ISSN 1178-2293 (Online)

University of Otago
Economics Discussion Papers
No. 1508

DECEMBER 2015

Evaluating ingenious instruments for fundamental determinants of long-run economic growth and development*

P. Dorian OWEN
Department of Economics, University of Otago, PO Box 56, Dunedin 9054, New Zealand

Address for correspondence:

Dorian Owen
Department of Economics
University of Otago
PO Box 56
Dunedin
NEW ZEALAND
Email: dorian.owen@otago.ac.nz
Telephone: 64 3 479 8655

December 2015

Evaluating ingenious instruments for fundamental determinants of long-run economic growth and development*

P. Dorian OWEN
Department of Economics, University of Otago, PO Box 56, Dunedin 9054, New Zealand

---

ABSTRACT
The empirical literature on the determinants of cross-country differences in long-run development is characterized by the ingenious nature of many of the instruments used. However, scepticism remains about their ability to provide a valid basis for causal inference. This paper examines the extent to which explicit consideration of the statistical adequacy of the underlying reduced form (RF), 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 RFs in influential studies reveals evidence of model misspecification, with parameter non-constancy and spatial dependence of the residuals almost ubiquitous. This feature, surprisingly not previously identified, potentially undermines inferences 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 Codes: O10, O40, C36

* Earlier versions of this paper were presented at the New Zealand Association of Economists Conference, Wellington; Otago Development and Growth Workshop, Dunedin; New Zealand Econometrics Study Group meeting, Hamilton; and seminars in the Department of Economics, University of Reading; the Research Institute for Applied Social Sciences (RIASS), Swansea University; and the Department of Mathematics and Statistics, University of Otago. I am grateful to participants, especially Mark Casson, Lorenzo Ductor, David Fielding, John Gibson, Dave Maré and Peter Phillips, for helpful comments and suggestions. Thanks also to Wilner Jeanty for providing updated versions of his Stata routines `spwmatrix` and `anketest` and Romain Wacziarg for sharing the data used in Spolaore and Wacziarg (2013).

Contact details:

Dorian Owen
Department of Economics
University of Otago
PO Box 56
Dunedin 9054
New Zealand

Email: Dorian.Owen@otago.ac.nz
Tel: (64) 3 479 8655

1. Introduction

Interpreting patterns of causation from growth regressions is fraught with difficulties. By the nature of the process of economic growth, key variables such as income per capita, human capital, physical capital and technology are interrelated and jointly determined.[1] One response is to step back from the evaluation of the effects of the 'proximate' determinants of economic growth, such as technological change and accumulation of physical and human capital, to investigate the 'deeper', more fundamental, determinants of long-term growth and 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 can demand and afford better quality institutions.[2]

Consequently, widespread use of instrumental variables (IV) estimation, more specifically two-stage least squares (2SLS), is a defining feature of the literature examining the fundamental determinants of cross-country differences in long-run 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

[1] North and Thomas (1973, p.2) point out that "[t]he factors we have listed (innovation, economies of scale, education, capital accumulation, etc.) are not causes of growth; they *are* growth" (emphasis in original).

[2] For some critics, the endogeneity of institutions is a fundamental weakness of this exercise. On this basis, they argue "[i]nstitutions are not a deeper cause than the supply of factors or technology" (Przeworski, 2004, p.184).

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".

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, which 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 (e.g., 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.

The aim of this paper is to apply the approach proposed by Spanos (1990, 2006, 2007, 2015) to focus more attention on 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 highlights the statistical 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. He emphasizes the desirability of probing the statistical adequacy of the RF (i.e., whether the probabilistic assumptions are valid for the data) 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 the fundamental determinants literature 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 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 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 cross-country variation in levels of long-run economic development, usually measured by income per capita. 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 as an explanatory variable 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 potential 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 their instruments for institutional quality 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 as a measure of distance from the equator (as Western Europeans were attracted to colonies with climates similar to their home countries), 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 and highly cited 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 (including political and property rights for the bulk of the population), 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 led to 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, however, 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.

Whereas Acemoglu et al. (2001, 2002) focus on the disease environment, 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 (in addition to settler mortality and latitude). 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.

Several of the early empirical studies (Acemoglu et al., 2001; Easterly and Levine, 2003; and Rodrik et al., 2004) conclude that geographic conditions affect development purely via their effect on institutions; after controlling for institutional quality, geography appears to have little direct effect on income. In response, Sachs (2003) shows that a measure of malaria transmission is statistically significant when added to representative specifications from these three studies, implying that geographical variables have a direct, as well as an indirect, effect, on GDP per capita.[3] 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 (temperature, mosquito abundance and vector specificity) 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

[3] Other studies (Olsson and Hibbs, 2005; Carstensen and Gundlach, 2006) also challenge the characterization of geographical effects as entirely indirect, typically providing evidence for additional direct effects of aspects of geography on income levels.

latitude and whether a country is landlocked) are also frequently included as control variables in fundamental determinants regressions, especially when checking robustness (e.g., Easterly and Levine, 2003, 2013). 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 (in particular, the average and standard deviation of east-west wind speeds) 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 the problem of selection for different degrees of colonial control. Iyer, therefore, uses the death of a ruler without a natural heir as an instrument for direct rule. She finds that 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 in the relevant period 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-gather 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 comparative levels of development.

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 review also highlights how justification for the various instrumentation strategies is based primarily on informal economic theory arguments.

3. IV 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 following generic framework:

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

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.[4] Subscript i denotes observations for country i.[5] $\boldsymbol{X}_i$ can be decomposed as $(\boldsymbol{X}_{1i}' \ \boldsymbol{X}_{2i}')'$ 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. In terms of 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) $\mathrm{E}(\boldsymbol{X}_{1i}\boldsymbol{Z}_i') = \boldsymbol{\Sigma}_{\boldsymbol{XZ}} \neq 0$; and (c) $\mathrm{E}(\boldsymbol{Z}_i\boldsymbol{Z}_i') = \boldsymbol{\Sigma}_{\boldsymbol{ZZ}} > 0$. Implicitly, it is also assumed, if $\boldsymbol{\alpha}_1 \neq 0$, that (d) $\mathrm{E}(\boldsymbol{Z}_i y_i) \neq 0$ (Spanos, 2007, p.38).[6]

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. Hence, exclusion restrictions are based on economic theoretical considerations, whether formal or more informal (Acemoglu, 2005). 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 (e.g., 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 their identification strategies rely primarily on the plausibility of their a priori theoretical arguments.

---

[4] 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).

[5] 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 a cross-sectional analysis.

[6] To simplify the notation, observed variables are assumed to have zero means. These 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)′: $\mathrm{plim}(N^{-1}\boldsymbol{X}_1'\boldsymbol{Z}) = \boldsymbol{\Sigma}_{\boldsymbol{XZ}} \neq 0$; (c)′: $\mathrm{plim}(N^{-1}\boldsymbol{Z}'\boldsymbol{Z}) = \boldsymbol{\Sigma}_{\boldsymbol{ZZ}} > 0$, 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 $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_N)'$ (Spanos, 2007, pp.37-38).

Statistical considerations are not entirely ignored. If the equation of interest is overidentified, i.e., there are more additional instruments than endogenous explanatory variables ($p > m_1$), then testing for overidentifying restrictions is feasible and 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.134-137). They therefore rely on the *untestable* validity of sufficient of the instruments to obtain at least exact identification; by themselves, 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 recent 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 (Stock et al., 2002; Andrews and Stock, 2007).[7] However, as Spanos (2007) emphasizes, weak instrumentation is only one of several potential deviations from the underpinning assumptions of IV estimation; other more basic statistical aspects are largely ignored.

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):

[7] Consequently, tests of instrument relevance have been proposed (Stock and Yogo, 2005) and inference methods robust to weak instrumentation have been developed (Moreira, 2003; Kleibergen, 2007).

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

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

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

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.[8] 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 framework within which the structural model is embedded. A key insight of Spanos's 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):

$$y_i = \alpha_0' \boldsymbol{X}_i + \gamma_0' \boldsymbol{Z}_i + \varepsilon_{0i} \tag{3a}$$

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

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

Spanos proves that imposing the (non-testable) identification restriction $\gamma_0 = 0$, in conjunction with $\boldsymbol{B}_1 \neq 0$ and $\boldsymbol{\beta}_1 \neq 0$, 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 $\mathrm{E}(\mathrm{Z}_i\varepsilon_i) = 0$, and with conditions (b)-(d) holding (Spanos, 2007, pp.42-45).[9] Hence, although $\mathrm{E}(\mathrm{Z}_i\varepsilon_i) = 0$ is not testable, by embedding the structural equation in (1) in the MLR/RF in equation (2), the conditions $\mathrm{E}(\boldsymbol{X}_{1i}\varepsilon_i) \neq 0$ and $\mathrm{E}(\mathrm{Z}_i\varepsilon_i) = 0$ are 'operationalized' via the reparameterization/ restriction on the MLR/RF; moreover, the derived assumptions in the MLR/RF are testable.

[8] Spanos (2007) refers to the reduced form in equation (2) as the 'implicit reduced form', in contrast with an explicit RF arising in a more fully specified simultaneous system; here we adopt the more common usage.

[9] 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).

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).[10] In contrast, misspecification of the MLR/RF model will potentially invalidate IV-based inference. Consequently, whether inference using an IV strategy is reliable depends on whether the assumptions underlying the MLR/RF, including distributional assumptions in equation (2c), are valid for the observed data being analysed.[11] Inference based on conventional formulae requires normality of the error terms, correct functional form, homoskedasticity, parameter constancy (across the cross-sectional units) and error independence (cross-sectional independence in the case of cross-country data) (Spanos, 2007, Table 2.2). Assessment of statistical adequacy of the MLR/RF requires testing these assumptions.[12] If the MLR/RF is misspecified, this suggests a need to respecify the model, with any additional (exogenous) variables added to ensure statistical adequacy becoming part of the extended instrument set.

From this perspective, the statistical adequacy of the RF is an essential prerequisite for 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. This approach is 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 and ignores that the MLR/RF, specified in terms of the joint distribution of the observable variables, provides the framework within which the structural equation is embedded. Although instrument exogeneity is not directly testable, it is reflected in the parameterizations for the structural parameters in the context of the MLR/RF.

---

[10] In the exactly identified case, Spanos (2007, p.55) argues that "the statistical adequacy of the MLR model is sufficient to secure the reliability of inference based on the IV estimators". In the overidentified case, "statistical adequacy of the statistical model is not sufficient"; the overidentifying restrictions also need to be valid.

[11] Hendry and Nielsen (2007, p.220) make the same point: "The reduced-form assumptions are implied by the structural assumptions, so that if the reduced-form assumptions fail, the structural assumptions fail … If, for instance, the normality assumption fails, then the structural normality assumption … would fail."

[12] Other practical recommendations for utilizing information in the RFs are more limited in their scope. Murray (2006) and Angrist and Pischke (2009) suggest checking the signs and statistical significance of reduced-form coefficients, in particular to see if they are at odds with a priori intuition. For the case of a single endogenous explanatory variable in equation (1), Chernozhukov and Hansen (2008) suggest using the equivalent of equation (2a) to conduct valid inference (under the usual assumptions) even if instruments are weak.

Overall, therefore, the bottom line in Spanos's 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 underlying MLR/RF, which explicitly depends on *both* the specification of the structural model and the instrumentation strategy. In most fundamental determinants (and growth) 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. More importantly, testing for misspecification of the RF 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.[13]

[13] One response to concerns about validity of underlying statistical assumptions is the development and application of Generalized Method of Moments (GMM) estimation or non-parametric methods, which require 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, the convention in the

In general, this literature places little emphasis on reporting evidence on statistical adequacy. For example, although over 200 regression models are fitted in the studies by Acemoglu et al. (2001), Easterly and Levine (2003) and Rodrik et al. (2004), the only diagnostic test reported is a test for overidentifying restrictions and the null is rejected for very few of the different model/instrument combinations considered. Although it has a role to play in helping to assess instrument validity, it seems uncontroversial to argue that it is asking too much of this one test to discriminate between different models. Instead, the response to model uncertainty in these studies is to conduct a robustness/sensitivity analysis by adding control variables, singly or in sets, to regressions that include the key explanatory variable(s) of interest. Without explicit misspecification testing, however, there is no guarantee that all, or indeed any, of these models are statistically adequate.

In a cross-sectional context, the statistical assumptions underlying the MLR/RF in equation (2), on which the standard formulae for the sampling distribution of the 2SLS estimator depend (i.e., normality, homoskedasticity and independence of the error terms, correct functional form, and parameter constancy) can be tested for each of the replicated studies. Normality of the errors is relevant given the typical sample sizes in this literature (ranging from $N = 21$ to less than 100 in the regressions examined), precluding appeal to the Central Limit Theorem. Doornik and Hansen's (2008) test for normality (denoted *Norm*) is reported. This is based on the skewness and kurtosis of the OLS residuals and is approximately $\chi^2(2)$ distributed under the null of normal errors. For ease of evaluating test results, the tables report *p*-values for all the diagnostic tests, with *p*-values less than 0.05 in bold.

Heteroskedasticity is widely regarded as a natural feature of cross-sectional data and use of heteroskedastic-consistent standard errors is common (without reporting tests for heteroskedasticity or consideration of whether heteroskedastic-consistent and conventional standard errors differ). However, such standard-error corrections are valid only asymptotically and their finite-sample properties can be unsatisfactory; given the small sample sizes in most of the studies, this is therefore a concern. More importantly, *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.133-134; Sims, 2010;

fundamental determinants literature, including all the studies examined, is to rely on 2SLS estimation, applied to relatively small samples, to fit simple linear-in-parameters models with additive errors and constant parameters across countries.

King and Roberts, 2015).[14] Widespread use of standard-error corrections has tended to lead to this being ignored.

Two versions of White's (1980) test for heteroskedasticity are reported. The first test statistic (*Hetero*) is based on an auxiliary regression of the squared residuals on a constant, the original regressors and their squares; the second statistic (*HeteroX*) also includes cross-products of the regressors and is reported only if there are sufficient observations. Both test statistics are distributed as finite-sample-adjusted F approximations to asymptotic χ^2 distributions under the null of unconditional homoskedasticity.

Fundamental determinants studies typically specify the logarithm of the development proxy as the dependent variable, with explanatory variables entered linearly, in logs or, occasionally, as quadratics. Testing for functional form misspecification (i.e., neglected nonlinearities) is implemented using a RESET test that include squares and cubes of the fitted values from the original regression as additional regressors; this is denoted *RESET* and is approximately F-distributed under the null that the coefficients on these additional regressors are zero. Functional form misspecification may also be reflected in rejection of the normality and homoskedasticity tests and apparent parameter non-constancy.

Given the MLR nature of the RFs, system misspecification tests, multivariate equivalents of the single-equation tests, are also reported. The vector normality test, denoted *NormVec*, is distributed as $\chi^2(2M)$ under the null of normality, where M (= m_1 + 1) is the number of equations in the MLR. Vector heteroskedasticity tests involve auxiliary multivariate regressions of all residual variances and covariances on the original regressors and their squares (and, where relevant and feasible, their cross-products). These are denoted *HeteroVec* and *HeteroXVec* respectively and are asymptotically distributed as χ^2 ($sM(M+1)/2$) under the null of homoskedasticity, where s is the number of non-redundant regressors in the auxiliary regression, but p-values are reported for F-adjusted finite-sample approximations. The vector RESET test, denoted *RESETVec*, is approximately F-distributed.[15]

With cross-country data, lack of independence of the errors is likely to manifest itself as spatial dependence, where 'spatial' may be interpreted broadly to involve socio-economic as

[14] In dynamic time-series models, a common factor test can help statistically distinguish between residual autocorrelation due to autocorrelated errors and residual autocorrelation due to model misspecification (Hendry, 1995, Ch. 7). Unfortunately, there is no analogue for heteroskedastic residuals.

[15] Further details of the tests, implemented using OxMetrics 7, are given in Doornik and Hendry (2013, Ch. 11).

well as geographical distance. Surprisingly, relatively few studies (e.g., Conley and Ligon, 2002) have explored spatial dependence in economic growth and development arising from cross-country spillovers in the growth process. To test for spatial dependence, *p*-values for Moran's *I* statistic (Moran, 1950) and a Lagrange Multiplier (LM) test applied to the residuals of the fitted RFs are reported. The standardized Moran's *I* statistic is asymptotically normally distributed under the null of no spatial autocorrelation but has reasonable small-sample properties (Anselin and Florax, 1995). The LM test reported, denoted $LM_{\rho\lambda}$, is asymptotically $\chi^2(2)$ distributed under the null of absence of both spatial error and spatial lag dependence, and has good finite-sample properties (Anselin et al., 1996).[16] These tests require specification of an a priori weights matrix based on plausible assumptions about the extent of potential spatial linkages. The results obtained depend on this choice, although if the errors are spatially independent then this property should hold for any reasonable choice of weights matrix. 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.[17] 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 Le Gallo, 2008).

The parameters in B_1, B_2, α, and hence β_1 and β_2, are usually assumed to be invariant to *i*. Parameter constancy is explored by recursive graphical analysis of coefficient estimates (with ±2 standard errors bands plotted as dotted lines in the figures) for each of the variables in the RF and of break-point Chow tests calculated at different points in the sample (e.g., Hendry and Nielsen, 2007, pp.195-197).[18] Whereas the normality, heteroskedasticity, RESET and spatial dependence tests are independent of the ordering of the data, different orderings will

[16] $LM_{\rho\lambda}$ has lower power compared to the appropriate one-directional test if only one type of spatial dependence is present (Anselin, 2006) but results are reported for the two-directional joint test given the absence of a clear prior indication of the form of any potential spatial dependence.

[17] 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 endogenous 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`, both Stata routines written by Wilner Jeanty.

[18] 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.

affect the recursive plots and Chow tests. Parameter constancy should imply lack of statistical significance (apart from Type I errors) for Chow tests for all possible orderings.[19] The recursive plots for the coefficient estimates and the Chow tests are based on the observations ordered by the size of the development proxy, log of 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 model misspecification, e.g., omitted variables.

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 the degree of dependence between the tests is unknown, the Bonferroni inequality implies the probability of rejecting one or more of the valid 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 $\alpha = 0.01$. The distributional assumptions and reported *p*-values are valid only if the model is correctly specified, so that rejections, especially rejections for more than one test for the same model, do not provide a clear guide to the direction of required respecification (Hendry and Nielsen, 2007, p.135). The diagnostics are therefore interpreted holistically as an overall check of statistical adequacy.

Tests for overidentification and weak instruments are also reported, although their validity is conditional on the statistical adequacy of the RFs. *Sargan-p* is the *p*-value for Sargan's (1958) test of overidentifying restrictions. Under the null that the instruments are independent of the error term in equation (1), the Sargan test is asymptotically distributed as $\chi^2(q)$, where q is the number of overidentifying restrictions.[20] *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; entries in bold correspond to significant values based on a maximal size of 15%. Also reported are the partial R^2s between the endogenous regressors and the additional

[19] 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.

[20] 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.

instruments, and, where relevant, Shea (1997) partial R^2s, which take into account intercorrelations between the instruments and tend to be notably smaller than the former if instruments are weak.

5. Results

The criteria for selecting studies for replication and examination of RFs are influence, representativeness and ready availability of the relevant data (from authors' and journals' websites). On the basis of these criteria, the studies examined include: 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 (Spolaore and Wacziarg, 2009; Putterman and Weil, 2010; Easterly and Levine, 2013), reported by Spolaore and Wacziarg (2013) in their review article, are also replicated.

Hall and Jones (1999), in their main model explaining ln(*Y*/*L*), the natural logarithm of output per worker, include a measure of 'social infrastructure' (*SocInf*) as the sole explanatory variable in *X*. 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) (*lnFR*) as instruments for *SocInf*. Results of diagnostic testing of the RFs are reported in Table 1, columns (1) and (2), for a representative model (Hall and Jones, 1999, Table II, row 3) fitted to a complete data set for 79 countries (which avoids the need to impute data). 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, which, as later results demonstrate, is a common feature of the studies examined.

Columns (3) and (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.[21] 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 for the RFs. 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 2 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 in the models. Again, there is 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 2(a) (for the RF in Table 2, column (1), based on the model in Acemoglu et al., 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.

[21] 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).

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 (calculated as the average of six World Bank Governance Indicators and labelled *Inst* in Table 3) 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 are reported in Table 3. There is evidence of heteroskedasticity and functional form misspecification in the RFs. The model in Easterly and Levine's (2003) Table 5, row 4 performs best on these tests (with only the multivariate RESET test having a *p*-value less than 0.05). 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).[22]

Diagnostic tests for the RFs of two representative models from Sachs (2003), which add an index of malarial ecology (*ME*) as an instrument to address the endogeneity of malarial risk, are reported in Table 4. 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*). The recursive estimates, as 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 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 population density (*pd*) for different years (1, 1000 and 1500). The

[22] In the RF for institutions, *YrsOpen* is the only coefficient to be statistically significant; Easterly and Levine treat this variable as exogenous, whereas in several other studies (e.g. Rodrik et al., 2004) it is assumed to be endogenous and instrumented.

explanatory variables are the (logarithm of the) number of years since the Neolithic transition (*yst*) and a common set of geographical controls (land productivity, absolute latitude, mean distance to the nearest coast or river, the percentage of land within 100 km of the coast or river, and continent dummies for Africa, Europe and Asia). Although they point out that reverse causality from population density to *yst* is not a problem, "the OLS estimates of the effect of the time elapsed since the transition to agriculture may suffer from omitted variable bias, reflecting spurious correlations with the outcome variable being examined" (p.2106). To address endogeneity, they use the numbers of prehistoric domesticable species of wild plants and animals from Olsson and Hibbs (2005) to instrument *yst*, arguing that their only effect on later population density is via their effect on the timing of the Neolithic transition.[23]

Diagnostic tests corresponding to Ashraf and Galor's IV regressions are reported in Table 5. As well as population density in different years, they also explore the effects of *yst* on subsequent technological sophistication, represented in column (9) and (11) by *natech*1*K* and *natech*1, respectively, a non-agricultural technology index in years 1000 and 1. Spatial dependence of the residuals is evident for all 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 (including both agricultural and non-agricultural technology) is examined, using prehistoric availability of domesticable plants and animal species as instruments, given the latter's role in determining the timing of the Neolithic transition (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.[24]

Other recent studies that focus on historical or intergenerational factors, such as Chanda and Putterman (2007), Spolaore and Wacziarg (2009), Putterman and Weil (2010) and

[23] 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 perhaps surprising given the emphasis on potential omitted variables as a source of endogeneity for *yst*; omitted variables may also be correlated with the geographical controls, which would potentially bias OLS estimates for *all* the coefficients.

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

Easterly and Levine (2013) are also less concerned with reverse causation and place more emphasis on reporting OLS estimates of equation (1).[25] If $E(\boldsymbol{X}_i\varepsilon_i) = 0$, then direct examination of statistical adequacy of the single-equation OLS estimates would be appropriate. From this perspective, Table 6 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 (2013), the models relating to columns (1) and (2) include ancestry-adjusted years of agriculture and ancestry-adjusted state history respectively, whereas columns (3), (4) and (5) include the share of descendants of Europeans. Following Spolaore and Wacziarg (2009), the models relating to columns (6), (7) and (8) include genetic distance, as a proxy for a wide range of intergenerationally transmitted characteristics. Although the normality, heteroskedasticity and RESET tests give less cause for concern, there is consistent evidence of spatial dependence and apparent parameter non-constancy (although less dramatic than in some of the earlier studies considered).

Ashraf and Galor (2013) regress the logarithm of population density in 1500 (ln*pd*1500 in Table 7), as a proxy for historical productivity, on observed genetic diversity, while controlling for the timing of the Neolithic transition (*yst*), the percentage of arable land (*arable*), absolute latitude (*AbsLat*), land suitability for agriculture (*agsuit*) and continent fixed effects. The 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.

Diagnostic tests corresponding to estimates in Ashraf and Galor's Table 2, columns (5) and (6) are reported in Table 7. Because of the small sample size, the *HeteroX* tests cannot be calculated. However, the other diagnostics reveal relatively few problems; apart from

[25] 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.

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 the RFs from any of the other studies considered, and this is the only study considered for which there is little evidence of spatial dependence of the residuals. Although the small sample results in relatively wide confidence bands, most coefficients are statistically significant over the full range of recursive samples, as illustrated in the plots for the RF for diversity (for Ashraf and Galor's Table 2, column 5) 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 and a need to amend the original models. Even if we discount 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 some evidence of non-normality or 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 (as with institutions) 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 difficult to come up with 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. As these variables are not usually orthogonal, omitted variables bias is a potentially serious problem.

Spatial dependence appears to be an almost universal feature of the residuals from the fitted models. Given the cross-country nature of the data, this is perhaps not a surprise, but it is a feature of the statistical models that has been almost entirely neglected. The only exception is a robustness analysis in the online appendix for Ashraf and Galor's (2013) baseline sample in which a correction for spatial autocorrelation is applied to the standard errors. None of the studies attempts 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, such as equation (1), 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. 152-158). However, 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). Alternatively, 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, 2014). 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 attempting to address potential endogeneity problems, several studies have proposed ingenious instruments. As emphasized by Acemoglu (2005) and Imbens (2010), economic theory (regardless of its degree of formalism) underpins the specification of the models, including the choice of relevant explanatory variables and the 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 RFs, specified in terms of the observable variables, provides an embedding framework for the structural equations of interest. Failure of the statistical assumptions underlying the RFs implies failure of the corresponding structural-equation assumptions.

While it is doubtful that any single generic method can provide cast-iron evidence of causality (Basu 2013), 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 in the literature on the fundamental determinants of development exhibit varying degrees of evidence of model misspecification. This feature, surprisingly not previously identified, potentially undermines the inferences drawn about the structural parameters, such as the quantitative and statistical significance 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 relevant time-series data. One possible interpretation of the lack of statistical adequacy for parsimonious models based on relatively narrowly defined sets of explanatory variables and instruments fitted to cross-sectional data is that these models are just too simple. Important factors (multiple fundamental determinants, different dimensions of the various determinants, interactions, dynamics and nonlinearities) may be missing. The more plausible instruments based on quasi-random variation from natural experiments may well be based on sound theoretical arguments, but the statistical adequacy of the empirical models may be undermined by the overly simplistic nature of these models. In addition, 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 reliability 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 explicit modelling of cross-section dependence offer avenues for potential additional insights.

Table 1: Testing statistical adequacy of RFs for Hall and Jones (1999)

	(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

Notes: 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): distance from the equator, fraction of the population speaking one of five major Western European languages, fraction speaking English as their first language, Frankel and Romer's (1999) (natural log of) predicted trade share. See text for explanation of tests; suffix '*p*' denotes *p*-value. $\beta = 0.25$ in the spatial weighting matrix.

Table 2: Testing statistical adequacy of RFs for Acemoglu et al. (2001)

	(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
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

Notes: 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 Table x, Column y of Acemoglu et al. (2001). Instrument sets: Exogenous regressors: absolute latitude (in T4C2, T4C8, T5C6, T5C8, T5C9), continent dummies for Asia, Africa and 'Other' (in T4C8), French legal origin dummy (in T5C6, T5C9), French colonial dummy (in T5C9), religion variables (in T5C7, T5C8, T5C9); Additional instrument: log of European settler mortality (all models, which are exactly identified). See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. $\beta = 0.25$ in the spatial weighting matrix.

Table 3: Testing statistical adequacy of RFs for Easterly and Levine (2003)

	(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	

Notes: 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 Table x, Row y of Easterly and Levine (2003). # corresponds to the model in T4R6 but excluding non-oil crops/minerals dummies in the IV set (not reported in EL). Instrument set in each column: Exogenous regressors: French legal origin dummy, religion dummies (Catholic, Muslim, other) and ethnolinguistic diversity (all models), oil dummy (in T4R6, T4R6#), years open (T5R4); Additional instrument: log of European settler mortality and absolute latitude (all models), landlocked (in T4R4, T4R6, T4R6#), set of 10 crops/minerals dummies (in T4R6). See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. $\beta = 0.2$ in the spatial weighting matrix.

Table 4: Testing statistical adequacy of RFs for Sachs (2003)

	(1)	(2)	(3)	(4)
		T1C10		T1C12
	*lcgdp*95	*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	
Sargan-p	0.404		0.560	
CD-F	**6.371**		11.592	
	Rule	*Mal*94*p*	*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

Notes: Dependent variables: *lcgdp*95 is log of GDP per capita in 1995 (from 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 Table x, Column y of Sachs (2003). Model T1C12 is for the three-equation MLR for *lcgdp*95, *Rule* and *Malfal*. Instrument set in each column (all additional instruments): log of European settler mortality; the share of the population in temperate ecozones; index of malarial ecology based on temperature, mosquito abundance and vector specificity. See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. $\beta = 0.2$ in the spatial weights matrix.

Table 5: Testing statistical adequacy of RFs for Ashraf and Galor (2011)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
	AG(2011)		AG(2011)		AG(2011)		SW(2013)		AG(2011)			AG(2011)		AG(2011)	
	T2C6		T3C6		T4C6		T2C4		T8C3		T8C6	T9C3		T9C6	
	*pd*1500	*yst*	*pd*1000	*yst*	*pd*1	*yst*	*pd*1500	*yst*	*natech*1*K*	*yst*	*natech*1	*pd*1000	*tech*1*K*	*pd*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	

Notes: Dependent variables (all entered in natural logarithms): *pd*1500, *pd*1000 and *pd*1 are, respectively, population density in years 1500, 1000 and 1; *yst* is 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: log of land productivity, log of absolute latitude, mean distance to nearest coast or river, percentage of land within 100 km of coast or river, continent dummies for Africa, Europe and Asia (except for columns (7) and (8), see fn. 23); Additional instruments: number of domesticable species of plants prehistorically native to relevant landmass; corresponding number of domesticable species of animals. TxCy denotes the model in Table x, Column y of the relevant study. See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. $\beta = 0.175$ in the spatial weights matrix (except $\beta = 0.15$ for columns (5) and (6), and (14) and (15)).

Table 6: Testing statistical adequacy of illustrative models from Spolaore and Wacziarg (2013)

	(1) T5C2 *lpci*05	(2) T5C4 *lpci05*	(3) T6C3 *lpci05*	(4) T6C4 *lpci05*	(5) T6C5 *lpci05*	(6) T7C1 *lpci05*	(7) T7C2 *lpci05*	(8) T7C3 *lpci05*
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

Notes: Dependent variable in all models is the logarithm of per capita income in 2005 (*lpci*05). All OLS regressions include a common set of control variables: absolute latitude, percentage of land area in the tropics, landlocked dummy, island dummy. Additional exogenous regressors for each column are: (1) ancestry-adjusted years of agriculture; (2) ancestry-adjusted state history; (3) share of dependants of Europeans, ancestry-adjusted years of agriculture; (4) share of dependants of Europeans, ancestry-adjusted state history; (5) share of dependants of Europeans, F_{ST} weighted genetic distance to the US (current); (6) F_{ST} genetic distance to the US (1500 match); (7) F_{ST} weighted genetic distance to the US (current); (8) F_{ST} weighted genetic distance to the US (current), share of dependants of Europeans. TxCy denotes the model in Table x, Column y of Spolaore and Wacziarg (2013). See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. $\beta = 0.25$ in the spatial weights matrix.

Table 7: Testing statistical adequacy of RFs for Ashraf and Galor (2013)

	(1)	(2)	(3)	(4)	(5)	(6)
		T2C5			T2C6	
	ln*pd*1500	*Div*	*DivSq*	ln*pd*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

Notes: Dependent variables: ln*pd*1500 is the natural log of population density in 1500; *Div* is (observed) genetic diversity and *DivSq* is its square. TxCy denotes the model in Table x, Column y of Ashraf and Galor (2013). Instrument sets: Exogenous regressors: log of Neolithic transition timing; log percentage of arable land; log absolute latitude; log land suitability for agriculture (in all models); continent dummies (Africa, Europe, Americas) in T2C6; Additional instruments (in all models): migratory distance from East Africa (*mdistAddis*); predicted genetic diversity squared (based on regression of genetic diversity on migratory distance and all second-stage control variables) (*divhatsq*). See text for explanation of diagnostic tests; suffix '*p*' denotes *p*-value. NF = not feasible due to small sample. $\beta = 0.1$ in spatial weighting matrix.

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

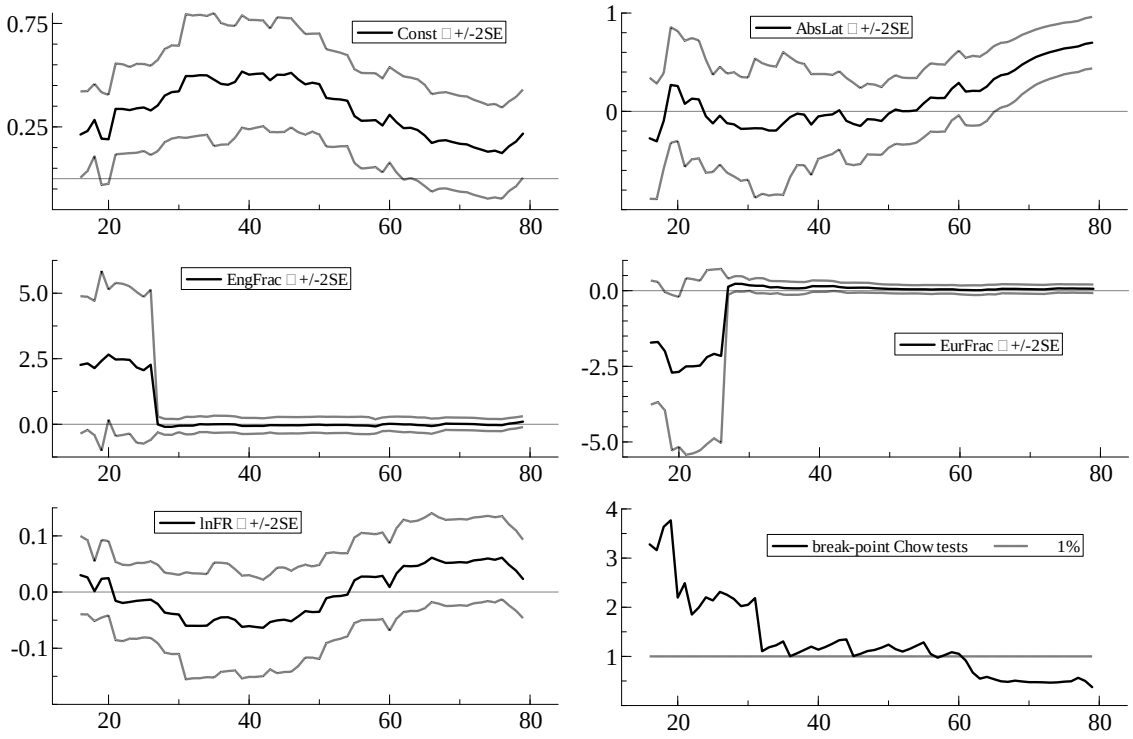
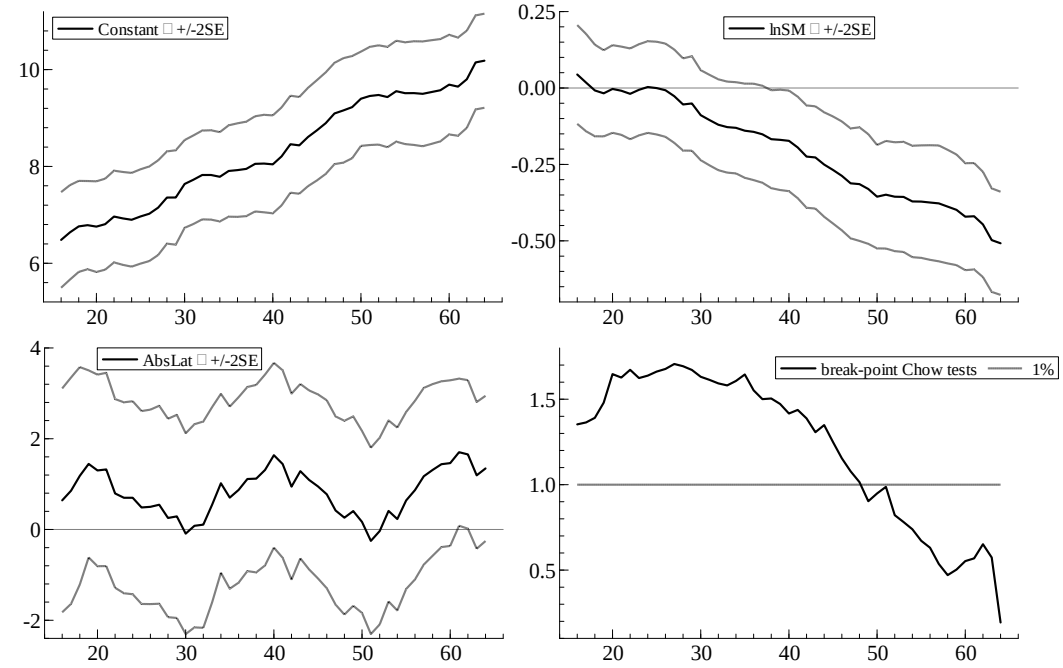


Fig. 2. Recursive coefficient estimates and break-point Chow tests for RFs for Acemoglu et al. (2001, Table 4, column 2)

(a) RF for ln*GDPpc*



(b) RF for *AvExpr*

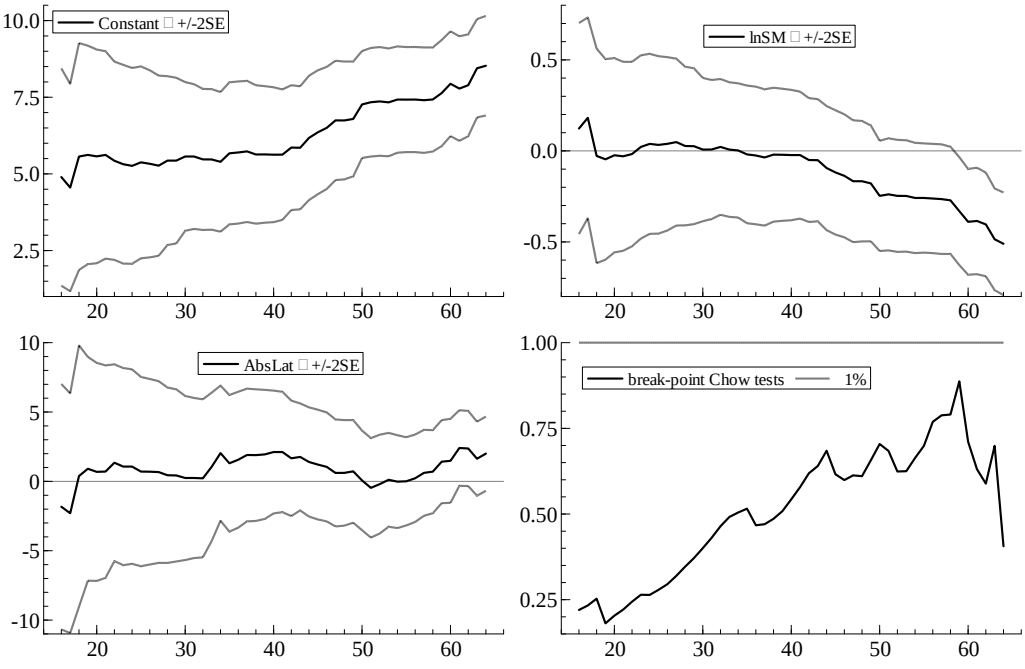
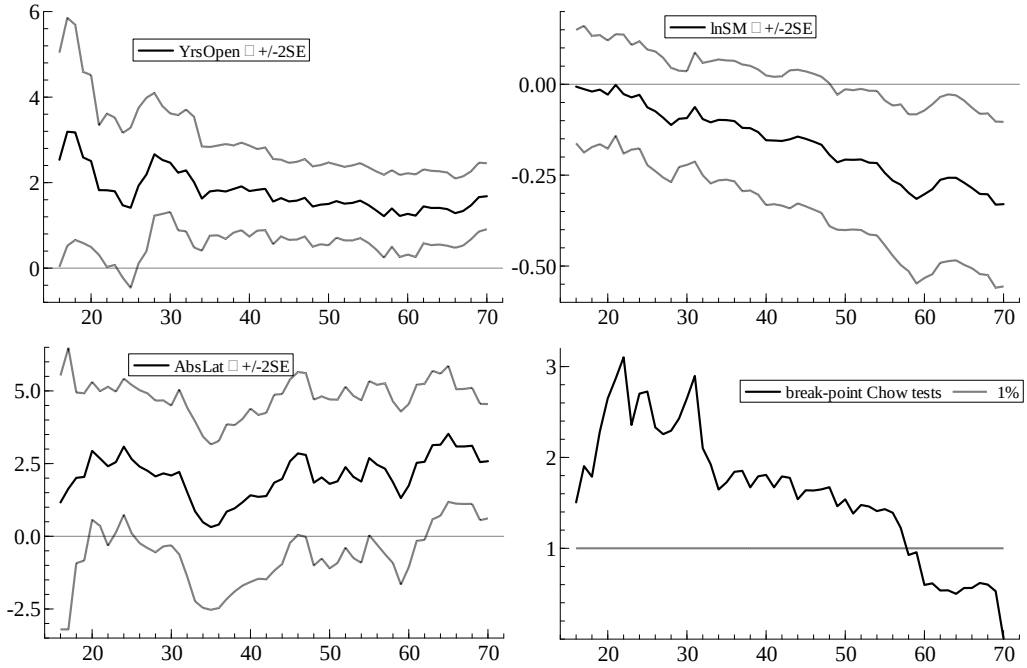


Fig. 3. Recursive estimates for selected coefficients and break-point Chow tests for RFs for Easterly and Levine (2003, Table 5, row 4)

(a) RF for ln*GDPpc*



(b) RF for *Inst*

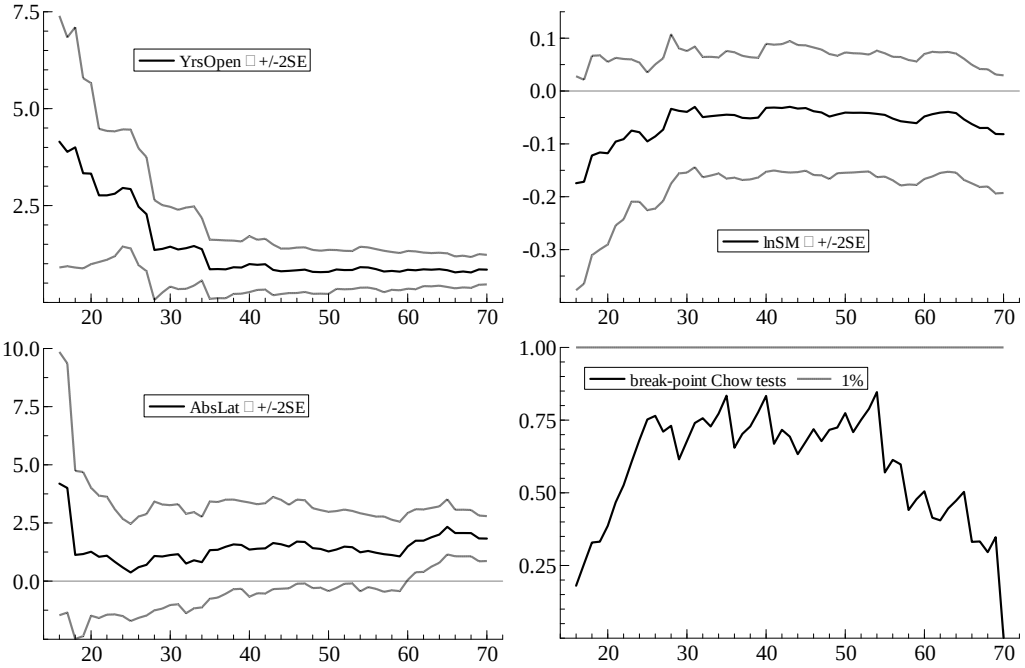
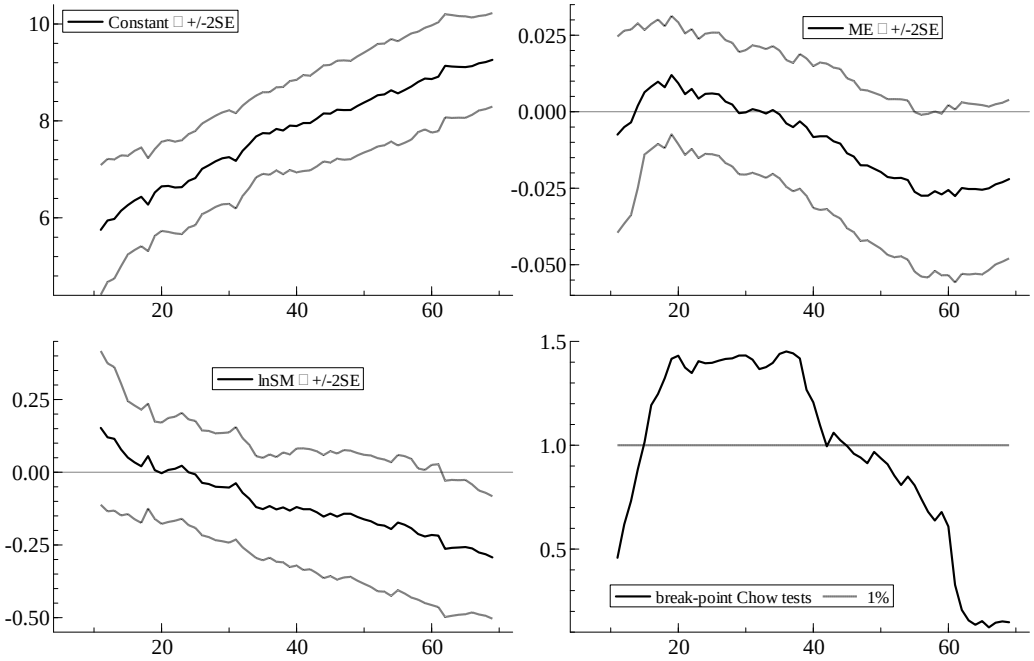


Fig. 4. Recursive estimates for selected coefficients and break-point Chow tests for RFs for Sachs (2003, Table 1, column 10)

(a) RF for *lcgdp95*



(a) RF for *Malfal*

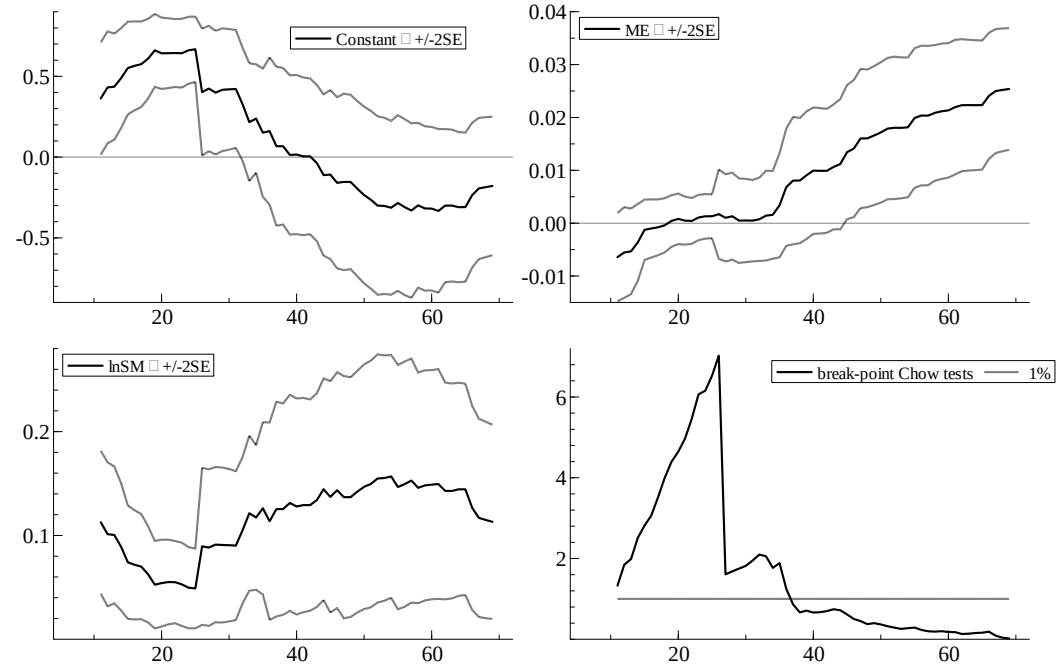
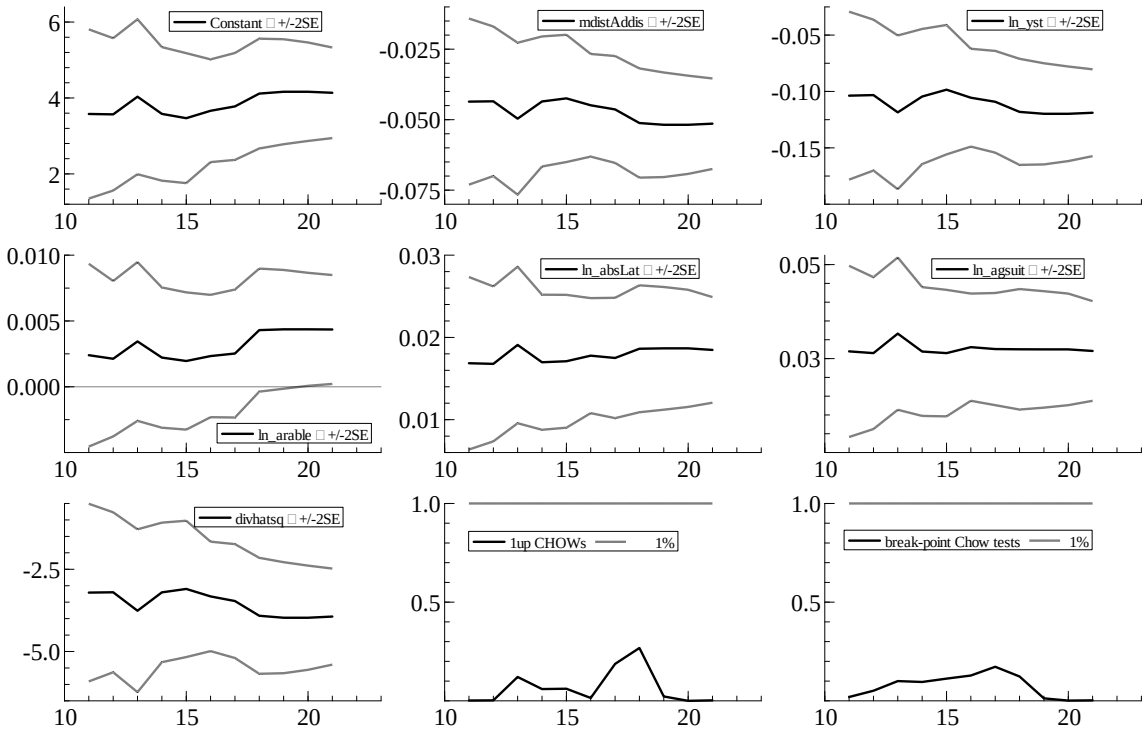


Fig. 5. Recursive coefficient estimates and break-point Chow tests for RF for Ashraf and Galor (2013, Table 2, column 5) for Diversity (*Div*)



References

Acemoglu, D., 2005. Constitutions, politics, and economics: a review essay on Persson and Tabellini's The Economic Effects of Constitutions. Journal of Economic Literature 43(4), 1025-1048.

Acemoglu, D., Johnson, S., Robinson, J.A., 2001. The colonial origins of comparative development: an empirical investigation. American Economic Review 91(5), 1369-1401.

Acemoglu, D., Johnson, S., Robinson, J.A. 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Quarterly Journal of Economics 117(4), 1231–1294.

Acemoglu, D., Johnson, S., Robinson, J.A., 2005. Institutions as a fundamental cause of long-run growth. In Aghion, P., Durlauf, S. (Eds), Handbook of Economic Growth, Volume 1A. Elsevier North-Holland, Amsterdam, pp. 385-472.

Andrews, D.W.K., Stock, J.H., 2007. Inference with weak instruments. In Blundell, R., Newey, W.K., Persson, T. (Eds), Advances in Economics and Econometrics, Theory and Applications (9th Congress of the Econometric Society), Vol. 3. Cambridge University Press, Cambridge, pp. 122-173.

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Angrist, J.D., Pischke, J.-S., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. Journal of Economic Perspectives 24(2), 3-30.

Anselin, L., 2006. Spatial econometrics. In Mills, T.C., Patterson, K. (Eds), Palgrave Handbook of Econometrics, Volume 1: Econometric Theory. Palgrave MacMillan, Basingstoke, pp. 3-58.

Anselin, L., Bera, J., Florax, R., Yoon, M., 1996. Simple diagnostic tests for spatial dependence. Regional Science and Urban Economics 26(1), 77–104.

Anselin, L., Florax, R.J.G.M., 1995. Small sample properties of tests for spatial dependence in regression models: Some further results. In Anselin, L., Florax, R.J.G.M. (Eds), New Directions in Spatial Econometrics. Springer, Berlin, pp. 21–74.

Ashraf, Q., Galor, O., 2011. Dynamics and stagnation in the Malthusian epoch. American Economic Review 101(5), 2003–2041.

Ashraf, Q., Galor, O., 2013. The “Out of Africa” hypothesis, human genetic diversity, and comparative economic development. American Economic Review 103(1), 1-46.

Basu, K., 2013. The method of randomization, economic policy, and reasoned intuition. World Bank Policy Research Working Paper 6722.

Bazzi, S, Clemens, M.A., 2013. Blunt instruments: avoiding common pitfalls in identifying the causes of economic growth. American Economic Journal: Macroeconomics 5(2), 152-186.

Bloom, D.E., Sachs, J. D., 1998. Geography, demography, and economic growth in Africa. Brookings Papers on Economic Activity, Issue 2, 207–273.

Bockstette, V., Chanda, A., Putterman, L. 2002. States and markets: the advantage of an early start. Journal of Economic Growth 7(4), 347-369.

Brock, W.A., Durlauf, S.N., 2001. Growth empirics and reality. World Bank Economic Review 15(2), 229-272.

Carstensen, K., Gundlach, E., 2006. The primacy of institutions reconsidered: direct income effects of malaria prevalence. World Bank Economic Review 20(3), 309-339.

Chanda, A., Putterman, L., 2007. Early starts, reversals and catch-up in the process of economic development. Scandinavian Journal of Economics 109(2), 387–413.

Chernozhukov, V., Hansen, C., 2008. The reduced form: a simple approach to inference with weak instruments. Economics Letters 100(1), 68-71.

Conley, T., Ligon, E., 2002. Economic distance and cross-country spillovers. Journal of Economic Growth 7(2), 157-187.

Cragg, J.G., Donald, S.G., 1993. Testing identifiability and specification in instrumental variable models. Econometric Theory 9(2), 222-240.

Deaton, A., 2010. Instruments, randomization, and learning about development. Journal of Economic Literature 48(2), 424-455.

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

Doornik, J.A., Hansen, H., 2008. An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics 70(s1), 927-939.

Doornik, J.A., Hendry, D.F., 2013. Modelling Dynamic Systems, PcGive 14: Volume II, Timberlake Consultants Ltd, London.

Durlauf, S.N., Johnson, P.A., Temple, J.R.W., 2005. Growth econometrics. In Aghion, P., Durlauf, S. (Eds), Handbook of Economic Growth, Volume 1A. Elsevier North-Holland, Amsterdam, pp. 555-677.

Easterly, W., 2007. Inequality does cause underdevelopment: insights from a new instrument. Journal of Development Economics 84(2), 755-776.

Easterly, W., Levine, R., 2003. Tropics, germs, and crops: how endowments influence economic development. Journal of Monetary Economics 50(1), 3-39.

Easterly, W., Levine, R., 2013. The European origins of economic development. Revised version of NBER Working Paper 18162 (http://williameasterly.org/publications/working-papers/).

Eberhardt, M., Teal, F., 2014. The magnitude of the task: productivity analysis with heterogeneous technology. Mimeograph, University of Nottingham, School of Economics (https://sites.google.com/site/medevecon/publications-and-working-papers).

The Economist, 2006. Winds of change, November 4th, p. 84.

Engerman, S.L., Sokoloff, K.L., 1997. Factor endowments, institutions, and differential paths of growth among New World economies. In Haber, S. (Ed.), How Latin America Fell Behind. Stanford University Press, Stanford, CA, pp. 260-304.

Feyrer, J., Sacerdote, B., 2009. Colonialism and modern income: islands as natural experiments. Review of Economics and Statistics 91(2), 245-262.

Fingleton, B., Le Gallo, J., 2008. Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: finite sample properties. Papers in Regional Science 87(3), 319-339.

Frankel, J.A., 2003. Comments and discussion on: Bosworth, B.P., Collins, S.M., The empirics of growth: An update. Brookings Papers on Economic Activity, 2003, 2, 189-199.

Frankel, J.A., Romer, D., 1999. Does trade cause growth? American Economic Review, 89(3), 379-399.

Fuchs-Schuendeln, N., Hassan, T.A., 2015. Natural experiments in macroeconomics. National Bureau of Economic Research, NBER Working Paper 21228.

Gallup, J.L., Sachs J.D., Mellinger, A.D., 1999. Geography and economic development. International Regional Science Review 22(2), 179-232.

Hall, R.E., Jones, C.I., 1999. Why do some countries produce so much more output per worker than others? Quarterly Journal of Economics 114(1), 83-116.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50(4), 1029-1054.

Heckman, J.J., Urzúa, S., 2010. Comparing IV with structural models: what simple IV can and cannot identify. Journal of Econometrics 156(1), 27-37.

Hendry, D.F., 1995. Dynamic Econometrics. Oxford University Press, Oxford.

Hendry, D.F., Nielsen, B., 2007. Econometric Modeling: A Likelihood Approach. Princeton University Press, Princeton, NJ.

Imbens, G.W., 2010. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). Journal of Economic Literature 48(2), 399-423.

Iyer, L., 2010. Direct versus indirect colonial rule in India: long-term consequences. Review of Economics and Statistics 92(4), 693-713.

King, G., Roberts, M., 2015. How robust standard errors expose methodological problems they do not fix, and what to do about it. Political Analysis, 23(2), 159–179.

Kleibergen F., 2007. Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics. Journal of Econometrics 139(1), 181–216.

Knowles, S., Owen, P.D., 2010. Which institutions are good for your health? The deep determinants of comparative cross-country health status. Journal of Development Studies 46(4), 701-723.

Kraay, A., 2015. Weak instruments in growth regressions: implications for recent cross-country evidence on inequality and growth. World Bank Group, Policy Research Working Paper 7494.

La Porta, R., Lopez-de-Silanes, F. Shleifer, A. Vishny, R. 1999. The quality of government. Journal of Law, Economics and Organization 15(1), 222-279.

Malik, A., Temple, J.R., 2009. The geography of output volatility. Journal of Development Economics 90(2), 163-178.

Mauro, P. 1995. Corruption and growth. Quarterly Journal of Economics 110(3), 681-712.

Mayer, T, Zignago, S., 2011. Notes on CEPII's distances measures: The GeoDist database. CEPII, Working Paper No 2011-25.

Moran, P.A.P., 1950. A test for the serial dependence of residuals. Biometrika 37(1/2), 178–181.

Moreira, M.J. 2003. A conditional likelihood ratio test for structural models. Econometrica 71(4), 1027–1048.

Murray, M.P., 2006. Avoiding invalid instruments and coping with weak instruments. Journal of Economic Perspectives 20(4), 111–132.

North, D.C., Thomas, R.P., 1973. The Rise of the Western World: A New Economic History. Cambridge University Press, Cambridge.

Olsson, O., 2005. Geography and institutions: plausible and implausible linkages. Journal of Economics 10, Supplement, 167-194.

Olsson, O., Hibbs, D.A., Jr., 2005. Biogeography and long-run economic development. European Economic Review 49(4), 909-938.

Przeworski, A., 2004. The last instance: Are institutions the primary cause of economic development? European Journal of Sociology 45(2), 165-188.

Putterman, L., Weil, D.N., 2010. Post-1500 population flows and the long-run determinants of economic growth and inequality. Quarterly Journal of Economics 125(4), 1627-1682.

Rodrik, D., Subramanian, A., Trebbi, F., 2004. Institutions rule: the primacy of institutions over geography and integration in economic development. Journal of Economic Growth 9(2), 131-165.

Sachs, J., 2003. Institutions don't rule: direct effects of geography on per capita income. National Bureau of Economic Research, NBER Working Paper 9490.

Sachs, J.D., Warner, A.M., 1995. Economic reform and the process of global integration. Brookings Papers on Economic Activity, Issue 1, 1-95.

Sargan, J.D., 1958. The estimation of economic relationships using instrumental variables. Econometrica 26(3), 393-415.

Shea, J., 1997. Instrument relevance in multivariate linear models: a simple measure. Review of Economics and Statistics 79(2), 348-352.

Sims, C.A., 2010. But economics is not an experimental science. Journal of Economic Perspectives 24(2), 59-68.

Spanos, A., 1990. The simultaneous-equations model revisited: statistical adequacy and identification. Journal of Econometrics 44(1-2), 87-105.

Spanos, A., 2006. Econometrics in retrospect and prospect. In Mills, T.C., Patterson, K. (Eds), Palgrave Handbook of Econometrics, Volume 1: Econometric Theory. Palgrave MacMillan, Basingstoke, pp. 3-58.

Spanos, A., 2007. The instrumental variables method revisited: on the nature and choice of optimal instruments. In Phillips, G.D.A., Tzavalis, E. (Eds), The Refinement of Econometric Estimation and Test Procedures: Finite Sample and Asymptotic Analysis. Cambridge University Press, Cambridge, pp. 34-59.

Spanos, A., 2015. Revisiting Haavelmo's structural econometrics: bridging the gap between theory and data. Journal of Economic Methodology 22(2), 171-196.

Spolaore, E., Wacziarg, R., 2009. The diffusion of development. Quarterly Journal of Economics 124(2), 469–529.

Spolaore, E., Wacziarg, R., 2013. How deep are the roots of economic development? Journal of Economic Literature 51(2), 325-369.

Spolaore, E., Wacziarg, R., 2014. Long-term barriers to economic development. In Aghion, P., Durlauf, S.N. (Eds), Handbook of Economic Growth, Volume 2A. Elsevier North-Holland, Amsterdam, pp. 121-176.

Stock, J.H., Wright, J.H., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. Journal of Business & Economic Statistics 20(4), 518-529.

Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In Andrews, D.W.K., Stock, J.H. (Eds), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg. Cambridge University Press, Cambridge.

White, H., 1980. A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4), 817-838.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data, Second Edition. MIT Press, Cambridge, MA.

Zietz, J., 2001. Heteroskedasticity and neglected parameter heterogeneity. Oxford Bulletin of Economics and Statistics 63(2), 263-273.