WISE WORKING PAPER SERIES

WISEWP0612

ARE “NEARLY EXOGENOUS” INSTRUMENTS RELIABLE？

Daniel Berkowitz And Mehmet Caner And Ying Fang

July 26, 2006

厦门大学王亚南经济研究院
The Wang Yanan Institute for Studies in Economics, Xiamen University

ARE "NEARLY EXOGENOUS" INSTRUMENTS RELIABLE?

Daniel Berkowitz,*
Mehmet Caner**
Ying Fang***

First draft: June 15, 2006

Revised: July 26, 2006

Abstract

Instrumental variable methods are widely used to make inferences about the impact of some variable on economic outcomes; for example, whether or not institutions influence long-term growth. The test-statistics used for making these inferences, however, are based on the generally unrealistic identifying assumption that the instruments are exogenous. We find that when carefully chosen instruments are more realistically modeled as "nearly" exogenous, the standard test-statistics are unreliable: the t-statistic substantially and unpredictably either overrejects or underrejects the null and the Anderson-Rubin test always overrejects. We show how an Anderson-Rubin test-statistic derived from the delete-d jackknife procedure developed by Wu [1986] can be used to make reliable inferences in small samples when instruments are "nearly exogenous." Our procedure adjusts the critical values according to the correlation between the instrument and structural error. We are able to do this both in exactly identified systems as well as overidentified ones. Furthermore, our test is robust to the weak instruments problem. We use this test to confirm and to correct inferences about the impact of institutions in the celebrated work of Acemoglu, Johnson and Robinson [2001].

*Department of Economics, University of Pittsburgh, dmberk@pitt.edu;
**Department of Economics, University of Pittsburgh, caner@pitt.edu;
***Department of Economics, University of Pittsburgh, yifst1@pitt.edu.

We have benefited from discussions with Daron Acemoglu, Karen Clay, Alexis Leon, and Nese Yildiz and all the seminar participants in Turkish Central Bank. We thank Debby Ziolkowski for excellent editing.

I. Introduction

Economists frequently apply instrumental variable methods to draw inferences about whether or not some variable influences an economic outcome. Labor economists employ varied instruments, including quarter and year of birth [Angrist and Krueger 1991], tuition and distance to nearest college [Kane and Rouse 1995], attending reform school [Meghir and Palene 1999] and birth year interacted with school buildings in region of birth [Dufflo 2001] to test for whether or not a person's education influences her salary and wages. In a more recent literature that combines macro-economics, political economy and comparative institutions, economists employ instruments including early settler mortality [Acemoglu, Johnson and Robinson 2001], ethnic capital [Hall and Jones 1999], ethno-linguistic fractionalization [Mauro 1995] and legal families [Djankov et al. 2003, and Acemoglu and Johnson 2006] to determine whether or not the quality of institutions influences long-term growth and investment.

If long-term growth is regressed on institutions and other relevant variables using ordinary least squares (OLS), then inferences can be made about whether or not institutions and long-term growth are correlated; however, we would not necessarily be able to infer whether or not institutions drive long-term growth. One reason for this is that long-term growth could in fact partially be responsible for the quality of institutions since a country that is wealthy can afford good institutions, while a poor country typically cannot [Glaeser et al. 2004]. Or, there may be an unobserved variable or noisily measured factor such as culture or the education level of early settlers that simultaneously drives the quality of institutions and long-term growth [Guiso, Sapienza and Zingales 2006]. In either case, institutions are endogenous, and inferences about causality cannot be made.

Instrumental variable methods are employed to make causal connections. Researchers pick relevant instruments: they should be related to the endogenous explanatory variable both on the basis of a priori argument and statistically. For example, Acemoglu et al. [2001], for herein sometimes denoted AJR [2001], argue that early settler mortality in colonies is strongly related to the quality of contemporary institutions that restrain the government from expropriating private assets. The a priori argument, roughly speaking, is that settlers who believed that they would live for a long time in their colony were more likely to invest in institutions that limit expropriation; settlers who anticipated that they could not survive very long in their colony would tend to set up extractive institutions; and the quality of institutions set up by all settlers tended to be persistent. Anticipated early settler mortality is proxied by using the disease environment in colonies around the time of settlement. This strong statistical relationship between the instrument (early disease environment) and endogenous explanatory variable (institutions hundreds of years later) is verified in a reduced form regression.[1]

Instruments must also be exogenous; that is, they are not related to the outcome variable after controlling for relevant explanatory variables. For example, early settler mortality is exogenous if it is not systematically related to long-term growth after controlling for institutions and other relevant variables such as population and latitude. This requirement, however, is very strong because it means that settler mortality can only influence long-term growth indirectly through the quality of contemporary institutions. The exogeneity of early settler mortality, however, is controversial: for example, as noted by Glaeser et al. [2004], early settler mortality could also influence long-term growth through its impact on the unobservable human capital of the early settlers. There are many other seemingly exogenous instruments that are also controversial. For example, Angrist [1990] argues that draft lottery numbers are instruments for testing whether serving in Vietnam affects the earnings of men in the civilian sector because these numbers influence earnings purely through military service. Wooldridge [2002, p.88] argues, however, that this is not necessarily true because civilian employers are more likely to invest in job training for employees who have low draft numbers. Therefore, lottery numbers could also influence earnings through job training, which is unobservable.

In this paper we develop a simple technique for making inferences about whether or not an endogenous variable matters for some outcome when instruments are "nearly exogenous." Nearly exogenous instruments influence outcome variables primarily through the endogenous explanatory variable, but they also plausibly and weakly influence the outcome through other unobserved channels. They are therefore weakly correlated with the error term in the structural equation. Once we model instruments as nearly exogenous and not perfectly exogenous, there is both bad news and good news. The bad news is that standard test-statistics for making inferences are unreliable: even when the instrument is very close to being exogenous, the t-test-statistic grossly and unpredictably overrejects or underrejects the null hypothesis and the one-sided Anderson-Rubin test overrejects. The good news is that we can make accurate inferences in small samples using an Anderson-Rubin statistic derived from the delete-d jackknife procedure (see Wu [1986]). Even though none of the resampling methods are consistent in this case, the delete-d jackknife method comes arbitrarily close to the true distribution in large samples. We also show that this method works well in small samples, and is better than any other method used so far in terms of size properties. More generally, our technique allows practitioners to use instrumental variable methods for carefully chosen instruments that, while not perfectly exogenous, are more realistically modeled as nearly exogenous.

This test-statistic corrects for correlations between instruments and the structural error term by adjusting the critical values according to the degree of correlation. Researchers often employ the Sargan test and Hansen's J-test to validate exogeneity in overidentified systems. It is well known, however, that both the Sargan and J-tests have low power and are unreliable for providing guidance about the validity of instruments (Bound et al. [1995]). Han and Hausman [2002] provide another test for validity that works when there are many instruments. It is, however, often difficult to find just one valid instrument. Our test can be used in exactly identified systems, and it is also robust to weak instruments.

In the next section we show that when instruments are relevant and nearly exogenous, inferences drawn from the t-test and the Anderson-Rubin test in two-stage least square systems are unreliable in small samples; and in section 3 we show that these problems hold in large samples. In section 4 we show that the t-statistic cannot be repaired, but the Anderson-Rubin test can be partially fixed using the delete-d jackknife procedure. In section 5 we use Monte Carlo simulations to understand how the delete-d jackknife Anderson-Rubin test can be reliably constructed in small samples. We show that the delete-d jackknife AR-test is less size distorted than the standard AR-test and t-test. In section 6 we use this test to confirm and correct inferences drawn about the impact of institutions on long-run growth by AJR [2001]. This test-statistic can be implemented using STATA and the general program is available at http://www.at.edu/~dmberk/ddj-ARtest.txt. In section 7 we conclude.

II. Inference Using the Standard Test-statistics

In this section we relax the assumption that instruments must be exogenous and introduce a definition of "near exogeneity." This section then delivers the bad news that the standard two-stage least squares (TSLS) test-statistics are unreliable when carefully chosen instruments are "nearly" exogenous. Subsequent sections, fortunately, report the good news that jackknife techniques can be used to derive a reliable test-statistic.

Suppose we want to check for whether or not an institution, say property rights enforcement, influences long-term growth in a sample of countries.[2] If we suspect that institutions are endogenous, and we also believe that a linear specification is appropriate, we would estimate and compute test-statistics for the following simple linear simultaneous equations model (Hausman [1984] and Phillips [1984]):

$$LRGr = \beta_0 + \beta_1 INST + u; \qquad (1)$$

$$INST = \Pi_0 + Z\Pi_1 + V. \qquad (2)$$

Equation (1) is the structural equation, where LRGr is an nx1 vector of long-run growth, INST is an nx1 vector of institutions, and *u* is an nx1 vector of structural error terms that have zero mean and finite variance $\sigma_u^2 < \infty$. Equation (2) is the reduced form, Z is an nxk matrix of instruments and *V* is an nx1 vector of reduced form errors that has a zero mean and finite variance. $\sigma_V^2 < \infty$. The error terms *u* and *V* may be correlated and n represents the number of countries. The parameters β_0, β_1, Π_0 *and* Π_1, are unknowns, and, for notational conventional, we denote $\beta = \{\beta_0, \beta_1\}$, $\Pi = \{\Pi_0, \Pi_1\}$. Other covariates, for example, population, latitude and education, can be added to the system in equations (1) and (2) without loss of generality.[3]

In order to determine whether or not institutions matter, we estimate the unknown parameter β_1 and use test-statistics to check whether $\beta_1 = 0$. To do this properly, we need valid instruments that are both relevant and exogenous. As previously discussed, relevant instruments are picked on the basis of a theoretical, institutional and/or historical argument, and are validated ex post by estimating the reduced form. Staiger and Stock [1997] propose an F-statistic of at least 10 for the null that $\Pi_1 = 0$ as ex post validation of relevance. The second criterion for validity is that instruments are exogenous, which implies they are orthogonal to the error term in the structural equation:

(3) $$Exogenous \Rightarrow \ Cov\, Z_i^{'} u_i = 0\,.$$

It is generally difficult, as we have previously argued, to find instruments that satisfy this strong condition. We want to check, then, if we can make reliable inferences about institutions when instruments are relevant but, as in the case of early settler mortality, may not be exogenous. In particular, while these instruments influence long-run growth in the structural equation primarily through institutions, they may also be weakly correlated with unobserved factors that can also influence long-term growth. We model this potential small correlation as "nearly exogenous," which is a local to zero setup:

(4) $$Nearly\ Exogenous \Rightarrow \ Cov\ Z_i^{'} u_i = C/\sqrt{n}\ \ is\ small$$

where C is an nx1 vector of constants.

If we choose $Cov\ Z_i^{'} u_i = C$ to capture near exogeneity, then the test-statistics always diverge in the limit. Thus, this assumption does not provide any guidance for finite sample behavior when there is some mild correlation between the instrument and error.

In what follows, small sample simulation methods are used to show that even a slight relaxation of the exogeneity assumption in equation (3) makes the standard test-statistics unreliable. Suppose we employ the TSLS t-test to determine whether or not institutions matter. Denoting the H_0 and H_1 as the null and the alternative, and $\hat{\beta}_{1,TSLS}$ as the TSLS estimator of β_1, we use the t-statistic to test

$H_0 : \beta_1 = 0$, against

$H_1 : \beta_1 \neq 0$, where the t-statistic is given by

$$(5) \qquad t = \hat{\beta}_{1,TSLS} / \sqrt{a \operatorname{var} \hat{\beta}_{1,TSLS}} \ .$$

In figures 1-2, we use standard methods to simulate the distribution of the t-statistic for a sample of 100 countries with instruments that are exogenous and nearly exogenous. For simplicity and no loss of generality, the intercept coefficients $\beta_0\ and\ \Pi_0$ are both set at 0 and the true value of the coefficients $\beta_1\ and\ \Pi_1$ are set at 0 and 1, respectively. Thus, institutions are identified by a strong instrument and the true null hypothesis is that institutions do not matter.

We generate i.i.d. data for the one instrument, the structural error term and reduced form, (Z,u,V), from a joint normal distribution N(0, Λ) and

$$(6) \qquad \Lambda = \begin{pmatrix} 1 & Cov\, Z_i u_i & 0 \\ Cov\, Z_i u_i & 1 & Cov\ V_i u_i \\ 0 & Cov V_i u_i & 1 \end{pmatrix}.$$

where Cov Z_i'u_i measures the correlation between the instrument Z and the error term u, and Cov V_i'u_i measures the endogeneity of institutions, which is set to 0.25 in all simulations. When the i.i.d. data (Z,u,V) are generated, we can derive the observation of and INST and LRGr by using equations (1) and (2) and specified true values of β_1 *and* Π_1. Based on the information of LRGr, INST and Z, we compute the t-statistic and then test whether the null of β_1= 0 can be rejected at the 5% level by using the critical value 1.95. We replicate the simulation by 1000 times to derive the distribution of the t-statistic and calculate the actual rejection probability which is reported in Table 1.

Figure 1 illustrates the distribution of the t-statistic when the instrument is exogenous, and nearly exogenous with small positive correlation: Cov $Z_i u_i$ = 0.10. The distribution under exogeneity is close to a standard normal distribution, and the distribution under near exogeneity shifts to the right and is close to a normal distribution with a nonzero mean. This shift implies that the null is falsely rejected 19.2% of the time from the right-hand tail, which is much higher than the appropriate 2.5% rate. The null is falsely rejected at the 0.2% rate from the left-hand tail, which is conservative; and, the two-sided test falsely rejects at 19.4% rate, which is almost quadruple the nominal 5% rate.

Figure 2 compares distributions when the instrument is exogenous and then nearly exogeneous: Cov Z_i'u_i = - 0.10. The t-statistic is conservative on the right-hand side and falsely rejects roughly 0.3% of time. It overrejects from the left-hand side at a 14.0% rate; and, the two-sided test has size problems and falsely rejects 14.3% of the time.

Table 1 reports rates of right-hand side and left-hand false rejection when the instrument is more weakly correlated with the error term: Cov Z_i'u_i = 0.06 or -0.06 and illustrates that as the absolute value of the correlation decreases, the size problems of the two-sided t-test are mitigated. When the correlation is positive there is a 9.4% false rejection rate on the right-hand side, a conservative 0.4% rate from the left-hand side and an overall 9.8% false rejection rate. When the correlation is negative, the rates of false rejection on the right-hand and left-hand sides are 0.6% and 7.2%, respectively, and the overall false rejection rate is 7.9%.

Suppose we test the null against the alternative using the one-sided Anderson-Rubin (Anderson and Rubin [1949]) test:

(7)[4] $$AR(\beta_1 = 0) = LRGr' P_z \, LRGr \, / \, (LRGr' M_z \, LRGr) / (n-2) .$$

Here, $AR(\beta_1 = 0)$ is the test-statistic for the null, $P_z = Z(Z'Z)^{-1} Z$ is the projection matrix and $M_z = I - P_z$.

Figure 3 compares simulations of the small sample distributions of the Anderson-Rubin statistic when the instrument is exogenous and nearly exogenous. Under exogeneity the distribution is close to a standard chi-square and, when Cov Z_i'u_i = 0.10, the distribution shifts to the right and is close to a non-centered chi-square. Because this is a one-sided test, the shift depends only on the *absolute value* of the correlation. If we set the critical value at 3.85, the nominal probability of falsely rejecting is 5%, and the actual rate under near exogeneity is 17.5% so that near exogeneity creates small sample problems.

Table 1 illustrates that the small sample problems of the Anderson-Rubin test (for herein, denoted the AR-test) are also diminished when the instrument is less endogenous. When the correlation decreases to 0.06, the AR-test falsely rejects 9.4% of the time. Since it is not possible to calculate the absolute value of the correlation between the instruments and structural error, it is not possible to adjust for this small sample distortion and the AR-test is also unreliable.

III. Large Sample Distributions

This section adds to the bad news: we show that the shifts in test-statistic distributions observed in the small sample simulations for nearly exogenous instruments also hold in limit. For the next three sections of the paper, we generalize the simultaneous equations system equations (1) and (2) to model a more general system with m ≥ 1 endogenous explanatory variables, and k ≥ m instruments:

$$\text{(1*)} \qquad y = Y\beta + u$$

$$\text{(2*)} \qquad Y = Z\Pi + V$$

where y is an nx1 vector of some outcome variable, Y is an nxm matrix of endogenous explanatory variables, Z is an nxk matrix of instruments, u is an nx1 vector of structural errors, V is an nxm matrix of reduced form errors; the errors have zero means and finite variance, and u and V are correlated with each other. As noted before, other exogenous covariates can be added to the system.

In the next theorem, we show that near exogeneity shifts the asymptotic distribution of the t-statistic to normal with nonzero mean.

Theorem 1: Suppose that the instrument is nearly exogenous according to (4), and the standard assumption 2 in the appendix holds. Then,

$$t \overset{d}{\rightarrow} N[\sigma_u^{-1}(\Pi' Q_{zz} \Pi)^{-1/2} \Pi' C, 1] \tag{8}$$

where σ_u is the square root of σ_u^2, and Q_{zz} is the second moment matrix of instruments.

Proof. See the Appendix.

According to Theorem 1, the mean of the distribution depends upon the parameter C, which, by equation (4), is related to the small correlation between structural error and instruments. When C=0 and the instruments are exogenous, the t-statistic converges to the standard normal distribution. When C>0 (given $\Pi > 0$), the distribution shifts to the right. When C<0 (given $\Pi > 0$), the distribution shifts to the left. Since we cannot consistently estimate C let alone know its sign, we cannot use this large sample theorem to improve inference.

The next theorem characterizes the impact of near exogeneity on the distribution of the AR-test, which is now more generally defined from equation (7) for k instruments and m endogenous explanatory variables:

(7*) $AR(\beta_0)= (y - Y\beta_0)'P_z(y - Y\beta_0)/(y - Y\beta_0)'M_z(y - Y\beta_0)/(n-k-m)$

We use this statistic to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ where β_0 is the true value.

Theorem 2: Suppose that the instrument is nearly exogenous according to (4), and the standard assumption 2 in the appendix holds. If the null hypothesis is $\beta = \beta_0$, then

(9) $AR(\beta_0) \overset{d}{\rightarrow} \chi^2_K(\varsigma)$

where $\chi^2_K(\varsigma)$ is a non-central chi-square distribution with k degrees of freedom and the non-centrality parameter $\varsigma = C'\Omega^{-1}C$, where $\Omega = \sigma^2_u \otimes Q_{zz}$.

Proof. See Caner [2006] and Fang [2006].

According to Theorem 2, the mean of the non-centrality parameter is quadratic in parameter C. Therefore, when C=0 the AR-test converges to the centered chi-square distribution, and when C≠0 the distribution shifts to the right. Again, since we do not know C, we cannot use these theorems to obtain appropriate critical values.

IV. Reliable Inference under Near Exogeneity

This section contains the good news that it is possible to make reliable inferences when instruments are nearly exogenous. When large sample results are problematic, a standard remedy is to employ resampling methods including the bootstrap, the jackknife and subsampling. While we cannot repair the t-statistic, we show that we can adjust the Anderson-Rubin test using the delete-d jackknife procedure developed by Wu [1986] and get very close to the true limiting distribution and, more importantly, obtain good small sample properties.

The reason why the t-statistic cannot be fixed using resampling methods is that it contains the TSLS estimator $\hat{\beta}_{1,TSLS}$ (see equation (5)). Resampling techniques that are designed to pick up correlations between instruments and the structural error term will fail because the TSLS estimator uses the estimated residual vector which, *by construction,* is orthogonal to the instrument. Thus, resampling procedures are forced essentially to ignore correlations between instruments and structural error terms.

More formally, let t_S denote the delete-d jackknife t-statistic (for herein, denoted as the ddj t-statistic):

$$(10) \qquad t_S = \hat{\beta}_{1S,TSLS} - \hat{\beta}_{1,TSLS} / \sqrt{a\,\text{var}\,\hat{\beta}_{1S,TSLS}}$$

where $\hat{\beta}_{1S,TSLS}$ is the ddj estimator, $a\,\text{var}\,\hat{\beta}_{1S,TSLS}$ is its estimated variance and $\hat{\beta}_{1,TSLS}$ is the TSLS estimator for the full sample. The calculation of the ddj test-statistic at the 10% level is implemented using the following algorithm:

Step 1: Pick d observations to be deleted: $d = \gamma n, \textit{ where } 0 < \gamma < 1,$ where n is the sample size, and then delete d randomly chosen observations from the sample;

Step 2: For the block size b = n-d, compute the TSLS estimator and its corresponding estimated variance, $\hat{\beta}_{1S,TSLS}$ and $a\,\text{var}\,\hat{\beta}_{1S,TSLS}$, and then compute the ddj t-statistic as defined in (10);

Step 3: Put the d observations back into the sample and then repeat steps 1 and 2 at least 1000 times and then sort these computed ddj t-statistics (sampling without replacement);

Step 4: Use the 90% percentile ddj t-statistic as the data-dependent critical value;

Step 5: We reject the null hypothesis when the t-statistic from the full sample is larger than the data-dependent critical value found in step 4.

The next theorem characterizes the limiting distribution of the ddj t-statistic. To derive this, we set $d = \gamma n,\ where\ 0 < \gamma < 1,\ and\ n \to \infty$.

Theorem 3: If the instrument is nearly exogenous according to (4), and the standard assumption 3 in the Appendix holds, then

(11) $t \overset{d}{\to} N[0,\ 2 - \gamma - 2\sqrt{1-\gamma}\ ]$

$where\ \gamma = d/n\ and\ 0 < \gamma < 1.$

Proof. See the Appendix.

Theorem 3 shows that the limiting distribution of the ddj t-statistic deviates from the true distribution in Theorem 1: $t \overset{d}{\to} N[\sigma_u^{-1}(\Pi' Q_{zz}\Pi)^{-1/2}\Pi' C,\ 1]$; the mean is not zero and the variance is not one. Thus, the delete-d jackknife procedure fails to correct for C≠0. Because we cannot estimate the sign or size of C, we cannot pick critical values that allow us to make reliable inferences. We have shown in finite sample simulations (available upon request) that the ddj t-test has massive size problems when instruments are exogenous and nearly exogenous.

We can, however, compute a delete-d jackknife Anderson-Rubin test-statistic (for, herein denoted the ddj AR-test) to account for the noncentral chi-square distribution that emerges under near exogeneity. Let y_b, Y_b and Z_b denote, respectively, subvectors or submatrices of y, Y and Z, where d is the number of observations randomly deleted (without replacement), and b = n – d is the block size: y_b is a bx1 vector Y_b is a bxm matrix and Z_b is a bxk matrix, and the AR-test-statistic for any block is denoted $AR_S(\beta_0)$:

$$(11)\ AR_S\,(\beta_0) = (y_b - Y_b\beta_0)' P_{zb}\,(y_b - Y_b\beta_0) / (y_b - Y_b\beta_0)' M_{z_b}\,(y_b - Y_b\beta_0) / (b - k - m)\,.$$

We compute the ddj AR-test using the similar five steps for computing the ddj t-test except that at step 2 we compute the ddj AR-test defined in equation (11), and test the null that $\beta = \beta_0$. However, in steps 1-5 to compute the ddj AR-test, we use β_0 rather than the estimator of β. Again, we reject when the full sample AR-test-statistic exceeds the data-dependent critical value. This delete-d jackknife procedure partially accounts for the correlation between structural errors and instrument.

It is important to note that the bootstrap procedure cannot solve the near exogeneity problem because it requires that the correlation between the instruments and structural errors be estimated in the bootstrap samples, and it is impossible to obtain estimates that are consistent. Subsampling also does not work because it requires very small block sizes, and therefore it cannot replicate these correlations. These results are established in Caner [2006] and Fang [2006].

The next theorem characterizes the limiting distribution of $AR_S\,(\beta_0)$.

Theorem 4: Suppose the instrument is nearly exogenous according to (4), and the standard assumption 3 in the Appendix holds. If the null is $\beta = \beta_0$, then

$$(12)\qquad AR_S\,(\beta_0) \overset{d}{\to} \chi^2_K\,(\tilde{\varsigma})$$

where $\chi^2_K(\tilde{\varsigma})$ *is a noncentral chi-square distribution with k degrees of freedom and* $\tilde{\varsigma}$ *is the noncentrality parameter:* $\tilde{\varsigma} = (1-\gamma)C'\Omega^{-1}C$, *and* $\Omega = \sigma^2_u \otimes Q_{zz}$, *where* $1-\gamma$ = *b/n is the share of the observations that is resampled.*

Proof. See Caner [2006] and Fang [2006].

Theorem 4 shows the delete-d jackknife procedure generates a large sample chi-square distribution with a noncentrality parameter that is equal to b/n times the noncentrality parameter in Theorem 2. The distribution of the ddj AR-test in Theorem 4 is very close to the true distribution in Theorem 2. Thus, a large block size is appropriate for obtaining an accurate limiting distribution.

Regarding small samples, Wu [1990] argues that a block size between 1/4th and 3/4th's of the sample size is desirable for reducing size distortion. Block sizes that are less than 1/4th are only relevant for subsampling, and block sizes greater than 3/4th's are very conservative, and therefore, have severe power problems. We use Monte Carlo simulations in the next section and find that a block size of 1/4th appears to be most appropriate for the small sample of 64 countries employed in AJR [2001], and a block of size of 3/8th's provides a conservative robustness check. More extensive simulation studies, however, are required for the choice of block size.

V. Monte Carlo Simulations

In this section we conduct Monte Carlo simulations showing that the ddj AR-test has good small sample properties when the block size is set at 1/4th the sample size. We simulate the linear simultaneous equations model defined in (1*) and (2*) with one endogenous variable and one instrument. All of our results are robust when we over-identify using two instruments. The true value of the structural parameter β is $\beta_0 = 0$. We set the sample size equal to 64 in order to conduct comparisons of various tests' performance with AJR [2001]. The *i.i.d.* data (Z_i, u_i, V_i) are generated from a joint normal distribution $N(0, \Lambda)$ which is described in (6), and there is endogeneity: $\text{cov}\, V_i u_i = 0.25$. The measure of near exogeneity, $\text{cov}\, Z_i u_i$ can take on values of 0.10 or 0.15. We set Π (the regressor for the instrument) at either 0.1 or 1 in all cells of the vector to represent a weak and a strong instrument. The nominal size is 10%.[5]

We have shown that only the Anderson-Rubin test can be repaired with the delete-d-jackknife procedure. Table II Panel A reports the rate of false rejection for the full sample ("unrepaired") AR-test when the instrument is strong; Panel B reports these rates when the instrument is weak. The AR-test, as predicted by Theorem 2, clearly has poor small sample properties. It is striking that the small sample properties are virtually similar for the strong and weak instruments; under exogeneity, the false rejection rate is roughly 10%; when $\text{cov}\, Z_i u_i$ = 0.10, the rate is about 23%; and when $\text{cov}\, Z_i u_i$ = 0.25, the false rejection rate is 34%. The reason that the distinction between strong and weak instruments does not matter is that the AR-test does not rely on estimates of the reduced form parameter Π (see equations (7) and (7*)).

Table III reports the small sample properties of the ddj AR- test for block sizes covering 1/4th to1/2 the sample: $b = \{16, 18, 20, 22, 24, 28, 30, 32\}$. Because the ddj AR-test also does not rely on estimates of the reduced form parameter (see eq (11)), it has similar small sample properties for strong and weak instruments. Thus, with no loss of generality, we discuss results for the case of strong instruments. When the block size is large, for example, b=32, the rates of false rejection are 1.3% and 2.3% respectively, when the correlation between instruments and structural errors are 0.10 and 0.15, this result is highly conservative since the nominal size is 10%. As the block size shrinks, there are more rejections. When b=16 covering 1/4th the sample, the false rejection rate is 7.4% and 11.8% when $\operatorname{cov} Z_i u_i = 0.10 \textit{ and } 0.15$, and the small sample properties are quite good. When b = 24 covering 3/8th's the sample, the false rejection rate is 4.0% and 9.5% when $\operatorname{cov} Z_i u_i = 0.10 \textit{ and } 0.15$, and the test is more conservative.

By comparing Tables II and III, we see that regardless of the choice of block size, the ddj AR-test is less size distorted than the AR-test. Our advice for practitioners is to first pick a block size that is 1/4th the sample; then use a block size that is 3/8th's the sample as a conservative robustness check. If the results are similar, then the inferences are reliable.

VI. Implementation using Early Settler Mortality

In this section, we use the ddj AR-test to check inferences made about the impact of institutions, INST, on long-run growth, LRGr, where Z is the instrument (early settler mortality) from AJR [2001]. We add X, an nxh vector of controls to equations (1) and (2), where X is the null set in some of the regressions, and includes combinations of variables such as latitude, continent dummies, colonial and legal origins, etc.:

$$LRGr = \beta_0 + \beta_1 INST + \beta_2 X + u \qquad (1)$$

$$INST = \Pi_0 + Z\,\Pi_1 + X \prod_2 + V \qquad (2)$$

We test the null $H_0 : \beta_1 = 0$ against the alternative $H_1 : \beta_1 \neq 0$. Tables IV-VI contain sets of control variables used in AJR [2001]. Panel A contains point estimates and standard errors (in parentheses); panel B contain test-statistics including the regular t-statistic and associated p-values, the regular (full-sample) AR-test statistic, the p-values for the ddj AR-test when the block size is 16 (1/4th the sample size), 24 (3/8th's the sample size) and for the full sample. AJR [2001] use the t-statistic for making inferences; we check these inferences primarily with the ddj AR-test with block size 16 and then with block size 24 as a conservative robustness check. Finally, we also compare p-values for the ddj AR-test and the full sample AR-test to get a sense of the endogeneity of the instrument.

Table IVa replicates and then checks inferences made in the baseline regressions in AJR [2001], Table IV. In column (1) there are no control variables; the p-value of the ddj AR-test when b=16 is 0.012, and institutions are significant at 10% the level. In column (2) we control for latitude and institutions continue to be significant at the 10% level. In column (3) we add Asia, Africa and "other" continent dummy variables: latitude is included in column (4). The ddj AR-tests have p-values of 0.082 and 0.094, respectively. The p-values of the ddj AR-test when b=24 marginally exceed 0.10 in two out of four cases; however, this is a conservative robustness test. Generally, we can say that at the 10% level we find evidence that institutions matter for long-run growth.

In Tables V and VI we check for the significance of institutions with additional controls (see AJR [2001] Tables V and VII). The ddj AR-test-statistics confirm that institutions matter. In Table V, the British and French colonial dummies or the French legal origin dummy are included as controls: because the ddj AR-test has p-values between 0.022 and 0.051 when b=16 and the p-values are never greater than 0.06 for the conservative test with b=24, we always reject the null at the 1 level.

Table VI includes contemporary health-related variables, including malaria in 1994, life expectancy in 1995 and infant mortality in 1995. The standard t-test and AR-test always reject the null at the 10% level. The more reliable ddj AR-test with b =16, however, fails to significantly reject the null in all cases. When we control for malaria in column (1), the p-value of ddj AR-test is 0.130. When we control for both malaria and latitude in column (2), the p-value is 0.147. When we control for life expectancy in column (3), the p-value is 0.152; when we control for both life expectancy and latitude in column (4), the p-value is 0.147. In column (5) we add infant mortality and in column (6) we add both infant mortality and latitude; and, the p-values are 0.161 and 0.202.

The reason why institutions are marginally significant in Table VI is that they are also correlated with the contemporary health variables. For example, the correlation between log settler mortality and malaria risk is 0.67 (Gallup and Sachs [2001]; Glaeser et al. [2004]). Thus, the correlation between the instrument and the structural error terms depends upon the correlation between the instrument and control variables in the structural equation. The higher this correlation, the less nearly exogenous is the instrument. From the simulations in Table III, it is clear that the larger the correlation between the structural errors and the instruments, the larger is the difference between the sizes of the regular AR-test based on a chi-square distribution and delete-d jackknife AR-test based on data dependent critical values. The same is true for differences in p-values for the AR-test and the ddj AR-test.

The difference between p-values of the ddj AR-test and the AR-test in Table VI are relatively large compared to those in Tables IV and V, hence leading to a failure to reject the null in Table VI. However, the level of near exogeneity is not enough to overturn most of the AJR [2001] findings.

VII. Conclusions

Instrumental variable methods have been used by economists to identify casual relations between variables such as institutions and long-run growth, or education and job market performance. It is clear, however, that it is difficult to find instruments that are truly exogenous. We have shown that once we relax the exogeneity assumption to allow for near exogeneity, the standard test-statistics are unreliable. More constructively, we find that it is also possible to use jackknife methods to repair the Anderson-Rubin test so that reliable inferences can be made. Our method is novel because it enables practitioners to validate near exogeneity in exactly identified as well as overidentified systems. It can also be used for weak instruments.

Appendix

In the beginning of this appendix, we first list near exogeneity assumption and some moment conditions that are required to obtain the theorems in the paper. Assumptions 1 and 2 are sufficient for Lemma 1, Theorem 1 and Theorem 2. Assumptions 1 and 3 are sufficient for Theorem 3 and Theorem 4.

Assumption 1: Near Exogeneity $E\left[Z_i^{'}u_i\right]=C/\sqrt{N}$, where C is a fixed $K\times 1$ vector.

Assumption 2: The following limits hold jointly when the sample size N converges to infinity:

(a) $(u'u/N, V'u/N, V'V/N)\overset{p}{\rightarrow}(\sigma_u^2, \Sigma_{Vu}, \Sigma_{VV})$, where σ_u^2 , Σ_{Vu} and Σ_{VV} are respectively a 1×1 scalar, an $m\times 1$ vector and an $m\times m$ matrix.

(b) $Z'Z/N\overset{p}{\rightarrow}Q_{ZZ}$ where Q_{ZZ} is a positive definite, finite $K\times K$ matrix.

(c) $(Z'u/\sqrt{N}, Z'V/\sqrt{N})\rightarrow(\overline{\Psi}_{Zu}, \Psi_{ZV})$, and

$$\begin{pmatrix}\overline{\Psi}_{Zu}\\ \Psi_{ZV}\end{pmatrix}\rightarrow N\left[\begin{pmatrix}C\\ 0\end{pmatrix}, \Sigma\otimes Q\right]$$

where $\Sigma=\begin{pmatrix}\sigma_u^2 & \Sigma_{Vu}^{'}\\ \Sigma_{Vu} & \Sigma_{VV}\end{pmatrix}$.

These convergences in Assumption 2 are not primitive assumptions but hold under weak primitive conditions. Parts (a) and (b) follow from the weak law of large numbers, and Part (c) follows from triangular arrays central limit theorem. Instead of a mean zero normal distribution in Staiger and Stock [1997], the $\overline{\Psi}_{Zu}$ in (c) is a normal distribution with nonzero mean, which is a drift term C coming from the near exogeneity assumption. For any independent sequence $Z_i'u_i$, if $E\left[Z_i'u_i\right]^{2+\delta} < \Delta < \infty$ for some $\delta > 0$ for all $i = 1, 2, 3, ..., N$, then Liapunov's theorem leads to the limiting results in (c); see James Davidson (1994).

Assumption 3: Define

$\sigma_b = E(u_b'u_b / b)$

and

$Q_b = E(Z_b'Z_b / b)$

Assume the following conditions hold jointly for $\delta > 0$,

(a) $E\left|z_{b,i}u_b\right|^{2+\delta} < \Delta_1 < \infty$ for all $b < N$ and all $1 \le i \le K$

(b) $E\left|z_{b,i}z_{b,j}\right|^{1+\delta} < \Delta_2 < \infty$ for all $b < N$ and all $1 \le i, j \le K$

(c) $E\left|u_b^2\right|^{1+\delta} < \Delta_3 < \infty$ for all $b < N$

(d) $\sigma_b \to \sigma_u^2 > 0$ uniformly as $b \to \infty$

(e) $Q_b \to Q_{ZZ}$ uniformly and uniformly positive definite as $b \to \infty$

Lemma 1. If Assumptions 1 and 2 hold for the model defined by (1**) and (2***), then the TSLS estimator* $\overset{\wedge}{\beta}_{TSLS}$ *is consistent and*

$$\sqrt{N}(\overset{\wedge}{\beta}_{TSLS}-\beta_0)\xrightarrow{d}N((\Pi'Q_{ZZ}\Pi)^{-1}\Pi'C,\sigma_u^2(\Pi'Q_{ZZ}\Pi)^{-1})$$

where $u'u/N\rightarrow E(u_i^2)=\sigma_u^2$, $Z'Z/N\rightarrow E(Z_i^{'}Z_i)=Q_{ZZ}$.

Proof of Lemma 1: We know that

$$\overset{\wedge}{\beta}_{TSLS}=(Y'P_ZY)^{-1}(Y'P_Zy).$$

So we have

$$\sqrt{N}(\overset{\wedge}{\beta}_{TSLS}-\beta_0)$$
$$=[(\frac{Y'Z}{N})(\frac{Z'Z}{N})^{-1}(\frac{Z'Y}{N})]^{-1}[(\frac{Y'Z}{N})(\frac{Z'Z}{N})^{-1}(\frac{Z'u}{\sqrt{N}})]$$

By Assumption 2, we can obtain that

$$[(\frac{Y'Z}{N})(\frac{Z'Z}{N})^{-1}(\frac{Z'Y}{N})]^{-1}$$
$$\xrightarrow{p}(\Pi'Q_{ZZ}\Pi)^{-1}$$

Now, we consider

$$\frac{Z'u}{\sqrt{N}}=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[Z_i^{'}u_i-E(Z_i^{'}u_i)]+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}E(Z_i^{'}u_i)$$

By the triangular array central limit theorem, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[Z_i^{'}u_i-E(Z_i^{'}u_i)]\xrightarrow{d}N[0,\sigma_u^2Q_{ZZ}].$$

By the triangular array weak law of large number and Assumption 1, we have

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}E(Z_i^{'}u_i)\xrightarrow{p}C.$$

Combining the above results, we obtain

$$\frac{Z'u}{\sqrt{N}} \xrightarrow{d} N[C, \sigma_u^2 Q_{ZZ}].$$

Then the result in the lemma follows directly. *Q.E.D.*

Lemma 1 summarizes the limiting results of the TSLS estimator under near exogeneity. The reason why we can obtain a consistent estimator under near exogeneity is because the correlation between instruments and structural errors shrinks toward zero asymptotically. When C=0, we can obtain the regular results of the TSLS estimator under the orthogonality condition. Instead of a normal distribution with a zero mean, near exogeneity can shift the distribution away from the zero mean. The nonzero mean depends on an unknown local to zero parameter C which is impossible to be estimated consistently [Andrews 2000].

Proof of Theorem 1: The result in the theorem directly follows from Lemma 1. *Q.E.D.*

Proof of Theorem 3: As defined in (10),

$$t_S = \frac{\overset{\wedge}{\beta}_{S,TSLS} - \overset{\wedge}{\beta}_{TSLS}}{\sqrt{a\,\mathrm{var}(\overset{\wedge}{\beta}_{S,TSLS})}}$$

where

$$a\,\mathrm{var}(\hat{\beta}_{S,TSLS}) = \hat{\sigma}^2_{\mu,b}[(Y_b{}'Z_b)(Z_b{}'Z_b)^{-1}(Z_b{}'Y_b)]^{-1}$$

and

$$\hat{\sigma}^2_{u,b} = (y_b - Y_b\hat{\beta}_{S,TSLS})'(y_b - Y_b\hat{\beta}_{S,TSLS})/(b-K-m)$$

By Assumption 3 and weak law of large number [Fang 2006], we have

$\hat{\sigma}_{u,b}^2 \to \sigma_u^2$ in probability,

and

$$[(\frac{Y_b' Z_b}{b})(\frac{Z_b' Z_b}{b})^{-1}(\frac{Z_b' Y_b}{b})]^{-1}$$
$$\xrightarrow{p} (\Pi' Q_{ZZ} \Pi)^{-1}.$$

The t_S-statistic can be rewritten as

$$t_S = \frac{(\hat{\beta}_{S,TSLS} - \beta_0) - (\hat{\beta}_{TSLS} - \beta_0)}{\sqrt{a\,\mathrm{var}(\hat{\beta}_{S,TSLS})}}$$

Consider the first term in the above equation,

$$\sqrt{b}(\hat{\beta}_{S,TSLS} - \beta_0)$$
$$= [(\frac{Y_b' Z_b}{b})(\frac{Z_b' Z_b}{b})^{-1}(\frac{Z_b' Y_b}{b})]^{-1}[(\frac{Y_b' Z_b}{b})(\frac{Z_b' Z_b}{b})^{-1}(\frac{Z_b' u_b}{\sqrt{b}})]$$

By Assumption 3 and the triangular array central limit theorem, we can obtain

$$\frac{Z_b' u_b}{\sqrt{b}} = \frac{1}{\sqrt{b}}\sum_{i=1}^{b}[Z_{b,i}u_{b,i} - E(Z_{b,i}u_{b,i})] + \frac{1}{\sqrt{b}}\sum_{i=1}^{b}E(Z_{b,i}u_{b,i})$$
$$\xrightarrow{d} N[0, \sigma_u^2 Q_{ZZ}] + (\sqrt{1-\gamma})C$$
$$= N[(\sqrt{1-\gamma})C, \sigma_u^2 Q_{ZZ}].$$

So we have

$$\frac{\sqrt{b}(\hat{\beta}_{S,TSLS} - \beta_0)}{\sqrt{\sigma_u^2 (\Pi' Q_{ZZ} \Pi)^{-1}}} \xrightarrow{d} N[\delta_C, 1]$$

where

$$\delta_C = \sigma_u (\Pi' Q_{ZZ} \Pi)^{-1/2} \Pi' (\sqrt{1-\gamma})C$$

By the similar method, noting that $\sqrt{b} = \sqrt{1-\gamma} \times \sqrt{N}$ we can obtain that

$$\frac{\sqrt{b}(\hat{\beta}_{TSLS} - \beta_0)}{\sqrt{\sigma_u^2 (\Pi' Q_{ZZ} \Pi)^{-1}}} \xrightarrow{d} N[\delta_C, (1-\gamma)]$$

Then the result in the theorem follows from above. *Q.E.D.*

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson (2001), “The Colonial rigins of Comparative Development: An Empirical Investigation,” *American Economic Review,* 91, 1369-1401.

Acemoglu, Daron and Simon Johnson (2006), “Unbundling Institutions,” *Journal of Political Economy*, 113, 949-995.

Anderson, Theodore Wilbur and Herman Rubin (1949), “Estimators of the Parameters of a Single Equation in a Complete Set of Stochastic Equations,” *The Annals of Mathematical Statistics*, 20, 46-63.

Andrews, Donald W.K. (2000), “Inconsistency of the Bootstrap with a Parameter Is on the Boundary of the Parameter Space,” *Econometrica*, 68,399-405.

Angrist, Joshua D. (1990), “Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records,” *American Economic Review*, 80, 313-336.

Angrist, Joshua D. and Alan B. Krueger (1991), “Does Compulsory School Attendance Affect Schooling and Earnings,” *Quarterly Journal of Economics,* 106, 979-1014.

Bound, John, David A. Jaeger, and Regina M. Baker (1995), “Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90, 443-450.

Caner, Mehmet (2006), “Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Fixed and Many Moment Asymptotics”, working paper, University of Pittsburgh, http://www.pitt.edu/~caner/wegel.pdf.

Davidson, James (1994), *Stochastic Limit Theory: An Introduction for Econometricians*, Cambridge: Oxford University Press.

Davidson, Russell. and James G. McKinnnon (1993), “ Estimation and Inference in Econometrics,” New York: Oxford University Press.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer (2003), “Courts,” *Quarterly Journal ofEconomics,* 118, 453-517.

Duflo, Esther (2001), “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment,” *American Economic Review*, 91, 795-813.

Fang, Ying (2006), “Instrumental Variables Regression with Weak Instruments and Near Exogeneity,” working paper, University of Pittsburgh.

Gallup, John L. and Jeffrey D. Sachs (2001), “The Economic Burden of Malaria,” *The American Journal of Tropical Medicine and Hygiene*, 64(1-2), (Suppl.), 85-96.

Glaeser, Edward, Rafael La Porta, Florencio Lopez-de-Silanes and Andrei Shleifer (2004), “Do Institutions Cause Growth?”, *Journal of Economic Growth*, 9, 271-303.

Guiso, Luigi, Paola Sapienza and Luigi Zingales (2006), “Does Culture Affect Economic Outcomes,” forthcoming in the *Journal of Economic Perspectives.*

Hall, Robert E. and Charles I. Jones (1999), “Why Do Some Countries Produce So Much More Output per Worker than Others”? *Quarterly Journal of Economics,* 114, 83-116.

Hausman, Jerry A. (1984), “Specification and Estimation of Simultaneous Equations Models,” Ch.7 in *Handbook of Econometrics*, Amsterdam: North-Holland.

Hausman, Jerry A. and Jinyong Hahn (2002), “A New Specification Test for the Validity of Instrumental Variables,” *Econometrica,* 70, 163-189.

Kane, Thomas J. and Cecelia E. Rouse (1995), “Labor-Market Returns to Two- and Four-Year Colleges,” *American Economic Review*, 85, 600-614.

Mauro, Paolo (1995), “Corruption and Growth,” *Quarterly Journal of Economics,* 110, 681-712.

Meghir, Costas and Marten Palme (1999), “Assessing the Effect of Schooling on Earnings Using a Social Experiment,” Unpublished Working Paper, University College London.

Phillips, Peter C.B. (1984), “Exact Small Sample Theory in the Simultaneous Equations Model,” Ch8 in *Handbook of Econometrics*, Amsterdam: North-Holland.

Staiger, Douglas and James H. Stock (1997), “Instrumental Variables Regression with Weak Instruments,” *Econometrica,* 65, 557-586.

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, 518-529.

Woolridge, Jeffey M. (2002), *Econometric Analysis of Cross Section and Panel Data*. London and Cambridge: MIT Press.

Wu, Chien-Fu Jeff (1986), "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis" *The Annals of Statistics*, 14, 1261-1295.

Wu, Chien-Fu Jeff (1990), "On the Asymptotics Properties of the Jackknife Histogram," *The Annals of Statistics*, 18, 1438-1452.

Table I: Test-statistics
Sample Size = 100, and 1,000 simulations
Truth is that Institutions Do Not Matter

Test-statistic	Nominal 5% Critical Values	Cov $Z_i{}'u_i$	Actual rejection rate	Actual rejection rate (RHS)	Actual rejection rate (LHS)
t-statistic	±1.95	0.06	9.8%	9.4%	0.4%
t-statistic	±1.95	-0.06	7.9%	0.6%	7.2%
AR-test	3.85	±0.06	9.4%	n.a.	n.a.
t-statistic	±1.95	0.10	19.4%	19.2%	0.2%
t-statistic	±1.95	-0.10	14.3%	0.3%	14.0%
AR-test	3.85	±0.10	17.7%	n.a.	n.a.

Table II: Sizes of Anderson-Rubin test

	$CovZ_i'u_i = 0$	$CovZ_i'u_i = 0.10$	$CovZ_i'u_i = 0.15$
Π=1 (strong instrument)			
Rate of false rejection	9.7	21.8	33.5
Π=0.1 (weak instrument)			
Rate of false rejection	10.1	22.6	34.4

Note: The data generating process of the simulation is based on (6). The sample size is N=64 and the nominal size is 10%. The Anderson-Rubin is defined in (7).

Table III:
Size properties of the ddj Anderson Rubin test

Part A: Π=1 (strong instrument)

Rate of false rejection of the null

Block size	$Cov\,Z_i^{'}u_i = 0.10$	$Cov\,Z_i^{'}u_i = 0.15$
16	7.4	11.8
18	6.0	13.5
20	5.3	10.4
22	4.0	9.5
24	3.3	8.4
26	2.5	6.0
28	1.9	3.7
30	1.5	3.2
32	1.3	2.3
Part B: Π=0.1 (weak instrument)		
16	7.2	13.0
18	6.7	9.4
20	3.8	10.7
22	3.1	7.0
24	3.4	7.0
26	3.2	5.2
28	1.7	3.6
30	0.8	3.3
32	1.2	2.6

Note: The data generating process of the simulation is based on (6). The sample size is n=64 and the nominal size is 10%. The parameter b represents block size and b=64-d, where d = the deleted observations. We compute the delete-d jackknife Anderson-Rubin test defined in (11) with b = {16, 18, 20, 22, 24, 26, 28, 30, 32}.

Table IV: Baseline regressions				
	(1)	(2)	(3)	(4)
Table IVa: Two-Stage Least Squares				
Average protection against expropriation risk 1985-1995	0.94	1.00	0.98	1.10
	(0.16)	(0.22)	(0.30)	(0.46)
Latitude		−0.65		−1.20
		(1.34)		(1.8)
Asia dummy			−0.92	−1.10
			(0.40)	(0.52)
Africa dummy			−0.46	-0.44
			(0.36)	(0.42)
“Other” continent dummy			−0.94	−0.99
			(0.85)	(1.0)
Table IVb: test-statistics for significance of expropriation risk				
t-statistic and p-values				
	6.03	4.49	3.28	2.39
	[< 0.000]	[< 0.000]	[0.001]	[0.017]
Full sample AR-statistic, full sample and delete-d jackknife p-values				
AR(β_0)	56.602	36.838	20.321	14.492
b = 16	[0.012]	[0.028]	[0.082]	[0.094]
b = 24	[0.029]	[0.054]	[0.102]	[0.134]
Full sample	[<0.000]	<0.000]	[<0.000]	[0.006]

Notes: Tables 4-6 were generated using STATA 9. In tables 4-6 the dependent variable is log GDP per capita in 1995. The numbers in parentheses are standard errors of coefficient estimators. The numbers in brackets in panel B in tables 4-7 are these p-values for the test-statistics. We use b=16 (1/4th the sample size) and b=24 (3/8th’s the sample size) to compute the delete-d jackknife Anderson-Rubin test, where b=16 has the best small sample properties and b=24 is very conservative. The results in this table are based on AJR (2001), p1386. AR (β_0) is calculated from the full sample. “Full Sample” shows the p-value when the AR(β_0) and chi-square (Theorem 2) critical values are used.

Table V: Controls for Colonial and Legal Origin				
	(1)	(2)	(3)	(4)
Table Va: Two-Stage Least Squares				
Average protection against expropriation risk 1985-1995	1.08	1.16	1.08	1.18
	(0.22)	(0.34)	(0.19)	(0.29)
Latitude		-0.75		-1.12
		(1.70)		(1.56)
British colonial dummy	-0.78	-0.80		
	(0.35)	(0.39)		
French colonial dummy	-0.12	-0.06		
	(0.35)	(0.42)		
French legal origin dummy			0.89	-0.96
			(0.32)	(0.39)
Table Vb: test-statistics for significance of expropriation risk				
t-statistic and p-values				
	4.95	3.43	5.65	4.06
	[<0.000]	[<0.000]	[<0.000]	[<0.000]
Full sample AR-statistic, full sample and delete-d jackknife p-values				
AR(β_0)	46.302	27.466	56.702	37.349
b = 16	[0.022]	[0.051]	[0.015]	[0.034]
b = 24	[0.029]	[0.060]	[0.028]	[0.044]
Full sample	[<0.000]	[<0.000]	[<0.000]	[<0.000]

Notes: Results are based on AJR [2001], p1389.

Table VI: Controls for Contemporary
Health Environment

	(1)	(2)	(3)	(4)	(5)	(6)
Table VIa: Two-Stage Least Squares						
Average protection against expropriation risk 1985-1995	0. 69	0. 72	0. 63	0.72	0. 55	0. 56
	(0.25)	(0.30)	(0.28)	(0.29)	(0.24)	(0.31)
Latitude		– 0.57		-0.56		-0.10
		(1.04)		(1.04)		(0.95)
Malaria in 1994	-0.58	– 0.60				
	(0.47)	(0.47)				
Life expectancy in 1995			0. 03	-0.60		
			(0.02)	(0.47)		
Infant mortality in 1995					– 0.01	– 0.01
					(0.005)	(0.006)
Table VIb: test-statistics for significance of expropriation risk						
t-statistic and p-values						
	2.73	2.43	2.28	2.43	2.30	1.79
	[0.008]	[0.018]	[0.026]	[0.018]	[0.025]	[0.079]
Full sample AR-statistic, full sample and delete-d jackknife p-values						
AR(β_0)	8.364	7.290	7.003	7.290	5.513	3.593
b = 16	[0.130]	[0.147]	[0.152]	[0.147]	[0.161]	[0.202]
b = 24	[0.193]	[0.197]	[0.182]	[0.197]	[0.201]	[0.230]
Full sample	[0.004]	[0.009]	[0.009]	[0.009]	[0.017]	[0.061]

Notes: Results are based on AJR [2001], p1392.

[1] Instruments that marginally satisfy this requirement are denoted weak and are the subject of a large and growing literature [see Staiger and Stock, 1997; Stock et al. 2000]. This paper focuses primarily on strong instruments that satisfy the relevance criteria. Weak instruments are briefly discussed in section 5.

[2] We just consider one kind of institution and, hence, one endogenous variable for expositional simplicity. Our method also works for multiple endogenous variables. See Acemoglu and Johnson (2006) for an analysis of how instrumental variables can be used to identify how two endogenous institutions, property rights (measured by a survey of risk of expropriation) and efficiency of contracts (measured by an index of legal formalism), can affect long-run growth.

[3] By the Frisch-Waugh-Lovell Theorem, we can always project out these covariates and obtain the system in equations (1) and (2) (see Davidson and McKinnnon, 1993, p.19).

[4] We can generalize this test-statistic to allow for multiple endogenous explanatory variables and as least as many instruments.

[5] The simulation results reported in this section are robust to different values of endogeneity.

Figure 1: The t-test with an Almost Exogenous Instrument (Positively correlated)

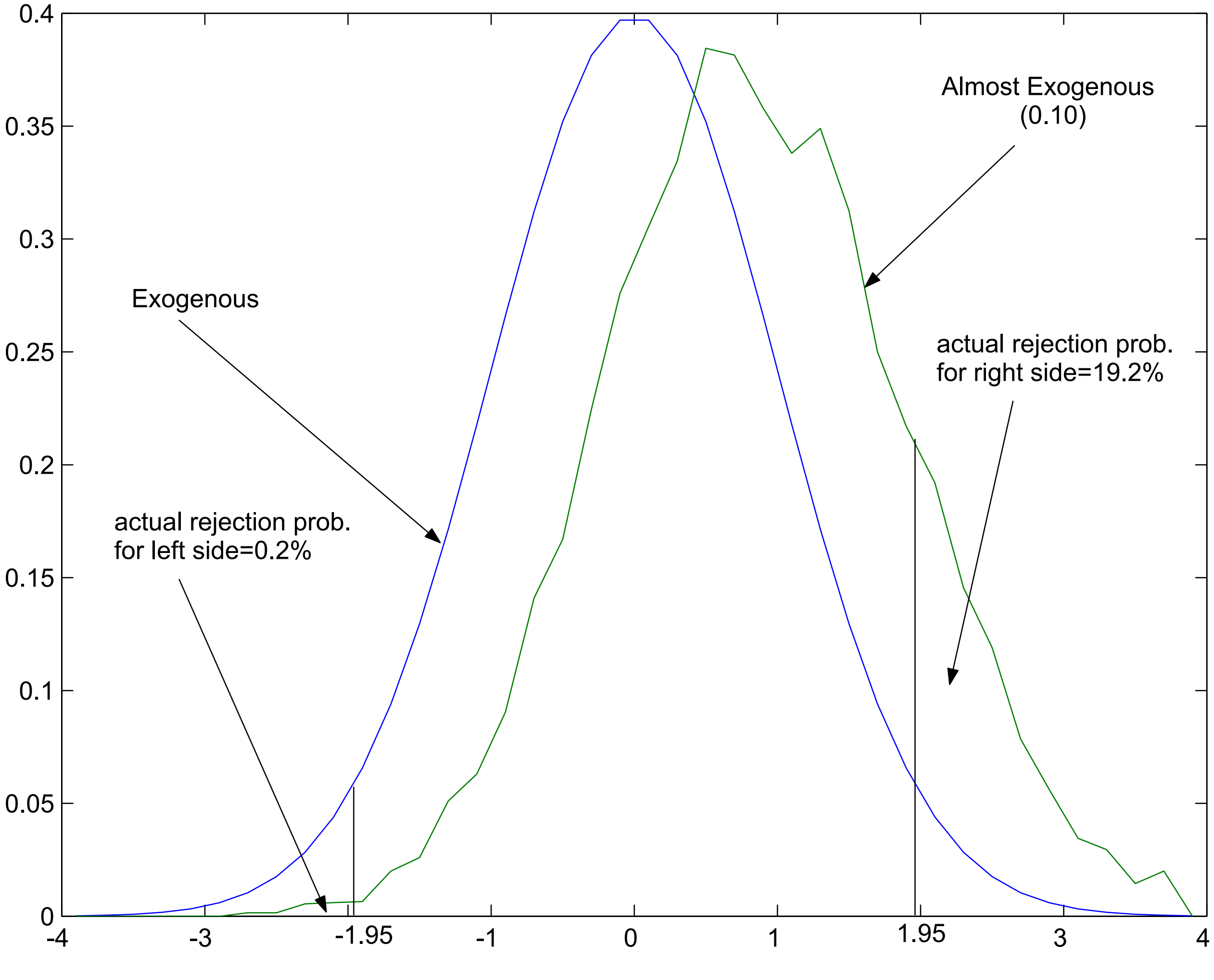


Figure 2: The t-test with an Almost Exogenous Instrument (Negatively correlated)

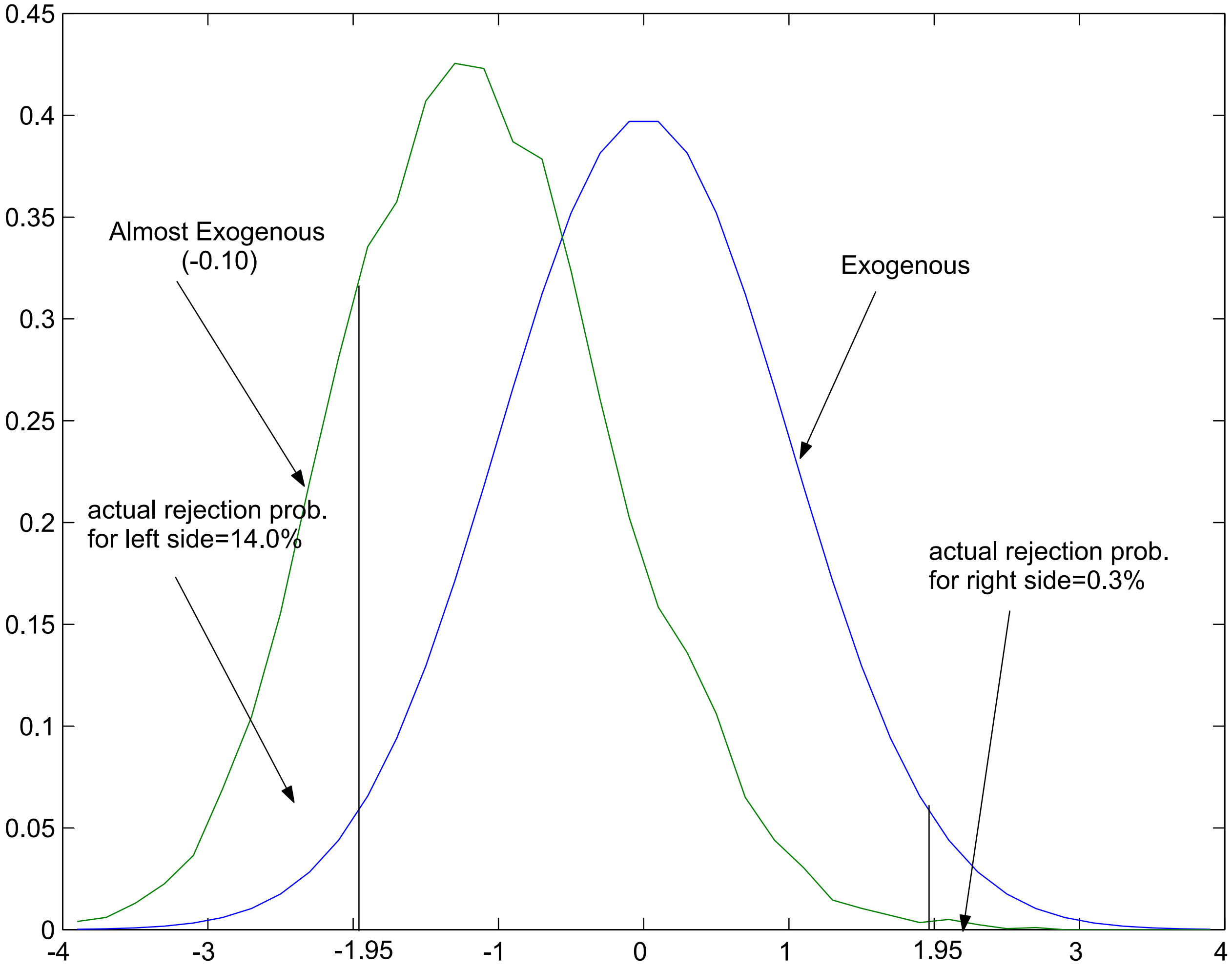
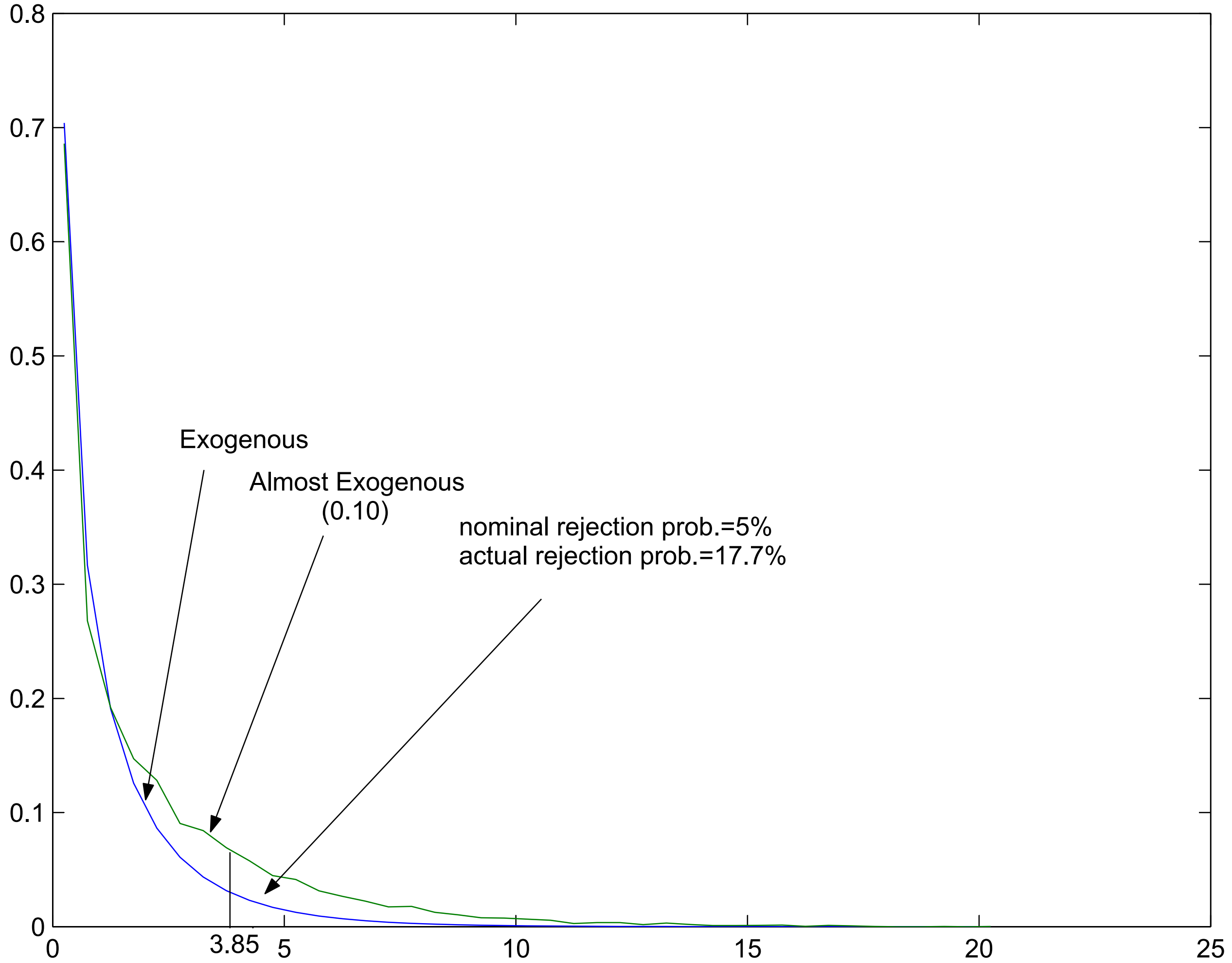


Figure 3: The Anderson-Rubin test with an Almost Exogenous Instrument (0.10)


Exogenous
Almost Exogenous
(0.10)
nominal rejection prob.=5%
actual rejection prob.=17.7%