# Institutions and Growth: Evidence from Estimation Methods Robust to Weak Instruments

Alexandre Dmitriev*

September 4, 2011

**Abstract**

This paper focuses on the empirical approach proposed by Hall and Jones (1999) to estimate the effect of what they call "social infrastructure" on productivity across countries. We consider the issue of weak identification in their linear instrumental variables model. The evidence obtained from partially robust estimators, such as the $k$-class and jackknife estimators, is interpreted on the basis of Monte Carlo studies. Our findings suggest that using certain $k$-class estimators allows exclusive reliance on the linguistic variables to instrument for institutional quality despite their low correlation with the endogenous regressor in question.

*JEL* Classification: C15; O40

*Keywords* Institutions; Growth; Weak Instruments; Robust Estimators

*School of Economics, Australian School of Business, The University of New South Wales, Sydney NSW 2052, Australia; Tel.: (+61) 2 9385 3351; email: `a.dmitriev@unsw.edu.au`

# 1 Introduction

This paper revisits the empirical methodology of Hall and Jones (1999) (henceforth HJ99), who estimate the contribution of institutions in accounting for income differences across countries.[1] To address the issues of endogeneity and reverse causality, HJ99 estimate the effect of social infrastructure on per capita income using two-stage least squares (TSLS). They choose instruments that reflect the extent of Western European influence, which is suggested to be one of the forces behind the adoption of favorable infrastructure. The instruments include the fraction of population speaking English at birth, the fraction of population speaking one of the five major European languages at birth, the distance from the equator (latitude) and Frankel and Romer (1999) geography-predicted trade intensity. This study focuses on how the institutional quality is instrumented for in the HJ99 linear instrumental variables (IV) model.

Linear IV estimation requires that the instruments are sufficiently correlated with the endogenous regressor and uncorrelated with the error term. These requirements are often referred to as instrument relevance and exogeneity. Simple diagnostic tests show that HJ99's instruments are weakly correlated with their measure for institutions.[2] As a result, HJ99's instrument relevance and therefore their TSLS estimation results might be questionable. This paper attempts to address the issue of week identification in HJ99 by answering two questions: (i) is the fraction of population speaking the major European languages at birth a relevant instrument for social infrastructure? (ii) do institutions have a positive causal effect on growth if they are instrumented with languages? To answer these question we investigate if the performance of TSLS can be improved upon by using estimators that are partially robust to weak instruments and what estimators are preferable in the context of the HJ99 model.

This paper contributes to the literature in three ways. First, it highlights that all of HJ99's instruments are weakly correlated with the included endogenous regressor. Hence, TSLS point estimates and hypothesis tests could be unreliable. Second, it suggests that instrumenting for institutional quality with linguistic instruments alone is feasible. However, to be able do

[1] Hall and Jones (1999) and Acemoglu, Johnson and Robinson (2001) are the major contributions to the literature that addresses the origins of observed cross-country income disparities.

[2] See Table 1.

so, one must carefully choose among the IV estimators robust to weak instruments. Based on Monte Carlo experiments that replicate certain characteristics of the data, this study advocates the use of the Fuller-k estimator. Finally, we show that the results og HJ99 are not driven by their reliance on geographic and trade instruments. Estimations that exclusively use linguistic instruments indicate on a positive causal effect of institutions on prosperity. This result addresses the critique directed towards HJ99 that they rely on latitude and trade intensity as instruments.

There are several reasons for focusing on linguistic instruments for institutions. First, several researchers express dissatisfaction with HJ99's latitude and Frankel and Romer (1999) trade intensity as instruments for institutional quality.[3] Second, instrumenting institutions with languages instead of Acemoglu *et al.* (2001) (henceforth AJR) settler mortality allows empirical researchers to use a longer sample. HJ99's sample covers 127 countries whereas AJR's covers only 63. Second, having an additional instrument for institutions could be useful for identification purposes in multivariate specifications that attempt to identify partial effects on income of geography, trade and institutions. Finally, the validity of AJR's instruments is not beyond doubt. For instance, Glaeser *et al.* (2004) argue that AJR's settler mortality rate fails to fulfill the exclusion criterion for an instrumental variable.

This work is related to two strands of the literature. The first aims to establish the causal relationship between per capita income and the fundamental determinants of growth. Contributions to this strand quantitatively assess the causal effect on income of institutions (Hall and Jones, 1999; Acemoglu *et al.*, 2001), geography (Sachs, 2001), openness (Frankel and Romer, 1999; Alcalá and Ciccone, 2004) and culture (Tabellini, 2010). The second strand investigates the small-sample performance of various estimators in linear IV models when instruments are weak. Contributions related to this strand include Flores-Lagunes (2007), Davidson and MacKinnon (2006), Blomquist and Dahlberg (1999), and Angrist *et al.* (1999) among others.

The rest of the paper is organized as follows. In the next section we apply some diagnostic tools to HJ99's linear IV model to determine if weak identification is an issue. Section 3 ranks

[3]Acemoglu *et al.* (2001, p. 1373) express certain skepticism about HJ99 using "latitude" as an instrument for institutions noting that "...the theoretical reasoning behind these instruments is not entirely convincing". Some scholars e.g. Rodrik *et al.* (2004) go on to claim that the instruments proposed by Acemoglu *et al.* (2001) are preferable to those of HJ99.

the estimators partially robust to weak instruments based on Monte Carlo evidence. Section 4 estimates the effect of institutions on growth using partially robust estimators. It considers whether results in HJ99 are driven by their reliance on geographic and trade instruments. Section 5 concludes.

## 2 Institution, Growth and Instrument Weakness

To find a source of exogenous variation in institutions that will not have a direct effect on current output levels, scholars have had to consider the geographical and historical determinants of institutions. For instance, Hall and Jones (1999) rely on distance from the equator, whereas Acemoglu *et al.* (2001) utilize historical settler mortality rates to instrument for institutional quality. Thus, it is natural to expect that the instruments proposed would only be weakly correlated with the endogenous variable of interest. This low correlation, often referred to as 'instrument weakness', might constitute a source of severe problems for estimation and inference purposes.[4] Stock *et al.* (2002) emphasize that the least applied researchers should do is to use the basic tools for detecting weak instruments such as first-stage $F$-statistics. Based on this suggestion, we report the values for the first-stage $F$-statistics for the six specifications in Table 1. Specification i relies on all four instruments used by HJ99. Specification ii uses the latitude and predicted trade intensity to instrument for social infrastructure. Specification iii instruments for institutional quality with linguistic variables alone. The remaining three specifications differ from specifications i-iii in that they rely on the sample without imputed data.

[ Insert Table 1 about here. ]

Based on the results shown in Table 1, two observations can be made. First, all specifications free of imputed data exhibit low $F$-statistics.[5] The same can be said about both specifications that rely on linguistic variables as instruments. Second, none of the reported $F$-statistics is much larger than ten. These observations suggest that weak instruments are potentially a problem. Before proceeding to the next section, we should comment on the role

[4]See, for example, Hahn and Hausman (2003) and references therein.

[5]The rule of thumb proposed by Stock *et al.* (2002) is that one should be concerned about weak instruments should the value of the first-stage $F$-statistic be lower than ten.

of $F$-statistics as a diagnostic tool. The literature on weak instruments suggests that one should interpret the first-stage $F$-statistics with a certain care. As noted by Stock and Yogo (2002) the Staiger-Stock rule of thumb is too conservative if limited information maximum likelihood or Fuller-$k$ estimators are used, but not conservative enough to ensure that the TSLS Wald test does not suffer from size distortions.

In this respect, Blomquist and Dahlberg (1999) express a certain skepticism concerning the excessive reliance on $F$-statistics as a diagnostic tool. They state that "...it has become somewhat of folklore that if the first stage $F-$statistics is large, the TSLS performs well. However, the folklore is not correct". Their claim is supported by simulations in which having average $F$-statistics of 29 with a sample size of 2408 does not preclude the TSLS from displaying the average bias of $-51.6\%$. In line with this argument Stock *et al.* (2002), suggest using $k$-class or Jackknife estimators even if the $F$-statistics are in excess of ten. However, Staiger and Stock (1997) show that even if the $F$-statistics are low (less that five), the instruments do not have to be irrelevant. To overcome this potential problem, they suggest using alternative estimators that are more robust to the presence of weak instruments, such as $k$-class estimators.

## 3 Monte-Carlo Simulations

As shown by Blomquist and Dahlberg (1999), in the Monte Carlo experiments, the ranking of the partially robust estimators heavily depends on the nature of the data generating process (DGP). Hence, they advocate complementing the estimates with a Monte-Carlo study for the relevant sample size and DGP. Based on this suggestion, I investigate the relative performance of the five $k$-class estimators and two Jackknife estimators in the context of the linear IV model used by HJ99.

### 3.1 DGP and evaluation criteria

The linear IV regression model estimated by HJ99 is given by

$$\begin{aligned} Y_L &= \alpha + \beta S + \varepsilon \\ S &= Z\delta + u, \end{aligned}$$

where $Y_L$ is an $N \times 1$ vector of log income per capita, $Z$ is an $N \times K_2$ matrix of instruments, $S$ is an $N \times 1$ vector that proxies for social infrastructure, and $\alpha$ is a scalar.

The DGP is designed so that the generated data replicates certain features of the actual observations. The dependent variable is generated as

$$y = a + bS + e,$$

where $a = 7$, $b = 3$, and $S$ is the proxy for social infrastructure from HJ99. Following the idea of Friedman (1984), the error term is generated by $e = k\widehat{u} + \xi$, where $\xi \sim N(0, \sigma_\xi^2)$ and $\widehat{u}$ is a projection of $S$ onto the space orthogonal to the space spanned by the instruments, i.e. $\widehat{u} = (I - Z(Z'Z)^{-1}Z')S$. The values of the parameters $\sigma_\xi^2 = 0.1$, and $k = 0.8$ are set so that the first two empirical moments of the generated $y$ are be close to (or, more specifically, within 10% in absolute value from) those of the observed $Y_L$ for a chosen extent of correlation between the regressor and error.

The advantage of the proposed DGP is that the generated data resemble the original sample in terms of the sample size, the first two empirical moments and, more importantly, the extent of the correlation of the instruments with the endogenous regressor. In fact, in each replication of the experiment, the first-stage $F$-statistics are identical to those reported in Table 1.

The disadvantages of this formulation include rather restrictive assumptions regarding fixed instruments and Gaussian errors. Measurement error is not modeled due to its unknown form. Hence, we implicitly assume that measurement error will not change the ranking of the partially robust estimators. Finally, for the purpose of this exercise, it is assumed that the instruments are orthogonal to the error term. The orthogonality assumption can be tested using the asymptotic methods suggested by Staiger and Stock (1997) or sample re-use methods. This strategy will be used in the section 5.1, which re-examines the results presented in HJ99.

A brief note should be made concerning the choice of measures used to compare the estimators. Some of the estimators considered in this study (e.g., JIVE or UJIVE) do not have first or second finite sample moments. For this reason, Angrist *et al.* (1999) advocate the use of median estimates and median absolute errors to evaluate the performance of the estimators. However, as pointed by Hausman *et al.* (2001), the absence of finite sample moments is an issue of practical concern. As Hausman *et al.* (2001) showed in the Monte Carlo experiments, the "moments problem" may cause certain estimators to exhibit extremely high mean estimates or root mean square error (RMSE). On these grounds, they advocate the use of mean estimates and RMSE to evaluate the finite sample performance of the estimators.

Following Angrist *et al.* (1999) and Blomquist and Dahlberg (1999), I evaluate performance of the estimators based on five statistics: percentage bias, RMSE, the median absolute error and the quantiles around the true parameter value. Furthermore, I report the coverage rates, which are computed as a fraction of the replications when the calculated confidence interval covers the true parameter value. Confidence intervals were estimated using the following bootstrap procedure with 1,000 replications. First, for a sample size $N$, I drew uniformly with replacement $N$ observations. Then, I used instrumental variables on the generated data to obtain a new estimate, $\widetilde{\beta}$. The bootstrapped 95% confidence interval was calculated as $\left[\widetilde{\beta}_{0.025}, \widetilde{\beta}_{0.975}\right]$ where $\widetilde{\beta}_{0.025}$, and $\widetilde{\beta}_{0.975}$ are the 2.5% and 97.5% percentiles, respectively, of the sampling distribution obtained.

### 3.2 Estimators

This study investigates the performance of the seven estimators that are partially robust to weak instruments. To choose the estimators, we follow Stock *et al.* (2002) who advocate use of $k$-class estimators and Jackknife estimators when weak identification is a concern. There are several reasons for doing so. First, partially robust estimators are more reliable than TSLS when instruments are weak. Second, they are relatively easy to compute. Third, these methods allow inference using conventional normal asymptotic approximation (Stock *et al.*, 2002). Finally, these estimators readily provide point estimates even when there are several endogenous regressors.

In this study we examine relative performance of five $k$-class estimators: OLS, TSLS, the limited information maximum likelihood estimator (LIML), an estimator of the family of Fuller -$k$ estimators, and bias-adjusted TSLS (BTSLS).[6] In addition, we consider two jackknife estimators: the Jackknife Instrumental Variable Estimator (JIVE) originally proposed by Angrist *et al.* (1995) and the Unbiased Jackknife Instrumental Variable Estimator

[6]The definition of these estimators given by Stock and Yogo (2002, p. 7).

(UJIVE) developed by Blomquist and Dahlberg (1999).

### 3.3 Simulation results

Simulations were performed for the six chosen specifications.[7] The results are reported in Figure 1 and in Tables 2 and 3. The conclusions are summarized as follows.

First, the performance of the TSLS estimator does not seem to be severely affected by the weak correlation of the instruments with the endogenous regressor. The results are robust across specifications.

Second, the jackknife estimators are outranked by the $k$-class estimators across all specifications. In particular, JIVE suffers from both severe average bias and size distortion. For instance, in specification iv, which relies on 79 observations and four instruments, JIVE exhibits a negative average bias of 24.8%, while the calculated 95% confidence interval covers the true parameter in only 67% of the cases.

Third, shifting from the four instruments to the linguistic instruments results in mote imprecise estimates. The RMSE almost doubles for the $k$-class estimators.

Finally, in specification vi, the performance of the OLS, LIML and jackknife estimators is inferior in terms of bias, RMSE, and median absolute error. The Fuller - $k$ estimator outranks the remaining estimators in terms of median square error, RMSE and size distortion. Furthermore, its performance remains stable across the specifications. Hence, the Fuller - $k$ estimator will be used for inference purposes.

[ Insert Figure 1, Table 2, and Table 3 about here. ]

These results parallel the findings of Hausman *et al.* (2001), who report a reduction in bias and MSE using the Fuller-$k$ estimator relative to TSLS and LIML when the instruments are weak. Furthermore, using the Monte-Carlo design of Hahn and Hausman (2002), Hahn *et al.* (2004) reached the conclusion that the TSLS, jackknife TSLS (UJIVE) and Fuller estimators often perform better than, for instance JIVE, or LIML. One potential explanation for this involves the notion of the "moments problem". Even though it has been long recognized that partially robust estimators, such as JIVE or LIML, do not have finite sample moments,

[7] All the simulations and estimations in this paper were performed using MATLAB. The code is available upon request.

Hahn *et al.* (2004) show that this feature can create problems in weak instrument situations. Although neither the TSLS nor the Fuller estimators suffers from ”moments problem”, jackknife TSLS has finite sample moments only up to the degree of overidentification. When I use the linguistic instruments only (specification vi), my simulation results are suggestive of the ”moments problem”. Both estimators with ”no moments” perform poorly, as does UJIVE due to the low degree of overidentification. The Fuller estimator and TSLS outrank the rest of the estimators, supporting the claim of Hahn and Hausman (2003) that ”instrument pessimism” is sometimes overstated for TSLS.

## 4 Reassessment of the Hall and Jones (1999) Results

This section discusses the estimation results of the HJ99 linear IV regression model using partially robust estimators. The results across the six specifications are reported in Table 4 and are briefly summarized below.

First, the point estimates obtained from TSLS, LIML, Fuller-$k$ and BTSLS do not differ substantially within particular specifications. The exception is JIVE, which gives somewhat lower point estimates than the rest of the partially robust estimators. This observation is consistent with the simulation results indicating that JIVE is plagued by negative bias while the $k$-class estimators are virtually free from it.

Second, in specifications i through v, most of the estimates obtained from the partially robust estimators are found to be significant at all conventional levela. The exceptions are the UJIVE estimates in specifications iii through v and the LIML estimates in v, which were found to be insignificant. However, the Monte-Carlo simulation results suggest that, in these cases, UJIVE has much lower degree of precision than the $k$-class estimators in terms of RMSE and median absolute error. Hence, for inference purposes TSLS, BTSTS and Fuller-$k$ are preferable.

Third, the specifications iii and vi, which rely exclusively on the linguistic instruments, produce somewhat higher point estimates than the rest of the specifications. The difference becomes more apparent in the sample with no imputed data.

Finally, specification vi, which relies on the linguistic instruments and the sample with no imputed data, is worth some special attention. In this specification, all of the estimates with the notable exception of the Fuller-$k$, are insignificant.[8] To interpret these findings, one should utilize the simulations results. They suggest that the LIML and jackknife estimators suffer from bias and size distortion. Furthermore, as shown in Figure 1 and Table 2, these estimators are more imprecise than the TSLS and Fuller-$k$ estimators. The latter conclusion follows from both the RMSE and the median absolute error criteria. Relying on the Fuller-$k$ estimator yields a point estimate of 7.08, which is significant at any conventional level.

# 5 Robustness Checks

The criterion of instrument relevance was considered above in detail. The issue that still requires some attention is instrument exogeneity. The problem is that the exogeneity and overidentification tests, which are equivalent under conventional asymptotics, are no longer equivalent under weak instruments. As shown by Staiger and Stock (1997), the tests may suffer from size distortions and lower power against violations of the orthogonality condition. This issue constitutes a rationale for scepticism about the overidentification tests reported by HJ99. Hence I used the guidelines provided by Staiger and Stock (1997) to address this problem. In addition, I considered a version of a bootstrap test of overidentifying restrictions.

## 5.1 Tests of overidentifying restrictions based on asymptotic approximations

Following Staiger and Stock (1997), I used two test statistics: $TR^2$ from the regression of the $IV$ residuals on the instruments and exogenous variables ($\phi_{reg}$), and Basmann's statistics ($\phi_{Bas}$). Both tests use the residuals from the following $k-$ class regressions: TSLS, LIML and Fuller - $k$. The results of the overidentification tests across specifications are reported in Table 5. Some comments on the results are as follows. First, the null of instrument exogeneity cannot be rejected at any conventional level of significance across most specifications and tests proposed. Second, the notable exception is specification iv (shorter sample, all four instruments), for which both the Basmann test and the $TR^2$ test reject the null at the 10% level. However, for the purpose of this studythe issue does not seem to be a major concern, because the null is not rejected for any specification involving only language instruments.

[8] With $t$-statistics of 1.27 obtained from the TSTL and BTSLS, some researchers (e.g. Frankel and Romer 1999) would call an estimate "marginally" significant.

Furthermore, according to the Monte Carlo results of Staiger and Stock (1997), the Basmann test is preferable for inference. They conclude that even though both test have size distortions, under the null, the TSLS version tends to overreject, while the LIML version tends to underreject. Thus, Staiger and Stock (1997) advocate reliance on the Basmann-LIML test, which in specification iv fails to reject the null.

In summary, the overidentifying restrictions cannot be rejected based on the tests suggested by Staiger and Stock (1997). This finding provides a justification for the orthogonality assumptions made in the Monte Carlo experiments. Interpreting these results, one should take into account that under certain parameter values, the Basmann-LIML test advocated by Staiger and Stock (1997) might have low power against small violations of the orthogonality condition. This suggests the need to reexamine the results with sample re-use methods instead of relying on asymptotic approximations.

### 5.2 Bootstrapping Basmann - LIML test

To overcome the potential problems associated with the Basmann-LIML test based on asymptotic approximation, the bootstrap principle can be applied. This claim can be justified on two grounds. First, as shown by Staiger and Stock (1997), the asymptotic distribution of the Basmann-LIML test statistics does not depend on any unknown parameter and therefore it is asymptotically pivotal. Thus, the bootstrap method is likely to converge faster than the corresponding asymptotic approximation and display lesser degree of size distortion (Horowitz, 2001). Second, bootstrap tests might be particularly advantageous relative to the asymptotic ones under weak identification. For instance, as argued by Wong (1996), who investigated the properties of the bootstrap Hausman test relative to the asymptotic approximation, the advantage of using bootstrapping increases as the correlation between the instruments and the endogenous regressors decreases.

The bootstrap procedure utilized in this section is similar to that of Wong (1996). However, the statistics bootstrapped here is based on the LIML version of the Basmann test. Furthermore, the procedure uses Friedman's (1984) orthogonalization idea. In matrix notation, the model considered is

$$
\begin{aligned}
y &= Y\beta + X\gamma + u, && (1)\\
Y &= Z\Pi + X\Phi + V, && (2)
\end{aligned}
$$

where (1) is the structural equation of interest, $y$ and $Y$ are $T \times 1$ vectors of the endogenous variables, and (2) is a reduced form equation for $Y$. $X$ is a $T \times K_1$ matrix of exogenous regressors, $Z$ is a $T \times K_2$ matrix of instruments, $u$ and $V$ are $T \times 1$ vectors of error terms. The errors $\begin{pmatrix} u_t & V_t \end{pmatrix}'$, where $u_t$ denotes the $t$-th observation on $u$, are assumed to be zero-mean, serially uncorrelated, and homoskedastic. Furthermore, it is assumed that both the exogenous variables and the instruments are orthogonal to the error terms.

In the context of HJ99 study, $y$ is the vector of log income per capita, $Z$ is a matrix of instruments, $Y$ is a vector that proxies for social infrastructure, $X$ is a $T \times 1$ vector of ones, and $\beta$ is the main parameter of interest.

The bootstrap procedure follows a simple algorithm. First, the model in (1) is estimated using a $k$-class estimator to obtain the residual vector $\widehat{u}(k) = y - Y\widehat{\beta}(k) - X\widehat{\gamma}(k)$. Those residuals will not be exactly orthogonal to the set of instruments; that is, $\frac{1}{T}Z'\widehat{u}(k) \neq 0$. This assumption, however, is imposed by the model. Following Friedman (1984), one might utilize the component of the residual vector orthogonal to the set of instruments given by $\widetilde{u}(k) = (I - Z'(Z'Z)^{-1}Z')\widehat{u}(k)$, where $I$ denotes a $T$-dimensional identity matrix.

Second, given a quadruple of $(y, Y, Z, X)$, the bootstrap method is used to re-sample with replacement from the empirical distribution of $(\widetilde{u}(k), Y, Z, X)$. Denoting the values drawn from the empirical distribution by $(\widetilde{u}(k)^*, Y^*, Z^*, X^*)$, $y^*$ is generated for each bootstrap replication as $y^* = Y^*\widehat{\beta}(k) + X^*\widehat{\gamma}(k) + \widetilde{u}(k)^*$.

Next, using the bootstrap sample the model, was estimated using a $k$-class estimator to obtain a residual vector $\widehat{u}^*(k) = y^* - Y^*\widehat{\beta}^*(k) - X^*\widehat{\gamma}^*(k)$. Then the Basmann's statistic was calculated as

$$\phi^*_{Bas}(k) = \left[\widehat{u}^*(k)' P^*_{Z^\perp}\widehat{u}^*(k)\right] / \left[\widehat{u}^*(k)' M^*_{Z^\perp}\widehat{u}^*(k) / (T - K_1 - K_2)\right],$$

where $Z^{*\perp} = (I - X^*(X^{*\prime}X^*)^{-1}X^*)Z^*$, $P^*_{Z^\perp}$ and $M^*_{Z^\perp}$ are the corresponding projection and residualizing matrices given by $P^*_{Z^\perp} = Z^{*\perp\prime}(Z^{*\perp\prime}Z^{*\perp})^{-1}Z^{*\perp\prime}$ and $M^*_{Z^\perp} = I - P^*_{Z^\perp}$, respectively.

The Basmann's statistic $\phi^*_{Bas}(k)$ was calculated for each of $10,000$ bootstrap replications. The 90%, 95% and 99% percentiles in the obtained empirical distribution represent the corresponding bootstrap critical values reported in Table 6.

For convenience, the Basmann-LIML statistics $\phi_{Bas}(k_{LIML})$ for all six specifications are repeated in Table 6 along with the bootstrap critical values. The conclusion about instrument orthogonality drawn from the bootstrapped Basmann-LIML test coincides with those based on the asymptotic approximations. In each case, the Basmann-LIML test statistic falls short of the tabulated critical values, which implies failure to reject the overidentifying restrictions.

In summary, the overidentifying restrictions cannot be rejected based on the tests robust to the presence of weak instruments for all specifications. The conclusion holds for tests based on asymptotic approximations and as the test relying on small sample properties. The evidence obtained provides further support for the orthogonality assumption made in section 3.

# 6 Conclusion

This paper focuses on the empirical approach used by Hall and Jones (1999) to estimate the effect of institutions on income across countries. We show that instruments proposed by HJ99, and in particular linguistic instruments, are weakly correlated with their proxy for institutional quality. The basic diagnostic criteria for weak identification, such as low values of the first stage $F$-statistics, indicate that the performance of the TSLS estimator might be poor.

Our Monte Carlo experiments suggest that in the context of HJ99 model the TSLS and Fuller-$k$ estimators are not plagued by either severe size distortion or bias despite low values of first-stage $F$-statistics. When social infrastructure is instrumented with languages, inference should be made on the basis of the Fuller-$k$ estimator. We find the linguistic variables to be relevant instruments for the institutional quality. When we instrument institutional quality with the fraction of the population speaking English and European languages at birth, social infrastructure remains a significant determinant of income differences.

## References

Acemoglu, D., Johnson, S. and Robinson, J. A. (2001) The colonial origins of comparative development: An empirical investigation, *American Economic Review*, **91**, 1369–1401.

Alcalá, F. and Ciccone, A. (2004) Trade and productivity, *The Quarterly Journal of Economics*, **119**, 612–645.

Angrist, J. D., Imbens, G. W. and Krueger, A. (1995) Jackknife instrumental variables estimation, NBER Technical Working Papers 0172, National Bureau of Economic Research, Inc.

Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999) Jackknife instrumental variables estimation, *Journal of Applied Econometrics*, **14**, 57–67.

Blomquist, S. and Dahlberg, M. (1999) Small sample properties of LIML and jackknife IV estimators: Experiments with weak instruments, *Journal of Applied Econometrics*, **14**, 69–88.

Davidson, R. and MacKinnon, J. G. (2006) The case against jive, *Journal of Applied Econometrics*, **21**, 827–833.

Flores-Lagunes, A. (2007) Finite sample evidence of iv estimators under weak instruments, *Journal of Applied Econometrics*, **22**, 677–694.

Frankel, J. A. and Romer, D. (1999) Does trade cause growth?, *American Economic Review*, **89**, 379–399.

Friedman, D. A. (1984) On bootstrapping two-stage least-squares estimates in stationary linear models, *The Annals of Statistics*, **12**, 827–842.

Glaeser, E. L., Porta, R. L., de Silanes, F. L. and Shleifer, A. (2004) Do institutions cause growth?, *Journal of Economic Growth*, **9**, 271–303.

Hahn, J. and Hausman, J. (2002) A new specification test for the validity of instrumental variables, *Econometrica*, **70**, 163–189.

Hahn, J. and Hausman, J. (2003) Weak instruments: Diagnosis and cures in empirical econometrics, *American Economic Review*, **93**, 118–125.

Hahn, J., Hausman, J. and Kuersteiner, G. (2004) Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations, *Econometrics Journal*, **7**, 272–306.

Hall, R. E. and Jones, C. I. (1999) Why do some countries produce so much more output per worker than others?, *The Quarterly Journal of Economics*, **114**, 83–116.

Hausman, J. A., Hahn, J. and Kuersteiner, G. (2001) Higher order MSE of jackknife 2SLS, Tech. rep., MIT, Department of Economics.

Horowitz, J. L. (2001) The bootstrap, in *Handbook of Econometrics* (Eds.) J. Heckman and E. Leamer, Elsevier, chap. 52, pp. 3159–3228.

Rodrik, D., Subramanian, A. and Trebbi, F. (2004) Institutions rule: The primacy of institutions over geography and integration in economic development, *Journal of Economic Growth*, **9**, 131–165.

Sachs, J. D. (2001) Tropical underdevelopment, Working Paper 8119, National Bureau of Economic Research.

Staiger, D. and Stock, J. H. (1997) Instrumental variables regression with weak instruments, *Econometrica*, **65**, 557–586.

Stock, J. H., Wright, J. H. and Yogo, M. (2002) A survey of weak instruments and weak identification in generalized method of moments, *Journal of Business and Economic Statistics*, **20**, 518–29.

Stock, J. H. and Yogo, M. (2002) Testing for weak instruments in linear IV regression, NBER Technical Working Papers 0284, National Bureau of Economic Research, Inc.

Tabellini, G. (2010) Culture and institutions: Economic development in the regions of europe, *Journal of the European Economic Association*, **8**, 677–716.

Wong, K.-f. (1996) Bootstrapping Hausman's exogeneity test, *Economics Letters*, **53**, 139–143.

Table 1: First-stage $F$-statistics across specifications

| | Specification | First-stage $F$-statistic |
|---|---|---|
| i) | Full sample (127 countries), four instruments | 12.843 |
| ii) | Full sample (127 countries), latitude & Frankel-Romer trade share | 13.118 |
| iii) | Full sample (127 countries), linguistic instruments | 6.499 |
| iv) | No imputed data (79 countries), four instruments | 6.721 |
| v) | No imputed data (79 countries), latitude & Frankel-Romer trade share | 9.066 |
| vi) | No imputed data (79 countries), linguistic instruments | 2.118 |

Table 2: Monte Carlo simulation results

| | $k-$class estimators | | | | | Jackknife estimators | |
|---|---|---|---|---|---|---|---|
| | OLS | TSLS | LIML | Fuller$-k$ | BTSLS | JIVE | UJIVE |
| Specification i) | (HJ99 full sample, 127 obs., 4 instruments); $\rho = 0.391$; | | | | | | |
| Av. % Bias | 0.187 | 0.001 | 0.013 | 0.007 | 0.011 | $-0.141$ | $-0.031$ |
| RMSE | 0.573 | 0.205 | 0.216 | 0.210 | 0.215 | 0.470 | 0.248 |
| SIZE | 0.001 | 0.974 | 0.980 | 0.976 | 0.976 | 0.790 | 0.980 |
| Specification ii) | (HJ99 full sample, Distance/FR tr. share); $\rho = 0.452$; | | | | | | |
| Av. % Bias | 0.219 | 0.002 | 0.005 | 0.005 | 0.002 | $-0.170$ | $-0.037$ |
| RMSE | 0.668 | 0.274 | 0.281 | 0.270 | 0.274 | 0.577 | 0.334 |
| SIZE | 0.000 | 0.972 | 0.983 | 0.978 | 0.972 | 0.836 | 0.985 |
| Specification iii) | (HJ99 full sample, languages); $\rho = 0.485$; | | | | | | |
| Av. % Bias | 0.238 | 0.002 | 0.019 | 0.003 | 0.002 | $-0.308$ | $-0.084$ |
| RMSE | 0.723 | 0.372 | 0.395 | 0.361 | 0.372 | 0.995 | 0.550 |
| SIZE | 0 | 0.974 | 0.970 | 0.972 | 0.974 | 0.750 | 0.996 |
| Specification iv) | (Sample with no imputed data, 79 obs., 4 instruments); $\rho = 0.415$ | | | | | | |
| Av. % Bias | 0.196 | 0.002 | 0.026 | 0.015 | 0.022 | $-0.249$ | $-0.061$ |
| RMSE | 0.604 | 0.266 | 0.294 | 0.275 | 0.295 | 0.791 | 0.377 |
| SIZE | 0.017 | 0.974 | 0.976 | 0.976 | 0.977 | 0.676 | 0.980 |
| Specification v) | (Sample with no imputed data, Distance/FR tr. share); $\rho = 0.453$ | | | | | | |
| Av. % Bias | 0.216 | 0.0002 | 0.011 | 0.004 | 0.0002 | $-0.236$ | $-0.058$ |
| RMSE | 0.662 | 0.309 | 0.322 | 0.303 | 0.309 | 0.771 | 0.416 |
| SIZE | 0.080 | 0.981 | 0.984 | 0.977 | 0.981 | 0.814 | 0.989 |
| Specification vi) | (Sample with no imputed data, languages); $\rho = 0.516$ | | | | | | |
| Av. % Bias | 0.251 | $-0.010$ | $-0.055$ | 0.017 | $-0.010$ | $-0.887$ | $-0.616$ |
| RMSE | 0.766 | 0.599 | 0.912 | 0.535 | 0.599 | 2.724 | 2.699 |
| SIZE | 0.000 | 0.998 | 1.000 | 0.989 | 0.998 | 0.543 | 1.000 |

Note: Average per cent bias (Av % Bias), RMSE and SIZE (coverage rate for a 95% confidence interval calculated as a proportion of replications when the confidence interval covers the true parameter value) for the estimates of $\beta$. Reported $\rho$ indicates the average sample correlation coefficient between the endogenous regressor and the error term. The experiments rely on 1000 Monte Carlo replications and 1000 Bootstrap iterations.

Table 3: Quantiles around $\beta$ and median absolute errors: six specifications

| Specification | Estimator | Quantiles around $\beta$ | | | | | Median |
|---|---|---|---|---|---|---|---|
| | | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 | Absolute Error |
| Specification i) | OLS | 3.419 | 3.480 | 3.562 | 3.645 | 3.709 | 0.562 |
| Full HJ99 sample, | TSLS | 2.732 | 2.857 | 2.991 | 3.135 | 3.258 | 0.139 |
| 127 observations, | LIML | 2.686 | 2.816 | 2.953 | 3.106 | 3.240 | 0.144 |
| 4 instruments | Fuller$-k$ | 2.710 | 2.836 | 2.971 | 3.120 | 3.252 | 0.140 |
| | BTSLS | 2.690 | 2.820 | 2.963 | 3.109 | 3.239 | 0.142 |
| | JIVE | 2.315 | 2.434 | 2.573 | 2.717 | 2.835 | 0.426 |
| | UJIVE | 2.610 | 2.745 | 2.902 | 3.065 | 3.197 | 0.165 |
| Specification ii) | OLS | 3.510 | 3.579 | 3.660 | 3.735 | 3.809 | 0.660 |
| Full HJ99 sample, | TSLS | 2.662 | 2.822 | 2.996 | 3.195 | 3.370 | 0.184 |
| 127 observations, | LIML | 2.631 | 2.794 | 2.973 | 3.177 | 3.355 | 0.192 |
| 2 instruments: | Fuller$-k$ | 2.678 | 2.833 | 3.004 | 3.202 | 3.370 | 0.184 |
| latitude and | BTSLS | 2.662 | 2.822 | 2.996 | 3.195 | 3.370 | 0.184 |
| trade intensity | JIVE | 2.151 | 2.306 | 2.482 | 2.673 | 2.845 | 0.517 |
| | UJIVE | 2.496 | 2.675 | 2.879 | 3.100 | 3.301 | 0.237 |
| Specification iii) | OLS | 3.556 | 3.633 | 3.716 | 3.793 | 3.861 | 0.716 |
| Full HJ99 sample, | TSLS | 2.514 | 2.743 | 2.996 | 3.246 | 3.445 | 0.253 |
| 127 observations, | LIML | 2.440 | 2.681 | 2.941 | 3.217 | 3.420 | 0.264 |
| 2 linguistic, | Fuller$-k$ | 2.541 | 2.766 | 3.004 | 3.259 | 3.446 | 0.254 |
| instruments | BTSLS | 2.514 | 2.743 | 2.996 | 3.246 | 3.446 | 0.253 |
| | JIVE | 1.596 | 1.820 | 2.077 | 2.321 | 2.527 | 0.993 |
| | UJIVE | 2.113 | 2.410 | 2.751 | 3.074 | 3.345 | 0.359 |
| Specification iv) | OLS | 3.408 | 3.492 | 3.586 | 3.681 | 3.768 | 0.586 |
| No imputed data, | TSLS | 2.648 | 2.823 | 3.000 | 3.175 | 3.326 | 0.176 |
| 79 observations, | LIML | 2.554 | 2.733 | 2.928 | 3.111 | 3.270 | 0.192 |
| 4 instruments | Fuller$-k$ | 2.602 | 2.772 | 2.964 | 3.141 | 3.293 | 0.182 |
| | BTSLS | 2.564 | 2.750 | 2.937 | 3.126 | 3.294 | 0.190 |
| | JIVE | 1.920 | 2.088 | 2.262 | 2.433 | 2.590 | 0.737 |
| | UJIVE | 2.399 | 2.609 | 2.286 | 3.040 | 3.235 | 0.248 |
| Specification v) | OLS | 3.461 | 3.556 | 3.644 | 3.740 | 3.825 | 0.644 |
| No imputed data, | TSLS | 2.613 | 2.777 | 3.001 | 3.208 | 3.398 | 0.217 |
| 79 observations, | LIML | 2.569 | 2.734 | 2.964 | 3.186 | 3.373 | 0.231 |
| 2 instruments: | Fuller$-k$ | 2.640 | 2.792 | 3.012 | 3.214 | 3.403 | 0.209 |
| latitude and | BTSLS | 2.613 | 2.777 | 3.001 | 3.208 | 3.398 | 0.217 |
| trade intensity | JIVE | 1.915 | 2.079 | 2.287 | 2.504 | 2.681 | 0.713 |
| | UJIVE | 2.361 | 2.564 | 2.820 | 3.088 | 3.306 | 0.287 |
| Specification vi) | OLS | 3.576 | 3.656 | 3.753 | 3.843 | 3.928 | 0.752 |
| No imputed data, | TSLS | 2.229 | 2.564 | 2.975 | 3.373 | 3.743 | 0.400 |
| 79 observations, | LIML | 1.965 | 2.340 | 2.843 | 3.293 | 3.723 | 0.477 |
| 2 linguistic, | Fuller$-k$ | 2.389 | 2.673 | 3.052 | 3.410 | 3.745 | 0.373 |
| instruments | BTSLS | 2.229 | 2.564 | 2.975 | 3.373 | 3.743 | 0.400 |
| | JIVE | −0.408 | −0.033 | 0.361 | 0.714 | 1.073 | 2.639 |
| | UJIVE | −1.388 | −0.111 | 1.228 | 2.432 | 3.652 | 1.938 |

Table 4: Estimation results, six specifications

| | $k-$class estimators | | | | | Jackknife estimators | |
|---|---|---|---|---|---|---|---|
| | OLS | TSLS | LIML | Fuller$-k$ | BTSLS | JIVE | UJIVE |
| i) 127 obs.; 4 inst | 3.289 | 5.085 | 5.300 | 5.243 | 5.186 | 4.772 | 5.382 |
| | (0.196) | (0.508) | (0.676) | (0.633) | (0.556) | (0.455) | (0.705) |
| ii) 127; Dist/FR | | 4.670 | 4.760 | 4.692 | 4.670 | 4.234 | 4.912 |
| | | (0.676) | (1.435) | (0.756) | (0.676) | (0.632) | (2.247) |
| iii) 127; Languages | | 5.798 | 6.210 | 5.951 | 5.798 | 5.149 | 6.818 |
| | | (1.600) | (3.202) | (1.326) | (1.600) | (1.129) | (16.162) |
| iv) 79; 4 instruments | 3.074 | 4.661 | 5.268 | 5.142 | 4.828 | 4.090 | 5.110 |
| | (0.253) | (0.612) | (3.726) | (0.996) | (0.735) | (0.559) | (3.047) |
| v) 79; Dist./FR tr.sh | | 3.939 | 3.975 | 3.915 | 3.939 | 3.344 | 4.124 |
| | | (0.705) | (7.983) | (0.733) | (0.705) | (0.770) | (9.654) |
| vi) 79; Languages | | 6.538 | 8.645 | 7.082 | 6.538 | 4.528 | 15.417 |
| | | (5.130) | (26.941) | (2.259) | (5.129) | (4.435) | (27.715) |

Note: The dependent variable is log of income per capita. The regressors are a constant and a proxy for social infrastructure. Depending on specification the instruments used include fraction of population speaking a English at birth, fraction of population speaking a European language, distance from the equator and Frankel and Romer (1999) geography predicted trade intensity. Standard errors are given in the parenthesis. The standard errors are computed using a Bootstrap procedure described in the text.

Table 5: Testing overidentifying restrictions, six specifications

| | $TR^2$ test ($\phi_{reg}$) | | | Basmann test ($\phi_{Bas}$) | | |
|---|---|---|---|---|---|---|
| | TSLS | LIML | Fuller$-k$ | TSLS | LIML | Fuller$-k$ |
| i) 127 obs.; 4 instruments | 4.177 | 4.035 | 4.044 | 4.149 | 4.003 | 4.013 |
| | (0.243) | (0.258) | (0.257) | (0.246) | (0.261) | (0.260) |
| ii) 127 obs.; Dist/FR tr.sh | 1.377 | 1.356 | 1.370 | 1.360 | 1.342 | 1.352 |
| | (0.241) | (0.244) | (0.244) | (0.244) | (0.247) | (0.245) |
| iii) 127 obs.; Languages | 1.823 | 1.698 | 1.745 | 1.806 | 1.681 | 1.727 |
| | (0.177) | (0.193) | (0.187) | (0.179) | (0.195) | (0.189) |
| iv) 79 obs.; 4 instruments | 6.431 | 5.824 | 5.846 | 6.558 | 5.890 | 5.914 |
| | (0.092) | (0.121) | (0.119) | (0.087) | (0.117) | (0.116) |
| v) 79 obs.; Dist/FR tr.sh | 0.616 | 0.613 | 0.620 | 0.597 | 0.594 | 0.601 |
| | (0.433) | (0.434) | (0.431) | (0.440) | (0.441) | (0.438) |
| vi) 79 obs.; Languages | 2.192 | 1.578 | 1.856 | 2.169 | 1.549 | 1.828 |
| | (0.139) | (0.209) | (0.173) | (0.141) | (0.214) | (0.177) |

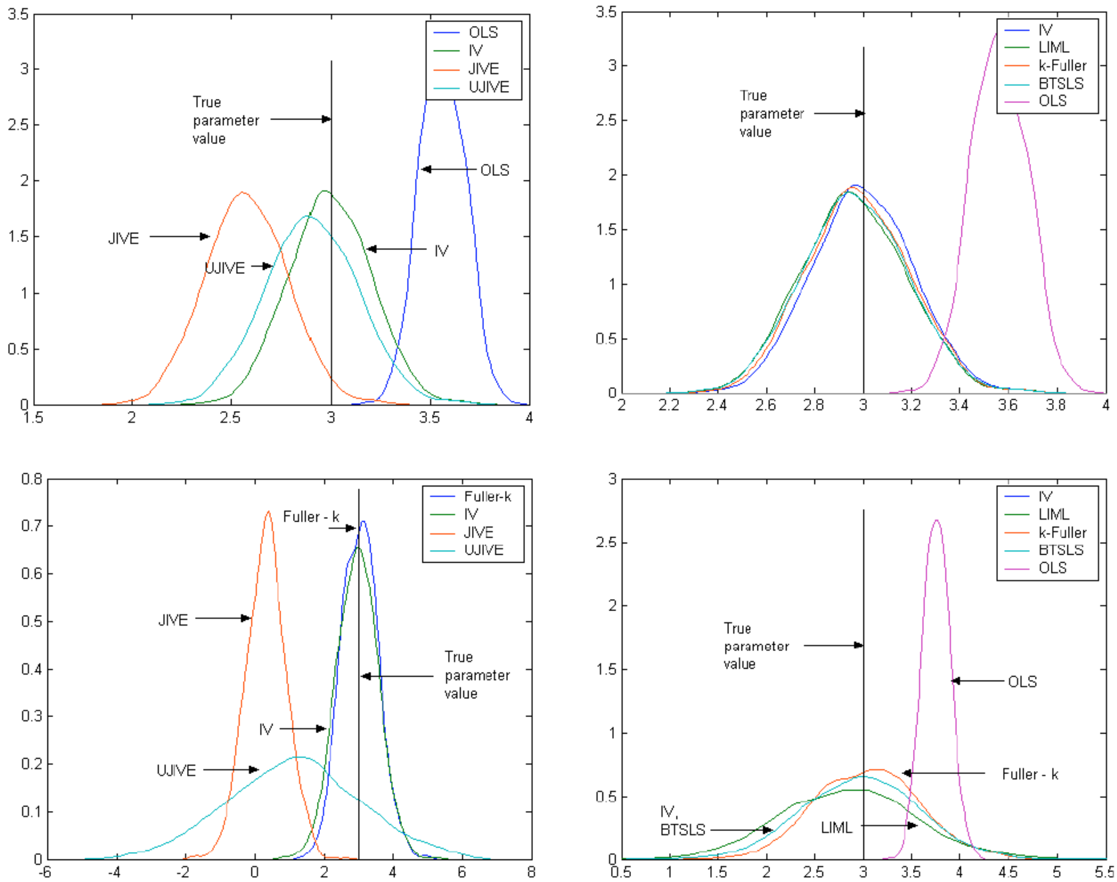
Note: The table reports $TR^2$ over-identification test statistics and Basmann's test statistics in the $\chi^2$ form. The tests results are reported for the residuals of the k-class regressions namely TSLS, LIML and Fuller$-k$ respectively. The $p$-values reported in the parenthesis are based on asymptotic approximation. The dependent variable in the second stage IV regression is log of income per capita. The regressors are a constant and a proxy for social infrastructure. Depending on specification the instruments used include fraction of population speaking English at birth, fraction of population speaking a European language, distance from the equator, and Frankel and Romer (1999) predicted trade share.

Table 6: Bootstrap critical values for Basmann–LIML test of overidentifying restrictions

| Specification | 90% critical values | 95% critical values | 99% critical values | Basmann statistic $\phi_{Bas}(k_{LIML})$ |
|---|---|---|---|---|
| i) Full sample, 4 instruments | 7.375 | 9.353 | 14.423 | 4.003 |
| ii) Full sample, latitude & FR tr.share | 3.217 | 4.564 | 7.760 | 1.342 |
| iii) Full sample, languages | 3.017 | 4.400 | 7.469 | 1.681 |
| iv) No imputations, 4 instruments | 6.293 | 7.925 | 11.602 | 5.890 |
| v) No imputations, latitude & FR tr.share | 2.578 | 3.644 | 6.489 | 0.594 |
| vi) No imputations, languages | 2.432 | 3.509 | 5.989 | 1.549 |

Note: The table repots a set of critical values for the LIML version of the Basmann's test of overidentifying restrictions and the corresponding test statistics for each of the six specifications considered. The bootstrap procedure is described in the text.

Figure 1: Kernel density estimates of beta: specifications i) and vi)




Note: The figure reports distributions for the estimators of the coefficient on institutional quality based on the kernel density estimation. The upper panel corresponds to the specification involving all four instruments and the larger sample (127 observations). The lower panel reports the results from the specification relying on the two linguistics instruments only and the sample free from imputed data (79 observations). Left panels serve for comparison of the jackknive estimators with some of the $k$-class estimators. The right panels present the distributions of all of the k-class estimators considered. The details of the DGP and the estimators used are described in the text.