise.} \end{cases} \tag{5}$$

Any confounder that explains $RV_{q,\alpha}$% of the residual variance of *both* the treatment and of 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, sufficiently strong to make the adjusted $1-\alpha$ confidence interval include $(1-q)|\hat{\tau}_{\text{res}}|$). Likewise, a confounder with associations

[3] The partial Cohen's f^2 can be written as $f^2_{Y\sim D|\mathbf{X}} = R^2_{Y\sim D|\mathbf{X}}/(1 - R^2_{Y\sim D|\mathbf{X}})$

lower than $RV_{q,\alpha}$ is not capable of overturning the conclusion of such a test. Setting $\alpha = 1$ returns the robustness value for the point estimate. Further details on how to interpret the robustness value in practice are given in the next sections.

2.3 Bounds on the strength of confounding using observed covariates

Consider a confounder orthogonal to the observed covariates, ie., $Z \perp \mathbf{X}$, or, equivalently, consider only the part of Z not linearly explained by $\mathbf{X}$. Now denote by X_j a specific covariate of the set $\mathbf{X}$ and define

$$k_D := \frac{R^2_{D\sim Z|\mathbf{X}_{-j}}}{R^2_{D\sim X_j|\mathbf{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|\mathbf{X}_{-j},D}}{R^2_{Y\sim X_j|\mathbf{X}_{-j},D}}. \tag{6}$$

where $\mathbf{X}_{-j}$ represents the vector of covariates $\mathbf{X}$ excluding X_j. That is, the terms k_D and k_Y represent how strong the confounder Z is relative to observed covariate X_j, where "strength" is measured by how much residual variation they explain of the treatment (for k_D) and of the outcome (for k_Y). Given k_D and k_Y, we can rewrite the strength of the confounders as (Cinelli and Hazlett, 2020),

$$R^2_{D\sim Z|\mathbf{X}} = k_D f^2_{D\sim X_j|\mathbf{X}_{-j}}, \qquad R^2_{Y\sim Z|D,\mathbf{X}} \leq \eta^2 f^2_{Y\sim X_j|\mathbf{X}_{-j},D}, \tag{7}$$

where η is a scalar which depends on k_Y, k_D and $R^2_{D\sim X_j|\mathbf{X}_{-j}}$.

These equations allow the investigator to assess the maximum bias that a hypothetical confounder at most "k times" as strong as a particular covariate X_j could cause. This can be used to explore the relative strength of confounding necessary for bias to have changed the research conclusion. Furthermore, when the researcher has domain knowledge to argue that a certain covariate X_j is particularly important in explaining treatment or outcome variation, and that omitted variables cannot explain as much residual variance of D or Y as that observed covariate, these results can be used to set plausible bounds in the total amount of confounding. The same inequalities hold if one uses a group of variables for benchmarking, by simply replacing the individual partial R^2 with the group partial R^2 of those variables (we provide a novel derivation for this result in the appendix). Finally, we note that an intuitive but informal benchmarking approach involves directly plugging in $R^2_{D\sim Z|\mathbf{X}} = R^2_{D\sim X_j|\mathbf{X}_{-j}}$ and $R^2_{Y\sim Z|D,\mathbf{X}} = R^2_{Y\sim X_j|\mathbf{X}_{-j},D}$ instead of using (7). However, this method can significantly underestimate the bias caused by a Z as strong as X_j. Section 4.1.3 discusses this issue further and provides a numerical example.

2.4 Multiple or non-linear confounders

Suppose that, instead of a single unobserved confounder Z, there are *multiple* unobserved confounders $\mathbf{Z} = [Z_1, Z_2, \ldots, Z_k]$. In this case, the regression the investigator wished she had run becomes:

$$Y = \hat{\tau} D + \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma} + \hat{\varepsilon}_{\text{full}}. \tag{8}$$

As Cinelli and Hazlett (2020) show, the previous results considering a single unobserved confounder are in fact *conservative* when considering the impact of multiple confounders, barring an adjustment in the degrees of freedom of Equation 4. Moreover, since the vector $\mathbf{Z}$ is arbitrary, this can also accommodate non-linear confounders or even misspecification of the functional form of the observed covariates $\mathbf{X}$. In other words, to assess the maximum bias that multiple, non-linear confounders could cause in our current estimates, it suffices to think in terms of the *maximum explanatory power* that $\mathbf{Z}$ could have in the treatment and outcome regressions, as parameterized by $R^2_{D\sim\mathbf{Z}|\mathbf{X}}$ and $R^2_{Y\sim\mathbf{Z}|D,\mathbf{X}}$. To put it in another way, it suffices to examine the impact of a single unmeasured confounder with that postulated strength, as this scenario represents the maximum bias that multiple confounders of equal strength could induce.

2.5 "Non-standard" standard errors

The previous results provide exact algebraic identities of how point estimates and *classical* (homoskedastic) standard errors change due to the inclusion of Z. Researchers, however, may also perform statistical inference using "non-standard" standard errors, such as using robust standard errors or the nonparametric bootstrap.

Robust standard errors can be computed using the influence function of the bounds, as derived in Chernozhukov et al. (2022). Inference using the nonparametric bootstrap is also straightforward. It suffices to: (1) resample the data with replacement, and (2) compute the bias adjusted estimate, i.e., $\hat{\tau}_{\pm} := \hat{\tau}_r \pm |\widehat{\text{bias}}|$, for each bootstrap resample. Confidence intervals for $\tau_{\pm}$ can then be constructed by using the desired percentile of the bootstrapped samples (or, computing bootstrap standard errors and using a normal approximation). Simple analytical expressions for $RV_{q,\alpha}$, such as Equation 5 for the classical case, are unavailable when using robust standard errors or the bootstrap, thus $RV_{q,\alpha}$ must be evaluated numerically.

For simplicity of exposition, in the main text of this tutorial we focus on the classical standard error case.[4] In the appendix we showcase the use of "non-standard" standard errors by replicating the running example using both the regular nonparametric bootstrap and the cluster bootstrap.

3. `sensemakr` for `R`: Basic functionality

In this section we illustrate the basic functionality of `sensemakr` for `R`. Given that sensitivity analysis requires contextual knowledge to be properly interpreted, we illustrate these tools with a real example. We use `sensemakr` to reproduce all results found in Section 5 of Cinelli and Hazlett (2020), which estimates the effects of exposure to violence on attitudes towards peace, in Darfur, Sudan. Further details about this application and the data can be found in Hazlett (2019).

3.1 Violence in Darfur: data and research question

In 2003 and 2004, the Darfurian government orchestrated a horrific campaign of violence against civilians, killing an estimated two hundred thousand people. This application asks

[4] In large samples, the bias term dominates sampling uncertainty, thus making the choice of standard errors a second-order concern relative to the magnitude of omitted variable bias. This happens because standard errors decrease at a $\sqrt{n}$ rate, while the bias remains constant, regardless of sample size.

whether, on average, being directly injured or maimed in this episode made individuals more likely to feel "vengeful" and unwilling to make peace with those who perpetrated this violence. Or, might those who directly suffered such violence be motivated to see it end, supporting calls for peace?

The `sensemakr` package provides the data required for this example based on a survey among Darfurian refugees in eastern Chad (Hazlett, 2019). The data were constructed from a survey conducted between April and June of 2009 by the "Darfurian Voices" team with support of the US Department of State, with the purpose of representing refugee voices in the ongoing political processes. The full survey, initially including 1872 civilians, was representative of adult refugees (eighteen years or older) from Darfur living in the twelve Darfurian refugee camps in eastern Chad at the time of sampling. Due to the identification approach taken here (see below), we must restrict analysis to the 1276 observations representing civilians who reported being present in their villages in Darfur during the time of village attack, and thus subject to possibly being injured.

The "treatment" variable of interest is `directlyharmed`, which indicates whether the individual was physically injured or maimed during the attack on her or his village in Darfur. The main outcome of interest is `peacefactor`, an index measuring pro-peace attitudes. Other covariates in the data include: `village` (a factor variable indicating the original village of the respondent), `female` (a binary indicator of gender), `age`, `herder_dar` (whether they were a herder in Darfur), `farmer_dar` (whether they were a farmer in Darfur), and `past_voted` (whether they report having voted in an earlier election, prior to the conflict). For further details, see `?darfur`.

To get started we first need to install the package. From within `R`, the `sensemakr` package can be installed from the Comprehensive `R` Archive Network (CRAN).

```
install.packages("sensemakr")
```

After loading the package, the data can be loaded with the command `data("darfur")`.

```
library(sensemakr)
data("darfur")
```

Hazlett (2019) argues that the purpose of these attacks was to punish civilians from ethnic groups presumed to support the opposition and to kill or drive these groups out so as to reduce this support. Violence against civilians included aerial bombardments by the government as well as assaults by the *Janjaweed*, a pro-government militia. For this example, suppose a researcher argues that, while some villages were more or less intensively attacked, *within village* violence was largely indiscriminate. The bombings were crude, could not be finely targeted below the level of village, and the strategic purpose of the attacks was not kill or capture specific individuals. Similarly, the *Janjaweed* had no reason to target certain individuals rather than others, and no information with which to do so, with one major exception—women were targeted and often subjected to sexual violence.

Supported by these considerations, this researcher may argue that adjusting for `village` and `female` is sufficient for control of confounding, and run the following linear regression model (in which other pre-treatment covariates, although not necessary for identification, are also included):

	Dependent variable:
	peacefactor
directlyharmed	0.097*** (0.023)
female	−0.232*** (0.024)
Observations	1,276
R^2	0.512
Residual Std. Error	0.310 (df = 783)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1: OLS results for `darfur.model`. To conserve space, only the results for `directlyharmed` and `female` are shown.

```
darfur.model <- lm(peacefactor ~ directlyharmed  + village +  female +
                                 age + farmer_dar + herder_dar +
                                 pastvoted + hhsize_darfur,
                   data = darfur)
```

This regression model results in the estimates shown in Table 1. According to this model, those who were directly harmed in violence were on average more "pro-peace," not less.

3.1.1 The threat of unobserved confounders

The previous estimate requires the assumption of *no unobserved confounders* for unbiasedness. While supported by the claim that there is no targeting of violence *within village and gender strata*, other investigators may challenge this account. For example, although the bombing was crude, perhaps bombs were still more likely to hit the center of the village, and those in the center were also likely to hold different attitudes towards peace. Or, it could be the case that the *Janjaweed* observed signals that indicate individual characteristics such as wealth, and targeted using this information. Or perhaps an individual's (prior) political attitudes could have led them to take actions that exposed them to greater risk during the attack. To complicate things, all these factors could interact with each other or otherwise have other non-linear effects.

These concerns suggest that, instead of the previous linear model (`darfur.model`), we should have run the ideal, but infeasible model:

```
darfur.complete.model <- lm(peacefactor ~ directlyharmed  + village +
                              female + age + farmer_dar + herder_dar +
                              pastvoted + hhsize_darfur +
                              center*wealth*political_attitudes,
                            data = darfur)
```

Where center*wealth*political_attitudes indicates *fully interacted* terms for these three variables. However trying to fit the model darfur.complete.model will result in error, as none of the variables center, wealth or political_attitudes were measured.

Given an assumption on how strongly omitted variables relate to the treatment and the outcome, how would including them have changed our inferences regarding the coefficient of directlyharmed? Or, what is the *minimal strength* that these unobserved confounders (or *all* remaining unobserved confounders) need to have to change our previous conclusions? Additionally, how can we leverage our contextual knowledge about the attacks to judge how plausible such confounders are? For instance, given the limited opportunities for targeting and the special role of gender in this case, if we assumed that unobserved confounding cannot explain more than female, what would this imply about the maximum possible strength of confounding? We show next how to use sensemakr to answer each of these questions.

3.2 Violence in Darfur: sensitivity analysis

The main function in sensemakr for R is sensemakr(). This function performs the most commonly required sensitivity analyses and returns an object of class sensemakr, which can then be further explored with the print, summary and plot methods (see details in ?print.sensemakr and ?plot.sensemakr). We begin the analysis by applying sensemakr() to the original regression model, darfur.model.

```
darfur.sensitivity <- sensemakr(model = darfur.model,
                                treatment = "directlyharmed",
                                benchmark_covariates = "female",
                                kd = 1:3,
                                ky = 1:3,
                                q = 1,
                                alpha = 0.05,
                                reduce = TRUE)
```

The arguments of this call are:

- **model**: the lm object with the outcome regression. In our case, darfur.model.
- **treatment**: the name of the treatment variable. In our case, "directlyharmed".
- **benchmark_covariates**: the names of covariates that will be used to bound the plausible strength of the unobserved confounders. Here, we put "female", which one could argue to be among the main determinants of exposure to violence *within village*. It was also found to be among the strongest determinants of attitudes towards peace empirically. Variables considered as separate benchmarks can be passed as a single character vector; variables that should be treated jointly as a group for benchmarks should be passed as named list of character vectors.
- **kd** and **ky**: these arguments parameterize how many times stronger the confounder is related to the treatment (kd) and to the outcome (ky) in comparison to the observed benchmark covariate ("female"). In our example, setting kd = 1:3 and ky = 1:3

means we want to investigate the maximum strength of a confounder once, twice, or three times as strong as female (in explaining treatment and outcome variation). When both `kd` and `ky` are vectors, as in the example, the vectors are evaluated considering each coordinate pair in sequence. For example, setting `kd = 1:3` and `ky = 1:3` is equivalent to evaluating the pairs (`kd=1, ky=1`), (`kd=2, ky=2`) and (`kd=3, ky=3`). If only `kd` is given, `ky` will be set equal to it by default.

- **q**: this allows the user to specify what fraction of the effect estimate would have to be explained away to be problematic. Setting `q = 1` means that a reduction of 100% of the current effect estimate (i.e. a true effect of zero) would be deemed problematic. The default is `q = 1`.
- **alpha**: significance level of interest for making statistical inferences. The default is `alpha = 0.05`.
- **reduce**: should we consider confounders acting towards *increasing* or *reducing* the absolute value of the estimate? The default is `reduce = TRUE`, which means we are considering confounders that pull the estimate towards (or through) zero. Setting `reduce = FALSE` will consider confounders that pull the estimate *away* from zero.

Using the default arguments, one can simplify the previous call to

```
darfur.sensitivity <- sensemakr(model = darfur.model,
                                treatment = "directlyharmed",
                                benchmark_covariates = "female",
                                kd = 1:3)
```

After running `sensemakr()`, we can explore the sensitivity analysis results. We note that the function `sensemakr()` also has `formula` and `numeric` methods. See `?sensemakr` for details.

3.2.1 Sensitivity statistics for routine reporting

The print method for `sensemakr` provides the original (observed) estimate along with three summary sensitivity statistics suited for routine reporting: (1) the partial R^2 of the treatment with the outcome; (2) the robustness value (RV) required to reduce the estimate entirely to zero (i.e. $q = 1$); and, (3) the RV beyond which the estimate would no longer be statistically distinguishable from zero at the 5% level ($q = 1$, $\alpha = 0.05$).

```
 print(darfur.sensitivity)

Sensitivity Analysis to Unobserved Confounding

Model Formula: peacefactor ~ directlyharmed + village + female + age +
    farmer_dar + herder_dar + pastvoted + hhsize_darfur

Observed Estimates of ' directlyharmed ':
  Coef. estimate: 0.097
  Standard Error: 0.023
  t-value: 4.18

Sensitivity Statistics:
  Partial R2 of treatment with outcome: 0.022
  Robustness Value, q = 1 : 0.139
  Robustness Value, q = 1 alpha = 0.05 : 0.076

For more information, check summary.
```

The package also provides a function that creates a latex or html table with these results, as shown in Table 2 (for the html table, simply change the argument to `format = "html"`).

```
ovb_minimal_reporting(darfur.sensitivity, format = "latex")
```

Outcome: *peacefactor*

Treatment:	Est.	S.E.	t-value	$R^2_{Y\sim D\|\mathbf{X}}$	$RV_{q=1}$	$RV_{q=1,\alpha=0.05}$
directlyharmed	0.097	0.023	4.184	2.2%	13.9%	7.6%
df = 783		*Bound (1x female)*: $R^2_{Y\sim Z\|\mathbf{X},D} = 12.5\%$, $R^2_{D\sim Z\|\mathbf{X}} = 0.9\%$				

Table 2: Minimal sensitivity analysis reporting.

Together these three sensitivity statistics provide the ingredients for a standard reporting template proposed in Cinelli and Hazlett (2020). More precisely:

- The robustness value for bringing the point estimate of `directlyharmed` exactly to zero ($RV_{q=1}$) is 13.9%. This means that unobserved confounders that explain 13.9% of the residual variance *both* of the treatment and of the outcome are sufficiently strong to explain away all the observed effect. On the other hand, unobserved confounders that *do not* explain at least 13.9% of the residual variance *both* of the treatment and of the outcome are not sufficiently strong to do so.

- The robustness value for testing the hypothesis that the coefficient of `directlyharmed` is zero ($RV_{q=1,\alpha=0.05}$) falls to 7.6%. This means that unobserved confounders that explain 7.6% of the residual variance *both* of the treatment and of the outcome are sufficiently strong to bring the lower bound of the confidence interval to zero (at the chosen significance level of 5%). On the other hand, unobserved confounders that *do not* explain at least 7.6% of the residual variance *both* of the treatment and of the outcome are not sufficiently strong to do so.

- Finally, the partial R^2 of `directlyharmed` with `peacefactor` means that, in an *extreme scenario*, in which we assume that unobserved confounders explain *all* of the left out variance of the outcome, these unobserved confounders would need to explain at least 2.2% of the residual variance of the treatment to fully explain away the observed effect.

These quantities summarize what we need to know in order to safely rule out confounders that are deemed to be problematic. Researchers can then argue as to whether they fall within plausible bounds on the maximum explanatory power that unobserved confounders could have in a given application.

Where investigators are unable to offer strong arguments limiting the *absolute strength* of confounding, it can be productive to consider *relative claims*, for instance, by arguing that unobserved confounders are likely not multiple times stronger than a certain observed covariate. In our application, this is indeed the case. One could argue that, given the nature of the attacks, it is hard to imagine that, within village, unobserved confounding could explain much more of the residual variance of targeting than what is explained by the observed variable `female`. The lower corner of the table, thus, provides bounds on confounding as strong as female, $R^2_{Y\sim Z|\mathbf{X},D} = 12.5\%$, and $R^2_{D\sim Z|\mathbf{X}} = 0.9\%$. Since both of those are below the robustness value of $RV_{q=1} = 13.9\%$, confounders as strong as `female` are not sufficient to explain away the observed point estimate. Moreover, the bound on $R^2_{D\sim Z|\mathbf{X}}$ is below the partial R^2 of the treatment with the outcome, $R^2_{Y\sim D|\mathbf{X}} = 2.2\%$. This means that even an extreme confounder explaining *all* residual variation of the outcome and as strongly associated with the treatment as `female` is also not logically capable of bringing the point estimate down to zero. For cases where one association is above the RV and the other below it, such as for the case of the $RV_{q=1,\alpha=0.05}$ of 7.5%, we conduct additional analyses as illustrated by the sensitivity contour plots we show next. As noted in Section 2.4, these results are exact for a single unobserved confounder, and conservative for multiple confounders, possibly acting non-linearly.

Finally, the summary method for `sensemakr` provides an extensive report with verbal descriptions of all these analyses. Entering the command `summary(darfur.sensitivity)` produces verbose output similar to the text explanations in the last several paragraphs (and thus not reproduced here), so that researchers can directly cite or include such text in their reports.

3.2.2 Sensitivity contour plots of point estimates and t-values

The minimal report of sensitivity results provided by Table 2 offers a useful summary of how robust the current estimate is to unobserved confounding. Researchers can extend and refine sensitivity analyses through plotting methods for `sensemakr` that visually explore the whole range of possible estimates that confounders with different strengths could cause. These plots can also represent different bounds on the plausible strength of confounding based on different assumptions on how they compare to observed covariates.

We begin by examining the default plot type, contour plots for the point estimate.

```
plot(darfur.sensitivity)
```

The resulting plot is shown in the left panel of Figure 1. The horizontal axis shows the residual share of variation of the treatment that is hypothetically explained by unobserved confounding, $R^2_{D\sim Z|\mathbf{X}}$. The vertical axis shows the hypothetical partial R^2 of unobserved confouding with the outcome, $R^2_{Y\sim Z|\mathbf{X},D}$. The contours show what estimate for `directlyharmed` would have been obtained in the full regression model including unobserved confounders with such hypothetical strengths. Note the plot is parameterized in way

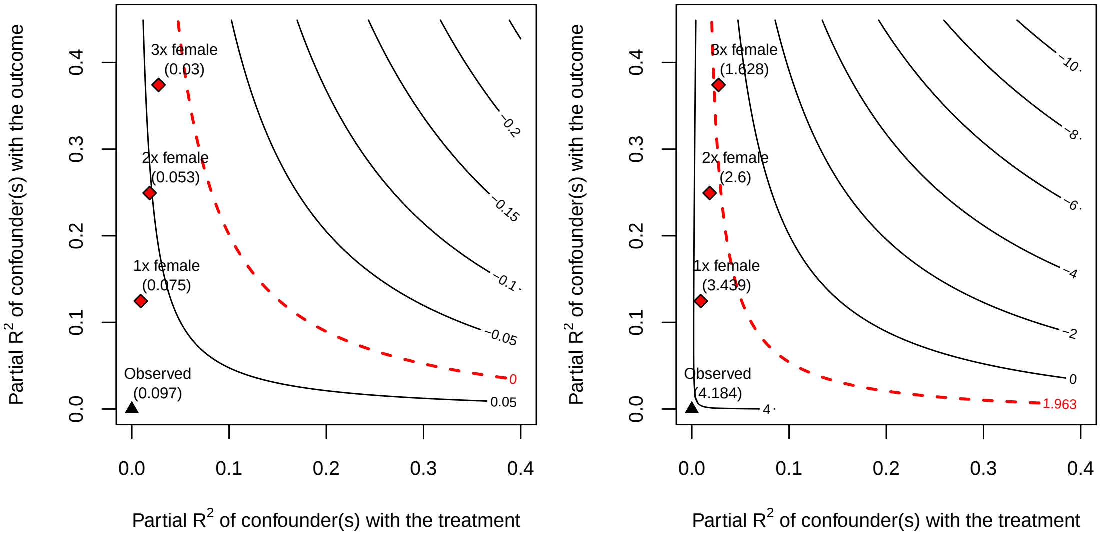

Figure 1: Sensitivity contour plots of point estimate (left) and t-value (right).

that hurts our preferred hypothesis, by pulling the estimate towards zero. Recall that the direction of the bias was determined by the argument `reduce = TRUE` of the `sensemakr()` call.

The bounds on the strength of confounding, determined by the parameter `kd = 1:3` in the call for `sensemakr()`, are also shown in the plot. The plot reveals that the direction of the effect (positive) is robust to confounding once, twice or even three times as strong as the observed covariate `female`, although in this last case the magnitude of the effect is reduced to a third of the original estimate.

We now examine the sensitivity of the *t-value* for testing the null hypothesis of zero effect by choosing the option `sensitivity.of = "t-value"` of the `plot()` method.

```
plot(darfur.sensitivity, sensitivity.of = "t-value")
```

The resulting plot is shown in the right of Figure 1. At the 5% significance level, the null hypothesis of zero effect would still be rejected given confounders once or twice as strong as `female`. However, while the point-estimate remains positive, accounting for sampling uncertainty now means that the null hypothesis of zero effect *would not* be rejected with the inclusion of a confounder three times as strong as `female`.

3.2.3 Sensitivity plots of extreme scenarios

Sometimes researchers may be better equipped to make plausibility judgments about the strength of determinants of the treatment assignment mechanism, and have less knowledge about the determinants of the outcome. In those cases, sensitivity plots using *extreme scenarios* are a useful option. These are produced with the option `type = "extreme"`. Here one assumes confounding explains all or some large fraction of the residual variance of the outcome, then vary how strongly such confounding is hypothetically related to the treatment to see how this affects the resulting point estimate. One way to interpret this

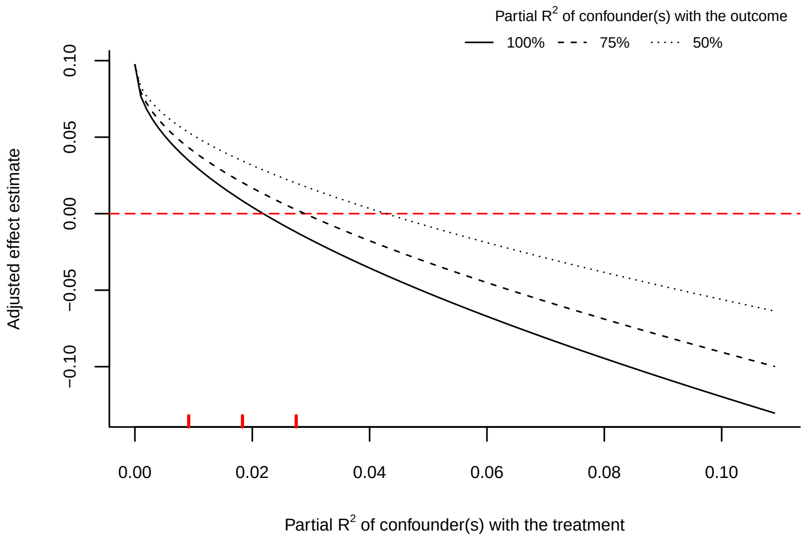

Figure 2: Sensitivity analysis to extreme scenarios.

plot in relation to the sensitivity contour plots of Figure 1 is to view it as fixing the vertical axis at a specific high value (such as its maximum of 1) while varying the horizontal axis.

```
plot(darfur.sensitivity, type = "extreme")
```

Figure 2 shows the produced plot. By default these plots consider confounding that explains 100%, 75%, and 50% of the residual variance of the outcome, producing three separate curves. This is equivalent to setting the argument `r2yz.dx = c(1, .75, .5)`. The bounds on the strength of association of a confounder once, twice or three times as strongly associated with the treatment as `female` are shown as red ticks in the horizontal axis. As the plot shows, even in the most extreme case ($R^2_{Y\sim Z|\mathbf{X},D} = 100\%$), confounders would need to be more than twice as strongly associated with the treatment as `female` to fully explain away the point estimate. Moving to the scenarios $R^2_{Y\sim Z|\mathbf{X},D} = 75\%$ and $R^2_{Y\sim Z|\mathbf{X},D} = 50\%$, confounders would need to be more than three times as strongly associated with the treatment as `female` to fully explain away the point estimate.

3.2.4 Grouped benchmarks

Users can also use a *group* of variables collectively as benchmarks, by providing a *named list* of character vectors to the `benchmark_covariates` argument. Each character vector of the list forms its own group. For example, an important application of grouping variables is to benchmark categorical variables, such as `village`, which, in our case, consists of 485 dummy variables. The code below demonstrates how to perform this analysis. As previously discussed, `village` and `female` are the main confounders in this application, with `village` being the most important one. In fact, certain villages were singled out for more or less violence and `village` also explain most of the differences in attitudes towards peace. Thus, in this particular application, it is unlikely to imagine confounders even a fraction as strong as `village` in explaining treatment and outcome variation (e.g, 10%-30% as strong). However, if such confounder did exist, as shown on the left side of Figure 3,

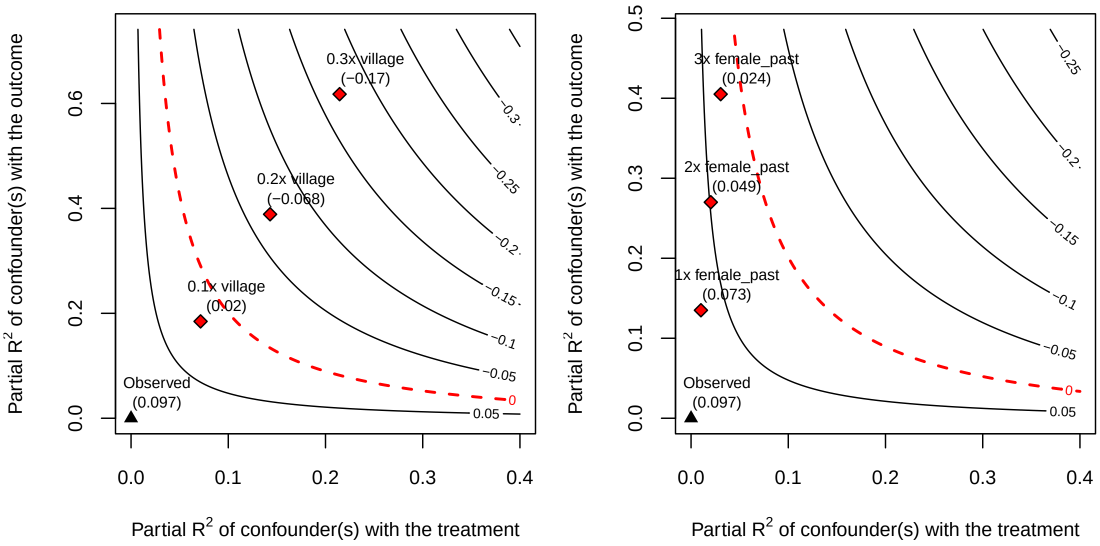

Figure 3: Sensitivity contours of the point estimate using grouped benchmarks: `village` (left); `female` and `pastvoted` (right).

they would be strong enough to completely overturn the results (note `kd` and `ky` are set to 1/10, 2/10 and 3/10).

```
village <- grep(names(coef(darfur.model)), pattern = "village", value = T)
group.sens <- sensemakr(model = darfur.model,
                        treatment = "directlyharmed",
                        benchmark_covariates = list(village = village),
                        kd = c(.1, .2, .3))
plot(group.sens)
```

Another important application of grouped benchmarks is when using interaction terms. In the appendix, we show how interaction terms can be incorporated by using the recentering strategy suggested in Hirano and Imbens (2001). This recentering transforms the model in a way such that the target parameter of inference is still a single regression coefficient. In that setting, to fully capture the strength of `female`, we now need to consider both its main term as well as its interaction term.

Finally, grouping can also be used to combine arbitrary confounders into a bundle. This approach can be useful for constructing different, more conservative scenarios in which one considers latent confounders of similar strength to the covariates in the group. For example, the command below computes bounds on the maximum strength of confounding once, twice or three times as strong as the *combined* explanatory power of the covariates `female` and `pastvoted`. The names of the list are used for setting the benchmark labels in plots and tables. As we can see in the right of Figure 3, though the impact of such a confounder would be greater than `female` alone, the main conclusions of the study would still hold under this scenario.

```
group.sens2 <- sensemakr(model = darfur.model,
                         treatment = "directlyharmed",
                         benchmark_covariates =
                             list(female_past = c("female", "pastvoted")),
                         kd = 1:3)
plot(group.sens2)
```

4. sensemakr for R: Advanced use

The standard functionality demonstrated in the previous section will suffice for most users, most of the time. More flexibility can be obtained when needed by employing additional functions, particularly:

- *functions for computing the bias, adjusted estimates and standard errors:* these comprise, among others, the functions `bias()`, `adjusted_estimate()`, `adjusted_se()` and `adjusted_t()`. They take as input the original (observed) estimate (in the form of a linear model or numeric values) and a pair of sensitivity parameters (the partial R^2 of the omitted variable with the treatment and the outcome), and return the new quantity adjusted for omitted variable bias.

- *functions for computing sensitivity statistics*: these comprise, among others, the functions `partial_r2()`, `robustness_value()`, and `sensitivity_stats()`. These functions compute sensitivity statistics suited for routine reporting, as proposed in Cinelli and Hazlett (2020). They take as input the original (observed) estimate (in the form of a linear model or numeric values), and return the corresponding sensitivity statistic.

- *sensitivity plots*: `ovb_contour_plot()` and `ovb_extreme_plot()` allow estimation and plotting of the contour and extreme scenario plots, respectively. The convenience function `add_bound_to_contour()` allows the user to place manually computed bounds on contour plots. All plot functions return invisibly the data needed to replicate the plot, so users can produce their own plots if preferred. The default options for plots work best with width and height around 4 to 5 inches.

- *bounding functions*: `ovb_bounds()` computes bounds on the maximum strength of confounding "k times" as strong as certain observed covariates. The auxiliary function `ovb_partial_r2_bound()` computes bounds for confounders by passing the values of the partial R^2 of the benchmarks directly.

We demonstrate the use of these functions below through examples chosen to illustrate important features of sensitivity analysis.

4.1 Formal versus informal benchmarking: customizing bounds

Informal "benchmarking" procedures have been suggested as aids to interpretation for numerous sensitivity analyses. These approaches are usually described as revealing how an unobserved confounder Z "not unlike" some observed covariate X_j would alter the results of a study (Imbens, 2003; Blackwell, 2013; Hosman et al., 2010; Carnegie et al., 2016; Dorie et al., 2016; Hong et al., 2018). As argued in Cinelli and Hazlett (2020), these informal proposals may lead users to erroneous conclusions, even when they make correct suppositions about how unobserved confounders compare to observed covariates. Here we replicate Section 6.1 of Cinelli and Hazlett (2020) using `sensemakr` and provide a numerical example illustrating the potential for misleading results from informal benchmarking. This example also demonstrates advanced usage of the package, including how to construct sensitivity contour plots with customized bounds.

4.1.1 Data and model

We begin by simulating the data generating process which will be used in our example, as given by Equations 9 to 12 below. Here we have a treatment variable D, an outcome variable Y, one observed confounder X, and one *unobserved* confounder Z. All disturbance variables U are standardized mutually independent normals. Note that in this case, the treatment D has no causal effect on Y.

Model 1:

$$Z = U_z \tag{9}$$

$$X = U_x \tag{10}$$

$$D = X + Z + U_d \tag{11}$$

$$Y = X + Z + U_y \tag{12}$$

Also note that, in this model: (i) the unobserved confounder Z is independent of X; and, (ii) the unobserved confounder Z is *exactly like* X in terms of its strength of association with the treatment and the outcome. The code below draws 100 samples from this data generating process. We use the function `resid_maker()` to make sure the residuals are standardized and orthogonal, thus all properties that we describe here hold exactly even with finite sample size. This function is provided by the package.

```
n <- 100
X <- scale(rnorm(n))
Z <- resid_maker(n, X)
D <- X + Z + resid_maker(n, cbind(X, Z))
Y <- X + Z + resid_maker(n, cbind(X, Z, D))
```

In this example, the investigator knows that she needs to adjust for the confounder Z but, unfortunately, does not observe Z. Therefore, she is forced to fit the restricted linear model adjusting for X only.

```
model.ydx <- lm(Y ~ D + X)
```

Results from this regression are shown in the first column of Table 3, showing a large and statistically significant coefficient estimate for both D and X.

4.1.2 Formal benchmarks

Suppose the investigator *correctly* knows that: (i) Z and X have the same strength of association with D and Y; and, (ii) Z is independent of X. How can she leverage this

	Dependent variable:	
	Y	
	Restricted OLS	Full OLS
	(1)	(2)
D	0.500***	0.000
	(0.088)	(0.102)
X	0.500***	1.000***
	(0.152)	(0.144)
Z		1.000***
		(0.144)
Observations	100	100
R^2	0.500	0.667
Residual Std. Error	1.240 (df = 97)	1.020 (df = 96)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3: First column: results of the restricted regression adjusting for X only. Second column: results of the full regression adjusting for X and Z.

information to understand how much bias a confounder Z "not unlike" X could cause? As shown in Section 2.3, Equation 7 can be used to bound the maximum amount of confounding caused by an unobserved confounder Z as strongly associated with the treatment D and with the outcome Y as the observed covariate X.

Separately from the main `sensemakr()` function, these bounds can be computed with the function `ovb_bounds()`. In this function one needs to specify the linear model being used (`model = model.ydx`), the treatment of interest (`treatment = "D"`), the observed variable used for benchmarking (`benchmark_covariates = "X"`), and how many times stronger Z is in explaining treatment (`kd = 1`) and outcome (`ky = 1`) variation, as compared to the benchmark variable X.

```
formal_bound <- ovb_bounds(model = model.ydx,
                           treatment = "D",
                           benchmark_covariates = "X",
                           kd = 1,
                           ky = 1)
```

We can now inspect the output of `ovb_bounds()`.

```
formal_bound[1:6]
```

```
  bound_label r2dz.x r2yz.dx treatment adjusted_estimate adjusted_se
1        1x X    0.5   0.333         D                 0       0.102
```

As we can see, the results of the bounding procedure correctly shows that an unobserved confounder Z, that is truly "not unlike X," would: (1) explain 50% of the residual variation of the treatment and 33% of the residual variation of the outcome; (2) bring the point estimate exactly to zero; and, (3) bring the standard error to 0.102. This is precisely what one obtains when running the full regression model adjusting for *both* X and Z, as shown in the second column of Table 3.

4.1.3 Informal benchmarks

We now demonstrate an "informal benchmark" to show its dangers. Computing the bias due to the omission of Z requires two sensitivity parameters: its partial R^2 with the treatment D and its partial R^2 with the outcome Y. Informal approaches follow from the intuition that we can simply take the observed associations of X with D and Y, found directly from regressions for the treatment and the outcome, to "calibrate" the magnitude of the sensitivity parameters of an unobserved confounder "not unlike" X. Unfortunately, as formalized in Cinelli and Hazlett (2020), these observed associations are themselves affected by the omission of the omitted variable, making naive comparisons potentially misleading.

What happens if we nevertheless attempt to use those observed statistics for benchmarking? To compute the informal benchmarks, we first need to obtain the observed partial R^2 of X with the outcome Y. This can be done using the `partial_r2()` function of `sensemakr` in the `model.ydx` regression.

```
r2yx.d <- partial_r2(model.ydx, covariates = "X")
```

We next need to obtain the partial R^2 of X with the treatment D. For that, we need to fit a new regression of the treatment D on the observed covariate X here denoted by `model.dx`.

```
model.dx <- lm(D ~ X)
r2dx   <- partial_r2(model.dx, covariates = "X")
```

We then determine what would be the implied adjusted estimate due to an unobserved confounder Z with this pair of partial R^2 values. This can be computed using the `adjusted_estimate()` function.

```
informal_adjusted_estimate <- adjusted_estimate(model     = model.ydx,
                                                treatment = "D",
                                                r2dz.x    = r2dx,
                                                r2yz.dx   = r2yx.d)
```

Let us now compare those informal benchmarks with the formal bounds. To prepare, we first plot sensitivity contours with the function `ovb_contour_plot()`. Next, we add the informal benchmark to the contours, using the numeric method of the function `add_bound_to_contour()`. Finally, we use `add_bound_to_contour()` again to add the previously computed formal bounds.

```
# draws sensitivity contours
ovb_contour_plot(model = model.ydx, treatment = "D", lim = .6)
```

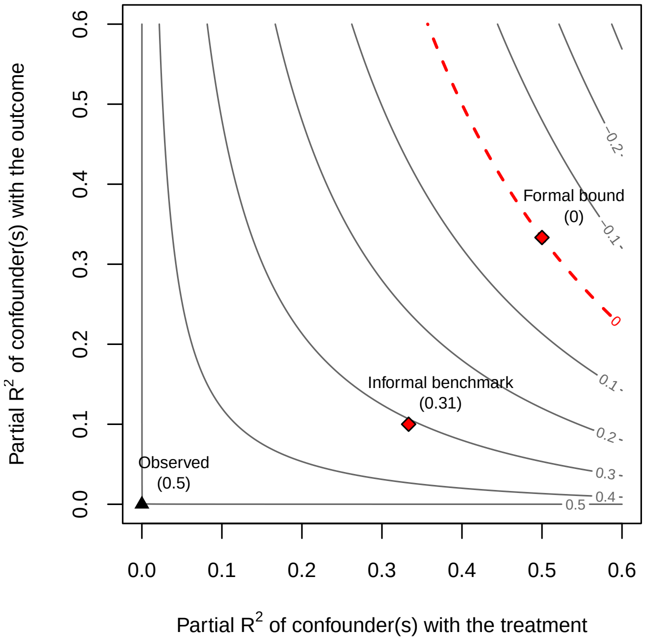

Figure 4: Informal benchmarking *versus* proper bounds.

```
# adds informal benchmark
add_bound_to_contour(r2dz.x = r2dx, r2yz.dx = r2yx.d,
                     bound_value = informal_adjusted_estimate,
                     bound_label = "Informal benchmark")

# adds formal bound
add_bound_to_contour(bounds = formal_bound,
                     bound_label = "Formal bound")
```

Note how the results from informal benchmarking are misleading: the benchmark point is still far from zero, which would suggest that an unobserved confounder Z "not unlike" X is unable to explain away the observed effect, when in fact it is, as it was shown in Table 3. This incorrect conclusion occurs despite the investigator *correctly* assuming both that: (i) Z and X have the same strength of association with D and Y; and, (ii) Z is independent of X. Therefore, we do not recommend using informal benchmarks for sensitivity analysis, and suggest researchers use formal approaches such as the ones provided with `ovb_bounds()`. For further details and discussion, see Sections 4.4 and 6.1 of Cinelli and Hazlett (2020).

4.2 Assessing the sensitivity of existing regression results

We conclude this section by demonstrating how to replicate Section 3 using only the statistics found in the regression table along with the individual functions available in the package.

4.2.1 Sensitivity statistics

The robustness value and the partial R^2 are key sensitivity statistics, useful for standardized sensitivity analyses reporting. Beyond the main `sensemakr()` function, these statistics can be computed directly by the user with the functions `robustness_value()` and `partial_r2()`. With a fitted `lm` model in hand, the most convenient way to compute the RV and partial R^2 is by employing the `lm` methods for these functions, as in

```
robustness_value(model = darfur.model, covariates = "directlyharmed")
partial_r2(model = darfur.model, covariates = "directlyharmed")
```

However, when one does not have access to the data in order to run this model, simple summary statistics such as: (i) the point estimate for the `directlyharmed` (0.097); (ii) its estimated standard error (`0.023`); and, (ii) the degrees of freedom of the regression (`783`) are sufficient to compute the RV and the partial R^2.

```
robustness_value(t_statistic = 0.097/0.023, dof = 783)
partial_r2(t_statistic = 0.097/0.023, dof = 783)
```

The convenience function `sensitivity_stats()` also computes all sensitivity statistics for a regression coefficient of interest and returns them in a `data.frame`.

4.2.2 Plotting functions

All plotting functions can be called directly with `lm` objects or numerical data. For example, the code below uses the function `ovb_contour_plot()` to replicate Figure 1 (without the bounds) using only the summary statistics of Table 1.

```
ovb_contour_plot(estimate = 0.097, se = 0.023, dof = 783)
ovb_contour_plot(estimate = 0.097, se = 0.023, dof = 783,
                 sensitivity.of = "t-value")
```

The extreme scenario plots (as in Figure 2) can also be reproduced from summary statistics using the function `ovb_extreme_plot()`,

```
ovb_extreme_plot(estimate = 0.097, se = 0.023, dof = 783)
```

All plotting functions return (invisibly) the data needed to reproduce them, allowing users to create their own plots if they prefer.

4.2.3 Adjusted estimates, standard errors and t-values

These functions allow users to compute the adjusted estimates given different postulated degrees of confounding. For instance, suppose a researcher has reasons to believe a confounder explains 10% of the residual variance of the treatment and 15% of the residual variance of the outcome. If the underlying data are not available, the investigator can still compute the adjusted estimate and t-value that one would have obtained in the full regression adjusting for such confounder.

	Dependent variable:
	directlyharmed
female	-0.097^{***} (0.036)
Observations	1,276
R^2	0.426
Residual Std. Error	0.476 (df = 784)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Treatment regression for the Darfur example. To conserve space only the results for `female` are shown, which will be used for benchmarking.

```
 adjusted_estimate(estimate = 0.097, se = 0.023, dof = 783,
                   r2dz.x = .1, r2yz.dx = 0.15)
```

```
[1] 0.0139
```

```
 adjusted_t(estimate = 0.097, se = 0.023, dof = 783,
            r2dz.x = .1, r2yz.dx = 0.15)
```

```
[1] 0.622
```

The results show this confounder is not strong enough to bring the estimate to zero, but it is sufficient to bring the t-value below the usual 5% significance threshold of 1.96.

4.2.4 Computing bounds from summary statistics

Finally, we show how users can compute bounds on the strength of confounding using only summary statistics, if the paper also provides a *treatment* regression table, i.e., a regression of the treatment on all observed covariates. Such regressions are sometimes shown in published works as part of efforts to describe the "determinants" of the treatment, or as "balance tests" in which the investigator assesses whether observed covariates predict treatment assignment. For the Darfur example, this regression is shown in Table 4 (only the results for the coefficient of `female` is shown). Note that, for benchmarking purposes, one needs to consider the regression of D on all $\mathbf{X}$, and not only of D on X_j separately.

Using the results of Tables 1 and 4 we can compute the bounds on confounding 1, 2 and 3 times as strong as `female`, as we have done before. First we compute the partial R^2 of `female` with the treatment and the outcome

```
 r2yxj.dx <- partial_r2(t_statistic = -0.232/0.024, dof = 783)
 r2dxj.x <- partial_r2(t_statistic = -0.097/0.036, dof = 783)
```

Next, we compute the bounds on the partial R^2 of the unobserved confounder using the `ovb_partial_r2_bound()` function.

```
bounds <- ovb_partial_r2_bound(r2dxj.x = r2dxj.x, r2yxj.dx = r2yxj.dx,
                               kd = 1:3, ky = 1:3,
                               bound_label = paste(1:3, "x", "female"))
```

Finally, the `adjusted_estimate()` function computes the estimates implied by these hypothetical confounders.

```
bound.values <- adjusted_estimate(estimate = 0.0973, se = 0.0232, dof = 783,
                                  r2dz.x = bounds$r2dz.x,
                                  r2yz.dx = bounds$r2yz.dx)
```

This information along with the numeric methods for the plot functions, allow us to reproduce the contour plots of Figure 1 using only summary statistics. Note that, since we are performing all calculations manually, appropriate limits of the plot area need to be set by the user.

```
ovb_contour_plot(estimate = 0.0973, se = 0.0232, dof = 783, lim = 0.45)
add_bound_to_contour(bounds, bound_value = bound.values)
```

5. `sensemakr` for `Stata`

For `Stata` users, we have also developed a homonymous package `sensemakr`, which is available for download on SSC. The package can be installed as follows:

```
ssc install sensemakr, replace all
```

The main function of the `Stata` package is `sensemakr`, which is called using the format:

```
sensemakr depvar covar [if] [in], treat(varlist)
```

For consistency with the syntax of the well-known `regress` command, the first variable is assumed to be the dependent variable, while the subsequent treatment variable and covariates can appear in any order. The required argument is `treat(varlist)`, which indicates the treatment variable for which sensitivity analysis is conducted.

By default, `sensemakr` displays sensitivity statistics for routine reporting, as well as a text interpretation of the results. Specifically, the output table reports three key values: the partial R^2 of the treatment with the outcome (R2yd.x), the robustness value (RV) required to reduce the point estimate entirely to zero (if q= 1), and the RV beyond which the estimate would no longer be statistically distinguishable from zero at the 5% level (q= 1, α= 0.05).

Should users wish to bound the plausible strength of unobserved confounders relative to existing covariates, they can specify the option `benchmark(varlist)`. `benchmark()` can accept multiple covariates from the main specification, including time-series and factor variables. If a benchmark is specified, `sensemakr` displays a bounds table. By default, this bounds table displays estimates for a hypothetical confounder that is 1, 2, and 3 times as strong as each benchmark covariate in explaining residual variation in both the treatment and the outcome, as well as adjusted coefficient estimates for the treatment if such a confounder were present. In addition to these bounds, the table displays treatment coefficients under an "extreme scenario," in which the confounder is assumed to have the same relationship to the treatment (R2dz.x) as each benchmark, but explains *all* the residual variance of the outcome (R2yz.dx=1).

5.1 Violence in Darfur

In this section, we briefly demonstrate how to replicate the analysis of Section 3, using the dataset `darfur.dta` included with `sensemakr` for `Stata`.

Users can investigate the sensitivity of the `directlyharmed` treatment estimate, as well as bounds using the benchmark covariate `female`, via the following call:

```
  use darfur.dta, clear
  sensemakr peacefactor directlyharmed age farmer herder pastvoted hhsize ///
  female i.village_, treat(directlyharmed) benchmark(female)
```

Grouped benchmarks can be assessed using the `gbenchmark(varlist)` option. For instance, the following code adds the joint benchmark `female` and `pastvoted`. Note that while the options `gbenchmark()` and `benchmark()` can be used in tandem, only a single grouped benchmark, consisting of all the variables specified in `gbenchmark()`, can be evaluated per `sensemakr` call.

```
. sensemakr peacefactor directlyharmed age farmer herder pastvoted hhsize ///
female i.village_, treat(directlyharmed) benchmark(female) ///
gbenchmark(female pastvoted)
```

Users can modify the output using the following options:

- **alpha(real)**: the significance level. The default is `0.05`.
- **gname(string)**: enables the user to specify a custom name for the grouped benchmark specified in `gbenchmark()` (if used). By default, names for grouped benchmarks are constructed by appending variables with '-'.
- **kd(numlist)** and **ky(numlist)**: these arguments parameterize how many times stronger the confounder is related to the treatment (`kd`) and to the outcome (`ky`), in comparison to the benchmark covariate. By default, kd and ky are set to (1 2 3), so provides estimates for a hypothetical confounder that is 1, 2, and 3 times as strong as each benchmark covariate. If only option `kd(numlist)` is provided, `ky` will be set equal to `kd` by default. If the user opts to specify `kd` *and* `ky`, the number of elements within each option must be equivalent.
- **latex(filename)**: saves a condensed version of the reporting outputs in filename.tex.
- **noreduce**: the default functionality assumes that confounders *reduce* the absolute value of the estimate. If the user wishes to assume that confounders pull the estimate away from zero, they can specify the `noreduce` flag.

- **q(real)**: this option enables the user to specify what fraction of the effect estimate would have to be explained away to be problematic. Defaults to 1, implying that a reduction of 100% of the current effect estimate (true effect of 0) would be problematic.
- **r2yz(numlist)**: Allows the user to specify alternative scenarios for the extreme bounds table. For instance, inputting `(.5 .75)` would display the expected treatment coefficients if a confounder explained 50% and 75% of the residual variance of the outcome. By default `r2yz` is set to 1.
- **suppress**: eliminates verbose description of sensitivity statistics.

Should users wish to design their own custom exports, all reported estimates are accessible within the `e()` class.

5.1.1 Sensitivity contour plots of point estimates and t-values

Sensitivity plots for point estimates and t-values can be generated by appending the options `contourplot` and `tcontourplot`, respectively, to the `sensemakr` call. The contour plots can be customized with the following display options:

- **clines**: the number of contour lines to display on each plot. Defaults to 7.
- **clim(numlist)**: the symmetric axis limits for the contour plots. Max range is `(0 1)`

In addition, advanced users can generate their own plots by accessing the raw contour data within `e(contourgrid)` or `e(tcontourgrid)`.

5.1.2 Sensitivity plots of extreme scenarios

Plots for extreme confounding scenarios are generated using the `extremeplot` option. By default these plots consider confounding that explains 100%, 75%, and 50% of variation in the residual outcome, producing three separate curves for each scenario. The extreme scenario plot can be customized with the following display options:

- **r2yz(numlist)**: enables the user to specify custom values for the extreme plot. Users can specify a maximum of four custom values.
- **elim(numlist)**: adjusts the x-axis limits of the plot. Max range is `(0 1)`. Note that limits for the y-axis are set automatically to include the critical value.

5.2 Differences between R and Stata implementations

The `Stata` package produces analogous outputs to the `R` implementation. There are two main differences between the packages. First, the additional functions described in section 4 are implemented in Mata, and are thus inaccessible from the command line in the `Stata` package. Users who wish to access these helper functions should favor the `R` implementation. Second, the plots produced by the `R` package can be more easily customized than those produced by the `Stata` package. However, if advanced users wish to customize the contour plots produced by the `Stata` package, they can reconstruct them using the raw data provided in the e() class output: e(contourgrid) and e(tcontourgrid) for the coefficient and t-contour plots, respectively.

6. Discussion

We recognize that the tools we present here have the potential to be misused, and that it may be tempting to use sensitivity analyses as "robustness tests" that should be "passed," in way similar to the current abuse we observe, for instance, with statistical significance testing (Ziliak and McCloskey, 2008; Cinelli, 2012; Benjamin et al., 2018; Amrhein and Greenland, 2018). We thus conclude this tutorial with brief remarks regarding the appropriate use of sensitivity analysis in general and as applied to the tools provided by `sensemakr` in particular.

What sensitivity analyses can and cannot tell us

The quantities and graphics computed by `sensemakr` tell us what we need to be *prepared to believe* in order to sustain that a given conclusion is not due to confounding. For instance, in the applied example discussed here, `sensemakr` reveals that, even in a worst case scenario where the unobserved confounder explains all the residual variation of the outcome, this unobserved confounder would need to be more than twice as strongly associated with the treatment as the covariate `female` to fully explain away the observed estimated effect of `directlyharmed`. This is a true quantitative statement that describes the strength of confounding needed to overturn the research conclusions.

Note, however, that sensitivity analyses cannot tell us whether such confounder is likely to exist. The role of sensitivity analysis is, therefore, to discipline the discussion regarding the causal interpretation of the effect estimate. Ultimately, this discussion needs to rely on domain knowledge, and is beyond the realm of statistics alone. To illustrate using our example:

1. A causal interpretation of the research conclusion may be defended by claiming that, given the way injuries (the "treatment") occurred, the scope for targeting particular types of individuals was quite limited; aircraft dropped makeshift and unguided bombs and other objects over villages, and militia raided without concern for who they would attack—the only known major exception to this, due to sexual assaults, was targeting gender, which is also one of the most visually apparent characteristics of an individual. Thus, a confounder twice as strong as `female` would be indeed surprising.

2. Similarly, for the causal conclusion to be persuasively dismissed, it does not suffice to argue that *some* confounding might exist. Helpful skepticism must articulate why a confounder that explains more than twice of the variation of the treatment assignment than the covariate `female` is plausible. Otherwise, the putative confounder cannot logically account for all the observed association, even if it explains all or some large portion of the residual outcome variation.

Robustness to confounding is thus claimed to the extent one agrees with the arguments articulated in point 1, while the results can be deemed fragile insofar as alternative stories meeting the requirements in point 2 can be offered. Both types of arguments need to rely on domain knowledge as to how the attacks occurred and what could presumably influence the outcome variable.

In sum, sensitivity analyses should not be used to obviate discussions about confounding by engaging in automatic procedures; rather, they should be used to stimulate a disciplined, quantitative argument about confounding, in which such statements are made and debated. The tools provided by `sensemakr` allow users to easily and transparently report the sensitivity of their causal inferences to unobserved confounding, thereby enabling this disciplined discussion as to what can be concluded from imperfect observational studies.

Acknowledgments

We would like to thank the associate editor and three anonymous reviewers for their helpful comments. This work is supported in part by the Royalty Research Fund at the University of Washington, and by the National Science Foundation under Grant No. MMS-2417955.

Appendix A. Grouped Benchmarks

Here we show that using the observed partial R^2 of a *group* of variables along with the formulas provided for a single variable provides conservative bounds on the strength of unobserved confounding as strong as that group of variables.

Proposition 1 *Let the observed covariates* $\mathbf{X} = [X_1, \ldots, X_p]$ *be orthogonal to* Z*, or consider the part of* $\boldsymbol{X}$ *not linearly explained by* Z*. Denote by* $\mathbf{X_{(1\ldots j)}}$ *the group (vector) of variables* $[X_1, \ldots, X_j]$*. Denote the complement of this set by* $\mathbf{X_{-(1\ldots j)}}$*. Define,*

$$k_D := \frac{R^2_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}}}{R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}, \qquad k_Y := \frac{R^2_{Y\sim Z|D,\boldsymbol{X}_{-(1\ldots j)}}}{R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}}. \tag{13}$$

Then,

$$R^2_{D\sim Z|\boldsymbol{X}} = k_D f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}, \qquad R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}, \tag{14}$$

where,

$$f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}} := \frac{R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}{1 - R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}},$$

$$f^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}} := \frac{R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}}{1 - R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}},$$

$$\eta := \left(\frac{\sqrt{k_Y} + \left| f_{K_D,(1,\ldots j)} \times f_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}} \right|}{\sqrt{1 - f^2_{K_D,(1,\ldots j)} \times f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}} \right),$$

and,

$$f^2_{K_D,(1,\ldots j)} := \frac{k_D R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}{1 - k_D R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}.$$

Proof We start with the first equality in Equation 14, namely, $R^2_{D\sim Z|\boldsymbol{X}} = k_D f^2_{D\sim \mathbf{X_{(1\ldots j)}}|\mathbf{X_{-(1\ldots j)}}}$. This result was first proved by Cinelli and Hazlett (2020), and here we present an alternative argument that might be simpler to follow. Consider the full treatment regression,

$$D = \hat{\delta}_{ZD} Z + \boldsymbol{X}_{(1\ldots j)} \hat{\theta}_{(1\ldots j)} + \boldsymbol{X}_{-(1\ldots j)} \hat{\theta}_{-(1\ldots j)} + \hat{\varepsilon}_D, \tag{15}$$

and the restricted treatment regression,

$$D = \boldsymbol{X}_{(1\ldots j)} \hat{\theta}_{(1\ldots j),r} + \boldsymbol{X}_{-(1\ldots j)} \hat{\theta}_{-(1\ldots j),r} + \hat{\varepsilon}_D. \tag{16}$$

Note $R^2_{D\sim Z|\boldsymbol{X}}$ is the partial R^2 of Z in Equation 15, whereas $R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}$ is the partial R^2 of $\boldsymbol{X}_{(1\ldots j)}$ in Equation 16. Now define the index $X_{D,j} := \boldsymbol{X}_{(1\ldots j)} \hat{\theta}_{(1\ldots j)}$. Note that, since $\boldsymbol{X} \perp Z$, we have that $X_{D,j} = \boldsymbol{X}_{(1\ldots j)} \hat{\theta}_{(1\ldots j),r}$. We can rewrite the full treatment regression as

$$D = \hat{\delta}_{ZD} Z + X_{D,j} + \boldsymbol{X}_{-(1\ldots j)} \hat{\theta}_{-(1\ldots j)} + \hat{\varepsilon}_D. \tag{17}$$

We can likewise re-express $R^2_{D\sim Z|\boldsymbol{X}}$ as,

$$R^2_{D\sim Z|\boldsymbol{X}} = R^2_{D\sim Z|X_{D,j},\boldsymbol{X}_{-(1\ldots j)}}. \tag{18}$$

Applying the recursive definition of partial correlations, the absolute value of $R_{D\sim Z|X_{D,j},\boldsymbol{X}_{-(1\ldots j)}}$ becomes,

$$\left| R_{D\sim Z|X_{D,j},\boldsymbol{X}_{-(1\ldots j)}} \right| = \left| \frac{R_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}} - R_{D\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}} R_{Z\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}}}{\sqrt{1 - R^2_{D\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1 - R^2_{Z\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}}}} \right| \tag{19}$$

$$= \left| \frac{R_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}}}{\sqrt{1 - R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}} \right| \tag{20}$$

$$= \left| \frac{\sqrt{k_D} R_{D\sim \boldsymbol{X}_{(1\ldots j)}|\mathbf{X_{-(1\ldots j)}}}}{\sqrt{1 - R_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}} \right|. \tag{21}$$

Here the first equality in Equation 20 stems from the orthogonality of $\boldsymbol{X}$ and Z, which implies $R_{Z\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}} = 0$ and $X_{D,j} = \boldsymbol{X}_{(1\ldots j)} \hat{\theta}_{(1\ldots j),r}$, and the second equality of Equation 21 is due to the definition of k_D. We thus have that,

$$R^2_{D\sim Z|X_{D,j},\boldsymbol{X}_{-(1\ldots j)}} = k_D \times f^2_{D\sim \mathbf{X_{(1\ldots j)}}|\mathbf{X_{-(1\ldots j)}}}, \tag{22}$$

as desired.

We now prove the second inequality. While Cinelli and Hazlett (2020) had originally proposed iteratively using the recursive definition of partial correlations, we show that one can simply use the same formula for benchmarking against a single variable X_j to obtain valid upper bounds when using the partial R^2 of a group of variables $\boldsymbol{X}_{(1\ldots j)}$.

First, we we need to re-express $R_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}$. Using again the recursive definition of partial correlations we have,

$$\left|R_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}\right| = \left|\frac{R_{Z\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}} - R_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}} R_{D\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}}}{\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-R^2_{D\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}}}}\right| \tag{23}$$

$$= \left|\frac{R_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}} R_{D\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}}}{\sqrt{1-R^2_{D\sim Z|\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-R^2_{D\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}}}}\right| \tag{24}$$

$$\leq \left|\frac{\sqrt{k_D} R_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}} R_{D\sim \boldsymbol{X}_{(1\ldots j)}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}{\sqrt{1-k_D R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}}}\right| \tag{25}$$

$$= \left|f_{K_D,(1,\ldots j)} \times f_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}\right|. \tag{26}$$

The equality in Equation 24 comes from the orthogonality $\boldsymbol{X} \perp Z$ which implies that $R_{Z\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}} = 0$. The inequality in Equation 25 uses the definition of k_D, which implies $R^2_{Z\sim X_{D,j}|\boldsymbol{X}_{-(1\ldots j)}} = k_D R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}}$, and the fact that $R^2_{D\sim X_{Y,j}|\boldsymbol{X}_{-(1\ldots j)}} \leq R^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}$.

Now going back to our target, $R^2_{Y\sim Z|D,\boldsymbol{X}}$, it can be written in terms of $X_{Y,j}$ and $\boldsymbol{X}_{-(1\ldots j)}$,

$$R^2_{Y\sim Z|D,\boldsymbol{X}} = R^2_{Y\sim Z|D,X_{Y,j},\boldsymbol{X}_{-(1\ldots j)}} \tag{27}$$

Using the recursive definition of partial correlations we can re-express $R^2_{Y\sim Z|D,X_{Y,j},\boldsymbol{X}_{-(1\ldots j)}}$ as,

$$R_{Y\sim Z|D,X_{Y,j},\boldsymbol{X}_{-(1\ldots j)}} = \frac{R_{Y\sim Z|D,\boldsymbol{X}_{-(1\ldots j)}} - R_{Y\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}} R_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}}{\sqrt{1-R^2_{Y\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-R^2_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}}} \tag{28}$$

Finally, note that: (i) $R^2_{Y\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}} \leq R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}$, since the linear combination $X_{Y,j}$ is not necessarily the one that maximizes the R^2 in the restricted outcome regression; (ii) by the definition of k_Y, we have $R^2_{Y\sim Z|D,\boldsymbol{X}_{-(1\ldots j)}} = k_Y R^2_{Y\sim \mathbf{X}_{(\mathbf{1}\ldots\mathbf{j})}|\mathbf{D},\mathbf{X}_{-(\mathbf{1}\ldots\mathbf{j})}}$; and, (iii) as we showed before, $\left|R_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}\right| = \left|f_{K_D,(1,\ldots j)} \times f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}\right|$. Combining these we have,

$$\begin{aligned}
\left|R_{Y\sim Z|D,X_{Y,j},\boldsymbol{X}_{-(1\ldots j)}}\right| &\leq \frac{\left|R_{Y\sim Z|D,\boldsymbol{X}_{-(1\ldots j)}}\right| + \left|R_{Y\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}} R_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}\right|}{\sqrt{1-R^2_{Y\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-R^2_{Z\sim X_{Y,j}|D,\boldsymbol{X}_{-(1\ldots j)}}}} \\
&\leq \frac{\left|\sqrt{k_Y} R_{Y\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}\right| + \left|R_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}} \times f_{K_D,(1,\ldots j)} \times f_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}\right|}{\sqrt{1-R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}}\sqrt{1-f^2_{K_D,(1,\ldots j)} \times f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}} \\
&= \left(\frac{\sqrt{k_Y} + \left|f_{K_D,(1,\ldots j)} \times f_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}\right|}{\sqrt{1-f^2_{K_D,(1,\ldots j)} \times f^2_{D\sim \boldsymbol{X}_{(1\ldots j)}|\boldsymbol{X}_{-(1\ldots j)}}}}\right)\left(\frac{\left|R_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}\right|}{\sqrt{1-R^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}}}\right) \\
&= \eta\left|f_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}}\right|.
\end{aligned} \tag{29}$$

Therefore,

$$R^2_{Y\sim Z|D,\boldsymbol{X}} \leq \eta^2 \left| f^2_{Y\sim \boldsymbol{X}_{(1\ldots j)}|D,\boldsymbol{X}_{-(1\ldots j)}} \right|, \tag{30}$$

as we wanted to show. ■

Appendix B. Interaction and Bootstrap Example

In this section we illustrate both how to incorporate interaction terms and also how to use the nonparametric bootstrap for statistical inference using `sensemakr`. As an example, suppose a researcher wants to incorporate an interaction term of the covariate `female` with the treatment assignment `directlyharmed`. Since `sensemakr` performs sensitivity analysis of a regression coefficient, the first step is to reparameterize the model such that the effect of interest is represented by a single regression coefficient. This can be done using the approach suggested in Hirano and Imbens (2001), which consists of recentering the covariate `female` and including both the main term and the interaction term of `female` and `directlyharmed`. By performing the transformation in this way, the target coefficient of interest is still given by a single regression coefficient, the main term of `directlyharmed`. Notice that, when performing this transformation, traditional standard errors would underestimate uncertainty, as it does not account for the centering process. This can be easily overcome by using the nonparametric bootstrap, and performing the centering in each bootstrap resample. This example is also useful to illustrate grouped benchmarks, as here, to fully capture the strength of `female`, we now need to consider both its main effect and its interaction effect. The code below performs this analysis, and constructs confidence intervals for the bias-adjusted estimate if confounders were as strong as the observed covariate `female`.

```
# Bootstrap
B <- 1e3; # number of bootstrap samples
n <- nrow(darfur) # number of observations in the full data
adjusted_estimate_boot <- rep(NA, B) # vector to store results

# boostrap loop
for(i in 1:B){
  cat(i,"out of", B, "\n")

  # resample data
  idx_boot <- sample(1:n, size = n, replace = T)
  dat_boot <- darfur[idx_boot, ]

  # center covariate female
  dat_boot$female <- dat_boot$female - mean(dat_boot$female)

  # fit model with interaction term
  my.ols_boot <- lm(peacefactor ~ directlyharmed*female + age + farmer_dar +
                      herder_dar + pastvoted + hhsize_darfur + village,
                      data = dat_boot)

  # sensemakr
  sense.out_boot <- sensemakr(my.ols_boot, treatment = "directlyharmed",
                              benchmark_covariates =
                               list(female = c("female","directlyharmed:female")),
                              kd = 1)

  # save the estimate
  adjusted_estimate_boot[i] <- sense.out_boot$bounds[1,"adjusted_estimate"]
}

# percentile interval
sig.level <- 0.05
quantile(adjusted_estimate_boot, c(sig.level/2, 1-sig.level/2))
```

The 95% confidence interval using the bootstrap with an interaction term is [0.014, 0.12]. The traditional interval using classical standard errors and ignoring centering is [0.031, 0.12].

Appendix C. Cluster Bootstrap Example

We now showcase how to perform the same Darfur analysis we have performed in the main text using `sensemakr` and the cluster bootstrap. We use use `village` as a cluster. The bootstrap procedure is relatively simple to implement: instead of resampling cases with replacement, we resample clusters with replacement, and keep all cases within the cluster. The code below performs this analysis to construct confidence intervals for the bias-adjusted estimate if confounders were as strong as the observed covariate `female`.

```
# Cluster Bootstrap
B <- 1e3; # number of bootstrap samples
n <- nrow(darfur) # number of observations in the full data
adjusted_estimate_boot <- rep(NA, B) # vector to store results

# bootstrap loop
for(i in 1:B){
  cat(i,"out of", B, "\n")

  # get the vector of clusters
  clusters <- unique(darfur$village)
  # sample clusters with replacement
  sample_cluster <- sample(clusters, size = length(clusters), replace = T)
  # pick observations that are in the sampled clusters
  idx_boot <- darfur$village %in% sample_cluster
  # bootsrap sample
  dat_boot <- darfur[idx_boot, ]

  # fit model
  my.ols_boot <- lm(peacefactor ~ directlyharmed + age + farmer_dar + herder_dar +
                      pastvoted + hhsize_darfur + female + village,
                    data = dat_boot)

  # sensemakr
  sense.out_boot <- sensemakr(my.ols_boot, treatment = "directlyharmed",
                               benchmark_covariates = "female",
                               kd = 1)

  # save the estimate
  adjusted_estimate_boot[i] <- sense.out_boot$bounds[1,"adjusted_estimate"]
}

# percentile interval
sig.level <- 0.05
quantile(adjusted_estimate_boot, c(sig.level/2, 1-sig.level/2))
```

The 95% confidence interval using cluster bootstrap is $[0.035, 0.118]$. The results are very similar to the classical 95% confidence interval of $[0.032, 0.118]$.

References

Alexander D'Amour AlexanderM. Franks and Avi Feller. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 115(532):1730–1746, 2020.

Valentin Amrhein and Sander Greenland. Remove, rather than redefine, statistical significance. *Nature Human Behaviour*, 2(1):4–4, 2018.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2008.

Daniel J Benjamin, James O Berger, Magnus Johannesson, Brian A Nosek, E-J Wagenmakers, Richard Berk, Kenneth A Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, et al. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6, 2018.

Matthew Blackwell. A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, 22(2):169–182, 2013.

Babette A Brumback, Miguel A Hernán, Sebastien JPA Haneuse, and James M Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in medicine*, 23(5):749–767, 2004.

Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Long story short: Omitted variable bias in causal machine learning. Technical report, National Bureau of Economic Research, 2022.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67, 2020. doi: 10.1111/rssb.12348.

Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Available at SSRN 4217915*, 2022.

Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. *International Conference on Machine Learning*, 2019.

Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. *`sensemakr`: sensitivity analysis tools for OLS*, 2020a. URL `https://CRAN.R-project.org/package=sensemakr`. `R` package version 0.3.

Carlos Cinelli, Jeremy Ferwerda, and Chad Hazlett. Sensemakr: Stata module to provide sensitivity tools for ols, 2020b. URL `https://EconPapers.repec.org/RePEc:boc:bocode:s458773`.

Carlos Leonardo Kulnig Cinelli. Inferência estatística e a prática econômica no brasil: os (ab) usos dos testes de significância, 2012. URL `https://repositorio.unb.br/handle/10482/11230`.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *journal of National Cancer Institute*, 1(23):173–203, 1959.

Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

Kenneth A Frank. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, 29(2):147–194, 2000.

Kenneth A Frank, Gary Sykes, Dorothea Anagnostopoulos, Marisa Cannata, Linda Chard, Ann Krause, and Raven McCrory. Does nbpts certification affect the number of colleagues a teacher helps with instructional matters? *Educational Evaluation and Policy Analysis*, 30(1):3–30, 2008.

Kenneth A Frank, Spiro J Maroulis, Minh Q Duong, and Benjamin M Kelcey. 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, 2013.

Chad Hazlett. Angry or weary? how violence impacts attitudes toward peace among darfurian refugees. *Journal of Conflict Resolution*, page 0022002719879217, 2019.

MA Hernán and JM Robins. Causal inference: What if. *Boca Raton: Chapman & Hill/CRC*, 2020.

Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2:259–278, 2001.

Guanglei Hong, Xu Qin, and Fan Yang. Weighting-based sensitivity analysis in causal mediation studies. *Journal of Educational and Behavioral Statistics*, 43(1):32–56, 2018.

Carrie A Hosman, Ben B Hansen, and Paul W Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, pages 849–870, 2010.

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

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

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Joel A Middleton, Marc A Scott, Ronli Diakow, and Jennifer L Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323, 2016.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

Paul R Rosenbaum and Donald B Rubin. 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, 1983.

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

Steve Ziliak and Deirdre Nansen McCloskey. *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press, 2008.