Are Nearly “Exogenous Instruments” Reliable?

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

October 3, 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 or whether or not education affects job market performance. This paper is designed to help economists who typically pick instruments that are not perfectly exogenous. Our new theoretical result is that when instruments only slightly violate the assumption that structural errors are orthogonal to instruments, the widely used t-statistic substantially and unpredictably either over-rejects or under-rejects the null even when instruments are strong. We establish this result in the limit and in small sample simulations. Furthermore, another new result is that re-sampling methods cannot repair the t-statistic. We show that we can improve inference by employing an Anderson-Rubin test statistic derived from the delete-d jackknife procedure developed by Wu (1986). Our procedure adjusts the critical values according to the correlation between the instrument and structural error. We use this test to confirm and to correct inferences about the impact of institutions on long term in the celebrated work of Acemoglu, Johnson and Robinson (2001).

Keywords: exogeneity, jackknife, institutions.
Daniel Berkowitz,* Mehmet Caner** and Ying Fang***

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

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

***Department of Economics, Xiamen University, China, yifst1@gmail.com.

We have benefited from discussions with Daron Acemoglu, Karen Clay, and Nese Yildiz.
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 quality of contemporary institutions that restrain the government from expropriating private assets. The a priori argument, roughly speaking, is 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.\footnote{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.}

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. However, Wooldridge (2002, p.88) argues that this is not necessarily true: because civilian employers are more likely to invest in job training for employees who have high draft numbers, they 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 good and bad news for applied economists who use instruments. The bad news is that the standard t-test statistic is unreliable: even when the instrument is very close to being exogenous, the t-test grossly and unpredictably over-rejects or under-rejects the null. Furthermore, re-sampling methods including the bootstrap, the jackknife
and subsampling cannot salvage the t-statistic. We prove these results in the limit and in small samples. And, to our knowledge, these are new theoretical results.

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). Furthermore, in this paper we supply simple STATA code that enables applied researchers to implement this procedure. Even though none of the resampling methods are consistent for the Anderson-Rubin test, the delete-d jackknife method comes arbitrarily close to true distribution in large samples. We also show that this method works well in small samples and is better than any method used until so far in terms of size properties. Thus, 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 over-identified systems. It is well known, however, that the 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. Our test depends on the choice of block size and we talk about these issues in the paper. However, regardless of the block size choice delete d jackknife Anderson-Rubin test has better size than regular Anderson-Rubin and t-tests.
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 standard AR test and t-tests. 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 which can be obtained from the authors. In section 7 we conclude.

2. 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 that bad news that 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 not an institution, say property rights enforcement, influences long term growth in a sample of countries.\(^2\) If we suspect that

\(^2\) 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
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; Phillips, 1984):

\[
  LGR = \beta_0 + \beta_1 INST + u \tag{1}
\]

\[
  INST = \Pi_0 + Z \Pi_1 + V \tag{2}
\]

Equation (1) is the structural equation, where LGR 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 have zero means 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 \( \beta_0, \beta_1, \Pi_0 \) and \( \Pi_1 \), 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 \( \beta_1 \) 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 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 McKinnmon, 1993, p.19).
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:

$$Exogenous \Rightarrow \text{Cov} Z_i' u_i = 0$$

(3)

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:

$$Nearly \ \text{Exogenous} \Rightarrow \text{Cov} Z_i' u_i = C / \sqrt{n} \ is \ small$$

(4)

where $C$ is an $nx1$ vector of constants that is contained in compact set.

If we choose $\text{Cov} Z_i' u_i = C$ 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 $\hat{\beta}_{1,\text{TLS}}$ as the TSLS estimator of $\beta_1$, we use the t-statistic to test

$$H_0 : \beta_1 = 0 \text{, against}$$

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

$$t = \frac{\hat{\beta}_{1,\text{TLS}}}{\sqrt{\text{var} \hat{\beta}_{1,\text{TLS}}}}$$

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 $N(0, \Lambda)$ and

$$\Lambda = \begin{pmatrix} 1 & \text{Cov} Z_i u_i & 0 \\ \text{Cov} Z_i u_i & 1 & \text{Cov} V_i u_i \\ 0 & \text{Cov} V_i u_i & 1 \end{pmatrix}. \quad (6)$$
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 iid data \((Z,u,V)\) are generated, we can derive the observation of \( \text{INST} \) and LRGr by using equations (1) and (2) and specified true values of \( \beta_1 \) and \( \Pi_1 \). Based on the information of \((\text{LRGr}, \text{INST}, Z)\), we compute the t-statistic and then test whether the null of \( \beta_1 = 0 \) 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 then nearly exogenous with small positive correlation: Cov \( Z_i' u_i = 0.10 \). The distribution under exogeneity is close to standard normal distribution, and the distribution under near exogeneity shifts to right and is close to a normal distribution with non-zero 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 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 over-rejects 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:

\[
AR(\beta_i = 0) = \frac{\text{LRGr'Pz LRGr} / (\text{LRGr'Mz LRGr})}{(n - 2)}
\]

(7)

Here, \(AR(\beta_i = 0)\) 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 standard chi-square; and, when \(\text{Cov } Z_i'u_i = 0.10\) the distribution shifts to the right and is close to 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 17.47 and near exogeneity creates small sample problems.

---

4 We can generalize this test statistic to allow for multiple endogenous explanatory variables and as least as many instruments.
Table 1 illustrates that the small sample problems of 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.

3. 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 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 \( m \geq 1 \) endogenous explanatory variables, and \( k \geq m \) instruments:

\[
y = Y\beta + u \tag{1^*}
\]

\[
Y = Z\Pi + V \tag{2^*}
\]

where \( y \) and \( Y \) are respectively and \( nx1 \) vector and \( nxm \) 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, and 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 non-zero mean. This theorem is not derived or implied in Caner (2006).
Theorem 1: Suppose that the instrument is nearly exogenous according to (4), and the standard assumption 2 in the appendix holds. Then,

\[ t \xrightarrow{d} N[\sigma_u^{-1} (\Pi' Q_{zz} \Pi)^{-1/2} \Pi' C, 1] \]  

(8)

where \( \sigma_u \) is the square root of \( \sigma_u^2 \), 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:

\[ AR(\beta_0)= (y - Y\beta_0)' P_z (y - Y\beta_0) / (y - Y\beta_0)' M_z (y - Y\beta_0) / (n-k-m) \]  

(7*)

We use this statistic to test \( H_0 : \beta = \beta_0 \) against \( H_1 : \beta \neq \beta_0 \) where \( \beta_0 \) 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

\[
AR(\beta_0) \xrightarrow{d} \chi^2_k (\varsigma)
\]  

(9)

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 \( \Omega = \sigma_u^2 \otimes Q_z \).


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 \( C \neq 0 \) the distribution shifts to the right. Again, since we do not know \( C \), we cannot use these theorems to obtain appropriate critical values. The convergence is uniform.

4. Reliable Inference under Near Exogeneity

This section contains the good news that it possible to make reliable inferences when instruments are nearly exogenous. When large sample results are problematic, a standard remedy is to employ re-sampling 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 re-sampling methods is that it contains the TSLS estimator \( \hat{\beta}_{1,TLS} \) (see equation (5)). Re-sampling 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, re-sampling procedures are forced essentially to ignore correlations between instruments and structural error term.

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

\[
t_s = \frac{\hat{\beta}_{15,TLS} - \hat{\beta}_{1,TLS}}{\sqrt{a \text{ var } \hat{\beta}_{15,TLS}}}
\]  

(10)

where \( \hat{\beta}_{15,TLS} \) is the ddj estimator, \( a \text{ var } \hat{\beta}_{15,TLS} \) is its estimated variance and \( \hat{\beta}_{1,TLS} \) 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: \( d = \gamma n \), 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, \( \hat{\beta}_{15,TLS} \) and \( a \text{ var } \hat{\beta}_{15,TLS} \), 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 \( d = \gamma n \), \( 0 < \gamma < 1 \), and \( n \to \infty \). This result is not covered in Caner (2006). To our knowledge, this is a new theoretical result.

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

\[
t^{d} \rightarrow N[0, 2 - \gamma - 2 \sqrt{1 - \gamma}]
\]

(11)

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: \( t^{d} \rightarrow N[\sigma^{-1}_u (\Pi'Q_{zz} \Pi)^{-1/2} \Pi'C, 1] \); the mean is not zero and the variance is not one. Thus, the delete-d jackknife procedure fails to correct for \( C \neq 0 \). Because we cannot estimate the sign or size \( C \), 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 non-central chi-distribution that emerges under near exogeneity. Let \( y_b \), \( Y_b \) and \( Z_b \) denote, respectively, sub-vectors or sub-matrices 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 \( b \times m \) matrix and \( Z_b \) is a \( b \times k \) matrix, and the AR test statistic for any block is denoted \( AR_S(\beta_0) \):

\[
AR_S(\beta_0) = (y_b - Y_b \beta_0)'P_z (y_b - Y_b \beta_0)/(y_b - Y_b \beta_0)'M_z (y_b - Y_b \beta_0)/(b - k - m) \tag{11}
\]

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 \( \beta_0 \) rather than the estimator of \( \beta \). 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 is unable to replicate these correlations. These results are established in Caner (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

\[
AR_S(\beta_0) \xrightarrow{\text{d}} \chi_k^2(\xi) \tag{12}
\]
where $\chi_k^2 (\tilde{\varsigma})$ is a non-central chi-square distribution with $k$ degrees of freedom and $\tilde{\varsigma}$ is the non-centrality parameter: $\tilde{\varsigma} = (b/n) C^* \Omega^{-1} C$, and $\Omega = \sigma_u^2 \otimes Q_{\pi}$.


Theorem 4 shows the delete-d jackknife procedure generates a large sample chi-square distribution with non-centrality parameter that is equal b/n times the non-centrality 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. Convergence in Theorem 4 is uniform.

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 size greater than three-quarters 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 employed in AJR (2001) with 64 countries and a block of size of 3/8th's is useful as a conservative robustness check. More extensive simulation studies, however, are required for the choice of block size.

5. 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. 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 $\beta$ is $\beta_0 = 0$.

We set the sample size equal to 64 in order to conduct comparisons of 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 $\Pi$ (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 2 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 $\Pi$ (see equations (7) and (7*)).

Table 3 reports rates small sample properties of the ddj AR-test for block sizes covering $1/4$th to $1/2$ 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

---

$^5$ The simulation results reported in this section are robust to different values of endogeneity.
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 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 \( \text{cov } Z, u_i = 0.10 \) \text{ and } 0.15 \text{ and the small sample properties are quite}
good. When \( b = 24 \) covering 3/8ths the sample, the false rejection rate is 4.0% and 9.5%
when \( \text{cov } Z, u_i = 0.10 \) \text{ and } 0.15 \text{, and the test is more conservative.}

By comparing Tables 2 and 3 we see that regardless of the choice of block size,
the ddj AR test is less size distorted than 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 3/8ths the sample as
a conservative robustness check. If the results are similar, then the inferences are reliable.

### 6. 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.:

\[
LRGr = \beta_0 + \beta_1 \text{INST} + \beta_2 X + u
\]  

(1)
\[ \text{INST} = \Pi_0 + Z \Pi_1 + X \Pi_2 + V \] (2)

We test the null \( H_0 : \beta_1 = 0 \) against the alternative \( H_1 : \beta_1 \neq 0 \). Tables 4-6 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 (1/4\textsuperscript{th} the sample size), 24 (3/8\textsuperscript{th}’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 will 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 4a replicates and then checks inferences made in the baseline regressions in AJR (2001), Table 4. 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 the Asia, Africa and “other” continent dummy variables and 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 5-6 we check for the significance of institutions with additional controls (see AJR (2001) Tables 5 and 7). The ddj AR-test statistics confirm that institutions
matter. In Table 5, 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 10% level.

Table 6 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. However, the more reliable ddj AR-test with b =16 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 6 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 3, 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 6 are relatively large compared to those in Tables 4-5, hence leading to non-rejection of the null in Table 6. However the level of near exogeneity is not enough to overturn most of the AJR (2001) findings.

7. 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 in that it enables practitioners to validate near exogeneity in exactly identified as well as over-identified 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 $E[Z'u] = C / \sqrt{N}$, 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) $(u'u / N, V'u / N, V'V / N) \xrightarrow{p} (\sigma^2_u, \Sigma_{yu}, \Sigma_{v'y})$, where $\sigma^2_u$, $\Sigma_{yu}$ and $\Sigma_{v'y}$ are respectively a $1 \times 1$ scalar, an $m \times 1$ vector and an $m \times m$ matrix.

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

(c) $(Z'u / \sqrt{N}, Z'V / \sqrt{N}) \xrightarrow{p} (\Psi_{zu}, \Psi_{z'y})$, and

$$
\begin{pmatrix}
\Psi_{zu} \\
\Psi_{z'y}
\end{pmatrix} \xrightarrow{N} \begin{pmatrix}
C \\
0
\end{pmatrix} \Sigma \otimes Q
$$

where $\Sigma = \begin{pmatrix}
\sigma^2_u & \Sigma'_{yu} \\
\Sigma_{yu} & \Sigma_{v'y}
\end{pmatrix}$.

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 \( \Psi_{zu} \) 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 \( E[Z_i' u_i]^{2+\delta} < \Delta < \infty \) 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) \( E[z_{b,i} u_b]^{2+\delta} < \Delta_1 < \infty \) for all \( b < N \) and all \( 1 \leq i \leq K \)

(b) \( E[z_{b,i} z_{b,j}]^{1+\delta} < \Delta_2 < \infty \) for all \( b < N \) and all \( 1 \leq i, j \leq K \)

(c) \( E[u_b^2]^{1+\delta} < \Delta_3 < \infty \) for all \( b < N \)

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

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

Lemma 1. If Assumptions 1 and 2 hold for the model defined by (1*) and (2*), then the TSLS estimator \( \hat{\beta}_{TLS} \) is consistent and
\[ \sqrt{N}(\hat{\beta}_{\text{TLS}} - \beta_0)^d \rightarrow N((\Sigma'Q_{\text{ZZ}}\Sigma)^{-1}\Sigma'C, \sigma_a^2(\Sigma'Q_{\text{ZZ}}\Sigma)^{-1}) \]

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

Proof of Lemma 1: We know that

\[ \hat{\beta}_{\text{TLS}} = (Y'P_{\text{Z}}Y)^{-1}(Y'P_{\text{Z}}Y). \]

So we have

\[ \sqrt{N}(\hat{\beta}_{\text{TLS}} - \beta_0) \]
\[ = [\left(\frac{Y'Z}{N}\right)(\frac{Z'Z}{N})^{-1}\left(\frac{Z'Y}{N}\right)^{-1}] \left[\frac{Y'Z}{N}(\frac{Z'Z}{N})^{-1}(\frac{Z'u}{\sqrt{N}})\right] \]

By Assumption 2, we can obtain that

\[ [\frac{Y'Z}{N}(\frac{Z'Z}{N})^{-1}\left(\frac{Z'Y}{N}\right)^{-1}]^p \rightarrow (\Sigma'Q_{\text{ZZ}}\Sigma)^{-1} \]

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

\[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [Z_i'u_i - E(Z_i'u_i)] \overset{d}{\rightarrow} N[0, \sigma_a^2 Q_{\text{ZZ}}]. \]

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

\[ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} E(Z_i'u_i) \overset{p}{\rightarrow} C. \]

Combining the above results, we obtain

\[ \frac{Z'u}{\sqrt{N}} \overset{d}{\rightarrow} N[C, \sigma_a^2 Q_{\text{ZZ}}]. \]
Then the result in the lemma follows directly. \textit{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 (Donald W.K. Andrews, 2000).

Proof of Theorem 1: The result in the theorem directly follows from Lemma 1. \textit{Q.E.D.}

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

\[
t_s = \frac{\hat{\beta}_{S,TLS} - \hat{\beta}_{TSL}}{\sqrt{a \operatorname{var}(\hat{\beta}_{S,TLS})}}
\]

where

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

and

\[
\hat{\sigma}^2_{u,b} = (y_b - Y_b\hat{\beta}_{S,TLS})(y_b - Y_b\hat{\beta}_{S,TLS})/(b-K-m)
\]

By Assumption 3 and weak law of large number, we have

\[
\hat{\sigma}^2_{u,b} \to \sigma^2_u \text{ in probability},
\]
and
\[
\left[ \left( \frac{Y_b' Z_b}{b} \right) \left( \frac{Z_b' Z_b}{b} \right)^{-1} \left( \frac{Z_b' Y_b}{b} \right) \right]^{-1} \rightarrow (\Pi' Q_{ZZ} \Pi)^{-1}.
\]

The $t_S$-statistic can be rewritten as
\[
t_S = \frac{(\hat{\beta}_{S,TLS} - \beta_0) - (\hat{\beta}_{TLS} - \beta_0)}{\sqrt{a \text{ var}(\hat{\beta}_{S,TLS})}}
\]

Consider the first term in the above equation,
\[
\sqrt{b}(\hat{\beta}_{S,TLS} - \beta_0) = \left[ \left( \frac{Y_b' Z_b}{b} \right) \left( \frac{Z_b' Z_b}{b} \right)^{-1} \left( \frac{Z_b' Y_b}{b} \right) \right]^{-1} \left[ \left( \frac{Y_b' Z_b}{b} \right) \left( \frac{Z_b' Z_b}{b} \right)^{-1} \left( \frac{Z_b' u_b}{b} \right) \right]
\]

By Assumption 3 and the triangular array central limit theorem, we can obtain
\[
\frac{Z_b' u_b}{\sqrt{b}} = \frac{1}{\sqrt{b}} \sum_{i=1}^{b} [Z_{b,i} u_{b,i} - E(Z_{b,i} u_{b,i})] + \frac{1}{\sqrt{b}} \sum_{i=1}^{b} E(Z_{b,i} u_{b,i})
\]
\[
 \rightarrow N[0, \sigma_u^2 Q_{ZZ}] = N[(\sqrt{1-\gamma})C, \sigma_u^2 Q_{ZZ}]
\]

So we have
\[
\sqrt{b}(\hat{\beta}_{S,TLS} - \beta_0) \rightarrow N[\delta_C, I]
\]

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

28
\[
\frac{\sqrt{b}(\beta_{\text{TLS}} - \beta_0)}{\sqrt{\sigma_u^2 (\Pi'Q_{zz}\Pi)^{-1}}} \overset{d}{\rightarrow} N[\delta_C, (1 - \gamma)]
\]

Then the result in the theorem follows from above. \(Q.E.D.\)
References

rigins of Comparative Development: An Empirical Investigation,” American Economic
Review, 91, 1369-1401.

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

the Boundary of the Parameter Space,” Econometrica, 68, 399-405.

Evidence from Social Security Administrative Records,” American Economic Review, 80,
313-336.


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.

Empirical Likelihood Estimators: Fixed and Many Moment Asymptotics”, working


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

<table>
<thead>
<tr>
<th>Test-statistic</th>
<th>Nominal 5% Critical Values</th>
<th>Cov Z_i'\hat{u}_i</th>
<th>Actual rejection rate</th>
<th>Actual rejection rate (RHS)</th>
<th>Actual rejection rate (LHS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-statistic</td>
<td>±1.95</td>
<td>0.06</td>
<td>9.8%</td>
<td>9.4%</td>
<td>0.4%</td>
</tr>
<tr>
<td>t-statistic</td>
<td>±1.95</td>
<td>-0.06</td>
<td>7.9%</td>
<td>0.6%</td>
<td>7.2%</td>
</tr>
<tr>
<td>AR test</td>
<td>3.85 ±0.06</td>
<td></td>
<td>9.4%</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
<tr>
<td>t-statistic</td>
<td>±1.95</td>
<td>0.10</td>
<td>19.4%</td>
<td>19.2%</td>
<td>0.2%</td>
</tr>
<tr>
<td>t-statistic</td>
<td>±1.95</td>
<td>-0.10</td>
<td>14.3%</td>
<td>0.3%</td>
<td>14.0%</td>
</tr>
<tr>
<td>AR test</td>
<td>3.85 ±0.10</td>
<td></td>
<td>17.7%</td>
<td>n.a.</td>
<td>n.a.</td>
</tr>
</tbody>
</table>

Table 2: Sizes of Anderson-Rubin test

<table>
<thead>
<tr>
<th>CovZ_i'\hat{u}_i = 0</th>
<th>CovZ_i'\hat{u}_i = 0.10</th>
<th>CovZ_i'\hat{u}_i = 0.15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Π=1 (strong instrument)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate of false rejection</td>
<td>9.7</td>
<td>21.8</td>
</tr>
<tr>
<td>Π=0.1 (weak instrument)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate of false rejection</td>
<td>10.1</td>
<td>22.6</td>
</tr>
</tbody>
</table>

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 3: Size properties of the ddj Anderson Rubin test

<table>
<thead>
<tr>
<th>Block size</th>
<th>$Cov Z_i u_i = 0.10$</th>
<th>$Cov Z_i u_i = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>7.4</td>
<td>11.8</td>
</tr>
<tr>
<td>18</td>
<td>6.0</td>
<td>13.5</td>
</tr>
<tr>
<td>20</td>
<td>5.3</td>
<td>10.4</td>
</tr>
<tr>
<td>22</td>
<td>4.0</td>
<td>9.5</td>
</tr>
<tr>
<td>24</td>
<td>3.3</td>
<td>8.4</td>
</tr>
<tr>
<td>26</td>
<td>2.5</td>
<td>6.0</td>
</tr>
<tr>
<td>28</td>
<td>1.9</td>
<td>3.7</td>
</tr>
<tr>
<td>30</td>
<td>1.5</td>
<td>3.2</td>
</tr>
<tr>
<td>32</td>
<td>1.3</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Part B: $\Pi=0.1$ (weak instrument)

<table>
<thead>
<tr>
<th>Block size</th>
<th>$Cov Z_i u_i = 0.10$</th>
<th>$Cov Z_i u_i = 0.15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>7.2</td>
<td>13.0</td>
</tr>
<tr>
<td>18</td>
<td>6.7</td>
<td>9.4</td>
</tr>
<tr>
<td>20</td>
<td>3.8</td>
<td>10.7</td>
</tr>
<tr>
<td>22</td>
<td>3.1</td>
<td>7.0</td>
</tr>
<tr>
<td>24</td>
<td>3.4</td>
<td>7.0</td>
</tr>
<tr>
<td>26</td>
<td>3.2</td>
<td>5.2</td>
</tr>
<tr>
<td>28</td>
<td>1.7</td>
<td>3.6</td>
</tr>
<tr>
<td>30</td>
<td>0.8</td>
<td>3.3</td>
</tr>
<tr>
<td>32</td>
<td>1.2</td>
<td>2.6</td>
</tr>
</tbody>
</table>

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 4: Baseline regressions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average protection against expropriation risk 1985-1995</td>
<td>0.94</td>
<td>1.00</td>
<td>0.98</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>(0.16)</td>
<td>(0.22)</td>
<td>(0.30)</td>
<td>(0.46)</td>
</tr>
<tr>
<td>Latitude</td>
<td></td>
<td></td>
<td>−0.65</td>
<td>−1.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.34)</td>
<td>(1.8)</td>
</tr>
<tr>
<td>Asia dummy</td>
<td></td>
<td></td>
<td></td>
<td>−0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.40)</td>
</tr>
<tr>
<td>Africa dummy</td>
<td></td>
<td></td>
<td></td>
<td>−0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.36)</td>
</tr>
<tr>
<td>“Other” continent dummy</td>
<td></td>
<td></td>
<td></td>
<td>−0.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.85)</td>
</tr>
</tbody>
</table>

Table 4a: Two-Stage Least Squares

Table 4b: Test statistics for significance of expropriation risk

<table>
<thead>
<tr>
<th>t-statistic and p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.03</td>
</tr>
<tr>
<td>[&lt; 0.000]</td>
</tr>
</tbody>
</table>

Full sample AR-statistic, full sample and delete-d jackknife p-values

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR($\beta_0$)</td>
<td>56.602</td>
<td>36.838</td>
<td>20.321</td>
<td>14.492</td>
</tr>
<tr>
<td>b = 16</td>
<td>[0.012]</td>
<td>[0.028]</td>
<td>[0.082]</td>
<td>[0.094]</td>
</tr>
<tr>
<td>b = 24</td>
<td>[0.029]</td>
<td>[0.054]</td>
<td>[0.102]</td>
<td>[0.134]</td>
</tr>
<tr>
<td>Full sample</td>
<td>[&lt;0.000]</td>
<td>&lt;0.000</td>
<td>[&lt;0.000]</td>
<td>[0.006]</td>
</tr>
</tbody>
</table>

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($\beta_0$) is calculated from the full sample. “Full Sample” shows the p-value when the AR($\beta_0$) and chi-square (Theorem 2) critical values are used.
### Table 5: Controls for Colonial and Legal Origin

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 5a: Two-Stage Least Squares</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average protection against expropriation risk 1985-1995</td>
<td>1.08</td>
<td>1.16</td>
<td>1.08</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.34)</td>
<td>(0.19)</td>
<td>(0.29)</td>
</tr>
<tr>
<td>Latitude</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.75</td>
<td>-1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.70)</td>
<td>(1.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>British colonial dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.78</td>
<td>-0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td>(0.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French colonial dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.12</td>
<td>-0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.35)</td>
<td>(0.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>French legal origin dummy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.89</td>
<td>-0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.39)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5b: Test statistics for significance of expropriation risk

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>t-statistic and p-values</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4.95</td>
<td>3.43</td>
<td>5.65</td>
<td>4.06</td>
</tr>
<tr>
<td></td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
</tr>
<tr>
<td><strong>Full sample AR-statistic, full sample and delete-d jackknife p-values</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(β₀)</td>
<td>46.302</td>
<td>27.466</td>
<td>56.702</td>
<td>37.349</td>
</tr>
<tr>
<td></td>
<td>[0.022]</td>
<td>[0.051]</td>
<td>[0.015]</td>
<td>[0.034]</td>
</tr>
<tr>
<td>b = 24</td>
<td>[0.029]</td>
<td>[0.060]</td>
<td>[0.028]</td>
<td>[0.044]</td>
</tr>
<tr>
<td>Full sample</td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
<td>[&lt;0.000]</td>
</tr>
</tbody>
</table>

Notes: Results are based on AJR (2001), p1389.
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 6a: Two-Stage Least Squares</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average protection against expropriation risk 1985-1995</td>
<td>0.69</td>
<td>0.72</td>
<td>0.63</td>
<td>0.72</td>
<td>0.55</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.30)</td>
<td>(0.28)</td>
<td>(0.29)</td>
<td>(0.24)</td>
<td>(0.31)</td>
</tr>
<tr>
<td>Latitude</td>
<td>-0.57</td>
<td>-0.56</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.04)</td>
<td>(1.04)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malaria in 1994</td>
<td>-0.58</td>
<td>-0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.47)</td>
<td>(0.47)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Life expectancy in 1995</td>
<td></td>
<td></td>
<td>0.03</td>
<td>-0.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant mortality in 1995</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.006)</td>
</tr>
<tr>
<td><strong>Table 6b: test statistics for significance of expropriation risk</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-statistic and p-values</td>
<td>2.73</td>
<td>2.43</td>
<td>2.28</td>
<td>2.43</td>
<td>2.30</td>
<td>1.79</td>
</tr>
<tr>
<td></td>
<td>[0.008]</td>
<td>[0.018]</td>
<td>[0.026]</td>
<td>[0.018]</td>
<td>[0.025]</td>
<td>[0.079]</td>
</tr>
<tr>
<td>Full sample AR-statistic, full sample and delete-d jackknife p-values</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR(β₀)</td>
<td>8.364</td>
<td>7.290</td>
<td>7.003</td>
<td>7.290</td>
<td>5.513</td>
<td>3.593</td>
</tr>
<tr>
<td></td>
<td>[0.130]</td>
<td>[0.147]</td>
<td>[0.152]</td>
<td>[0.147]</td>
<td>[0.161]</td>
<td>[0.202]</td>
</tr>
<tr>
<td>b = 16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.193]</td>
<td>[0.197]</td>
<td>[0.182]</td>
<td>[0.197]</td>
<td>[0.201]</td>
<td>[0.230]</td>
</tr>
<tr>
<td>b = 24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.004]</td>
<td>[0.009]</td>
<td>[0.009]</td>
<td>[0.009]</td>
<td>[0.017]</td>
<td>[0.061]</td>
</tr>
<tr>
<td>Full sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Results are based on AJR (2001), p1392.
Figure 1: The t-test with an Almost Exogenous Instrument (Positively correlated)

- **Exogenous**
- **Almost Exogenous** (0.10)

- Actual rejection prob. for right side = 19.2%
- Actual rejection prob. for left side = 0.2%
Figure 2: The t-test with an Almost Exogenous Instrument (Negatively correlated)

Almost Exogenous 
(-0.10)

actual rejection prob. for left side=14.0%

Exogenous

actual rejection prob. for right side=0.3%
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%