econometrics

MDPI

Article

Evaluating Ingenious Instruments for Fundamental Determinants of Long-Run Economic Growth and Development

P. Dorian Owen

Department of Economics, University of Otago, P.O. Box 56, Dunedin 9054, New Zealand; dorian.owen@otago.ac.nz; Tel.: +64-3-479-8655

Academic Editors: Gilles Dufrénot, Fredj Jawadi and Alexander Mihailov
Received: 25 November 2016; Accepted: 18 August 2017; Published: 5 September 2017

Abstract: Empirical studies of the determinants of cross-country differences in long-run development are characterized by the ingenious nature of the instruments used. However, scepticism remains about their ability to provide a valid basis for causal inference. This paper examines whether explicit consideration of the statistical adequacy of the underlying reduced form, which provides an embedding framework for the structural equations, can usefully complement economic theory as a basis for assessing instrument choice in the fundamental determinants literature. Diagnostic testing of the reduced forms in influential studies reveals evidence of model misspecification, with parameter non-constancy and spatial dependence of the residuals being almost ubiquitous. This feature, surprisingly not previously identified, potentially undermines the inferences drawn about the structural parameters, such as the quantitative and statistical significance of different fundamental determinants.

Keywords: fundamental determinants of economic development; long-run economic growth; instrumental variables; reduced form; statistical adequacy

JEL Classification: O10; O40; C36

1. Introduction

Interpreting patterns of causation from growth regressions is fraught with difficulties. By the nature of the process of economic growth, 'proximate' determinants, such as human capital, physical capital, and technology are interrelated and jointly determined with income per capita. One response is to step back from the evaluation of the effects of these 'proximate' determinants and investigate the 'deeper', more fundamental, determinants of long-term growth and hence levels of development. The search for fundamental determinants has concentrated on relatively slowly changing factors that have a pervasive effect on economies over long periods, with the initial focus on the relative importance of institutions and geography, and, more recently, history, biology, and culture (Acemoglu et al. 2005; Spolaore and Wacziarg 2013, 2014). Whether a variable is considered to be exogenous or endogenous has not, however, been used as a criterion to distinguish proximate from fundamental determinants. For example, whereas many aspects of geography, history, and biology are temporally predetermined, institutions are more obviously endogenous, if only because more highly developed economies demand and can afford better quality institutions.

Consequently, widespread use of instrumental variables (IV) estimation, specifically two-stage least squares (2SLS), is a defining feature of the literature examining the fundamental determinants of cross-country differences in levels of development. As The Economist (2006, p. 84) pointedly observes, "all of the fun in the recent spate of papers is in the instruments themselves. Economists are outdoing each other with ever more curious instruments, ranging from lethal mosquitoes (Sachs 2003) to heirless maharajahs (Iyer 2010), or ... wind speeds and sea currents (Feyrer and Sacerdote 2009) ... [i]ndeed, 'reverse causality', which was once a frustrating problem, is now seen as a chance to demonstrate ingenuity."

Despite the ingenious nature of many of these instruments, there is scepticism about their ability to provide a convincing basis for causal inference. Durlauf et al. (2005, p. 638) express this view forcefully: " ... the belief that it is easy to identify valid instrumental variables in the growth context is deeply mistaken. We regard many applications of instrumental variable procedures in the empirical growth literature to be undermined by the failure to address properly the question of whether these instruments are valid, i.e., whether they may be plausibly argued to be uncorrelated with the error term in a growth regression". Some authors (e.g., Freedman 2006; Qin 2015; Swamy et al. 2015) express despair at the usefulness of IV estimation methods more generally.

Justification of instrument validity conventionally relies on 'telling a good story' and on the a priori degree of realism of any counter-example (Frankel 2003). This is usually supported by reporting results of tests of overidentifying restrictions, although these cannot test the validity of the overall instrumentation strategy. Concerns about the validity and relevance of instruments have led to practical suggestions for strengthening the basis for causal inference based on IV estimation (Murray 2006; Angrist and Pischke 2009; Bazzi and Clemens 2013; Kraay 2015), but these focus mainly on assessing the plausibility of estimates or addressing weak instrumentation.

This paper applies the approach proposed by Spanos (1990, 2006, 2007, 2015) to consider the statistical dimensions of the instrumentation strategies used in the fundamental determinants literature, as a complement to assessing instrument choice primarily on the basis of economic theory. Spanos' approach emphasizes that behind every structural model there exists a statistical model, which comprises the totality of the probabilistic assumptions imposed on the data; consequently, the validity of these statistical assumptions ('statistical adequacy') is crucial for securing reliable inference. He focuses on the probabilistic underpinnings of IV estimation by explicit consideration of the implicit reduced form (RF) as the statistical model that summarizes the information in the observed data. This highlights the need to probe the statistical adequacy of the RF (i.e., whether the probabilistic assumptions are valid for the data under consideration) by misspecification testing. This step is a prerequisite for testing overidentification restrictions and whether instruments are weak, and, ultimately, for reliable inference on structural parameters. In contrast, standard practice in the application of 2SLS estimation in this literature (and more widely) is to focus on these latter characteristics and ignore the statistical adequacy of the overall framework.

Section 2 contains an overview of the nature of the instruments used in the literature on the fundamental determinants of comparative development. Section 3 discusses the contributions of economic theory and statistics in devising valid instrumentation strategies in this context and outlines Spanos' arguments on the role of the RF. Section 4 outlines the tests used to assess the statistical adequacy of RFs and Section 5 reports and discusses the results for a representative selection of influential studies. Section 6 concludes.

2. Ingenious Instruments for Fundamental Determinants of Economic Development

Empirical studies in the fundamental determinants literature use parsimonious models to evaluate the relevance of different fundamental determinants in explaining comparative levels of long-run economic development, usually measured by income per capita. The slowly evolving nature of variables identified as fundamental determinants and the lack of long runs of relevant time-series data lead to reliance on exploiting cross-country variation in cross-sectional analyses. Most of the earlier studies focus on competing claims about the primacy of the quality of institutions (Hall and Jones 1999; Acemoglu et al. 2001, 2002; Easterly and Levine 2003; Rodrik et al. 2004) versus the role of geographical endowments (Bloom and Sachs 1998; Gallup et al. 1999; Sachs 2003; Olsson and Hibbs 2005). The multiple mechanisms by which geography and institutions can affect income are discussed in detail in many of the original papers and later reviews (Easterly and Levine 2003; Rodrik et al. 2004;

Acemoglu et al. 2005; Olsson 2005; Spolaore and Wacziarg 2013); the following comments, therefore, concentrate on the nature of the instruments used in this literature.

Institutional quality is likely to be endogenous in a model explaining income per capita for several reasons: reverse causality (higher levels of income per capita provide the resources to enhance institutional quality), omitted variables correlated with both income and institutions, and measurement error. Finding appropriate instruments for institutions is therefore a priority in order to obtain consistent estimates of the partial effect of institutions on income per capita. In contrast, it has been argued that geography is "as exogenous a determinant as an economist can ever hope to get" (Rodrik et al. 2004, p. 133). However, the predetermined nature of variables reflecting aspects of geography (or biology or history) does not necessarily imply they are exogenous, i.e., orthogonal to the error term in the structural model. Error terms in models fitted to observational data are 'derived' variables, reflecting model specification (Hendry and Nielsen 2007, p. 160). Consequently, omitted relevant explanatory variables correlated with geographical, biological, or historical variables may induce econometric endogeneity, and hence bias and inconsistency. In a similar vein, Deaton (2010, p. 431) emphasizes the crucial difference between exogenous variables and variables that are 'external' (i.e., not caused by variables in the model): "[w]hether any of these instruments is *exogenous* (or satisfies the exclusion restrictions) depends on the specification of the equation of interest and is not guaranteed by its *externality*" (emphasis in original).

Hall and Jones (1999), in an early empirical contribution demonstrating the importance of institutional quality, choose instruments on the basis that societies more strongly influenced by Western Europeans were more likely to adopt favourable institutions. Their proxies for Western European influence include absolute latitude, the fraction of the population speaking one of the five major Western European languages as their first language, and the fraction speaking English as their first language. Their identification strategy relies on these variables being correlated with their measure of institutional quality but having no direct effect on current output per worker (especially for latitude) and not reflecting targeting of Western influence to areas with higher present-day output per worker (especially for the language fractions).

Acemoglu et al. (2001), in the most influential study in the fundamental determinants literature, instrument institutional quality, specifically the strength of property rights, using historical European settler mortality. Favourable disease environments (lower settler mortality) initially led to 'settler colonies' with higher-quality institutions, whereas unfavourable disease environments (higher settler mortality) led to 'extractive colonies' with poorer-quality institutions geared to expropriating returns from local resources. The persistence of institutions after colonization resulted in these choices having long-lasting effects on current institutions and current living standards. Acemoglu et al. (2001, 2002) argue that settler mortality satisfies the required exclusion restriction for a valid instrument because the effect of historical disease environment on current living standards is entirely indirect, via its effect on historical and current institutions. The restriction would be questionable if historical and current disease environments are correlated and the latter has a direct effect not controlled for in the model, or if institutional quality is correlated with other persistent settler characteristics (e.g., human capital or culture) that have important impacts on development.

Engerman and Sokoloff (1997) emphasize mineral and crop endowments as the driving force behind the mode of colonization. Abundance of minerals and of crops such as sugarcane, tobacco, and cotton, combined with high indigenous population density, encouraged the use of plantation agriculture and slave labour to exploit economies of scale, and led to inequality and poor-quality institutions. In contrast, endowments suited to grain and livestock, combined with sparse population, promoted more egalitarian family farming, development of a sizeable middle class, and good-quality institutions. Thus, a distinctive aspect of Easterly and Levine's (2003) instrumentation strategy is the inclusion of a set of crop and mineral endowment dummies. Similarly, Easterly (2007) proposes the ratio of the share of arable land suitable for growing wheat to the corresponding share suitable for growing sugarcane as the basis for an instrument for inequality.

Early empirical studies in this literature (Acemoglu et al. 2001; Easterly and Levine 2003; Rodrik et al. 2004) conclude that geographic conditions affect development purely via their indirect effect on institutions. In response, Sachs (2003) shows that a measure of malaria transmission is statistically significant when added to representative specifications from these studies, implying that geographical variables also have a direct effect on GDP per capita. Because richer countries can marshal more resources to eradicate malaria, malarial risk is treated as endogenous, so Sachs adds an index of malarial ecology, based on external bio-geographical variables, to his set of instruments.

Bockstette et al. (2002) propose state antiquity, measuring the historical depth of experience with state-level institutions, as a possible instrument for institutional quality and demonstrate its positive association with Hall and Jones' (1999) measure of institutional quality. More recently, it has been included in equations explaining income per capita or population density as a potential historical fundamental determinant (Chanda and Putterman 2007; Putterman and Weil 2010). Classification of legal origin, especially English common law versus French civil law, has been widely used as an instrument for institutional quality and financial market development, with common law regarded as providing greater protection for investors' rights (La Porta et al. 1999). Measures of ethnolinguistic diversity of populations have been used to instrument for corruption, or institutions more broadly (Mauro 1995). However, legal origin, ethnolinguistic fractionalization, and other instruments (such as latitude and whether a country is landlocked) are also frequently included as control variables in fundamental determinants regressions, especially when checking robustness (Easterly and Levine 2003, 2016). Whether a variable is used as an instrument or included as a control variable is therefore often not consistent across different studies (Bazzi and Clemens 2013). Exogenous control variables enter the instrument set in first-stage regressions (for all endogenous explanatory variables), but if they are relevant control variables this precludes them counting as additional instruments required for identification of the effect of the endogenous fundamental determinant(s).

As well as European settler mortality, the colonization process of different locations yielded natural experiments that have been exploited to provide other plausible instrumentation strategies for institutional quality. Feyrer and Sacerdote (2009) report evidence that current development outcomes for a sample of island colonies are positively associated with the length of time as a colony. They use variations in prevailing wind patterns as instruments for centuries of colonial rule or the first year as a colony. Wind speed and direction were crucial in determining which islands were colonized in the age of sail but would not have a direct effect on their current levels of income per capita or infant mortality.

Iyer (2010) compares development outcomes for Indian states that were under direct British rule compared to indirect rule. The 'Doctrine of Lapse' between 1848 and 1856, whereby the death of native rulers without a natural heir led to direct rule, provides a natural experiment that avoids selection problems. Iyer uses the death of a ruler without a natural heir as an instrument for direct rule and finds states that experienced direct rule have poorer post-colonial development outcomes. Identification is based on the plausible assumption that the death of an heirless maharajah would have no direct effect on modern outcomes.

Olsson and Hibbs (2005) use an index of biogeographic conditions, based on the numbers of domesticable native species of plants and animals in different parts of the world, as an explanatory variable in regressions explaining income per capita and the number of years since the Neolithic transition (from hunter-gatherer to agricultural societies). Ashraf and Galor (2011) subsequently use these biogeographic components as instruments for the timing of the transition in regressions explaining population density and technology levels in years 1, 1000, and 1500. Their findings support Diamond's (1997) arguments on the importance of biogeographical factors for the timing of the Neolithic transition, with an earlier transition leading to positive long-term effects on levels of non-agricultural technology and population density.

Recent studies emphasize the effects of genetic diversity (Ashraf and Galor 2013) and genetic distance (Spolaore and Wacziarg 2009, 2013) on economic development. According to Ashraf and Galor's (2013) 'out of Africa' hypothesis, a settlement's migratory distance from East Africa affects its degree of genetic diversity, which, in turn, has a long-lasting hump-shaped effect on productivity. Because genetic diversity could be endogenous in regressions explaining productivity, they use migratory distance from East Africa as an instrument for genetic diversity.

Overall, considerable imagination and ingenuity have been demonstrated in identifying natural experiments that provide plausible instruments for endogenous regressors in empirical studies of the fundamental determinants of comparative development. This brief review also highlights how justification for the various instrumentation strategies is based primarily on informal economic theory arguments.

3. Instrumental Variables Estimation and Reduced Forms

IV estimation is designed to provide consistent estimates when explanatory variables are endogenous, i.e., correlated with the error term in the structural model. Implementation requires the selection of a set of instruments sufficient to ensure identification. To obtain consistent estimates, the instruments need to be exogenous, i.e., uncorrelated with the error term (at least asymptotically), and relevant, i.e., have high (partial) correlations with the endogenous explanatory variables.

Existing cross-country empirical studies of the fundamental determinants of levels of development can be characterized in the generic framework

$$y_i = \alpha' \boldsymbol{X_i} + \varepsilon_i \quad \varepsilon_i \sim \mathrm{N}(0, \sigma^2) \quad i = 1, 2, \ldots, N \tag{1}$$

where y is, conventionally, the natural logarithm of income per capita (or output per worker) or, for earlier historical dates, population density, and $\boldsymbol{X}_i$ a $m \times 1$ vector of explanatory variables representing the fundamental determinants and relevant control variables.[1] Subscript i denotes observations for country i. $\boldsymbol{X}_i$ can be decomposed as $\left(\boldsymbol{X}'_{1i}\ \boldsymbol{X}'_{2i}\right)'$, where $\boldsymbol{X}_{1i}$ and $\boldsymbol{X}_{2i}$ are, respectively, $m_1 \times 1$ and $m_2 \times 1$ vectors of endogenous and exogenous determinants of income levels, and $\boldsymbol{\alpha}' = (\boldsymbol{\alpha}'_1\ \boldsymbol{\alpha}'_2)$ is an appropriately dimensioned parameter vector. For the stochastic error term, ε_i, this categorization assumes $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_i) \neq 0$ and $\mathrm{E}(\boldsymbol{X}_{2i}\varepsilon_i) = 0$.

To deal with the endogeneity of $\boldsymbol{X}_{1i}$, IV estimation introduces $\boldsymbol{Z}_i$, a $p \times 1$ vector of additional instruments ($p \geq m_1$) that satisfy exclusion restrictions, i.e., are not included in Equation (1). $\boldsymbol{Z}_i$ is assumed to satisfy: (a) $\mathrm{E}(\boldsymbol{Z}_i\varepsilon_i) = 0$; (b) $\mathrm{E}(\boldsymbol{X}_{1i}\boldsymbol{Z}'_i) = \boldsymbol{\Sigma}_{XZ} \neq 0$; and (c) $\mathrm{E}(\boldsymbol{Z}_i\boldsymbol{Z}'_i) = \boldsymbol{\Sigma}_{ZZ} > 0$. The crucial exogeneity requirement in (a), without which IV estimates are not consistent, is essentially non-verifiable because of the unobservable nature of the error term. Implicitly, if $\alpha_1 \neq 0$, it is also assumed that (d) $\mathrm{E}(\boldsymbol{Z}_i y_i) \neq 0$, i.e., the additional instruments need to be correlated with the dependent variable as well as the endogenous explanatory variable(s) (Spanos 2007, p. 38).[2] If the theory-based story motivating the choice of instruments conflicts with the implication that $\boldsymbol{Z}$ needs to be correlated with y, then this is a matter for concern. From this perspective, Angrist and Pischke (2009, p. 213) recommend inspecting the signs and significance of the coefficients on excluded instruments in the reduced form for y, noting that "if you can't see the causal relation of interest in the reduced form, it's probably not there".

---

1 A small minority of studies adopt other measures of development as the dependent variable, either as a complement to examining income per capita, e.g., infant mortality (Feyrer and Sacerdote 2009), or as an alternative, e.g., life expectancy (Knowles and Owen 2010) or output volatility (Malik and Temple 2009).

2 To simplify the notation, observed variables are assumed to have zero means. Assumptions (a)–(d) are the relevant finite-sample conditions; most formal treatments of the properties of IV estimation focus on the corresponding asymptotic conditions: (a)′: $\mathrm{plim}(N^{-1}\boldsymbol{Z}'\boldsymbol{\varepsilon}) = 0$; (b)′: $\mathrm{plim}(N^{-1}\boldsymbol{X}_1{}'\boldsymbol{Z}) = \boldsymbol{\Sigma}_{XZ} \neq 0$; (c)′: $\mathrm{plim}(N^{-1}\boldsymbol{Z}'\boldsymbol{Z}) = \boldsymbol{\Sigma}_{ZZ} > 0$, and (d)′: $\mathrm{plim}(N^{-1}\boldsymbol{Z}'\boldsymbol{y}) \neq 0$, where $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_N)'$, $\boldsymbol{X}_1 = (\boldsymbol{X}_{11}, \boldsymbol{X}_{12}, \ldots, \boldsymbol{X}_{1N})'$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)'$ (Spanos 2007, pp. 37–8).

IV estimation is sometimes characterized as an atheoretical strategy (Deaton 2010; Heckman and Urzúa 2010), in part because only the structural equation of interest, such as Equation (1), is usually specified explicitly. However, exclusion restrictions "are motivated by subject matter, that is economic, rather than statistical, knowledge" (Imbens 2010, p. 403), as is evident from the review in Section 2. The most influential studies in the literature on the fundamental determinants of development (Acemoglu et al. 2001, 2002) are regarded as providing good examples of historical natural experiments generating quasi-random variation in fundamental determinants (Angrist and Pischke 2010; Fuchs-Schuendeln and Hassan 2015). Judgements on the plausibility of identification strategies rely primarily on the plausibility of their a priori theoretical arguments.

Statistical considerations are not entirely ignored. If the equation of interest is overidentified (i.e., $p > m_1$), testing for overidentifying restrictions is commonly implemented. Overidentification tests (Sargan 1958; Hansen 1982) implicitly compare whether alternative sets of just-identified IV estimates, corresponding to different subsets of instruments, are equal (Wooldridge 2010, pp. 134–37). They therefore rely on the *untestable* validity of sufficient of the instruments to obtain at least exact identification; although informative, such tests cannot provide definitive evidence on instrument validity, as non-rejection is possible even if none of the instruments is exogenous.

In contrast, assumptions (b)–(d) can be checked directly using observable sample data, but, as Spanos (2006, p. 48) points out, this is "pitiably inadequate from the statistical viewpoint because there will be thousands of instruments whose sample second moments would seem to satisfy [these requirements]". The implications of using instruments only weakly correlated with the endogenous regressors have received considerable attention. If instruments are weak, IV estimates can be badly biased and their finite-sample distribution may be very different from their asymptotic distribution, even for large samples, distorting the size of tests and the coverage of confidence intervals (Andrews and Stock 2007). However, as Spanos (2007) emphasizes, weak instrumentation is only one of several potential deviations from the underpinning probabilistic assumptions of IV estimation.

A justification for instrument choice based solely (or primarily) on economic theory is not sufficient for valid inference because (a)–(d) are probabilistic conditions that apply to the vector stochastic process of the observable random variables. "[T]heory-based concepts like structural parameters, structural errors, orthogonality and non-orthogonality conditions, gain statistical 'operational meaning' when embedded into a statistical model specified exclusively in terms of the joint distribution of the *observable* random variables involved" (Spanos 2007, p. 39, emphasis in original). In this context, the relevant statistical model, specified in terms of the observable variables, is the full RF, equivalent to the multivariate linear regression (MLR)

$$y_i = \boldsymbol{\beta}_1' \boldsymbol{Z}_i + \boldsymbol{\beta}_2' \boldsymbol{X}_{2i} + u_{1i} \tag{2a}$$

$$\boldsymbol{X}_{1i} = \boldsymbol{B}_1' \boldsymbol{Z}_i + \boldsymbol{B}_2' \boldsymbol{X}_{2i} + \boldsymbol{u}_{2i} \tag{2b}$$

$$\text{with} \begin{pmatrix} u_{1i} \\ \boldsymbol{u}_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_{11} & \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} \right). \tag{2c}$$

Equations (2a) and (2b) are, respectively, the RFs for the dependent variable and endogenous right-hand-side variables. $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ are appropriately dimensioned matrices and vectors of reduced-form parameters. The MLR explicitly considers both the 'first-stage' regression(s) in Equation (2b) and the "now rarely considered regression of the variable of interest on the instrument[s]" (Deaton 2010, p. 428) in Equation (2a).

The MLR/RF provides the implicit framework within which the structural model is embedded. A key insight of Spanos' analysis is that Equation (1), subject to $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_i) \neq 0$, $\mathrm{E}(\boldsymbol{X}_{2i}\varepsilon_i) = 0$ and conditions (a)–(d), is equivalent to imposing restrictions on Equation (3), which is a reparameterized version of the reduced form in Equation (2)

$$y_i = \boldsymbol{\alpha}_0' \boldsymbol{X}_i + \gamma_0' \boldsymbol{Z}_i + \varepsilon_{0i} \tag{3a}$$

$$\boldsymbol{X}_{1i} = \boldsymbol{B}_1' \boldsymbol{Z}_i + \boldsymbol{B}_2' \boldsymbol{X}_{2i} + \boldsymbol{u}_{2i} \tag{3b}$$

$$\text{with} \begin{pmatrix} \varepsilon_{0i} \\ \boldsymbol{u}_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & \boldsymbol{\Omega}_{22} \end{pmatrix} \right). \tag{3c}$$

Spanos proves that imposing the (non-testable) identification restriction $\gamma_0 = 0$ in Equation (3a), in conjunction with $\boldsymbol{B}_1 \neq 0$ (in Equations (2b) and (3b)) and $\boldsymbol{\beta}_1 \neq 0$ (in Equation (2a)), triggers a reparameterization/restriction on the MLR/RF, maintaining $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_i) \neq 0$ (in contrast to $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_{0i}) = 0$ in Equation (3a)) and $\mathrm{E}(\boldsymbol{Z}_i\varepsilon_i) = 0$, and with conditions (b)–(d) holding (Spanos 2007, pp. 42–5).[3] Hence, although $\mathrm{E}(\boldsymbol{Z}_i\varepsilon_i) = 0$ is not directly testable, by embedding structural Equation (1) in the MLR/RF in Equation (2), the conditions $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_i) \neq 0$ and $\mathrm{E}(\boldsymbol{Z}_i\varepsilon_i) = 0$ are 'operationalized' via the reparameterization/restriction on the MLR/RF.[4]

Because the structural model in Equation (1) constitutes a reparameterization/restriction of the statistical model, i.e., the MLR/RF, "the statistical adequacy of the latter ensures the reliability of inference in the context of the former" (Spanos 2007, p. 48). Inference, based on conventional formulae, will be appropriate if the following probabilistic assumptions apply to the MLR/RF in Equations (2a) and (2b) (Spanos 2007, Table 2.2), where D(.) denotes the joint distribution, and $\boldsymbol{y_i} = (y_i, \boldsymbol{X}_{1i}')'$, $\boldsymbol{\Omega}$: is the error covariance matrix in Equation (2c) and $\boldsymbol{\Theta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{B}_1', \boldsymbol{B}_2', \boldsymbol{\Omega}\right)$

Normality	$\mathrm{D}(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i}, \boldsymbol{\Theta})$ is normally distributed	(4a)
Linearity	$\mathrm{E}(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i})$ is linear in $\boldsymbol{Z}_i$ and $\boldsymbol{X}_{2i}$	(4b)
Homoskedasticity	$\mathrm{Var}(\boldsymbol{y_i} \mid \boldsymbol{Z}_i, \boldsymbol{X}_{2i}) = \boldsymbol{\Omega}$ is homoskedastic (free of $\boldsymbol{Z}_i, \boldsymbol{X}_{2i}$)	(4c)
Independence	$(\boldsymbol{y_i} \mid \mathrm{Z}_\mathrm{i}, \mathrm{X}_{2\mathrm{i}})$, $i = 1, 2, \ldots, N$ are independent random variables	(4d)
i-invariance	$\boldsymbol{\Theta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{B}_1', \boldsymbol{B}_2', \boldsymbol{\Omega}\right)$ is constant for all i	(4e)

Assessment of statistical adequacy of the MLR/RF requires testing these assumptions. As well as a potentially destructive role in probing the theory-based selection of instruments, this process is also constructive. If the MLR/RF is misspecified, this suggests a need to respecify the model to take account of statistical information in the data that is not accounted for in the original statistical model (i.e., the original MLR/RF), with any additional (exogenous) variables added to ensure statistical adequacy becoming part of the extended instrument set.

Estimates of the RF parameters can be obtained using ordinary least squares (OLS); if assumptions (4a)–(4e) are valid, then these estimates are unbiased and efficient, even in finite samples. The corresponding 2SLS estimates of the structural parameters in Equation (1) will still be biased because the parameters in Equation (2b) have to be estimated (Hahn and Hausman 2002); however, these estimates will be consistent and, because they are based on a statistically adequate model, will provide a valid basis for inference (as long as any overidentifying restrictions are also valid).

[3] If the structural model is exactly identified ($p = m_1$), this involves a pure reparameterization with a one-to-one correspondence between reduced form and structural parameters. If the structural model is overidentified ($p > m_1$), it involves a reparameterization/restriction; in this case, Equation (3a), despite its 'reduced-form' label, is more general than the structural model in Equation (1).

[4] Details of the mapping between structural and MLR/RF parameters are provided by Spanos (2007, pp. 41–4). Note that the restrictions $\gamma_0 = 0$ are non-testable identifying restrictions imposed together with $\boldsymbol{B}_1 \neq 0$ and $\boldsymbol{\beta}_1 \neq 0$ to identify the structural parameters in $\boldsymbol{\alpha}_1$. Imposing $\gamma_0 = 0$ independent of $\boldsymbol{B}_1 \neq 0$ and $\boldsymbol{\beta}_1 \neq 0$ is equivalent to treating $\boldsymbol{X}_{1\mathrm{i}}$ as exogenous, which leads to a contradiction.

From this perspective, the statistical adequacy of the RF underpins the testing that conventionally occurs in most IV applications, i.e., testing overidentifying restrictions, testing for weak instruments, Hausman-type exogeneity tests and, ultimately, inference on the key parameters of interest in the structural model. The results from such tests are potentially misleading if prior testing reveals the MLR/RF to be misspecified.

Spanos' approach stands in stark contrast to common practice in applications of IV estimation, which treats fitting a linear projection in the first-stage regression in Equation (2b) as no more than a pure predictive exercise. This ignores the fact that the MLR/RF, the statistical model specified in terms of the joint distribution of the observable variables, provides the framework within which the structural equation is embedded and which reflects instrument exogeneity assumptions in the parameterizations for the structural parameters in the context of the MLR/RF.

Standard textbook treatments of 2SLS/IV estimation instead tend to emphasize that the asymptotic properties of 2SLS estimation and inference are retained under weaker assumptions. For example, consistency does not require normality of the errors, instead relying on asymptotic normality of the IV estimator, based on the central limit theorem under the assumption $\mathrm{E}(\boldsymbol{Z}_i\varepsilon_i) = 0$ and finite second-moment assumptions (Wooldridge 2010, Chapter 5). In addition, an asymptotic variance-covariance matrix for IV estimators can be derived assuming the error terms are heteroskedastic (Wooldridge 2010, pp. 106–7); this has led to widespread use of heteroskedastic-robust inference.[5] However, all these results are only asymptotic, i.e., they require large samples for their validity, and the central limit theorems used to derive them rely on untested dependence and heterogeneity restrictions (Spanos 2017).

Leamer (2010) provides a trenchant and compelling critique of over-reliance on methods (including heteroskedastic-robust standard errors) "which promise results without assumptions, as if we were already in Asymptopia where data are so plentiful that no assumptions are needed. But like procedures that rely explicitly on assumptions, these new methods work well in the circumstances in which explicit or hidden assumptions hold tolerably well and poorly otherwise" (Leamer 2010, p. 32). In the final reckoning, inference is based solely on the N data points available, so its reliability will depend on the approximate validity of such probabilistic assumptions with respect to the sample data (Spanos 2017). An appeal to asymptotic normality, for example, would be problematic if the relevant sampling distributions of the estimators are not approximately normal for the N available data points. From this perspective, "limit theorems 'as [N] tends to infinity' are logically devoid of content about what happens at any particular [N]. All they can do is suggest certain approaches whose performance must then be checked on the case at hand." (Le Cam 1986, p. xiv). In the fundamental determinants regressions examined in Section 5, typical sample sizes range from $N = 21$ to less than 100, with, for example, about 64 observations if Acemoglu et al.'s (2001) settler mortality instrument is used. In Leamer's (2010, p. 43) terms, this is distinctly "in the Land of the Finite Sample, infinitely far from Asymptopia", so that reliance on asymptotic results requires a good deal of faith. Particularly in such contexts, inference based on a statistical framework subject to a comprehensive set of explicit and non-rejected assumptions is more appealing than relying on asymptotic properties that depend on a weaker set of implicit and untested (or untestable) assumptions.

Overall, the bottom line in Spanos' approach is that instrument choice cannot be based solely on theoretical considerations (including the design of natural experiments) but also has an important statistical dimension, i.e., testing for the statistical adequacy of the statistical model, the underlying

[5] Although rarely discussed explicitly in the fundamental determinants literature, some applied researchers also weaken the linearity assumption, interpreting linear regression as providing a best linear predictor that approximates some nonlinear conditional expectation function (CEF) of the observable variables (Angrist and Pischke 2009). On this interpretation, how well the regression line fits the nonlinear CEF will vary with the values of the explanatory variables, with the consequent residual heteroskedasticity regarded as a natural feature of this characterization. This further shifts the focus to reliance on asymptotic results and heteroskedastic-robust inference.

MLR/RF, which depends on both the specification of the structural model and the instrumentation strategy. In most fundamental determinants studies, the full RF is not usually explicitly reported; some studies report the first-stage regressions for the endogenous explanatory variable(s), i.e., $\boldsymbol{X}_1$, but the corresponding reduced form for y is rarely reported. Testing for misspecification of the RF (or, indeed, the structural equation) is not evident in any of the studies. Emphasis on the statistical adequacy of the RF is consistent with Deaton's (2010, p. 435) broader argument that "the reduced form ... contains substantive information about the relationship between growth and the instruments. ... direct consideration of the reduced form is likely to generate productive lines of enquiry."

4. Testing Statistical Adequacy

Models in the fundamental determinants literature are highly parsimonious. They vary in terms of what is included in $\boldsymbol{X}$, which explanatory variables are assumed to be endogenous (i.e., in $\boldsymbol{X}_1$), and the additional instruments included in $\boldsymbol{Z}$. Brock and Durlauf (2001) emphasize that growth theories are 'open-ended', i.e., the relevance of one growth determinant does not normally preclude the relevance of other potential determinants. This makes choosing relevant instruments difficult; the risk of potential omitted variables, arising from the parsimonious nature of the models, and the likely correlations between these omitted variables and the instruments cast doubt on the exogeneity assumption for the instruments. Because this assumption is not directly testable, more emphasis on assessment of the statistical adequacy of the embedding statistical model of the observable variables may provide useful insights into the validity of the overall model/instrumentation combinations.

In general, this literature places little emphasis on reporting evidence on statistical adequacy. For example, over 200 regression models are fitted in the studies by Acemoglu et al. (2001), Easterly and Levine (2003), and Rodrik et al. (2004), but the only diagnostic test reported is a test for overidentifying restrictions and the null is rejected for few of the different model/instrument combinations considered. However, the sampling distribution of an overidentification test, and hence its size and power, depends on the statistical adequacy of the statistical model, so that misspecification of the RFs would potentially undermine the reliability of such a test. Instead of misspecification testing, the response to model/instrument uncertainty is to conduct a robustness analysis by adding control variables, singly or in sets, to regressions that include the key explanatory variable(s) of interest. Without explicit misspecification testing, there is no guarantee that all—or indeed any—of these models are statistically adequate.[6]

Inference, based on conventional formulae, will be appropriate if probabilistic assumptions (4a)–(4e) apply to the MLR/RF, which forms the relevant statistical model specified in terms of the observable random variables, and if overidentification restrictions are also satisfied. These assumptions, which underwrite the validity of inference with IV estimation, are tested for each of the replicated studies. As noted in Section 3, common practice regards assumptions (4a) and (4c) as not necessary, but instead relies on weaker assumptions to deliver consistent point estimates and asymptotic standard errors. For example, heteroskedasticity is widely regarded as a natural feature of cross-sectional data and use of heteroskedastic-consistent standard errors (White 1980), as a default, is common in the fundamental determinants literature, both for OLS and IV estimates. However, such corrections are valid only asymptotically and their finite-sample properties can be unsatisfactory

6 One response to concerns about validity of underlying statistical assumptions is the development and application of Generalized Method of Moments (GMM) estimation, which requires less restrictive assumptions. However, as Spanos (2015, p. 183) argues, this comes at a price: "weaker premises will always give rise to less precise inferences without any guarantee that they will be more adequate for the particular data, especially when the inference is unduly reliant on asymptotics ... Even if one has to rely on asymptotic results, the adequacy of the premises renders such results a lot more reliable for the given n. In contrast, asymptotic properties such as [consistent and asymptotically normal], stemming from nonvalidated premises, provide no guarantee for reliable inferences in practice." In any case, all the studies examined rely on 2SLS estimation, applied to relatively small samples, fitting simple linear-in-parameters models with additive errors and constant parameters across countries.

(MacKinnon and White 1985; Angrist and Pischke 2009, Chapter 8). Although degrees of freedom adjustments can yield heteroskedasticity-robust standard errors with improved finite-sample properties (MacKinnon 2013; Imbens and Kolesár 2016), it is important to appreciate that *residual* heteroskedasticity can be a symptom of model misspecification (e.g., neglected nonlinearity or heterogeneity) rather than heteroskedastic *errors* (Zietz 2001; Hendry and Nielsen 2007, pp. 133–34; Sims 2010; King and Roberts 2015), and such statistical misspecification does not, in general, disappear as N goes to infinity. Widespread use of standard-error corrections has tended to lead to this being ignored. Moreover, ignoring departures from normality and relying on heteroskedastic-consistent standard errors can lead to sizeable distortions in size and power of 'robust' F-tests (Spanos and Reade 2016).

Table 1 summarizes the misspecification tests applied to the models examined. These correspond to tests of assumptions (4a)–(4e), i.e., normality (denoted *Norm*), functional form (*RESET*), absence of heteroskedasticity (*Hetero* and *HeteroX*), lack of spatial dependence (Moran's I and $LM_{\rho\lambda}$), and parameter constancy. For ease of evaluating test results, subsequent tables report p-values for the diagnostic tests, with p-values less than 0.05 in bold. Given the MLR nature of the RFs, system misspecification tests, multivariate equivalents of the single-equation tests (with the suffix *Vec*), are also reported. Some tests have power against other alternatives, e.g., functional form misspecification may also be reflected in rejection of the normality and homoskedasticity tests and apparent parameter non-constancy.

With cross-country data, lack of independence (failure of assumption (4d)) is likely to manifest itself as spatial dependence, where 'spatial' may be interpreted broadly to involve socio-economic as well as geographical distance. Surprisingly, relatively few studies (e.g., Moreno and Trehan 1997; Conley and Ligon 2002; Ertur and Koch 2007) have explored spatial dependence in economic growth and development arising, for example, from cross-country spillovers in the growth process. To test for spatial dependence, p-values for Moran's I statistic (Moran 1948; Cliff and Ord 1973) and a Lagrange Multiplier (LM) test (Anselin et al. 1996) applied to the residuals of the fitted RFs are reported. Both have asymptotic distributions under the null, but have reasonable small-sample properties (Anselin and Florax 1995; Anselin et al. 1996). These tests require specification of an a priori weights matrix based on plausible assumptions about the extent of potential spatial linkages. The results reported are for economic distance, measured as a negative exponential function of geographical distance between countries i and j based on latitude and longitude (d_{ij}) and on the development proxy (y) used in each study.[7] Elements of the spatial weights matrix are defined as $W_{ij} = y_i y_j \exp(-\beta d_{ij})$ with $\beta = 0.25$ (unless otherwise indicated) and are row-standardized, so that each row's weights sum to one (Fingleton and Gallo 2008).

The parameters in $\boldsymbol{B}_1$, $\boldsymbol{B}_2$, $\boldsymbol{\alpha}$, and hence $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, are (usually implicitly) assumed to be invariant to i, as in assumption (4e). Parameter constancy is explored by recursive graphical analysis of coefficient estimates for the variables in the RF and of break-point Chow tests at different points in the sample (Hendry and Nielsen 2007, pp. 195–97).[8] The normality, heteroskedasticity, RESET, and spatial dependence tests are invariant to the ordering of the data, as are the full-sample coefficient estimates. However, the ordering of the data can affect the recursive plots and Chow tests, unless assumption (4d) holds and randomly selected cross-section observations are completely independent

---

7 This choice is consistent with Conley and Ligon's (2002) finding of positive spillovers of GDP per capita on neighbours' growth performance. Qualitatively similar results are obtained if the study's main explanatory variable, e.g., institutional quality, is used as the economic variable in the weighting scheme. Latitude and longitude data are from CEPII's database of geographical variables (Mayer and Zignago 2011). The spatial weights matrices are constructed using spwmatrix and the tests computed using anketest (Jeanty 2010), both Stata routines written by Wilner Jeanty.

8 With parameter constancy, the sequence of coefficient estimates should stabilize, with no sharp breaks, as N increases; the ideal is to be able to see, from left to right, through the 'tunnel' formed by the narrowing standard error bands. Qin et al. (2016), for example, provide a recent example of the usefulness of recursive plots as a diagnostic device. In the recursive graphs, the Chow test statistic values are scaled by the relevant critical values from the F-distribution at the 1% significance level; scaled test values greater than unity in the graphs (represented by the dotted line) therefore indicate statistical significance at the 1% level.

(Hendry 2009, p. 31). The lack of a unique natural ordering for cross-sectional data does not imply that heterogeneity and dependence are not a concern with such data; often several natural orderings are worth considering.[9] The recursive plots for the coefficient estimates and the Chow tests reported in Section 5 are based on the observations ordered by the size of the development proxy, income per capita, or population density. Results are summarized in the tables by indicating parameter non-constancy (NC) or constancy (C); where the classifications are marginal, such cases are labelled as 'C/NC'. If estimates of the parameters apparently vary with i, this may be indicative of outliers or model misspecification, e.g., omitted variables, or hidden dependence in the cross-sectional observations.[10]

This approach involves multiple testing of different hypotheses. Multiple testing increases the Type 1 error probability of the overall testing procedure; for example, with R tests and a significance level of α for each test, if all the null hypotheses are valid and the degree of dependence between the tests is unknown, the Bonferroni inequality implies the probability of rejecting one or more of the null hypotheses is $\leq R\alpha$ (Hendry 1995, pp. 490–1). Focusing on $R = 5$ key diagnostic tests (*Norm*, *Hetero*, *HeteroX*, *RESET* and Moran's I), the upper bound, $R\alpha$ equals 0.25 for $\alpha = 0.05$, and 0.05 for $\alpha = 0.01$. Spanos (2010, 2017) considers and dismisses criticisms directed against misspecification testing, such as double use of data and pre-test bias. He emphasizes the difference between misspecification testing of different dimensions and combinations of assumptions, involving testing outside the boundaries of the specified statistical model, compared to multiple testing of hypotheses within the boundaries of the statistical model. The distributional assumptions and reported p-values of the misspecification tests are based on a common null hypothesis that all the assumptions of the statistical model are valid, so that rejections, especially rejections for more than one test for the same model, may not provide a clear guide to the direction of required respecification (Hendry and Nielsen 2007, p. 135; Spanos 2017). The diagnostics are therefore interpreted holistically as an overall check of statistical adequacy.[11]

Tests for overidentification and weak instruments are also reported although, as previously noted, their validity is conditional on the statistical adequacy of the RFs. *Sargan-p* is the p-value for Sargan's (1958) test.[12] *CD-F* is the F-statistic form of Cragg and Donald's (1993) test for weak instruments, which is compared to Stock and Yogo's (2005) critical values based on a maximal size of 15%; entries in bold correspond to non-significant values of *CD-F*, indicative of weak instruments. Also reported are the partial R^2s between the endogenous regressors and the additional instruments, and, where relevant, Shea (1997) partial R^2s, which take into account intercorrelations between the instruments and tend to be notably smaller than standard partial R^2s if instruments are weak.

9 As Hendry (2009, p. 31) emphasizes, "[s]uitable tests for the absence of dependence would seem essential before too great weight is placed on results that [are based] on the claim of random sampling, especially when the units are large entities like countries."

10 Although examination of the different potential sources of parameter non-constancy in the individual models is beyond the scope of the current paper, identification of influential observations and outliers using, for example, jackknife estimation and associated DFBETAS (Belsley et al. 2005) could provide additional insights on the sensitivity of coefficient estimates to individual observations.

11 As an alternative to separate tests of different assumptions, Spanos (2017) suggests combining terms representing departures from several assumptions in a single auxiliary regression and jointly testing the potential violations.

12 Several studies report Hansen's (1982) J statistic, which is consistent in the presence of heteroskedasticity. However, in almost all cases, this makes no qualitative difference to the results.

Table 1. Diagnostic tests for misspecification.

Test	Reference	Null	Description
Norm *NormVec*	Doornik and Hansen (2008)	Normality	$Norm = z_1^2 + z_2^2 \underset{app}{\sim} \chi^2(2)$ under the null, where z_1 and z_2 are transformed skewness and kurtosis measures correcting for finite-sample dependence between sample skewness and kurtosis (computed using OxMetrics 7 (Doornik and Hendry 2013a, p. 276)). *NormVec* is the multivariate equivalent (Doornik and Hansen 2008, Section III) (computed using OxMetrics 7 (Doornik and Hendry 2013b, p. 227)).
Hetero *HeteroVec*	White (1980)	Homoskedasticity	Degrees-of-freedom-adjusted F-approximation to an asymptotically distributed χ^2 test statistic under the null, calculated as NR^2 from an auxiliary regression of the squared residuals on a constant, the original regressors, and their squares (computed using OxMetrics 7 (Doornik and Hendry 2013a, p. 277)). *HeteroVec* is obtained from a multivariate regression of all error variances and covariances on the original regressors and their squares (computed using OxMetrics 7 (Doornik and Hendry 2013b, p. 227)).
HeteroX *HeteroXVec*	White (1980)	Homoskedasticity	As for *Hetero*, but also including cross-products of the regressors in the auxiliary regression; reported only if there are sufficient observations (computed using OxMetrics 7 (Doornik and Hendry 2013a, p. 277)). *HeteroXVec* is the multivariate equivalent (computed using OxMetrics 7 (Doornik and Hendry 2013b, p. 227)).
RESET *RESETVec*	Ramsey (1969)	Correct functional form	Includes squares and cubes of the fitted values from the original regression as additional regressors. Under the null of zero coefficients on these additional regressors, the F-test is approximately F distributed (computed using OxMetrics 7 (Doornik and Hendry 2013a, p. 278)). *RESETVec* is the multivariate equivalent (computed using OxMetrics 7).
Moran's I	Moran (1948)	Lack of spatial autocorrelation	$I = (e'We/S)/(e'e/N)$, where e is a vector of OLS residuals, W is a spatial weights matrix and $S = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}$. A standardized version of I is approximately normally distributed under the null of no spatial autocorrelation (Cliff and Ord 1973) (computed using Jeanty's (2010) anketest routine in Stata 14.2).
$LM_{\rho\lambda}$	Anselin et al. (1996)	Lack of spatial autocorrelation	$LM_{\rho\lambda}$ is a joint test of lack of spatial error and spatial lag dependence and is asymptotically $\chi^2(2)$ distributed under the null of absence of both spatial error and spatial lag dependence ((Anselin et al. 1996, Equation (15))) (computed using Jeanty's (2010) anketest routine in Stata 14.2).
Parameter Constancy	Doornik and Hendry (2013b)	i-invariance of parameters	Recursive graph of estimated coefficients $\hat{\beta}_{ji} \pm 2\mathrm{se}(\hat{\beta}_{ji})$ for coefficient j for $i = 1, \ldots, M$, with M increasing to N. Sequences of break-point Chow tests assessing whether model based on first M observations yields good forecasts of the remaining observations ((Doornik and Hendry 2013a, Equation (18.2))) (computed using OxMetrics 7 (Doornik and Hendry 2013a, p. 264–5)).

5. Results

The criteria for selecting studies for replication are influence, representativeness and ready availability of the relevant data (from authors' and journals' websites; see Appendix A). The studies examined include those by Hall and Jones (1999), Acemoglu et al. (2001), Easterly and Levine (2003), Sachs (2003), Ashraf and Galor (2011), and Ashraf and Galor (2013). Illustrative models from other key studies by Spolaore and Wacziarg (2009), Putterman and Weil (2010), and Easterly and Levine (2016), reported by Spolaore and Wacziarg (2013) in their review article, are also replicated.

Table 2 provides a summary of the replicated structural equations from the selected studies, listing the dependent and explanatory variables in the models, the classification of explanatory variables as endogenous or exogenous, additional instruments included in the instrument set for IV estimation, and any diagnostic tests reported. In general, point estimates obtained for structural parameters and, where reported, parameters in first-stage regressions match those in the original studies with a few exceptions. Standard errors (usually heteroskedasticity-robust) in some instances differed in the second decimal place, but this has little effect on the interpretation of significance. The replicated results differ for some of the models from Acemoglu et al. (2001), for which there are marginal differences in point estimates and standard errors, especially for the model in their Table 5, column (9). However, there are no substantive changes in the conclusions and the replicated models used for testing for statistical adequacy are obtained directly from the data and Stata program files obtained from https://economics.mit.edu/faculty/acemoglu/data/ajr2001. The results for the two models replicated from Sachs (2003) differ slightly from the original results due to the inclusion of marginally updated data on the malaria variable in the data file obtained.

Table 3 reports a summary of the results of the misspecification tests applied to the set of RFs for each model, in order to provide an overview of the general pattern of test results, which are presented in more detail in subsequent tables. Entries in the test columns show the proportion of tests that reject the relevant null hypothesis at the 5% significance level. For each structural equation, the number of tests varies depending on the number of RFs (for the dependent variable and each of the endogenous explanatory variables) and the number of tests corresponding to each test type (including tests for individual RF equations and multivariate tests, so there is a degree of overlap between the individual-RF and multivariate tests for some assumptions). For the parameter constancy tests, classifications judged as marginal, denoted 'C/NC' in subsequent tables, are conservatively treated as consistent with non-rejection, although a clear pattern of parameter constancy is relatively rare. To provide a somewhat arbitrary visual overview, dark-shaded (red) cells with bold entries represent rejections in two-thirds or more of the misspecification tests of the relevant type for that set of RFs; cells lightly shaded (orange) represent proportions of rejections in the range between one-third (inclusive) and less than two-thirds. Despite the conservative treatment of the parameter constancy classifications, the preponderance of rejections in the spatial dependence and parameter constancy columns is striking.

Hall and Jones (1999), in their main model explaining $\ln(Y/L)$, the natural logarithm of output per worker, include 'social infrastructure' (*SocInf*) as the sole explanatory variable. This contains two equally weighted components: an index of the quality of institutions ('government antidiversion policies', *GADP*) and Sachs and Warner's (1995) measure of the degree of trade openness (*YrsOpen*). They use absolute latitude (*AbsLat*), the fraction of the population speaking one of the five major Western European languages as their first language (*EurFrac*), the fraction speaking English as their first language (*EngFrac*) and Frankel and Romer's (1999) (natural logarithm of) predicted trade share (based on a trade model including exogenous gravity variables) ($\ln FR$) as instruments for *SocInf*. Results of diagnostic testing of the RFs are reported in Table 4, columns (1) and (2), for a representative model (Hall and Jones 1999, Table II, row 3). Heteroskedasticity is evident in the residuals of the fitted RF for $\ln(Y/L)$ and there is some evidence of parameter non-constancy, especially for the coefficient on *AbsLat*. For the RF for *SocInf* there is evidence of non-normality of the errors, functional form misspecification and parameter non-constancy, as can be seen in the recursive plots in Figure 1. Lack of spatial dependence is also strongly rejected.

Table 2. Summary of replicated models.

Study	Table/Row or Column	Estimation Method (*N*)	DependentVariable	Explanatory Variables (Endogenous Variables in Bold)	Additional Instruments	Diagnostics
Hall and Jones (1999)	Table II, row 3	2SLS (*N* = 79)	ln(*Y*/*L*)	***SocInf***	*AbsLat*, *EurFrac*, *EngFrac*, ln*FR*	OverID
Hall and Jones (1999)	Table II, row 3#	2SLS (*N* = 79)	ln(*Y*/*L*)	***GADP*, *YrsOpen***	*AbsLat*, *EurFrac*, *EngFrac*, ln*FR*	
Acemoglu et al. (2001)	Table 4, column 2	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, *AbsLat*,	ln*SM*	
Acemoglu et al. (2001)	Table 4, column 8	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, *AbsLat*, continent dvs	ln*SM*	
Acemoglu et al. (2001)	Table 5, column 6	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, *AbsLat*, *FrLO*	ln*SM*	
Acemoglu et al. (2001)	Table 5, column 7	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, Religion variables	ln*SM*	
Acemoglu et al. (2001)	Table 5, column 8	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, *AbsLat*, Religion variables	ln*SM*	
Acemoglu et al. (2001)	Table 5, column 9	2SLS (*N* = 64)	ln*GDPpc*	***AvExpr***, *AbsLat*, *FrC*, *FrLO*, Religion variables	ln*SM*	
Easterly and Levine (2003)	Table 4, row 4	2SLS (*N* = 72)	ln*GDPpc*	***Inst***, *FrLO*, Religion variables, *EthDiv*	ln*SM*, *AbsLat*, *Landlock*	OverID, First-stage *F*
Easterly and Levine (2003)	Table 4, row 6	2SLS (*N* = 72)	ln*GDPpc*	***Inst***, *FrLO*, Religion variables, *EthDiv*, *Oil*	ln*SM*, *AbsLat*, *Landlock*, Crops/minerals dvs	OverID, First-stage *F*
Easterly and Levine (2003)	Table 4, row 6##	2SLS (*N* = 72)	ln*GDPpc*	***Inst***, *FrLO*, Religion variables, *EthDiv*, *Oil*	ln*SM*, *AbsLat*, *Landlock*	na
Easterly and Levine (2003)	Table 5, row 4	2SLS (*N* = 70)	ln*GDPpc*	***Inst***, *YrsOpen*, *FrLO*, Religion variables, *EthDiv*	ln*SM*, *AbsLat*	OverID
Sachs (2003)	Table 1, column 10	2SLS (*N* = 69)	*lcgdp*95	***Rule*, *Mal94p***	ln*SM*, *KGPTemp*, *ME*	
Sachs (2003)	Table 1, column 12	2SLS (*N* = 69)	*lcgdp*95	***Rule*, *Malfal***	ln*SM*, *KGPTemp*, *ME*	
Ashraf and Galor (2011)	Table 2, column 6	2SLS (*N* = 96)	*lpd*1500	***lyst***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Ashraf and Galor (2011)	Table 3, column 6	2SLS (*N* = 94)	*lpd*1000	***lyst***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Ashraf and Galor (2011)	Table 4, column 6	2SLS (*N* = 83)	*lpd*1	***lyst***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Spolaore and Wacziarg (2013)	Table 2, column 4	2SLS (*N* = 98)	*lpd*1500	***lyst***, *AbsLat*, *LandTropics*, *Landlock*, *Island*	*Plants*, *Animals*	
Ashraf and Galor (2011)	Table 8, column 3	2SLS (*N* = 93)	*natech*1*K*	***lyst***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Ashraf and Galor (2011)	Table 8, column 6	2SLS (*N* = 93)	*natech*1	***lyst***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Ashraf and Galor (2011)	Table 9, column 3	2SLS (*N* = 92)	*lpd*1000	***tech*1*K***, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Ashraf and Galor (2011)	Table 9, column 6	2SLS (*N* = 83)	*lpd*1	***tech*1**, ln(*AbsLat*), ln(*LandProd*), *distcr*, *Land100km*, continent dvs	*Plants*, *Animals*	OverID, First-stage *F*
Spolaore and Wacziarg (2013)	Table 5, column 2	OLS (*N* = 148)	*lpci*05	*AdjYrsAg*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 5, column 4	OLS (*N* = 135)	*lpci*05	*AdjStateHist*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 6, column 3	OLS (*N* = 147)	*lpci*05	*AdjYrsAg*, *EuroShare*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 6, column 4	OLS (*N* = 134)	*lpci*05	*AdjStateHist*, *EuroShare*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 6, column 5	OLS (*N* = 149)	*lpci*05	*EuroShare*, *WtGenDist*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 7, column 1	OLS (*N* = 155)	*lpci*05	*IndGenDist*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 7, column 2	OLS (*N* = 154)	*lpci*05	*WtGenDist*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Spolaore and Wacziarg (2013)	Table 7, column 3	OLS (*N* = 149)	*lpci*05	*WtGenDist*, *EuroShare*, *AbsLat*, *LandTropics*, *Landlock*, *Island*	na	
Ashraf and Galor (2013)	Table 2, column 5	2SLS (*N* = 21)	*lpd*1500	***Div*, *DivSq***, ln(*AbsLat*), *lyst*, ln(*Arable*), ln(*AgSuit*)	*mdistAddis*, *divhatsq*	
Ashraf and Galor (2013)	Table 2, column 6	2SLS (*N* = 21)	*lpd*1500	***Div*, *DivSq***, ln(*AbsLat*), *lyst*, ln(*Arable*), ln(*AgSuit*), continent dvs	*mdistAddis*, *divhatsq*	

Dependent variables: ln(*Y*/*L*) is log of output per worker in 1988; ln*GDPpc* is log of GDP per capita in 1995; *lcgdp*95 is log of GDP per capita in 1995 (Rodrik et al. 2004); *lpd*1500, *lpd*1000 and *lpd*1 are, respectively, log of population density in years 1500, 1000 and 1; *natech*1*K* and *natech*1 are, respectively, a non-agricultural technology index in 1000 and 1; *lpci*05 is log of per capita income in 2005; *Endogenous explanatory variables*: *SocInf* is a measure of 'social infrastructure' with two equally weighted components: *GADP* (government anti-diversion policies) and *YrsOpen* (Sachs and Warner's (1995) openness measure); *AvExpr* is average protection against expropriation risk (1985–1995); *Inst* is the average of six World Bank Governance Indicators; *Rule* is a Rule of Law index; *Mal94p* is the proportion of the population at risk of malaria transmission in 1994; *Malfal* is the proportion at risk of falciparum malaria transmission; *lyst* is log of years since the Neolithic transition; *tech*1*K* and *tech*1 are, respectively, a technology index in 1000 and 1; *Div* is genetic diversity and *DivSq* is its square; *Exogenous explanatory variables and additional instruments*: *AbsLat* is distance from the equator; *EurFrac* is the fraction of the population speaking one of five major Western European languages as their first language; *EngFrac* is the fraction speaking English as their first language; ln*FR* is Frankel and Romer's (1999) (natural log of) predicted trade share; ln*SM* is log of European settler mortality; continent dummy variables (dvs); *FrLO* is a French legal origin dummy; *FrC* is a French colonial dummy; Religion variables (%s Catholic, Muslim, 'Other'); *EthDiv* is ethnolinguistic diversity; *Oil* is an oil-producer dummy; *Landlock* is a dummy for countries with no access to the sea; Crops/minerals dvs is a set of dummies for whether the country has ever had bananas, coffee, copper, maize, millet, rice, rubber, silver, sugarcane, or wheat; *KGPTemp* is the share of the population in temperate ecozones; *ME* is an index of malarial ecology based on temperature, mosquito abundance and vector specificity; ln(*LandProd*) is log of land productivity; *distcr* is the mean distance to nearest coast or river; *Land100km* is percentage of land within 100 km of coast or river; *LandTropics* is the percentage of land area in the tropics; *Island* is an island dummy; *Plants* is the number of domesticable species of plants prehistorically native to relevant landmass and *Animals* is the corresponding number of domesticable species of animals; *AdjYrsAg* is ancestry-adjusted years of agriculture; *AdjStateHist* is ancestry-adjusted state history; *EuroShare* is the share of dependants of Europeans; *WtGenDist is* F_{ST} ancestry-adjusted weighted genetic distance to the US; *IndGenDist is* F_{ST} genetic distance to the US (1500 match); ln(*Arable*) is log percentage of arable land; ln(*AgSuit*) is log land suitability for agriculture; *mdistAddis* is migratory distance from East Africa; *divhatsq* is predicted genetic diversity squared (based on a regression of genetic diversity on migratory distance and all second-stage control variables). Detailed variable definitions are provided in the original studies; # and ## denote amended versions of the models in the original studies (see main text); *OverID* denotes that an overidentification test is reported; 'na' denotes not applicable.

Table 3. Summary of misspecification test results.

Study	Table/Row or Column	RF Dependent Variables	Normality	Heteroskedasticity	RESET	Spatial Dependence	Parameter Constancy
Hall and Jones (1999)	Table II, row 3	ln(Y/L), *SocInf*	1/3	2/6	**2/3**	**4/4**	**2/2**
Hall and Jones (1999)	Table II, row 3#	ln(Y/L), *GADP*, *YrsOpen*	1/4	2/8	2/4	**5/6**	**3/3**
Acemoglu et al. (2001)	Table 4, column 2	ln*GDPpc*, *AvExpr*	1/3	0/6	1/3	**4/4**	1/2
Acemoglu et al. (2001)	Table 4, column 8	ln*GDPpc*, *AvExpr*	1/3	2/6	1/3	**3/4**	1/2
Acemoglu et al. (2001)	Table 5, column 6	ln*GDPpc*, *AvExpr*	**2/3**	0/6	0/3	**4/4**	1/2
Acemoglu et al. (2001)	Table 5, column 7	ln*GDPpc*, *AvExpr*	1/3	1/6	**2/3**	**4/4**	1/2
Acemoglu et al. (2001)	Table 5, column 8	ln*GDPpc*, *AvExpr*	1/3	2/6	1/3	**4/4**	1/2
Acemoglu et al. (2001)	Table 5, column 9	ln*GDPpc*, *AvExpr*	0/3	1/6	0/3	**4/4**	1/2
Easterly and Levine (2003)	Table 4, row 4	ln*GDPpc*, *Inst*	0/3	1/6	**3/3**	**3/4**	1/2
Easterly and Levine (2003)	Table 4, row 6	ln*GDPpc*, *Inst*	0/3	0/6	**2/3**	**3/4**	1/2
Easterly and Levine (2003)	Table 4, row 6##	ln*GDPpc*, *Inst*	0/3	1/6	**2/3**	**3/4**	**2/2**
Easterly and Levine (2003)	Table 5, row 4	ln*GDPpc*, *Inst*	0/3	0/6	1/3	2/4	1/2
Sachs (2003)	Table 1, column 10	*lcgdp*95, *Rule*, *Mal94p*	1/4	3/8	2/4	**4/6**	**2/3**
Sachs (2003)	Table 1, column 12	*lcgdp*95, *Rule*, *Malfal*	1/4	2/8	2/4	**4/6**	**2/3**
Ashraf and Galor (2011)	Table 2, column 6	*lpd*1500, *yst*	**2/3**	3/6	0/3	**4/4**	**2/2**
Ashraf and Galor (2011)	Table 3, column 6	*lpd*1000, *yst*	**2/3**	3/6	1/3	**4/4**	**2/2**
Ashraf and Galor (2011)	Table 4, column 6	*lpd*1, *yst*	**3/3**	1/6	0/3	**4/4**	1/2
Spolaore and Wacziarg (2013)	Table 2, column 4	*lpd*1500, *yst*	**2/3**	**6/6**	**2/3**	**4/4**	1/2
Ashraf and Galor (2011)	Table 8, column 3	*natech*1*K*, *yst*	**3/3**	**4/6**	1/3	**4/4**	1/2
Ashraf and Galor (2011)	Table 8, column 6	*natech*1, *yst*	**2/3**	2/6	0/3	**4/4**	0/2
Ashraf and Galor (2011)	Table 9, column 3	*lpd*1000, *tech*1*K*	**2/3**	**5/6**	**3/3**	**4/4**	1/2
Ashraf and Galor (2011)	Table 9, column 6	*lpd*1, *tech*1	1/3	**4/6**	0/3	**4/4**	**2/2**
Spolaore and Wacziarg (2013)	Table 5, column 2	*lpci*05	0/1	0/2	0/1	**2/2**	**1/1**
Spolaore and Wacziarg (2013)	Table 5, column 4	*lpci*05	0/1	0/2	0/1	**2/2**	0/1
Spolaore and Wacziarg (2013)	Table 6, column 3	*lpci*05	0/1	0/2	**1/1**	**2/2**	**1/1**
Spolaore and Wacziarg (2013)	Table 6, column 4	*lpci*05	0/1	0/2	0/1	1/2	0/1
Spolaore and Wacziarg (2013)	Table 6, column 5	*lpci*05	0/1	**2/2**	0/1	**2/2**	**1/1**
Spolaore and Wacziarg (2013)	Table 7, column 1	*lpci*05	0/1	0/2	0/1	**2/2**	**1/1**
Spolaore and Wacziarg (2013)	Table 7, column 2	*lpci*05	0/1	0/2	0/1	**2/2**	**1/1**
Spolaore and Wacziarg (2013)	Table 7, column 3	*lpci*05	0/1	**2/2**	0/1	**2/2**	**1/1**
Ashraf and Galor (2013)	Table 2, column 5	*lpd*1500, *Div*, *DivSq*	0/4	1/4	1/4	0/6	0/3
Ashraf and Galor (2013)	Table 2, column 6	*lpd*1500, *Div*, *DivSq*	2/4	0/3	1/4	2/6	0/3

See Table 2 for definitions of variables. # and ## denote amended versions of the models in the original studies (see main text). Entries in the test columns show the proportion of tests that reject the relevant null hypothesis at the 5% significance level (including tests for individual RF equations and multivariate tests). Parameter constancy tests judged as C/NC are treated as consistent with non-rejection. Dark-shaded (red) cells represent rejections in $\geq 2/3$ of the relevant tests; cells lightly shaded (orange) represent $1/3 \leq$ proportion of rejections $< 2/3$.

Table 4. Testing statistical adequacy of RFs for Hall and Jones (1999).

Test	(1)		(2)	(3)		(4)
	Table II, Row 3			*SocInf* Components		
	ln(Y/L)		*SocInf*	*GADP*		*YrsOpen*
Norm-p	0.285		**0.046**	0.782		**0.001**
NormVec-p		0.953			0.632	
Hetero-p	**0.022**		0.774	0.613		0.791
HeteroVec-p		0.147			0.179	
HeteroX-p	**0.021**		0.760	0.397		0.941
HeteroXVec-p		0.071			0.208	
RESET-p	0.114		**0.011**	**0.000**		0.180
RESETVec-p		**0.000**			**0.000**	
Moran's *I-p*	**0.000**		**0.001**	**0.000**		**0.009**
$LM_{\rho\lambda}$-*p*	**0.002**		**0.017**	**0.003**		0.080
Parameter Constancy	NC		NC	NC		NC
R^2	0.614		0.328	0.535		0.167
N		79			79	
Sargan-p			0.232		0.151	
CD-F			**9.028**		**0.488**	
Partial R^2			0.328	0.535		0.167
Shea partial R^2			0.328	0.084		0.026

Dependent variables: ln(Y/L) is log of output per worker in 1988; *SocInf* is a measure of 'social infrastructure', made up of two equally weighted components: *GADP* (government anti-diversion policies) and *YrsOpen* (Sachs and Warner's (1995) measure of openness). Instrument set in each column (all additional instruments): *AbsLat*, *EurFrac*, *EngFrac*, and ln*FR* (defined in the text and notes to Table 2). See Table 1 for explanation of tests; suffix '*p*' denotes *p*-value. $\beta = 0.25$ in the spatial weighting matrix.

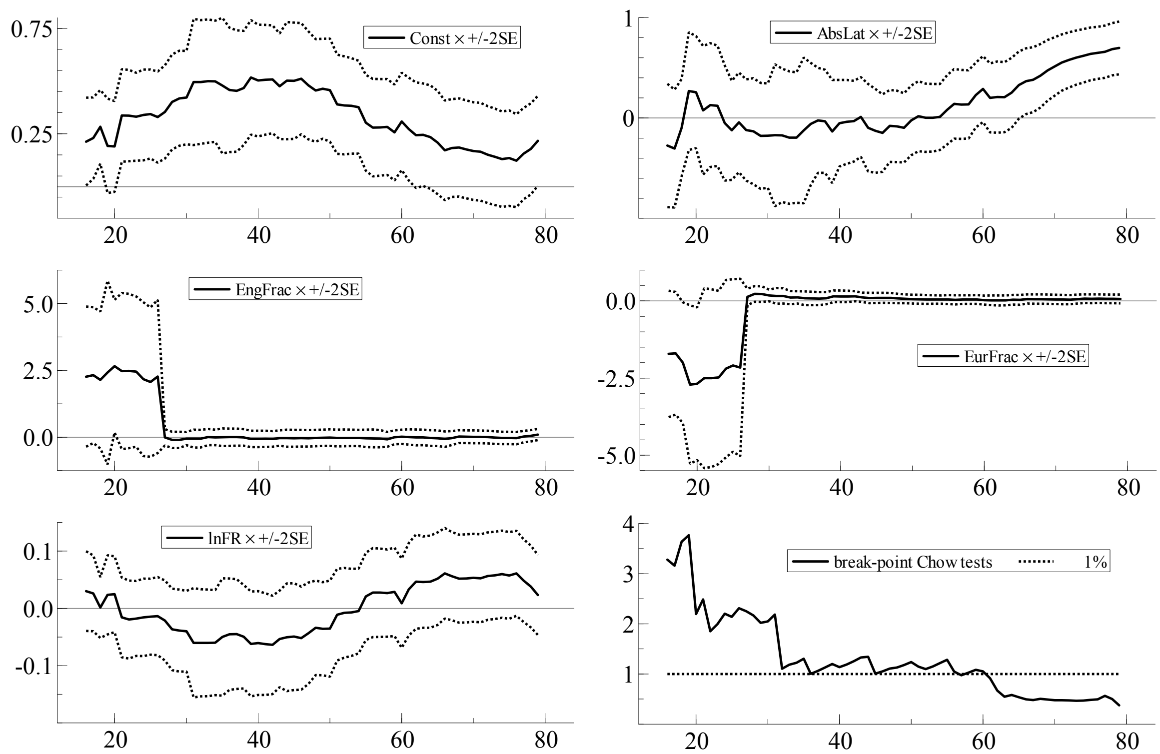


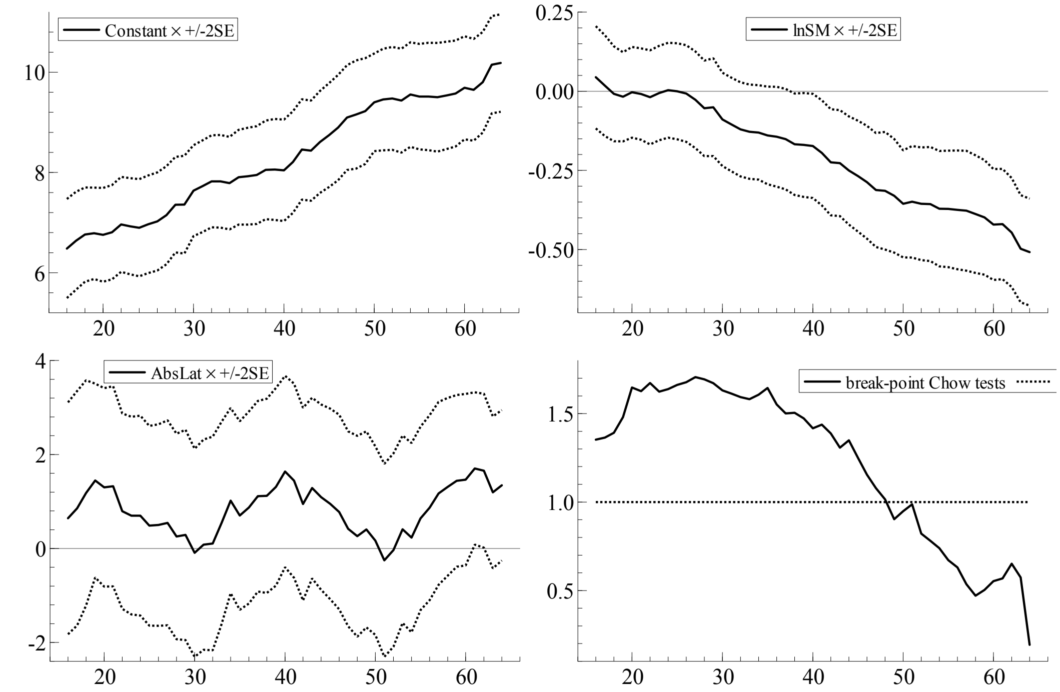
Figure 1. Recursive coefficient estimates and break-point Chow tests for Hall and Jones' RF for *SocInf* (Hall and Jones 1999, Table II, row 3).

Columns (3) and (4) in Table 4 report results for the components of *SocInf* separately, corresponding to a three-equation MLR including $\ln(Y/L)$, *GADP* and *YrsOpen* as dependent variables. Entries in columns (3) and (4) for the system tests therefore refer to the three-equation system, including the RF for $\ln(Y/L)$, for which the individual-equation test results are the same as in column (1). Again, *RESET* results suggest misspecification of the RF for *GADP*, whereas the RF for *YrsOpen* has non-normal errors and a poor fit. The recursive graphs also indicate parameter non-constancy. The apparent weakness in the instruments in the three-variable MLR (reflected in the tabulated results by a very low *CD-F* value and sizeable differences between the conventional and Shea partial R^2 values) may have motivated the use of equally weighted components for *SocInf*. Hall and Jones (1999, Table II) report the results of testing equality of the coefficients on *GADP* and *YrsOpen* in the structural equation for $\ln(Y/L)$. This restriction is not rejected; however, this result may not be reliable given the evidence of lack of statistical adequacy of the underlying RFs.

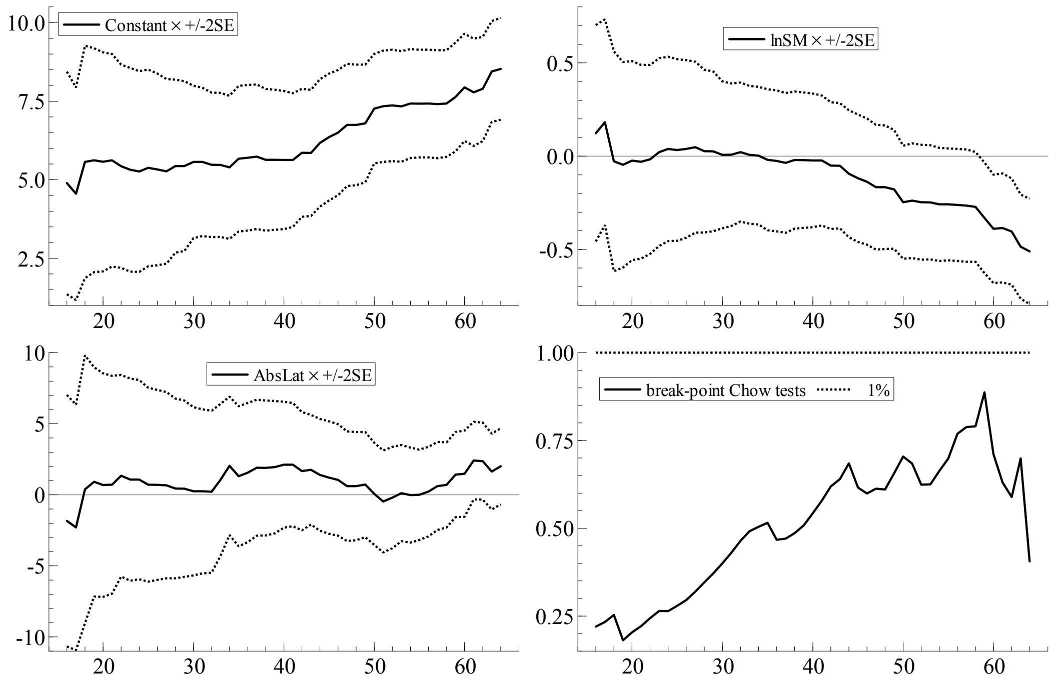
Settler mortality, the instrument for institutional quality proposed by Acemoglu et al. (2001), has been widely adopted by other studies. Table 5 contains diagnostic test results for the RFs for several representative models in Acemoglu et al. (2001, Tables 4 and 5) fitted to their base sample of 64 ex-colonies. These results reveal some evidence of non-normality, heteroskedasticity, and functional form misspecification. More importantly, there is again strong evidence of spatial dependence for all models.

Another recurring pattern is lack of parameter constancy in the recursive plots of the estimated coefficients, especially for the RF for ln*GDPpc*. This is illustrated in Figure 2a (for the RF in Table 5, column (1), based on the model in Acemoglu et al. (2001, Table 4, column 2)). The extensive set of significant break-point Chow test values and the drifting patterns in the intercept term and the coefficient on the crucial additional instrument, logarithm of settler mortality (ln*SM*), imply parameter non-constancy for the RF of ln*GDPpc*. None of the break-point Chow test values for the RF for Acemoglu et al.'s institutional quality variable, average expropriation risk (*AvExpr*), is significant, but the parameters for *AvExpr* are less precisely estimated. In particular, the coefficient on ln*SM* is not statistically significant in either RF until countries at higher levels of development are included; thereafter, the negative coefficients on ln*SM* in the RFs for both ln*GDPpc* (in panel a) and *AvExpr* (in panel b) continue to increase in absolute value as additional higher income countries are added to the sample.

Easterly and Levine (2003) fit several different models incorporating the effects of institutional quality, crop and mineral endowments, and policy outcomes. They regress the logarithm of GDP per capita in 1995 (ln*GDPpc*) on institutional quality (*Inst*, calculated as the average of six World Bank Governance Indicators) and control variables (including French legal origin, religion dummies, and ethnolinguistic fractionalization). The instrument set for *Inst* includes various subsets of settler mortality, latitude, landlocked, and crop/mineral endowment dummies. Diagnostic test results for representative models, reported in Table 6, reveal evidence of heteroskedasticity and functional form misspecification in the RFs, and spatial independence is strongly rejected (especially for ln*GDPpc*). The model in Easterly and Levine's (2003) Table 5, row 4 performs best on the misspecification tests. However, for this model, the recursive plots suggest that coefficient estimates for individual variables are either not statistically significant through the full set of recursive samples or are not constant. For example, Figure 3 shows the recursive plots for the coefficient on ln*SM* in the equation for ln*GDPpc* in panel (a) (demonstrating non-constancy) and in the equation for *Inst* in panel (b) (demonstrating non-significance). In the RF for *Inst*, the coefficient on *YrsOpen* is highly statistically significant; Easterly and Levine treat *YrsOpen* as exogenous, whereas in several other studies (e.g., Rodrik et al. 2004), it is assumed to be endogenous and is itself instrumented.



(**a**)



(**b**)

Figure 2. Recursive coefficient estimates and break-point Chow tests for Acemoglu et al.'s (2001, Table 4, column 2) RFs. (**a**) RF for ln*GDPpc*; (**b**) RF for *AvExpr*.

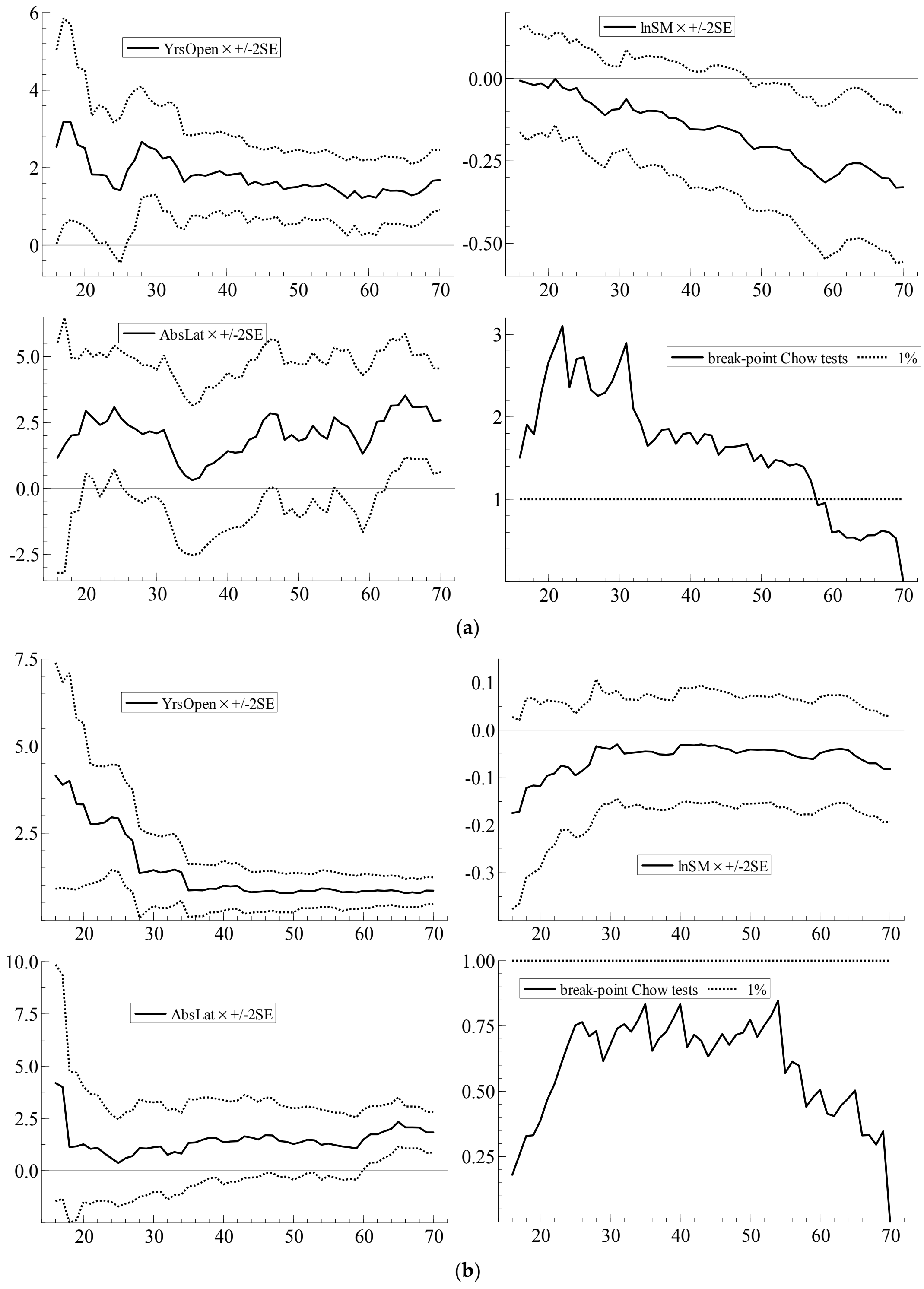


Figure 3. Recursive estimates for selected coefficients and break-point Chow tests for Easterly and Levine's (2003, Table 5, row 4) RFs. (**a**) RF for ln*GDPpc*; (**b**) RF for *Inst*.

Table 5. Testing statistical adequacy of RFs for Acemoglu et al. (2001).

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	T4C2		**T4C8**		**T5C6**		**T5C7**		**T5C8**		**T5C9**	
	ln*GDPpc*	***AvExpr***	**ln*GDPpc***	***AvExpr***	**ln*GDPpc***	***AvExpr***	**ln*GDPpc***	***AvExpr***	**ln*GDPpc***	***AvExpr***	**ln*GDPpc***	***AvExpr***
Norm-p	0.070	0.975	0.064	0.999	**0.046**	0.879	0.177	0.769	0.149	0.887	0.358	0.998
NormVec-p	**0.050**		**0.037**		**0.014**		**0.026**		**0.030**		0.074	
Hetero-p	0.253	0.513	**0.017**	0.831	0.377	0.642	0.312	0.814	0.083	0.800	0.187	0.823
HeteroVec-p	0.585		0.146		0.765		0.453		0.220		0.279	
HeteroX-p	0.272	0.654	**0.030**	0.859	0.345	0.733	0.079	0.333	**0.035**	0.727	0.066	0.698
HeteroXVec-p	0.641		0.209		0.740		**0.035**		**0.022**		**0.010**	
RESET-p	0.407	**0.006**	0.061	**0.014**	0.198	0.068	**0.042**	**0.044**	0.093	**0.026**	0.064	0.063
RESETVec-p	0.196		0.068		0.369		0.100		0.163		0.103	
Moran's *I-p*	**0.003**	**0.002**	**0.006**	**0.002**	**0.006**	**0.009**	**0.004**	**0.001**	**0.004**	**0.001**	**0.002**	**0.002**
$LM_{\rho\lambda}$-*p*	**0.023**	**0.005**	0.083	**0.021**	**0.035**	**0.008**	**0.019**	**0.005**	**0.029**	**0.005**	**0.020**	**0.008**
Parameter Constancy	NC	C/NC	NC	C/NC	NC	C/NC	NC	C/NC	NC	C/NC	NC	C/NC
R^2	0.500	0.296	0.584	0.328	0.505	0.345	0.562	0.321	0.588	0.354	0.591	0.369
N	64		64		64		64		64		64	
CD-F		13.093		**3.456**		9.886		19.841		**8.613**		**5.277**
Partial R^2		0.177		0.056		0.142		0.252		0.129		0.086

Dependent variables: ln*GDPpc* is log of GDP per capita in 1995; *AvExpr* is average protection against expropriation risk (1985–1995). TxCy denotes the model in Acemoglu et al. (2001), Table x, Column y. Instrument sets: Exogenous regressors: *AbsLat* (in T4C2, T4C8, T5C6, T5C8, T5C9), continent dummies for Asia, Africa and 'Other' (in T4C8), *FrLO* (in T5C6, T5C9), *FrC* (in T5C9), religion variables (in T5C7, T5C8, T5C9); Additional instrument: ln*SM* (all models, which are exactly identified). Variables are defined in the notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. β = 0.25 in the spatial weighting matrix.

Table 6. Testing statistical adequacy of RFs for Easterly and Levine (2003).

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	T4R4		T4R6		T4R6#		T5R4	
	ln*GDPpc*	*Inst*	ln*GDPpc*	*Inst*	ln*GDPpc*	*Inst*	ln*GDPpc*	*Inst*
Norm-p	0.268	0.842	0.908	0.157	0.374	0.931	0.349	0.958
NormVec-p	0.119		0.903		0.173		0.425	
Hetero-p	0.123	0.494	0.938	0.970	0.088	0.657	0.485	0.872
HeteroVec-p	0.537		0.971		0.614		0.832	
HeteroX-p	**0.006**	0.301	0.489	0.884	**0.001**	0.235	0.944	0.638
HeteroXVec-p	0.252		0.947		0.158		0.784	
RESET-p	**0.010**	**0.016**	**0.001**	0.284	**0.005**	0.059	0.071	0.133
RESETVec-p	**0.008**		**0.004**		**0.016**		**0.015**	
Moran's *I-p*	**0.000**	**0.017**	**0.015**	0.124	**0.001**	**0.012**	**0.004**	0.201
$LM_{\rho\lambda}$-*p*	**0.011**	0.150	**0.005**	**0.006**	**0.012**	0.121	**0.026**	0.393
Parameter Constancy	NC	C	NC	C/NC	NC	NC	NC	C/NC
R^2	0.615	0.573	0.787	0.729	0.632	0.593	0.686	0.674
N	72		72		72		70	
Sargan-p	0.066		0.429		0.145		0.097	
CD-F	**11.743**		**5.155**		**10.898**		12.131	
Partial R^2	0.359		0.563		0.345		0.285	

Dependent variables: ln*GDPpc* is log of GDP per capita in 1995; *Inst* is the average of six World Bank Governance Indicators. TxRy denotes the model in Easterly and Levine (2003), Table x, Row y. # corresponds to the model in T4R6 but excluding non-oil crops/minerals dummies in the IV set (not reported in Easterly and Levine (2003)). Instrument set in each column: Exogenous regressors: *FrLO*, religion dummies (Catholic, Muslim, other) and *EthDiv* (all models), *Oil* (in T4R6, T4R6#), *YrsOpen* (T5R4); Additional instruments: ln*SM* and *AbsLat* (all models), *Landlock* (in T4R4, T4R6, T4R6#), set of 10 crops/minerals dummies (in T4R6). Variables are defined in the notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. β = 0.2 in the spatial weighting matrix.

Table 7. Testing statistical adequacy of RFs for Sachs (2003).

Test	(1)	(2)	(3)	(4)
		T1C10		T1C12
	lcgdp95	*Rule*	*Mal94p*	*Malfal*
Norm-p	0.147	0.420	0.303	0.072
NormVec-p		**0.002**		**0.001**
Hetero-p	0.654	0.727	**0.000**	**0.000**
HeteroVec-p		**0.018**		0.162
HeteroX-p	0.757	0.651	**0.000**	**0.000**
HeteroXVec-p		0.093		0.356
RESET-p	0.274	0.148	**0.003**	**0.000**
RESETVec-p		**0.018**		**0.000**
Moran's *I-p*	**0.001**	0.817	**0.004**	**0.000**
$LM_{\rho\lambda}$-*p*	**0.001**	0.487	**0.027**	**0.001**
Parameter Constancy	NC	C	NC	NC
R^2	0.603	0.541	0.581	0.637
N	69		69	
	Rule	*Mal94p*	*Rule*	*Malfal*
Partial R^2	0.541	0.581	0.541	0.637
Shea partial R^2	0.253	0.272	0.367	0.432
Sargan-p	0.404		0.560	
CD-F	**6.371**		11.592	

Dependent variables: *lcgdp95* is log of GDP per capita in 1995 (Rodrik et al. 2004); *Rule* is a Rule of Law index; *Mal94p* is the proportion of the population at risk of malaria transmission in 1994; *Malfal* is the proportion at risk of falciparum malaria transmission. TxCy denotes the model in Sachs (2003), Table x, Column y. Model T1C12 is for the three-equation MLR for *lcgdp95*, *Rule*, and *Malfal*. Instrument set in each column (all additional instruments): ln*SM*, *KGPTemp*, and *ME*. Variables are defined in the notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. β = 0.2 in the spatial weights matrix.

Diagnostic tests for the RFs of two representative models from Sachs (2003), which add an index of malarial ecology (ME)—based on temperature, mosquito abundance, and vector specificity—as an instrument to address the endogeneity of malarial risk, are reported in Table 7. These raise concerns about non-normality, heteroskedasticity, and functional form, especially for the RFs for the malarial risk variables, *Mal94p* (the proportion of the population at risk of malaria transmission in 1994) and *Malfal* (the proportion at risk of malaria transmission involving the fatal species *Plasmodium falciparum*). Spatial independence is strongly rejected in the RFs for the logarithm of GDP per capita and the malarial risk variables. The recursive estimates, represented by selected plots in Figure 4, also indicate sometimes severe cases of parameter non-constancy. The lack of statistical adequacy of the RFs is consistent with Sachs' (2003, pp. 3–4) concern that "the model ... is worryingly oversimplified in any case" and that it is "very doubtful that a process as complex as economic development can possibly be explained by two or three variables alone".

To test the Malthusian theory that improvements in technology in the preindustrial era increased population density but not living standards, Ashraf and Galor (2011) fit a number of models explaining the logarithm of population density (*lpd*) for different years (1, 1000, and 1500). The explanatory variables include the logarithm of the number of years since the Neolithic transition (*lyst*). Although they point out that reverse causality from population density to *lyst* is not a problem, OLS estimates of *lyst*'s coefficient may suffer from omitted variables bias. To address endogeneity, they use the numbers of prehistoric domesticable species of wild plants and animals, from Olsson and Hibbs (2005), to instrument *lyst*, arguing that their only effect on later population density is via their effect on the timing of the Neolithic transition.[13] As well as population density in different years, they also explore the effects of *lyst* on subsequent technological sophistication, represented by a non-agricultural technology index in years 1000 and 1 (*natech*1*K* and *natech*1, respectively).

Diagnostic tests corresponding to Ashraf and Galor's IV regressions are reported in Table 8. Spatial independence of the residuals is strongly rejected for all the fitted models. There is also evidence of non-normality, heteroskedasticity, functional form misspecification, and parameter non-constancy. Similar results apply to the RFs for models of population density in which the effect of contemporaneous technology is examined (columns (12)–(15)). Significant diagnostic statistics are also apparent (columns (7) and (8)) for IV estimates of the illustrative version of Ashraf and Galor's model that Spolaore and Wacziarg (2013) report in their review paper.[14]

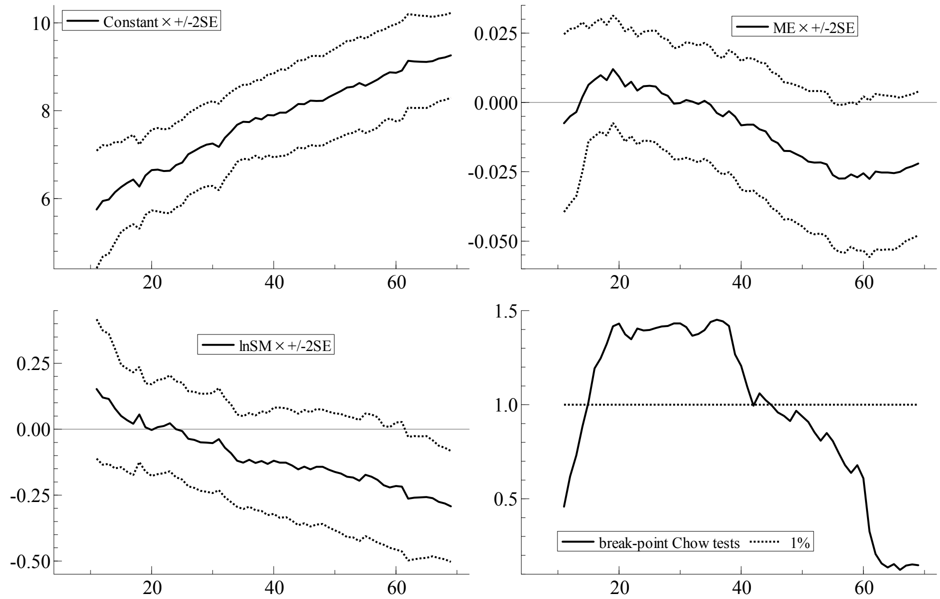
Other recent studies that focus on historical or intergenerational factors (Chanda and Putterman 2007; Spolaore and Wacziarg 2009; Putterman and Weil 2010; Easterly and Levine 2016) are also less concerned with reverse causation and place more emphasis on reporting OLS estimates of Equation (1).[15] If $\mathrm{E}(\boldsymbol{X}_i\varepsilon_i) = 0$, then direct examination of statistical adequacy of the single-equation OLS estimates would be appropriate. Table 9 reports diagnostic test results for a selection of illustrative models explaining the logarithm of per capita income in 2005 (*lpci*05), reported in Spolaore and Wacziarg (2013). Following Putterman and Weil (2010) and Easterly and Levine (2016), these include ancestry-adjusted years of agriculture (*AdjYrsAg*) and ancestry-adjusted state history models (*AdjStateHist*) (in columns (1) and (2)) and the share of descendants of Europeans (*EuroShare*) (in columns (3)–(5)). Following

13 Ashraf and Galor (2011, p. 2016) express the view that "variations in land productivity and other geographical characteristics are *inarguably exogenous* to the cross-country variation in population density" (emphasis added). This is surprising given the emphasis on potential omitted variables as a source of endogeneity for *lyst*; omitted variables may also be correlated with the geographical controls, which would potentially bias OLS estimates for all the coefficients.

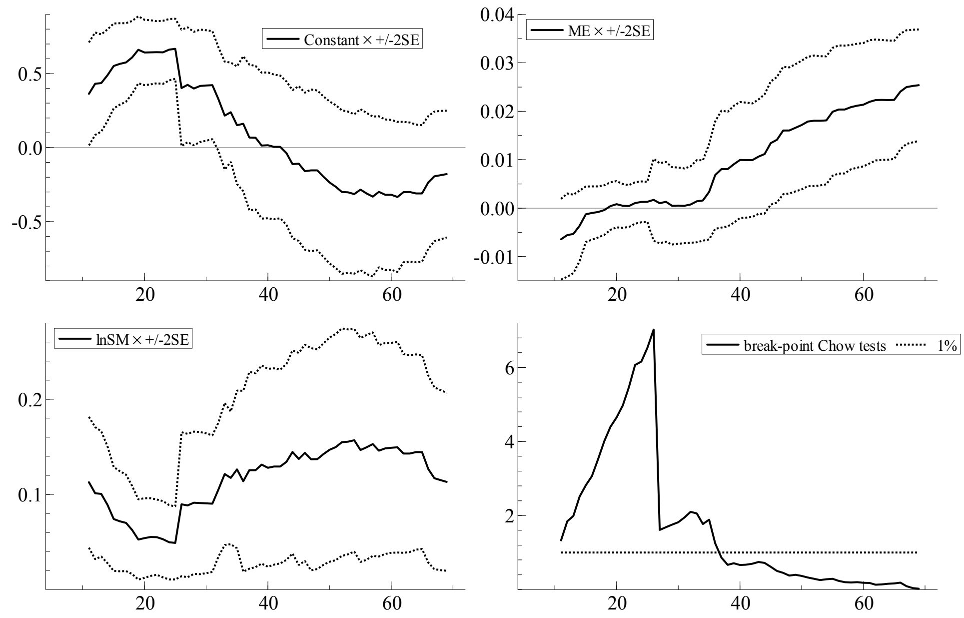
14 The version of the model fitted by Spolaore and Wacziarg (2013) includes different geographical control variables (absolute latitude, percentage of land area in the tropics, a landlocked dummy and an island dummy). These are therefore included with the additional instruments, *Plants* (the number of prehistoric wild grasses) and *Animals* (the number of prehistoric domesticable large mammals), in the instrument set appearing in each RF.

15 Correlation of explanatory variables with omitted variables is, however, still a source of endogeneity, which is considered to varying degrees. Spolaore and Wacziarg (2009) use genetic distance as of 1500 to instrument for current genetic distance in their bilateral income difference regressions. Putterman and Weil (2010) emphasize the importance of including appropriate controls to reduce the possibility of omitted variables bias.

Spolaore and Wacziarg (2009), the models in columns (6)–(8) include genetic distance, as a proxy for a wide range of intergenerationally transmitted characteristics. There is consistent evidence of spatial dependence and apparent parameter non-constancy (although the latter is less dramatic than in some of the earlier studies considered).



(**a**)



(**b**)

Figure 4. Recursive estimates for selected coefficients and break-point Chow tests for Sachs' (2003, Table 1, column 10) RFs. (**a**) RF for *lcgdp95*; (**b**) RF for *Malfal*.

Table 8. Testing statistical adequacy of RFs for Ashraf and Galor (2011).

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	Ashraf and Galor (2011)		**Ashraf and Galor (2011)**		**Ashraf and Galor (2011)**		**Spolaore and Wacziarg (2013)**		**Ashraf and Galor (2011)**			**Ashraf and Galor (2011)**		**Ashraf and Galor (2011)**	
	T2C6		**T3C6**		**T4C6**		**T2C4**		**T8C3**		**T8C6**	**T9C3**		**T9C6**	
	***lpd*1500**	***lyst***	***lpd*1000**	***lyst***	***lpd*1**	***lyst***	***lpd*1500**	***lyst***	***natech*1*K***	***lyst***	***natech*1**	***lpd*1000**	***tech*1*K***	***lpd*1**	***tech*1**
Norm-p	0.360	**0.010**	0.121	**0.015**	**0.029**	**0.002**	0.461	**0.001**	**0.004**	**0.004**	0.073	0.061	**0.003**	**0.023**	0.643
NormVec-p	**0.027**		**0.010**		**0.001**		**0.006**		**0.001**		**0.006**	**0.001**		0.075	
Hetero-p	0.323	0.096	0.283	0.085	**0.039**	0.425	**0.001**	**0.001**	**0.000**	0.150	**0.049**	0.329	**0.002**	0.050	**0.001**
HeteroVec-p	**0.011**		**0.002**		0.069		**0.000**		**0.000**		0.113	**0.000**		**0.001**	
HeteroX-p	**0.031**	0.082	**0.034**	0.083	0.064	0.346	**0.000**	**0.002**	**0.021**	0.067	0.149	**0.038**	**0.045**	0.113	**0.011**
HeteroXVec-p	**0.001**		**0.001**		0.185		**0.000**		**0.001**		**0.041**	**0.000**		**0.000**	
RESET-p	0.055	0.308	**0.010**	0.460	0.282	0.678	**0.035**	0.251	**0.016**	0.454	0.242	**0.013**	**0.008**	0.194	0.059
RESETVec-p	0.152		0.059		0.077		**0.020**		0.200		0.140	**0.010**		0.065	
Moran's *I-p*	**0.000**	**0.000**	**0.001**	**0.000**	**0.001**	**0.000**	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.001**	**0.002**
$LM_{\rho\lambda}$-*p*	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.012**
Parameter Constancy	NC	NC	NC	NC	NC	C/NC	NC	C/NC	NC	C/NC	C/NC	NC	C	NC	NC
R^2	0.686	0.685	0.650	0.698	0.617	0.712	0.474	0.721	0.720	0.674	0.555	0.624	0.711	0.614	0.511
N	96		94		83		98		93		93	92		83	
Sargan-p		0.358		0.159		0.587		0.216		0.343	0.254		0.938		0.250
CD-F		16.299		16.067		12.458		69.911		14.484	14.484		**8.595**		**7.105**
Partial R^2		0.275		0.277		0.255		0.606		0.259	0.259		0.173		0.163

Dependent variables: *lpd*1500, *lpd*1000, and *lpd*1 are, respectively, the log of population density in years 1500, 1000, and 1; *lyst* is the log of years since the Neolithic transition; *natech*1*K* and *natech*1 are, respectively, a non-agricultural technology index in 1000 and 1; *tech*1*K* and *tech*1 are, respectively, a technology index in 1000 and 1. Instrument set in each column: Exogenous regressors: ln(*LandProd*), ln(*AbsLat*), *distcr*, *Land100km*, continent dummies for Africa, Europe, and Asia (except for columns (7) and (8), see footnote 14); Additional instruments: *Plants*, *Animals*. TxCy denotes the model in Table x, Column y of the relevant study. Variables are defined in the notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. β = 0.175 in the spatial weights matrix (except β = 0.15 for columns (5), (6), (14), and (15)).

Table 9. Testing statistical adequacy of illustrative models from Spolaore and Wacziarg (2013).

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	T5C2	**T5C4**	**T6C3**	**T6C4**	**T6C5**	**T7C1**	**T7C2**	**T7C3**
	***lpci*05**	***lpci*05**	***lpci*05**	***lpci*05**	***lpci*05**	***lpci*05**	***lpci*05**	***lpci*05**
Norm-p	0.917	0.499	0.438	0.322	0.072	0.148	0.269	0.072
Hetero-p	0.249	0.431	0.115	0.214	**0.034**	0.097	0.097	**0.034**
HeteroX-p	0.130	0.237	0.128	0.150	**0.042**	0.146	0.058	**0.042**
RESET-p	0.636	0.739	**0.025**	0.531	0.220	0.590	0.365	0.220
Moran's *I-p*	**0.000**	**0.028**	**0.000**	**0.028**	**0.000**	**0.001**	**0.000**	**0.000**
$LM_{\rho\lambda}$-*p*	**0.000**	**0.006**	**0.000**	0.098	**0.000**	**0.001**	**0.001**	**0.000**
Parameter Constancy	NC	C/NC	NC	C/NC	NC	NC	NC	NC
R^2	0.523	0.588	0.580	0.656	0.545	0.499	0.496	0.545
N	148	135	147	134	149	155	154	149

Dependent variable in all models is the log of per capita income in 2005 (*lpci*05). All OLS regressions include a common set of control variables: *AbsLat*, *LandTropics*, *Landlock*, *Island*. Additional exogenous regressors for each column are: (1) *AdjYrsAg*; (2) *AdjStateHist*; (3) *EuroShare*, *AdjYrsAg*; (4) *EuroShare*, *AdjStateHist*; (5) *EuroShare*, *WtGenDist*; (6) *IndGenDist*; (7) *WtGenDist*; (8) *WtGenDist*, *EuroShare*. TxCy denotes the model in Spolaore and Wacziarg (2013), Table x, Column y. Variables are defined in the notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. β = 0.25 in the spatial weights matrix.

Ashraf and Galor (2013) regress the logarithm of population density in 1500 (*lpd*1500 in Table 10), as a proxy for historical productivity, on observed genetic diversity, controlling for the logarithm of the timing of the Neolithic transition (*lyst*), the logarithm of the percentage of arable land (ln(*Arable*)), the logarithm of absolute latitude (ln(*AbsLat*)), the logarithm of land suitability for agriculture (ln(*AgSuit*)), and continent fixed effects. Initial results are for a limited sample of 21 countries for which the required data can be compiled. Ashraf and Galor instrument observed genetic diversity using migratory distance from East Africa (*mdistAddis*). To test the hump-shaped effect of genetic diversity on productivity, they also include genetic diversity squared in their model; following Wooldridge (2010, p. 267), they use the squared value of predicted genetic diversity (*divhatsq*), from a preliminary regression of diversity on migration distance and controls, as an additional instrument.

Table 10. Testing statistical adequacy of RFs for Ashraf and Galor (2013).

Test	(1)	(2)	(3)	(4)	(5)	(6)
		T2C5			**T2C6**	
	***lpd*1500**	***Div***	***DivSq***	***lpd*1500**	***Div***	***DivSq***
Norm-p	0.545	0.947	0.930	0.876	**0.019**	**0.007**
NormVec-p		0.909			0.224	
Hetero-p	0.847	**0.044**	0.071	0.136	0.521	0.669
HeteroVec-p		0.286			NF	
RESET-p	0.415	0.816	0.750	0.591	0.060	0.284
RESETVec-p		**0.003**			**0.013**	
Moran's *I-p*	0.156	0.680	0.719	0.213	0.485	0.499
$LM_{\rho\lambda}$-*p*	0.130	0.207	0.235	0.080	**0.031**	0.028
Parameter Constancy	C	C	C	C	C	C
R^2	0.900	0.988	0.986	0.900	0.993	0.993
N		21			21	
CD-F		19.283			18.861	
Partial R^2		0.986	0.983		0.896	0.883
Shea partial R^2		0.740	0.738		0.815	0.804

Dependent variables: *lpd*1500 is the log of population density in 1500; *Div* is (observed) genetic diversity and *DivSq* is its square. TxCy denotes the model in Ashraf and Galor (2013), Table x, Column y. Instrument sets: Exogenous regressors: *lyst*, ln(*Arable*), ln(*AbsLat*), ln(*AgSuit*) (in all models); continent dummies (Africa, Europe, Americas) in T2C6; Additional instruments (in all models): *mdistAddis*, *divhatsq*. Variables are defined in the text and notes to Table 2. See Table 1 for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. NF = not feasible due to small sample size. β = 0.1 in spatial weighting matrix.

Diagnostic tests corresponding to estimates in Ashraf and Galor's (2013) Table 2, columns (5) and (6) are reported in Table 10. The diagnostics reveal relatively few problems; apart from marginal heteroskedasticity in the RF for genetic diversity, the only other potential problem is the multivariate RESET result, which is significant despite the individual equations passing this test. Adding continental dummies (in their Table 2, column (6)) appears to cause problems with the assumption of normal errors. The RFs (for both models) display less evidence of parameter non-constancy than any of the other studies examined, and this is the only study considered for which there is little evidence of spatial dependence of the residuals. Although the small sample leads to relatively wide confidence bands, most coefficients are statistically significant over the full range of recursive samples (e.g., see the plots for the RF for diversity in Figure 5).

However, the replicated models from Ashraf and Galor's (2013) study are the exception. In general, diagnostic testing of the RFs in these representative studies of the fundamental determinants of development provides evidence of varying degrees of failure of the underlying assumptions upon which conventional inference is based, which is suggestive of model misspecification. Even if we discount evidence of non-normality and concerns over heteroskedasticity as a possible indicator of misspecification, and are prepared to rely on corrections to standard errors as a default (even though sample sizes are not large in these studies), parameter non-constancy and spatial dependence in the residuals are almost ubiquitous, while several models also show evidence of functional form misspecification.

All the empirical studies of the fundamental determinants of development adopt a broadly similar approach, i.e., fitting simple, essentially static, highly parsimonious models with explanatory variables that are potentially endogenous, due to reverse causation and/or omitted variables. Despite the ingenuity displayed in identifying plausible natural experiments delivering quasi-random variation in the fundamental determinants, the highly parsimonious nature of the models makes it problematic to achieve statistically adequate RFs. The open-ended nature of growth theories (Brock and Durlauf 2001) also applies, if to a lesser degree, to the list of potential fundamental determinants (including different dimensions of institutional quality, as well as historical, geographical, and biological factors), so it is difficult to ensure that all relevant variables are included in the model. This is likely to be reflected in lack of statistical adequacy. As these variables are not usually orthogonal, omitted variables bias is a potentially serious problem.

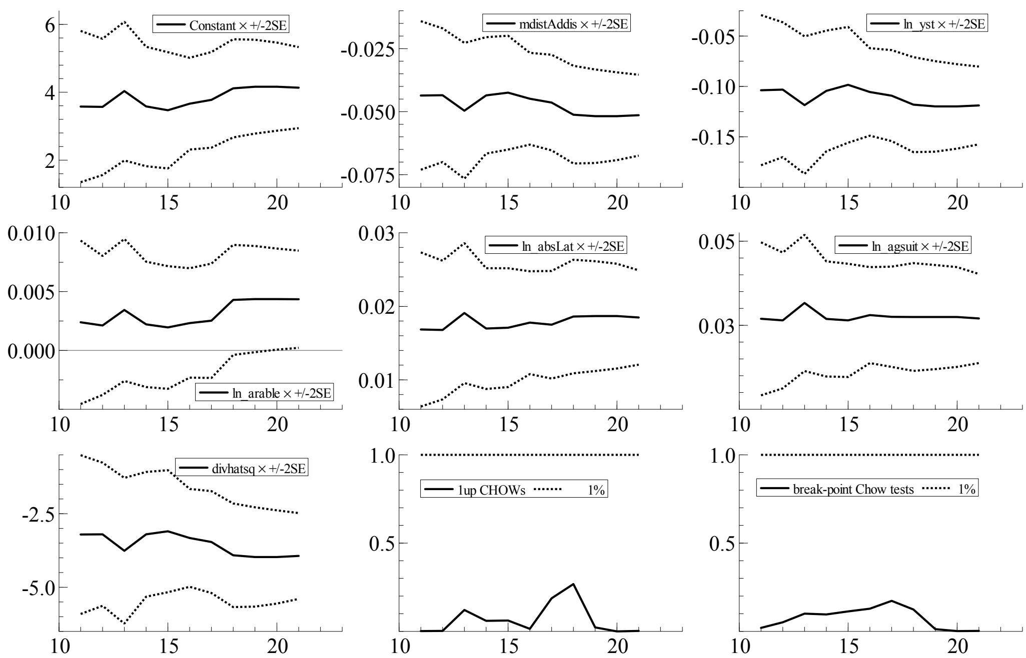


Figure 5. Recursive coefficient estimates and break-point Chow tests for Ashraf and Galor's (2013, Table 2, column 5) RF for Diversity (*Div*).

Spatial dependence appears to be an almost universal feature of the residuals from the fitted models. Given the cross-country nature of the data, evidence of hidden dependencies is perhaps not surprising, but it casts doubt on the underlying assumption of random sampling.[16] This is cause for concern given that omitted variables bias and inconsistency are amplified in the presence of spatial dependence. For example, LeSage and Pace (2009, Section 3.3) present Monte Carlo results demonstrating the possibility of serious bias when OLS is used in situations where omitted variables are combined with spatial dependence in regressors and/or errors; they also note that spatial dependence in the dependent variable further increases bias. Spatial dependence is a feature of the statistical models that has been almost entirely neglected in this literature. The only exception in the studies examined is a robustness analysis in the online appendix for Ashraf and Galor's (2013) study in which the standard errors are corrected for spatial autocorrelation in the error terms (their Appendix A Tables D2 and D3); they also report that their results are robust to the use of estimators that allow for a first-order spatial autoregressive model with first-order spatial autoregressive errors (their Appendix A Table D18). Ironically, the models from their study are the only ones for which there is little evidence of spatial autocorrelation in the residuals. None of the other studies attempt to model spatial dependence explicitly in the structural equation.

The apparent lack of parameter constancy in these studies is related to concerns expressed by Deaton (2010) that equations in the growth and development literature are really not structural equations in which the parameters are constant. Instead, Deaton argues that variation in the parameters across cross-sectional units is likely and is affected by the choice of instruments. If parameter heterogeneity across countries is relevant, the focus shifts to estimating a local average treatment effect, which requires stronger assumptions (e.g., Angrist and Pischke 2009, pp. 152–58). Rejection of the null of parameter constancy does not necessarily imply acceptance of the alternative of varying parameters (in an otherwise appropriately specified model), because apparent parameter non-constancy is often a symptom of a misspecified model (Hendry 1995). However, parameter heterogeneity across different countries at different stages of development is consistent with evidence from panel time-series estimation of production relationships in different countries (Eberhardt and Teal 2017); this interpretation suggests that the effects of the fundamental determinants are likely to vary at different stages of development.

6. Conclusions

Empirical analysis in the growing literature on the fundamental determinants of cross-country comparative development relies heavily on 2SLS estimation of structural parameters in highly parsimonious models. In addressing potential endogeneity problems, several studies have proposed a series of ingenious instruments. Economic theory (regardless of its degree of formalism) underpins model specification, including the choice of relevant explanatory variables and exclusion restrictions. Instrumentation strategies in this literature are therefore not atheoretical. Rather, following Spanos's (2007) arguments, a greater concern is the lack of attention paid to the statistical adequacy of the underlying statistical model, as summarized in the system's reduced-form equations. Whereas most applications of IV/2SLS estimation treat the fitting of the first-stage regression as purely a prediction exercise, Spanos emphasizes that the full set of RFs, specified in terms of the

[16] In the fundamental determinants literature, the country is the usual unit of geographical aggregation, as is the case for all the studies considered here. An interesting question, left for future investigation, is whether spatially correlated residuals are also present in sub-national empirical studies, such as Michalopoulos's (2012) exploration of the determinants of ethnolinguistic diversity at different levels of spatial aggregation, including 'virtual countries' and adjacent regions. Noting that cross-country studies are based on relatively small sample sizes, a referee poses the question of whether statistical assumptions are more likely to be violated with small samples. Sub-national studies would also provide a larger number of observations, which may shed light on this question, but, in general, misspecification is just as likely with models fitted to large as to small samples and its adverse effects (e.g., due to heterogeneity) may be even more damaging in large samples (Spanos 2017).

observable variables, provides an embedding framework for the structural equations, and reflects instrument exogeneity assumptions in the parameterizations for the structural parameters in the context of the MLR/RF. Failure of the statistical assumptions underlying the RFs implies failure of the corresponding structural-equation assumptions. Lack of statistical adequacy, i.e., invalid underlying probabilistic assumptions for the data being analysed, can lead to inconsistent estimators and/or unreliable inference, due to actual Types I and II error probabilities deviating from their nominal values. Spanos' approach emphasizes that a statistical framework subject to a comprehensive set of explicit, non-rejected assumptions is better placed to provide valid inference in comparison to asymptotic properties that depend on a weaker set of implicit and untested (or untestable) assumptions. Particularly given the sample sizes common in the literature on the fundamental determinants of development, the former is a more appealing strategy.

Both a sound theoretical justification for exclusion restrictions *and* statistical adequacy of the RFs are desirable features of a credible instrumentation strategy. However, when subject to diagnostic testing for misspecification of their RFs, influential representative studies of the fundamental determinants of development exhibit varying degrees of evidence of lack of statistical adequacy. The most concerning departures from the underlying statistical assumptions involve lack of parameter constancy across countries and spatial dependence, which appear to be almost ubiquitous. These features, surprisingly not previously identified, potentially undermine inferences drawn about the structural parameters and hence the effects of particular fundamental determinants. In addition, lack of statistical adequacy across a wide range of different variants of the models suggests that the typical sensitivity analysis reported in this literature may not be sufficient to ensure robustness and reliability of inference.

Empirically identifying the fundamental determinants of long-run development is an ambitious research agenda, made doubly difficult by the long spans of time over which the relevant processes operate and by the lack of long runs of time-series data. One possible interpretation of the lack of statistical adequacy for these parsimonious models fitted to cross-sectional data is that these models are just too simple. Important factors (multiple fundamental determinants and their different dimensions, interactions, dynamics, and nonlinearities) may be missing.[17] The more plausible instruments relying on quasi-random variation from natural experiments may well be based on sound theoretical arguments, but while theory and natural experiments can point to 'candidate' instruments, their statistical appropriateness needs to be tested for the data in question. In this context, the statistical adequacy of the empirical models may be undermined by the overly simplistic nature of these models. In particular, evidence of parameter non-constancy, whether symptomatic of misspecification and/or reflecting heterogeneity in responses across countries, and hidden spatial dependence in cross-section data, require more attention than they have previously received.

Overall, there appear to be sufficient concerns about the statistical adequacy of the IV regressions fitted in most existing fundamental determinants studies to cast doubt on the ability of such parsimonious models to identify the fundamental determinants of development, notwithstanding the ingenious nature of many of the instruments used. On a more positive note, further investigation of the reasons for apparent parameter non-constancy and cross-section dependence offers avenues for potential additional insights.

Acknowledgments: Earlier versions of this paper were presented at the NZAE Conference, Wellington; Otago Development and Growth Workshop, Dunedin; the NZESG meeting, Hamilton; and seminars in the Department of Economics, University of Reading; the Research Institute for Applied Social Sciences, Swansea

[17] Casey and Klemp's (2016) recent study is an exception to the widespread neglect of dynamics in this literature. They examine the use of historical instruments for contemporary endogenous explanatory variables and explicitly consider the persistence of the latter, in aiming to quantify the long-run causal effect of historical values of the endogenous fundamental determinant, such as institutional quality, on contemporary levels of development. Their approach, using system GMM, suggests that conventional IV estimation overestimates the effect of institutions on development.

University; and the Department of Mathematics and Statistics, University of Otago. I am grateful to participants and colleagues, especially Mark Casson, Lorenzo Ductor, David Fielding, John Gibson, Dave Maré, Peter Phillips, and Aris Spanos, and to four referees for helpful comments and suggestions. Thanks also to Wilner Jeanty for providing updated versions of his Stata routines spwmatrix and anketest, and Romain Wacziarg and Jeffrey Sachs for sharing the data used in Sachs (2003) and Spolaore and Wacziarg (2013).

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Data Sources

Hall and Jones (1999): https://web.stanford.edu/~chadj/HallJones400.asc
Acemoglu et al. (2001): https://economics.mit.edu/faculty/acemoglu/data/ajr2001
Easterly and Levine (2003): http://faculty.haas.berkeley.edu/ross_levine/papers.htm
Sachs (2003): obtained on request from Jeffrey Sachs (December 2003)
Ashraf and Galor (2011): https://www.aeaweb.org/articles?id=10.1257/aer.101.5.2003
Spolaore and Wacziarg (2013): obtained on request from Romain Wacziarg (October 2013)
Ashraf and Galor (2013): https://www.aeaweb.org/articles?id=10.1257/aer.103.1.1

References

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2001. The colonial origins of comparative development: An empirical investigation. *American Economic Review* 91: 1369–401. [CrossRef]

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2002. Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics* 117: 1231–94. [CrossRef]

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. Institutions as a fundamental cause of long-run growth. In *Handbook of Economic Growth*. Edited by Philippe Aghion and Steven N. Durlauf. Amsterdam: Elsevier North-Holland, Volume 1A, pp. 385–472.

Andrews, Donald W.K., and James H. Stock. 2007. Inference with weak instruments. In *Advances in Economics and Econometrics, Theory and Applications: 9th Congress of the Econometric Society, vol 3*. Edited by Richard Blundell, Whitney K. Newey and Torsten Persson. Cambridge: Cambridge University Press, pp. 122–73.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives* 24: 3–30. [CrossRef]

Anselin, Luc, and Raymond J.G.M. Florax. 1995. Small sample properties of tests for spatial dependence in regression models: Some further results. In *New Directions in Spatial Econometrics*. Edited by Luc Anselin and Raymond J.G.M. Florax. Berlin: Springer, pp. 21–74.

Anselin, Luc, Anil K. Bera, Raymond Florax, and Mann J. Yoon. 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26: 77–104. [CrossRef]

Ashraf, Quamrul, and Oded Galor. 2011. Dynamics and stagnation in the Malthusian epoch. *American Economic Review* 101: 2003–41. [CrossRef] [PubMed]

Ashraf, Quamrul, and Oded Galor. 2013. The "out of Africa" hypothesis, human genetic diversity, and comparative economic development. *American Economic Review* 103: 1–46. [CrossRef] [PubMed]

Bazzi, Samuel, and Michael A. Clemens. 2013. Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth. *American Economic Journal: Macroeconomics* 5: 152–86. [CrossRef]

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 2005. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken: John Wiley.

Bloom, David E., and Jeffrey D. Sachs. 1998. Geography, demography, and economic growth in Africa. *Brookings Papers on Economic Activity* 1998: 207–73. [CrossRef]

Bockstette, Valerie, Areendam Chanda, and Louis Putterman. 2002. States and markets: The advantage of an early start. *Journal of Economic Growth* 7: 347–69. [CrossRef]

Brock, William A., and Steven N. Durlauf. 2001. What have we learned from a decade of empirical research on growth? Growth empirics and reality. *World Bank Economic Review* 15: 229–72. [CrossRef]

Casey, Gregory, and Marc Klemp. 2016. Instrumental variables in the long run. MPRA Working Paper 68696. University Library of Munich, Munich, Germany.

Chanda, Areendam, and Louis Putterman. 2007. Early starts, reversals and catch-up in the process of economic development. *Scandinavian Journal of Economics* 109: 387–413. [CrossRef]

Cliff, Andrew D., and J.K. Ord. 1973. *Spatial Autocorrelation*. London: Pion.

Conley, Timothy G., and Ethan Ligon. 2002. Economic distance and cross-country spillovers. *Journal of Economic Growth* 7: 157–87. [CrossRef]

Cragg, John G., and Stephen G. Donald. 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory* 9: 222–40. [CrossRef]

Deaton, Angus. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature* 48: 424–55. [CrossRef]

Diamond, Jared. 1997. *Guns, Germs, and Steel: The Fates of Human Societies*. New York: W.W. Norton.

Doornik, Jurgen A., and Henrik Hansen. 2008. An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* 70: 927–39. [CrossRef]

Doornik, Jurgen A., and David F. Hendry. 2013a. *Empirical Econometric Modelling, PcGive 14, Volume I*. London: Timberlake Consultants Ltd.

Doornik, Jurgen A., and David F. Hendry. 2013b. *Modelling Dynamic Systems, PcGive 14, Volume II*. London: Timberlake Consultants Ltd.

Durlauf, Steven N., Paul A. Johnson, and Jonathan R.W. Temple. 2005. Growth econometrics. In *Handbook of Economic Growth*. Edited by Philippe Aghion and Steven N. Durlauf. Amsterdam: Elsevier North-Holland, Volume 1A, pp. 555–677.

Easterly, William. 2007. Inequality does cause underdevelopment: Insights from a new instrument. *Journal of Development Economics* 84: 755–76. [CrossRef]

Easterly, William, and Ross Levine. 2003. Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics* 50: 3–39. [CrossRef]

Easterly, William, and Ross Levine. 2016. The European origins of economic development. *Journal of Economic Growth* 21: 225–57. [CrossRef]

Eberhardt, Markus, and Francis Teal. 2017. The magnitude of the task ahead: Macro implications of heterogeneous technology. Unpublished manuscript, School of Economics, University of Nottingham, UK.

Engerman, Stanley L., and Kenneth L. Sokoloff. 1997. Factor endowments, institutions, and differential paths of growth among New World economies. In *How Latin America Fell Behind*. Edited by Stephen Haber. Stanford: Stanford University Press, pp. 260–304.

Ertur, Cem, and Wilfried Koch. 2007. Growth, technological interdependence and spatial externalities: Theory and evidence. *Journal of Applied Econometrics* 22: 1033–62. [CrossRef]

Feyrer, James, and Bruce Sacerdote. 2009. Colonialism and modern income: Islands as natural experiments. *Review of Economics and Statistics* 91: 245–62. [CrossRef]

Fingleton, Bernard, and Julie Le Gallo. 2008. Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties. *Papers in Regional Science* 87: 319–39. [CrossRef]

Frankel, Jeffrey A. 2003. Comments and discussion on: Bosworth, B.P., Collins, S.M., The empirics of growth: An update. *Brookings Papers on Economic Activity* 2003: 189–99.

Frankel, Jeffrey A., and David Romer. 1999. Does trade cause growth? *American Economic Review* 89: 379–99. [CrossRef]

Freedman, David A. 2006. Statistical models for causation: What inferential leverage do they provide? *Evaluation Review* 30: 691–713. [CrossRef] [PubMed]

Fuchs-Schuendeln, Nicola, and Tarek Alexander Hassan. 2015. Natural experiments in macroeconomics. National Bureau of Economic Research Working Paper 21228, National Bureau of Economic Research, Cambridge, MA, USA.

Gallup, John Luke, Jeffrey D. Sachs, and Andrew D. Mellinger. 1999. Geography and economic development. *International Regional Science Review* 22: 179–232. [CrossRef]

Hahn, Jinyong, and Jerry Hausman. 2002. Notes on bias in estimators for simultaneous equation models. *Economics Letters* 75: 237–41. [CrossRef]

Hall, Robert E., and Charles I. Jones. 1999. Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics* 114: 83–116. [CrossRef]

Hansen, Lars Peter. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–54. [CrossRef]

Heckman, James J., and Sergio Urzúa. 2010. Comparing IV with structural models: What simple IV can and cannot identify. *Journal of Econometrics* 156: 27–37. [CrossRef] [PubMed]

Hendry, David F. 1995. *Dynamic Econometrics*. Oxford: Oxford University Press.

Hendry, David F. 2009. The methodology of empirical econometric modeling: Applied econometrics through the looking-glass. In *Palgrave Handbook of Econometrics: Volume 2: Applied Econometrics*. Edited by Terence C. Mills and Kerry Patterson. Basingstoke: Palgrave Macmillan, pp. 3–67.

Hendry, David F., and Bent Nielsen. 2007. *Econometric Modeling: A Likelihood Approach*. Princeton: Princeton University Press.

Imbens, Guido W. 2010. Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature* 48: 399–423. [CrossRef]

Imbens, Guido W., and Michal Kolesár. 2016. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics* 98: 701–12. [CrossRef]

Iyer, Lakshmi. 2010. Direct versus indirect colonial rule in India: Long-term consequences. *Review of Economics and Statistics* 92: 693–713. [CrossRef]

Jeanty, P. Wilner. 2010. Anketest: Stata module to perform diagnostic tests for spatial autocorrelation in the residuals of OLS, SAR, IV, IV-SAR, probit, and logit models. Available online: ideas.repec.org/c/boc/bocode/s457113.html (accessed on 25 March 2014).

Kraay, Aart C. 2015. Weak instruments in growth regressions: Implications for recent cross-country evidence on inequality and growth. World Bank Policy Research Working Paper 7494, World Bank Group, Washington, D.C., USA.

Knowles, Stephen, and P. Dorian Owen. 2010. Which institutions are good for your health? The deep determinants of comparative cross-country health status. *Journal of Development Studies* 46: 701–23. [CrossRef]

King, Gary, and Margaret E. Roberts. 2015. How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis* 23: 159–79. [CrossRef]

La Porta, Rafael, Florencio Lopez-de-Silanes, Andrei Shleifer, and Robert Vishny. 1999. The quality of government. *Journal of Law, Economics, and Organization* 15: 222–79. [CrossRef]

Le Cam, Lucien. 1986. *Asymptotic Methods in Statistical Decision Theory*. New York: Springer.

Leamer, Edward E. 2010. Tantalus on the road to Asymptopia. *Journal of Economic Perspectives* 24: 31–46. [CrossRef]

LeSage, James, and Robert Kelley Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton: CRC Press.

MacKinnon, James G. 2013. Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. Edited by Xiaohong Chen and Norman R. Swanson. New York: Springer, pp. 437–461.

MacKinnon, James G., and Halbert White. 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 29: 305–25. [CrossRef]

Malik, Adeel, and Jonathan R.W. Temple. 2009. The geography of output volatility. *Journal of Development Economics* 90: 163–78. [CrossRef]

Mauro, Paolo. 1995. Corruption and growth. *Quarterly Journal of Economics* 110: 681–712. [CrossRef]

Mayer, Thierry, and Soledad Zignago. 2011. Notes on CEPII's distances measures: The GeoDist database. CEPII Working Paper 2011-25, CEPII, Paris, France.

Michalopoulos, Stelios. 2012. The origins of ethnolinguistic diversity. *American Economic Review* 102: 1508–39. [CrossRef] [PubMed]

Moran, P.A.P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10: 243–51.

Moreno, Ramon, and Bharat Trehan. 1997. Location and the growth of nations. *Journal of Economic Growth* 2: 399–418. [CrossRef]

Murray, Michael P. 2006. Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives* 20: 111–32. [CrossRef]

Olsson, Ola. 2005. Geography and institutions: Plausible and implausible linkages. *Journal of Economics* 86: 167–94.

Olsson, Ola, and Douglas A. Hibbs Jr. 2005. Biogeography and long-run economic development. *European Economic Review* 49: 909–38. [CrossRef]

Putterman, Louis, and David N. Weil. 2010. Post-1500 population flows and the long-run determinants of economic growth and inequality. *Quarterly Journal of Economics* 125: 1627–82. [CrossRef] [PubMed]

Qin, Duo. 2015. Resurgence of the endogeneity-backed instrumental variable methods. *Economics: The Open-Access, Open-Assessment E-Journal* 9: 1–35. [CrossRef]

Qin, Duo, Sophie van Huellen, and Qing-Chao Wang. 2016. How credible are shrinking wage elasticities of married women labour supply? *Econometrics* 4: 1. [CrossRef]

Ramsey, J.B. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society. Series B* 31: 350–71.

Rodrik, Dani, Arvind Subramanian, and Francesco Trebbi. 2004. Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth* 9: 131–65. [CrossRef]

Sachs, Jeffrey D. 2003. Institutions don't rule: Direct effects of geography on per capita income. National Bureau of Economic Research Working Paper 9490, National Bureau of Economic Research, Cambridge, MA, USA.

Sachs, Jeffrey D., and Andrew M. Warner. 1995. Economic reform and the process of global integration. *Brookings Papers on Economic Activity* 1995: 1–95. [CrossRef]

Sargan, J.D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415. [CrossRef]

Shea, John. 1997. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics* 79: 348–52. [CrossRef]

Sims, Christopher A. 2010. But economics is not an experimental science. *Journal of Economic Perspectives* 24: 59–68. [CrossRef]

Spanos, Aris. 1990. The simultaneous-equations model revisited: Statistical adequacy and identification. *Journal of Econometrics* 44: 87–105. [CrossRef]

Spanos, Aris. 2006. Econometrics in retrospect and prospect. In *Palgrave Handbook of Econometrics*. Edited by T.C. Mills and K. Patterson. Basingstoke: Palgrave MacMillan, Volume 1, Econometric Theory; pp. 3–58.

Spanos, Aris. 2007. The instrumental variables method revisited: On the nature and choice of optimal instruments. In *The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis*. Edited by G.D.A. Phillips and E. Tzavalis. Cambridge: Cambridge University Press, pp. 34–59.

Spanos, Aris. 2010. Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics* 158: 204–20. [CrossRef]

Spanos, Aris. 2015. Revisiting Haavelmo's structural econometrics: Bridging the gap between theory and data. *Journal of Economic Methodology* 22: 171–96. [CrossRef]

Spanos, Aris. 2017. Mis-specification testing in retrospect. *Journal of Economic Surveys*. [CrossRef]

Spanos, Airs, and James J. Reade. 2016. Heteroskedasticity/autocorrelation consistent standard errors and reliability of inference. Working Paper for the 17th Annual Oxmetrics Conference, George Washington University, Washington, D.C., USA.

Spolaore, Enrico, and Romain Wacziarg. 2009. The diffusion of development. *Quarterly Journal of Economics* 124: 469–529. [CrossRef]

Spolaore, Enrico, and Romain Wacziarg. 2013. How deep are the roots of economic development? *Journal of Economic Literatur* 51: 325–69. [CrossRef]

Spolaore, Enrico, and Romain Wacziarg. 2014. Long-term barriers to economic development. In *Handbook of Economic Growth*. Edited by P. Aghion and S.N. Durlauf. Amsterdam: Elsevier North-Holland, Volume 2A, pp. 121–76.

Stock, James H., and Motohiro Yogo. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Edited by D.W.K. Andrews and J.H. Stock. Cambridge: Cambridge University Press.

Swamy, P.A.V.B., George S. Tavlas, and Stephen G. Hall. 2015. On the interpretation of instrumental variables in the presence of specification errors. *Econometrics* 3: 55–64. [CrossRef]

The Economist. 2006. Winds of change. *The Economist*. November 4th, p. 84. Available online: http://www.economist.com/node/8104169 (accessed on 26 October 2013).

White, Halbert. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–38. [CrossRef]

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge: MIT Press.
Zietz, Joachim. 2001. Heteroskedasticity and neglected parameter heterogeneity. *Oxford Bulletin of Economics and Statistics* 63: 263–73. [CrossRef]