Incomplete Draft. Cite at your own peril.

Implausibly Exogenous? Sensitivity Analysis for Instrumental Variables Methods*

Soumyajit Mazumder†
Harvard University
smazumder@g.harvard.edu

Date last updated: December 21, 2017
Date first created: December 3, 2017

Natural experiments to assess the causal effects of politically relevant variables have proliferated in recent years. Often times, researchers will use an instrumental variable to generate plausibly exogenous variation in some key variable of interest so as to identify causal effects. While researchers tend to focus on the validity of the exclusion restriction, I highlight the limitations of this design when the instrument itself might deviate from ignorability. Within a sensitivity analysis framework, I show how the bias generated from deviations from exogeneity of the instrument can have an a priori unknown direction and lead to weak instrument problems. I apply the method developed in this paper to revisit the long-run impacts of historical institutions. I also provide an R package to easily implement this sensitivity analysis procedure for applied work.

*I would like to thank Matt Blackwell and Fabrizia Mealli for helpful discussions on this project. All errors are, of course, the author's own.

†Ph.D. Candidate, Department of Government, Harvard University, web: http://smazumder.me

INTRODUCTION

Establishing causality is a fundamental goal across most of the social sciences. Knowing the answer to questions such as whether institutions cause growth, if foreign aid causes democratization, and whether the removal of campaign spending limits causes political polarization, for example, is crucial to testing existing theories and speaking to relevant debates. Unfortunately (or fortunately), it is nearly impossible to run randomized control trials that vary these characteristics for myriad reasons. Instead, researchers tend to rely on "natural experiments" whereby some idiosyncratic feature of the world generates *plausibly exogenous* variation in some key variable of interest. Using this "natural experiment" setting, researchers often use an instrumental variables research design–designs where one or more variables create exogenous variation in some other endogenous variable of interest–to identify the causal effect of that endogenous variable.

While instrumental variables designs are increasingly popular in political science, many point out the potential flaw of this design from the exclusion restriction (Conley, Hansen, and Rossi 2012). Though this is an important assumption to justify, I re-emphasize in this article the need to pay closer attention to the (conditional) ignorability assumption required for instrumental variables designs as well. I show that violations of this assumption can have important consequences for causal inferences by biasing the effects in an ex ante unknown direction and by potentially inducing weak instruments problems. Since instrumental variables designs are fundamentally ratio estimands, bias from exogeneity violations can affect both the reduced-form and first-stage effects in different ways. I then go onto provide a sensitivity analysis procedure that builds on Blackwell (2014)'s confounding function approach to directly quantify the amount of unmeasured confounding needed in both the first-stage and reduced-form effects to invalidate a given instrumental variables design. I also provide an `R` package for applied researchers to easily implement the sensitivity analysis procedure.

A BRIEF OVERVIEW OF EXISTING SENSITIVITY ANALYSIS METHODS

Quantifying the sensitivity of one's results to core identification assumptions should be a key step in any analysis of "natural experiments." To date, there are a number of prominent approaches to sensitivity analysis especially with respect to the (conditional) ignorability assumption–the assumption that some treatment is as-good-as randomly assigned conditional on some set of observable covariates. The first approach, typified by Rosenbaum (2002), approaches sensitivity analysis from the following thought experiment. Suppose that units differed in their treatment assignment probabilities. How extreme would these probability differences have to be to invalidate some estimated effect? Using propensity score methods, researchers can then go through these different thought experiments to ascertain how sensitive one's results are to differential treatment

probabilities. The main limitation of this approach, however, is that one cannot directly evaluate *directional* hypotheses about the strength of unmeasured confounding.

The other core approach developed by Imbens (2003) approaches sensitivity analysis through a different re-parameterization of the above thought experiment. Given some level of association between some confounder and the outcome, how much of an effect would that confounder have to have on the explanatory variable to alter inferences? Relatedly given some level of association between the confounder and the explanatory variable, how large would the association between that confounder and the outcome have to be to change inferences? Blackwell (2014), working in this spirit, extends Imbens (2003)'s approach by introducing the *confounding function*, which quantifies to which potential outcomes vary depending on treatment status. While many of these approaches are tailored to selection-on-observables designs, the core contribution of this article is to extend these approaches to instrumental variables designs.

A third set of approaches attempt to do sensitivity analysis for ignorability without invoking any assumptions or invoking minimal assumptions (Ding and VanderWeele 2016). In these analyses, researchers relax assumptions about interactions between confounders and treatments, binary versus continuous variables, and the types of hypotheses under consideration. This idea is also close in spirit to Manski (1990) who advocates for deriving bounds on estimated treatment effects using no or minimal assumptions.

Existing work on sensitivity analysis for instrumental variables tends to focus on the exclusion restriction. Particularly, Conley, Hansen, and Rossi (2012) provide model-based sensitivity analysis procedures for the exclusion restriction that either rely on structural estimation or bayesian setups. Instead, I show that even if the exclusion restriction holds, deviations from ignorability for the instrument can still pose serious problems for causal inferences.

A SELECTION BIAS APPROACH FOR WHEN INSTRUMENTS DEVIATE FROM EXOGENEITY

Notation and Estimands

Throughout, I rely on the potential outcome framework for causal effects as developed by Rubin (1978). Let D_i be some potentially endogenous/confounded treatment of interest for some unit or individual i. Since D_i might be confounded, we might use an instrumental variables design that uses some variable/encouragement Z_i to generate plausibly exogenous variation in D_i to understand the effect of D_i on some outcome Y_i. Potential outcomes, then, can be written in as $Y_i(d, z)$ whereby the hypothetical outcome that i would have is potentially a function of both the treatment and the instrument. Moreover, we can also write D_i in terms of potential outcomes as a function

of Z_i, which gives us $D_i(z)$. From this notation, one can define causal effects simply as being contrasts between various potential outcomes.

Assumptions for Instrumental Variables

Allowing for heterogeneous effects of the instrument across units i in some population N requires us to make several assumption to identify causal effects using instrumental variables designs. Using principle stratification, one can split the overall causal effect of the encouragement Z_i into a weighted average of strata-specific causal effects among compliers, defiers, never-takers, and always takers (Angrist, Imbens, and Rubin 1996).

Assumption I (Stable Unit Treatment Value Assumption/SUTVA): $Y_i(d, z) = Y_i(d', z')$ and $Y_i(d, z) = Y_i, \forall d \in D_i, z \in Z_i$.

This assumption states that potential outcomes are only a function of i's own treatment and encouragement status and that potential outcomes are consistent across all units in the population. Essentially, this rules out the presence of spillover effects across units. Additionally, the consistency component of this assumption implies that the instrument and treatments are comparable across all units.

Assumption II (Ignorability of Instrument): $Y_i(d, z), D_i(z) \perp\!\!\!\perp Z_i | X_i$.

Ignorability is a key condition for identification in instrumental variables design. This assumption states that the instrument/encouragement Z_i must be as-if random conditional on some set of covariates X_i. When researchers state that they use a "plausibly exogenous" variable as an instrument, this is what they usually mean. With this assumption along with SUTVA, one can identify the *Intention-to-Treat* (ITT) effect of the instrument on the outcome. Moreover, ignorability also implies the identification of the *encouragement* effect of the instrument on treatment takeup.

Assumption III (Encouragement Effect): $E[D_i|Z_i = 1, X_i] - E[D_i|Z_i = 0, X_i] > 0$.

This assumption simply states that the instrument must have an effect on treatment uptake. This is the Encouragement Effect (EE) of the instrument. Assuming ignorability holds, one can directly test this assumption in the data.

Assumption IV (Monotonicity of Instrument): $D_i(1) \geq D_i(0), \forall i \in N$.

This assumption simply states that the instrument/encouragement can only move treatment takeup in one direction. That is, there can be no *defiers* who respond by not taking the treatment if encouraged to do so. This assumption is important to narrow down the set of principle-strata specific causal effects to compliers, never-takers, and always-takers.

Assumption V (Exclusion Restriction) $Y_i(d, 1) = Y_i(d, 0), \forall i \in N$.

Finally, the exclusion restriction states that the instrument Z_i can only affect the outcome Y_i *through the endogenous treatment* D_i *of interest*. Within the principal-

stratification framework, this assumption states that the ITT among never-takers and always-takers is equal to zero.

Identification Result

From these above assumptions, one can identify the Local Average Treatment Effect (LATE) τ_{LATE} among the subpopulation of units who respond to the instrument by taking up the treatment:[1]

$$\tau_{LATE} = \frac{E[Y_i|Z_i = 1, X_i = x] - E[Y_i|Z_i = 0, X_i = x]}{E[D_i|Z_i = 1, X_i = x] - E[D_i|Z_i = 0, X_i = x]} \tag{1}$$

From this, we can see that the causal effect τ_{LATE} is a *ratio* of two causal effects. To estimate this, one could use Wald Estimator that uses the sample analogues to estimate τ_{LATE}.[2] In the numerator, we have the ITT–the causal effect of instrument on the outcome. In the denominator, we have the encouragement effect–the causal effect of the instrument on treatment uptake. Separately, these terms are only identified under ignorability of the instrument. This suggests that if ignorability does not hold, either of these terms can be biased in *ex ante* unknown directions. In the following section, I derive the bias function that follows from deviations from ignorability *even if* exclusion and monotonicity hold.

Sensitivity Analysis using Confounding Functions

In this section, I extend Blackwell (2014) and introduce the *confounding function* within the context of instrumental variables designs. As noted above, bias can propagate into the identification of τ_{LATE} in both the numerator and the denominator when ignorability does not hold. Equation 2 directly quantifies the amount and direction of confounding that would occur if potential outcomes for both the reduced-form and first-stage differed by encouragement status:

$$q(z, x) = \frac{E[Y_i(z)|Z_i = z, X_i = x] - E[Y_i(z)|Z_i = 1 - z, X_i = x]}{E[D_i(z)|Z_i = z, X_i = x] - E[D_i(z)|Z_i = 1 - z, X_i = x]} \tag{2}$$

$$\implies q(z, x) = \frac{q(z, x)_{ITT}}{q(z, x)_{EE}} \tag{3}$$

Equation 2 directly models what deviations from ignorability would mean in terms of bias in the first-stage and reduced-forms captured by the decomposition in Equation 3. This decomposition allows us to separately examine biases from the ITT and EE

[1] I refer readers to Angrist, Imbens, and Rubin (1996) for derivation of this result.

[2] One can generalize the Wald-type estimator to regression analysis simply by taking the ratio of the reduced form and first-stage coefficients: $\beta_{LATE} = \frac{\beta_{reduced}}{\beta_{first}}$.

results as a result of deviations from ignorability. When ignorability holds, $Y_i(d, z)$ and $D_i(z)$ are mean independent of Z_i. In this case, there is no confounding so both confounding functions $q(z, x)_{ITT}$ and $q(z, x)_{EE}$ are zero. To further build intuition, the confounding functions $q(z, x)_{ITT}$ and $q(z, x)_{EE}$ allow researchers to investigate thought experiments of the form: what would the bias be if mean potential outcomes were higher or lower for encouraged versus unencouraged units? How would this affect the reduced-form results? How would this affect the first-stage results?

To make this more concrete, consider the setting of Acemoglu, Johnson, and Robinson (2001) who look at the impact of institutions on economic development using historical settler mortality as an instrumental variable. The main idea is that high settler mortality induced colonial powers to set up bad institutions in colonized countries that these bad institutions persisted until today. Further suppose that we can define the instrument as a binary variable of high settler mortality ($Z_i = 1$) versus low settler mortality ($Z_i = 0$) places and the endogenous variable institutions as good institutions ($D_i = 1$) versus bad institutions ($D_i = 0$). Then, the confounding functions presented in Equation 3 correspond to the following settings. If $q(1, x)_{ITT} < 0$, then average potential outcomes of high settler mortality places have lower levels of development than if these places would have been low settler mortality places instead. Relatedly, $q(1, x)_{EE} < 0$ implies that average potential outcomes in the quality of institutions are greater in high settler mortality places than if these were low settler mortality areas instead. What these confounding functions represent, then, is the potential for units to be fundamentally different beyond any causal effect of settler mortality through institutions even barring the potential for exclusion restriction violations.

One can provide additional structure to these confounding functions through a variety of different parameterizations (Blackwell 2014). For instance, we can parameterize the confounding in both the ITT and EE with the parameters α_{ITT} and α_{EE}:[3]

$$q(z, x; \alpha_{ITT}) = \alpha_{ITT} \tag{4}$$

$$q(z, x; \alpha_{EE}) = \alpha_{EE} \tag{5}$$

When α_{ITT} or α_{EE} are non-zero, then expected potential outcomes of Y_i and D_i are on average higher or lower than their counterfactual values within given strata X_i depending on whether these values are positive or negative respectively. Importantly, this implies that bias from deviations from ignorability can have an unknown direction *ex ante*. If either α_{ITT} or α_{EE} are negative, then the sign can flip. Moreover, this also

[3] One could also use a number of different parameterizations of the confounding function depending on the quantity of interest. For example, one could respecify the confounding function to look at how differences in potential outcomes might vary by treatment status.

points out another problem that can result from deviations from ignorability: weak instruments issues. This happens when the observed EE equals α_{EE}:

$$\tau_{LATE} = \frac{E[Y_i(z)|Z_i = z, X_i = x] - E[Y_i(z)|Z_i = 1 - z, X_i = x] - q(z, x; \alpha_{ITT})}{E[D_i(z)|Z_i = z, X_i = x] - E[D_i(z)|Z_i = 1 - z, X_i = x] - q(z, x; \alpha_{EE})} \tag{6}$$

Not only, then, can violations of ignorability invalidate inferences through affecting the ITT, but they can also lead to violations of Assumption III (the presence of encouragement effects). Especially in small samples, bias in the first-stage can lead to weak instruments estimation problems, which provide unreliable inferences about causal parameters.

Implementation

Drawing from Blackwell (2014), one can derive a sensitivity analysis by varying the confounding functions by readjusting the observed outcome or treatment status simply by estimating a propensity score and applying a specific parameterization of the confounding function:[4]

$$Y_i^q = Y_i - q(Z_i, X_i)Pr(1 - Z_i|X_i) \tag{7}$$

$$D_i^q = D_i - q(Z_i, X_i)Pr(1 - Z_i|X_i) \tag{8}$$

By adjusting the outcomes and treatment status by subtracting the hypothetical amount of confounding resultant of deviations from ignorability, we can directly assess the sensitivity of estimates of τ_{LATE} when relaxing both the ignorability assumption for the ITT and EE. Using this adjustment, researchers can simply just re-run their analyses using Wald or two-stage least squares regression estimators.

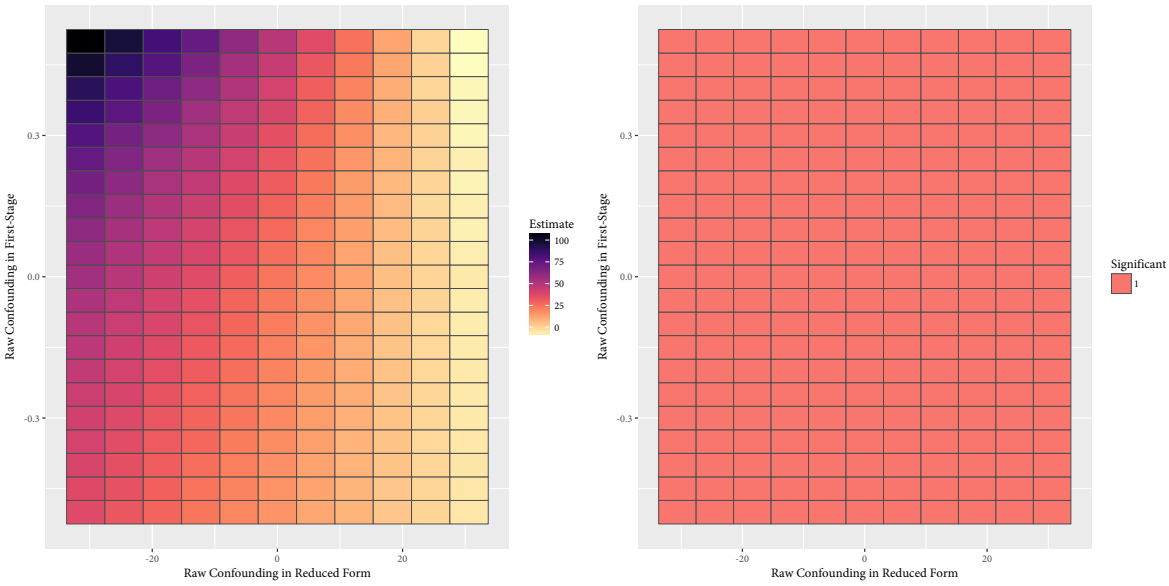
REPLICATION AND RESULTS

Revisiting Iyer (2010)

Is direct colonial rule bad for development? This is a key question in the literatures within comparative politics and political economy. To assess this question, Iyer (2010) uses a natural experiment in the context of India where the death of Indian rulers without an heir led to direct rule by the British during the Doctrine of Lapse (1848-1856). Because the direct versus indirect rule can be confounded by many factors including

[4] See Blackwell (2014) for proofs of these results.

Figure 1: Sensitivity Analysis of Iyer (2010)



culture, human capital, and other unobservable features, Iyer (2010) argues that "plausibly exogenous" variation in the death of rulers of Princely States during this Doctrine of Lapse creates an instrumental variable for the presence of direct rule in a district. To potentially make this assumption tenable, Iyer (2010) controls for some baseline geographic characteristics such as soil quality, latitude, and whether the district is coastal.

Though the exclusion restriction seems plausible within setting, it may be the case that deaths are not actually random. For example, it could be the case that places that already have poor institutions or have lower levels of development have worse health and that these places subsequently have higher death rates. Thus, it is entirely possible that the instrument–death of a ruler during the period of Lapse–deviates from (conditional) ignorability. As shown above, this could bias both reduced-form and the first-stage results. Even more, bias in the first-stage results can lead to weak instruments problems in the estimation of causal effects.

To assess the sensitivity of Iyer (2010)'s instrumental variables results, I apply a one-sided parameterization of the confounding functions of the form $q(z, x) = \alpha(2z - 1)$ where we allow the confounding to vary by encouragement status Z_i.

Figure 1 plots the estimated effect of direct rule on one proxy for contemporary development–infant mortality–varying the α_{ITT} and α_{EE} parameters across a range

of values over the support of both the outcome and endogenous variable on the left panel. Moreover, the right panel of Figure 1 plots an indicator for whether the estimate was statistically significant from zero at the $p < 0.05$ level using the non-parametric bootstrap to generate uncertainty estimates. From this exercise, we can see that even large deviations from ignorability never make the point estimate of τ_{LATE} switch signs. Moreover, $\widehat{\tau_{LATE}}$ is always statistically significant at the $p < 0.05$ even over this fairly large grid of α parameters. These results, then, show that *even if* deaths of rulers are not totally (conditionally) random, the inferences that we make about the adverse effects of direct colonial rule on long-run economic development in India are unlikely to change.

CONCLUSION AND EXTENSIONS

In this paper, I present a framework for understanding the consequences of implausibly exogenous instrumental variables. Bias from deviations from ignorability are two-fold: (1) they affect the potential sign of estimated causal effects in unknown directions and (2) they can lead to weak instruments issues. Recognizing this issue, I provide a sensitivity analysis procedure that extends confounding functions developed by Blackwell (2014) to ascertain the conditions under which unmeasured confounding can overturn conclusions about causal relations in instrumental variables designs. I also allow applied researchers to easily implement these tools in an easy to use `R` package.

Natural experiments that rely on instrumental variables are fundamental to overcoming the problem of selection bias in observational studies. Though instrumental variables designs are a potentially productive avenue for making progress, it is important to understand how our results change when assumptions such as ignorability and the exclusion restriction do not hold. Thus, this paper can be viewed as a natural extension of Sovey and Green (2011) and Conley, Hansen, and Rossi (2012) who argue for applied researchers to discuss the plausibility of their assumptions in-depth. This paper contributes to this literature by providing a direct characterization of the bias resultant from violations of ignorability and how to conduct a useful sensitivity analysis of these violations.

References

Acemoglu, Daron, Simon Johnson, and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *American Economic Review* 91 (5): 1369–1401.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin. 1996. "Identification of Causal Effects using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–455.

Blackwell, Matthew. 2014. "A Selection Bias Approach to Sensitivity Analysis or Causal Effects." *Political Analysis* 22 (2): 169–182.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94 (1): 260–272.

Ding, Peng, and Tyler J. VanderWeele. 2016. "Sensitivity Analysis Without Assumptions." *Epidemiology* 27 (3): 368–377.

Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review* 93 (2): 126–132.

Iyer, Lakshmi. 2010. "Direct versus Indirect Colonial Rule in India: Long-Term Consequences." *Review of Economics and Statistics* 92 (4): 693–713.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80 (2): 319–323.

Rosenbaum, Paul R. 2002. *Observational Studies*. Springer-Verlag.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.

Sovey, Allison J., and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55 (1): 188–200.