An Omitted Variable Bias Framework for Sensitivity Analysis of Instrumental Variables

PRELIMINARY DRAFT
(download latest version here)

Carlos Cinelli[*] Chad Hazlett[†]

September 18, 2020

Abstract

We develop an "omitted variable bias" framework for sensitivity analysis of instrumental variable (IV) estimates that is immune to "weak instruments," naturally handles multiple "side-effects" (violations of the exclusion restriction assumption) and "confounders" (violations of the ignorability of the instrument assumption), exploits expert knowledge to bound sensitivity parameters, and can be easily implemented with standard software. Conveniently, we show that many pivotal conclusions regarding the sensitivity of the IV estimate (e.g. tests against the null hypothesis of zero causal effect) can be reached simply through separate sensitivity analyses of two familiar auxiliary OLS estimates, namely, the effect of the instrument on the treatment (the "first stage") and the effect of the instrument on the outcome (the "reduced form"). More specifically, we introduce sensitivity statistics for routine reporting, such as *robustness values* for IV estimates, describing the minimum strength that omitted variables need to have to invalidate the conclusions of an IV study. Next we provide visual displays that fully characterize the sensitivity of IV point-estimates and confidence intervals to violations of the standard IV assumptions. Finally, we offer formal bounds on the worst possible bias under the assumption that the maximum explanatory power of omitted variables are no stronger than a multiple of the explanatory power of observed variables. We apply our methods in a running example that uses instrumental variables to estimate the returns to schooling.

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

[†] Assistant Professor, Departments of Statistics and Political Science, University of California Los Angeles. Email: chazlett@ucla.edu. URL: chadhazlett.com.

[CC: Table of contents here temporarily.]

Contents

1 Introduction

Unobserved confounding often complicates efforts to make causal claims from observational data (Pearl, 2009; Imbens and Rubin, 2015; Rosenbaum, 2017). Instrumental variable (IV) regression offers a powerful and widely used tool to address unobserved confounding, by exploiting "exogenous" sources of variation of the treatment (Wright, 1928; Bowden and Turkington, 1990; Angrist et al., 1996; Angrist and Pischke, 2009). IV methods have also become a vital tool in the analysis of randomized experiments with imperfect compliance. These qualities have made instrumental variables "a central part of the econometrics canon since the first half of the twentieth century" (Imbens, 2014, p.324). Beyond economics, instrumental variables are prominent tools in the arsenal of investigators seeking to make causal claims across the social sciences, epidemiology, medicine, and other fields (see e.g. Hernán and Robins, 2006; Didelez and Sheehan, 2007; Baiocchi et al., 2014).

Yet, IV carries its own set of demanding assumptions. Principally, (i) an instrumental variable must not itself be confounded with the outcome, and (ii) it should influence the outcome *only* by influencing uptake of the treatment. These assumptions can be violated by either omitted confounders (violating the first assumption), or omitted "side-effects" of the treatment that then influence the outcome (violating the second).[1] Although in special cases these assumptions may entail testable implications (Pearl, 1995; Swanson et al., 2018; Kédagni and Mourifié, 2020), they are often unverifiable and must instead be defended by appealing to domain knowledge and theoretical arguments. Whether a given IV study identifies the causal effect of interest, then, turns on debates as to whether these assumptions hold.

Particularly in recent years, economists and other scholars have adopted a more skeptical posture towards IV methods, emphasizing the importance of both defending the credibility of these assumptions as well as assessing the consequences of its failures (see e.g., Deaton, 2009; Heckman and Urzua, 2010). More worrisome, if the IV assumptions fail to hold, it is well known that the bias of the IV estimate may be *worse* than the original confounding bias of the simple regression estimate that IV was supposed to address (Bound et al., 1995). Therefore, researchers are also advised to perform sensitivity analyses to assess the degree of violation of the IV assumptions that would be required to alter the conclusions of an IV study. Although a variety of sensitivity methods for IV have been proposed (DiPrete and Gangl, 2004; Altonji et al., 2005; Small, 2007; Small and Rosenbaum, 2008; Conley et al., 2012; Wang et al., 2018; Jiang et al., 2018; Cinelli et al., 2019),

[1] In the recent IV literature, the first assumption is usually called *exogeneity*, *ignorability*, *unconfoundedness* or *independence* of the instrument, whereas the second assumption is called the *exclusion* restriction (Angrist and Pischke, 2009; Pearl, 2009; Imbens and Rubin, 2015; Swanson et al., 2018). In earlier econometric works, these two assumptions were often combined into one, also labeled the "exclusion restriction" (Imbens, 2014).

such sensitivity analyses are still rare in practice.

In this paper, we develop an omitted variable bias (OVB) framework for assessing the sensitivity of IV estimates against violations of its underlying assumptions.[2] Building on recent developments of the OVB framework for OLS estimates (Cinelli and Hazlett, 2020), we provide a suite of sensitivity analysis tools for IV that: (i) has correct test size (or confidence interval coverage) regardless of instrument strength; (ii) naturally handles violations due to multiple "side-effects" and "confounders;" (iii) exploits expert knowledge to bound sensitivity parameters; and, (iv) can be easily implemented with standard software.

In particular, we introduce two sensitivity statistics for IV estimates: (i) the *robustness value* for IV describes the minimum strength of association (in terms of partial R^2) that side-effects or confounders need to have, both with the instrument and with the untreated potential outcome, such that they are capable of changing the conclusions of the study; and (ii) the *extreme robustness value*, which describes the minimal strength of association that side-effects or confounders need to have with the *instrument alone* (regardless of their association with the untreated potential outcome), as otherwise such omitted variables cannot be problematic. We propose the routine reporting of those quantities to improve the transparency and facilitate the assessment of the credibility of IV studies. Next, we offer intuitive graphical tools for investigators to assess how postulated confounding of any degree would alter the IV point estimates, t-values and lower or upper limits of confidence intervals. These tools can also be supplemented with formal bounds on the worst possible bias that side-effects or confounders could cause, under the assumption that the maximum explanatory power of these omitted variables are no stronger than a multiple of the explanatory power of observed variables.

Conveniently, considering the fact that investigators are already well advised to carefully examine their "first stage" (the effect of the instrument on the treatment) and "reduced form" (the effect of the instrument on the outcome) (e.g. Angrist and Krueger, 2001; Angrist and Pischke, 2009), we show that many pivotal conclusions regarding the sensitivity of the IV estimate can be reached simply through separate sensitivity analyses of these two familiar auxiliary OLS estimates. Specifically, *examining the sensitivity of the reduced form and first stage allows a sensitivity analysis for IV that will be sufficient for many purposes.* First, if researchers are interested in the null hypothesis of *zero effect,* all the OVB tools developed for OLS in Cinelli and Hazlett (2020) and Cinelli et al. (2020) can

[2] We focus on the "just-identified" case with one treatment and one instrument. One reason for our focus is that a thorough consideration of the identification assumptions and how they may be violated is already complicated enough with a single instrument (Angrist and Pischke, 2009). Second, and relatedly, in most applied settings, the single-instrument and single-treatment setup is the most common. For example, in a broad review of papers in the AER and 15 other AEA journals, Young (2018) finds that 80% of IV regressions were of this type. Finally, in many "multiple instrument" studies, it is not uncommon for researchers to also report and give special focus on the analysis of their "best" instrument (Angrist and Pischke, 2009), or to combine instruments into a single instrument, such as, for example, constructing an allele score in Mendelian Randomization (Burgess and Thompson, 2015).

simply be directly applied to the reduced form regression, and confounders or side-effects shown to be problematic there are equally problematic for IV. Second, for researchers interested in assessing not just the null of zero, but biases towards *zero or worse*, we show that *the robustness of the IV estimate formally reduces to the minimum of the robustness of the reduced form and the robustness of the first stage regressions.*

Finally, while developing OVB tools for IV, we refined and extended the original OVB tools for OLS (Cinelli and Hazlett, 2020), resulting in a new perspective on how to perform sensitivity analysis that is extremely simple to implement, and can thus aid in its widespread adoption. We show that, to perform sensitivity analysis to omitted variables, one can simply replace traditional critical values by an "OVB-adjusted" critical value. This value depends only on the hypothetical partial R^2 of the omitted variables with the dependent and independent variables of the OLS regression, the degrees of freedom, and the significance level of the tests. Traditional critical values can then be seen as simply adjusted critical values assuming *zero* confounding. Researchers can thus easily perform sensitivity analysis with *any* standard regression software, by simply substituting traditional thresholds with these new thresholds, such as comparing observed t-values with adjusted critical values, or constructing confidence intervals using adjusted critical values.

In what follows, Section 2 introduces the running example which will be used throughout the paper, and also provides the essential background on the main IV estimators—which, at their core, all leverage OLS. Section 3 thus reviews and extends the results of OVB for OLS (Cinelli and Hazlett, 2020), which not only improves the sensitivity of OLS on its own, but greatly simplifies its use in the IV setting. Section 4 then extends the OVB framework specifically for IV, by first showing what can be gleaned from the first stage and reduced form regressions alone, and then establishing the necessary OVB-type results in the Anderson-Rubin framework. Section 5 shows how these results can can be deployed in practice, and in Section 6 we conclude by discussing how our proposal differs from previous approaches to sensitivity analysis for IV, and by offering further guidance on the interpretation of sensitivity results. Open-source software for `R` and `Stata` implements the methods discussed in this paper.[3]

2 Background

In this section we introduce the running example and use it to briefly review the required background on instrumental variables as a solution to the omitted variable bias problem, as well as the main

[3] Sensitivity analysis of the reduced form, first stage, and Anderson-Rubin regression for a specific null hypothesis can already be performed using the `R` and `Stata` package `sensemakr` (Cinelli et al., 2020). Additional functionality, such as contour plots with lower and upper limits of the Anderson-Rubin confidence interval, is forthcoming.

approaches to IV estimation.

2.1 Running example—estimating the returns to schooling

2.1.1 Ordinary least squares and the OVB problem

Many observational studies have established a positive and large association between educational achievement and earnings using regression analysis (Card, 1999). Here we consider the work of Card (1993), which employed a sample of 3,010 individuals from the National Longitudinal Survey of Young Men (NLSYM). Considering the multivariate linear regression below,

$$\text{Earnings} = \hat{\tau}_{\text{OLS,res}}\text{Education} + \boldsymbol{X}\hat{\beta}_{\text{OLS,res}} + \hat{\varepsilon}_{\text{OLS,res}} \tag{1}$$

where *Earnings* measures the log transformed hourly wages of the individual,[4] *Education* is an integer-valued variable indicating the completed years of education of the individual and the matrix $\boldsymbol{X}$ comprises race, experience, and a set of regional factors, Card concluded that each additional year of schooling was associated with approximately 7.5% higher wages (see Table 5).

Educational achievement, however, is not randomly assigned; perhaps individuals who obtain more education have higher wages due to other reasons, such as coming from wealthier families, or having higher levels of some unobserved characteristic, such as "ability" or "motivation." If data on these variables were available, then multivariate regression, further adjusting for such variables, would be able to capture the causal effect of educational attainment on schooling, as in

$$\text{Earnings} = \hat{\tau}_{\text{OLS}}\text{Education} + \boldsymbol{X}\hat{\beta}_{\text{OLS}} + \boldsymbol{U}\hat{\gamma}_{\text{OLS}} + \hat{\varepsilon}_{\text{OLS}} \tag{2}$$

where $\boldsymbol{U}$ denotes a set of variables that, along with $\boldsymbol{X}$, is sufficient to eliminate confounding concerns. Such detailed information on individuals, however, is not available, and researchers will not even agree upon which variables $\boldsymbol{U}$ are needed. In the absence of such variables, regression estimates that adjust for only a partial list of characteristics (such as $\boldsymbol{X}$) may suffer from "omitted variable bias" (Angrist and Pischke, 2009; Cinelli and Hazlett, 2020) and are likely to overestimate the "true" returns to schooling.

2.1.2 Instrumental variables as a solution to the OVB problem

Instrumental variable methods offer an alternative route to estimate the causal effect of schooling on earnings without having data on the unobserved variables $\boldsymbol{U}$. The key for such methods to work is to

[4] In this case, regression coefficients can be conveniently interpreted, approximately, as percent changes in earnings.

find a new variable (the "instrument") that changes the incentives to educational achievement, but is associated with earnings *only through* its effect on education. To that end, Card (1993) proposed exploiting the role of geographic differences in college accessibility. Students who grow up far from the nearest college may face higher educational costs, discouraging them from pursuing higher level studies. For instance, for such students, living at home while attending college may not be an option. Next, Card (1993) argued that, conditional on the set of observed variables $\boldsymbol{X}$, available on the NLSYM, whether one lives near a college is not itself confounded with earnings, nor does proximity to college affect earnings apart from its effect on years of education.

Let's call such variable *Proximity*. If, conditionally on $\boldsymbol{X}$, we believe that *Proximity* is a valid instrument—and under further functional restrictions on the data-generating process, such as monotonicity[5]—we can recover a valid estimate of the (local) average treatment effect of *Education* on *Earnings* by simply taking the ratio of two OLS coefficients, one measuring the effect of proximity on earnings, and another measuring the effect of proximity on educational achievement. More precisely, consider the two linear regression equations,

$$\text{Education} = \hat{\theta}_{\text{res}}\text{Proximity} + \boldsymbol{X}\hat{\psi}_{\text{res}} + \hat{\varepsilon}_{d,\text{res}} \tag{3}$$

$$\text{Earnings} = \hat{\lambda}_{\text{res}}\text{Proximity} + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}} \tag{4}$$

where *Proximity* is our instrumental variable, an indicator of whether the individual grew up in an area with a nearby accredited 4-year college. Throughout the paper we refer to these equations as the "first stage" (Equation 3) and the "reduced form" (Equation 4), as these are now common usage (Angrist and Pischke, 2009, 2014; Imbens and Rubin, 2015; Andrews et al., 2019).[6] The results of both regressions are shown in Table 5. The coefficient for *Proximity* on the first-stage regression, $\hat{\theta}_{\text{res}} \approx 0.32$, reveals that those who grew up near a college indeed have higher educational attainment, having completed an additional 0.32 years of education, on average. Likewise, the coefficient for *Proximity* on the reduced-form regression, $\hat{\lambda}_{\text{res}} \approx 0.042$, suggests that those who grew up near a college have 4.2% higher earnings.The IV estimate is then given by the ratio of these two coefficients,

$$\hat{\tau}_{\text{res}} := \frac{\hat{\lambda}_{\text{res}}}{\hat{\theta}_{\text{res}}} \approx \frac{0.042}{0.319} \approx 0.132 \tag{5}$$

[5] Conditions that allow a causal interpretation of the IV estimand are extensively discussed elsewhere, and we will not review them here. See Angrist et al. (1996); Angrist and Pischke (2009); Imbens (2014) for further discussion.

[6] Though now well established, these labels abuse the original meaning of the terminology, since both regressions are in their "reduced form." Equation 3 is called the "first stage" due to its operational role on two-stage least squares estimation, as we will see next. See also Imbens (2014) and Andrews et al. (2019).

The value of $\hat{\tau}_{\text{res}} \approx 0.132$ suggests that, contrary to the OLS estimate of 7.5%, and perhaps surprisingly, each additional year of schooling instead raises wages by much more—13.2%.

2.1.3 What if the instrumental variable estimate itself suffers from OVB?

The previous IV estimates rely on the assumption that *Proximity* is a valid instrument for the effect of *Education* on *Earnings*. This means that, conditional on $\boldsymbol{X}$, *Proximity* and *Earnings* must be unconfounded, and the effect of *Proximity* on *Earnings* must go entirely through *Education.* As is often the case, neither assumption is easy to defend in this setting.

First, some of the same factors we were initially worried might confound the relationship between *Education* and *Earnings* could similarly confound the relationship of *Proximity* and *Earnings* (such family wealth or family connections). Second, as argued in Card (1993), the presence of a college nearby may be associated with high primary and secondary school quality, which in its turn also affects earnings, and there is no information regarding the quality of the schools attended by the individuals of the survey. Finally, other geographic confounders can make some localities likely to both have colleges nearby and lead to higher earnings. These are only coarsely conditioned on by the regional indicators included in $\boldsymbol{X}$, thus residual biases may still remain.

In other words, we can summarize the previous arguments by claiming that, although we proposed an instrumental variable approach for dealing with the OVB problem, our IV analysis itself might suffer from OVB. For instance, instead of adjusting only for $\boldsymbol{X}$ as in the previous Equations 4 and 3, we should have adjusted for *both* the observed covariates $\boldsymbol{X}$ *and unobserved* covariates $\boldsymbol{W} = [\textit{Family Wealth}, \textit{High School Quality}, \textit{Place of Residence}]$ as in

$$\text{Education} = \hat{\theta}\text{Proximity} + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_d \tag{6}$$

$$\text{Earnings} = \hat{\lambda}\text{Proximity} + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_y \tag{7}$$

and the IV estimate we wished we had would be given by

$$\hat{\tau} := \frac{\hat{\lambda}}{\hat{\theta}} \tag{8}$$

Thus, our original IV estimate $\hat{\tau}_{\text{res}}$ deviates from our ideal IV estimate $\hat{\tau}$. How strong would $\boldsymbol{W}$ have to be so that it would change our original conclusions? To develop a precise algebraic answer to this question, we must first review the mechanics of IV estimation, and particularly the role of OLS in the most popular IV estimation approaches.

2.2 A brief review on (the mechanics of) IV estimation

Let Y denote the $(n \times 1)$ vector of the outcome of interest with n observations, D the $(n \times 1)$ treatment variable, and Z the $(n \times 1)$ instrumental variable. We further denote by $\boldsymbol{X}$ an $(n \times p)$ matrix of observed covariates, and by $\boldsymbol{W}$ an $(n \times l)$ matrix of *unobserved* covariates. The target quantity of IV estimation consists of a ratio of two population regression coefficients,

$$\tau := \frac{\lambda}{\theta} \tag{9}$$

where θ and λ are the population regression coefficients of Z on D (the first-stage) and Z on Y (the reduced-form) both adjusting for $\boldsymbol{X}$ and $\boldsymbol{W}$. We call the ratio τ the *IV estimand.* Here we briefly review the commonly used approaches to make inferences regarding this ratio.

2.2.1 Indirect Least Squares and Two-Stage Least Squares

Indirect Least Squares. The first and perhaps most straightforward approach to instrumental variable estimation was already outlined in the previous section. It consists of first running two OLS models capturing the effect of the instrument on the treatment (the first stage) and the effect of the instrument on the outcome (the reduced form),

$$\textbf{First-stage:} \quad D = \hat{\theta}Z + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_d \tag{10}$$

$$\textbf{Reduced-form:} \quad Y = \hat{\lambda}Z + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_y \tag{11}$$

The estimator for τ is constructed by simply using the plug-in principle,

$$\hat{\tau}_{\text{ILS}} := \frac{\hat{\lambda}}{\hat{\theta}} \tag{12}$$

The ratio $\hat{\tau}_{\text{ILS}}$ may be called the *indirect least squares* (ILS) estimator, or more straigthforwardly, the "ratio of coefficients estimator." Inference in the ILS framework can be performed using the delta-method, resulting in the estimated variance

$$\widehat{\text{var}}(\hat{\tau}_{\text{ILS}}) := \frac{1}{\hat{\theta}^2}\left(\widehat{\text{var}}(\hat{\lambda}) + \hat{\tau}_{\text{ILS}}^2\widehat{\text{var}}(\hat{\theta}) - 2\hat{\tau}_{\text{ILS}}\widehat{\text{cov}}(\hat{\lambda}, \hat{\theta})\right) \tag{13}$$

Where $\widehat{\text{var}}(\cdot)$ and $\widehat{\text{cov}}(\cdot)$ are the usual OLS variance and covariance estimators (see appendix for details).

Two-Stage Least Squares. A closely related approach for instrumental variable estimation is denoted by "two-stage least squares" (2SLS). As its name suggests, this involves two nested steps of OLS estimation. One first runs the first-stage regression given by Equation 10. Next, we regress the outcome on the fitted values of the first-stage regression, here denoted by $\widehat{D}$,

$$\textbf{Second-stage:} \quad Y = \hat{\tau}_{\text{2SLS}}\widehat{D} + \boldsymbol{X}\hat{\beta}_{\text{2SLS}} + \boldsymbol{W}\hat{\gamma}_{\text{2SLS}} + \hat{\varepsilon}_{\text{2SLS}} \tag{14}$$

The 2SLS estimate corresponds to the coefficient $\hat{\tau}_{\text{2SLS}}$ in Equation 14, called the "second-stage" regression. By appealing to the Frisch-Waugh-Lovell (FWL) theorem (Frisch and Waugh, 1933; Lovell, 1963, 2008), one can readily show that $\hat{\tau}_{\text{2SLS}}$ and $\hat{\tau}_{\text{ILS}}$ are numerically identical,

$$\hat{\tau}_{\text{2SLS}} = \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, \widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\hat{\theta} \times \text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\hat{\theta}^2 \times \text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} = \frac{\hat{\lambda}}{\hat{\theta}} \tag{15}$$

Where cov(·) and var(·) denote the sample covariance and variance; $Y^{\perp \boldsymbol{X},\boldsymbol{W}}$,$\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}}$ and $D^{\perp \boldsymbol{X},\boldsymbol{W}}$ denote the variables Y, $\widehat{D}$ and D after removing the components linearly explained by $\boldsymbol{X}$ and $\boldsymbol{W}$. As with ILS, inference in 2SLS can be performed by appealing to the asymptotic normality of the ratio, with estimated variance

$$\widehat{\text{var}}(\hat{\tau}_{\text{2SLS}}) := \frac{\text{var}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \hat{\tau}_{\text{2SLS}} D^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(\widehat{D}^{\perp \boldsymbol{X},\boldsymbol{W}})} \times (n - p - l - 1)^{-1} \tag{16}$$

Using the FWL theorem one can further show that $\widehat{\text{var}}(\hat{\tau}_{\text{2SLS}})$ and $\widehat{\text{var}}(\hat{\tau}_{\text{ILS}})$ are also numerically identical (see appendix).

2.2.2 Anderson-Rubin regression and Fieller's theorem

The methods of ILS and 2SLS make use of a normal approximation to the sampling distribution of the ratio $\hat{\lambda}/\hat{\theta}$, which may prove unreliable when θ is "close" to zero, relative to the sampling variability of $\hat{\theta}$—this is known as the "weak instrument" problem. Two simple and recommended alternatives that allows constructing confidence intervals with correct coverage, regardless of the "strength" of the first stage, are the proposals of Anderson and Rubin (1949) and Fieller (1954) (e.g. see Andrews et al., 2019).

Anderson-Rubin regression. Suppose we would like to test whether the causal effect of the treatment D on the outcome Y is equal to a specific value, say, $\tau = \tau_0$. Now recall that, if Z is indeed a valid instrumental variable for the effect on D on Y, then it should be associated with Y

only through its effect on D. The Anderson-Rubin approach exploits this idea for inference. Create the random variable $Y_{\tau_0} := Y - \tau_0 D$ in which we subtract from Y the putative causal effect of D. Then, for a valid instrument, under the null hypothesis of $H_0 : \tau = \tau_0$, we should not see an association between Y_{τ_0} and Z, conditional on $\boldsymbol{X}$ and $\boldsymbol{W}$. In other words, if we run the linear regression

$$\textbf{Anderson-Rubin:} \quad Y_{\tau_0} = \hat{\phi}_{\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\tau_0} + \boldsymbol{W}\hat{\gamma}_{\tau_0} + \hat{\varepsilon}_{\tau_0} \tag{17}$$

then we should find that $\hat{\phi}_{\tau_0}$ is equal to zero, but for sampling variation. To test the null hypothesis $H_0 : \phi_{\tau_0} = 0$ in the Anderson-Rubin regression is thus to test the null hypothesis $H_0 : \tau = \tau_0$. A valid $1 - \alpha$ confidence interval can be constructed by collecting all values τ_0 such that the null hypothesis $H_0 : \phi_{\tau_0} = 0$ is not rejected at the chosen significance level α:

$$\text{CI}_{1-\alpha}(\tau) = \{\tau_0;\ t^2_{\phi_{\tau_0}} \leq t^{*2}_{\alpha,n-p-l-1}\} \tag{18}$$

Where $t_{\phi_{\tau_0}}$ is the t-value for testing the coefficient $\hat{\phi}_{\tau_0}$, and $t^*_{\alpha,n-p-l-1}$ the usual α level critical threshold for the t statistic, with the appropriate degrees of freedom. It is also convenient to define the point estimate $\hat{\tau}_{\text{AR}}$ as the value τ_0 which makes $\hat{\phi}_{\tau_0}$ exactly equal to zero

$$\hat{\tau}_{AR} := \{\tau_0;\ \hat{\phi}_{\tau_0} = 0\} \tag{19}$$

By the FWL theorem, note we can write $\hat{\phi}_{\tau_0}$ as a linear combination of $\hat{\lambda}$ and $\hat{\theta}$,

$$\hat{\phi}_{\tau_0} = \frac{\text{cov}(Y^{\perp \boldsymbol{X},\boldsymbol{W}} - \tau_0 D^{\perp \boldsymbol{X},\boldsymbol{W}}, Z^{\perp \boldsymbol{X},\boldsymbol{W}})}{\text{var}(Z^{\perp \boldsymbol{X},\boldsymbol{W}})} = \hat{\lambda} - \tau_0 \hat{\theta} \tag{20}$$

Thus resulting in $\hat{\tau}_{AR} = \frac{\hat{\lambda}}{\hat{\theta}}$, a point estimate numerically identical to the previous estimators.

Fieller's theorem. The connection between Fieller's theorem and the Anderson-Rubin approach follows from Equation 20. The central statistic of Fieller's theorem is the linear combination $\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0 \hat{\theta}$. Under the null hypothesis $H_0 : \tau = \tau_0$, if the estimators $\hat{\lambda}$ and $\hat{\theta}$ are asymptotically normal, it follows that $\hat{\phi}_{\tau_0}$ is also asymptotically normal with mean zero, and estimated variance

$$\widehat{\text{var}}(\hat{\phi}_{\tau_0}) = \widehat{\text{var}}(\hat{\lambda}) + \tau_0^2 \widehat{\text{var}}(\hat{\theta}) - 2\tau_0 \widehat{\text{cov}}(\hat{\lambda}, \hat{\theta}) \tag{21}$$

Confidence intervals are then constructed exactly as in Equation 18. This confidence interval has some properties that are important for our purposes. As we will further discuss in Section 4, the confidence interval can take three forms: (i) a connected interval; (ii) a union of two disjoint unbounded (more precisely, half-bounded) intervals; and, (iii) the whole real line. Further, the interval will be unbounded if, and only if, the null hypothesis that the first-stage is zero is not rejected; likewise, the interval will include zero if, and only if, the null hypothesis that the reduced-form is zero is not rejected.

2.3 Problem statement

As we have seen, all main approaches for IV estimation result in the same point estimate—the ratio of the reduced-form and first-stage regression coefficients. They differ only in how to perform inference, with ILS/2SLS resorting to the asymptotic normality of the ratio estimator, and the Anderson-Rubin/Fieller approach inverting the test of the linear combination of both coefficients. The equations presented above, replicated in the third column of Table 1, are the IV regressions our analyst *wished* she had run, adjusting for both $\boldsymbol{X}$ and $\boldsymbol{W}$. However, since $\boldsymbol{W}$ is *unobserved*, the investigator is forced to run instead the restricted models in the second column of Table 1.

	Restricted IV regressions (omitting $\boldsymbol{W}$)	Full IV regressions (including $\boldsymbol{W}$)
First-stage	$D = \hat{\theta}_{\text{res}} Z + \boldsymbol{X}\hat{\psi}_{\text{res}} + \hat{\varepsilon}_{d,\text{res}}$	$D = \hat{\theta} Z + \boldsymbol{X}\hat{\psi} + \boldsymbol{W}\hat{\delta} + \hat{\varepsilon}_d$
Reduced-form	$Y = \hat{\lambda}_{\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}}$	$Y = \hat{\lambda} Z + \boldsymbol{X}\hat{\beta} + \boldsymbol{W}\hat{\gamma} + \hat{\varepsilon}_y$
Anderson-Rubin	$Y_{\tau_0} = \hat{\phi}_{\tau_0,\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\tau_0,\text{res}} + \hat{\varepsilon}_{\tau_0,\text{res}}$	$Y_{\tau_0} = \hat{\phi}_{\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\tau_0} + \boldsymbol{W}\hat{\gamma}_{\tau_0} + \hat{\varepsilon}_{\tau_0}$

Table 1: the omitted variable bias problem for instrumental variable estimation.

Our task is thus to characterize how point estimates and confidence intervals for the IV estimate given by these regressions would have changed due to the inclusion of $\boldsymbol{W}$. Since, at their core, all these IV approaches rely on OLS estimation, we can leverage all OVB tools for OLS (Cinelli and Hazlett, 2020) for examining the sensitivity of IV methods.

3 Omitted variable bias with the partial R^2 parameterization

In this section, we review and extend the results of Cinelli and Hazlett (2020) regarding the partial R^2 parameterization of the OVB formula for OLS. In particular, we introduce the notion of *OVB-adjusted* critical values, and show how sensitivity analysis can be performed by simply substituting traditional critical values with the adjusted ones. We also introduce the idea of a set of compatible inferences given bounds on the strength of confounding, and formalize sensitivity statistics for routine reporting as answering an inverse question regarding those sets. These extensions are not only useful for the sensitivity of OLS estimates themselves, but will greatly simplify the generalization of these results to the IV setting. To fix ideas, here we discuss the OVB framework in the context of the reduced form regression coefficient, but the reader should have in mind that all results presented here are algebraic, and thus hold for *any* OLS estimate.

3.1 Sensitivity in an omitted variable bias framework

The OVB framework starts with a target coefficient obtained from a *full* regression equation that the analyst wished she could have estimated (such as those in the third column of Table 1). For concreteness, suppose we are interested in the coefficient $\hat{\lambda}$ of the regression equation of the outcome Y on the instrument Z, adjusting for a set of observed covariates $\boldsymbol{X}$ and a single *unobserved* covariate W (we generalize to multivariate W below),

$$Y = \hat{\lambda} Z + \boldsymbol{X}\hat{\beta} + \hat{\gamma} W + \hat{\varepsilon}_y \tag{22}$$

However, when W is unobserved, estimating the full regression equation is infeasible. Instead, the investigator is forced to estimate the *restricted* model given by

$$Y = \hat{\lambda}_{\text{res}} Z + \boldsymbol{X}\hat{\beta}_{\text{res}} + \hat{\varepsilon}_{y,\text{res}} \tag{23}$$

Where $\hat{\lambda}_{\text{res}}$ and $\hat{\beta}_{\text{res}}$ are the coefficients of the restricted OLS adjusting for Z and $\boldsymbol{X}$ alone, and $\hat{\varepsilon}_{y,\text{res}}$ its corresponding residual. The OVB framework seeks to answer the following question: how do the inferences for λ_{res} from the restricted OLS model (omitting W), compare with the inferences for λ from the full OLS model (adjusting for W)?

3.1.1 Adjusted estimates and standard errors

Let $R^2_{Y\sim W|Z,\boldsymbol{X}}$ denote the partial R^2 of W with Y, after controlling for Z and $\boldsymbol{X}$, and let $R^2_{Z\sim W|\boldsymbol{X}}$ denote the partial R^2 of W with Z after adjusting for $\boldsymbol{X}$. Given the estimates of the restricted model, $\hat{\lambda}_{\text{res}}$ and $\widehat{\text{se}}(\hat{\lambda}_{\text{res}})$, the values $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$ are sufficient to recover $\hat{\lambda}$ and $\widehat{\text{se}}(\hat{\lambda})$ (Cinelli and Hazlett, 2020). More precisely, define $\widehat{\text{bias}} := \hat{\lambda}_{\text{res}} - \hat{\lambda}$ as the difference between the restricted estimate and the full estimate. We then have,

$$|\widehat{\text{bias}}| = \sqrt{\frac{R^2_{Y\sim W|Z,\boldsymbol{X}} R^2_{Z\sim W|\boldsymbol{X}}}{1 - R^2_{Z\sim W|\boldsymbol{X}}}\text{df}} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) = \text{BF}\sqrt{\text{df}} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \tag{24}$$

Where $\text{df} = n - p - 1$ stands for the degrees of freedom of the restricted model actually run. For notational convenience, and to aid interpretation, we define the term $\text{BF} := \sqrt{\frac{R^2_{Y\sim W|Z,\boldsymbol{X}} R^2_{Z\sim W|\boldsymbol{X}}}{1-R^2_{Z\sim W|\boldsymbol{X}}}}$ as the "bias factor" of W, which is the part of the bias solely determined by $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$. Likewise, the standard error of the full model can be recovered with

$$\widehat{\mathrm{se}}(\hat{\lambda}) = \sqrt{\frac{1-R^2_{Y\sim W|Z,\boldsymbol{X}}}{1-R^2_{Z\sim W|\boldsymbol{X}}}\left(\frac{\mathrm{df}}{\mathrm{df}-1}\right)} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) = \mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \tag{25}$$

Where again, for convenience, we define $\mathrm{SEF} := \sqrt{\frac{1-R^2_{Y\sim W|Z,\boldsymbol{X}}}{1-R^2_{Z\sim W|\boldsymbol{X}}}}$ as the "standard error factor" of W, summarizing the factor of the adjusted standard error which is solely determined by the sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$. Note that SEF consists of the square-root of the product of the familiar "variance inflation factor," $1/\left(1-R^2_{Z\sim W|\boldsymbol{X}}\right)$ and what could be labeled the "variance reduction factor," $1-R^2_{Y\sim W|Z,\boldsymbol{X}}$. Cinelli and Hazlett (2020, Sec. 4) provide further discussion. Although simple, Equations 24 and 25 form the basis of a rich set of sensitivity exercises regarding point estimates, standard errors and t-values in terms of sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$.

Multiple unobserved variables. For simplicity of exposition, throughout the text we will usually refer to a single omitted variable W. These results, however, can be used for performing sensitivity analyses considering multiple omitted variables $\boldsymbol{W} = [W_1, W_2, \ldots, W_n]$, and thus also non-linearities and functional form mispecification of observed variables. In such cases, barring an adjustment in the degrees of freedom, the equations are conservative, and reveal the maximum biases that a multivariate $\boldsymbol{W}$ with such pair of partial R^2 values could cause (Cinelli and Hazlett, 2020).

3.1.2 Adjusted lower and upper limits of confidence intervals

We now closely examine how the confidence interval of a regression coefficient changes due to the inclusion of W. Traditional confidence intervals account for sampling uncertainty, and are constructed by multiplying the standard error of the coeficient by a critical value (for example, 1.96 for a 95% confidence level). We show that replacing this traditional critical value with an *OVB-adjusted critical value*, which we introduce here, accounts for both sampling uncertainty and systematic biases due to the omission of W. Although simple, this perspective will prove useful for deriving and understanding OVB-type results for OLS in general, and for instrumental variables in particular, such as in the Anderson-Rubin framework of Section 4.

Specifically, let $t^*_{\alpha,\mathrm{df}-1}$ denote the critical value for a standard t-test with significance level α and $\mathrm{df}-1$ degrees of freedom. Now let $\mathrm{LL}_{1-\alpha}(\lambda)$ be the lower limit and $\mathrm{UL}_{1-\alpha}(\lambda)$ be the upper limit of a $1-\alpha$ confidence interval for λ in the full model, i.e.,

$$\mathrm{LL}_{1-\alpha}(\lambda) := \hat{\lambda} - t^*_{\alpha,\mathrm{df}-1} \times \widehat{\mathrm{se}}(\hat{\lambda}), \quad \mathrm{UL}_{1-\alpha}(\lambda) := \hat{\lambda} + t^*_{\alpha,\mathrm{df}-1} \times \widehat{\mathrm{se}}(\hat{\lambda}), \tag{26}$$

Considering the direction of the bias that further reduces the lower limit, or, alternatively, a direction that further increases the upper limit, Equations 24 and 25 imply that both quantities can be written as a function of the restricted estimates and a new multiplier

$$\mathrm{LL}_{1-\alpha}(\lambda) = \hat{\lambda}_{\mathrm{res}} - t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}), \quad \mathrm{UL}_{1-\alpha}(\lambda) = \hat{\lambda}_{\mathrm{res}} + t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} \times \widehat{\mathrm{se}}(\hat{\lambda}_{\mathrm{res}}) \tag{27}$$

where $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ stands for the *OVB-adjusted critical value*

$$t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2} := \mathrm{SEF}\sqrt{\mathrm{df}/(\mathrm{df}-1)} \times t^*_{\alpha,\mathrm{df}-1} + \mathrm{BF}\sqrt{\mathrm{df}}. \tag{28}$$

The subscript $\boldsymbol{R}^2 = \{R^2_{Y\sim W|Z,\boldsymbol{X}}, R^2_{Z\sim W|\boldsymbol{X}}\}$ conveys the fact that it depends on both sensitivity parameters. Inferences using the critical threshold $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ now take into account both sampling uncertainty as well as systematic biases due to confounding.[7] The adjusted critical value $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ *uniquely determines* the extreme points of inference for λ that one could obtain after adjusting for an omitted variable W with a given pair of partial R^2. Finally, given the equivalence of hypothesis testing and confidence intervals, it thus follows that, given any hypothetical strength of W, to test the general null hypothesis of a change of $(100 \times q^*)\%$ of the current estimate $\hat{\lambda}_{\mathrm{res}}$, it suffices to rescale the original t-value by q^* and compare this to the adjusted critical threshold $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$.

3.1.3 Compatible inferences given bounds on partial R^2

Given hypothetical values for $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$, the previous results allow us to determine the exact changes in inference regarding a parameter of interest due to the inclusion of W with such strength. Often, however, the analyst does not know the exact strength of omitted variables, and wishes to investigate the *worst* possible inferences that could be induced by a confounder with bounded strength, for instance, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\,\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\,\max}_{Z\sim W|\boldsymbol{X}}$. That is, we wish to find the maximum adjusted critical value due to an omitted variable W with *at most* such strength. Writing $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ as a function of the sensitivity parameters $R^2_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}}$, we

[7] For a numerical example, consider a case with 100 degrees of freedom and a significance level of $\alpha = 5\%$. The traditional critical value, assuming zero confounding biases, is $t^*_{.05,100} \approx 1.98$. If we now allow for an omitted variable with strength given by $R^2_{Y\sim W|Z,\boldsymbol{X}} = R^2_{Z\sim W|\boldsymbol{X}} = .1$, this leads to an increased OVB-adjusted critical value of $t^{\dagger}_{.05,100,.1,.1} \approx 3.05$. Further note $t^{\dagger}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ *increases* the larger the sample size—for instance, if the degrees of freedom were instead 1,000, the adjusted critical value would increase to approximately 5.30.

solve the maximization problem

$$\max_{R^2_{Y\sim W|Z,\boldsymbol{X}},R^2_{Z\sim W|\boldsymbol{X}}} t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2} \quad \text{s.t.} \quad R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\max}_{Y\sim W|Z,\boldsymbol{X}}, \quad R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\max}_{Z\sim W|\boldsymbol{X}} \tag{29}$$

Note that, although this maximum is often reached at the extrema of both coordinates, this is not always the case. Due to the variance reduction factor, sometimes increasing $R^2_{Y\sim W|Z,\boldsymbol{X}}$ reduces the standard error more than enough to compensate for the increase in bias, resulting in tighter confidence intervals. Denoting the solution to the optimization problem in expression (29) as $t^{\dagger\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2}$, the *most extreme possible* lower and upper limits after adjusting for W are given by

$$\text{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \hat{\lambda}_{\text{res}} - t^{\dagger\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}), \quad \text{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2} = \hat{\lambda}_{\text{res}} + t^{\dagger\max}_{\alpha,\text{df}-1,\boldsymbol{R}^2} \times \widehat{\text{se}}(\hat{\lambda}_{\text{res}}) \tag{30}$$

And the interval composed of such limits

$$\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda) = \left[\text{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda), \quad \text{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)\right] \tag{31}$$

retrieves *all inferences for λ which are compatible with an omitted variable with such strengths.* In other words, without imposing further constraints on W, for any value λ_0 inside $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$, we can find a W such that $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq R^{2\max}_{Y\sim W|Z,\boldsymbol{X}}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\max}_{Z\sim W|\boldsymbol{X}}$ and the confidence interval for λ after adjusting for W includes λ_0. Moreover, if the true partial R^2 of W lies within the posited bounds, then $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ is the union of all confidence intervals that would be obtained with confounders with that strength or less, and thus constitutes itself a confidence interval with *at least* $1-\alpha$ coverage.

3.2 Sensitivity statistics for routine reporting

Given plausibility judgments on the strength of the omitted variable W, the previous results allow one to perform detailed sensitivity analyses by examining the exact (or the worst possible) inferences that one would have obtained with the full regression model. However, widespread adoption of sensitivity analysis benefits from simple and interpretable sensitivity statistics that quickly convey the overall robustness of an estimate. To that end, Cinelli and Hazlett (2020) proposed two sensitivity metrics for routine reporting: (i) the partial R^2 of the dependent variable Z with the independent variable Y, $R^2_{Y\sim Z|\boldsymbol{X}}$; and, (ii) the *robustness value* (RV). We generalize the notion of a partial R^2 as a measure of robustness to extreme scenarios, by introducing the *extreme robustness value* (XRV), for which the partial R^2 is a special case. We also recast both robustness metrics as a solution to an "inverse" question regarding the interval of compatible inferences, $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$—that is, given a threshold of inference for λ deemed to be of scientific importance (say, zero), what is the *minimum* strength of the sensitivity parameters $\boldsymbol{R}^2$ that could lead $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ to include such threshold? This framework facilitates extending these metrics to other contexts, including the IV setting.

3.2.1 The extreme robustness value

One benefit of the partial R^2 parameterization is that the parameter $R^2_{Y\sim W|Z,\boldsymbol{X}}$ can be left completely unconstrained. In other words, in the optimization problem in expression 29, one can set the bound for $R^2_{Y\sim W|Z,\boldsymbol{X}}$ to its trivial bound of 1, and this still results in non-trivial bounds on the set of possible inferences. By contrast, if $R^2_{Z\sim W|\boldsymbol{X}}$ were left unconstrained, as it approaches unity, the confidence interval will eventually run from minus to plus infinity. This leads to our first inverse question: what is the *bare minimum* strength of association of the omitted variable W with the independent variable Z that could bring its estimated coefficient to zero, or to a region where it is no longer statistically different than zero (or any other threshold of substantive interest)?

To answer this question, we can see $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ as a function of the bound $R^{2\max}_{Z\sim W|\boldsymbol{X}}$ alone, obtained from maximizing the adjusted critical value in expression 29 where: (i) the parameter $R^2_{Y\sim W|Z,\boldsymbol{X}}$ is left completely unconstrained (i.e, $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq 1$); and, (ii) the parameter $R^2_{Z\sim W|\boldsymbol{X}}$ is bounded by XRV (i.e, $R^{2\max}_{Z\sim W|\boldsymbol{X}} \leq \mathrm{XRV}$). The *Extreme Robustness Value* $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ is defined as the greatest lower bound XRV such that the null hypothesis that a change of $(100 \times q)\%$ of the original estimate, $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\text{res}}$, is not rejected at the α level,

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) := \inf\Big\{\mathrm{XRV};\ (1-q^*)\hat{\lambda}_{\text{res}} \in \mathrm{CI}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\lambda)\Big\} \tag{32}$$

The solution to this problem gives,

$$\mathrm{XRV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \ f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \dfrac{f^2_{q^*}(\lambda) - f^{*2}_{\alpha,\mathrm{df}-1}}{1+f^2_{q^*}(\lambda)}, & \text{otherwise.} \end{cases} \tag{33}$$

Where $f_{q^*}(\lambda) := q^*|f_{Y\sim Z|\boldsymbol{X}}|$ (here $f_{Y\sim Z|\boldsymbol{X}}$ stands for the partial Cohen's f and we define $f^*_{\alpha,\mathrm{df}\,-1} := t^*_{\alpha,\mathrm{df}\,-1}/\sqrt{\mathrm{df}\,-1}$).[8] Note $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ can be interpreted as an "adjusted partial R^2" of Z with Y. To see why, let us first consider the case of the minimal strength to bring the point estimate ($\alpha = 1$) to

[8] Cohen's f^2 can be written as $f^2 = R^2/(1-R^2)$.

exactly zero ($q^* = 1$). We then have that $f^*_{\alpha=1,\mathrm{df}-1} = 0$ and $f^2_{q^*=1}(\lambda) = f^2_{Y\sim Z|\boldsymbol{X}}$, resulting in

$$\mathrm{XRV}_{q^*=1,\alpha=1}(\lambda) = \frac{f^2_{Y\sim Z|\boldsymbol{X}}}{1+f^2_{Y\sim Z|\boldsymbol{X}}} = R^2_{Y\sim Z|\boldsymbol{X}} \tag{34}$$

This recovers the result of Cinelli and Hazlett (2020), and shows that, for an omitted variable W to bring down the estimated coefficient to zero, it needs to explain at least as much residual variation of the dependent variable Z, as Z explains of the independent variable Y. For the general case, we simply perform two adjustments that dampens the "raw" partial R^2 of Z with Y. First we adjust it by the proportion of reduction deemed to be problematic q^* through $f_{q^*} = q^*|f_{Y\sim Z|\boldsymbol{X}}|$; next, we subtract the threshold for which statistical significance is lost at the α level (via $f^{*2}_{\alpha,\mathrm{df}-1}$).

The extreme robustness value is thus the equivalent of a "Cornfield condition" (Cornfield et al., 1959) for OLS estimates, and delineates the bare minimum strength of omitted variables necessary to overturn a certain conclusion—if W cannot explain at least $\mathrm{XRV}_{q^*,\alpha}(\lambda)$ of the residual variation of Z, then such variable *is not* strong enough to bring about a bias of $(100 \times q^*)\%$ on the original estimate, at the significance level of α. One could also allow the maximum bound on $R^2_{Y\sim W|Z,\boldsymbol{X}}$ to be a less extreme value, different than unity—the solution to this problem is provided in the appendix.

3.2.2 The robustness value

Placing no constraints on the association of the omitted variable W with Y may be too conservative an exercise. An alternative measure of robustness of the OLS estimate is to consider the minimal strength of association that the omitted variable needs to have, *both* with Z and Y, so that a $1-\alpha$ confidence interval for λ will include a change of $(100 \times q^*)\%$ of the current restricted estimate.

Write $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\lambda)$ as a function of both bounds varying simultaneously, that is, construct $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\lambda)$ by maximizing the adjusted critical value with bounds given by $R^2_{Y\sim W|Z,\boldsymbol{X}} \leq \mathrm{RV}$ and $R^2_{Z\sim W|\boldsymbol{X}} \leq \mathrm{RV}$. The *robustness value* $\mathrm{RV}_{q^*,\alpha}(\lambda)$ for not rejecting the null hypothesis that $H_0 : \lambda = (1-q^*)\hat{\lambda}_{\mathrm{res}}$, at the significance level α, is defined as

$$\mathrm{RV}_{q^*,\alpha}(\lambda) := \inf\left\{\mathrm{RV};\ (1-q^*)\hat{\lambda}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\lambda)\right\} \tag{35}$$

We then have that,

$$\mathrm{RV}_{q^*,\alpha}(\lambda) = \begin{cases} 0, & \text{if } \; f_{q^*}(\lambda) \leq f^*_{\alpha,\mathrm{df}-1} \\ \frac{1}{2}\left(\sqrt{f^4_{q^*,\alpha}(\lambda) + 4f^2_{q^*,\alpha}(\lambda)} - f^2_{q^*,\alpha}(\lambda)\right), & \text{if } \; f^*_{\alpha,\mathrm{df}-1} < f_{q^*}(\lambda) < f^{*-1}_{\alpha,\mathrm{df}-1} \\ \mathrm{XRV}_{q^*,\alpha}(\lambda), & \text{otherwise.} \end{cases} \tag{36}$$

Where $f_{q^*,\alpha}(\lambda) := q^*|f_{Y\sim Z|\boldsymbol{X}}| - f^*_{\alpha,\mathrm{df}\,-1}$. In the appendix we show the conditions of Equation 36 are equivalent to those first derived in Cinelli and Hazlett (2020), with the advantage of being simpler to verify. The first case occurs when the confidence interval already includes $(1-q^*)\hat{\lambda}_{\mathrm{res}}$ or the mere change of one degree of freedom achieves this. The second case occurs when both associations of W reach the bound. Finally, in the last case the solution is an interior point—this happens when the bound is large enough such that the constraint on the association with the outcome is not binding; in this case the RV reduces to the XRV.

The robustness value offers a simple interpretable measure that summarizes the strength of omitted variables necessary to change the estimate in problematic ways. If W explains $\mathrm{RV}_{q^*,\alpha}(\lambda)$ of the residual variance of both Z and Y, then such variable is sufficiently strong to bring about a $(100 \times q)\%$ change in the estimate at the significance level of α, while any omitted variable that does not explain $\mathrm{RV}_{q^*,\alpha}(\lambda)$ of the residual variance, neither of Z nor of Y, is not sufficiently strong to do so.

3.3 Bounding the strength of the omitted variable using observed covariates

One further result is required before turning to the sensitivity of IV estimates. Let the omitted variable W be orthogonal to the observed covariates $\boldsymbol{X}$, ie., $W \perp \boldsymbol{X}$ (or, equivalently, consider the part of W not linearly explained by $\boldsymbol{X}$), and let X_j be a specific covariate of the set $\boldsymbol{X}$. Now define

$$k_Z := \frac{R^2_{Z\sim W|\boldsymbol{X}_{-j}}}{R^2_{D\sim X_j|\boldsymbol{X}_{-j}}}, \qquad k_Y := \frac{R^2_{Y\sim W|Z,\boldsymbol{X}_{-j}}}{R^2_{Y\sim X_j|Z\boldsymbol{X}_{-j}}}. \tag{37}$$

where $\boldsymbol{X}_{-j}$ represents the vector of covariates $\boldsymbol{X}$ excluding X_j. These new parameters, k_Z and k_Y, stand for how much "stronger" W is relatively to the observed covariate X_j in terms of residual variation explained of Z and Y. As shown in Cinelli and Hazlett (2020), given k_Z and k_Y, our previous sensitivity parameters can be rewritten as

$$R^2_{Z\sim W|\boldsymbol{X}} = k_Z f^2_{Z\sim X_j|\boldsymbol{X}_{-j}}, \qquad R^2_{Y\sim W|D,\boldsymbol{X}} \leq \eta^2 f^2_{Y\sim X_j|Z,\boldsymbol{X}_{-j}} \tag{38}$$

where η is a function of both parameters k_Y, k_Z and $R^2_{Z\sim X_j|\boldsymbol{X}_{-j}}$.

These equations allow investigators to leverage knowledge of *relative importance* of variables to make plausibility judgments regarding sensitivity parameters. For instance, if researchers have domain knowledge to argue that a certain observed covariate X_j is supposed to be a strong determinant of the treatment and the outcome variation, and that the omitted variable W is not likely to explain as much residual variance of Z and Y as that observed covariate, these results can be used to set plausible bounds on the maximum bias that W could cause. In the appendix we further show these same inequalities are conservative when a set of covariates $\boldsymbol{X}_j$ is used for benchmarking, by simply using the joint partial R^2 of those variables.

4 An omitted variable bias framework for the sensitivity of IV

Having established the tools for analyzing the sensitivity of conventional OLS estimates, we are now in a position to extend them for instrumental variable analyses. As explained, an OVB-approach to sensitivity analysis of IV estimates begins by assuming that the researcher measured and included observed covariates $\boldsymbol{X}$, but would also have liked to adjust for W in order for the IV conditions to hold. In this section, we first show how assessing the sensitivity of the reduced form and first stage is sufficient to draw valuable conclusions regarding the sensitivity of IV. We then show how to construct a straightforward sensitivity analysis framework for the IV estimate within the Anderson-Rubin approach, allowing one to investigate the sensitivity of point estimates, standard errors, and confidence intervals.

4.1 Sensitivity analysis of the reduced-form and of the first-stage

The recent literature on instrumental variables places strong emphasis on the first stage and the reduced form estimates. Not only are the first stage and reduced form often substantively meaningful on their own, but their critical examination plays an important role for motivating the causal story behind a particular instrumental variable. For example, in the "local average treatment effect" interpretation of the IV estimand, *both* the first stage and the reduced form must be unconfounded so that the resulting estimate can be interpreted as the average causal effect among compliers (Angrist et al., 1996). Therefore, beyond a means to the final IV estimate, researchers are advised to report and to interpret the first stage and the reduced form by, for example, assessing whether their magnitudes, patterns and signs are in accordance to the postulated mechanisms that justify the choice of instrument (Angrist and Krueger, 2001; Angrist and Pischke, 2009; Imbens, 2014; Angrist and Pischke, 2014; Imbens and Rubin, 2015). While investigating these separate regressions, researchers can deploy all results discussed on the previous section.

Fortunately, sensitivity analyses of the first stage and reduced form also provide answers to many pivotal sensitivity questions regarding the IV estimate itself. In particular, as we shall see, if the investigator is interested in assessing the strength of confounding needed to bring the IV point estimate to zero, or to not reject the null hypothesis of zero effect, then the results of the sensitivity analysis of the reduced form is all that is needed. If interest lies in also determining whether the IV estimate could be arbitrarily large, then the sensitivity of the first stage must also be assessed, as confounders capable of changing the direction of the first stage could also lead to unbounded IV estimates.

4.1.1 What the reduced-form and first-stage reveal about the IV point estimate

First let us consider the sensitivity of the point estimate, since they numerically coincide for all estimators. That is, recall that all the three main estimators under consideration are algebraically equivalent, and can be seen as the ratio of the reduced form and the first stage coefficients,

$$\hat{\tau} := \hat{\tau}_{\text{ILS}} = \hat{\tau}_{\text{2SLS}} = \hat{\tau}_{AR} = \frac{\hat{\lambda}}{\hat{\theta}} \tag{39}$$

This simple algebraic fact allows us to draw two important conclusions regarding the sensitivity of $\hat{\tau}$ from the sensitivity of $\hat{\lambda}$ and $\hat{\theta}$ alone. First, confounders can bring the IV point estimate to zero *if, and only if,* they can bring the reduced-form point estimate to zero. Therefore, if sensitivity analysis of the reduced form reveals that omitted variables are not strong enough to explain away $\hat{\lambda}$, then they also cannot explain away the IV point estimate. Or, more worrisome, if analysis reveals that it takes weak confounding or side-effects to explain away $\hat{\lambda}$, the same holds for the IV estimate $\hat{\tau}$. In sum, for any of the IV estimators considered here, to assess the strength of confounding needed to bring the IV point estimate to zero, all one needs to do is to perform a sensitivity analysis on the reduced form regression coefficient. Second, if we cannot rule out confounders that are sufficiently strong to *change the sign of the first stage* point estimate, then we also cannot rule out that the IV point estimate could be *arbitrarily large* in either direction, even if not exactly equal to zero. Therefore, whenever we are interested in biases as large *or larger* than a certain amount, then *the robustness of the first-stage point estimate puts an upper bound on the robustness of the IV point estimate.*

4.1.2 What the reduced-form and first-stage reveal about IV hypothesis tests

Contrary to the point estimate, different estimation approaches can lead to different conclusions regarding how omitted variables would have changed inferences. Let us start by examining the Anderson-Rubin/Fieller approach, as not only it has nominal coverage regardless of instrument strength, but its conclusions match the intuition of current guidelines when assessing the first-stage and reduced form estimates (Angrist and Krueger, 2001; Angrist and Pischke, 2009, 2014).

Consider again the IV estimand

$$\tau = \frac{\lambda}{\theta}$$

Note that the same arguments we used before for the estimator hold for the estimand. Logically, provided the ratio is well defined ($\theta \neq 0$), we have that $\tau = 0 \iff \lambda = 0$. Therefore, a test of the null hypothesis $H_0 : \lambda = 0$ in the reduced form regression is *logically equivalent* to a test of the null hypothesis $H_0 : \tau = 0$ for the IV estimand. Similarly, for a fixed λ, if we cannot rule out that θ is arbitrarily close to zero in either direction, then, logically, we also cannot rule out that τ is arbitrarily large in either direction—a test for the null hypothesis $H_0 : \theta = 0$ is thus *logically equivalent* to testing whether arbitrarily large sizes for τ can be ruled out.

One virtue of the Anderson-Rubin/Fieller approach is that it is coherent with respect to these logical implications. Recall the Anderson-Rubin test for the null hypothesis $H_0 : \tau = \tau_0$ is based on the test of $H_0 : \phi_{\tau_0} = 0$. The point estimate and (estimated) standard error for $\hat{\phi}_{\tau_0}$ are given by

$$\hat{\phi}_{\tau_0} = \frac{\text{cov}(Y_{\tau_0}^{\perp \boldsymbol{X},W}, Z^{\perp \boldsymbol{X},W})}{\text{var}(Z^{\perp \boldsymbol{X},W})}, \qquad \widehat{\text{se}}(\hat{\phi}_{\tau_0}) = \frac{\text{sd}(Y_{\tau_0}^{\perp Z,\boldsymbol{X},W})}{\text{sd}(Z^{\perp \boldsymbol{X},W})}\sqrt{\frac{1}{\text{df}-1}} \tag{40}$$

Which, as we have seen, can be expressed in terms of the first-stage and reduced-form estimates

$$\hat{\phi}_{\tau_0} = \hat{\lambda} - \tau_0\hat{\theta}, \qquad \widehat{\text{se}}(\hat{\phi}_{\tau_0}) = \sqrt{\widehat{\text{var}}(\hat{\lambda}) + \tau_0^2\widehat{\text{var}}(\hat{\theta}) - 2\tau_0\widehat{\text{cov}}(\hat{\lambda},\hat{\theta})} \tag{41}$$

Testing $H_0 : \phi_{\tau_0} = 0$ requires comparing the t-value for $\hat{\phi}_{\tau_0}$ with a critical threshold $t^*_{\alpha,\text{df}-1}$, and the null hypothesis is not rejected if $|t_{\hat{\phi}_{\tau_0}}| \leq t^*_{\alpha,\text{df}-1}$. Squaring and rearranging terms we obtain the quadratic inequality which must hold for non-rejection:

$$\underbrace{\left(\hat{\theta}^2 - \widehat{\text{var}}(\hat{\theta}) \times t^{*2}_{\alpha,\text{df}-1}\right)}_{a}\tau_0^2 + \underbrace{2\left(\widehat{\text{cov}}(\hat{\lambda},\hat{\theta}) \times t^{*2}_{\alpha,\text{df}-1} - \hat{\lambda}\hat{\theta}\right)}_{b}\tau_0 + \underbrace{\left(\hat{\lambda}^2 - \widehat{\text{var}}(\hat{\lambda}) \times t^{*2}_{\alpha,\text{df}-1}\right)}_{c} \leq 0 \tag{42}$$

When considering the null hypothesis $H_0 : \tau_0 = 0$, only the term c remains, and c is less or equal to zero if, and only if, one cannot reject the null hypothesis $H_0 : \lambda = 0$ in the reduced form regression.

The Anderson-Rubin approach thus comports with the recommendation of Angrist and Krueger (2001) that "if you can't see the causal relation of interest in the reduced form, it's probably not there." Also note that arbitrarily large values for τ_0 will satisfy the inequality in Equation 42 if, and only if, $a < 0$, meaning that we cannot reject the null hypothesis $H_0 : \theta = 0$ in the first stage regression. This supports the recommendation that, if one is unsure about the direction of the first-stage, it is likely that very little can be said about the magnitude of the IV estimate.

Within the Anderson-Rubin framework, we thus reach analogous conclusions regarding hypothesis testing as those regarding the point estimate: (i) when interest lies in the *zero null hypothesis, the sensitivity of the reduced form is exactly the sensitivity of the IV*—no other analyses are needed. Confounders or side-effects sufficiently strong to bring the reduced form to a region where it is not statistically different than zero can also bring the IV estimate to a region where it is not statistically different than zero, and only confounders with such strength are capable of doing so; (ii) if one is interested in biases of a certain amount, *or larger,* then the sensitivity of the first-stage needs also to be assessed. Specifically, for any null hypothesis of interest, confounding that makes the first stage not statistically significant from zero will also cause us to not reject values arbitrarily worse than that specific null hypothesis for the IV estimate.[9]

As it is well known, it is not uncommon for frequentist statistical tests to lead to logically incoherent decisions (Gabriel, 1969; Schervish, 1996; Patriota, 2013; Fossaluza et al., 2017). While inferences made in the Anderson-Rubin framework have the expected behavior in this setting, inferences using the asymptotic approximations of ILS or 2SLS, however, do not necessarily comply with these logical expectations. Cases can be found for ILS and 2SLS where, for instance, one fails to reject the null hypothesis $H_0 : \lambda = 0$, yet still rejects the null hypothesis $H_0 : \tau = 0$ (and vice-versa).[10] Such claims could lead researchers to contradictions if they are not careful in interpreting the results of these significance tests, and they also do not conform to current guidelines for interpreting the first-stage and reduced-form regressions (Angrist and Pischke, 2009).

4.2 Sensitivity analysis of the IV in the Anderson-Rubin approach

We now apply the OVB framework for assessing the sensitivity of the IV estimate directly. We focus on the Anderson-Rubin approach for this task because: (i) as a single level regression (for a given choice of τ_0), all OVB results from Section 3 can be directly applied; (ii) relatedly, as we shall see, we can perform sensitivity analysis with only two interpretable sensitivity parameters; (iii) it has correct

[9] Similar observations regarding the importance of the robustness of the first-stage for hidden biases have been made before in the context of randomization inference (Small and Rosenbaum, 2008; Rosenbaum, 2017).

[10] See appendix for numerical examples.

test size regardless of "instrument strength" (the strength of the first stage); and, (iv) as explained in the last section, its conclusions conform to current recommendations regarding the interpretation of the first stage and reduced form regressions.

4.2.1 Sensitivity of the t-value

We begin by examining the sensitivity of the t-value for testing a specific null hypothesis $H_0 : \tau = \tau_0$, as this is the simplest and most straightforward application of the tools of Section 3. Recall that, in the Anderson-Rubin framework, a test for the null hypothesis $H_0 : \tau = \tau_0$ is simply a test for the null hypothesis $H_0 : \phi_{\tau_0} = 0$ in the regression of Y_{τ_0} on the instrument Z and covariates $\boldsymbol{X}$ and W. Therefore, a standard OLS sensitivity analysis for testing the null hypothesis $H_0 : \phi_{\tau_0} = 0$ on the Anderson-Rubin regression gives the desired results for $H_0 : \tau = \tau_0$.

In detail, a sensitivity analysis for the null hypothesis H_0 that the IV estimate τ equals some τ_0 can be performed as follows:

1. Construct the "putative potential outcome" at the null value τ_0, $Y_{\tau_0} = Y - \tau_0 D$;

2. Run the OLS model $Y_{\tau_0} = \hat{\phi}_{\text{res},\tau_0} Z + \boldsymbol{X}\hat{\beta}_{\text{res},\tau_0} + \hat{\varepsilon}_{\tau_0,\text{res}}$;

3. Perform sensitivity analysis on the observed t-value, $t_{\hat{\phi}_{\text{res},\tau_0}}$.

This procedure tells the users how confounding no worse than $\mathbf{R}^2 = \{R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}},\ R^2_{Z\sim W|\boldsymbol{X}}\}$ would alter inferences for the test $H_0 : \tau = \tau_0$. While the meaning of the parameter $R^2_{Z\sim W|\boldsymbol{X}}$ is straightforward—the share of residual variation of the instrument explained by the confounder—the meaning of the parameter $R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}}$ is less familiar. Recall, however, that under the null hypothesis $H_0 : \tau = \tau_0$, we have that $Y_{\tau_0} = Y_0$ (assuming a constant treatment effects model), and thus $R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}}$ can be interpreted as the share of residual variance of the *untreated potential outcome* Y_0 explained by W.

Bounds on the strength of confounders. The bounds discussed in Section 3.3 work without modification in the Anderson-Rubin setting. The plausibility judgment one is making here is that of how strong unobserved confounders or side-effects are, relative to observed covariates, in explaining the residual variance of the untreated potential outcome and of the instrument, *under the null hypothesis* that $H_0 : \tau = \tau_0$. Since the judgment is made under a specific null, the bounds will be different when testing different null hypothesis. Therefore, it may be useful to compute bounds under a slightly more *conservative* assumption, that confounders are no stronger than (a multiple of) the *maximum* explanatory power of an observed covariate, over all possible values of τ_0—this has the useful property of giving the same bounds for any hypothesis test. We provide such formulation in the appendix.

4.2.2 Compatible inferences given bounds on partial R^2

Instead of assessing the sensitivity of the test statistic for specific null hypothesis, investigators may be interested in recovering the whole set of inferences compatible with plausibility judgments on the maximum strength of confounding, and a chosen significance level. As discussed in Section 2, for a critical threshold $t^*_{\alpha,\mathrm{df}-1}$, the confidence interval for τ in the Anderson-Rubin framework is given by

$$\mathrm{CI}_{1-\alpha}(\tau) = \{\tau_0;\ t^2_{\phi_{\tau_0}} \leq t^{*2}_{\alpha,\mathrm{df}-1}\} \tag{43}$$

Now consider bounds on sensitivity parameters $R^2_{Y_{\tau_0}\sim W|Z,\boldsymbol{X}} \leq R^{2\max}_{Y_0\sim W|Z,\boldsymbol{X}}$ (which should be judged to hold *regardless* of the value of τ_0) and $R^2_{Z\sim W|\boldsymbol{X}} \leq R^{2\max}_{Z\sim W|\boldsymbol{X}}$. Let $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$ denote the maximum OVB-adjusted critical value under the posited bounds on the strength of W. The set of compatible inferences for τ, $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ is then simply given by

$$\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau) = \left\{\tau_0;\ t^2_{\hat{\phi}_{\mathrm{res},\tau_0}} \leq \left(t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right\} \tag{44}$$

This interval can be found analytically using the same inequality as in Equation 42, now with the parameters of the restricted regression actually run, and the traditional critical value replaced by the OVB-adjusted critical value $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$

$$\begin{aligned}\underbrace{\left(\hat{\theta}^2_{\mathrm{res}} - \widehat{\mathrm{var}}(\hat{\theta}_{\mathrm{res}}) \times \left(t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right)}_{a}\tau_0^2 + \underbrace{2\left(\widehat{\mathrm{cov}}(\hat{\lambda}_{\mathrm{res}},\hat{\theta}_{\mathrm{res}}) \times \left(t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2 - \hat{\lambda}_{\mathrm{res}}\hat{\theta}_{\mathrm{res}}\right)}_{b}\tau_0 \\ + \underbrace{\left(\hat{\lambda}^2_{\mathrm{res}} - \widehat{\mathrm{var}}(\hat{\lambda}_{\mathrm{res}}) \times \left(t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}\right)^2\right)}_{c} \leq 0\end{aligned} \tag{45}$$

Solving the quadratic equation analytically recovers the set of compatible inferences for τ. Note that users can easily compute $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ with any software that computes Anderson-Rubin or Fieller's confidence intervals by simply providing the modified critical threshold $t^{\dagger\max}_{\alpha,\mathrm{df}-1,\boldsymbol{R}^2}$.

It is now useful to discuss the possible shapes of $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}$ as this will help understanding the robustness values for IV we derive next. Let $\mathbf{r} = \{r_{\min}, r_{\max}\}$ denote the roots of the quadratic equation, which can be written as $\mathbf{r} = -b \pm \sqrt{\Delta}/2a$, with $\Delta = b^2 - 4ac$. If $a > 0$ (i.e, we have a statistically significant first-stage), the quadratic equation will be convex, and thus only the values between the roots will be non-positive. This leads to the connected confidence interval $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = [r_{\min}, r_{\max}]$. When $a < 0$, the curve is concave and, as we discussed, this leads to unbounded confidence intervals. Here we have two sub-cases: (i) when $\Delta < 0$, the quadratic curve never touches zero, and thus the confidence interval is simply the whole real line $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = (-\infty, +\infty)$; when $\Delta > 0$ the confidence interval will be union of two disjoint intervals $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2} = (-\infty, r_{\min}] \cup [r_{\max}, +\infty)$.[11]

4.2.3 Sensitivity statistics for routine reporting

Armed with the notion of a set of compatible inferences for IV, $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$, we are now able to formally define (extreme) robustness values for instrumental variable estimates.

Extreme robustness values for IV. The extreme robustness value for the IV estimate, $\text{XRV}_{q^*,\alpha}(\tau)$, is defined as the minimum strength of association of confounders with the instrument so that we cannot reject a reduction of $(100 \times q^*)\%$ of the original IV estimate; that is,

$$\text{XRV}_{q^*,\alpha}(\tau) := \inf\left\{\text{XRV};\ (1-q^*)\hat{\tau}_{\text{res}} \in \text{CI}^{\max}_{1-\alpha,1,\text{XRV}}(\tau)\right\} \tag{46}$$

It then follows immediately from Equation 44 that

$$\text{XRV}_{q^*,\alpha}(\tau) = \text{XRV}_{1,\alpha}(\phi_{\tau^*}) \tag{47}$$

where $\tau^* = (1-q^*)\hat{\tau}_{\text{res}}$. As in the general case, the extreme robustness value can be interpreted as a "dampened" partial R^2 of the instrument Z with the "putative" untreated potential outcome Y_{τ_0}. Also of interest is the special case of the minimum strength to bring the IV estimate to a region where it is no longer statistically different than zero ($q^* = 1$), in which we obtain $\text{XRV}_{1,\alpha}(\tau) = \text{XRV}_{1,\alpha}(\lambda)$. That is, for the null hypothesis of $H_0 : \tau = 0$, the extreme robustness value of the IV estimate equals the extreme robustness value of the reduced form estimate, as we discussed in the last section.

The $\text{XRV}_{q^*,\alpha}(\tau)$ computes the minimal strength of W required to not reject a particular null hypothesis of interest. We might be interested, instead, in asking about the minimal strength of omitted variables to not reject a specific value *or worse*. When confidence intervals are connected, such as the case of standard OLS, the two notions coincide. But in the Anderson-Rubin case, as we have seen, confidence intervals for the IV estimate can sometimes consist of disjoint intervals. Therefore, let the upper and lower limits of $\text{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ be $\text{LL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ and $\text{UL}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau)$ respectively. The extreme robustness value $\text{XRV}_{\geq q^*,\alpha}(\tau)$ for the IV estimate is defined as the minimum

[11] See Mehlum (2020) for an intuitive graphical characterization of Fieller's solutions using polar coordinates.

strength of association that confounders need to have with the instrument so that we cannot reject a change of $(100 \times q^*)\%$ *or worse* of the original IV estimate;

$$\mathrm{XRV}_{\geq q^*,\alpha}(\tau) := \inf\left\{\mathrm{XRV};\ (1-q^*)\hat{\tau}_{\mathrm{res}} \in \left[\mathrm{LL}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\tau),\quad \mathrm{UL}^{\max}_{1-\alpha,1,\mathrm{XRV}}(\tau)\right]\right\} \tag{48}$$

Now note that, whenever $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{df}-1}(\tau)$ is connected, we must have that $\mathrm{XRV}_{\geq q^*,\alpha}(\tau) = \mathrm{XRV}_{q^*,\alpha}(\tau)$. On the other hand, recall that $\mathrm{CI}^{\max}_{1-\alpha,\mathrm{df}-1}(\tau)$ will be disjoint only if $t^2_{\hat{\theta}_{\mathrm{res}}} \leq (t^{\dagger\max}_{\alpha,\mathrm{df}-1})^2$, which is precisely the condition for the extreme robustness value of the first-stage. Therefore,

$$\mathrm{XRV}_{\geq q^*,\alpha}(\tau) = \min\{\mathrm{XRV}_{1,\alpha}(\phi_{\tau^*}),\quad \mathrm{XRV}_{1,\alpha}(\theta)\} \tag{49}$$

This corroborates our previous conclusion that, when we are interested in biases as large or larger than a certain amount, the robustness of the IV estimate is bounded by the robustness of the first-stage.

Robustness values for IV. The definitions for the IV robustness value follow the same logic discussed above, but now considering both bounds on $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}$ varying simultaneously. That is,

$$\mathrm{RV}_{q^*,\alpha}(\tau) := \inf\left\{\mathrm{RV};\ (1-q^*)\hat{\tau}_{\mathrm{res}} \in \mathrm{CI}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\tau)\right\} \tag{50}$$

Again from Equation 44 we have that

$$\mathrm{RV}_{q^*,\alpha}(\tau) = \mathrm{RV}_{1,\alpha}(\phi_{\tau^*}) \tag{51}$$

Which for the special case of $q^* = 1$ simplifies to $\mathrm{RV}_{1,\alpha}(\tau) = \mathrm{RV}_{1,\alpha}(\lambda)$, as before. We can also define robustness values for not rejecting the null hypothesis of a reduction of $(100 \times q^*)\%$ *or worse*

$$\mathrm{RV}_{\geq q^*,\alpha}(\tau) := \inf\left\{\mathrm{RV};\ (1-q^*)\hat{\tau}_{\mathrm{res}} \in \left[\mathrm{LL}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\tau),\quad \mathrm{UL}^{\max}_{1-\alpha,\mathrm{RV},\mathrm{RV}}(\tau)\right]\right\} \tag{52}$$

By the same arguments articulated above, $\mathrm{RV}_{\geq q^*,\alpha}(\tau)$ must be the minimum of the robustness value of the Anderson-Rubin regression evaluated at $\tau^* = (1-q^*)\hat{\tau}_{\mathrm{res}}$ and the robustness value of the first-stage regression

$$\mathrm{RV}_{\geq q^*,\alpha}(\tau) = \min\{\mathrm{RV}_{1,\alpha}(\phi_{\tau^*}),\quad \mathrm{RV}_{1,\alpha}(\theta)\} \tag{53}$$

Which for the special case of $q^* = 1$ (zero null hypothesis) simplifies to the minimum of the robustness value of the first-stage and of the reduced-form, $\text{RV}_{\geq q^*=1,\alpha}(\tau) = \min\{\text{RV}_{1,\alpha}(\lambda), \quad \text{RV}_{1,\alpha}(\theta)\}$.

5 Using the OVB framework for the sensitivity analysis of IV

We now return to our running example to show how these tools can be deployed to assess the robustness of those findings to violations of the IV assumptions, and to facilitate the debate on the basis of these results.

5.1 Minimal reporting and sensitivity plots of the first stage and reduced form

We propose investigators begin their sensitivity analysis by examining the robustness of the first-stage and reduced-form estimates. Not only are these analyses usually important on their own right, but in many cases—including this one—this exercise will be sufficient to establish that the instrumental variable estimate is not very informative of the causal effect of interest, since one is not in a position to rule out confounders or side-effects that can explain away those auxiliary estimates.

Sensitivity of the reduced-form: assessing the zero null hypothesis

We start by examining the sensitivity of the reduced-form estimate of our running example, namely, the effect of *Proximity* on *Earnings*. Recall that if we cannot rule out that the reduced-form is zero, we also cannot rule out the IV estimate is zero.

Table 2 shows the minimal sensitivity reporting initially proposed in Cinelli and Hazlett (2020), but now incorporating the new results of Section 3. Beyond the usual statistics such as the point estimate, standard-error and t-value, we recommend that researchers also report the: (i) partial R^2 of the instrument with the outcome ($R^2_{Y\sim Z|\boldsymbol{X}} = 0.18\%$), as well as (ii) the robustness value ($\text{RV}_{q^*,\alpha} = 0.67\%$), and (iii) the extreme robustness value ($\text{XRV}_{q^*,\alpha} = 0.05\%$), both for where the confidence interval would cross zero ($q^* = 1$), at a chosen significance level (here, $\alpha = 0.05$).

Outcome: *Earnings* (log)

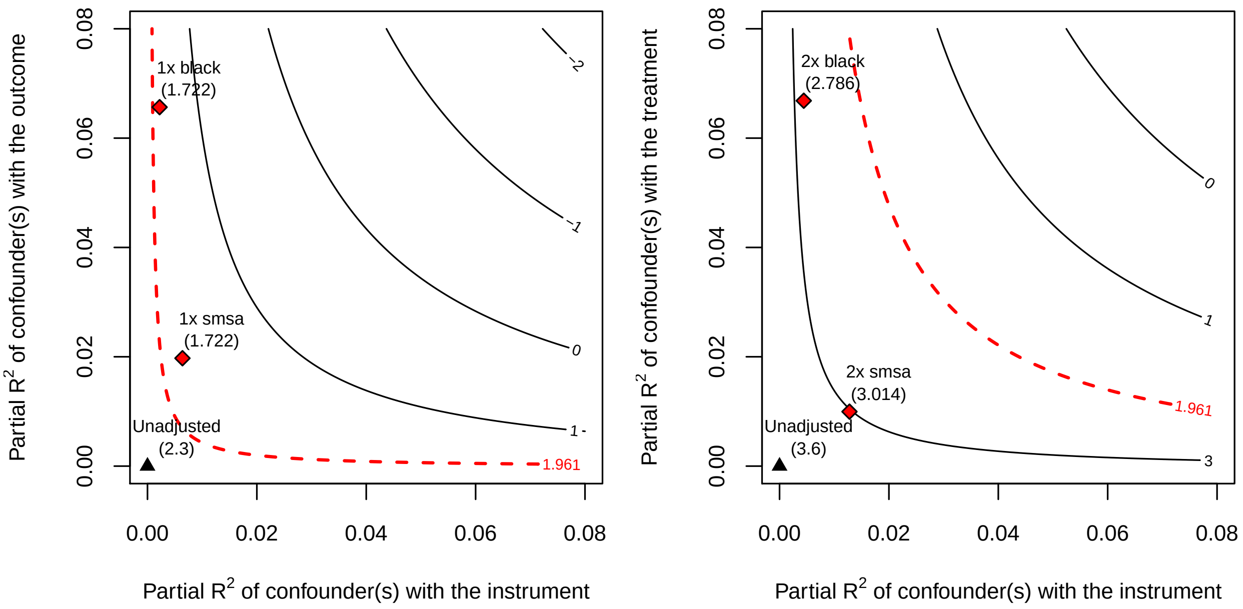
Instrument	Estimate	S.E.	t-value	$R^2_{Y\sim Z\|\mathbf{X}}$	$\text{XRV}_{q^*,\alpha}$	$\text{RV}_{q^*,\alpha}$
Proximity	0.042	0.018	2.33	0.18%	0.05%	0.67%

Bound (1x smsa): $R^2_{Y\sim W|Z,\mathbf{X}} = 2\%$, $R^2_{W\sim Z|\mathbf{X}} = 0.6\%$, $t^\dagger_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.55$

Note: df = 2994, $q^* = 1$, $\alpha = 0.05$

Table 2: Minimal sensitivity reporting of the reduced form regression

For our running example, the robustness value reveals that confounders that explain 0.67%



(a) Sensitivity contours of the reduced form

(b) Sensitivity contours of the first stage

Figure 1: Sensitivity contour plots for the t-value of the reduced form (a) and first stage (b) with benchmark bounds.

of the residual variation both of *proximity* and of (log) *Earnings* are sufficiently strong to make the reduced form estimate statistically insignificant, whereas confounders that explain less than 0.67% of the residual variation of both the instrument and of the outcome are not strong enough to do so. The extreme robustness value and the partial R^2 show that, if we are not willing to impose constraints on the strength of confounders with the outcome, then they would need to explain less than 0.05% or 0.18% of the instrument to escape concerns of eliminating statistical significance or fully eliminating the point estimate, respectively. To aid users in making plausibility judgments, the note of Table 2 provides the maximum strength of unobserved confounding if it were as strong as *smsa* (an indicator variable for whether the individual lived in a metropolitan region) along with the confounding-adjusted critical value for a confounder with such strength, $t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.55$. Since the observed t-value (2.33) is less than the adjusted critical threshold of 2.55, the table reveals that confounding as strong as *smsa* (for example, in the form of residual geographic confounding) is sufficiently strong to be problematic.

Beyond the results of Table 2, we also advise researchers to provide a sensitivity contour plot of the t-value for testing the null hypothesis of zero effect, while also showing different bounds on strength of confounding, under different assumptions of how they compare to the observed variables. This is shown in Figure 1a. The horizontal axis describes the partial R^2 of the confounder with the instrument whereas the vertical axis describes the partial R^2 of the confounder with the outcome. The contour lines show the t-value one would have obtained, had a confounder with such postulated strength been included in the reduced-form regression. The red dashed line shows the statistical significance threshold, and the red diamonds places bounds on strength of confounding as strong as *black* (an indicator for race) and, again, *smsa*. As we can see, confounders as strong as either *black* or *smsa* are sufficient to bring the reduced-form, and hence also the IV estimate, to a region which is not statistically different from zero. Since it is not very difficult to imagine residual confounders as strong or stronger than those (e.g., parental income, finer grained geographic location, etc.), these results for the reduced-form are already sufficient to call into question the reliability of the instrumental variable estimate.

Sensitivity of the first-stage: assessing the stability of IV

We now turn to the sensitivity analysis of the first-stage regression. Table 3 performs the same sensitivity exercises as before, but now for the regression of *Education* (treatment) on *Proximity* (instrument). As expected, the association of proximity to college with years of education is stronger than its association with earnings, and this is also reflected in the robustness statistics, which are slightly higher ($R^2_{D\sim Z|\mathbf{X}} = 0.44\%$, $\text{XRV}_{q^*,\alpha} = 0.31\%$ and $\text{RV}_{q^*,\alpha} = 3.02\%$). As the note of Table 3 shows, confounding as strong as *smsa* would not be sufficiently strong to bring the first-stage estimate to a region where it is not statistically different than zero.

Outcome: *Education* (years)

Instrument	Estimate	S.E.	t-value	$R^2_{D\sim Z\|\mathbf{X}}$	$\text{XRV}_{q^*,\alpha}$	$\text{RV}_{q^*,\alpha}$
Proximity	0.32	0.088	3.64	0.44%	0.31%	3.02%

Bound (1x smsa): $R^2_{D\sim W|Z,\mathbf{X}} = 0.5\%$, $R^2_{Z\sim W|\mathbf{X}} = 0.6\%$, $t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.26$

Note: $\text{df} = 2994$, $q^* = 1$, $\alpha = 0.05$

Table 3: Minimal sensitivity reporting of the first stage regression

Figure 1b supplements those analysis with the sensitivity contour plot for the t-value of the first-stage regression. Here the horizontal axis still describes the partial R^2 of the confounder with the instrument, but now the vertical axis describes the partial R^2 of the confounder with the treatment. The plot reveals that, contrary to the reduced form, the first stage survives confounding once or twice as strong as *black* or *smsa*. Considering that there are practical reasons to believe that proximity to college should indeed affect years of schooling to some degree, even if one cannot rule out with certainty that stronger confounding could exist, the first-stage does not seem to be the main source of concern in our running example. As revealed in the previous section, the the most prominent risk to the validity of the IV estimate seems to stem from plausible sources of residual confounding on the reduced form estimate.

5.2 Minimal reporting and sensitivity plots of the IV estimate

Finally, we turn our attention to the sensitivity analysis of the IV estimate itself, and Table 4 shows our proposed minimal sensitivity reporting. We start with the IV point estimate (0.132), as well as the lower limit ($\text{LL}_{1-\alpha} = 0.025$) and the upper limit ($\text{UL}_{1-\alpha} = 0.285$) of the Anderson-Rubin confidence interval. The t-value for testing the null hypothesis of zero effect is also shown (2.33). Next, we propose researchers to report the extreme robustness value $\text{XRV}_{\geq q^*,\alpha}$ and the robustness value $\text{RV}_{\geq q^*,\alpha}$ for bringing the lower limit of the confidence interval to *or beyond* zero (or another meaningful threshold), at the 5% significance level.

Outcome: *Earnings* (log)

Treatment	Estimate	$\text{LL}_{1-\alpha}$	$\text{UL}_{1-\alpha}$	t-value	$\text{XRV}_{\geq q^*,\alpha}$	$\text{RV}_{\geq q^*,\alpha}$
Education (years)	0.132	0.025	0.285	2.33	0.05%	0.67%

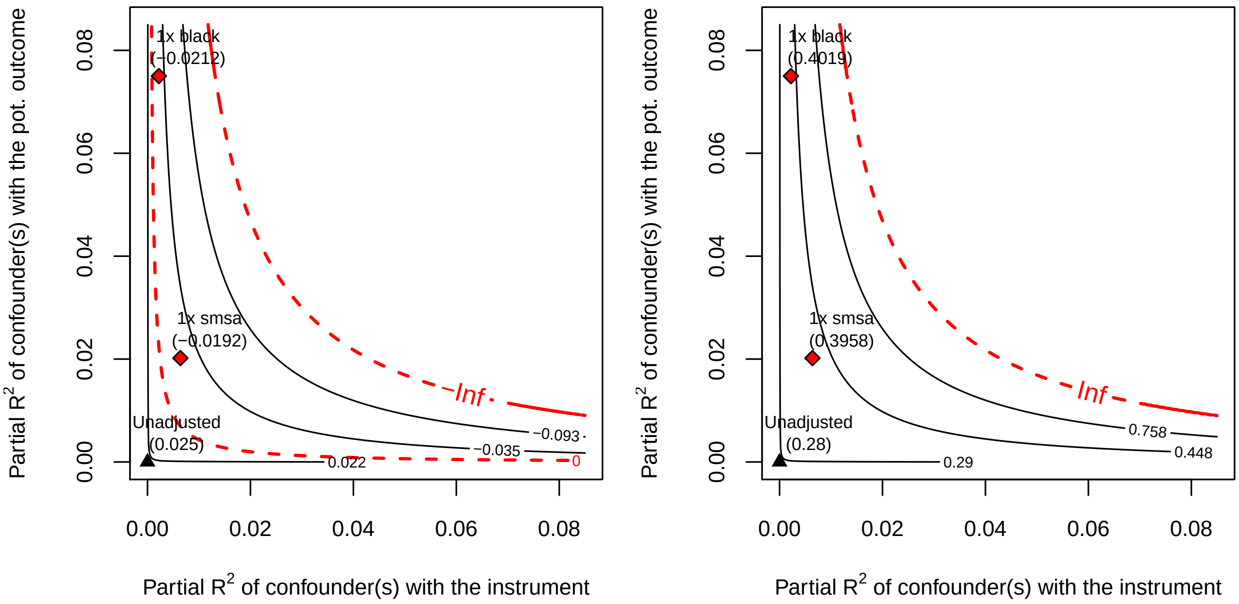
Bound (1x smsa): $R^2_{Y_0\sim W|Z,\mathbf{X}} = 2\%$, $R^2_{W\sim Z|\mathbf{X}} = 0.6\%$, $t^{\dagger}_{\alpha,\text{df}-1,\boldsymbol{R}^2} = 2.55$

Note: df = 2994, $q^* = 1$, $\alpha = 0.05$

Table 4: Minimal sensitivity reporting of IV estimate (Anderson-Rubin)

As derived in Section 4.2.3, we have that the (extreme) robustness value of the IV estimate for bringing the lower limit of the confidence interval to or below zero is the minimum of either the (extreme) robustness value of the reduced form and the (extreme) robustness value of the first stage. Therefore, the sensitivity statistics of Table 4 essentially reproduce the results of Table 2, and all conclusions are the same.

If researchers have already performed the sensitivity of the first-stage and reduced-form regressions, sensitivity analyses for the IV estimate itself are, thus, more informative for null hypotheses *other than zero*. To that end, investigators may wish to examine sensitivity contour plots similar to those of Figure 1, but with contours now showing the adjusted *lower and upper limits* of the confidence interval. These contours are shown Figure 2. Here, as usual, the horizontal axis describes the partial R^2 of the confounder with the instrument, but now the vertical axis describes the partial R^2 of the confounder with the untreated *potential* outcome Y_0. The contour line shows the worst lower (or upper) limit of the set of compatible inferences considering confounders bounded by such strength. Red dashed lines shows a critical contour line of interest (such as zero) as well as the boundary beyond confidence intervals become unbounded. As the plot reveals, even confounding as strong as *smsa* could lead to an interval of compatible inferences for the causal effect of $\mathrm{CI}^{\max}_{1-\alpha,\boldsymbol{R}^2}(\tau) = [-0.02, 0.40]$, which both includes zero and is too wide for any meaningful conclusions regarding the "true" returns to schooling. That is, if we are concerned that omitted variables explaining at least as much of the instrument and untreated potential outcome as *smsa* might exist, then we are unable to reject any estimates in this range.



(a) Sensitivity contours for the lower limit

(b) Sensitivity contours for the upper limit

Figure 2: Sensitivity contour plots for the lower (a) and upper (b) limits of the 95% confidence interval for the IV estimate.

6 Discussion

6.1 The traditional approach to sensitivity for IV

[CC: Discuss prior work in context here.]

6.2 On interpreting robustness values

[CC: Include Angrist and Krueger here for comparison with Card. Discuss the difference.]

6.3 Final remarks

[CC: For convenience, below a Summary of what we did in the paper. Later change to final remarks.]

For an IV estimate to identify a causal effect, it must satisfy the assumptions that, conditionally on included covariates $\boldsymbol{X}$, the instrument is not confounded with the outcome, and it affects the outcome only through its effect on the treatment. In recent years, economics and other social sciences have emphasized the need to scrutinize whether these assumptions are defensible in a given setting. However, the central question in evaluating any empirical IV estimate is not the binary question of *whether these assumptions hold exactly*, but rather, whether these assumptions are violated severely enough to substantively change key research conclusions. Sensitivity analyses provide the tools to reveal precisely what strength of violation is required to change conclusions, enabling improved transparency in how results are reported, together with improved debate around whether they can be believed.

In this work we offer flexible and accessible tools for sensitivity analysis of IV estimates. Building on recent developments of OVB for OLS, we developed an OVB framework for IV that is immune to "weak instrument" concerns, employs a partial R^2 parameterization that naturally handles multiple "side-effects" and confounders of the instrument, and allows exploiting expert knowledge to bound sensitivity parameters, by giving formal bounds on the worst possible bias under the assumption that the maximum explanatory power of omitted variables are no stronger than a multiple of the explanatory power of observed variables. We also developed new sensitivity statistics for routine reporting in the IV setting, such as (extreme) robustness values for IV estimates, describing the minimum strength that omitted variables need to have, both with the instrument, as well with the untreated potential outcome, to invalidate the conclusions of an IV study. Finally, we showed how to deploy visual displays that fully characterize the sensitivity of IV point-estimates and confidence intervals to violations of the standard IV assumptions.

In extending the OVB framework for IV, we have also developed a new perspective for sensitivity analysis of OLS, which can be important for disseminating its use. Specifically, we showed how to construct an "OVB-adjusted" critical t-value that accounts for confounding of any postulated strength. Researchers can then use this new critical value for hypothesis tests or constructing confidence intervals that correctly recover the set of inferences that are compatible with confounding up to the specified strength. These types of analyses can be easily implemented in any software by simply providing the new adjusted threshold.

Our recommendations for sensitivity analysis of IV can be summarised as follows. Reinforcing current prescriptions by Angrist and Pischke (2009) and others, we first recommend that users examine the sensitivity of the first stage and reduced form regressions. These analyses are important to the causal logic of IV questions in their own right. Moreover, a lot can be gleaned about the sensitivity of the ultimate IV estimate from the sensitivity of these two regressions alone. First, for inquiries regarding what confounding would eliminate the IV estimate altogether, the robustness of the reduced form is precisely the robustness of the (Anderson-Rubin) IV estimate. Second, confounding that makes the first stage indistinguishable from zero make IV estimates with arbitrarily large sizes possible. To this end, we recommend that users employ the minimal reporting style we illustrate for the sensitivity of the reduced form and first stage (as in Tables 2 and 3), as well as contour plots to further examine the sensitivity of these components (as in Figure 1). In both cases, it is also useful to examine bounds associated with confounding (multiple times) as strong as observed covariates to aid in building arguments. Finally, the sensitivity of the IV estimate itself can then be examined. Here again, we propose a minimal reporting style (Table 4) that shows the original estimate together with the lower and upper bounds of the $1 - \alpha$ Anderson-Rubin confidence interval, together with the XRV and RV. Contour plots of the lower limit and upper limit of the confidence interval, as in Figure 2, then allow the user to assess the set of compatible inferences considering confounding bounded by the strengths of association shown on the horizontal and vertical axes.

Software able to produce sensitivity analysis of the first stage and reduced form is currently available in the `R` and `Stata` package `sensemakr`. Additional functionality for sensitivity analysis of the IV estimate itself is forthcoming.

	Dependent variable:			
	educ	lwage		
	FS	RF	OLS	IV
	(1)	(2)	(3)	(4)
nearc4	0.320***	0.042**		
	(0.088)	(0.018)		
educ			0.075***	0.132**
			(0.003)	(0.055)
black	−0.936***	−0.270***	−0.199***	−0.147***
	(0.094)	(0.019)	(0.018)	(0.054)
smsa	0.402***	0.165***	0.136***	0.112***
	(0.105)	(0.022)	(0.020)	(0.032)
other covariates	yes	yes	yes	yes
Observations	3,010	3,010	3,010	3,010
R^2	0.477	0.195	0.300	0.238
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 5: Card (1993) results

References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human resources*, 40(4):791–821.

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.

Andrews, I., Stock, J. H., and Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11:727–753.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect.* Princeton University Press.

Baiocchi, M., Cheng, J., and Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450.

Bowden, R. J. and Turkington, D. A. (1990). *Instrumental variables*, volume 8. Cambridge university press.

Burgess, S. and Thompson, S. G. (2015). *Mendelian randomization: methods for using genetic variants in causal estimation.* CRC Press.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research.

Card, D. (1999). The causal effect of education on earnings. In *Handbook of labor economics*, volume 3, pages 1801–1863. Elsevier.

Cinelli, C., Ferwerda, J., and Hazlett, C. (2020). sensemakr: Sensitivity analysis tools for OLS in R and Stata. *Working Paper.*

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).*

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. (2019). Sensitivity analysis of linear structural causal models. *International Conference on Machine Learning.*

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203.

Deaton, A. S. (2009). Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. Technical report, National bureau of economic research.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16(4):309–330.

DiPrete, T. A. and Gangl, M. (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological methodology*, 34(1):271–310.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185.

Fossaluza, V., Izbicki, R., da Silva, G. M., and Esteves, L. G. (2017). Coherent hypothesis testing. *The American Statistician*, 71(3):242–248.

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

Gabriel, K. R. (1969). Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250.

Heckman, J. J. and Urzua, S. (2010). Comparing iv with structural models: What simple iv can and cannot identify. *Journal of Econometrics*, 156(1):27–37.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, pages 360–372.

Imbens, G. (2014). Instrumental variables: An econometrician's perspective. Technical report, National Bureau of Economic Research.

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

Jiang, Y., Kang, H., and Small, D. S. (2018). ivmodel: An r package for inference and sensitivity analysis of instrumental variables models with one endogenous variable. *R package vignette.*

Kédagni, D. and Mourifié, I. (2020). Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*.

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

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

Mehlum, H. (2020). The polar confidence curve for a ratio. *Econometric Reviews*, 39(3):234–243.

Patriota, A. G. (2013). A classical measure of evidence for general null hypotheses. *Fuzzy Sets and Systems*, 233:74–88.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443. Morgan Kaufmann Publishers Inc.

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

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

Schervish, M. J. (1996). P values: what they are and what they are not. *The American Statistician*, 50(3):203–206.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058.

Small, D. S. and Rosenbaum, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. (2018). Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947.

Wang, X., Jiang, Y., Zhang, N. R., and Small, D. S. (2018). Sensitivity analysis and power for instrumental variable studies. *Biometrics*.

Wright, P. G. (1928). *Tariff on animal and vegetable oils*. Macmillan Company, New York.