Advance Access publication December 11, 2012

Political Analysis (2013) 21:86–96
doi:10.1093/pan/mps037

Robust Estimation with Nonrandom Measurement Error and Weak Instruments

Timm Betz
Department of Political Science, University of Michigan, Ann Arbor
e-mail: timmbetz@umich.edu

Edited by Jonathan Katz

Two common problems in applications of two-stage least squares (2SLS) are nonrandom measurement error in the endogenous variable and weak instruments. In the presence of nonrandom measurement error, 2SLS yields inconsistent estimates. In the presence of weak instruments, confidence intervals and *p*-values can be severely misleading. This article introduces a rank-based estimator, grounded in randomization inference, which addresses both problems within a unified framework. Monte Carlo studies illustrate the deficiencies of 2SLS and the virtues of the rank-based estimator in terms of bias and efficiency. A replication of a study of the effect of economic shocks on democratic transitions demonstrates the practical implications of accounting for nonrandom measurement error and weak instruments.

1 Introduction

In situations where ordinary least squares performs poorly, instrumental variable techniques are a popular alternative, with two-stage least squares (2SLS) being the most commonly used estimator (Sovey and Green 2011). This article introduces a rank-based instrumental variables estimator, grounded in randomization inference, which provides a unified framework to address two problems for the use of 2SLS. The first is nonrandom measurement error, such that the instrument is correlated with the measurement error in an explanatory variable. In this case, coefficient estimates obtained from 2SLS are inconsistent. The second is the presence of weak instruments, in which case 2SLS yields incorrect measures of statistical uncertainty.

The estimator presented in this article is robust with respect to both problems. The estimator was first developed by Rosenbaum (1996, 2002); Imbens and Rosenbaum (2005) demonstrate that the estimator is robust to weak instruments. This article expands on Imbens and Rosenbaum by showing that variants of the estimator can also accommodate nonrandom measurement error. At the same time, this article provides the instrumental variables estimator for the techniques described in Keele, McConnaughy, and White (2012), who advocate the use of randomization inference in political science.[1]

The article proceeds in three parts. The first part briefly describes the problems for 2SLS arising from nonrandom measurement error and weak instruments. The second part introduces randomization inference and a rank-based instrumental variables estimator. Monte Carlo studies illustrate the deficiencies of 2SLS in the presence of nonrandom measurement error and show the superiority of the rank-based estimator. The third part presents a replication of Brückner and Ciccone (2011), who evaluate the effects of economic shocks on democratization. The comparison between the rank-based estimates and 2SLS suggests that the latter overestimates the size of the causal

Author's note: I thank Vincent Arel-Bundock, Amy Pond, Adam Sales, and seminar participants at the University of Michigan for their comments and many helpful discussions, and the editor, Jonathan Katz, and two reviewers for comments that greatly improved the manuscript. I am particularly indebted to Walter R. Mebane, Jr., for his comments, guidance, and advice. Thanks also to the Department of Political Science at the University of Michigan, which provided resources for conducting the simulation studies. Replication materials for the simulations and data analyses are available online at the *Political Analysis* dataverse: http://hdl.handle.net/1902.1/18952.

[1]For another application of randomization inference in political science, see Hansen and Bowers (2009).

© The Author 2012. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

relationship as well as its statistical significance. As the conclusion emphasizes, the results in this article underscore the necessity of considering the possibility of both nonrandom measurement error and weak instruments in applications of 2SLS.

2 Nonrandom Measurement Error and Weak Instruments

Measurement error is a common problem in political science. Estimators robust to various forms of measurement error have been suggested, among others, for linear models (Western 1995), probit models (Hug 2010), the Cox proportional hazards model (Desmarais and Harden 2012), and overdispersed binomial (Wand et al. 2001) as well as multinomial models (Mebane and Sekhon 2004).

None of these estimators accommodates endogenous variables, which are ubiquitous in political science and often addressed by instrumental variables estimators, 2SLS especially (Sovey and Green 2011). Measurement error is not a problem for 2SLS if it exists in the instrumented variable exclusively and is uncorrelated with the instrument. Indeed, classical measurement error is a common justification for the use of 2SLS (Angrist and Pischke 2009; Sovey and Green 2011). Yet, if measurement error in the explanatory variable is correlated with the instrument, 2SLS yields inconsistent coefficient estimates.[2] The possibility of such nonrandom measurement error is not far-fetched, but often neglected in applications. In particular, measurement error that is a function of the endogenous variable is likely to be correlated with the instrument as well, since the instrument must be correlated with the endogenous variable to be a valid instrument. For instance, poor countries tend to receive more foreign remittances than rich countries, yet are also more prone to misreport remittance flows (Ahmed 2012). An instrument for remittance flows, therefore, is likely to be correlated with the measurement error. 2SLS does not yield consistent estimates in the presence of such systematic measurement error.[3]

Another example can be found in studies of the effect of foreign direct investment (FDI) on various economic and political variables. Kerner and Lawrence (forthcoming) show that data on FDI based on balance of payments data are plagued by a myriad of problems and mismeasured in systematic ways. One of these problems is that it is "entirely possible that failure to conduct the required surveys or report data that conforms with the IMF's guidelines is correlated with some of the same political and institutional factors that are thought to affect political risk, suggesting [. . .] a source of both random and systematic error" (Kerner and Lawrence forthcoming, 8) in FDI data. Consequently, studies using domestic political institutions as instruments for FDI flows are likely to rely on instruments that are correlated with the measurement error in FDI flows, resulting in inconsistent 2SLS estimates.

This problem arising from nonrandom measurement error is confounded by the sensitivity of the mean-based 2SLS to even a few mismeasured observations. In the extreme case, incorrect data on a single observation can ruin the estimate, since the asymptotic breakdown point of the mean is zero. Estimators based on ranks, such as the one introduced in this article, are much less subject to this problem.[4]

Another set of problems is introduced by the presence of weak instruments. Weak correlation between the instrument and the endogenous variable reinforces the small sample bias of 2SLS (Angrist and Pischke 2009). Moreover, with weak instruments, the asymptotic theory to derive

[2]Formally, this point is most easily shown by writing the 2SLS coefficient estimate as a function of sample covariances and comparing the expression using the true data with the expression using the mismeasured data. The two expressions are equivalent, such that 2SLS is consistent, only if the measurement error in the endogenous variable is uncorrelated with the instrument.

[3]I use the terms systematic and nonrandom measurement error synonymously throughout the article to describe measurement error that is correlated with the instrument, possibly because the measurement error is a function of the instrumented variable. This definition does not necessarily imply that the instrumented variable is measured with bias.

[4]An alternative rank-based instrumental variables estimator is developed by Honoré and Hu (2004). These authors propose a "robust" estimator, but they are primarily concerned with robustness to nonnormal distributions of the dependent variable and not with robustness to systematic measurement error. As shown by Imbens and Rosenbaum (2005), the randomization-based estimator introduced in this article is robust to nonnormally distributed data as well. In fact, and in contrast to the estimator developed in Honoré and Hu, it is distribution free, and Imbens and Rosenbaum provide a proof that all distribution-free instrumental variable tests are permutation tests, that is, are based on randomization inference.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

the distribution of the 2SLS estimate breaks down, as pointed out by Bound, Jaeger, and Baker (1995) and Staiger and Stock (1997). Then, 2SLS yields incorrect measures of statistical uncertainty, which may result in unwarranted rejections of null hypotheses (Stock, Wright, and Yogo 2002). Weak instruments are pervasive (Staiger and Stock 1997, 557), and especially the most credible instruments are often generated by some form of natural experiment and only weakly correlated with the endogenous variable. The rank-based estimator presented in the following yields accurate confidence intervals when instruments are weak.

3 Randomization Inference

Randomization inference is based on random assignment of treatments.[5] Denote this assignment by z_i and the treatment by x_i. The treatment elicits some outcome, y_i, and the goal is to estimate the effect of the treatment on the outcome captured by a coefficient β. Absent treatment, such that $x_i = 0$, the response is y_c, some constant. Assuming constant, linearly additive treatment effects, the relationship between treatment and outcome is described by[6]

$$y_i = y_{ci} + \beta x_i. \tag{1}$$

This setup follows the potential outcomes framework as introduced in Rubin (1974). Holland (1986) provides an excellent discussion, and Angrist, Imbens, and Rubin (1996) emphasize the links between this framework and structural equation models.

3.1 *Randomized Treatment*

In the simplest case, the treatment itself is randomized. A first question to ask is whether there is any effect of the treatment on the response. If there is an effect, then the treatment and the response should exhibit some statistical relationship; in the absence of an effect, treatment and response should be independent. One way to evaluate the null hypothesis of no effect, H_0:$\beta_0 = 0$, is to describe the data by some test statistic, $t(x_i,y_i)$, such as the difference in means or a correlation coefficient. The value of the test statistic computed from the data can then be compared to the distribution of the test statistic under the null hypothesis of no effect. Since for many test statistics, the (exact or approximate) distribution under the null hypothesis of no effect is known, a wide range of test statistics is amenable to this method.

Hodges and Lehmann (1963) provide a method to obtain a point estimate for β. To arrive at the Hodges–Lehmann (HL) estimator, rewrite equation (1) as

$$y_{ci} = y_i - \beta x_i. \tag{2}$$

The difference on the right-hand side is the *adjusted response* (Rosenbaum 2002): It is the response net of the treatment effect. Since the left-hand side is a constant, for the true value of β, the adjusted response does not vary with the treatment x_i. To obtain a coefficient estimate, instead of testing H_0: $\beta_0 = 0$, one tests H_0: $\beta_0 = \beta$. If the null hypothesis is true, then $\beta_0 = \beta$, and the adjusted response would not exhibit any relationship with x_i. A test statistic $t(x_i,y_i - \beta x_i)$ can be used again. As before, suppose the expected value of this test statistic under the null hypothesis is known. Then, a coefficient estimate for β can be obtained by solving the estimating equation

$$t(x_i,y_i - \beta x_i) = \mathrm{E}[T]. \tag{3}$$

Using the variance of the test statistic and its exact or asymptotic distribution allows constructing a confidence interval for the point estimate. Rosenbaum (2002, 48–49) provides further discussion. Keele, McConnaughy, and White (2012) offer a more thorough introduction to randomization inference with applications from political science.

[5] This section draws on the discussion in Rosenbaum (2002).

[6] This model of effects is the most commonly used model in political science applications of 2SLS. Depending on the specific application, one might wish to consider other models of effects. Hansen and Bowers (2009), for instance, offer an application of randomization inference with additive but individually varying effects.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

3.2 *Randomized Instrument*

The previous discussion assumed that x_i is assigned randomly. When this assumption is violated, x_i is labeled endogenous. As in the 2SLS framework, suppose an instrument z_i exists, which is assigned randomly, which does not elicit a response directly (has no independent effect on y_i, i.e., satisfies the exclusion restriction), and which has some nonzero effect on x_i on average (z_i may be a weak instrument, but not an irrelevant one). As before, define the adjusted responses as $y_i - \beta x_i$. Then, the hypothesis H_0: $\beta_0 = \beta$ can be tested in a similar way as above, only now the adjusted responses are not compared to the treatment x_i, but to the instrument z_i: While the treatment is used to calculate the adjusted response, the computation of the test statistic is based on the instrument. The HL estimator thus finds the value of β that equates $t(z_i, y_i - \beta x_i)$ to its expected value under the null hypothesis of no effect.

When z_i is a binary instrument and $t(\cdot,\cdot)$ is the difference in means, the HL estimate is algebraically identical to 2SLS, which provides a convenient link between 2SLS and the HL estimator. The HL estimator is much more flexible than this, however, and allows for a wide range of test statistics other than the difference in means. To obtain an estimator more robust to measurement error than the mean-based 2SLS, test statistics based on ranks are particularly useful. While means can fall prey to even a single wrong data point, medians are resilient to wrong data on up to one half of the observations, but their robustness comes at the cost of ignoring a large amount of information. Rank statistics provide a middle ground for most practical applications, as they are situated between the mean and the median in terms of sensitivity to measurement error, but are still relatively efficient. This does not imply that rank statistics return unbiased estimates; rather, the bias is smaller than that of mean-based estimators (Coakley and Hettmansperger 1994; Huber and Ronchetti 2009; Croux and Dehon 2010).[7]

Before turning to a general treatment, consider a binary instrument and the Wilcoxon rank sum for an illustration of how to construct a confidence interval. Let m be the number of treated observations (those with $z_i = 1$), let N be the total number of observations, and let $\text{rnk}(y_i - \beta x_i)$ be a function that ranks the adjusted responses from 1 to N. The Wilcoxon rank sum then is calculated as

$$W = t(z_i, y_i - \beta x_i) = \sum_{i=1}^{N} \text{rnk}(y_i - \beta x_i) z_i, \tag{4}$$

and its expectation, variance, and asymptotic distribution under the null hypothesis of no relationship between z_i and $\text{rnk}(y_i - \beta x_i)$ are (see, e.g., Lehmann 1975, 14):

$$\text{E}[W] = \mu_W = \frac{m(N+1)}{2}, \tag{5}$$

$$\text{Var}[W] = \sigma_W^2 = \frac{m(N-m)(N+1)}{12}, \tag{6}$$

$$\frac{W - \mu_W}{\sigma_W} \xrightarrow{d} \mathcal{N}(0,1). \tag{7}$$

To obtain a 95% confidence interval, the Wilcoxon rank sum W and its standardized value $(W - \mu_W)/\sigma_W$ are computed from the data. For a 95% confidence interval, the critical value for a standard normal distribution is 1.96. Thus, all values of β associated with test statistics

[7] Depending on the data at hand, different rank statistics are preferable. With a binary instrument, the Wilcoxon rank sum is readily calculated. Rank sums can be applied to stratified data by using aligned ranks, subtracting the stratum-specific mean from the adjusted responses before calculating the Wilcoxon rank sum from the aligned data. Notice that for the aligned ranks, the means are calculated for the adjusted responses, $y_i - \beta x_i$, not for y_i or x_i alone. Notice also the similarity to fixed effects regressions, where the data are demeaned by the inclusion of unit-specific dummy variables. Continuous instruments can be accommodated by Spearman's rank correlation coefficient. For all of these statistics, the first and the second moments under the null hypothesis of no effect are readily obtained from standard statistical texts, such as Conover (1999) or Lehmann (1975), which allows for the calculation of the HL estimate of β.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

satisfying $|(W - \mu_W)/\sigma_W| < 1.96$ are included in the confidence set, and a confidence interval can be defined as the shortest interval covering the confidence set.

It is straightforward to apply the same steps to other scenarios and test statistics. Most rank test statistics are asymptotically normally distributed (Hájek and Sidak 1967, chap. 5), and for several rank statistics the exact distribution is known for small sample sizes (see, e.g., Conover 1999) and can be simulated for larger samples, which allows for the calculation of exact confidence sets. Most generally, Imbens and Rosenbaum (2005) show that the HL estimate of β and confidence intervals robust to weak instruments can be obtained for any test statistic

$$T = q(y - \beta x)'\rho(z) = \sum_{i=1}^{N} q_i \rho_i, \tag{8}$$

where $q(\cdot)$ scores the adjusted responses and $\rho(\cdot)$ scores the instrument settings.[8] Under the null hypothesis, the instrument is independent of the adjusted response, which facilitates the computation of the expectation and variance of the test statistic. These are

$$\mathrm{E}[T] = \mu = \bar{q}\bar{\rho}, \tag{9}$$

$$\mathrm{Var}[T] = \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(q_i - \bar{q})^2 \sum_{i=1}^{N}(\rho_i - \bar{\rho})^2, \tag{10}$$

where $\bar{q} = N^{-1}\sum_i q_i$ and $\bar{\rho} = N^{-1}\sum \rho_i$. Stratified data can be accommodated by calculating the stratum-specific expressions and then summing over strata. Using the above first and second moments, Imbens and Rosenbaum (2005, 117) show that the standardized test statistic is asymptotically normally distributed, such that

$$\frac{T - \mu}{\sigma} \xrightarrow{d} \mathcal{N}(0,1). \tag{11}$$

Then, the 95% confidence set and confidence interval are defined as

$$C_{95\%} = \left\{\beta : \left|\frac{T - \mu}{\sigma}\right| < 1.96\right\}, \tag{12}$$

$$CI_{95\%} = [\inf\{C_{95\%}\}, \sup\{C_{95\%}\}]. \tag{13}$$

A statistic comparable to the standard error of a least squares coefficient estimate can be obtained by dividing the length of the confidence interval by $2 \cdot 1.96$.

This approach inverts the procedure of 2SLS, where a standard error is calculated for the coefficient estimate and then confidence intervals are constructed, relying on assumptions about the asymptotic distribution of the coefficient estimate. Weak instruments may invalidate these assumptions, resulting in confidence intervals with wrong coverage. In contrast, confidence intervals for HL point estimates are derived from the distribution of the test statistic. Since the confidence interval for the point estimate is calculated from the distribution of the test statistic, rather than the distribution of the parameter estimate, the rank-based randomization inference confidence intervals maintain correct coverage regardless of the strength of the instrument. Imbens and Rosenbaum (2005) show that the confidence intervals derived from the rank-based test statistic have correct coverage regardless of the distribution of the data, and provide a proof that all distribution-free instrumental variable tests are based on randomization inference. When nonrandom measurement error is present in the data, the confidence intervals may no longer maintain exact coverage, but, due to the more accurate point estimates, in general have better coverage than the confidence intervals obtained from 2SLS.

[8] In the example of the Wilcoxon rank sum, $q_i \in \{1, \ldots, N\}$ ranks the adjusted responses and $\rho_i = z_i$ returns the binary instrument. If both $q(\cdot)$ and $\rho(\cdot)$ return ranks, equation (8) returns a test statistic that is, for testing purposes, equivalent to Spearman's rank correlation coefficient (Hájek and Sidak 1967, 113); adjustments for tied ranks are easily incorporated (Lehmann 1975, 300).

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

3.3 *Monte Carlo Simulations*

To compare the performance of the rank-based HL estimator and 2SLS in the presence of nonrandom measurement error, this subsection presents a series of Monte Carlo simulations.[9] The true model is

$$\begin{aligned} y_i &= -0.5 + 3x_i + u_i, \\ x_i &= 1 + 2z_i + v_i, \end{aligned} \tag{14}$$

where $\beta = 3$ is the coefficient of interest. u_i and v_i are bivariate normal random variables with mean zero and covariance 0.7. This correlation between the two error terms induces endogeneity of x_i and is a typical justification for invoking 2SLS. To evaluate the performance of the estimators in various settings, z_i is either binary or a continuous random variable drawn from a standard normal distribution.

Table 1 shows average coefficient estimates and standard errors from estimating the system in equation (14), when there is no measurement error. The simulations are based on data sets with 250 and 10,000 observations, respectively. The difference between the average coefficient estimate and the true coefficient $\beta = 3$ indicates the amount of bias. The range of coefficient estimates is also reported. Ideally, not only will the distribution of coefficient estimates be unbiased, but it should also be closely centered around the true coefficient β, such that each single coefficient estimate is a reliable estimate of the true coefficient.

The upper panel of Table 1 shows the results for a binary instrument. In the lower panel, the instrument is continuous. In all the four scenarios, the average 2SLS and the HL coefficient estimates are identical to the fourth digit, and their standard errors and the range of the distribution of coefficient estimates are very similar. Although the HL estimator is slightly less precise in data sets of 250 observations (as evidenced by the larger standard error), the differences to 2SLS are marginal. Absent systematic measurement error, therefore, the HL estimator appears to be on par with 2SLS.

To evaluate the consequences of systematic measurement error for the 2SLS and HL estimators, Fig. 1 compares 2SLS and HL point estimates as a function of the share of the data being affected by measurement error, again for a binary and a continuous instrument, respectively, in data sets with 250 observations. The solid lines show the bias. The dashed lines show the deviation of the largest and the smallest point estimates from the true parameter value. To introduce nonrandom measurement error, a variable w_i is created as

$$w_i = m_i - 2z_i, \tag{15}$$

where m_i is drawn from a standard normal distribution.[10] After x_i and y_i are calculated as above, a subset of the data x_i is changed to

$$x_i' = x_i + 2w_i. \tag{16}$$

On average x_i' is unbiased, since by construction $E[w_i] = 0$. However, the measurement error in x_i' is, also by construction, correlated with the instrument z_i.[11]

Figure 1 allows for up to 30% of the data to be affected by measurement error.[12] For 2SLS, the results are particularly worrisome with a binary instrument, as shown in the upper panel. When less than a quarter of the observations are affected by measurement error, the bias of 2SLS is positive and substantial—as shown in Fig. 1, the bias reaches >600% of the true parameter value.

[9] Replication files are available at the *Political Analysis* dataverse at http://hdl.handle.net/1902.1/18952. See Imbens and Rosenbaum (2005) for simulations that demonstrate the superior performance of the HL estimator when instruments are weak and under departures from normality.

[10] If z_i is a binary variable, w_i is constructed from the latent variable that was used to create z_i.

[11] While this model of measurement error is of a fairly general form, the ratio of noise to real data is rather large. Modifications to other data structures and models of measurement error are straightforward with the replication materials.

[12] When >30% of the data are affected by measurement error as in equations (15) and (16), the 2SLS estimates are biased by >1000%. The performance of the HL estimator decreases as well, albeit not nearly as sharply.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

Table 1 Monte Carlo simulations: Clean data

	Coefficient	*Range*	*SE*
Binary instrument			
250 observations			
2SLS	2.996	2.749–3.198	0.064
HL	2.996	2.744–3.214	0.067
10,000 observations			
2SLS	3.000	2.956–3.036	0.010
HL	3.000	2.960–3.036	0.010
Continuous Instrument			
250 observations			
2SLS	2.999	2.870–3.098	0.032
HL	2.999	2.865–3.108	0.034
10,000 observations			
2SLS	3.000	2.976–3.019	0.005
HL	3.000	2.977–3.019	0.005

Note. 5000 iterations of estimating equation (14).

When more than a quarter of the observations are mismeasured, the bias of 2SLS oscillates and is as large as 130 times the true parameter value.[13] The HL estimator, in contrast, performs well. It displays a narrow spread in the distribution of coefficient estimates and relatively little bias (the bias is never larger than 0.6). While not shown in the graph, the spread in the coefficient estimates increases substantially for HL only once more than a third of the data are affected by measurement error; and even then, the HL estimator is always less biased than 2SLS. A similar picture emerges from the lower panel of Fig. 1, where the instrument is continuous. While the HL estimator is never biased by more than 0.4, the bias of 2SLS is as large as 5. The distribution of the 2SLS estimates is characterized by an immense spread with some of the coefficient estimates being as large as 10 times the true parameter value. The HL estimates, in contrast, are tightly centered on the average coefficient estimate.

The HL estimator outperforms 2SLS also in terms of the coverage of the 95% confidence intervals. With a binary instrument and data sets of 250 observations, even when as little as 5% of the observations are affected by measurement error, only 35.9% of the confidence intervals obtained from 2SLS include the true parameter value. In contrast, 86.1% of the HL confidence intervals include the true parameter value. With a continuous instrument, only 15.1% of the 2SLS confidence intervals cover the true parameter value, whereas 84.9% of the HL confidence intervals include the true parameter value. Thus, even though with nonrandom measurement error the confidence intervals of the HL estimator no longer have correct coverage, the coverage is far better than that of the 2SLS confidence intervals.

Judging from the simulations in this section, the rank-based HL estimator clearly outperforms the 2SLS estimator in the presence of systematic measurement error, both in terms of bias and reliability. Moreover, in the absence of measurement error, the differences in terms of efficiency are marginal, whereas there is virtually no difference in terms of bias.

4 Rainfall and Democratization: An Application

This section replicates some of the results in Brückner and Ciccone (2011) (henceforth BC), who examine the effect of income shocks on democratic transitions.[14] The effect of income shocks is not

[13]For better readability, in Fig. 1, the maximal bias was capped at 25 and –5.

[14]BC also provide reduced-form regressions, linking rainfall to democratic transitions, which are not discussed in the following. In a separate paper, I provide a more extended discussion as well as a replication of Miguel, Satyanath, and Sergenti (2004), who use rainfall as an instrument to evaluate the effect of economic shocks on the onset of civil conflict. Their main findings from 2SLS are corroborated by the rank-based HL estimates. See, however, Ciccone (2011), who points out several problems with the specification of Miguel, Satyanath, and Sergenti (2004).

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

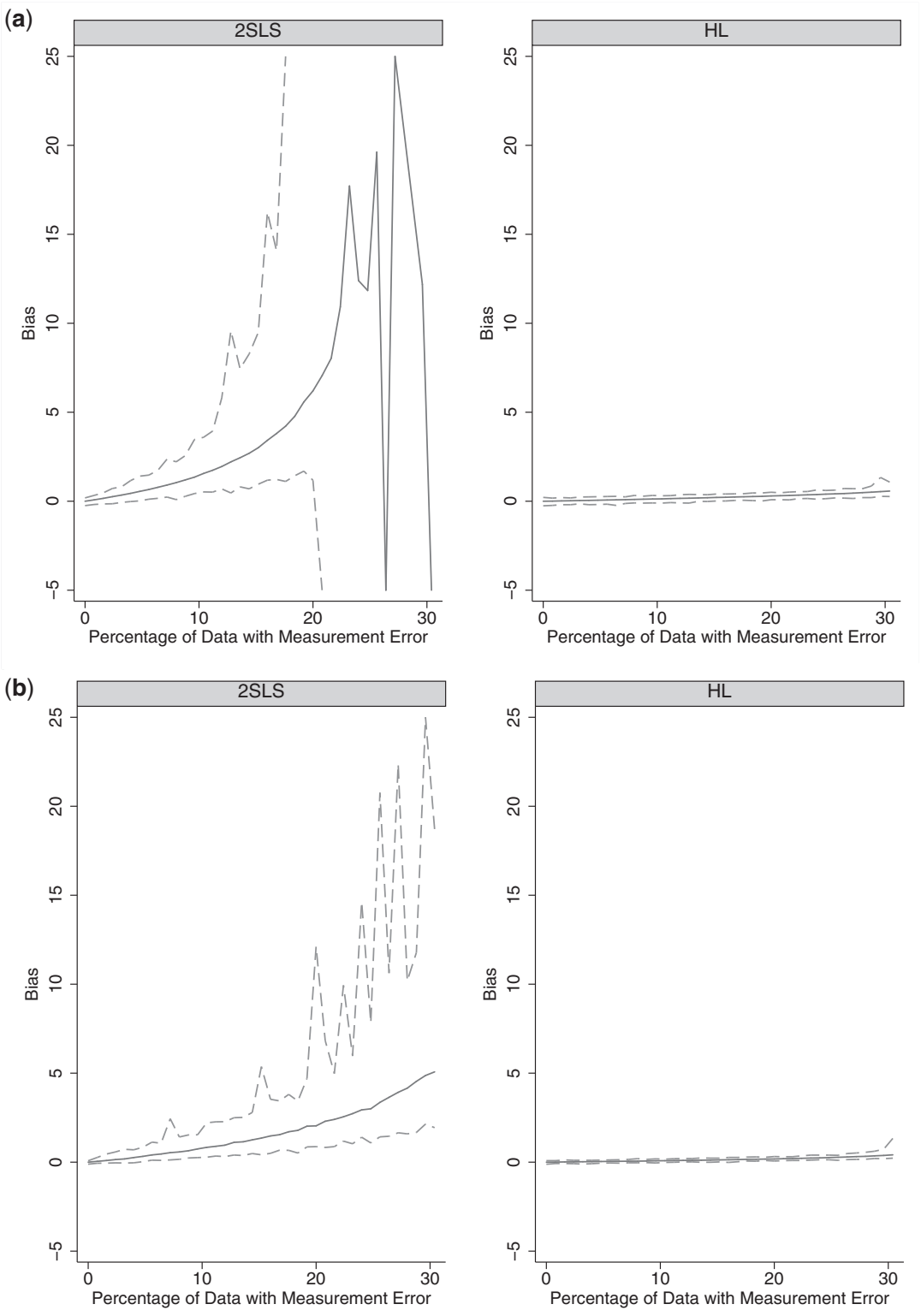


Fig. 1 Binary instrument (**a**), continuous instrument (**b**). Measurement error as in equation (16). Bias (solid line) and largest deviations from the true parameter value (dashed lines) for 2SLS and HL as a function of the percentage of data with measurement error. $\beta = 3$ is the true parameter value.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

easily identified, in part, because democratization itself, and expectations thereof, may drive economic performance. To address this problem, BC rely on rainfall shocks, which are plausibly random, as instrument for income shocks in 2SLS regressions. The link between rainfall and income is most plausible for sub-Saharan African countries to which their sample is restricted.[15]

BC rely on four different measures of democratization from Polity IV (Marshall and Jaggers 2006): changes in polity scores (*Polity2*), in executive constraints (*Exconst*), in competitiveness of political participation (*Polcomp*), and in openness and competitiveness of executive recruitment (*Exrec*). Lagged per capita income, the endogenous explanatory variable, is obtained from the Penn World Tables. The instrument is rainfall as measured by the National Aeronautics and Space Administration's Global Precipitation Climatology Project. All models include dummies to control for year- and country-specific effects as well as country-specific time trends. Thus, the variables measuring (log) GDP per capita and rainfall are capturing deviations of GDP per capita and rainfall, respectively, from country-specific and common time trends, and are interpreted as income and rainfall shocks.

Table 2 replicates Table V in BC, where the explanatory variable is log GDP per capita in the previous time period, net of common time trends, country-specific time trends, and country-specific effects. 2SLS and rank-based HL coefficient estimates are reported, together with standard errors and 95% confidence intervals. Since BC are concerned about the strength of their instrument (with first-stage *F*-statistics between 6 and 7), they report *p*-values based on an Anderson–Rubin Wald test, which in the presence of weak instruments are more accurate than *p*-values based on an asymptotic normal distribution.

Two results in Table 2 are noteworthy. First, the rank-based point estimates are of the same sign as the 2SLS estimates, but substantially smaller. The reduction in the size of the coefficient estimates is tremendous and essentially wipes out the practical significance of the results obtained by BC. This finding suggests that nonrandom measurement error may have biased the 2SLS estimates upward (in absolute terms), as was the case in the Monte Carlo simulations conducted in the previous section.[16]

BC briefly consider the possibility of systematic measurement error, but proceed under the assumption that any potential measurement error is classical and hence unproblematic for the 2SLS estimates. Yet, as Deaton (2005, 4) notes, nonrandom measurement error is pervasive in income data from sub-Saharan countries, the sample used by BC. Indeed, in their study of the effect of economic shocks on civil conflict, Miguel, Satyanath, and Sergenti (2004, fn 22) use rainfall as an instrument for economic shocks and attribute some theoretically unexpected findings to the possibility of "extensive nonclassical measurement error in national accounts data."

More specifically, the argument linking rainfall and economic conditions implies that large values in the rainfall data are associated with large values in the income data. Such outlier conditions are notoriously hard to measure. Similarly, if countries with smaller rainfall shocks tend to have higher levels of income, and wealthier countries tend to report their economic data more accurately, then nonrandom measurement error is a concern. The differences between the rank-based HL estimates and the 2SLS estimates indicate that such nonrandom measurement error is plausibly present in the data.

Second, from the Anderson–Rubin test, BC find a statistically significant effect of economic shocks on institutional change. For instance, they obtain a *p*-value of 0.049 when the dependent variable is the change in polity scores. Given that the confidence intervals based on the normal

[15] BC also evaluate and rule out several alternative causal mechanisms through which rainfall could affect democratic transitions, other than through economic conditions. See their note 16, in particular.

[16] An alternative explanation for the differences between the 2SLS and the HL point estimates would be the presence of unusually high outcomes on the dependent variable. If the "unusual" outcomes are considered aberrant, then the rank-based estimate is still preferable. However, Keele, McConnaughy, and White (2012, 489) point out that, if one has reason to believe that the "unusual" outcomes are accurately reflecting the relationship in the data (and hence are due to "unusual" values of the treatment), then a mean-based estimate might be preferable to a rank-based estimate, since the "unusual" observations may contain particularly valuable information. In the sample of BC, the difference between the rank-based HL estimates and the mean-based 2SLS cannot be explained by dropping the most extreme observations.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

Table 2 Replication of Table V in BC

	Coefficient	*SE*	*95% CI*	*Anderson–Rubin*
Dependent variable: change in polity scores				
2SLS	−18.02	11.35	−40.26 to 4.218	0.049
HL	−2.427	7.964	−29.43 to 1.790	–
Dependent variable: change in executive constraints				
2SLS	−5.809	3.940	−13.53 to 1.912	0.073
HL	−1.298	2.746	−10.42 to 0.343	–
Dependent variable: change in political competition				
2SLS	−7.680	4.673	−16.84 to 1.479	0.037
HL	−2.233	4.606	−17.41 to 0.648	–
Dependent variable: change in executive recruitment				
2SLS	−6.137	4.107	−14.185 to 1.912	0.054
HL	−1.087	3.262	−12.129 to 0.659	–

Note. Log GDP, lagged, with instrument log rainfall, lagged. Country- and year-fixed effects and country-specific time trends included. Anderson–Rubin is the *p*-value from the Anderson–Rubin test (χ^2). For 2SLS, confidence interval based on normal approximation.

approximation include zero, this finding underscores the importance of accounting for weak instruments in 2SLS—the more accurate Anderson–Rubin Wald test yields a statistically significant effect, where the normal-based approximation does not. However, the HL estimates suggest that even these results should be viewed with caution. The confidence intervals for all four HL estimates are asymmetric, wide, and include zero, such that none of the coefficients is different from zero at the 5% significance level.[17]

5 Conclusions

This article discussed two problems for applications of 2SLS. The first is the presence of nonrandom measurement error in an explanatory variable and its implications for point estimates. The second is the presence of weak instruments and the implications for measures of statistical uncertainty. Both of these problems can have substantively important consequences for inferences drawn from econometric results. Building on the work of Rosenbaum (2002) and Imbens and Rosenbaum (2005), this article introduced a rank-based HL estimator that addresses the problems arising from systematic measurement error and weak instruments within a unified framework. Monte Carlo simulations demonstrated the superior performance of the rank-based estimator, and the replication in the previous section underscored the importance of addressing these problems in applied research. The results in this article suggest that rank-based estimates should be used at the very least as a robustness check on 2SLS whenever systematic measurement error or weak instruments are of concern. Researchers should carefully consider the quality of their data when the two estimates differ.

References

Ahmed, Faisal Z. 2012. The perils of unearned foreign income: Aid, remittances, and government corruption. *American Political Science Review* 106:146–65.

Andrews, Bound, and James H. Stock. 2005. Discussion Paper No. 1530. *Inference with weak instruments*. New Haven, CT: Yale University, Cowles Foundation for Research in Economics.

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2007. Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics* 139:116–32.

[17]BC present results from an Anderson–Rubin Wald test based on a chi-square distribution. The results are less favorable (i.e., *p*-values are larger) when using an *F*-distribution. Ideally, one would want *p*-values invariant to such choices. Perhaps more importantly, Andrews and Stock (2005, 10) note that "an *F* critical value is typically employed with the AR [Anderson–Rubin] test." (For a review of testing procedures in instrumental variables regressions, see Andrews, Moreira, and Stock 2007).

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* 91:444–55.
Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
Bound, John, David A. Jaeger, and Regina M. Baker. 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogeneous explanatory variable is weak. *Journal of the American Statistical Association* 90:443–50.
Brückner, Markus, and Antonio Ciccone. 2011. Rain and the democratic window of opportunity. *Econometrica* 79:923–47.
Ciccone, Antonio. 2011. Economic shocks and civil conflict: A comment. *American Economic Journal: Applied Economics* 3:215–27.
Coakley, Clint W., and Thomas P. Hettmansperger. 1994. The maximum resistance of tests. *Australian Journal of Statistics* 36:225–33.
Conover, W. J. 1999. *Practical Nonparametric Statistics*. New York: John Wiley & Sons.
Croux, Christophe, and Catherine Dehon. 2010. Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods & Applications* 19:497–515.
Deaton, Angus. 2005. Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics* 87:1–19.
Desmarais, Bruce A., and Jeffrey J. Harden. 2012. Comparing partial likelihood and robust estimation methods for the Cox regression model. *Political Analysis* 20:113–35.
Hájek, Jaroslav, and Zbynek Sidak. 1967. *Theory of Rank Tests*. New York: Academic Press.
Hansen, Ben B., and Jake Bowers. 2009. Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association* 104:873–85.
Hodges, J. L., and E. L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34:598–611.
Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81:945–60.
Honoré, Bo E., and Luojia Hu. 2004. On the performance of some robust instrumental variables estimators. *Journal of Business and Economic Statistics* 22:30–39.
Huber, Peter J., and Elvezio Ronchetti. 2009. *Robust statistics*. Hoboken, NJ: Wiley.
Hug, Simon. 2010. The effect of misclassifications in probit models: Monte Carlo simulations and applications. *Political Analysis* 18:78–102.
Imbens, Guido W., and Paul R. Rosenbaum. 2005. Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A* 168:109–26.
Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. Strengthening the experimenters toolbox: Statistical estimation of internal validity. *American Journal of Political Science* 56:484–99.
Kerner, Andrew, and Jane Lawrence. Forthcoming. What's the risk? Bilateral investment treaties, political risk, and fixed capital accumulation. *British Journal of Political Science*.
Lehmann, E. L. 1975. *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.
Marshall, Monty G., and Keith Jaggers. 2006. *Polity IV dataset*. College Park, MD: Center for International Development and Conflict Management, University of Maryland.
Mebane, Walter R., Jr, and Jasjeet S. Sekhon. 2004. Robust estimation and outlier detection for overdispersed multinomial models of count data. *American Political Science Review* 48:392–411.
Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112:725–53.
Rosenbaum, Paul R. 1996. Comment on Angrist and Krueger. *Journal of the American Statistical Association* 91:465–68.
———. 2002. *Observational Studies*. New York: Springer.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 56:688–701.
Sovey, Allison J., and Donald P. Green. 2011. Instrumental variables estimation in political science: A readers' guide. *American Journal of Political Science* 55:188–200.
Staiger, Douglas O., and James H. Stock. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65:557–86.
Stock, James H., Jonathan H. Wright, and Motohiro Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics* 20:518–29.
Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr, Michael C. Herron, and Henry E. Brady. 2001. The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review* 95:793–810.
Western, Bruce. 1995. Concepts and suggestions for robust regression analysis. *American Journal of Political Science* 39:786–817.

https://doi.org/10.1093/pan/mps037 Published online by Cambridge University Press