

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/324133463>

Sensitivity analysis and power for instrumental variable studies

Article in *Biometrics* · April 2017

DOI: 10.1111/biom.12873

CITATIONS

22

READS

49

4 authors, including:




[Xuran Wang](#)

Icahn School of Medicine at Mount Sinai

16 PUBLICATIONS 442 CITATIONS

[SEE PROFILE](#)

Sensitivity Analysis and Power for Instrumental Variable Studies

Xuran Wang,^{*} Yang Jiang,^{**} Nancy R. Zhang,^{***} and Dylan S. Small^{****} 

The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.

^{*}*email:* xuranw@wharton.upenn.edu

^{**}*email:* yajiang@wharton.upenn.edu

^{***}*email:* nzh@wharton.upenn.edu

^{****}*email:* dsmall@wharton.upenn.edu

SUMMARY. In observational studies to estimate treatment effects, unmeasured confounding is often a concern. The instrumental variable (IV) method can control for unmeasured confounding when there is a valid IV. To be a valid IV, a variable needs to be independent of unmeasured confounders and only affect the outcome through affecting the treatment. When applying the IV method, there is often concern that a putative IV is invalid to some degree. We present an approach to sensitivity analysis for the IV method which examines the sensitivity of inferences to violations of IV validity. Specifically, we consider sensitivity when the magnitude of association between the putative IV and the unmeasured confounders and the direct effect of the IV on the outcome are limited in magnitude by a sensitivity parameter. Our approach is based on extending the Anderson–Rubin test and is valid regardless of the strength of the instrument. A power formula for this sensitivity analysis is presented. We illustrate its usage via examples about Mendelian randomization studies and its implications via a comparison of using rare versus common genetic variants as instruments.

KEY WORDS: Anderson–Rubin test; Instrumental variable (IV); Linear IV regression model; Measure of IV strength; Power function; Sensitivity analysis.

1. Introduction

In observational studies, it is challenging to make causal inference about treatment effects due to the potential presence of unmeasured confounding or reverse causation. One approach to address these challenges is the instrumental variable (IV) method, which uses an instrument to extract a quasi-random experimental study from an observational study. The method requires a valid IV, which is a variable that satisfies three conditions: (IV–C1) IV and exposure are associated either because IV is causing exposure and/or because they have a common parent; (IV–C2) there are no unmeasured confounders between the IV and outcome; (IV–C3) the IV affects the outcome only through its effect on the exposure. See Angrist et al. (1996), Hernán and Robins (2006), Brookhart and Schneeweiss (2007), Baiocchi et al. (2014), and Imbens (2014) for more discussions of IV.

Figure 1 depicts the three conditions for a variable being a valid IV and the relationship between the IV, exposure, outcome and unmeasured confounders. When applying the IV method in a real study, investigators need to evaluate if there are any variables which satisfy the three conditions for being a valid IV. (IV–C1) can be tested with observed data if we assume away the possibility that the exposure caused the IV but not otherwise. (IV–C2) and (IV–C3) cannot be completely tested, see Morgan and Winship (2007), Section 9. Therefore, it is often difficult to know whether an IV is perfectly valid in a study. Even when an IV is invalid, it may still be useful if the inferences from using the IV are not

sensitive to plausible magnitudes of invalidity, which can be assessed through a sensitivity analysis (Angrist et al., 1996; Imbens and Rosenbaum, 2005; Brookhart and Schneeweiss, 2007; Small and Rosenbaum, 2008). There is some previous work on sensitivity analysis for IV studies, see DiPrete and Gangl (2004), Small (2007), Kolesár et al. (2011) and Conley et al. (2012). These papers all use test statistics which are based on the two stage least squares estimator having an approximately normal distribution, which breaks down in the presence of weak instruments (instruments that are weakly associated with the exposure), see Nelson and Startz (1990). Weak IVs are very common in Mendelian randomization studies (Lawlor et al. (2008), Section 4.10), in which genetic variants are used as IVs.

The Anderson–Rubin (AR) test (Anderson and Rubin, 1949) has been shown to have good properties in a setting with one IV that is either weak or strong. Under the normal linear structural equation model setting that is reviewed in Section 2, the AR test uses the F-statistic in the regression of outcome on IV under the null hypothesis that exposure has no causal effect. It is an exact test regardless of the strength of the IV. When the covariance matrix of the structural errors is known, the AR test is uniformly most powerful among all unbiased tests (Moreira, 2001) and among all invariant similar tests (Andrews et al., 2006). When the covariance matrix is unknown, Andrews et al. (2006) showed that the AR test is asymptotically efficient for local alternatives under some regularity conditions.

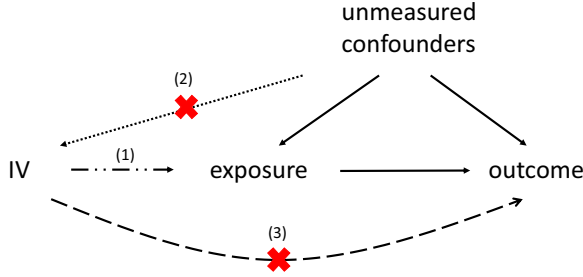


Figure 1. A valid IV requires three conditions. The dash-dotted line suggests that the IV is associated with the exposure, which is IV-C1. The non-existing (“X”) dotted line suggests that the assignment of IV is independent of the unmeasured confounders, which is IV-C2. Similarly, the non-existing dashed line represents IV-C3 that the IV affects the outcome only through its effect on the exposure.

Since the AR test has good performance for both weak and strong IVs, in this article, we develop a method of sensitivity analysis for instrumental variables based on the AR test. We demonstrate that the sensitivity analysis, unlike previous sensitivity analyses, is not strongly dependent on instrument strength. Another contribution we make is that we give a power formula for the sensitivity analysis. The power formula enables researchers to decide how large a sample to collect if the goal is to find evidence for an exposure effect that is insensitive to a specified amount of invalidity of the IV. We show that when considering sensitivity analysis, the concentration parameter (F statistic), which is a commonly used criterion for measuring IV strength, is no longer a good measure for achieving a large power. Instead, it is better to focus on the IV effect size. This has important implications for the design of Mendelian randomization studies, in particular the choice between focusing on common versus rare variants, which will be discussed in Section 5.

In Section 2, we formulate a potential outcome model with a possibly invalid IV. In Section 3, we review the original 2SLS estimator and the AR test. In Section 4, we present our sensitivity analysis approach and provide the power formula for sensitivity analysis. In Section 5, we present applications to Mendelian randomization studies. Section 6 provides conclusions.

2. Instrumental Variable Model with Possible Invalid Instruments

In this section, following Holland (1988), we formulate a causal potential outcomes model with a putative IV and connect it to the simultaneous equations model (Hausman, 1983).

This article considers the setting of one IV and one exposure. For individual i , Z_i , D_i , Y_i represent the observed IV, exposure and outcome accordingly. $Y_i^{(d,z)}$ denotes the potential outcome when the individual i is assigned the exposure d and IV z . $D_i^{(z)}$ is the “potential exposure,” which is the exposure of individual i if forcefully assigned to IV z . In total, n sets of outcome, exposure, and instruments are observed in an i.i.d fashion. We consider the Additive Linear, Constant Effects (ALICE) model of Holland (1988). The effect of the

IV on the exposure is linearly additive:

$$D_i^{(z)} - D_i^{(0)} = \eta z, \quad \forall i. \quad (1)$$

we can write the potential exposure $D_i^{(0)}$ as:

$$\begin{aligned} D_i^{(0)} &= \mathbb{E}(D_i^{(0)} | Z_i = 0) + \{\mathbb{E}(D_i^{(0)} | Z_i) - \mathbb{E}(D_i^{(0)} | Z_i = 0)\} \\ &\quad + \{D_i^{(0)} - \mathbb{E}(D_i^{(0)} | Z_i)\} = \mathbb{E}(D_i^{(0)} | Z_i = 0) + \kappa(Z_i) + v_i \\ &= \gamma_0 + \kappa Z_i + v_i, \quad \forall i. \end{aligned} \quad (2)$$

In equation (2), the effect of unmeasured confounders between the IV and exposure $\kappa(Z_i) = \mathbb{E}(D_i^{(0)} | Z_i) - \mathbb{E}(D_i^{(0)} | Z_i = 0)$ is linear $\kappa(Z_i) = \kappa Z_i$; we also assume the error term $v_i = D_i^{(0)} - \mathbb{E}(D_i^{(0)} | Z_i)$ is independent of IV Z_i . Combining (1) and (2), we get the following “first stage” model that relates the observed D to the observed Z :

$$\begin{aligned} D_i &= D_i^{(Z_i)} = \left(D_i^{(Z_i)} - D_i^{(0)} \right) + \mathbb{E}(D_i^{(0)} | Z_i = 0) + \kappa(Z_i) + v_i \\ &= \gamma_0 + \eta Z_i + \kappa Z_i + v_i \quad \forall i; \\ v_i &\text{ is i.i.d. with mean 0; } \quad v_i \perp Z_i. \end{aligned} \quad (3)$$

The causal effect for the exposure D on the potential outcome Y is linear and let $\delta_1^i(\cdot)$ be:

$$Y_i^{(d,z)} - Y_i^{(0,z)} = \beta d, \quad (4)$$

$$\delta_1^i(z) = Y_i^{(0,z)} - Y_i^{(0,0)}, \quad \forall i \quad (5)$$

Combining (4) and (5), we know that $\delta_1^i(\cdot)$ measures the direct effect of the IV on the outcome. We assume this effect is linear and homogeneous, $\delta_1^i(Z) = \delta_1 Z$.

For the potential outcome term $Y_i^{(0,0)}$:

$$\begin{aligned} Y_i^{(0,0)} &= \mathbb{E}(Y_i^{(0,0)} | Z_i = 0, D_i = 0) + \{\mathbb{E}(Y_i^{(0,0)} | Z_i, D_i = 0) \\ &\quad - \mathbb{E}(Y_i^{(0,0)} | Z_i = 0, D_i = 0)\} + \{Y_i^{(0,0)} - \mathbb{E}(Y_i^{(0,0)} | Z_i, D_i = 0)\} \\ &= \mathbb{E}(Y_i^{(0,0)} | Z_i = 0, D_i = 0) + \delta_2(Z_i) + u_i \\ &= \beta_0 + \delta_2 Z_i + u_i, \quad \forall i. \end{aligned} \quad (6)$$

In equation (6), $\delta_2(Z_i) = \mathbb{E}(Y_i^{(0,0)} | Z_i) - \mathbb{E}(Y_i^{(0,0)} | Z_i = 0)$, representing the effect of unmeasured confounders between the IV and outcome, is linear. We assume the error term $u_i = Y_i^{(0,0)} - \mathbb{E}(Y_i^{(0,0)} | Z_i, D_i = 0)$ is independent of Z_i . u_i is not independent of D_i because u_i may still be associated with v_i (see $\kappa(\cdot)$ and $\delta_1(\cdot)$ in Figure 2). Combining (4)–(6), the “second stage” model that relates the observed Y to the observed

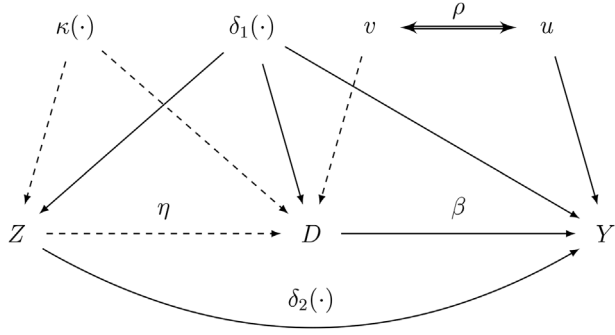


Figure 2. Complete DAG for the model in Section 2. The dashed arrows represent the first stage model and the solid arrows represent the second stage model.

Z is:

$$\begin{aligned} Y_i &= Y_i^{(D_i, Z_i)} = \left(Y_i^{(D_i, Z_i)} - Y_i^{(0, Z_i)} \right) + \left(Y_i^{(0, Z_i)} - Y_i^{(0, 0)} \right) \\ &\quad + \left(\mathbb{E}(Y_i^{(0, 0)} | Z_i = 0, D_i = 0) + \delta_2(Z_i) + u_i \right) \\ &= \beta_0 + \beta D_i + \delta_1 Z_i + \delta_2 Z_i + u_i \\ \forall i, \quad u_i &\text{ is i.i.d. with mean 0, } u_i \perp Z_i. \end{aligned} \quad (7)$$

Assume v_i and u_i are bivariate normal and combine (3) and (7), our complete model is:

$$\begin{aligned} Y_i &= \beta_0 + \beta D_i + (\delta_1 + \delta_2) Z_i + u_i, \quad D_i = \gamma_0 + (\eta + \kappa) Z_i + v_i; \\ (u_i, v_i) &\perp Z_i; \quad (u_i, v_i)^T \sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \end{aligned} \quad (8)$$

The parameter of interest is the causal effect of the exposure on the outcome β . The parameters δ_1 and δ_2 measure the violation of condition (IV-C3) and (IV-C2) for being a valid IV correspondingly.

The model (8) has the same structure as the models for sensitivity analysis in DiPrete and Gangl (2004), Kolesár et al. (2011) and Conley et al. (2012). As a summary, (8) relies on the following assumptions: 1) the observed subjects are an i.i.d. sample from a population; 2) the causal effects of the IV on the exposure and the exposure on the outcome are linearly additive; 3) the confounder function $\kappa(\cdot)$ and the IV violation functions $\delta_1^i(\cdot)$, $\delta_2(\cdot)$ are linear and $\delta_1^i(\cdot)$ is homogeneous for any individual i ; 4) the error terms (u_i, v_i) are bivariate normal and independent of Z_i .

We can reduce the number of parameters in (8) by defining $\gamma = \eta + \kappa$, $\delta = \delta_1 + \delta_2$. Z will satisfy condition (IV-C1) as long as $\gamma \neq 0$, (Hernán and Robins, 2006). The parameters δ_1 and δ_2 measure the violation of valid IV conditions and are combined into $\delta = \delta_1 + \delta_2$, so that δ can be treated as the sensitivity parameter in (8), which describes the amount of invalidity of the IV. δ is the correlation between the IV and outcome when given exposure,

$$\delta = \frac{\text{Cov}(Z, Y|X)}{\text{Var}(Z|X)}.$$

Other observed covariates (write as vector X_i of length k) can be added into (8):

$$\begin{aligned} Y_i &= \beta_0 + \beta_X^T X_i + \beta D_i + \delta Z_i + u_i, \\ D_i &= \gamma_0 + \gamma_X^T X_i + \gamma Z_i + v_i. \end{aligned}$$

Writing the vector form of the observations as $Y_{n \times 1} = (Y_1, \dots, Y_n)^T$, $D_{n \times 1} = (D_1, \dots, D_n)^T$, $X_{n \times k} = (X_1, \dots, X_n)^T$ etc, and also merging the intercept into the observed covariates X , we get an analogous model to (8):

$$\begin{aligned} Y &= X\beta_X + \beta D + \delta Z + u, \quad D = X\gamma_X + \gamma Z + v, \\ (u, v) &\perp Z; \quad (u_i, v_i)^T \sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}; \quad \text{rank}(X) = k. \end{aligned} \quad (9)$$

Since β_X and γ_X are not of interest, we can use the Frisch–Waugh–Lovell theorem (Davidson and MacKinnon, 1993; Wang and Zivot, 1998) to transform the model (9) by using the projection matrix $M_X = I_{N \times N} - X(X^T X)^{-1} X^T$. Also, to make the sensitivity parameter δ in the model more interpretable, we rescale δ as $\delta\sigma_1$. After this rescaling, a unit change in the invalid IV Z will lead to a change of δ standard deviations of the structural error $u = Y^{(0, 0)} - \mathbb{E}(Y^{(0, 0)} | Z)$. The final model becomes:

$$\begin{aligned} Y^* &= \beta D^* + \delta\sigma_1 Z^* + u^*, \quad D^* = \gamma Z^* + v^*, \quad Y^* = M_X Y; \quad D^* = M_X D; \\ Z^* &= M_X Z; \quad u^* = M_X u; \quad v^* = M_X v; \quad (u, v) \perp Z; \\ (u_i, v_i)^T &\sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}; \quad \text{rank}(X) = k. \end{aligned} \quad (10)$$

We will consider model (10) in the rest of the article. In the following sections, our work will mainly focus on inference for β , given restrictions on the range of the sensitivity parameter δ . In assessing plausible values of δ for the effect of the genetic variant on the unmeasured confounder, it may be useful to calibrate values of δ to the effect of the genetic variant on measured confounders X (Hsu and Small, 2013). Detailed interpretation of the models is discussed in Supplementary Materials.

3. The 2SLS Method and the Anderson Rubin Test

In this section, we consider model (10) with $\delta = 0$, which is the usual two stage IV regression model with a valid IV. We will briefly review the two stage least squares estimator (2SLS) that has an asymptotic normal distribution and the standard AR test.

The 2SLS estimator with single IV of β is found by first regressing D^* on Z^* to find \hat{D}^* , and then regressing Y^* on \hat{D}^* . The 2SLS estimator can be written as follows:

$$\hat{\beta}_{2SLS} = \frac{\text{cov}(Z^*, Y^*)}{\text{cov}(Z^*, D^*)} = \frac{Z^{*T} Y^*}{Z^{*T} D^*}. \quad (11)$$

As the sample size increases to infinity, $\hat{\beta}_{2SLS} \rightarrow \beta$. Also the asymptotic variance for $\hat{\beta}_{2SLS}$ (Nelson and Startz, 1990) is:

$$\text{Asymptotic Variance}(\sqrt{n} \times (\hat{\beta}_{2SLS} - \beta)) = \frac{\sigma_{u^*}^2 \cdot \text{Var}(Z^*)}{\text{Cov}^2(Z^*, D^*)}. \quad (12)$$

(11) and (12) can be used to construct an asymptotically valid t-test. However, when the IV is weak, the asymptotics of this test may provide a poor guide to the actual performance of the test even for moderately large sample sizes. (Nelson and Startz, 1990)

From Anderson and Rubin (1949), the AR test compares the null and alternative hypotheses: $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$. If the IV is valid, then under H_0 , $Y^* - \beta_0 D^* = u^*$ is independent of Z^* and the coefficient of regressing $Y^* - \beta_0 D^* = u^*$ in the population is 0. The AR test is an F test for this coefficient being 0 with the following expression:

$$AR(\beta_0) = \frac{(Y^* - \beta_0 D^*)^T P_{Z^*} (Y^* - \beta_0 D^*)}{(Y^* - \beta_0 D^*)^T M_{Z^*} (Y^* - \beta_0 D^*) / (n - k - 1)}, \quad (13)$$

where n is the number of samples, P_{Z^*} , M_{Z^*} are projection matrices $P_{Z^*} = Z^*(Z^{*T}Z^*)^{-1}Z^{*T}$ and $M_{Z^*} = I_n - P_{Z^*}$. Under $H_0 : \beta = \beta_0$, since $u^{*T}P_{Z^*}u^* \sim \sigma_1^2\chi_1^2$, $u^{*T}M_{Z^*}u^*/(n - k - 1) \sim \sigma_1^2\chi_{n-k-1}^2$ and they are independent (see more details in the Supplementary Materials):

$$AR(\beta_0) = \frac{u^{*T}P_{Z^*}u^*}{u^{*T}M_{Z^*}u^*/(n - k - 1)} \sim F_{1, n-k-1}. \quad (14)$$

We reject H_0 when $AR(\beta_0) > F_{1, n-k-1; 1-\alpha}$, where α is the significance level and $F_{1, n-k-1; 1-\alpha}$ is the $1 - \alpha$ quantile of the F distribution with degrees of freedom 1 and $n - k - 1$. (Notice that since u^* is the error after projecting out the effect of the covariates which include an intercept, there are only $n - k - 1$ degrees of freedom in the denominator.) In contrast with the t-test based on the 2SLS estimator, the AR test has correct size regardless of the sample size and strength of the IV.

Inverting the AR test, the $1 - \alpha$ confidence interval (CI) is constructed by solving the inequality:

$$CI_{1-\alpha} = \left\{ \beta : \frac{(Y^* - \beta D^*)^T P_{Z^*} (Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*} (Y^* - \beta D^*) / (n - k - 1)} \leq F_{1, n-k-1; 1-\alpha} \right\}. \quad (15)$$

We calculate the power functions for the AR test in the Supplementary Materials. Under the alternative hypothesis $H_1 : \beta - \beta_0 = \lambda$, the power formula is:

$$\begin{aligned} \text{Power} &= P_\lambda(AR(\beta_0) > F_{1, n-k-1; 1-\alpha}) \\ &= 1 - \Psi_{1, n-k-1, \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda}(F_{1, n-k-1; 1-\alpha}), \end{aligned} \quad (16)$$

where $\Lambda = \frac{\lambda^2}{(\sigma_1/\sigma_2)^2 + 2\rho\sigma_1/\sigma_2\lambda + \lambda^2}$, $\Psi_{a,b,k}(\cdot)$ is the CDF of the non-central F-distribution with degrees of freedom a , b , and non-centrality parameter k . The term $\gamma^2 Z^{*T} Z^* / \sigma_2^2$ is called the concentration parameter, the larger the concentration parameter is, the larger the power is. The concentration parameter

is the population value of the first stage F statistic for the IV when the treatment is regressed on it. It is a popular measure of instrument strength (Stock et al., 2002). Large values of the concentration parameter indicate strong instruments. See Rothenberg (1984), Section 6.1 for more information about the concentration parameter.

We want to point out that the strength of the IV poses a fundamental limit on the power in IV studies. Looking at equation (16), if the concentration parameter is fixed, then no matter how large the effect size λ/σ_1 is, Λ has a fixed upper bound $1/(1 - \rho^2)$ and the power has an upper bound $1 - \Psi_{1, n-k-1, \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2(1-\rho^2)}}(F_{1, n-k-1; 1-\alpha})$. Therefore the power

will not increase to 1 as the effect size increases to infinity. Thus, the concentration parameter imposes a fundamental limit on the power of the AR test which cannot be overcome with a large effect size of the treatment. However, the power of the AR test will increase to 1 when sample size n increases to infinity. See Section 4.2 Proposition 1 (d).

4. Sensitivity Analysis and Power of Sensitivity Analysis

For sensitivity analysis, we assume that $\delta \in (\underline{\delta}, \bar{\delta})$ in model (10) and make inference when the value of δ is unknown but its range $(\underline{\delta}, \bar{\delta})$ is known.

4.1. CI and Power Formula for Sensitivity Analysis Using AR Test

We first suppose the true value of the sensitivity parameter δ in $Y^* = \beta D^* + \delta \sigma_1 Z^* + u^*$ is known. The AR test statistic under $H_0 : \beta = \beta_0$ becomes:

$$\begin{aligned} AR(\beta_0) &= \frac{(Y^* - \beta_0 D^*)^T P_{Z^*} (Y^* - \beta_0 D^*)}{(Y^* - \beta_0 D^*)^T M_{Z^*} (Y^* - \beta_0 D^*) / (n - k - 1)} \\ &= \frac{\left(\delta \sigma_1 \sqrt{Z^{*T} Z^*} + \frac{Z^{*T} u^*}{\sqrt{Z^{*T} Z^*}} \right)^2}{u^{*T} M_{Z^*} u^* / (n - k - 1)} \sim F_{1, n-k-1, \delta^2 Z^{*T} Z^*}, \end{aligned} \quad (17)$$

where $F_{a,b,c}$ stands for the non-central F distribution with degrees of freedom a , b , and non-central parameter c . Therefore a $1 - \alpha$ CI can be obtained as:

$$CI_{1-\alpha}(\delta) = \left\{ \beta : \frac{(Y^* - \beta D^*)^T P_{Z^*} (Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*} (Y^* - \beta D^*) / (n - k - 1)} < F_{1, n-k-1, \delta^2 Z^{*T} Z^*; 1-\alpha} \right\}, \quad (18)$$

where $F_{a,b,c; 1-\alpha}$ stands for the $1 - \alpha$ quantile of the distribution $F_{a,b,c}$ defined as above. Define $\Delta = \max(|\underline{\delta}|, |\bar{\delta}|)$. Similar to the procedure in (Vansteelandt et al., 2006) for constructing uncertainty regions, we can construct a CI for β which will provide at least $1 - \alpha$ coverage by taking the union of

$CI_{1-\alpha}(\delta)$, for every $\delta \in (\underline{\delta}, \bar{\delta})$:

$$CI_{1-\alpha} = \cup_{\delta \in (\underline{\delta}, \bar{\delta})} CI_{1-\alpha}(\delta) = \left\{ \beta : \frac{(Y^* - \beta D^*)^T P_{Z^*} (Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*} (Y^* - \beta D^*) / (n - k - 1)} < F_{1, n-k-1, \Delta^2 Z^{*T} Z^*; 1-\alpha} \right\}. \quad (19)$$

We now consider the power for being able to reject $H_0 : \beta = \beta_0$ for all $\delta \in (\underline{\delta}, \bar{\delta})$ when the true δ is δ^* . For calculating the power, details are derived in the Supplementary Materials and here, we only show the main results. Under $H_1 : \beta - \beta_0 = \lambda \neq 0$, we have,

$$AR(\beta_0) \sim F_{1, n-k-1, \frac{(\gamma + \delta^* \sigma_1 / \lambda)^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda}. \quad (20)$$

therefore, the probability of correctly rejecting H_0 for a fixed value δ^* is:

$$Power_{\delta^*} = P_{\lambda}(AR(\beta_0) \notin CI_{1-\alpha}) = P_{\lambda}(AR(\beta_0) > F_{1, n-k-1, \Delta^2 Z^{*T} Z^*; 1-\alpha}) = 1 - \Psi_{1, n-k-1, \frac{(\gamma + \delta^* \sigma_1 / \lambda)^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda} (F_{1, n-k-1, \Delta^2 Z^{*T} Z^*; 1-\alpha}). \quad (21)$$

Rosenbaum (2010), Chapter 14.2 suggests calculating the power for the “favorable situation” in which there is a treatment effect and in fact there is no bias from unmeasured confounding ($\delta^* = 0$), but we do not know that there is no unmeasured confounding and want to be able to reject the null hypothesis of no treatment effect given a certain magnitude of unmeasured confounding, that is, $\delta \in (\underline{\delta}, \bar{\delta})$ in our setting. To calculate the power of sensitivity analysis under this favorable situation, we plug $\delta^* = 0$ into (21):

$$Power_0 = 1 - \Psi_{1, n-k-1, \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda} (F_{1, n-k-1, \Delta^2 Z^{*T} Z^*; 1-\alpha}). \quad (22)$$

Another type of power of sensitivity analysis calculation is to find the minimum power for rejecting the null hypothesis of no treatment effect under a sensitivity analysis that allows for unmeasured confounding in the range $\delta \in (\underline{\delta}, \bar{\delta})$. To calculate this minimum power, we take $\min_{\delta \in (\underline{\delta}, \bar{\delta})}$ on the right hand side of (21),

$$Power \geq \min_{\delta \in (\underline{\delta}, \bar{\delta})} Power_{\delta} = 1 - \Psi_{1, n-k-1, \frac{\min_{\delta \in (\underline{\delta}, \bar{\delta})} (\gamma + \delta \sigma_1 / \lambda)^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda} (F_{1, n-k-1, \Delta^2 Z^{*T} Z^*; 1-\alpha}). \quad (23)$$

The R package “ivpack,” available on CRAN, calculates the sensitivity analysis confidence interval (19) (function `ARSensitivity.ci`), the power of sensitivity analysis using formula (function `ARSensitivity.power`) and the minimum sample size needed for reaching a certain power in a sensitivity analysis (function `ARSensitivity.size`).

4.2. Effect of Different Parameters on Power of Sensitivity Analysis

We consider the effect of different parameters on the power of sensitivity analysis. Here, we will focus on analyzing the power formula (22) for the favorable situation.

The power formula (22) involves two different non-central F distributions with two non-centrality parameters:

$$ncp_1 = \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda; \quad ncp_2 = \Delta^2 Z^{*T} Z^*$$

To analyze the influence of different parameters in the power formula, we study how they affect the size of ncp_1 and ncp_2 . The Supplementary Materials proves the following properties:

PROPOSITION 1. In the power formula (22), we have

- (a) If $ncp_1 = ncp_2$, the power is always α .
- (b) For fixed ncp_2 , power increases as ncp_1 increases.
- (c) For fixed ncp_1 , power decreases as ncp_2 increases.
- (d) If $ncp_1 > ncp_2$, the power is larger than α and will increase to 1 as the sample size increases.
- (e) If $ncp_1 < ncp_2$, the power is smaller than α and will decrease to 0 as the sample size increases.

The $ncp_2 = \Delta^2 Z^{*T} Z^*$ is approximately equal to $n \cdot \Delta^2 \cdot SD(Z^*)^2$. Δ is determined by the allowance of sensitivity and n is the sample size. ncp_1 is affected many parameters:

$$ncp_1 = \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2} \Lambda = \frac{\gamma^2}{\sigma_2^2} \cdot Z^{*T} Z^* \cdot \frac{1}{(\frac{\sigma_1}{\sigma_2 \lambda} + \rho)^2 + 1 - \rho^2} = n \cdot \frac{\gamma^2}{\sigma_2^2} \cdot SD(Z^*)^2 \cdot \frac{1}{(\frac{\sigma_1}{\sigma_2 \lambda} + \rho)^2 + 1 - \rho^2}. \quad (24)$$

In Section 3, we stated that a large concentration parameter $\gamma^2 Z^{*T} Z^* / \sigma_2^2$ will lead to a large power of the AR test, assuming that the IV is valid. However, when conducting sensitivity analysis and considering the power formula (22), a large concentration parameter may be produced by a small $|\gamma|/\sigma_2$ and a large $Z^{*T} Z^*$ (or vice versa). This will result in both large ncp_1 and ncp_2 , for which the power of sensitivity analysis may not be large. On the other hand, if the IV effect size $|\gamma|/\sigma_2$ increases and the other parameters are fixed, then ncp_1 increases and ncp_2 stays the same, which leads to a larger power of sensitivity analysis. This suggests that if we want to have a large power of sensitivity analysis, we should focus on finding a large IV effect size $|\gamma|/\sigma_2$.

For the effect size λ/σ_1 , no matter how it varies, ncp_1 has an upper bound $\gamma^2 Z^{*T} Z^* / (\sigma_2^2 (1 - \rho^2))$ when $\lambda/\sigma_1 = -(\sigma_2 \rho)^{-1}$. Hence, the power of sensitivity analysis cannot go to 1 as the effect size increases to infinity. This is similar to the discussion at the end of Section 3, which is about the power property in the AR test.

If the effect size λ/σ_1 or the IV effect size $|\gamma|/\sigma_2$ is very small, then we may have

$$\frac{1}{(\frac{\sigma_1}{\sigma_2 \lambda} + \rho)^2 + 1 - \rho^2} \cdot \frac{\gamma^2}{\sigma_2^2} < \Delta^2. \quad (25)$$

This will result in $\text{ncp}_1 < \text{ncp}_2$ and the sensitivity analysis cannot have power larger than α for any sample size. Hence, the AR test is not unbiased.

To illustrate, we simulated a simple scenario and calculated the power of sensitivity analysis by varying different parameters. We consider the following base parameters: $\sigma_1^2 = 1$; $\sigma_2^2 = 4$; $\rho = 0.5$; $\gamma = 0.5$; $\lambda = -1$; $\text{SD}(Z) = 1$, and we want to use the power formula (22) under different allowance of sensitivity parameter interval $\delta \in [-0.05, 0.05]$, $[-0.08, 0.08]$, or $[-0.1, 0.1]$. We vary 1) the correlation parameter ρ ; 2) standard deviation of the instrument $\text{SD}(Z^*)$; 3) the effect size λ/σ_1 ; 4) IV effect size γ/σ_2 to see what's the effect upon the power of significant level $\alpha = 0.05$ with sample size $n = 100, 200$, and 1000 . Figure 3 shows the calculated powers. Each row contains 3 panels with different sample size $n = 100, 200$, and 1000 . The first row shows the power when ρ varies. We can see the power increases as ρ increases when the effect size λ/σ_1 is negative. Comparing the 3 panels, this trend is consistent and large sample size leads to bigger power. If effect size is positive, the power will decrease when ρ increases. The second row shows the power when $\text{SD}(Z^*)$ varies between $(-1, 1)$. As shown in the panels, the power increases when $\text{SD}(Z^*)$ increases and the trend is consistent across sample size. The third row varies the effect size λ/σ_1 by changing $\lambda \in (-2, 2)$. In general, we can see the power is large when the effect size is substantial. However, the upper bound for the power is when $\lambda/\sigma_1 = -(\rho\sigma_2)^{-1} = -1$. As λ/σ_1 moves below -1 , the power even starts to drop a little bit. This again corresponds to the previous discussion that no matter how large the effect size is, there's an upper bound for the power. The last row shows the panels of IV effect size γ/σ_2 varies between $(0.083, 0.25)$ by varying σ_2 between $(2, 6)$. We can see the power increases as the IV effect size increases with each fixed sample size. In general, the power increases as sample size increases. When power is less than the significant level $\alpha = 0.05$, the power will decrease as sample size n increases. This corresponds to the situation when the inequality (25) holds (See Figure 3 third row effect size near 0, last row when IV effect size is near 0 and the line $\delta \in [-0.1, 0.1]$).

The power is determined by two non-central parameters: ncp_1 and ncp_2 . When we vary correlation ρ , effect size λ/σ_1 , and IV effect size γ/σ_2 , ncp_2 is invariant while ncp_1 varies corresponds to equation (24). From Proposition 1, the power changes accordingly. When $\text{SD}(Z^*)$ increases, both ncp_1 and ncp_2 increases while the ratio $\text{ncp}_1/\text{ncp}_2$ stays the same. With a similar argument of Proposition 1 (d) and (e), if inequality (25) holds the power decreases when $\text{SD}(Z^*)$ increases (or vice versa).

The design sensitivity describes the asymptotic power of sensitivity analysis (Rosenbaum, 2004, 2010). See Supplemental Materials for more discussion.

5. Applications to Mendelian Randomization Studies

An important application area of the IV method is Mendelian randomization studies (Smith and Ebrahim, 2003, 2004; Ebrahim and Smith, 2008; Lawlor et al., 2008; Glymour et al., 2012). The basic idea of Mendelian randomization is to use inherited genetic variants as IVs to study the effect of an

exposure on an outcome. By Mendel's second law, the transmission of genetic variants between generations is independent of possible confounders like environment and lifestyle factors. This makes it plausible that genetic variants satisfy the condition (IV-C2) for being a valid IV. If a genetic variant is independent of unmeasured confounders and is also associated with the exposure and affects the outcome only through the exposure, then it is a valid IV. However, there are several ways that genetic variants could violate the conditions for being a valid IV (Didelez and Sheehan (2007) S.7 and Lawlor et al. (2008) S.4). If a genetic variant used as an IV is linked to another unmeasured genetic variant on the same chromosome that affects the outcome, there is linkage disequilibrium and the condition (IV-C2) is violated. Another way that a genetic variant could violate (IV-C2) is through population stratification (subpopulations which exhibit systematic differences in genotypes due to different ancestries) which is associated with both the IV and the outcome. Besides possibly violating (IV-C2), a genetic variant could violate (IV-C3) by being pleiotropic in such a way that the genetic variant influences both the exposure and the outcome through a pathway other than the exposure. Consequently, in most studies using Mendelian randomization, there is some concern about whether the proposed IVs are valid (e.g., see Nitsch et al. (2006)). It is useful to do a sensitivity analysis to examine how sensitive the analysis results are to the violation of (IV-C2) and (IV-C3).

5.1. Applications to a Mendelian Randomization Study

Here, we will use the same example in Freeman et al. (2013) to illustrate how to do sensitivity analysis and power calculation in a Mendelian randomization study.

The example concerns the causal effect of C-reactive protein (CRP), a marker of inflammation, on fibrinogen, a marker for coronary heart disease. The gene that makes CRP has several variations in the form of single nucleotide polymorphisms (SNPs). SNPs of the CRP gene have commonly been used to study the causal effects of CRP in Mendelian randomization studies, for example, see Lawlor et al. (2008). Although the CRP gene is believed to only directly affect CRP, it is possible that there is some unknown mechanism by which the CRP gene affects fibrinogen not through affecting CRP levels and it is also possible that there is population stratification. Consequently, we would like to consider a sensitivity analysis that allows for violations of the assumptions of the CRP gene being a valid IV.

We will use the same simulated data settings as Freeman et al. (2013), the setting is: $Z = \{1, 2, 3\}$ with $\text{prob.}(1/9, 4/9, 4/9)$; $U \sim N(0, 1.11 \times p \times 0.99)$, $X \sim N(U + 0.1(Z - 2)\sqrt{1.11 \times 9/4}, 1.11 \times (1 - p) \times 0.99)$; $Y = 0.234X + U/\sqrt{0.99}$, which gives $\beta = 0.234$; $\rho_{ZD}^2 = 0.01$; $\text{Var}(X) = 1.11$; $\text{Var}(Z) = 4/9$. We can rewrite this model in an equivalent form which fits our model setting (10) with:

$$\begin{aligned} \beta &= 0.234; \quad \delta = 0; \quad \gamma = 0.1 \cdot \sqrt{1.11 \cdot 9/4}; \quad \sigma_1^2 = 1.11 \cdot p; \\ \sigma_2^2 &= 1.11 \cdot 0.99; \quad \rho = \sqrt{p}, \end{aligned} \quad (26)$$

where the concentration parameter $\gamma^2 Z^T Z / \sigma_2^2 = 4.04 \times 10^{-3}n$. In Freeman et al. (2013), the larger p is, the more

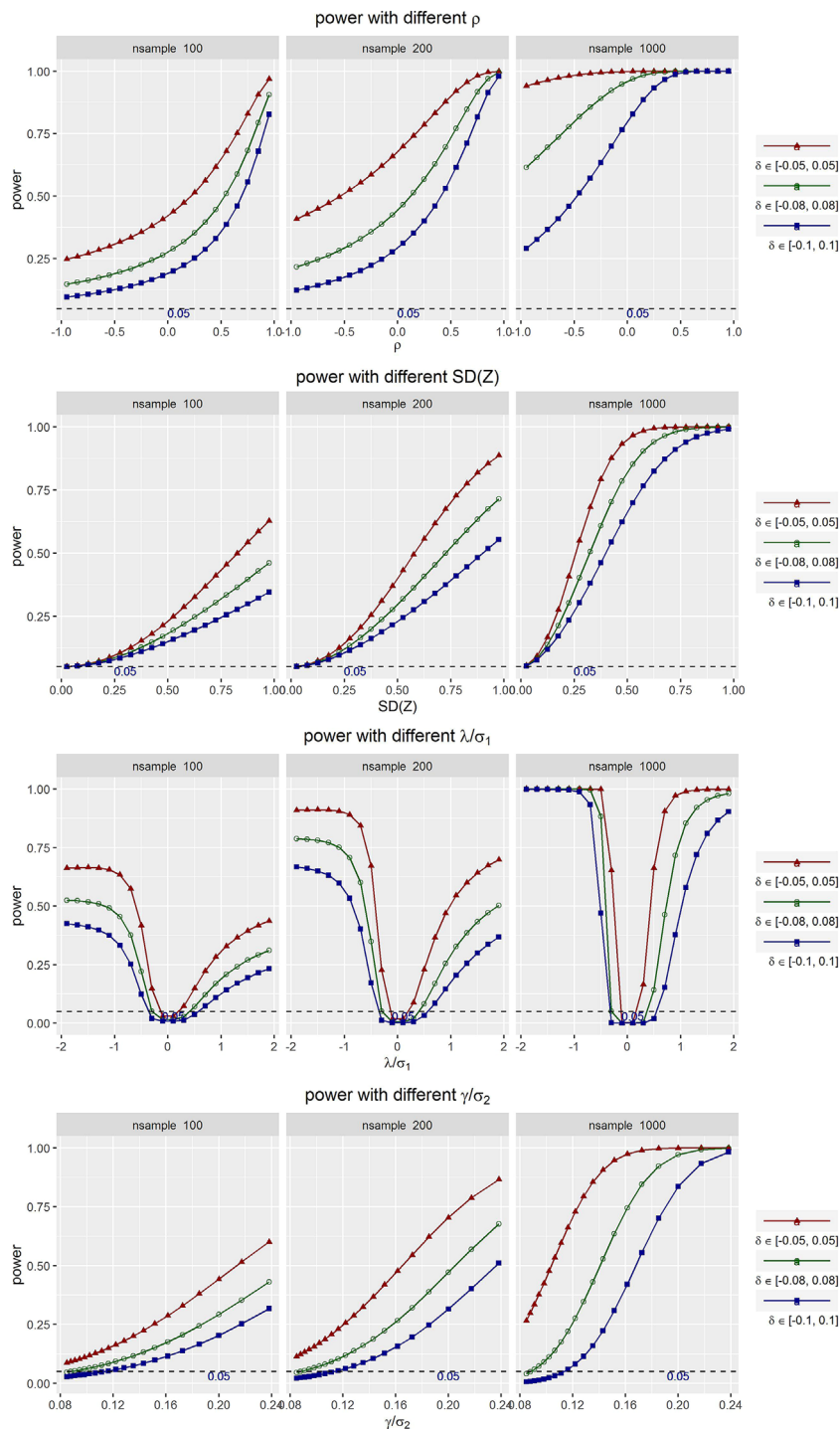


Figure 3. Power of sensitivity analysis in simulated scenario where the base parameters are $\sigma_1^2 = 1$, $\sigma_2^2 = 4$, $\rho = 0.5$, $\gamma = 0.5$, $\lambda = -1$, $sd(Z) = 1$. In each row, we vary one combination of parameters to observe the change of power when sample size $n = 100, 200$, and 1000 . The dashed line is the significant level $\alpha = 0.05$.

confounding there is. We will consider $p = 0.3$, a moderate asymptotic normal distribution, confounding effect, in most of the analysis below.

First, without sensitivity analysis, we calculate the necessary sample size needed for rejecting the null hypothesis with power greater than 0.8. If we use the t test based on

$$n = \frac{(z_{\alpha/2} + z_{\beta})^2 V}{\lambda^2 \rho_{ZD}^2}, \quad (27)$$

where $V = \sigma_1^2 / \text{Var}(D)$. By formula (27), we need a sample size of 4301 if we want the power greater than 0.8. However, if we use the AR test and plug the parameters (26) into the power formula (16), then we would need a sample size of 7085, larger than the sample size of 4301 suggested by (27). Although this would seem to suggest that using the t-test based on the asymptotic normal distribution can reduce the sample size needed compared to the AR test, the sample size needed for the t-test based on the asymptotic normal distribution from (27) cannot be trusted while the sample size needed for the AR test from (16) can be trusted. There are two reasons for this: (i) the nominal level of the t-test based on the asymptotic normal distribution is not reliable for finite samples and can be much greater than the actual level while the nominal level

of the AR test is exactly equal to the true level regardless of the sample size. (ii) the power formula (27) may not be accurate for finite samples while the formula (16) is an exact formula.

Figure 4 illustrates reason (i), in which the null hypothesis is tested both using the t-test based on the asymptotic normal distribution and the AR test where the null hypothesis is true. The top panel considers large confounding while the bottom panel considers small confounding. The strength of the IV is varied along the x-axis. 10,000 data sets are simulated with 5000 observations each and the rejection rate is displayed on the y-axis. The standard error of the rejection rate for a 0.05 level test from 10,000 simulations is smaller than 0.005. These simulations show that the AR test always has level about 0.05, equal to its nominal level, while the t-test based on the asymptotic normal distribution can have level way above its nominal level (an actual level of 0.25 compared to the nominal level of 0.05 in the top panel for a weak IV) or level way below its nominal level.

To illustrate reason (ii) about the power formula (27) being less accurate than the power formula (16), we consider the same setting described above and simulated 10,000 data sets with 4301 and 7085 observations from (26) with $p = 0.3$ and used the t-test. The power of 4301 observations is 0.7266 compared to the 0.8 that (27) says the power should be and the power of 7085 observations is 0.8747. In contrast, when simulating 10,000 data sets with 7085 observations from (26) with $p = 0.3$ and using the AR test, the power is 0.8002 compared to 0.8, as formula (16) said the power should be.

Now, we consider sensitivity analysis using the AR test for model (26) with $p = 0.3$. Suppose, we would like to conduct a sensitivity analysis for $\delta \in [-0.01, 0.01]$, which means a one unit change in the IV could lead to up to a 0.01 standard deviation change of the structural error. By the power formula (22), we need at least 8845 observations to achieve power at least 0.8 under the favorable situation $\delta = 0$. Figure 5 further explores the relationship between the sample size needed and the allowance of sensitivity. We can see the curve starts flat for a while and then turns steep sharply. This suggests that if the allowance of sensitivity is within a small range and we want to perform sensitivity analysis, we do not need to increase the sample size much to achieve the same power as without sensitivity analysis ($\delta = 0$). However, after a certain threshold, even allowing an extra little amount of sensitivity will result in a large increase of the sample size. In this scenario, the design sensitivity is 0.0499. If the range of sensitivity is greater than $(-0.0499, 0.0499)$, then the power will be close to zero no matter how large the sample size.

5.2. Sensitivity Analysis Using Rare versus Common Variants as IVs

Variants with minor allele frequency (MAF) $> 5\%$ are often referred to as common variants while those with MAF between 1 and 5% are referred to as low frequency variants and those with MAF $< 1\%$ are rare variants. Common variants tend to have small effects because a common variant with a large deleterious (beneficial) effect would be selected against (for) by evolution. Rare variants could have large deleterious effects and still not be selected against by evolution. Those rare variants that are associated with the

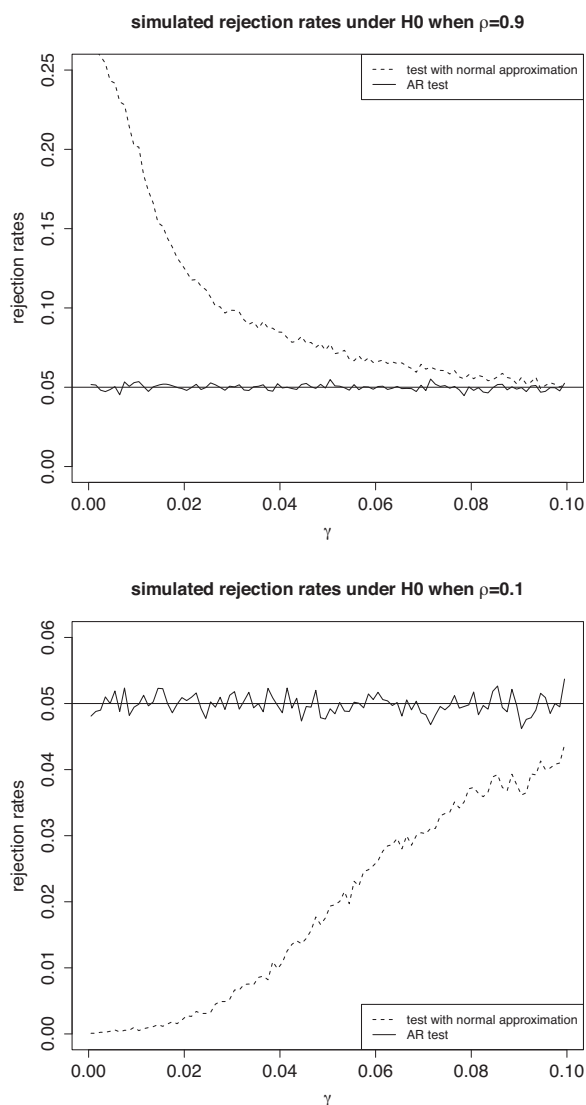


Figure 4. Data set is generated as $\sigma_1 = \sigma_2 = 1$, $\rho = 0.9(0.1)$, $\beta = 0$, $\gamma \in [0, 0.1]$, $\text{sd}(Z) = 1$ with sample size 5000. Test with normal asymptotic distribution and the standard AR test is performed with nominal significance level $\alpha = 0.05$. We calculate the average rejection rate among 20,000 simulated data sets.

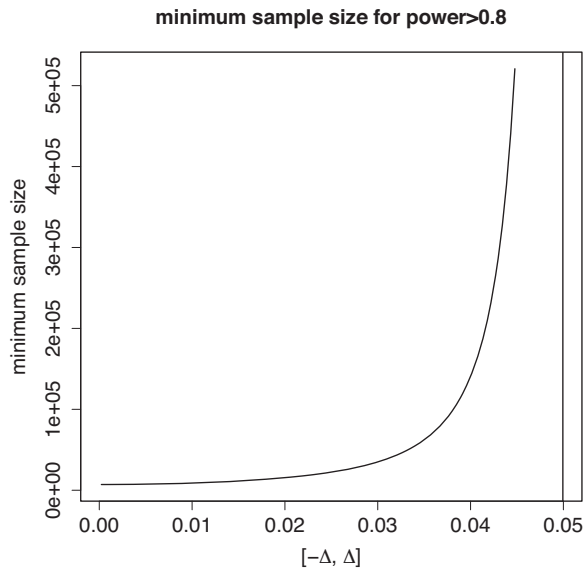


Figure 5. Sample size needed for achieving power >0.8 under different allowance of sensitivity. The vertical line stands for the design sensitivity 0.0499. Here, $\lambda = 0.234$, $\delta = 0$, $\gamma = 0.158$, $\sigma_1^2 = 0.333$, $\sigma_2^2 = 1.0989$, $\rho = 0.548$.

disease in a study with moderate sample size have to have a reasonably large effect. See Gibson (2012) and Zuk et al. (2014) for discussion about rare versus common variants. Most Mendelian randomization studies have focused on using common variants but there is increasing opportunity for using rare variants by making use of next generation sequencing Zuk et al. (2014). Furthermore, next generation sequencing techniques also facilitate the possibility of using structural variation (e.g., deletions, duplications, copy-number variants, insertions, inversions, and translocations) in Mendelian randomization studies, which often has larger effects than common SNP variants.

Here, we compare the power of using a common variant versus a rare variant. To make the comparisons, we assume $\sigma_1 = \sigma_2 = 1$, $\rho = 0.5$, $\beta = 1$. Suppose, the rare variant takes 0/1 with probability 0.995/0.005 and the IV effect is $\gamma_r = 0.142$ while the common variant takes 0/1 with probability 0.95/0.05 and the IV effect is $\gamma_c = 0.046$. By choosing these IV effect sizes, the rare and common variants have the same concentration parameter under the same sample size. We can use (22) to calculate the power under different sample size

and sensitivity. We investigate the scenarios where the sample size is $\{10^3, 10^4, 10^5, 10^6\}$ and the sensitivity allowance is $\{(0, 0), (-0.02, 0.02), (-0.05, 0.05)\}$. Results are in Table 1.

We see that if there's no concern about the IV being invalid ($\Delta = 0$), the power for the rare and common variants are exactly the same across different sample sizes since they have the same concentration parameter. However, if we allow for some of amount of IV invalidity and calculate the power of sensitivity analysis, then the rare variant has better power than the common variant. As discussed in Section 4.2, for power of sensitivity analysis, the IV effect size $|\gamma|/\sigma_2$ plays a more important role than the concentration parameter. The rare variant in Table 1 has a larger IV effect size and consequently a higher power of sensitivity analysis. Another thing to be noted is that when $\Delta = 0.02$, the power of sensitivity analysis increases as sample size increases for both rare and common variants, but when $\Delta = 0.05$, the power of sensitivity analysis decreases as sample size increases for the common variant. This is because the allowance of sensitivity is too large here such that the inequality (25) holds and the power of sensitivity analysis goes to zero.

In summary, if a rare variant has a larger effect size than a common variant such that the rareness and effect size balance each other to result in the same concentration parameter, then the rare variant has larger power of sensitivity analysis if there is concern about the IV being invalid. In Table 1, the magnitude Δ of IV invalidity for the common and rare variants is the same. Rare variants could tend to be more invalid because they are more likely to cluster in families, thus bringing about more population stratification, and variants with large are more susceptible to canalization Lawlor et al. (2008), meaning there is a direct effect of the IV on the outcome. Web Table 1 in Supplementary Materials examines the effect of the Δ for a rare variant being bigger than that for a common variant.

6. Discussion

We have developed a method of sensitivity analysis and a power formula of sensitivity analysis for causal studies using IVs based on the AR test. Compared to previously developed methods of sensitivity analysis for IVs, our method is not strongly dependent on instrument strength. We have shown that when designing causal studies using IVs in which there is concern that the IV might not be perfectly valid, the key strength parameter one should consider about the IV is not the IV's concentration parameter but instead the effect size of

Table 1

Power for rare(common) variants under different sample size and sensitivity. We set $\sigma_1 = \sigma_2 = 1$, $\rho = 0.5$, $\lambda = 1$. For rare variant, $\gamma_r = 0.142$, $SD(Z)_r = 0.071$ and for common variant, $\gamma_c = 0.046$, $SD(Z)_c = 0.218$. In doing so rare and common variants have the same concentration parameter under the same sample size. The numbers in parentheses represent the power for common variants.

	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
Concentration parameter	0.1	10	100	1000
$\Delta=0$	0.054 (0.054)	0.089 (0.089)	0.447 (0.447)	0.999 (0.999)
$\Delta=0.02$	0.054 (0.052)	0.086 (0.03)	0.377 (0.116)	0.997 (0.409)
$\Delta=0.05$	0.052 (0.042)	0.071 (0.016)	0.175 (0.001)	0.726 (0.000)

the IV on the exposure. An extension to a multiple IV model is presented in the Supplementary Materials.

7. Supplementary Materials

Web Appendices and Tables referenced in Sections 2, 3, 4, and 5 as well as a zip file with software and examples are available with this article at the *Biometrics* website on Wiley Online Library.

ACKNOWLEDGEMENTS

The authors would like to thank the editor, associate editor, and reviewers for their constructive comments and suggestions. Xuran Wang and Yang Jiang contributed equally to this work and are co-first authors.

REFERENCES

- 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**, 46–63.
- Andrews, D. W., Moreira, M. J., and Stock, J. H. (2006). Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica* **74**, 715–752.
- 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**, 444–455.
- Baiocchi, M., Cheng, J., and Small, D. (2014). Tutorial in biostatistics: Instrumental variable methods for causal inference. *Statistics in Medicine* **33**, 2297–2340.
- Brookhart, M. A. and Schneeweiss, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *The International Journal of Biostatistics* **3**, 14.
- Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *Review of Economics and Statistics* **94**, 260–272.
- Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. OUP Catalogue, Oxford, UK: Oxford University Press.
- Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research* **16**, 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**, 271–310.
- Ebrahim, S. and Smith, G. D. (2008). Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics* **123**, 15–33.
- Freeman, G., Cowling, B. J., and Schooling, C. M. (2013). Power and sample size calculations for mendelian randomization studies using one genetic instrument. *International Journal of Epidemiology* **42**, 1157–1163.
- Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics* **13**, 135–145.
- Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology* **175**, 332–339.
- Hausman, J. (1983). Specification and estimation of simultaneous equation models. *Handbook of Econometrics* **1**, 391–448.
- Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* **17**, 360–372.
- Holland, P. W. (1988). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology* **18**, 449–484.
- Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69**, 803–811.
- Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science* **29**, 323–358.
- Imbens, G. W. and Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **168**, 109–126.
- Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. L., and Imbens, G. W. (2011). Identification and inference with many invalid instruments. Technical report, National Bureau of Economic Research.
- Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163.
- Moreira, M. J. (2001). *Tests with Correct Size when Instruments can be Arbitrarily Weak*. Berkeley: Center for Labor Economics, University of California.
- Morgan, S. L. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press.
- Nelson, C. R. and Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business* **63**, 125–140.
- Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whitaker, J. C., and Leon, D. A. (2006). Limits to causal inference based on mendelian randomization: A comparison with randomized controlled trials. *American Journal of Epidemiology* **163**, 397–403.
- Rosenbaum, P. (2010). *Design of Observational Studies*. Springer Series in Statistics: Springer-Verlag New York.
- Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91**, 153–164.
- Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics* **2**, 881–935.
- Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102**, 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**, 924–933.
- Smith, G. D. and Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1–22.
- Smith, G. D. and Ebrahim, S. (2004). Mendelian randomization: Prospects, potentials, and limitations. *International Journal of Epidemiology* **33**, 30–42.
- Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics* **20**, 518–529.
- Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica* **16**, 953–979.

- Wang, J. and Zivot, E. (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* **66**, 1389–1404.
- Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., et al. (2014). Searching for missing heritability: Designing rare variant association studies.

Proceedings of the National Academy of Sciences **111**, 455–464.

Received April 2017. Revised February 2018.

Accepted February 2018.