Empir Econ (2015) 49:1153–1171
DOI 10.1007/s00181-015-0916-0

Sensitivity analysis for inference in 2SLS/GMM estimation with possibly flawed instruments

Richard A. Ashley · Christopher F. Parmeter

Received: 29 May 2012 / Accepted: 11 December 2014 / Published online: 12 February 2015

Abstract Credible inference requires attention to the possible fragility of the results (p values for key hypothesis tests) to flaws in the model assumptions, notably accounting for the validity of the instruments used. Past sensitivity analysis has mainly consisted of experimentation with alternative model specifications and with tests of overidentifying restrictions which actually presuppose instrument validity. We provide a feasible sensitivity analysis of two-stage least-squares and GMM estimation, quantifying the fragility/robustness of inference with respect to possible flaws in the exogeneity assumptions made, and also indicating which of these assumptions are most crucial. The method is illustrated via application to a well-known study of the education–earnings relationship.

Keywords Robustness · Invalid instruments · Flawed instruments · Instrumental variables · Sensitivity analysis · Two-stage least squares

JEL Classification C23

This paper has benefitted greatly from comments by Oscar Mitnik, Alfonso Flores-Lagunes and participants at MEG 2010.

R. A. Ashley
Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
e-mail: ashleyr@vt.edu

C. F. Parmeter (✉)
Department of Economics, University of Miami, Coral Gables, FL, USA
e-mail: cparmeter@bus.miami.edu

1 Introduction

The practice of econometrics begins with the specification and estimation of a model, or several models; it should then focus on assessing how the assumptions made impact the inferential results of greatest interest. This assessment can address functional form variants, additional covariates, alternative estimation methods and so forth; here we focus on the validity of the exogeneity assumptions made in the context of instrumental variables estimation and inference. This assessment process is a *sine qua non* for careful empirical work.

The present work is a generalization of a sensitivity analysis framework originally proposed in Ashley (2009), which addressed the robustness of an inference p value based on an exactly identified instrumental variables model with respect to instrument validity failures. Here we extend the reach of this type of sensitivity analysis to 2SLS and GMM model estimation. By 'inference p value'—here, and below—we mean the p value at which some null hypothesis of interest can be rejected; quite often, the main thrust of an empirical study is to test just one or two such hypotheses. Thus, the new version of this approach proposed here is applicable to a far wider class of econometric circumstances.

The basic idea here is to examine the sensitivity of key inference p values to a comprehensive set of explicit potential flaws in the exogeneity assumptions underlying 2SLS and GMM estimation/inference and to thereby explicitly quantify the degree to which the results one cares about most are robust (or fragile) with respect to reasonably likely departures from these assumptions. Specifically, how large a correlation between one or more of the instruments (or first-stage variables) and the original (structural) model errors is sufficient to overturn the statistical significance of a particular inference of interest? And does the fragility lie more with respect to flaws in one of the first-stage variables (instruments) or another?

More specifically, our sensitivity analysis works as follows. Suppose that the main point of the analysis is the finding that a particular null hypothesis can be rejected at, say, the 5 % level. And, for expository clarity only, consider the case where just one instrument ('z') used in the model estimation is possibly flawed. Thus, corr(z, u) is thought to possibly be nonzero, where u denotes the model structural error. The value of corr(z, u) is, of course, unobservable. For any given value of the *covariance* cov(z, u), however, we can obtain a consistent estimate of the sampling distribution of the model parameters, consistent estimates of the model parameters themselves and a concomitant consistent estimate of the variance of u. Conditional on this given value for cov(z, u), this resulting estimated sampling distribution yields an asymptotically valid rejection p value for the null hypothesis of interest, and this estimate of the variance of u yields a consistent estimate of corr(z, u). By randomly drawing a large number of values for cov(z, u), we are thus able to numerically delineate exactly how flawed the instruments need to be—as quantified by corr(z, u)—in order to overturn the finding that the original null hypothesis of interest can be rejected at the 5 % level. If this value of corr(z, u) is small in magnitude, then the original inference is 'fragile' with respect to instrument flaws, whereas this inference is 'robust' if the magnitude of corr(z, u) is substantial.

We should note here that our proposed sensitivity analysis differs from existing econometric approaches that allow for violations of the requisite exogeneity conditions necessary to have a valid instrument. A non-exhaustive list includes Hahn and Hausman (2006) and Berkowitz et al. (2008a) who impose a condition of 'near exogeneity' on the instrument-error correlation that ensures that the finite-sample correlation, while nonzero, is vanishingly small—i.e., of order $1/\sqrt{n}$, where n is the sample length. This type of condition has been termed 'local-to-zero' correlation. Berkowitz et al. (2008b) assume that the instrument-error correlation is nonzero and vanishingly small—although not of order $1/\sqrt{n}$ as in Berkowitz et al. (2008a)—and provide a data-driven resampling scheme to compute confidence intervals for the structural parameters of interest; this eliminates the need for a specific distributional assumption on the instrument-error correlation. However, a key concern with these local-to-zero approaches is: why is it reasonable to think that a larger quantity of sample data makes one's instruments any less flawed?[1]

The plan of the paper is as follows. Section 2 discusses why sensitivity analysis is an integral part of sound econometric research. Section 3 derives the straightforward modification of GMM (and 2SLS) analytics needed for the sensitivity analysis proposed here. Section 4 details precisely how to implement the proposed sensitivity analysis in practice. The application of this algorithm is then illustrated with several examples in Sect. 5. Section 6 summarizes the results and provides concluding remarks on the proposed sensitivity analysis.

2 Why sensitivity analysis?

Angrist and Pischke (2010) recently reviewed the practice of econometrics as developed following the seminal work of Leamer (1983), which called attention to the lack of credible econometric work being done at that time. They note that Leamer's (1983) recommendation of sensitivity analysis is no longer the only feasible way to lend credibility to empirical analysis, as important data sources based on organized and/or 'natural' experiments are now becoming available to an extent that he did not envisage (Angrist and Pischke 2010). Empirical analysis of non-experimental data can be seen as a complementary approach to the standard practice of conducting sensitivity analysis to the assumptions imposed in applied econometric work. However, we—in common with Stock (2010)—still see a hugely important role for sensitivity analysis with respect to the assumptions necessary for analyzing empirical data. Many of these assumptions are made in examining experimental data as well, as pointed out by Keane (2010b).

Subsequent to Leamer's (1983) biting criticism of the then-current practices in applied econometric work, the profession's response has been threefold. One response has been the development and widespread adoption of robust standard error estimators

[1] In a Bayesian context, Conley et al. (2007) and Kraay (2008) provide methods for working with instruments that do not fully satisfy the knife-edge orthogonally condition; these authors assume that the instruments are nearly exogenous and attach Bayesian priors to the correlation parameters. Other approaches in the literature include Murray (2006), Ebbes et al. (2009) and Small (2007). See Ashley and Parmeter (2013) for a more detailed literature review.

(White-Eicker and HAC), so as to make the analysis more resilient with respect to failures in some of the key model assumptions. The use of these methods has become increasingly feasible as typical sample sizes have grown. However, it should be noted, these methods have also handed analysts a perceived freedom to ignore the misspecification signals provided by heteroscedasticity and/or serial correlation in the model error term: neglected dynamics in the case of serial correlation and neglected nonlinearity (and/or neglected heterogeneity) in the case of heteroscedasticity.[2] A second response has been the use of 'extreme bounds' sensitivity analysis, which amounts to asking whether one's conclusions are robust with respect to trying out a variety of different covariates. This approach suffers from a fundamental flaw: The meaning of all of the population regression coefficients are contingent on the set of conditioning variables used. Inference stability with respect to such variation in the model specification is therefore irrelevant to the issue of the validity of the econometric assumptions and, in particular, the exogeneity assumptions underlying the inferences made in the original model. And, while instability with respect to variation in the instrument choices can be meaningful, it is very difficult to interpret; in particular, such instability does not necessarily imply that any of the instruments are invalid.

Finally, analysts now routinely make tests of the over-identifying assumptions in their models. However, per Keane (2010a, footnote 10) and Angrist and Pischke's (2010) comments [see also Stock's (2010) rejoinder to Angrist and Pischke], it is essential to note that the validity of any test of the over-identifying restrictions is still conditional on the assumed validity of the exogeneity assumptions made. If such a test rejects the null hypothesis that the over-identifying restrictions are correctly specified, this result only implies that something is misspecified, but it is not at all clear what, whereas if such a test fails to reject its null hypothesis, then—as with any hypothesis test—it is not really possible to draw any conclusion at all.

3 Estimation and inference with explicitly flawed exogeneity assumptions

The sensitivity analysis proposed here extends the work of Ashley (2009) by allowing for over-identification in IV models. While the approach can be easily extended to the setting of a nonlinear model, for expositional clarity the sensitivity analysis is described in the remainder of this section (and applied in Sect. 5) for the special case of a linear structural model with potentially heteroscedastic and/or serially correlated errors.

Consider, then, the standard linear model, with the structural equation

$$Y_1 = Y_2\alpha + Z_1\beta + \varepsilon, \tag{1}$$

where Y_2 is an $n \times g$ matrix of other endogenous variables, Z_1 is an $n \times k$ matrix of exogenous variables, α and β are $g \times 1$ and $k \times 1$ vectors of coefficients, respectively, and ε is the $n \times 1$ matrix of structural errors. Z_2, an $n \times j$ matrix of additional, purportedly exogenous, instrumental variables is assumed to be available, which—if

[2] See Sims (2010, pp. 66–67) with respect to the heterogeneity issue.

these instruments are valid—can be used to correct for the endogeneity of Y_2, thereby providing consistent estimates of α and β. It is assumed here that $j \geq g$, corresponding to either exact identification ($j = g$) or over-identification ($j > g$).

Accordingly, consistent least-squares estimation of Eq. 1 requires that the k conditions

$$E\left[Z_{1i}'\varepsilon_i\right] = 0 \tag{2}$$

and the j instrument validity conditions

$$E\left[Z_{2i}'\varepsilon_i\right] = 0, \tag{3}$$

hold for all $i \in [1, n]$. The condition in Eq. 2 incorporates the assumed exogeneity of the k variables in Z_1. The condition in Eq. 3 specifies that the j instruments are valid—i.e., uncorrelated with ε; this, of course, is equivalent to assuming that the j variables that compose Z_2 are actually exogenous.

Here, however, the point of the exercise is to obtain estimators for the case where one or more of the j instruments are flawed, i.e., not exogenous. In that case, Eq. 2 still holds, but Eq. 3 is modified to become

$$E\left[Z_{2i}'\varepsilon_i\right] = \Sigma_{Z_2\varepsilon}' \neq 0, \tag{4}$$

for all i.

Letting $\gamma' = [\alpha' \quad \beta']$ and $X = [Y_2 \quad Z_1]$, Eq. 1 can be written more compactly as:

$$Y_1 = X\gamma + \varepsilon. \tag{5}$$

And, letting $Z = [Z_1 \quad Z_2]$, the parameter vector γ in Eq. 5 can now be estimated via GMM, by minimizing:

$$\left[n^{-1}\sum_{i=1}^{n}\left(Z_i'\hat{\varepsilon}_i - \Sigma_{Z\varepsilon}'\right)\right]' \widehat{\mathcal{W}}^{-1} \left[n^{-1}\sum_{i=1}^{n}\left(Z_i'\hat{\varepsilon}_i - \Sigma_{Z\varepsilon}'\right)\right], \tag{6}$$

over $\hat{\gamma}$ where $\hat{\varepsilon}_i = Y_{1i} - X_i\hat{\gamma}$ and the subscript i denotes the ith row of the matrices Z, Y_1, and X defined above. Then, under the usual GMM assumptions, e.g., as given in Wooldridge (2010, Theorems 14.1 and 14.2),

$$\hat{\gamma}^{\text{flaw}} = (X'Z\widehat{\mathcal{W}}^{-1}Z'X)^{-1}X'Z\widehat{\mathcal{W}}^{-1}\left(Z'Y_1 - n\Sigma_{Z\varepsilon}'\right), \tag{7}$$

and the asymptotic sampling distribution of $\hat{\gamma}^{\text{flaw}}$ for given $\mathcal{W}$ and $\Sigma_{Z\varepsilon}$ is:

$$\sqrt{n}(\hat{\gamma}^{\text{flaw}} - \gamma) \xrightarrow{d} N\left(0, \left(X'Z\widehat{\mathcal{W}}^{-1}Z'X\right)^{-1}\mathcal{B}\left(X'Z\widehat{\mathcal{W}}^{-1}Z'X\right)^{-1}\right), \tag{8}$$

where

$$\mathcal{B} = X'Z\widehat{\mathcal{W}}^{-1}\Lambda\widehat{\mathcal{W}}^{-1}Z'X \tag{9}$$

and

$$\Lambda = E\left[\left(Z_i'\varepsilon_i - \Sigma_{Z\varepsilon}'\right)\left(Z_i'\varepsilon_i - \Sigma_{Z\varepsilon}'\right)'\right]. \tag{10}$$

Estimating the sampling distribution of $\hat{\gamma}^{\text{flaw}}$ is now straightforward, as a consistent estimator of Λ is given by:

$$\widehat{\Lambda} = n^{-1}\sum_{i=1}^{n}\left(Z_i'\hat{\varepsilon}_i - \Sigma_{Z\varepsilon}'\right)\left(Z_i'\hat{\varepsilon}_i - \Sigma_{Z\varepsilon}'\right)', \tag{11}$$

where $\hat{\varepsilon}_i$ is any consistent estimator of the structural error, ε_i. Substituting $\widehat{\mathcal{W}} = Z'Z$ into Eq. 7, then yields:

$$\hat{\gamma}^{\text{2SLS-flaw}} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}X'Z(Z'Z)^{-1}\left[Z'Y_1 - n\Sigma_{Z\varepsilon}'\right], \tag{12}$$

which is labelled '2SLS' here because it reduces to the usual 2SLS estimator of γ in the special case of valid instruments, where $\Sigma_{Z\varepsilon}'$ is zero. The fitting errors implied by $\hat{\gamma}^{\text{2SLS-flaw}}$—i.e., $Y_1 - X\hat{\gamma}^{\text{2SLS-flaw}}$—are analogously denoted $\hat{\varepsilon}^{\text{2SLS-flaw}}$ below. The estimator $\hat{\gamma}^{\text{2SLS-flaw}}$ is a consistent estimator of γ for a given value of the instrument flaw covariance vector (as shown above for any choice of $\widehat{\mathcal{W}}$) and also asymptotically normal. But it is not asymptotically efficient, because $\widehat{\mathcal{W}} = Z'Z$ is not the optimal weighting matrix.

The optimal weighting matrix is obtained, as usual, on a second pass, setting $\mathcal{W}^{\text{opt}} = \widehat{\Lambda}$ from Eq. 11, substituting $\hat{\varepsilon}^{\text{2SLS-flaw}}$ for $\hat{\varepsilon}$. Because $\widehat{\mathcal{W}}$ now equals $\widehat{\Lambda}$, Eqs. 7 and 8 through 11 now yield what we will call $\hat{\gamma}^{\text{GMM-flaw}}$ and its estimated sampling distribution:

$$\hat{\gamma}^{\text{GMM-flaw}} = \left(X'Z\widehat{\Lambda}^{-1}Z'X\right)^{-1}X'Z\widehat{\Lambda}^{-1}\left(Z'Y_1 - n\Sigma_{Z\varepsilon}'\right) \tag{13}$$

and

$$\sqrt{n}(\hat{\gamma}^{\text{GMM-flaw}} - \gamma) \xrightarrow{d} N\left(0, \left(X'Z\widehat{\Lambda}^{-1}Z'X\right)^{-1}\right), \tag{14}$$

for any specified value of $\Sigma_{Z\varepsilon}'$. This estimator is preferable to $\hat{\gamma}^{\text{2SLS-flaw}}$ because—per Wooldridge (2010) and Hall and Inoue (2003)—it is asymptotically efficient. We therefore use $\hat{\gamma}^{\text{GMM-flaw}}$ in the illustrative examples in Sect. 5 below; we note, however, that the sensitivity analysis proposed here in no way requires efficient parameter estimation.

It is then straightforward, using the estimated asymptotic sampling distribution of $\sqrt{n}(\hat{\gamma}^{\mathrm{GMM-flaw}} - \gamma)$ given in Eqs. 13 and 14, to calculate the rejection p value for any specific null hypothesis regarding the components (or functions of the components) of the structural parameter (γ), predicated upon any given value of the instrument flaw covariance vector, $\Sigma_{Z\varepsilon}$. The next section lays out how this result can be used to quantify the fragility (or robustness) of this particular rejection p value with respect to possible flaws in the instruments.

4 Implementation of the sensitivity analysis

In the description below, it is supposed (for expositional clarity) that a particular null hypothesis regarding γ has been rejected at the 5 % level, based on the usual 2SLS or GMM estimator and that the question at issue is how sensitive this result is to flaws in the exogeneity of one or more of the j variables in Z_2, leading to a failure in the population moment conditions given in Eq. 4 above as $E\left[Z_i'\varepsilon_i\right] = \Sigma'_{Z\varepsilon} \neq 0$.[3] Below, the number of instruments being analyzed is denoted 'm', but all k exogenous variables and all of the j 'actual' instruments are, of course, always used in the estimation of Eq. 1.

Recalling that the $k + j$-vector $\Sigma'_{Z\varepsilon}$ is the population covariance between the instruments and the model error term in Eq. 1, the sensitivity analysis consists of the following steps:[4]

1. Values for the m nonzero components of the instrument-error covariance vector $\Sigma_{Z\varepsilon}$ are randomly generated as an independent drawing from the multivariate normal distribution[5] with mean zero and variance-covariance matrix equal to σ^2 times the sample variance-covariance matrix of the m relevant variables; σ^2 is initially set equal to s^2, the sample estimate of the variance of ε under the assumption that the instruments are unflawed. The value of σ^2 is adjusted as needed at a later stage; this adjustment is described below, at the end of step 4.
2. Contingent on this value of $\Sigma'_{Z\varepsilon}$, the GMM parameter estimate ($\hat{\gamma}^{\mathrm{GMM-flaw}}$) and its sampling distribution are then obtained from Eqs. 11 to 14. Based on this sampling distribution, the new p value for the null hypothesis of interest is then calculated.
3. Also contingent on this value of the instrument-error covariance vector $\Sigma'_{Z\varepsilon}$, $\hat{\gamma}^{\mathrm{GMM-flaw}}$ provides a consistent estimator of γ for use in calculating $Y_1 -$

[3] The analysis would be essentially identical for a rejection at the 1 % (or any other) level: The description in this section is made definite for the 5 % level solely to enhance the clarity of the exposition. Similarly, the procedure described below can be readily modified to instead analyze the case where the null hypothesis is *not* rejected at the 5 % level and the issue is whether this *failure* to reject is due to a flaw in the exogeneity conditions, or where the null hypothesis is either joint or a nonlinear function of γ.

[4] This algorithm is implemented in R code (available from the authors), but the entire process is readily programmed in any matrix-oriented computer language.

[5] The multivariate normal distribution is used here for computational convenience only; the sensitivity analysis results are themselves insensitive to the alternative use of a diagonal variance-covariance matrix or even a different distribution (e.g., Wishart) entirely. Indeed, the instrument-error covariance vectors are randomly selected solely to sample the m-dimensional space in a computationally straightforward fashion. Non-random selection over a sufficiently dense grid would yield the same results with a sufficiently large number of drawings, but the computational burden of such a grid search does not scale well with m.

$X\hat{\gamma}^{\mathrm{GMM-flaw}}$, which are asymptotically equivalent to the structural model errors, ε. The sample variance of these fitting errors is thus a consistent estimator of the variance of ε. This estimate is then used (along with the sample variances of the instruments) to convert the instrument-error covariance vector $\Sigma_{Z\varepsilon}$ into a vector of instrument-error *correlations*, which are more interpretable than the corresponding covariances.

4. The foregoing calculation is not computationally burdensome, so it is quite feasible to repeat it for a large number (M_{rep}) of randomly drawn $\Sigma_{Z\varepsilon}$ instrument-error covariance vectors, at each such repetition generating a p value for the null hypothesis being analyzed and the associated vector of m instrument-error correlations. These m instrument-error correlations are written out to a 'results' file for each repetition generating a p value larger than 0.05—i.e., at each repetition for which the flaws in the instruments corresponding to $\Sigma_{Z\varepsilon}$ have 'overturned' the original inference result, assumed (at the beginning of this section) to be significant at the 5 % level. If the original inference result was a *failure* to reject the null hypothesis at the 5 % level, then this m-vector of implied instrument-error correlations is written out only for the repetitions in which the p value is *smaller* than 0.05.
   A reasonable value must be specified for σ^2, the variance of the distribution used to randomly generate the $\Sigma_{Z\varepsilon}$ vectors, so that the calculations make good use of the M_{rep} repetitions. If σ^2 is set too small, then an unnecessarily large value of repetitions is needed in order to generate a substantial 'sample' of inference overturns, whereas if σ^2 is set too large, then an unnecessarily large value of repetitions is needed in order to generate a sufficiently large number of inferences which are barely overturned. The value chosen for σ^2 is itself inconsequential, so long as M_{rep} is set large enough that the sensitivity results—e.g., the values of r_{min} discussed in the next step—are themselves insensitive to a repeat of the analysis with a new seed for the random number generator. In practice, we find that the sensitivity results are essentially invariant to the value chosen for σ^2—over several orders of magnitude—when M_{rep} is set in the range of 10^4–10^5. Consequently, we simply choose σ^2 to be sufficiently large as to yield at least several inference overturns with 1,000 repetitions. We then divide this value of σ^2 by ten, raise M_{rep} to 10^4–10^6 (depending on the size of m) and check to make sure that the results are robust to a repeat of the calculation with a new seed for the random number generator.
   Where the value of m exceeds one—i.e., where possible flaws in more than one instrument are being analyzed—then it is very useful to define r as the length of this vector of instrument-error correlations for each repetition and to write it to the 'results' file also. These length values facilitate the condensation and descriptive characterization of the results, described in the next step.[6]
5. This collection of M_{rep} calculated (m+1)-vectors—each of which consists of the vector of m instrument-error correlation components and its length (r)—can then be analyzed in several ways:

[6] Our code uses the Euclidean norm—the square root of the sum of the squares of the m components—for this length measure. While obviously not the only possible choice, this norm emphasizes the importance of the components which are largest in magnitude, which likely contributes to descriptive clarity.

(a) The first thing to look at is $r_{\min}$, the minimal value of r. What is the smallest instrument-error correlation vector length for which the null hypothesis is no longer rejected?
(b) Next, it is useful to also tabulate $r_{0.01} \ldots r_{0.20}$, where $r_{0.20}$, for example, is the length of the instrument-error correlation vector such that 20 % of all the repetitions leading to an 'overturn' of the original inference result are due to instrument-error correlation vectors no larger than this. These statistics address the question: What is the smallest instrument-error correlation vector length for which the original inference is overturned with some frequency?[7]
(c) The sensitivity analysis results can also be displayed graphically. The most useful ways in which this can be done are best illustrated in the empirical examples given in the next three sections, but they are described here, for completeness:
    i. First, in the special case where just two instrument-error covariances are being varied—i.e., where m equals two—it is both feasible and informative to simply plot a few thousand of these (two-dimensional) instrument-error correlation vectors for which the inference was 'overturned,' one component against the other.
    ii. Second, it is always feasible (and usually informative) to plot what might be called an 'Empirical Cumulative Distribution Function' or 'ECDF' of the lengths of the instrument-error correlation vectors corresponding to overturns of the original inference. The height of this plot is zero for instrument-error correlation vector lengths less than $r_{\min}$, its height reaches 0.01 for length equal to $r_{0.01}$, its height reaches 0.05 for length equal to $r_{0.05}$, and so forth.
(d) Finally, it is useful to examine the pattern of the m components of the vector of the instrument-error correlations corresponding to $r_{\min}$. If there is, in fact, some fragility to this inference regarding γ, then the pattern in these components points at which instruments are the primary sources of this fragility and which are not.

The next section illustrates the nature and usefulness of this sensitivity analysis algorithm via two examples. The first example uses artificial data generated by a very simple 'macroeconomics-like' system of simultaneous equations in which two of the three instruments are actually flawed. Two inferences with regard to the key structural coefficient in this model are analyzed. Under the usual assumption of valid instruments, one of these null hypotheses is rejected at the 5 % level and the other is not; but both inferences turn out to be fairly fragile with respect to instrument flaws. The next example is a replication of a well-known empirical study drawn from the labor economics literature—and its key inference result turns out to be arguably quite fragile. Based, in part, on these examples, the paper closes with some thoughts on how one might sensibly assign the subjective words 'fragile' and 'robust' to the objective

[7] The value of $r_{0.01}$ is essentially equivalent to that of $r_{\min}$ for all practical decision making as to the robustness or fragility of an inference result, but $r_{0.01}$ has much better sampling properties than does $r_{\min}$. That is, computed values of $r_{0.01}$ become independent of the random number generator seed for much smaller values of M_{rep}.

results produced by the algorithm described in this section and with some comments on the practical value of these subjective and objective results.

5 Illustration of the method

5.1 Illustrative example using simulated data

In this section, an example using simulated data illustrates how the sensitivity analysis procedure described above can be applied to evaluate the degree to which particular model inferences (hypothesis tests on a structural parameter) are sensitive to likely defects in the exogeneity assumptions made—i.e., to flaws in the instruments used.

Three hundred observations were generated from the simultaneous equation model,

$$c_i = 0.8y_i + 0.7d_i + \varepsilon_{i,1} \tag{15}$$

$$y_i = 0.7c_i + 0.5x_i + 0.6w_i + 0.2q_i + \varepsilon_{i,2}, \tag{16}$$

where $\varepsilon_{i,1}$ and $\varepsilon_{i,2}$ are both uncorrelated and generated (independently from one another) as niid(0,2) variates. The structural coefficient on y_i is denoted β below; the value of β is 0.80 for the data generated by Eqs. 15 and 16. The explanatory variables d_i and x_i are independently generated as $niid(0,1)$ variates and are hence exogenous. In contrast—so as to make them flawed instruments for y_i in Eq. 15—the variables w_i and q_i are generated from

$$w_i = 0.063\varepsilon_{i,1} + v_i^w \tag{17}$$

$$q_i = 0.063\varepsilon_{i,1} + v_i^q, \tag{18}$$

with v_i^w and v_i^q independently distributed as niid(0,1) variates; the particular linear relationships imposed in Eqs. 17 and 18 induce a correlation of 0.089 between $\varepsilon_{i,1}$ and each of w_i and q_i.

The structural equation for c_i, was then estimated, specifying the use of x_i, q_i, and w_i as instruments for y_i; the exogenous explanatory variable (d_i) is, as usual, automatically also included as an instrument. While w_i and q_i are flawed because of their correlation with $\varepsilon_{i,1}$, these two instruments are not weak: the sample correlations of w_i and q_i with y_i are 0.3705 and 0.2565, respectively.

Per the discussion in Sect. 3, the model is estimated using the 2-step GMM estimator, Eq. 13. The resulting estimated equation for c_i is:

$$\begin{array}{lll} c_i = 0.922y_i + 0.507d_i + \eta_i & \overline{R}^2 = 0.931 \\ \quad\quad (0.033) \quad\quad (0.089) & s^2 = 1.649 \end{array} \tag{19}$$

where the numbers in parentheses are estimated standard errors. Letting $\hat{\beta}^{\text{GMM}}$ denote the estimated coefficient on y_i in Eq. 19, observe that $\hat{\beta}^{\text{GMM}}$ exceeds its true value (0.80) by approximately four estimated standard errors. Thus, because of the flaws in

these two instruments, (w_i and q_i), the null hypothesis $H_o: \beta = 0.80$ is (incorrectly) rejected.

Assuming, as would be commonplace, that this sample of 300 observations is sufficiently large as to allow the use of asymptotic results like Eq. 14, the sensitivity analysis algorithm described in Sect. 4 is applied below to an assessment of the instrument flaw sensitivity of two different inferences regarding the model for these data. These inferences are simple hypothesis tests with regard to the value of the parameter β in Eq. 15, testing the two null hypotheses: $H_o: \beta = 0.80$ (versus $H_a: \beta \neq 0.80$) and $H_o: \beta = 0.90$ (versus $H_a: \beta \neq 0.90$). The null hypothesis $H_o: \beta = 0.80$ happens to be correct, but—of course—one would not know that in practice.[8]

The null hypothesis $H_o: \beta = 0.80$ is (incorrectly) rejected at the 5 % level and one (incorrectly) cannot reject the null hypothesis $H_o: \beta = 0.90$ at the 5 % level. With actual data, these inferential errors could be due to poor luck; here, we know that these inferential failures are for the most part due to the flaws in two of the instruments used in the estimation. The sensitivity analysis results below allow one to assess the robustness or fragility of these two inferences without already knowing that these two instruments are flawed.

The value of m was set to two in these two sensitivity analyses, and potential flaws in only w_i and q_i were considered. Sensitivity analysis results (with m equal to four) also examining these two inferences with regard to possible endogeneity of d_i and with regard possible flaws in x_i as an instrument could have been obtained just as easily, but an important objective for this first pair of examples is to illustrate the graphical depiction of the results described in Step 5.c(i) of Sect. 4, which is limited to the special case where m is two. An empirical example with $m = 3$ is analyzed in Sect. 5.2; Ashley and Parmeter (2013) illustrates the sensitivity analysis with $m = 30$.

M_{rep} randomly selected covariance pairs—[cov(w_i,$\varepsilon_{i,1}$), cov(q_i, $\varepsilon_{i,1}$)]—were generated; these correspond to the vector $\Sigma'_{Z\varepsilon}$ in Sect. 3. M_{rep} was set to 10,000 for this example, as the primary focus in this example is on the graphical display of the sensitivity analysis results.[9] Figure 1 displays all of the resulting correlation pairs [corr(w_i, $\varepsilon_{i,1}$), corr(q_i,$\varepsilon_{i,1}$)] for which the particular null hypothesis $H_o: \beta = 0.80$ can no longer be rejected at the 5 % level using the GMM estimates of Eq. 19 and the sampling distribution of Eq. 14. Figure 2, in contrast, displays all of the analogous correlation pairs for which the null hypothesis $H_o: \beta = 0.90$ *can* be rejected at the 5 % level.

Focusing first on Fig. 1, note that what is being plotted here are all of the generated [corr(w_i, $\varepsilon_{i,1}$), corr(q_i,$\varepsilon_{i,1}$)] pairs which are sufficiently flawed that $H_o: \beta = 0.80$—which *was* rejected at the 5 % level using $\hat{\beta}^{\text{GMM}}$—is no longer rejected. This set of pairs could usefully be called the 'No Longer Rejecting' set of instrumental flaws. Observe that this 'No Longer Rejecting' set does not include the origin, which corresponds to w_i and q_i both being assumed to be uncorrelated with $\varepsilon_{i,1}$. But it does include instrument-

[8] These two examples focus on a simple hypothesis test solely for expositional clarity; the sensitivity analysis can be applied equally well to null hypotheses involving multiple linear (or nonlinear) parameter restrictions; see Ashley and Parmeter (2013) for an application in this direction.

[9] One might want to use a larger value for M_{rep} in order to compute precise values for statistics like $r_{\min}$; one would surely want to use a larger value for M_{rep} for larger values of m.

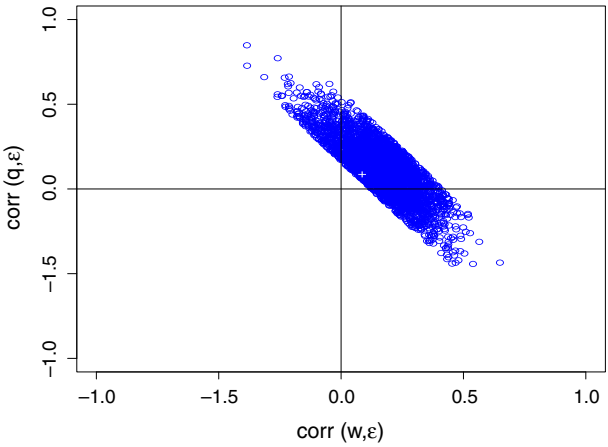

Fig. 1 Instrument-error correlation pairs in the 'No Longer Rejecting' set for testing $H_o : \beta = 0.80$ at the 5 % level in the model of Eqs. 15 and 16. The cross corresponds to the pair (0.089, 0.089), the actual value of $\Sigma_{Z_2\varepsilon}$ in this example

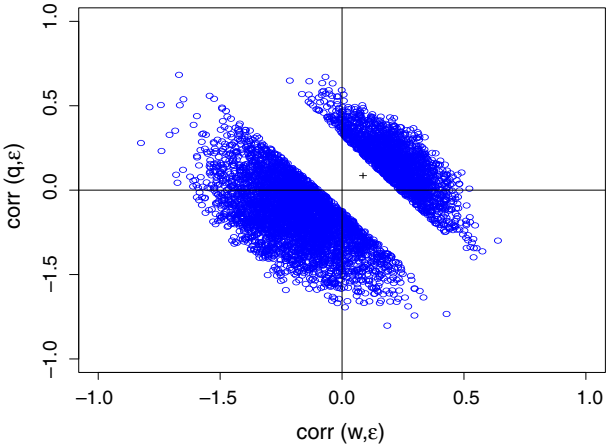

Fig. 2 Instrument-error correlation pairs in the 'No Longer Not Rejecting' set for testing $H_o : \beta = 0.90$ at the 5 % level in the model of Eqs. 15 and 16. The cross corresponds to the pair (0.089, 0.089), the actual value of $\Sigma_{Z_2\varepsilon}$ in this example

error pairs which are fairly close to the origin: the shortest ray from the origin to this 'No Longer Rejecting' set has a length of only 0.11; this is the '$r_{\min}$' defined in Step 5.a of Sect. 4. Thus, it requires a noticeable—but not all that large—set of flaws in these two instruments in order to overturn the sample rejection of $H_o : \beta = 0.80$.

One might therefore characterize this inference as somewhat 'fragile' to likely failures in these two instruments. On the other hand, looking at Fig. 1, if one had theoretical reasons for strongly suspecting that any flaws in the instruments w_i and q_i must be such as to induce a negative correlation with the structural error, $\varepsilon_{i,1}$, then one could instead conclude that this inference is actually fairly robust.

Finally, note that the instrument-error correlation vector corresponding to the minimal-length vector reaching the 'No Longer Rejecting' set is (0.089, 0.061). This result indicates that the inference regarding $H_o : \beta = 0.80$ is more or less equally fragile to flaws in either of the two instruments analyzed, w_i and q_i.

Turning to Fig. 2, what is plotted here are the points in a 'No Longer *Not* Rejecting' set of instrument-error correlations, as the null hypothesis $H_o : \beta = 0.90$ is—in this case incorrectly—*not* rejected in the GMM estimates. Note that the 'empty swath' of points in this plot consists of the instrument-error correlation pairs for which a failure to reject this null hypothesis still occurs. This inference (the failure to reject $H_o : \beta = 0.90$) is evidently somewhat more fragile than was the rejection of $H_o : \beta = 0.80$ analyzed in Fig. 1: here the value of $r_{\min}$ is only 0.086. Thus, for this second null hypothesis, one would conclude that this inference, also, is somewhat fragile.

The vector corresponding to the minimum length ray to the 'No Longer Not Rejecting' set of instrument-error correlation is $(-0.066, -0.054)$ for the null hypothesis $H_o : \beta = 0.90$. Consequently, the sensitivity analysis again indicates that the inference is more or less equally sensitive to flaws in either of the two instruments analyzed. A 'No Longer Not Rejecting' set of instrument-error correlations will typically look like Fig. 2 and have an empty swathe, containing the origin, running through it. Thus, in this instance, a theoretical presumption that any flaws in these two instruments must lead to positive instrument-error correlations would not reduce the apparent fragility of this inference.

Summarizing the results so far, then, the sensitivity analysis indicates that both of these inferences are fairly fragile to potential flaws in these two instruments; this result would alert an analyst to the likely possibility that these results could be artifacts of instrumental flaws. Here, where we artificially know that these data were generated in such a way that these two instruments actually *are* flawed—they are each, by construction, correlated with the model errors in amount 0.089—it is not surprising that both inferences are in fact incorrect.[10]

5.2 Empirical example: assessing the fragility/robustness of the Angrist and Krueger (1991) inference to likely instrument flaws

The sensitivity analysis proposed above (and applied, in Sect. 5.1, to inferences in a model based on artificially generated data) is of course also applicable to actual empirical models. For example, the procedure is applied in this section to the well-known Angrist and Krueger (1991) study testing the impact of schooling on earnings; the original model utilizes 329,509 sample observations on US men born between 1930 and 1939 and uses 30 instruments.

[10] Thus, it is appropriate that the actual correlation pair (0.089, 0.089)—denoted by '+'—lies outside of the 'No Longer Rejecting' set plotted in Fig. 1 and close to the 'No Longer Not Rejecting' set plotted in Fig. 2. Also, note that the length of the actual instrument-error flaw correlation vector $[\text{corr}(w_i, \varepsilon_{i,1}),\ \text{corr}(q_i, \varepsilon_{i,1})]$ is $0.126 = \sqrt{0.089^2 + 0.089^2}$. This length exceeds $r_{\min} = 0.11$—the minimum length for an instrument-error correlation vector to reach the 'No Longer Rejecting' set for no longer rejecting $H_o : \beta = 0.80$.

The basic Angrist–Krueger model is

$$log(Z_i) = X_i\beta_1 + YB_i\delta_1 + \rho E_i + \varepsilon_i, \tag{20}$$

where Z_i is the weekly wage of the ith individual, E_i is the total years of education of individual i, ρ represents the return to education, X_i is a vector of covariates for individual i, and YB_i is a 1×9 dummy variable vector containing a one for the component corresponding to the year in which the individual was born and zero for the other components.

The difficulty in estimating ρ in Eq. 20 is that unobserved taste and ability variables render E_i endogenous. Angrist and Krueger's contribution was to note that the interaction of school-entry requirements and compulsory schooling laws historically forced students born in certain months to stay in school longer, on average, than students born in other months. Angrist and Krueger therefore argued that an individual's birth month is correlated with his number of years of education—but uncorrelated with these unobserved personal attributes—yielding a source of valid instruments for E_i. In particular, Angrist and Krueger (1991) used interactions between three quarter-of-birth dummies and 10 year-of-birth dummies to create the 30 instruments alluded to above. Bound et al. (1995) re-analyzed the Angrist–Krueger model using only the three quarter-of-birth instruments, arguing that the full Angrist–Krueger information set included many weak instruments; we follow Bound et al. (1995) here solely for expositional clarity.[11]

Using this instrumentation strategy, the GMM estimates[12] of Eq. 20 are:

$$\underset{}{ln(Z_i)} = \underset{(0.250)}{4.592} + \underset{(0.020)}{0.105 E_i} + \cdots + \eta_i, \tag{21}$$

with $\overline{R}^2 = 0.091$ and $s^2 = 0.647$.[13] Using these instruments, $H_o : \rho = 0.0$ can be rejected with p value <0.0005. Moreover, the quarter-of-birth instruments collectively are not weak; the first-stage F-statistic for the excluded instruments is 32.269.

The actual validity of this clever Angrist–Krueger instrument choice was subsequently the subject of much debate: e.g., see arguments summarized in Bound and Jaeger (1996).[14] Consequentially, it is of substantial practical interest—even after a notable passage of time—to assess whether this inference result is fragile or robust with respect to possible flaws in these instruments.

[11] See Ashley and Parmeter (2013) for the analogous sensitivity analysis using the full 30-instrument setup of Angrist and Krueger (1991).

[12] The 2SLS estimates and associated standard errors are, to three decimal places, identical.

[13] The 9 year-of-birth coefficient estimates are not quoted in Eq. 21 as they are irrelevant to the present discussion.

[14] More recently, Buckles and Hungerman (2013) find that using season-of-birth instruments can produce inconsistent estimates across a wide array of empirical settings. In particular, later evidence indicates that birth-quarter is in fact correlated with a number of factors which affect wages but which, because they are typically omitted from the regression, make up part of the structural error term.

Letting m stand for the number of instruments examined in the sensitivity analysis, scatterplots such as Figs. 1 and 2 are not feasible to plot for values of m greater than two. Here $m = 3$. However, as noted Step 5.c(ii) of the algorithm—as defined in Sect. 4—the sensitivity of the ρ inference can be informatively displayed by means of a plot of the 'empirical cumulative distribution function,' here abbreviated to 'ECDF.' This is a plot of the proportion of the generated instrument-error correlation m-vectors for which the null hypothesis ($H_o : \rho = 0$) can no longer be rejected at some specified level and for which the length r of the implied instrument-error correlation m-vector is less than a given value.[15] The height of this plot first exceeds zero as r rises above $r_{\min}$, its height reaches 0.01 at r equal to $r_{.01}$, its height reaches 0.05 at r equal to $r_{.05}$, and so forth. Alternatively, one can just tabulate the values of $r_{\min}$, $r_{.01}$, $r_{.05}$, $r_{.10}$, and $r_{.20}$. The sensitivity results for the Bound et al. (1995) model inference on $H_o : \rho = 0.0$ are displayed in both of these ways below, using M_{rep} equal to 50, 000 and considering the inference result to be 'overturned' in any case for which $H_o : \rho = 0.0$ is no longer rejected at the 5 % level.[16]

Figure 3 displays the empirical cumulative distribution function plot for m equal to 3, corresponding to a simultaneous analysis of all three of the quarter-of-birth instruments used in Bound et al. (1995). This plot displays an empirical cumulative distribution function which rises quite sharply for very small values of the instrument-error correlation length. Evidently, the Angrist–Krueger rejection of $H_o : \rho = 0$, with the consequent conclusion that the number of years of education has a significant impact on log-wages, is quite fragile with respect to minor flaws in any of these three quarter-of-birth instruments.

Results for $r_{\min}$, $r_{.01}$, $r_{.05}$, $r_{.10}$, and $r_{.20}$ are given in Table 1 for both the $m = 3$ analysis plotted in Fig. 3 and for each instrument individually, assuming that the remaining two instruments are valid. In particular, results are tabulated which analyze the sensitivity of the inference on $H_o : \rho = 0.0$ with respect to possible flaws in each of the three instruments in Bound et al. (1995) separately and the 'All 3' column tabulates the results simultaneously allowing for possible flaws in all three instruments.

Table 1 follows Angrist and Krueger's original nomenclature for the instruments: The 'Qtr4' instrument, for example, is the dummy variable for men born in the 4th quarter of any year. The table row marked 'min t 1st stage' gives the magnitude of the estimated t-ratio for this instrument in the first-stage regression model of an ordinary 2SLS estimate of Eq. 20; the 'All 3' column of the row gives the minimum magnitude for this t-ratio over the corresponding group of instruments. These estimated t-ratios are substantial: The first-stage F-statistic for these three instruments is substantial in this Bound et al. (1995) reformulation of the model.

So as to put these instrument-error correlation statistics from the sensitivity analysis in perspective, the table row marked 'sup corr' tabulates the supremum of the magnitudes of the sample correlations between each instrument in the group considered

[15] The length 'r' is defined in Step 4 of the algorithm, as defined in Sect. 4; $r_{.01}$, $r_{.05}$, $r_{.10}$, and $r_{.20}$ are defined in Step 5.b.

[16] Setting $M_{\text{rep}} = 50, 000$ is ordinarily sufficient to make $r_{\min}$ stable to two decimal places with respect to re-running the analysis with a different initial seed for the random number generation in Step 1 of the sensitivity analysis algorithm, as outlined in Sect. 4.

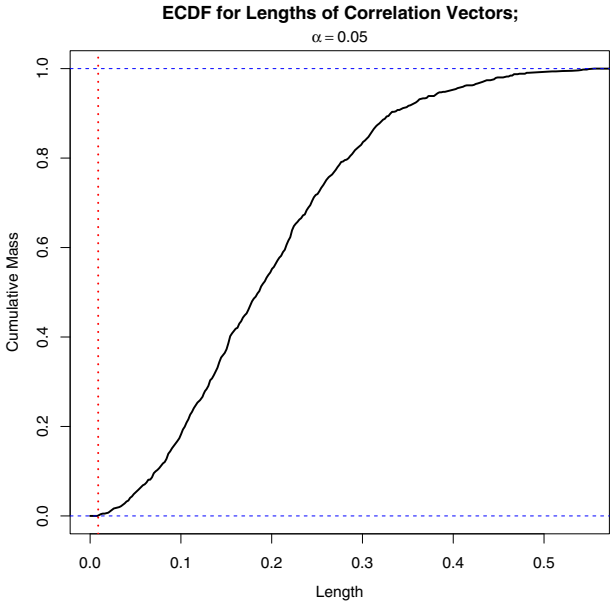

Fig. 3 Empirical cumulative distribution function for the sensitivity analysis of the Bound et al. (1995) inference on $H_o : \rho = 0.0$ with respect to the three quarter-of-birth instruments. The *dotted vertical line marks* the minimum length instrument-error correlation vector for which H_o is no longer rejected at the 5 % level

Table 1 Sensitivity analysis results on $H_o : \rho = 0.0$ in the Bound et al. (1995) model (Eq. 20)

Instrument	Qtr2	Qtr3	Qtr4	All 3
min t 1st stage	3.479	7.065	9.136	3.479
$r_{\min}$	0.013	0.006	0.005	0.009
$r_{0.01}$	0.013	0.007	0.005	0.023
$r_{0.05}$	0.014	0.007	0.005	0.049
$r_{0.10}$	0.015	0.007	0.006	0.073
$r_{0.20}$	0.016	0.008	0.006	0.104
sup corr†	0.013	0.006	0.005	0.009
inf corr	0.013	0.006	0.005	−0.002

† The value of 'sup corr' for each column is the supremum of the sample correlations between all of the instruments analyzed and all of the endogenous explanatory variables in the model; there is only one endogenous explanatory variable (E_i) in the Angrist–Krueger model. The value of 'inf corr' is the corresponding infimum; this is, of course, equal to the supremum for columns in which only one instrument is analyzed

and the endogenous explanatory variable (E_i) in the model; the table row marked 'inf corr' tabulates the analogous infima.

Table 1 bears much the same interpretation as the ECDF plot in Fig. 3, but quantifies the results more precisely. In particular, it indicates that the Bound et al. (1995) inference rejecting $H_o : \rho = 0.0$ is very sensitive to small instrument-error flaws: A

correlation of just 0.01 is sufficient to overturn the rejection of this null hypothesis at the 5 % level. Thus—without making any further assumptions about the original Angrist–Krueger model (as re-estimated using the Bound et al. (1995) instrument set) and without evaluating the criticisms of their instruments made by Bound and Jaeger (1996) and others—our sensitivity analysis results clearly indicate that the Angrist–Krueger inference result with respect to $H_o : \rho = 0.0$ is 'fragile' with respect to possible flaws in their instruments.

These sensitivity analysis results are also potentially informative as to which particular instruments contribute most heavily to whatever fragility is found. The $r_{\min}$ entries in Table 1 do not vary much across the different sets of instruments analyzed. But a better way to address this issue is to examine the components of the minimum length m-vector (ray) of instrument-error correlations to the 'No Longer Rejecting' set.

In the analysis here, we find that this vector is

$$(\text{Qtr2}, \text{Qtr3}, \text{Qtr4}) = (-0.002, -0.001, 0.009). \tag{22}$$

The fact that all three of these correlation components are quite small in magnitude echoes the conclusion reached above: An instrument-error correlation need only exceed around 0.01 in magnitude in order to overturn the original Angrist–Krueger inference. This vector allows us to go a bit further than that, however, and note that the Angrist–Krueger inference is a bit less sensitive to ('fragile with respect to') flaws in 'Qtr2' and 'Qtr4' than to flaws in 'Qtr3', for which the vector component is notably larger, albeit still tiny. Were these three instruments distinctly sourced, this result would suggest that the instrument 'Qtr3' might merit more scrutiny than the other instruments, and one might in that case conclude that the additional risk of parameter estimator inconsistency (due to possible deficiencies in the validity of these particular instruments) is not worth the additional estimation precision their use provides. These three components do not differ all that much in the present case, however. Moreover, in the Angrist–Krueger model setting, all of these instruments are conceptually on an equal footing. Consequently, the most appropriate conclusion to draw from this estimated vector in this case is that the inference with respect to $H_o : \rho = 0.0$ is fragile with respect to all three of these instruments.

6 Concluding remarks

This paper has proposed a feasible sensitivity analysis with regard to the exogeneity (i.e., instrumental variable validity) assumption made when IV regression is used to confront endogeneity in econometric models. Given that it is impossible to directly test for such instrumental flaws, our sensitivity analysis algorithm—in the spirit of Leamer (1983)—improves on existing methods in several ways. In particular, the proposed procedure features the following advantages:

1. It does not require *any* additional assumptions regarding either the underlying model *or* the size of the instrumental flaws. In particular, these flaws are *not*

assumed to be either small or (for some bizarre reason) diminishing with the sample length, as in methods based on 'local-to-zero' asymptotics.
2. It is computationally straightforward and efficient; therefore it is feasible to implement it in actual applied work with large data sets and/or a substantial number of instruments.
3. It is applicable in any setting where one is estimating a parametric model using instrumental variables.

The sensitivity analysis proposed here (to flawed instruments in the endogenous explanatory variable setting) is an important contribution to the 'applied econometrics toolkit', because this predicament is so common as to be an endemic feature of empirical economic analysis. For example, Keane (2010a, p. 6) notes that "…exogeneity assumptions are always a priori, and there is no such thing as an 'ideal' instrument that is 'obviously' exogenous."

Our proposed sensitivity analysis neither eliminates nor even ameliorates such instrumental flaws. What it *does* do is inform the analyst as to which model inferences are clearly 'robust' to likely flaws and which are clearly 'fragile.' Where an inference is fragile, then it should be interpreted with an appropriately large measure of caution. Where it is robust, we nevertheless support Keane (2010a, b) argument that one must still attend to the underlying *economics* of the model in order to properly interpret the coefficient estimates and inferential results relating to them. In particular, if a population coefficient does not mean what one thinks it means—e.g., because of failure to attend to the conditioning assumptions—then the fact that one's hypothesis test rejection p values are robust to likely instrumental flaws is of little consolation.

Still, for those cases where the model coefficients (and null hypotheses couched in terms of them) *are* well defined, it is definitely advantageous to be able to assess whether (and *which* of) the empirical inferences about them are robust to likely levels of instrumental flaws and which are fragile. A large value of the sensitivity analysis statistic $r_{\min}$ defined in the algorithm described above—i.e., values of 0.30, say, and up—pretty clearly corresponds to an inference which is fairly 'robust' to flaws in the instruments; such an inference is therefore relatively credible. In contrast, a small value of $r_{\min}$—such as a value less than around 0.05, say—corresponds to an inference which is fairly 'fragile' with respect to flaws in the instruments; such an inference is therefore not very credible. In less clear-cut cases, one must obviously exercise one's own, to some degree subjective, judgement. But—even in such an instance—the sensitivity analysis has given one something objective to work with and at least made it clear that this *is* an intermediate case.

Also, finding that a particular inference is fragile is not necessarily the end of the story. One might well, in such an instance, strongly consider dropping one or more of the instruments figuring most prominently in the m-vector of instrument-error correlations corresponding to $r_{\min}$: The additional estimation precision afforded by using this instrument is likely not worth the additional fragility it engenders.[17] On the other hand, if an inference is highly fragile in many or most instrument-error

[17] See also Donald and Newey (2001) and Donald et al. (2009).

correlation directions, then this inference is simply not as meaningful as it appears. It is better to know that.

Note also that 'fragility'/'robustness' is specific to each inference which one considers. For a given model—and data set on the explanatory variables, and selection of instruments—the sensitivity analysis can easily indicate that inferences with regard to some null hypotheses are quite fragile and that inferences with respect to other null hypotheses are quite robust. In either case, knowing the robustness of these particular inferences is of empirical importance.

References

Angrist J, Krueger A (1991) Does compulsory school attendance affect schooling and earnings? Q J Econ 106(4):979–1014

Angrist J, Pischke J-S (2010) The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. J Econ Perspect 24(2):3–30

Ashley R (2009) Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. J Appl Econom 24(3):325–337

Ashley R, Parmeter CF (2013) Sensitivity analysis of inference in GMM estimation with possibly-flawed moment conditions. University of Miami, Department of Economics Working Paper 2013-08

Berkowitz D, Caner M, Fang Y (2008a) The validity of instruments revisited. North Carolina State University Working Paper

Berkowitz D, Caner M, Fang Y (2008b) Are nearly exogenous instruments reliable? Econ Lett 101:20–23

Bound J, Jaeger D (1996) "On the validity of season of birth as an instrument in wage equations: a comment on Angrist and Krueger's "Does compulsory school attendance affect schooling and earnings?,"". National Bureau of Economic Research Working Paper no. 5835

Bound J, Jaeger D, Baker R (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. J Am Stat Assoc 90(430):443–450

Buckles K, Hungerman DM (2013) Season of birth and later outcomes: old questions, new answers. Rev Econ Stat 95(3):711–724

Conley T, Hansen C, Rossi P (2007). Plausibly exogenous. Mimeo, Chicago Graduate School of Business

Donald SG, Imbens GW, Newey WK (2009) Choosing instrumental variables in conditional moment restriction models. J Econom 152(1):28–36

Donald SG, Newey WK (2001) Choosing the number of instruments. Econometrica 69(5):1161–1191

Ebbes P, Wedel M, Böckenholt U (2009) Frugal IV alternatives to identify the parameter for an endogenous regressor. J Appl Econom 24(3):446–468

Hahn J, Hausman J (2006) IV estimation with valid and invalid instruments. http://econ-www.mit.edu/faculty/?prof_id=hausman

Hall AR, Inoue A (2003) The large sample behaviour of the generalized methods of moments estimator in misspecified models. J Econom 114:361–394

Keane M (2010a) Structural vs. atheoretic approaches to econometrics. J Econom 156:3–20

Keane M (2010b) A structural perspective on the experimentalist school. J Econ Perspect 24(2):47–58

Kraay A (2008) Instrumental variables regressions with honestly uncertain exclusion restrictions. The World Bank, Policy Research Working Paper 4632

Leamer E (1983) Let's take the con out of econometrics. Am Econ Rev 73(1):31–43

Murray MP (2006) Avoiding invalid instruments and coping with weak instruments. J Econ Perspect 20(4):111–132

Sims CA (2010) But economics is not an experimental science. J Econ Perspect 24(2):59–68

Small DS (2007) Sensitivity analysis for instrumental variables regression with overidentifying restrictions. J Am Stat Assoc 102(479):1049–1058

Stock JH (2010) The other transformation in econometric practice: robust tools for inference. J Econ Perspect 24(2):83–94

Wooldridge J (2010) Econometric analysis of cross section and panel data. MIT Press, Cambridge